Principles of Data Analysis

Rebecca Thorn

Prasenjit Saha

Published by Cappella Archive (ISBN 1-902918-11-8).

The text can be downloaded free from www.physik.uzh.ch/~psaha/pda/ in A4, letter, or paperback size. This web page also has some brief reviews.

The paperback edition (which is much nicer but not much more expensive than laser printing) is available directly from the publisher cappella-archive.com.

To everyone who has shared with me, in different times and places, the fun of the ideas herein.

You know who you are!

Contents

Preface

People who know me are probably asking: What is a theorist doing writing a book about data analysis?

Well, it started as a course for final-year maths and physics undergraduates, which I taught in 1996. Since then, some kind colleagues and friends have told me they found the course notes useful, and suggested I tidy them up into a short book. So here it is. It is also an affordable book: it is available free via the web, and if you care for a nicely-bound copy, Cappella Archive will sell you one at production cost.

What this book hopes to convey are ways of thinking (= principles) about data analysis problems, and how a small number of ideas are enough for a large number of applications. The material is organized into eight chapters:

1) Basic probability theory, what it is with the Bayesians versus the Frequentists, and a bit about why quantum mechanics is weird (Bell's theorem).
2) Binomial and Poisson distributions, and some toy problems introducing the key ideas of parameter fitting and model comparison.
3) The central limit theorem, and why it makes Gaussians ubiquitous, from counting statistics to share prices.
4) An interlude on Monte-Carlo algorithms.
5) Least squares, and related things like the χ^2 test and error propagation. [Including the old problem of fitting a straight line amid errors in both x and y.]
6) Distribution function fitting and comparison, and why the Kolmogorov-Smirnov test and variants of it with even longer names are not really arcane. [Sample problem: invent your own KS-like statistic.]
7) Entropy in information theory and in image reconstruction.
8) Thermodynamics and statistical physics reinterpreted as data analysis problems.

As you see, we are talking about data analysis in its broadest, most general, sense.

Mixed in with the main text (but set in smaller type) are many examples, problems, digressions, and asides. For problems, I've indicated levels of difficulty ([1]: trivial to [4]: seriously hard) but here individual experiences can be very different, so don't take these ratings too seriously. But the problems really are at the heart of the book—data analysis is nothing if isn't about solving problems. In fact, when I started preparing the original course, I first gathered a set of problems and then worked out a syllabus around them. Digressions are long derivations of some key results, while asides are useless but cute points, and both can be skipped without losing continuity.

The mathematical level is moderately advanced. For example, the text assumes the reader is used to matrix notation and Lagrange multipliers, but Fourier transforms are briefly explained in the Appendix. A few of the problems involve writing short programs, in any programming language.

References are in footnotes, but *Probability theory: The Logic of Science* by E.T. Jaynes (in press at Cambridge University Press [see also `bayes.wustl.edu`]) and *Data Analysis: A Bayesian Tutorial* by D.S. Sivia (Oxford University Press 1996) have been so influential that I must cite them here. My book is not a substitute for either of these, more of a supplement: if Sivia makes the basics crystal-clear and Jaynes shows how powerful the underlying ideas are, I have tried to illustrate the enormous range of applications and put them in context.

Although this book is small, the number of people who have distinctly contributed to it is not. James Binney first introduced me to the ideas of E.T. Jaynes, and he and Brian Buck explained to me why they are important. Vincent Macaulay has since continued what they had begun, and indeed been for fifteen years my guide to the literature, sympathetic reviewer, voice of reason. If Vincent had written this book it would have been much better. I remain grateful to Dayal Wickramasinghe for giving me the opportunity to design and teach a course on a random interesting topic,[1] and to the students for going through with it, especially Alexander Austin, Geoffrey Brent, and Paul Cutting, who not only demolished any problems I could concoct, but often found better solutions than I had. Several people posed interesting problems to me, my response to which is included in this book; Ken Freeman and Scott Tremaine posed the two hardest ones. Eric Grunwald, Abigail Kirk, Onuttom Narayan, Inga Schmoldt, Firoza Sutaria, Andrew Usher, and Alan Whiting all suggested improvements on early versions of the manuscript. Rebecca Thorn illustrated the front cover and Nigel Dowrick wrote the back cover. David Byram-Wigfield provided much-appreciated advice on typography and made the bound book happen. Naturally, none of these people is responsible for errors that remain.

I hope you will find this book entertaining and useful, but I have to include a health warning: this book is written by a physicist. I don't mean the fact that most of the examples come from physics and astronomy: I simply picked the examples I could explain best, and many of them have straightforward translations into other areas. The health warning has to do with physical scientists' notions of interestingness. If you have hung out with physicists you can probably spot a physicist even from the way they talk about the stock market. So if you are a statistician or a social scientist and find this book completely wrongheaded and abhorrent, my apologies, it's a cultural thing.

December 2002

[1] At the Australian National University in lovely Canberra, by Lake Burley Griffin amongst the kangaroos, go there if you can.

1. Probability Rules!

What do you think of when someone says 'Data'?

I often imagine a stream of tiny, slightly luminous numbers flowing from a telescope to a spinning disc, sometimes a frenetic man reading a heap of filled questionnaires while batting down any that try to get blown away. Some people see a troupe of chimpanzees observing and nimbly analyzing the responses of their experimental humans. Some only see a pale android with stupendous computing powers.

From such mental images we might abstract the idea that data are information *not yet in the form we want it*, and therefore needing non-trivial processing. Moreover, the information is *incomplete*, through errors or lack of some measurements, so probable reconstructions of the incomplete parts are desired. Schematically, we might view data analysis as

$$\left\{\begin{array}{c}\text{Incomplete}\\ \text{information}\end{array}\right\} \xrightarrow{\text{Probability theory}} \{\text{Inferences}\}\,.$$

This book is an elaboration of the scheme above, interpreted as broadly as possible while still being quantitative, and as we'll see, both data and inferences may take quite surprising forms.

Concrete situations involving data analysis, of which we will discuss many in this book, tend to fall cleanly into one of four groups of problems.

(1) First and most straightforward are the situations where we want to measure something. The measurement process may be very indirect, and involve much theoretical calculation. For example, imagine measuring the height of a mountain-top with a barometer; this would require not only reading off the air pressure, but theoretical knowledge of how air pressure behaves with height and how much uncertainty is introduced by weather-dependent fluctuations. But the relation between the numbers being observed and the numbers being inferred is assumed to be well understood. We will call such situations 'parameter-fitting' problems.

(2) The second group of problems involves deciding between two or more possibilities. An archetype of such problems is "Will it rain tomorrow?" We will take up similar (though much easier) questions as 'model comparison problems'.

(3) The third kind of situation is when we have a measurement, or perhaps we can see a pattern, and we want to know whether to be confident that what we see is really there, or to dismiss it as a mirage produced by noise. We will call these 'goodness-of-fit problems', and they will typically arise together with parameter-fitting problems.

(4) Finally, we will come across some curious situations where the data are few indeed, perhaps even a single number, but we do have some additional theoretical knowledge, and we still want to make what inferences we can.

Through probability theory we can pose, and systematically work out, problems of all these four kinds in a fairly unified way. So we will spend the rest of this chapter

developing the necessary probability theory from first principles and then some formalism for applying it to data analysis.

The formal elements of probability theory are quite simple. We write $\mathfrak{prob}(A)$ for the probability associated with some state A, $\mathfrak{prob}(\overline{A})$ is the probability of *not* A, $\mathfrak{prob}(AB)$ or $\mathfrak{prob}(A, B)$ is the probability of A *and* B, and $\mathfrak{prob}(A \mid B)$ is the so-called conditional probability of A given B.[1] Probabilities are always numbers in $[0, 1]$ and obey two fundamental rules:

$$\begin{aligned} &\mathfrak{prob}(A) + \mathfrak{prob}(\overline{A}) = 1 \quad \text{(sum rule)}, \\ &\mathfrak{prob}(AB) = \mathfrak{prob}(A \mid B)\, \mathfrak{prob}(B) \quad \text{(product rule)}. \end{aligned} \tag{1.1}$$

Note that the product rule is not a simple multiplication, because A and B may depend on each other. (Cats with green eyes or brown eyes are common, but cats with one green eye and one brown eye are rare.)

EXAMPLE [Pick a number...] Any natural number, and I will do likewise. What is the probability that the two numbers are coprime?

Say the two numbers are m, n. We want the probability that m, n do not have a common factor that equals $2, 3$, or any other prime. For conciseness let us write D_p to mean the state "p divides both m and n". Accordingly, $\overline{D}_p$ means "p divides at most one of m, n" (*not* "p divides neither m nor n"). We have

$$\mathfrak{prob}(\overline{D}_p) = \left(1 - p^{-2}\right). \tag{1.2}$$

The probability that m and n are coprime is

$$\mathfrak{prob}(m, n \text{ coprime}) = \mathfrak{prob}(\overline{D}_2, \overline{D}_3, \overline{D}_5, \ldots) \tag{1.3}$$

and using the product rule gives us

$$\mathfrak{prob}(m, n \text{ coprime}) = \mathfrak{prob}(\overline{D}_2)\, \mathfrak{prob}(\overline{D}_3 \mid \overline{D}_2)\, \mathfrak{prob}(\overline{D}_5 \mid \overline{D}_2, \overline{D}_3) \cdots. \tag{1.4}$$

This scary expression fortunately simplifies, because the $\mathfrak{prob}(\overline{D}_p)$ are in fact independent: $\frac{1}{3}$ of natural numbers are multiples of 3, and if you remove all multiples of 2, that's still the case. Hence we have

$$\begin{aligned} &\mathfrak{prob}(m, n \text{ coprime}) = \mathfrak{prob}(\overline{D}_2)\, \mathfrak{prob}(\overline{D}_3)\, \mathfrak{prob}(\overline{D}_5) \cdots \\ &= \left(1 - 2^{-2}\right)\left(1 - 3^{-2}\right)\left(1 - 5^{-2}\right) \cdots \end{aligned} \tag{1.5}$$

The infinite product in the last line equals (see page 72 for more on this formula) $6/\pi^2$. □

PROBLEM 1.1: Achilles and the Tortoise take turns rolling a die (Achilles first) until one of them rolls a six. What is the probability that the Tortoise rolls a six?

Try to get the answer in two different ways: (i) via an infinite series with one term per turn, and (ii) by invoking a symmetry. **[1]**

---

[1] In this book *all* probabilities are really conditional, even if the conditions are tacit. So when we write $\mathfrak{prob}(A)$, what we really mean is $\mathfrak{prob}(A \mid \{\text{tacit}\})$.

From the rules (1.1), two corollaries follow.

The first corollary is called the marginalization rule. Consider a set of possibilities M_k that are "exhaustive" and "mutually exclusive", which is to say one and only one of the M_k must hold. For such M_k

$$\sum_k \mathfrak{prob}(M_k \,|\, A) = 1, \tag{1.6}$$

for any A, and using the product rule and (1.6) gives the marginalization rule

$$\sum_k \mathfrak{prob}(M_k, A) = \mathfrak{prob}(A). \tag{1.7}$$

We will use the marginalization rule many times in this book; usually M_k will represent the possible values of some parameter.

The second corollary follows on writing $\mathfrak{prob}(AB)$ in two different ways and rearranging:

$$\mathfrak{prob}(B \,|\, A) = \frac{\mathfrak{prob}(A \,|\, B)\,\mathfrak{prob}(B)}{\mathfrak{prob}(A)}. \tag{1.8}$$

It is called Bayes' theorem,[1] and we will also use it many times in this book. For the moment, note that $\mathfrak{prob}(A \,|\, B)$ may be very different from $\mathfrak{prob}(B \,|\, A)$. (The probability of rain given clouds in the sky is not equal to the probability of clouds in the sky given rain.)

EXAMPLE [A classic Bayesian puzzle] This puzzle appears in books (and television game shows) in different guises.

You are in a room with three closed doors. One of the doors leads to good stuff, while the other two are bad. You have to pick a door and go to your fate. You pick one, but just as you are about to open it, someone else in the room opens another door and shows you that it is bad. You are now offered the choice of switching to the remaining door. Should you?

Let's say $\mathfrak{prob}(a)$ is the probability that the door you picked is good, and similarly $\mathfrak{prob}(b)$, $\mathfrak{prob}(c)$ for the other doors. And let's say $\mathfrak{prob}(B)$ is the probability that the second door is bad and gets opened. We want

$$\mathfrak{prob}(a \,|\, B) = \frac{\mathfrak{prob}(B \,|\, a)\,\mathfrak{prob}(a)}{\mathfrak{prob}(B)} \tag{1.9}$$

where

$$\begin{aligned} \mathfrak{prob}(B) = {} & \mathfrak{prob}(B \,|\, a)\,\mathfrak{prob}(a) + \mathfrak{prob}(B \,|\, b)\,\mathfrak{prob}(b) \\ & + \mathfrak{prob}(B \,|\, c)\,\mathfrak{prob}(c). \end{aligned} \tag{1.10}$$

We set

$$\mathfrak{prob}(a) = \mathfrak{prob}(b) = \mathfrak{prob}(c) = \tfrac{1}{3} \tag{1.11}$$

since we have no other information. We also have $\mathfrak{prob}(B|b) = 0$ because we have defined the states b and B as mutually exclusive. That leaves $\mathfrak{prob}(B|a)$ and $\mathfrak{prob}(B|c)$.

[1] Equation (1.8) isn't much like what Bayes actually wrote, but Bayes' theorem is what it's called.

Now a subtlety arises. If the other person opened a door at random then $\mathfrak{prob}(B \mid a) = \mathfrak{prob}(B \mid c)$, which gives $\mathfrak{prob}(a \mid B) = \frac{1}{2}$. But when this problem is posed as a puzzle, it is always clear that the other person knows which doors are bad, and is opening a bad door you have not chosen just to tease you. This gives $\mathfrak{prob}(B \mid a) = \frac{1}{2}$ (they could have opened B or C) and $\mathfrak{prob}(B \mid c) = 1$ (they had to open B because c is good). Hence $\mathfrak{prob}(a \mid B) = \frac{1}{3}$, and it is favourable to switch. □

In taking the sum and product rules (1.1) as fundamental, we have left unspecified what a probability actually is. In orthodox statistics, a probability is always a fraction of cases

$$\mathfrak{prob}(A) = \frac{\text{Number of cases where } A \text{ holds}}{\text{Number of possible cases}} \tag{1.12}$$

or a limit of such fractions, where the possible cases are all 'equally likely' because of some symmetry. The sum and product rules readily follow. But for most of the applications in this book a definition like (1.12)—the so-called "Frequentist" definition—is too restrictive. More generally, we can think of probabilities as a formalization of our intuitive notions of plausibility or degree-of-belief, which may or may not be frequency ratios, but which always obey sum and product rules. This interpretation, together with principles for assigning values to those probabilities which are not frequency ratios, dates back to Bayes and Laplace. In more recent times it was developed by Jeffreys,[1] and it is good enough for this book.

Still, the latter definition leaves us wondering whether there might be other formalizations of our intuition that don't always follow the sum and product rules. To address this question we need an axiomatic development of probability that will derive the sum and product rules instead of postulating them, and then we can compare the axioms with our intuition. There are several such axiomatic developments: Cox[2] relates his formulation to our intuition for plausible reasoning, de Finetti[3] relates his development to our intuition for risks and gambling, while Kolmogorov's axioms relate to measure theory. But all three allow probabilities that are not frequency ratios but which nevertheless obey the sum and product rules.

Whichever general definition one follows,[4] the important consequence for data analysis is that non-Frequentist probabilities open up a large area of applications involving Bayes' theorem. Hence the generic name "Bayesian" for probability theory beyond the Frequentist regime.

---

[1] H.S. Jeffreys, *Theory of Probability* (Oxford University Press 1961—first edition 1937).

[2] R.T. Cox, *The Algebra of Probable Inference* (Johns Hopkins Press 1961).

[3] B. de Finetti, *Theory of Probability* (Wiley 1974). Translated by A. Machí & A. Smith from *Teoria della Probabilitá* (Giulio Einaudi editore s.p.a. 1970).

[4] Mathematicians tend to follow Kolmogorov, Bayesian statisticians seem to prefer de Finetti, while physical scientists are most influenced by Cox.

DIGRESSION [Probability as state of knowledge] Cox's development of probability as a measure of our state of knowledge is particularly interesting from the data-analysis point of view. Here is a sketch of it, without proofs of two results now called Cox's theorems.[1]

Imagine a sort of proto-probability $p(A)$ which measures how confident we are (on the basis of available information) that A is true; $p(A) = 0$ corresponds to certainty that A is false and $p(A) = 1$ to certainty that A is true. As yet we have no other quantitative rules; we want to make up some reasonable rules for combining proto-probabilities.

If we have values for the proto-probabilities of A and of B-given-A, we desire a formula that will tell us the proto-probability of A-and-B. Let us denote the desired formula by the function F, that is to say

$$p(AB) = F(p(B\,|\,A), p(A)). \tag{1.13}$$

We require F to be such that if there is more than one way of computing a proto-probability they should give the same result. So if we have $p(A)$, $p(B\,|\,A)$, and $p(C\,|\,AB)$, we could combine the first two to get $p(AB)$ and then combine with the third to get $p(ABC)$, or we could combine the last two to get $p(BC\,|\,A)$ and then combine with the first to get $p(ABC)$, and both ways should give the same answer. In other words we require

$$F(F(x, y), z) = F(x, F(y, z)). \tag{1.14}$$

If we further require $F(x, y)$ to be non-decreasing with respect to both arguments, the general solution for F [though we will not prove it here] satisfies

$$w(F(x, y)) = w(x)w(y). \tag{1.15}$$

where w is an arbitrary monotonic function satisfying $w(0) = 0$ and $w(1) = 1$. Let us define

$$q(A) = w(p(A)). \tag{1.16}$$

Now, $q(A)$ has all the properties of a proto-probability, so we are free to redefine proto-probability as $q(A)$ rather than $p(A)$. If we do this, (1.15) gives

$$q(AB) = q(A)q(B\,|\,A) \tag{1.17}$$

which is the product rule.

If we have a value for $q(A)$ we also desire a formula that will tell us $q(\overline{A})$. Let S be that formula, i.e.,

$$q(\overline{A}) = S(q(A)). \tag{1.18}$$

Then $S(S(x))$ must be x. The general solution for S [though again we will not prove it here] is

$$S(x) = (1 - x^m)^{1/m} \tag{1.19}$$

[1] The following is actually based on Chapter 2 of Jaynes; Cox's own notation is different. Jaynes compares with Kolmogorov and de Finetti in his Appendix A.

where m is a positive constant. Hence we have

$$q^m(A) + q^m(\bar{A}) = 1, \tag{1.20}$$

which is to say $q^m(A)$ satisfies the sum rule. But from (1.17), $q^m(A)$ also satisfies the product rule. So we change definition again, and take $q^m(A)$ as the proto-probability.

The argument sketched here shows that, without loss of generality, we can make proto-probabilities satisfy the sum and product rules. At which point we may rename them as simply probabilities. □

We will now develop some formalism, based on the probability rules, for the four topics or problem-types mentioned at the start of this chapter. The rest of this book will be about developing applications of the formalism. (Though occasionally we will digress, and sometimes some aspect will develop into a topic by itself.) The details of the applications will sometimes be quite complicated, but the basic ideas are all simple.

First we consider parameter fitting. Say we have a model M with some parameters ω, and we want to fit it to some data D. We assume we know enough about the model that, given a parameter value we can calculate the probability of any data set. That is to say, we know $\mathfrak{prob}(D \mid \omega, M)$; it is known as the "likelihood". Using Bayes' theorem we can write

$$\mathfrak{prob}(\omega \mid D, M) = \frac{\mathfrak{prob}(D \mid \omega, M)\, \mathfrak{prob}(\omega \mid M)}{\mathfrak{prob}(D \mid M)}. \tag{1.21}$$

The left hand side, or the probability distribution of parameters given the data, is what we are after. On the right, we have the likelihood and then the strange term $\mathfrak{prob}(\omega \mid M)$, which is the probability distribution of the parameters without considering any data. The denominator, by the marginalization rule, is

$$\mathfrak{prob}(D \mid M) = \sum_{\omega'} \mathfrak{prob}(D \mid \omega', M)\, \mathfrak{prob}(\omega' \mid M), \tag{1.22}$$

and clearly just normalizes the right hand side. The formula (1.21) can now be interpreted as relating the probability distributions of ω before and after taking the data, via the likelihood. Hence the usual names "prior" for $\mathfrak{prob}(\omega \mid M)$ and "posterior" for $\mathfrak{prob}(\omega \mid D, M)$.[1] In applications, although we will always have the full posterior probability distribution, we will rarely display it in full. Usually it is enough to give the posterior's peak or mean or median (for parameter estimates) and its spread (for uncertainties); and since none of these depend on the normalization of the posterior, we can usually discard the denominator in (1.21) and leave the formula as a proportionality.

In equation (1.22), and elsewhere in this chapter, we are assuming that the parameters ω take discrete values; but this is just to simplify notation. In practice both data

[1] The presence of two different probability distributions for ω, with different conditions reminds us that all probabilities in this book are conditional.

and parameters can be continuous rather than discrete, and if necessary we can replace probabilities by probability densities and sums by integrals.

Probability theory tells how to combine probabilities in various ways, but it does not tell us how to assign prior probabilities. For that, we may use symmetry arguments, physical arguments, and occasionally just try to express intuition as numbers. If inferences are dominated by data, the prior will not make much difference anyway. So for the moment, let us keep in mind the simplest case: say ω can take only a finite number of values and we don't have any advance knowledge preferring any value, hence we assign equal prior probability to all allowed values (as we have already done in equation 1.11 in the three-doors puzzle); this is called the principle of indifference.[1]

If ω has a continuous domain, assigning priors gets a little tricky. We can assign a constant prior probability density for $\mathfrak{prob}(\omega)$, but then changing variable from ω to (say) $\ln \omega$ will make the prior non-constant. We have to decide what continuous parameter-variable is most appropriate or natural. In this book we will only distinguish between two cases. A 'location parameter' just sets a location, which may be positive or negative and largeness or smallness of the value has no particular significance; location parameters usually take a constant prior density. In contrast, a 'scale parameter' sets a scale, has fixed sign, and the largeness or smallness of the value does matter. Now, the logarithm of a scale parameter is a location parameter. Thus a scale parameter ω takes prior density $\propto 1/\omega$. This is known as a Jeffreys prior. (Of course, this argument is only a justification, not a derivation. There are several possible derivations, and we will come to one later, at the end of the next chapter.) Discrete parameters may also take Jeffreys priors; for example, if ω is allowed to be any integer from 1 to ∞.

Often, in addition to the parameters ω that we are interested in, there are parameters μ whose values we don't care about. A simple example is an uninteresting normalization present in many problems; we will come across more subtle examples too. The thing to do with so-called nuisance parameters is to marginalize them out:

$$\mathfrak{prob}(D \,|\, \omega, M) = \sum_{\mu} \mathfrak{prob}(D \,|\, \mu, \omega, M)\, \mathfrak{prob}(\mu \,|\, \omega, M). \tag{1.23}$$

Note that in the marginalization rule (1.7) the thing being marginalized is always to the *left* of the condition in a conditional probability. If what we want to marginalize is to the right, as is μ in $\mathfrak{prob}(D \,|\, \mu, \omega, M)$ in (1.23) we need to move it to the left by multiplying by a prior—in this case $\mathfrak{prob}(\mu \,|\, \omega, M)$—and invoking the product rule.

We now move on to model comparison. Say we have two models M_1 and M_2, with parameters ω_1 and ω_2 respectively, for the same data. To find out which model the data favour we first compute

$$\mathfrak{prob}(D \,|\, M_1) = \sum_{\omega_1} \mathfrak{prob}(D \,|\, \omega_1, M_1)\, \mathfrak{prob}(\omega_1 \,|\, M_1) \tag{1.24}$$

[1] No, really. The name comes from Keynes, though the idea is much older.

and similarly $\mathfrak{prob}(D\,|\,M_2)$, and then using Bayes' theorem we have

$$\frac{\mathfrak{prob}(M_1\,|\,D)}{\mathfrak{prob}(M_2\,|\,D)} = \frac{\sum_{\omega_1}\mathfrak{prob}(D\,|\,\omega_1,M_1)\,\mathfrak{prob}(\omega_1\,|\,M_1)}{\sum_{\omega_2}\mathfrak{prob}(D\,|\,\omega_2,M_2)\,\mathfrak{prob}(\omega_2\,|\,M_2)} \times \frac{\mathfrak{prob}(M_1)}{\mathfrak{prob}(M_2)} \tag{1.25}$$

since $\mathfrak{prob}(D)$ cancels. Here $\mathfrak{prob}(M_1)$ and $\mathfrak{prob}(M_2)$ are priors on the models, and the ratio of posteriors on the left hand side is called the "evidence", "odds ratio", or the "Bookmakers' odds" for M_1 versus M_2.

Note that model comparison makes inferences about models with all parameters marginalized out. In two respects it contrasts parameter fitting. First, when parameter fitting we can usually leave the prior and likelihood unnormalized, but when comparing models all probabilities need to be normalized. Second, when parameter fitting we are disappointed if the likelihood $\mathfrak{prob}(D\,|\,\omega,M)$ is a very broad distribution because it makes the parameters very uncertain, but in model comparison a broad likelihood distribution actually favours a model because it contributes more in the sum (1.24).

The third topic is goodness of fit. Model comparison tells us the best available model for a given data set, and parameter fitting tells us the best parameter values. But they do not guarantee that *any* of the available models is actually correct. (Imagine fitting data from an odd polynomial to different even polynomials. We would get a best-fit, but it would be meaningless.) So having found the best model and parameters, unless we know for other reasons that one of the models is correct, we still have to pose a question like "could these data plausibly have come from that model?" Testing the goodness of fit provides a way of addressing this question, though not a very elegant way. To do this we must choose a statistic or function (say ψ) which measures the goodness of fit, better fits giving lower ψ. A good example is the reciprocal of the likelihood: $1/\mathfrak{prob}(D\,|\,\omega,M)$. Then we compute the probability that a random data set (from the fixed model and parameter values being tested) would fit *less well* than the actual data:

$$\mathfrak{prob}(\psi > \psi_D), \tag{1.26}$$

ψ_D being what the actual data give. The probability (1.26) is called the p-value of the data for the statistic ψ. If the p-value is very small the fit is clearly anomalously bad and the model must be rejected.

A goodness-of-fit test needs a good choice of statistic to work well. If we choose a ψ that is not very sensitive to deviations of model from data, the test will also be insensitive. But there is no definite method for choosing a statistic. The choice is an ad hoc element, which makes goodness of fit a less elegant topic than parameter fitting and model comparison.

There is another important contrast between parameter fitting and model comparison on the one hand and goodness-of-fit testing on the other hand. The first two hold the data

fixed and consider probabilities over varying models and parameters, whereas the third holds the model and parameters fixed and considers probabilities over varying data sets. Hence the first two need priors for models and parameters, whereas the third does not. Now, priors are a very Bayesian idea, outside the Frequentist regime. Thus goodness-of-fit testing is part of Frequentist theory, whereas parameter fitting and model comparison (as developed here) are not expressible in Frequentist terms.

There is parameter estimation in Frequentist theory of course, but it works differently; it is based on "estimators". An estimator is a formula $\varpi(D, M)$ that takes a model and a data set and generates an approximation to the parameters. The innards of $\varpi(D, M)$ are up to the insight and ingenuity of its inventor. An example of an estimator is "ω such that $\mathfrak{prob}(D \,|\, \omega, M)$ is maximized"; it is called the maximum likelihood estimator. A desirable property of estimators is being "unbiased", which means that the estimator averaged over possible data sets equals the actual parameter value, i.e.,

$$\frac{\sum_D \varpi(D, M)\ \mathfrak{prob}(D \,|\, \omega, M)}{\sum_D \mathfrak{prob}(D \,|\, \omega, M)} = \omega. \tag{1.27}$$

Not all used estimators are unbiased; for example, maximum likelihood estimators are in general biased. Meanwhile, Frequentist model comparison sometimes compares the maximized values of the likelihoods for different models; sometimes it involves a goodness-of-fit statistic. Thus, Frequentist parameter fitting and model comparison both rely much more on ad hoc choices and much less on general principles than the Bayesian methods. This book will refer only occasionally to estimators.

The fourth and last topic is maximum entropy. In our discussion of the first three types of problem, assigning priors played a fairly minor role. We treated priors as just a way of making some tacit assumptions explicit, on the understanding that with good data any sensible prior would lead to the same inferences. This approach is fine for many problems, including most we will discuss in this book. For some problems, however, assigning probabilities is not a peripheral concern, it is the main thing. Such problems involve data that are incomplete in some severe way, though they may be highly accurate. Here is an artificial, but representative, example: suppose we have a biased six-sided die, and data tell us the average number of dots is not 3.5 but 4.5; with no further data can we make any predictions for the probability of 1 dot? Probability theory by itself does not tell us how to proceed, we need to add something. That something is the principle of maximum entropy.

The concept of entropy comes originally from physics but it has a more fundamental meaning in information theory as a measure of the uncertainty in a probability distribution. Now, a probability distribution $p_1, \ldots, p_N$ certainly implies uncertainty and lack of knowledge, but it is not obvious that the uncertainty can be usefully quantified by a single number. However, a fundamental theorem in information theory (Shannon's theorem) shows that if we suppose that there *is* a measure $S(p_1, \ldots, p_N)$ of the uncertainty and

moreover S satisfies certain desirable conditions, then S must be

$$S = -\sum_{i=1}^{N} p_i \log p_i. \tag{1.28}$$

Equation (1.28) is the information-theoretic entropy. We will discuss its derivation and its relation to physics in detail in chapters 7 and 8, but for now we can think of S (with $\log_2$) as the number of bits needed to change a probability distribution to a certainty.

The principle of maximum entropy is that when probabilities need to be assigned, they are assigned so as to maximize the entropy so far as data allow. Thus, in the case of the biased-die problem, we would maximize $S(p_1, \ldots, p_6)$ subject to $\sum_k kp_k = 4.5$. It is not hard to see that if there are no constraints (apart from $\sum_i p_i = 1$ of course) maximum entropy reduces to indifference.

The above information-theory arguments and the limiting case of indifference all motivate the maximum-entropy principle, but they do not require it. I should emphasize that maximum entropy is something new being added to the methodology to fill a need. It is the idea most associated with Jaynes, though the physicists Boltzmann and Gibbs anticipated aspects of it.

PROBLEM 1.2: A Bayesian arriving in an unfamiliar town sees two taxis, and their numbers are 65 and 1729. Assuming taxis in this town are numbered sequentially from 1 to N, what is the probability distribution for N? **[2]**

For many years Bayesian methods were controversial and promoted by only a small minority of enthusiasts. Look up Bayes' Theorem in a good old-fashioned probability and statistics book, and you may find disapproval[1] or horror[2] at the idea of using Bayes' theorem for anything other than some artificial examples. On the other hand, look up "Frequentist" in a Bayesian manifesto[3] and you will find example upon example of problems where Frequentist methods fail. But post 1990 or so, with Bayesian methods better known and the contentious arguments mostly already written, authors tend to be more relaxed. Thus a mainly Frequentist source will endorse some Bayesian ideas,[4] and a Bayesian book will recommend some Frequentist methods.[5] Jaynes once expressed the

---

[1] W. Feller, *An introduction to probability theory and its applications* (Wiley 1971).

[2] J.V. Uspensky, *Introduction to mathematical probability*, (McGraw-Hill 1937).

[3] For example, T.J. Loredo, *From Laplace to SN 1987A: Bayesian inference in astrophysics* (1990, available in `bayes.wustl.edu`) and above all Jaynes.

[4] The data analysis parts of *Numerical Recipes* by W.H. Press, S.A Teukolsky, W.T. Vetterling, & B.P. Flannery (Cambridge University Press 1992) are a good example.

[5] J.M. Bernardo & A.F.M. Smith *Bayesian Theory* (Wiley 1994) and A. Gelman, J.B. Carlin, H.S. Stern, & D.B. Rubin *Bayesian Data Analysis* (Chapman & Hall 1995) are two recent examples; Jeffreys (uncharacteristic of its generation in this respect as in others) is another.

hope that in time one should no more need the label of "Bayesian" to use Bayes' theorem than one needs to be a "Fourierist" to use Fourier transforms. Writing in 2002, perhaps that time has arrived.

PROBLEM 1.3: This problem is a long digression on yet another aspect of probability. However, the actual calculation you have to do is very simple, and when you have done it you can feel justly proud of having deduced Bell's theorem, a profound statement about the physical world.

At the sub-molecular level, nature becomes probabilistic in an important and even disturbing way. The theory of quantum mechanics (which predicts experimental results in this regime with incredible accuracy) has probability as a fundamental part. But even some of the founders of quantum mechanics, notably Einstein, were very uncomfortable about its probabilistic character, and felt that quantum mechanics must be a simplification of an underlying deterministic reality.

Then in 1964 J.S. Bell concocted a thought experiment to demonstrate that if quantum mechanics is correct, then an underlying deterministic theory won't work. Real experiments based on Bell, starting with those by A. Aspect and coworkers (c. 1980), show exactly what quantum mechanics predicts, though some problems of interpretation remain. Bell's thought experiment can be appreciated without a knowledge of quantum mechanics, and in this problem we will follow a version due to Mermin.[1]

There are three pieces of apparatus, a source in the middle and two detectors to the left and right. Each detector has a switch with three settings (1, 2, and 3 say) and two lights (red and green). The source has a button on it; when this button is pressed the source sends out two particles, one towards each detector. Each particle is intercepted by the detector it was sent towards, and in response the detector flashes one of its lights.

The experiment consists of repeatedly pressing the source button for random settings of the detector switches and recording which light flashed on which detector and what the switch settings were. Thus if at some button-press, the left-hand detector is set to '1' and flashes green while the right-hand detector is set to '3' and flashes red, we record `1G3R`. The observations look something like `1G3R 1G1G 3G2G 3G1R 3R2G 3R3R 1R3R 1R3G 1R2G...`

An important point is that, apart from the sending of particles from source to detector, there is no communication between the three pieces of apparatus. In particular, the source never knows what the detector settings are going to be (the settings can change while a particle is in transit).

Examining the observations after many runs, we find two things.[2]

(i) When both detectors happen to have the same switch settings they always flash the same colour, with no preference for red or green. Thus, we see `1R1R` and `1G1G` equally often, but never `1R1G` or `1G1R`.

---

[1] N.D. Mermim, *Boojums all the way through* (Cambridge University Press 1990).

[2] The particles are spin-half particles (e.g., electrons), and the source ensures that pairs emitted together have the same spin. The three settings on the detectors measure the spin component along one of three directions at 120° to each other: green light flashes for spin component $+\frac{1}{2}$, red light for $-\frac{1}{2}$. When the detectors have different switch settings, quantum mechanics predicts the colour coincidence rate to be $\cos^2\left(\frac{1}{2}\times 120^\circ\right)=\frac{1}{4}$. This is what is observed.

(ii) Considering all the data (regardless of switch settings) the colours coincide, on-average, half the time.

Neither of (i) or (ii) is remarkable in itself, but taken together they are fatal for a deterministic explanation. To see this, we consider what a deterministic explanation implies. To determine which colour flashes, each particle would have to carry, in effect, an instruction set of the type "green light if the setting is 1, red light otherwise" (say GRR). Since the source doesn't know what the detectors switch settings are going to be, *both* particles from any button-press must have the same instruction set. Different button-presses can give different instruction sets. In particular, the above data could have resulted from GRR GRR GGG RRG RGR GGR RGR RGG RGG...

Your job is to show that the average coincidence rate from instruction sets will be $\frac{5}{9}$ or more, not $\frac{1}{2}$. **[4]**

2. Binomial and Poisson Distributions

The binomial distribution is basically the number of heads in repeated tosses of a (possibly biased) coin. The Poisson distribution is a limiting case. We go into these two in some detail, because they come up in many applications, and because they are a nice context to illustrate the key ideas of parameter fitting and model comparison.

If some event (e.g., heads) has probability p, the probability that it happens n times in N independent trials is

$$\mathfrak{prob}(n\,|\,N) = {}^{N}C_n\, p^n q^{N-n}, \quad q = 1 - p. \tag{2.1}$$

The n occurrences could be combined with the $N - n$ non-occurrences in ${}^{N}C_n$ ways, hence the combinatorial factor. The sum $\sum_n \mathfrak{prob}(n\,|\,N)$ is just the binomial expansion of $(p + q)^N$, so the distribution is correctly normalized, and hence the name.[1]

A generalization which comes up occasionally is the multinomial distribution. Here we have K possible outcomes with probabilities p_k. The probability that the k-th outcome will happen n_k times in N independent trials is

$$\mathfrak{prob}(n_1, \ldots, n_K) = N! \prod_{k=1}^{K} \frac{p_k^{n_k}}{n_k!}, \quad \sum_{k=1}^{K} p_k = 1, \quad \sum_{k=1}^{K} n_k = N. \tag{2.2}$$

For the combinatorial factor, take the total number of possible permutations ($N!$) and reduce by the number of permutations among individual outcomes ($n_k!$). The sum of $\mathfrak{prob}(n_1, \ldots, n_K)$ over k is $\left(\sum_k p_k\right)^N$ so again the distribution is normalized.

Another variant is the negative binomial (also called Pascal) distribution. Here, instead of fixing the number of trials N, we keep trying till we get a pre-specified number n of events. The probability that we need N trials for n events is the probability of getting $n - 1$ events in $N - 1$ trials and then an event at the N-th trial, so

$$\mathfrak{prob}(N\,|\,n) = {}^{N-1}C_{n-1}\, p^n q^{N-n}. \tag{2.3}$$

Now $\sum_{N=n}^{\infty} {}^{N-1}C_{n-1}\, q^{N-n}$ is just the binomial expansion for $(1 - q)^{-n}$, so (2.3) is normalized.

ASIDE (A probably useless but certainly cute point.) If we let both $\mathfrak{prob}(n)$ and $\mathfrak{prob}(N)$ take Jeffreys priors then $\mathfrak{prob}(n\,|\,N)$ and $\mathfrak{prob}(N\,|\,n)$ derive from each other via Bayes' theorem. □

[1] Another term is 'Bernoulli trials' after Jakob Bernoulli, one of the pioneers of probability theory and author of *Ars Conjectandi* (published 1713).

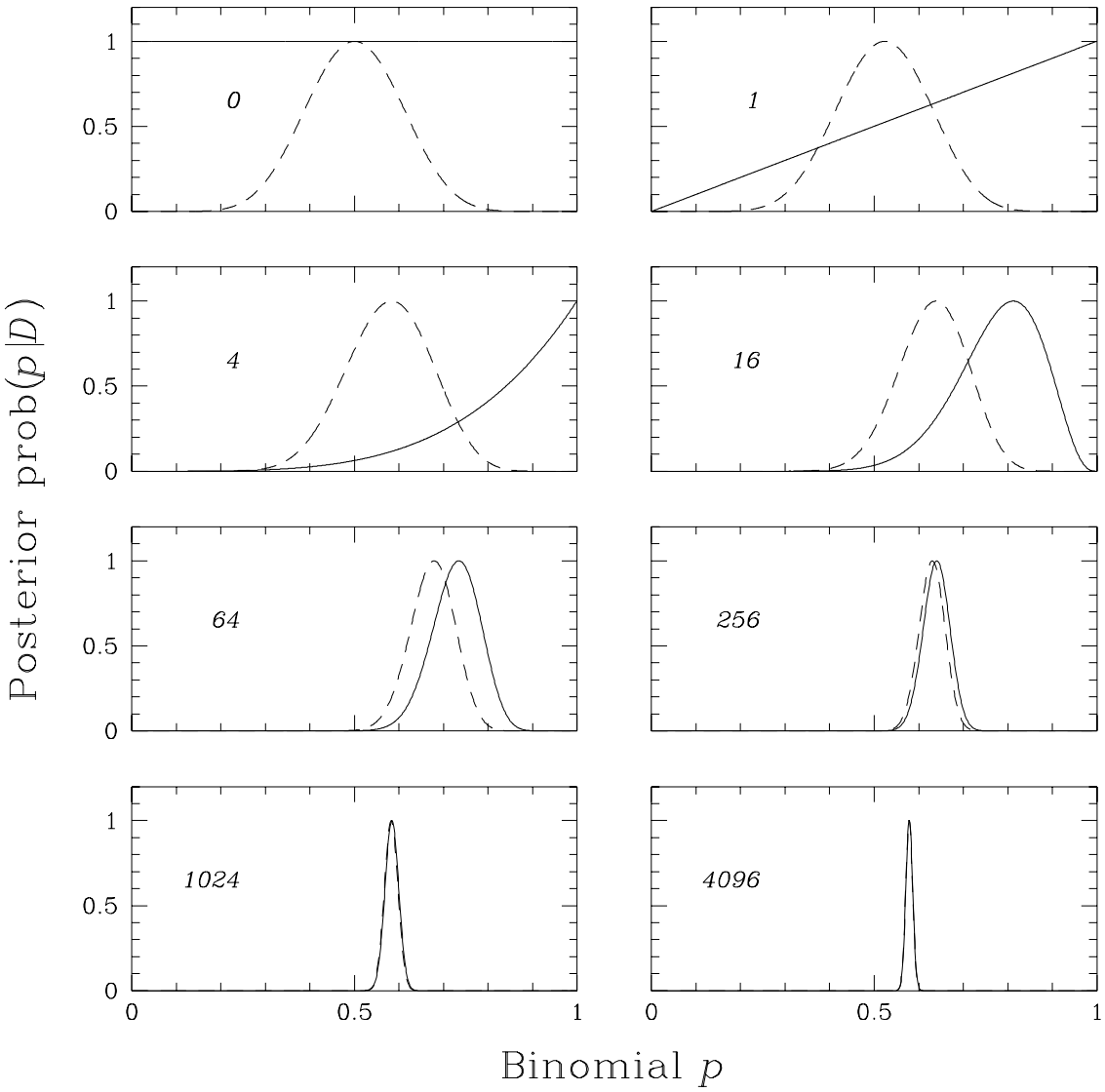

Figure 2.1: Posterior probability distribution after different numbers of simulated coin tosses; solid curve for the uninformative prior, dashed curve for the informative prior. For neatness, the curves are not normalized, just scaled to unit maximum.

EXAMPLE [A biased coin] This is a toy problem to illustrate how parameter fitting and model comparison work.

Imagine a coin which is possibly biased, i.e., $p \equiv \mathfrak{prob}(\text{heads})$ possibly $\neq \frac{1}{2}$. We proceed to toss it 4096 times to (a) examine if the coin is biased, and (b) find out what p is if it *is* biased. [This *is* an imaginary coin, only my computer thinks it exists.]

We will consider two models: (i) M_1, where p is a parameter to be fitted; and (ii) M_2, where $p = \frac{1}{2}$ and there are no parameters. If there are n heads after N tosses, the likelihoods from the two

models are

$$\begin{aligned} \mathrm{prob}(D\,|\,p, M_1) &= {}^{N}C_n\, p^n (1-p)^{N-n}, \\ \mathrm{prob}(D\,|\,M_2) &= {}^{N}C_n\, 2^{-N}. \end{aligned} \tag{2.4}$$

What to do for the prior for p in model M_1? To illustrate the possibilities, we try two priors: a flat prior and a prior broadly peaked around $p = \frac{1}{2}$. These would commonly be called uninformative and informative, the latter being based on a premise that makers of biased coins probably wouldn't try to pass off anything too obviously biased. We take

$$\mathrm{prob}(p\,|\,M_1) = \frac{(2K+1)!}{(K!)^2} p^K (1-p)^K \tag{2.5}$$

with $K = 0$ (uninformative) and $K = 10$ (informative). The factorials serve to normalize—see equation (M.3) on page 70.

The posterior for M_1 is then

$$\mathrm{prob}(p|D, M_1) \propto {}^{N}C_n\, p^{n+K} (1-p)^{N-n+K} \tag{2.6}$$

and is plotted in Figure 2.1 after different numbers of tosses. After 0 tosses, the posterior is of course the same as the prior. We see that with enough data the posteriors with both priors tend to the same curve.

Our fit for p really consists of the posterior distribution itself (for a given prior). For brevity, we just might calculate the mean and standard deviation of the posterior and not bother to plot the curve. But there is nothing special about the mean and standard deviation, one might prefer to quote instead the median, along with the 5th and 95th percentile values as a "90% confidence interval".[1]

To compare M_1 and M_2 we need to marginalize out the parameter dependence in M_1, that is, we need to multiply the likelihood $\mathrm{prob}(D\,|\,p, M_1)$ from (2.4) with the normalized prior $\mathrm{prob}(p\,|\,M_1)$ from (2.5) and integrate over p. For M_2 there are no parameters to marginalize. We get for the marginalized likelihood ratio:

$$\frac{\mathrm{prob}(D\,|\,M_1)}{\mathrm{prob}(D\,|\,M_2)} = \frac{(2K+1)!(n+K)!(N-n+K)!}{(K!)^2 (N+2K+1)!}\, 2^N. \tag{2.7}$$

In general, if $p = \frac{1}{2}$, the marginalized likelihood ratio will start at unity and slowly fall; if $p \neq \frac{1}{2}$, it will start at unity and hover or fall a little for a while, but will eventually rise exponentially. The Bookmakers' odds ratio is the ratio of posterior probabilities of the two models:

$$\frac{\mathrm{prob}(M_1\,|\,D)}{\mathrm{prob}(M_2\,|\,D)} = \frac{\mathrm{prob}(D\,|\,M_1)}{\mathrm{prob}(D\,|\,M_2)} \times \frac{\mathrm{prob}(M_1)}{\mathrm{prob}(M_2)}. \tag{2.8}$$

Here $\mathrm{prob}(M_1)$ and $\mathrm{prob}(M_2)$ are priors *on the models.* Figure 2.2 shows the odds with model priors set equal. One might prefer, say, $\mathrm{prob}(M_1) = 10^{-6}\,\mathrm{prob}(M_2)$ on the grounds that most coins are

[1] Bayesians tend to prefer "credible region", to avoid possible confusion with the Frequentist meaning of "confidence interval", which has to do with the distribution of an estimator.

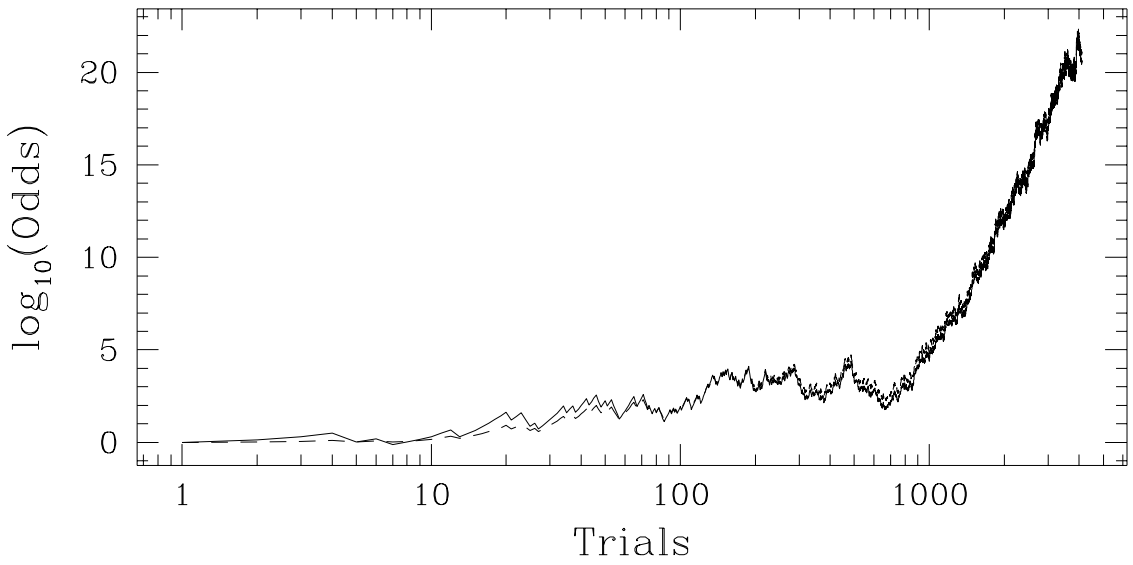

Figure 2.2: Bookmakers' odds for the coin being biased, as the tosses progress. Again solid curve for the uninformative prior and dashed curve for the informative prior, but they almost coincide.

unbiased, but as Figure 2.2 shows, even that would become pretty irrelevant when we are confronted with enough data.

In more serious examples, we will have to do more numerical work, but the general idea will remain much the same. Some things to take away from this example are the following.

(i) For parameter fitting, there's no need to normalize probabilities; but for model comparison normalization is essential—and remember that priors have to be normalized too.

(ii) Information on parameters comes from the moment we start to take data, but takes time to stabilize.

(iii) Model comparison takes more data than parameter fitting—there's a larger space of possibilities to consider!

(iv) Priors are important when we have little data, but with lots of data they become pretty irrelevant. □

The Poisson distribution is the limit of a binomial distribution as $p \to 0$ but $N \to \infty$ such that Np remains finite. Writing m for Np, we get for large N:

$$\begin{aligned} \mathfrak{prob}(n\,|\,m) &= \frac{N!}{n!(N-n)!} \times \left(\frac{m}{N}\right)^n \times \left(1-\frac{m}{N}\right)^{N-n} \\ &\simeq \frac{N^n}{n!} \times \left(\frac{m}{N}\right)^n \times \left(1-\frac{m}{N}\right)^{N} \end{aligned} \tag{2.9}$$

and hence (using the formula M.1 on page 70 for e)

$$\mathfrak{prob}(n\,|\,m) \to e^{-m}\frac{m^n}{n!}. \tag{2.10}$$

The Poisson distribution is clearly normalized since $\sum_n \mathrm{prob}(n|m) = 1$. The number of trials N is replaced by a waiting time—m is proportional to that waiting time. A common example in astronomy is the number of photons received from some source for given exposure time. In this example one bothers with (2.10) only if the source is faint enough that m is quite small—otherwise it goes over to another limiting form we'll come to in the next chapter. We'll call m the mean (and justify that name shortly).

PROBLEM 2.1: Show that, for a Poisson process, the probability distribution of the waiting time τ needed for $n+1$ events is

$$\mathrm{prob}(\tau|n+1,m) = e^{-m\tau}\frac{m^{n+1}\tau^n}{n!},$$

given that m is the mean for $\tau = 1$. **[2]**

EXAMPLE [Two Poisson processes] In this example we work out Bookmakers' odds for whether two Poisson processes have different means, given data on numbers of events. One process has produced $m_1, \ldots, m_K$ events over K different periods, and the second has produced $n_1, \ldots, n_L$ events over L different periods. All the waiting periods (for both processes) are equal.

We take two models: in M_1, both processes have mean a; in M_2, the means are a and b. Writing $M = \sum_{k=1}^{K} m_k$ and $N = \sum_{l=1}^{L} n_l$, the likelihoods are (using 2.10)

$$\begin{aligned}\mathrm{prob}(D\,|\,a,b,M_2) &\propto e^{-aK}e^{-bL}a^M b^N,\\ \mathrm{prob}(D\,|\,a,M_1) &\propto e^{-a(K+L)}a^{M+N}.\end{aligned} \tag{2.11}$$

(I've suppressed a product of $m_k!$ and $n_l!$ factors in the likelihoods, since they would always cancel in this problem.) Notice how the likelihoods have simplified to depend on the data effectively only through M and N.

Now we have to marginalize the likelihoods over the means a and b. So we need a normalized prior for these. Since a and b set scales they should each take a Jeffreys prior. But the simple $1/a$ Jeffreys prior is not normalized and thus cannot be used for model comparison, though it is fine for parameter fitting. We need to modify the prior in some way, so that we can normalize it.

The way out is not very elegant, but it works. We know that a is not going to be orders of magnitude different from the m_k (and similarly for b), and just by looking at the data, we can name some $c_{\min}$ and $c_{\max}$, between which we are very confident a and b must lie. We then let

$$\begin{aligned}&\mathrm{prob}(a) = (a\Lambda)^{-1}, \quad c_{\min} \le a \le c_{\max},\\ &\text{where } \Lambda = \ln(c_{\max}/c_{\min}),\end{aligned} \tag{2.12}$$

and zero outside $[c_{\min}, c_{\max}]$, and similarly for $\mathrm{prob}(b)$. These priors are normalized. Of course they depend on our (arbitrary) choice of $c_{\min}$ and $c_{\max}$, but only logarithmically, and since data-dependent terms tend to be factorial or exponential, we put up with this arbitrariness.

An alternative (sometimes called the Γ-prior) is

$$\mathrm{prob}(a) = (\Gamma(\epsilon))^{-1}\delta^{\epsilon}a^{\epsilon-1}e^{-\delta a} \tag{2.13}$$

with ϵ and δ being small numbers to be chosen by us. The Gamma-function formula (M.2) from page 70 shows this prior is normalized. It is similar to (2.12) except that the cutoffs are nice and smooth. The values of ϵ and δ determine in which regions the cutoffs appear, and we would want the cutoffs to be around $c_{\min}$ and $c_{\max}$. So we still need to think implicitly about $c_{\min}$ and $c_{\max}$.

Let us say we adopt the prior (2.12). We should then really calculate

$$\mathrm{prob}(D \,|\, M_1) = \int_{c_{\min}}^{c_{\max}} \mathrm{prob}(D \,|\, a, M_1)\,(a\Lambda)^{-1}\,da \tag{2.14}$$

and so on. But the integrands will become very small outside $[c_{\min}, c_{\max}]$ and we can comfortably use the approximations

$$\mathrm{prob}(D \,|\, M_1) \simeq \int_0^{\infty} \mathrm{prob}(D \,|\, a, M_1)\,(a\Lambda)^{-1}\,da \tag{2.15}$$

and so on. Using the Gamma-function formula (M.2) again we get

$$\frac{\mathrm{prob}(D \,|\, M_2)}{\mathrm{prob}(D \,|\, M_1)} = \frac{1}{\Lambda}\,\frac{(K+L)^{M+N}}{K^M L^N}\,\frac{(M-1)!(N-1)!}{(M+N-1)!}. \tag{2.16}$$

When M/K is not too different from N/L, these odds are close to $1/\Lambda$, and decrease slowly as M and N increase. But if M/K and N/L are very different, the odds favouring two different means become spectacularly large. □

PROBLEM 2.2: The following suggestion, applied to the previous example, would avoid having the odds depend on the cutoff. For M_1 use the prior (2.12) on a; for M_2, let $a = uc$ and $b = (1-u)c$, and then use a flat prior in $[0,1]$ for u and the prior (2.12) for c.

The odds can be worked out without a computer only for $K = L$. For this case show that the odds for different means are

$$2^{M+N}\,\frac{M!N!}{(M+N+1)!}.$$

This is similar to what (2.16) would give for $K = L$, but of course not identical, and the difference conveys some feeling for how much the results depend on the choice of prior. **[3]**

One thing we will often have to do later in this book is take expectation values. The expectation value of a function f of a probabilistic number x is

$$\langle f(x) \rangle \equiv \sum\nolimits_i f(x_i)\ \mathrm{prob}(x_i), \tag{2.17}$$

or an integral if appropriate. We can think of an expectation value as an average over the probability distribution, weighted by the probability; or we can think of it as marginalizing out the x dependence of $f(x)$. Another notation is $E(f(x))$. One minor caution: if there are several variables involved, we must not confuse expectation values over different variables.

If we have some $\mathrm{prob}(x)$, then $\langle x^n \rangle$ is called the n-th moment of x. The first two moments are particularly important: $\langle x \rangle$ is called the mean, $\langle x^2 \rangle - \langle x \rangle^2$ is called the variance, and the square root of the variance is called the standard deviation or dispersion.

PROBLEM 2.3: Show that a binomial distribution has a mean of Np and a variance of Npq, and that a Poisson distribution has both mean and variance equal to m. **[2]**

PROBLEM 2.4: Two proofreaders independently read a book. The first finds n_1 typos, the second finds n_2 typos, including n_{12} typos found by both. What is the expected number of typos not found by either? **[1]**

DIGRESSION [Jeffreys Prior, again] The idea of expectation values gives us a nice way to derive the Jeffreys prior, which we motivated on page 9 but have not yet properly derived.

Imagine that you and a sibling receive gift tokens in outwardly identical envelopes. The gift tokens have values x and x/N, where N is known but x is unknown. You open your envelope and find you have (in some units) 1. What is your expectation for your sibling's value?

After opening your envelope, you know that x must be either 1 or N, leaving your sibling with either $1/N$ or N. (For definiteness we suppose $N > 1$.) Let us first calculate the probability that you have the larger gift token. To do this, we use the notation $\mathrm{prob}(1)$ for the probability of $x = 1$, and $\mathrm{prob}(1^*)$ for the probability of the "data" (i.e., of you opening your envelope and finding a value of 1) and so on. The posterior probability that your gift token is larger is

$$\mathrm{prob}(1\,|\,1^*) = \frac{\mathrm{prob}(1^*\,|\,1)\,\mathrm{prob}(1)}{\mathrm{prob}(1^*\,|\,1)\,\mathrm{prob}(1) + \mathrm{prob}(1^*\,|\,N)\,\mathrm{prob}(N)}. \tag{2.18}$$

Assuming the envelopes are assigned at random, $\mathrm{prob}(1^*\,|\,1) = \mathrm{prob}(1^*\,|\,N)$, and hence

$$\mathrm{prob}(1\,|\,1^*) = \frac{\mathrm{prob}(1)}{\mathrm{prob}(1) + \mathrm{prob}(N)}. \tag{2.19}$$

From this we compute your expectation for your sibling's token as

$$\frac{(1/N)\,\mathrm{prob}(1) + N\,\mathrm{prob}(N)}{\mathrm{prob}(1) + \mathrm{prob}(N)}. \tag{2.20}$$

Now (2.20) had better be 1. Otherwise you would be concluding that your sibling was better or worse off, regardless of data (which, however well it may reflect the human condition, is bad data analysis). Equating (2.20) to 1 gives

$$\mathrm{prob}(N) = (1/N)\,\mathrm{prob}(1), \tag{2.21}$$

which is the Jeffreys prior.

Curiously, in order to make the expectation value (2.20) equal 1, the probability (2.19) has to be $> \frac{1}{2}$. In other words, with the Jeffreys prior you are in one sense neutral as to whether your sibling is better off, while in another sense concluding that you yourself are probably better off. □

3. Gaussian Distributions

A Gaussian (or normal) distribution is given by

$$\mathfrak{prob}(x) = \frac{1}{\sqrt{2\pi}\,\sigma}\,\mathfrak{exp}\left[-\frac{(x-m)^2}{2\sigma^2}\right]. \tag{3.1}$$

As we can easily verify (using the Gaussian integrals M.9 and M.10 from page 71) $\mathfrak{prob}(x)$ is normalized and has mean m and dispersion σ. A Gaussian's tails fall off quickly; 68% of the probability lies within $m \pm \sigma$, 95% lies within $m \pm 2\sigma$, and 99.7% lies within $m \pm 3\sigma$.

Gaussians are a generic limiting form for the probability of a sum where the terms in the sum are probabilistic. The detailed distributions of the terms tend to get washed out in the sum, leaving Gaussians.

To see why this detail-washing-out happens, we note first that when we have several independent trials from some $\mathfrak{prob}(x)$, the probability distribution of the sum is a convolution. Thus

$$\mathfrak{prob}(\text{sum of 2 trials} = x) = \int \mathfrak{prob}(y)\,\mathfrak{prob}(x-y)\,dy, \tag{3.2}$$

which is just a consequence of the marginalization rule, and so on for several trials. In other words, the probability of a sum is the convolution of the probabilities. Convolutions are conveniently manipulated using Fourier transforms; if we define

$$\mathfrak{cf}(k) = \int_{-\infty}^{\infty} e^{ikx}\,\mathfrak{prob}(x)\,dx \tag{3.3}$$

and use the convolution theorem (M.16) from page 72 we have

$$\mathfrak{prob}(\text{sum of N trials} = x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-ikx}\,\mathfrak{cf}^N(k)\,dk. \tag{3.4}$$

The Fourier transform $\mathfrak{cf}(k)$ of a probability distribution function $\mathfrak{prob}(x)$ is called the characteristic function, or moment generating function. The latter name comes because

$$\mathfrak{cf}(k) = \left\langle e^{ikx}\right\rangle = 1 + ik\langle x\rangle - \frac{k^2}{2!}\langle x^2\rangle + \cdots \tag{3.5}$$

and the n-th moment of x (if it exists) is i^{-n} times the n-th derivative of $\mathfrak{cf}(0)$.[1] Since $\mathfrak{prob}(x)$ is normalized, $\mathfrak{cf}(k)$ always exists, but derivatives need not. A counterexample is the Lorentzian (or Cauchy) distribution, for which

$$\begin{aligned} \mathfrak{prob}(x) &= \frac{1}{\pi(1+x^2)}, \\ \mathfrak{cf}(k) &= e^{-|k|}. \end{aligned} \tag{3.6}$$

[1] In statistics the moment generating function is usually defined as $\langle \mathfrak{exp}(kx)\rangle$ rather than $\langle \mathfrak{exp}(ikx)\rangle$ as here.

Here $\mathfrak{prob}(x)$ doesn't have a mean or higher moments, and correspondingly, $\mathfrak{cf}(0)$ is not smooth and doesn't generate any moments.

Suppose, though, that we have some $\mathfrak{prob}(x)$ for which $\langle x \rangle$ and $\langle x^2 \rangle$ do exist. We can assume without losing generality that the mean is 0 and the variance is 1. Then

$$\mathfrak{cf}(k) = 1 - \tfrac{1}{2}k^2 + k^2\theta(k), \tag{3.7}$$

where the remainder term $\theta(k) \to 0$ as $k \to 0$. For large-enough N

$$\mathfrak{cf}^N(k) \simeq e^{-Nk^2/2} \tag{3.8}$$

and using the Fourier transform identity (M.17) from page 72 we get

$$\mathfrak{prob}(\text{sum of } N \text{ trials} = x) \simeq \frac{1}{\sqrt{2\pi N}} e^{-x^2/(2N)}. \tag{3.9}$$

In other words, for N (many) independent trials, the probability distribution of the sum tends to a Gaussian with mean and variance scaled up from the original by N; equivalently, the mean of N trials tends to become Gaussian-distributed with the original mean and with the variance scaled down by N. This result is known as the central limit theorem.

From the central limit theorem it follows immediately that for large N, a binomial distribution tends to a Gaussian with mean Np and variance Npq, while a Poisson distribution tends to a Gaussian with equal mean and variance. But the small contributions involved must themselves have finite means and dispersions, otherwise the theorem doesn't apply; in particular, a sum over trials from a Lorentzian doesn't turn into a Gaussian.

PROBLEM 3.1: For a probability distribution constant in $[-1, 1]$ and zero elsewhere, calculate the characteristic function and verify that it gives the correct variance. **[1]**

EXAMPLE [The central limit theorem on ice] This is a rather artificial example, but I thought it would be interesting to see the central limit theorem do some rather surprising things.

Imagine a large frictionless ice rink, with many hockey pucks whizzing about on it. Each puck has negligible size and mass m. The surface mass density of the puck distribution is Σ, and the pucks have no preferred location. The pucks have speed distribution $f(v)$ but the velocities have no preferred direction.

We place on the ice, at rest, a disc of mass M and radius a (much larger than the pucks). We then wait for some time τ, during which many pucks collide with the disc, pushing it around. We want to calculate the probability distribution for how much the disc moves in that time.

In the following we will assume that most pucks are always moving much faster than the disc, and also that most pucks move a distance $\gg a$ in time τ.

The central limit theorem becomes applicable because there are many collisions with the disc, each contributing to the disc's net displacement. Thus the probability distribution for the disc's

position will be a Gaussian and the mean will clearly be zero. It is enough to calculate the variance, or the mean square displacement (call it Δ^2) in one direction from all collisions.

Consider a puck with speed v. It will hit the disc some time during the interval τ if (i) its initial distance r from the disc is less than $v\tau$, and (ii) it is aimed in the right direction. Since most pucks will come from distances $\gg a$, from distance r a fraction $a/(\pi r)$ of pucks will be aimed at the disc. If the puck hits the disc at a point θ on the disc circumference (relative to the x axis) and is travelling in a direction ψ relative to the normal at that point, the change in the disc's x-velocity will be

$$\delta v_x = -\frac{2mv}{M} \cos\theta \cos\psi \tag{3.10}$$

If the puck was initially at distance r, the collision will happen at time r/v. The induced disc displacement at time τ will be

$$\delta x = -\frac{2m}{M}(v\tau - r) \cos\theta \cos\psi. \tag{3.11}$$

Averaging over θ and ψ we get

$$\langle \delta x^2 \rangle = \frac{m^2}{M^2}(v\tau - r)^2. \tag{3.12}$$

Now we integrate $\langle \delta x^2 \rangle$ over the puck distribution. Remembering the puck density and the aiming factor, we have

$$\Delta^2 = \frac{2a\Sigma}{m} \int f(v)\, dv \int_0^{v\tau} \langle \delta x^2 \rangle\, dr \tag{3.13}$$

or

$$\Delta^2 = \frac{2\Sigma a m}{3M^2} \tau^3 \langle v^3 \rangle. \tag{3.14}$$

The probability distribution for the disc is

$$\mathrm{prob}(x, y) = (2\pi)^{-1} \Delta^{-2} e^{-(x^2+y^2)/(2\Delta^2)}, \tag{3.15}$$

or, in terms of the distance:

$$\mathrm{prob}(r) = \Delta^{-2}\, r\, e^{-r^2/(2\Delta^2)}. \tag{3.16}$$

We had to assume that the third moment of $f(v)$ exists, but the other details of $f(v)$ got washed out. □

PROBLEM 3.2: A drunk has been leaning against a lamppost in a large city square, and decides to take a walk. Our friend then walks very carefully, taking equal steps, but turns by a random angle after each step.[1] How far do they get after N steps? **[2]**

[1] For a truly charming discussion of this and many many other problems, see G. Gamow, *One two three... infinity* (Dover Publications 1988—first edition 1947).

EXAMPLE [Random walking share prices] Here is an application from finance.

Suppose that at time t remaining till some fiducial date (we will use a decreasing time variable in this example) we own some stock with share price $s(t)$. The share price may rise or fall, and we want to reduce our exposure to possible future loss by trading away some possible future profit. In order to do this we sell a 'call option' on part of our stock. A call option on a share is the option to buy that share at $t = 0$, not at $s(0)$ but at a predetermined 'strike price' s_0. At $t = 0$ the option is worth

$$f(s,0) = \begin{cases} s - s_0 & \text{if } s > s_0 \\ 0 & \text{otherwise} \end{cases} \tag{3.17}$$

because if $s > s_0$ at $t = 0$ the holder of the option will buy the stock at s_0 and make a gain of $s - s_0$, but if $s < s_0$ at $t = 0$ then the option is worth nothing. The problem is to find the current value $f(s,t)$ of the option.

Suppose now that we sell such an option, not on all the stock we own but on a fraction $(\partial f/\partial s)^{-1}$ of it. Our net investment per share is then

$$s - \left(\frac{\partial f}{\partial s}\right)^{-1} f \tag{3.18}$$

and it has the nice property that changing s by Δs and f by Δf (at fixed t) produces no net change in the investment. Accordingly, the combination (3.18) of stock owned and option sold is called a 'neutral hedge equity'. With time, however, the hedge equity is expected to grow at the standard risk-free interest rate r, i.e.,

$$\Delta s - \left(\frac{\partial f}{\partial s}\right)^{-1} \Delta f = \left(s - \left(\frac{\partial f}{\partial s}\right)^{-1} f\right) r\,(-\Delta t). \tag{3.19}$$

(Here we have $-\Delta t$ because t is decreasing.) We Taylor-expand Δf as

$$\Delta f = f(s+\Delta s, t+\Delta t) - f(s,t) = \frac{\partial f}{\partial s}\,\Delta s + \tfrac{1}{2}\,\frac{\partial^2 f}{\partial s^2}\,(\Delta s)^2 + \frac{\partial f}{\partial t}\,\Delta t \tag{3.20}$$

plus higher-order terms which we will neglect. The fractional share price is assumed to random walk, thus

$$(\Delta s)^2 = -\sigma^2 s^2\,\Delta t, \tag{3.21}$$

and the constant σ is called the 'volatility'. Substituting for $(\Delta s)^2$ in the Taylor expansion, and then inserting Δf in equation (3.19) and rearranging, we get

$$\frac{\partial f}{\partial t} = \tfrac{1}{2}\sigma^2 s^2\,\frac{\partial^2 f}{\partial s^2} + r\,s\,\frac{\partial f}{\partial s} - r\,f. \tag{3.22}$$

This is a partial differential equation for $f(s,t)$, subject to the boundary conditions (3.17). The solution (derived in the following digression) is called the Black-Scholes formula:

$$f(s,t) = s\,N(d_1) - e^{-rt} s_0\,N(d_2). \tag{3.23}$$

Here

$$N(x) \equiv \frac{1}{\sqrt{2\pi}} \int_{-x}^{\infty} e^{-t^2/2}\,dt \tag{3.24}$$

is the cumulative of a Gaussian (or alternatively, the error function from page 28 written differently), and the arguments d_1, d_2 are

$$d_1 = \frac{\ln(s/s_0) + rt}{\sigma\sqrt{t}} + \tfrac{1}{2}\sigma\sqrt{t}, \quad d_2 = \frac{\ln(s/s_0) + rt}{\sigma\sqrt{t}} - \tfrac{1}{2}\sigma\sqrt{t}. \tag{3.25}$$

The first term in the Black-Scholes formula (3.23) is interpreted as the expected benefit from acquiring a stock outright, the second term is interpreted as the present value of paying the strike price on the expiration day.

Black-Scholes is the starting point for much work in current financial models.[1] □

DIGRESSION [Green's function for Black-Scholes] The Black-Scholes differential equation (3.22) is equivalent to a diffusion equation, a beast long familiar in physics and solvable by standard techniques.

To simplify to a diffusion equation, we make a few changes of variable. Writing

$$f = e^{-rt}g, \quad x = \ln(s/s_0), \tag{3.26}$$

that is, factoring out the interest-rate growth and putting the share price on a log scale, the differential equation (3.22) becomes

$$\frac{\partial g}{\partial t} = \tfrac{1}{2}\sigma^2 \frac{\partial^2 g}{\partial x^2} + \left(r - \tfrac{1}{2}\sigma^2\right) \frac{\partial g}{\partial x}\,. \tag{3.27}$$

Now changing to moving coordinates

$$z = x + \left(r - \tfrac{1}{2}\sigma^2\right)t, \quad h(z,t) = g(x,t), \tag{3.28}$$

we get

$$\frac{\partial h}{\partial t} = \tfrac{1}{2}\sigma^2 \frac{\partial^2 h}{\partial z^2}\,. \tag{3.29}$$

This is a diffusion equation, and its solution gives the effect of the boundary condition diffusing back in time.

We can solve equation (3.29) using Fourier transforms and their properties (see page 72). Writing $H(k,t)$ for the Fourier transform of $h(z,t)$ we have

$$\frac{dH}{dt} = -\tfrac{1}{2}k^2\sigma^2 H, \tag{3.30}$$

which immediately integrates to

$$H = e^{-\frac{1}{2}k^2\sigma^2 t} + \text{const.}$$

If the integration constant is 0, we can inverse transform to get

$$\hat{h}(z,t) = \frac{1}{\sqrt{2\pi t}\sigma} e^{-\frac{1}{2}z^2/(\sigma^2 t)}. \tag{3.31}$$

[1] See e.g., www.nobel.se/economics/laureates/1997/press.html

Now $\hat{h}(z,t)$ is not the general solution, but a particular solution for the boundary condition $\hat{h}(z,0) = \delta(z)$. (This is clear from the δ function limit, equation M.18 on page 72.) Such particular solutions are known in mathematical physics as Green's functions. When this Green's function is convolved with the given boundary condition, it gives us the appropriate solution:

$$\begin{aligned} h(z,t) &= \frac{1}{\sqrt{2\pi t}\sigma} \int_0^\infty e^{-\frac{1}{2}(z'-z)^2/(\sigma^2 t)}\, h(z',0)\, dz', \\ h(z',0) &= s_0\left(e^{z'} - 1\right). \end{aligned} \tag{3.32}$$

To simplify the integral we change variable from z' to $v = (z'-z)/(\sigma\sqrt{t})$, which gives

$$h(z,t) = \frac{s_0}{\sqrt{2\pi}} \int_{-d_2}^\infty e^{-\frac{1}{2}v^2 + z + \sigma\sqrt{t}v}\, dv - \frac{s_0}{\sqrt{2\pi}} \int_{-d_2}^\infty e^{-\frac{1}{2}v^2}\, dv. \tag{3.33}$$

Now changing variable in the first integral to $u = v - \sigma\sqrt{t}$ gives

$$\begin{aligned} h(z,t) &= \frac{s_0}{\sqrt{2\pi}} e^{z+\frac{1}{2}\sigma^2 t} \int_{-d_1}^\infty e^{-\frac{1}{2}u^2}\, du - \frac{s_0}{\sqrt{2\pi}} \int_{-d_2}^\infty e^{-\frac{1}{2}v^2}\, dv \\ &= s_0 e^{z+\frac{1}{2}\sigma^2 t}\, N(d_1) - s_0\, N(d_2). \end{aligned} \tag{3.34}$$

Finally, putting back the variables s and f in place of z and h (via equations 3.26 and 3.28) gives the Black-Scholes formula (3.23). □

The distribution of experimental random errors is typically Gaussian, and is usually (safely) taken to be so without bothering about the detailed contributions that make it up. Posterior distributions in different problems may also often be well-approximated by Gaussians. People giving parameter estimates often quote 1σ or 2σ or 3σ uncertainties on the results. This can mean either of two things:

(i) The posterior probability distribution has been approximated by a Gaussian, and the stated σ is the dispersion; or

(ii) The posterior is not a Gaussian, but 65% (or 95% or 99.7%, as applicable) of the area under the posterior is within the stated range—in this case the σ notation is just a shorthand, and Gaussians don't really come into it.

Option (i) is a good idea if we can get a simple, even if approximate, answer. Option (ii) is safer, because it avoids a Gaussian approximation, but will almost certainly need a computer. We will find both options useful in later examples.

When we do want to approximate a posterior—or indeed any function $f(x)$—by a Gaussian $G(m,\sigma,x)$, there are two standard ways of doing so. The easy way is to simply set m and σ^2 to the mean and variance of $f(x)$. Alternatively, one can require $f(x)$ to have the same peak (i.e., m) as $G(m,\sigma,x)$, and require $\ln f(x)$ to have the same second derivative (i.e., $-\sigma^2$) as $\ln G(m,\sigma,x)$ at m. In other words, we can set

$$f'(m) = 0, \quad \sigma^{-2} = -\frac{f''(m)}{f(m)}. \tag{3.35}$$

We will use the formula (3.35) whenever possible, because the approximation is not affected by the tails of $f(x)$. But if (3.35) is intractable (because we cannot solve $f'(m) = 0$ easily) we will fall back on the easier first method.

EXAMPLE [Estimating m and σ from a sample] Say we are given some numbers $x_1, \ldots, x_N$ which (we happen to know) come from a Gaussian distribution, and we want to use these data to estimate m and σ.

We have for the likelihood (suppressing factors of 2π)

$$\mathrm{prob}(D \,|\, m, \sigma) \propto \sigma^{-N} \exp\left[-\tfrac{1}{2}\sigma^{-2} \sum\nolimits_i (x_i - m)^2\right]. \tag{3.36}$$

If we write

$$\overline{m} = \frac{1}{N} \sum\nolimits_i x_i, \qquad \overline{\sigma}^2 = \frac{1}{N} \sum\nolimits_i (x_i^2 - \overline{m}^2), \tag{3.37}$$

(the sample mean and variance), and rearrange the likelihood a bit, we get

$$\mathrm{prob}(D \,|\, m, \sigma) \propto \sigma^{-N} e^{-\frac{N}{2\sigma^2}\left[(m-\overline{m})^2 + \overline{\sigma}^2\right]}. \tag{3.38}$$

If we give m a flat prior then $\mathrm{prob}(m \,|\, D, \sigma)$ becomes a Gaussian with mean $\overline{m}$ and dispersion $\sigma/\sqrt{N}$, so $m = \overline{m} \pm \sigma/\sqrt{N}$. (We are using 1σ errors, as is the default.) The uncertainty depends on what we will estimate for σ, but since that's bound to be pretty close to $\overline{\sigma}$, we can safely say

$$m = \overline{m} \pm \overline{\sigma}/\sqrt{N}. \tag{3.39}$$

Estimating σ is a bit more complicated. First we marginalize out m, and then put a $1/\sigma$ prior on σ, getting the posterior

$$\mathrm{prob}(\sigma | D) \propto \sigma^{-N} e^{-\frac{N\overline{\sigma}^2}{2\sigma^2}}. \tag{3.40}$$

This is not a Gaussian in σ, but we can approximate it by one. Using approximation formula (3.35) we get

$$\sigma = \overline{\sigma} \pm \overline{\sigma}/\sqrt{2N}. \tag{3.41}$$

This answer is simple, but it isn't entirely satisfactory because it may admit negative values. So one may prefer something more elaborate if that is a worry.

Had we used a flat prior for σ, (3.41) would have been replaced by $\sigma = \sqrt{N/(N-1)}\,\overline{\sigma} \pm \sqrt{\frac{1}{2}N/(N-1)^2}\,\overline{\sigma}$. If, on the other hand, we had the $1/\sigma$ prior but knew m independently, we would have got $\sigma = \sqrt{N/(N+1)}\,\overline{\sigma} \pm \sqrt{\frac{1}{2}N/(N+1)^2}\,\overline{\sigma}$. □

One sometimes comes across the "error function" $\mathrm{erf}(x)$, which is essentially the probability of a number drawn from a Gaussian being less than some value. The definition is

$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2}\, dt, \tag{3.42}$$

so

$$\mathrm{prob}(y < x | \text{Gaussian}) = \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{x-m}{\sqrt{2}\,\sigma}\right)\right]. \tag{3.43}$$

There's also $\mathrm{erfc}(x) \equiv 1 - \mathrm{erf}(x)$.

PROBLEM 3.3: Digital cameras use a technology, originally developed for astronomical imaging, called a CCD (charge coupled device). The guts of a CCD is a matrix of pixels (picture elements, say 1024×1024 of them). During an exposure, each pixel collects a charge proportional to the number of photons reaching it. After the exposure, the collected charges are all measured, hence counting the number of photons. When working with a CCD one really is counting photons, and has to worry about counting statistics. The counts follow Poisson statistics, of course, but people generally use the Gaussian approximation. So if a pixel records N photons, it is taken to be drawn from a Gaussian distribution with $m = N$ and $\sigma = \sqrt{N}$.

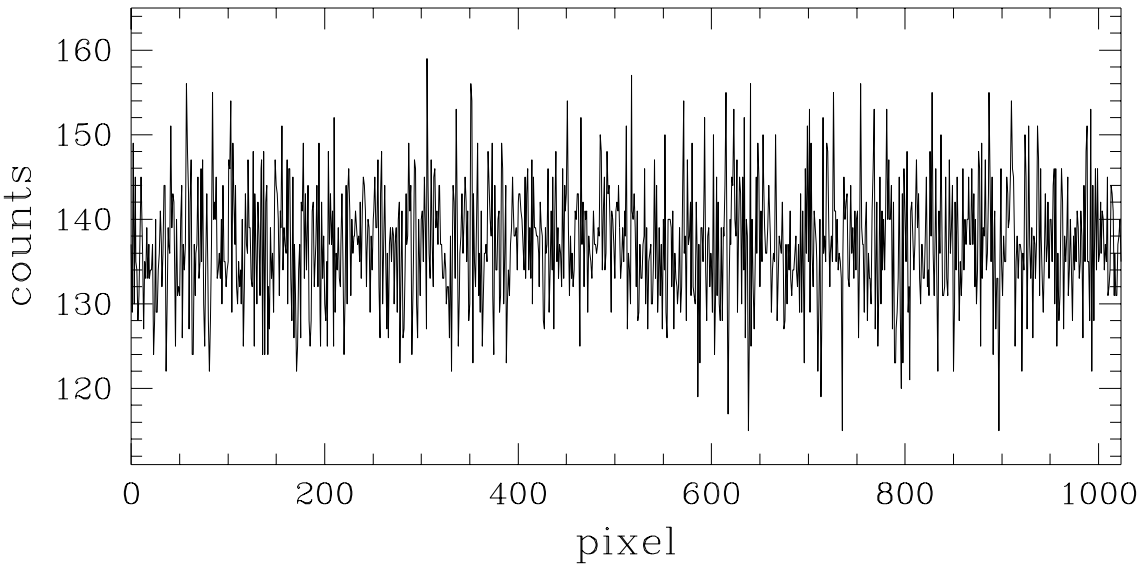

Figure 3.1: Counts along one row of an imaginary CCD.

The major source of instrumental error in CCDs comes because each pixel needs to have some initial charge before it can operate. In effect, there are some photons of internal origin. These (fictitious) internal photons can easily be calibrated for and their number subtracted out. But the internal photons are themselves subject to Poisson/Gaussian noise, which we need to know about. This is called readout noise.

To estimate the readout noise, one takes an exposure without opening the shutter, and estimates a Gaussian fit to the counts. Figure 3.1 is a plot with simulated counts for 1024 pixels along one row of a fictitious CCD. There's a constant term ('bias') added to all the counts, so it's only the variation in the counts which matters.

I showed the plot to the astronomer who first taught me the above, and asked him what he would estimate the readout noise to be, from this information. He looked at the plot for maybe 15 seconds, and said "I'd say it was between 6 and 7". How do you think he got that? **[2]**

PROBLEM 3.4: As well as being excellent photon detectors, CCDs are unfortunately also excellent detectors of cosmic rays. So any raw CCD image will have a few clusters of pixels with huge photon counts.

One way of getting around the problem of cosmic rays is to divide an exposure into a few shorter exposures and then take the pixel-by-pixel medians of the shorter exposures. (The median is very unlikely to come from an exposure during which the pixel in question was hit by a cosmic ray.)

Say we have five exposures, and consider one of the (vast majority of) pixels that were never hit by cosmic rays. The five counts for our pixel would represent five trials from the same Gaussian distribution, but the median of the five would not have a Gaussian distribution. What distribution would it have? **[3]**

PROBLEM 3.5: Alan M. Turing, one of the patron saints of computer science, wrote a philosophical article in 1950 on *Computing machinery and Intelligence.* In it he introduces what he calls the imitation game (now usually called the Turing test). In it, a human 'interrogator' is talking by teletype to two 'witnesses', one human and one a computer. The interrogator's job is to figure out which is which. The human witness's job is to help the interrogator, the computer's job is to confuse the interrogator (i.e., pretend to be the human). Turing proposes the ability (if achieved) to fool most human interrogators as an operational definition of intelligence.

Most of the article consists of Turing first considering, and then wittily rebutting, nine different objections to his proposal. The last of these is that humans are allegedly capable of ESP, but computers are not; Turing's discussion includes the following passage.

> Let us play the imitation game, using as witnesses a man who is good as a telepathic receiver, and a digital computer. The interrogator can ask such questions as 'What suit does the card in my right hand belong to?' The man by telepathy or clairvoyance gives the right answer 130 times out of 400 cards. The machine can only guess at random, and perhaps get 104 right, so the interrogator makes the right identification.

Nobody seems to know if Turing was serious in this passage, or just having the philosophers on. Certainly, '130 times out of 400 cards' sounds less than impressive. What do you think? **[2]**

4. Monte-Carlo Basics

With the material we have covered so far, we can write down likelihoods and posteriors in many kinds of examples. Mostly though, the probability distributions will depend on several parameters, and in some complicated way. So we need numerical methods to find peaks and widths of peaks, and to marginalize over some parameters.

In this chapter we will develop methods for generating a sample of ω from a probability distribution, say $\mathfrak{prob}(\omega\,|\,D)$. Of course, we can make this sample as large as we please, given enough computer time. If ω happens to be one-dimensional, we can then easily estimate the mean and variance, or percentile values, whatever. If ω is multi-dimensional, we consider one dimension at a time (*after* generating the multi-dimensional sample); this amounts to marginalizing over the other dimensions. In this way, we can estimate each parameter (with uncertainties) with all the other parameters marginalized out.

We need first to know a little about random number generators. For our purpose random number generators are functions on a computer that return a number with uniform probability in $[0, 1]$. They are quite sophisticated things, designed to make the output appear as truly probabilistic as possible. Actually, they are spewing out a pre-determined sequence of numbers, which is why some people say "pseudo-random" numbers. A state-of-knowledge view would be that if we don't know the sequence and can't, from the output, figure out anything about what is coming next, then it's random. Knuth's book[1] is the ultimate reference, but here are a couple of things to remember.

(a) The sequence a random number generator produces is determined by an initial 'seed'; usually, if you run the program again, you get the same sequence. If you want a different sequence, just skip a few numbers at the start of the program.

(b) A random sample of a uniform distribution is not completely uniform but has structure on all scales. As the sample gets larger, the structure gradually fades. If you take the output of a good random number generator and numerically take the power spectrum (mod-square of the Fourier transform), it will be flat.

If we have some $f(x)$, defined for x in $[0, 1]$ and $|f(x)| \leq 1$, it's easy to concoct random numbers distributed according to $f(x)$. We first call the uniform random number generator, say it returns r; we use r a fraction $f(r)$ of the time; otherwise we try again. This generalizes easily to any bounded f defined in a finite multi-dimensional space. It is called the rejection method. It always works (though if f has an integrable singularity or the region is infinite, one will need an extra transformation). But it can be very inefficient, especially in several dimensions.

A more efficient way, if tractable, is to go via the cumulative distribution function. In

[1] D.E. Knuth, *The Art of Computer Programming,* vol. 2 (Addison-Wesley 1979), which includes the gem "Random numbers should never be generated by a program written at random. Some theory should be used."

one dimension, consider

$$F(x) = \int_{-\infty}^{x} f(x')\,dx', \tag{4.1}$$

and for a uniform random number r, output $F^{-1}(r)$; it will be distributed according to $f(x)$. This can be generalized to higher dimensions, provided we can compute $F^{-1}(r)$ easily.

A variant of the inverse-cumulant method above is to divide the domain of x into K segments, each contributing equally to the cumulative distribution function, and then generate K random numbers at once, one in each segment. To make it sound more impressive, this procedure is called 'Latin hypercube sampling'.

PROBLEM 4.1: Consider the distribution function

$$f(r) \propto \frac{1}{r^2(1+r^2)}$$

where r is the radial coordinate in three dimensional space. How can we use the inverse-cumulant method to generate random space points (x, y, z) distributed according to $f(r)$? Suggest a method that will work for

$$g(r) \propto \frac{\sin^2(r)}{r^2(1+r^2)}.$$

(A description of the algorithms will do—no formal code-style notation required.) **[2]**

The inverse-cumulant and rejection methods won't do, though, for most of our applications. The distribution functions will generally be far too complicated for inverse-cumulants, and if they are sharply peaked in several dimensions, rejection is too inefficient. For such problems there is a powerful set of iterative schemes, sometimes called 'Markov chain Monte-Carlo algorithms'. The idea is that the function to be sampled, say $f(x)$, is approached through a sequence of approximations $f_n(x)$. These $f_n(x)$ are distribution functions (in multi-dimensions), not samples. Each $f_n(x)$ is related to the next iterate $f_{n+1}(x)$ by some "transition probabilities" $p(x \to x')$, thus:

$$f_{n+1}(x) = \sum_{x'} f_n(x')\,p(x' \to x), \quad \text{where} \sum_{x} p(x' \to x) = 1. \tag{4.2}$$

The transition probabilities (normalized as above) don't depend on n, hence the sequence $f_1, f_2, \ldots$ is a Markov chain. The key to making $f_n(x)$ converge to $f(x)$ is to choose the right transition probabilities, and they can be anything that satisfies

$$f(x)\,p(x \to x') = f(x')\,p(x' \to x), \tag{4.3}$$

known as detailed balance.

DIGRESSION [Convergence of the iterations] If $f_n(x)$ has already converged to $f(x)$, further iterates will stay there, because (using 4.3 and 4.2)

$$\begin{aligned} f_{n+1}(x) &= \sum_{x'} f_n(x')\, p(x' \to x) \\ &= \sum_{x'} f_n(x)\, p(x \to x') = f_n(x). \end{aligned} \tag{4.4}$$

To see why $f_n(x)$ will approach $f(x)$, we take

$$\begin{aligned} |f(x) - f_{n+1}(x)| &= \Big| f(x) - \sum_{x'} f_n(x')\, p(x' \to x) \Big| \\ &= \Big| \sum_{x'} \big(f(x') - f_n(x') \big)\, p(x' \to x) \Big| \\ &\le \sum_{x'} \big| f(x') - f_n(x') \big|\, p(x' \to x), \end{aligned} \tag{4.5}$$

where the last step uses the triangle inequality. Summing now over x, we get

$$\sum_x |f(x) - f_{n+1}(x)| \le \sum_{x'} \big| f(x') - f_n(x') \big|, \tag{4.6}$$

implying convergence. □

We can use the convergence property of $f_n(x)$ to generate a Markov chain $x_1, x_2, \ldots$ that eventually samples $f(x)$, as follows. At any x_n, we pick a trial x'_n, and set

$$x_{n+1} = \begin{cases} x'_n, & \text{with } p(x_n \to x'_n), \\ x_n, & \text{otherwise}, \end{cases} \tag{4.7}$$

where the transition probability $p(x_n \to x'_n)$ satisfies the detailed balance condition (4.3). In other words, we transit to x'_n with probability $p(x_n \to x'_n)$, otherwise we stay at x_n. Then x_1 (which we choose at random) may be considered a sample of $f_1(x)$, x_2 a sample of $f_2(x)$, and so on. For large-enough N, the chain $x_N, x_{N+1}, \ldots$ will be a sample of $f(x)$.

The easiest choice for the transition probability is

$$p(x \to x') = \min\left[f(x')/f(x), 1 \right], \tag{4.8}$$

and the resulting algorithm is called the Metropolis algorithm. There are many other algorithms which improve on it in various ways,[1] but plain Metropolis is still the most popular.

There are some technical issues to think about when implementing the Metropolis algorithm. The first is how to pick the new trial x'. One is as a step: $x' = x + r\delta x$, where

[1] The original context of these algorithms was condensed-matter physics, and the most demanding applications are still in that area—see for example, J.J. Binney, N.J. Dowrick, A. Fisher, & M.E.J. Newman, *The Theory of Critical Phenonena,* (Oxford University Press 1992).

δx is a stepsize and r a random number in $[-1, 1]$. Ultimate convergence won't depend on the choice of δx, but efficiency will. The second issue is how long to iterate for. Now Metropolis operates by accepting all trial steps that would increase f and accepting some but not all trial steps that would decrease f. So there will be an initial period (sometimes called 'burn-in') when the algorithm seeks out the maximal region of $f(x)$, and a later 'equilibrium' period when it wanders around in the maximal region, with occasional brief forays into lower regions. Our sample should come from the equilibrium period, but how do we recognize that it has begun? There are more sophisticated things one can do, but one simple way is to look for the disappearance of an increasing trend in, say, 100 iterations.

PROBLEM 4.2: Consider our friend

$$f(r) \propto \frac{1}{r^2(1+r^2)}$$

again, and this time write a Metropolis program to generate a sample of (x, y, z) distributed according to $f(r)$. Plot a histogram of the x values in $[-5, 5]$, and on it overplot the exact marginalized distribution, which is

$$\int f(r)\, dy\, dz = \frac{1}{2\pi} \ln\left(1 + \frac{1}{x^2}\right).$$

The size of the Metropolis steps, the size of the sample, and the scaling for the histogram are all up to you, but you should choose them to show that the histogram is a believable approximation to the exact distribution. **[4]**

5. Least Squares

The most bread-and-butter thing in data analysis is when we have a signal $F(\omega, t)$ depending on some parameters ω and an independent variable t, and some Gaussian noise $n(\sigma(t))$, and we measure the sum at certain discrete points. In other words, the data are

$$d_i = F(\omega, t_i) + n(\sigma(t_i)), \quad i = 1, \ldots, N. \tag{5.1}$$

The form of $F(\omega, t)$ is known, and the noise $n(\sigma(t))$ has a Gaussian distribution with zero mean and known time-dependent dispersion. The problem is to infer the values of the ω. This (for reasons that will be obvious shortly) is called least-squares. The key requirements are that the noise must be Gaussian and additive on the signal—not a universal situation, but a very common one, and the reason for spending a lot of effort on this problem in its many forms.[1]

By assumption and equation (5.1), $d_i - F(\omega, t_i)$ will have a Gaussian distribution, hence the likelihood is

$$\begin{aligned} \mathfrak{prob}(D \,|\, \omega) &= (2\pi)^{-N/2} \left(\prod\nolimits_i \sigma_i^{-1} \right) \mathfrak{exp}\left[-\sum\nolimits_i Q_i^2 \right], \\ Q_i^2 &= \frac{(d_i - F(\omega, t_i))^2}{2\sigma^2(t_i)}. \end{aligned} \tag{5.2}$$

By rescaling

$$d_i \leftarrow \frac{\sigma}{\sigma(t_i)} d_i, \quad F_i(\omega) \leftarrow \frac{\sigma}{\sigma(t_i)} F(\omega, t_i) \tag{5.3}$$

we can write

$$\begin{aligned} \mathfrak{prob}(D \,|\, \omega, \sigma) &= (2\pi)^{-N/2} \sigma^{-N} \, \mathfrak{exp}\left[-\frac{Q^2}{2\sigma^2} \right], \\ Q^2 &= \sum\nolimits_i (d_i - F_i(\omega))^2. \end{aligned} \tag{5.4}$$

In (5.4) we left the overall noise scale σ as a parameter; sometimes it is known, sometimes it must be inferred along with the ω.

Typically, some of the parameters will enter F linearly and some nonlinearly. For example if F is a tenth degree polynomial, we have $F_i(c_n) = \sum_{n=0}^{10} c_n t_i^n$ and all the parameters c_n are linear. Or F_i might be something like $c_0 + c_1 \cos(\alpha t_i) + c_2 \cos(2\alpha t_i)$, with one non-linear and three linear parameters. Linear parameters are much easier to deal with, so let us change our notation slightly to use $a_1, \ldots, a_L$ for the linear parameters (amplitudes) and ω for all the nonlinear parameters. Thus we replace $F_i(\omega)$ by

[1] This chapter is a mixture of well-known and comparatively little-known results. The parameter fitting and model comparison parts are based mostly on G.L. Bretthorst, *Bayesian Spectrum Analysis and Parameter Estimation* (Springer-Verlag 1988).

$\sum_{l=1}^{L} a_l f_{li}(\omega)$, and the likelihood (5.4) by

$$\begin{aligned} \mathrm{prob}(D \,|\, a_l, \omega, \sigma) &= (2\pi)^{-N/2} \sigma^{-N} \exp\left[-\frac{Q^2}{2\sigma^2}\right], \\ Q^2 &= \sum_{i=1}^{N} \left(d_i - \sum\nolimits_{l=1}^{L} a_l \, f_{li}(\omega) \right)^2 . \end{aligned} \tag{5.5}$$

Typically, the sum over l (the amplitudes) will have just a few terms but the sum over i (the data) will have many terms.

EXAMPLE [Fitting a straight line] Let us work through this simple example before dealing with the general case. We have some data $y_1, \ldots, y_N$ measured at x_i with Gaussian errors σ_i, and we want to fit these to a straight line $y = mx + c$. There are no errors in x.

First, as usual, we rescale each y_i and the corresponding x_i to change all the σ_i to σ. The likelihood is then

$$\begin{aligned} \mathrm{prob}(D \,|\, m, c) &\propto \sigma^{-N} \exp\left[-\frac{Q^2}{2\sigma^2}\right], \\ Q^2 &= \sum_{i=1}^{N} (y_i - m x_i - c)^2 . \end{aligned} \tag{5.6}$$

Let us put uniform priors on the parameters m and c. Then the posterior probability $\mathrm{prob}(m, c \,|\, D)$ is proportional to the likelihood.

The posterior, as well as being Gaussian in the data, is also a two-dimensional Gaussian in m, c; the reason, as we can verify from (5.6), is that m, c are linear parameters (amplitudes). To find the values $(\overline{m}, \overline{c}$, say) that maximize the posterior we equate the relevant partial derivatives of Q^2 to zero. Doing this, we have

$$\begin{pmatrix} \sum_i x_i^2 & \sum_i x_i \\ \sum_i x_i & N \end{pmatrix} \begin{pmatrix} \overline{m} \\ \overline{c} \end{pmatrix} = \begin{pmatrix} \sum_i x_i y_i \\ \sum_i y_i \end{pmatrix} \tag{5.7}$$

a matrix equation whose form

$$\mathbf{C}^{-1} \cdot \mathbf{a} = \mathbf{P} \tag{5.8}$$

we will meet again. Solving the matrix equation for $(\overline{m}, \overline{c})$, we get

$$\begin{aligned} \overline{m} &= \frac{N \sum_i x_i y_i - \sum_i x_i \sum_i y_i}{N \sum_i x_i^2 - \left(\sum_i x_i\right)^2}, \\ \overline{c} &= \frac{\sum_i x_i^2 \sum_i y_i - \sum_i x_i y_i \sum_i x_i}{N \sum_i x_i^2 - \left(\sum_i x_i\right)^2}. \end{aligned} \tag{5.9}$$

Expanded around its minimum, Q^2 is just a constant plus its quadratic part, hence the posterior is

$$\begin{aligned} \mathrm{prob}(m, c \,|\, D) &\propto \sigma^{-N} \exp\left[-\frac{Q^2}{2\sigma^2}\right], \\ Q^2 &= \mathrm{const} + (m - \overline{m})^2 \textstyle\sum_i x_i^2 + (c - \overline{c})^2 N \\ &\quad + 2(m - \overline{m})(c - \overline{c}) \textstyle\sum_i x_i . \end{aligned} \tag{5.10}$$

To find the uncertainty in either m or c, we marginalize out the other from the posterior. The marginal posteriors turn out to be (using formula M.11 on page 71) Gaussian as well

$$\begin{aligned}
&\mathrm{prob}(m\,|\,D) \propto \exp\left[-\frac{(m-\overline{m})^2}{2\sigma_m^2}\right] \quad \sigma_m^2 = \frac{\sigma^2 N}{\Delta},\\
&\mathrm{prob}(c\,|\,D) \propto \exp\left[-\frac{(c-\overline{c})^2}{2\sigma_c^2}\right] \qquad \sigma_c^2 = \frac{\sigma^2 \sum_i x_i^2}{\Delta}\\
&\Delta = N\sum_i x_i^2 - \left(\sum_i x_i\right)^2.
\end{aligned} \tag{5.11}$$

Though these expressions look messy when written out in full, all one really needs to remember is the matrix equation (5.7/5.8). The least-squares values (5.9) are the solution of the matrix equation, and the expressions for σ_m^2 and σ_c^2 in (5.11) are the diagonal elements of $\sigma^2\mathbf{C}$. □

PROBLEM 5.1: A popular way of estimating errors in problems where it's difficult to do the probability theory is called bootstrap. This problem is to test bootstrap in a simple context, by comparing its results with what we have calculated from probability theory.

In the following table y is linear in x but with some Gaussian noise.

x	y	x	y	x	y	x	y
0.023	0.977	0.034	0.928	0.047	0.951	0.059	1.206
0.070	1.094	0.080	1.002	0.082	0.769	0.087	0.979
0.099	1.043	0.129	0.686	0.176	0.638	0.187	0.808
0.190	0.728	0.221	0.760	0.233	0.770	0.246	0.869
0.312	0.631	0.397	0.575	0.399	0.735	0.404	0.571
0.461	0.679	0.487	0.415	0.530	0.299	0.565	0.410
0.574	0.509	0.669	0.291	0.706	0.167	0.850	-0.067
0.853	0.023	0.867	0.128	0.924	0.332	1.000	0.045

Fit for the coefficients in $y = mx + c$. Let's say $\overline{m}$ and $\overline{c}$ are the estimated values. Then invent a new data set, by sampling the given data N times (32 times in this case) *with replacement*, and fit again for m and c—say you get $\hat{m}$ and $\hat{c}$. Repeat a number of times, and calculate the matrix

$$\begin{pmatrix} \langle(\hat{m}-\overline{m})(\hat{m}-\overline{m})\rangle & \langle(\hat{m}-\overline{m})(\hat{c}-\overline{c})\rangle \\ \langle(\hat{c}-\overline{c})(\hat{m}-\overline{m})\rangle & \langle(\hat{c}-\overline{c})(\hat{c}-\overline{c})\rangle \end{pmatrix}$$

(where the averages are over the generated data sets). This is the procedure for bootstrap, and the matrix is taken as an estimate of the matrix $\sigma^2\mathbf{C}$. **[3]**

We continue with the general case. Let us rewrite (5.5) in a more convenient notation. If we define

$$\begin{aligned}
&d^2 = \sum_{i=1}^{N} d_i^2,\\
&P_l = \sum_{i=1}^{N} d_i f_{li}(\omega), \quad C_{kl}^{-1} = \sum_{i=1}^{N} f_{ki}(\omega) f_{li}(\omega),
\end{aligned} \tag{5.12}$$

then the likelihood (5.5), after expanding and rearranging, becomes

$$\begin{aligned} \mathfrak{prob}(D \,|\, a_l, \omega, \sigma) &= (2\pi)^{-N/2} \sigma^{-N} \, \mathfrak{exp}\left[-\frac{Q^2}{2\sigma^2}\right], \\ Q^2 &= d^2 + \sum\nolimits_{k,l} a_k a_l C^{-1}_{kl} - 2 \sum\nolimits_l a_l P_l. \end{aligned} \tag{5.13}$$

Equation (5.13) is just asking to be written in matrix notation, so let us introduce

$$\mathbf{a} \leftarrow a_l, \quad \mathbf{P} \leftarrow P_l, \quad \mathbf{C} \leftarrow C_{kl}, \tag{5.14}$$

using which (5.13) becomes

$$\begin{aligned} \mathfrak{prob}(D \,|\, \mathbf{a}, \omega, \sigma) &= (2\pi)^{-N/2} \sigma^{-N} \, \mathfrak{exp}\left[-\frac{Q^2}{2\sigma^2}\right] \\ Q^2 &= d^2 + \mathbf{a}^T \cdot \mathbf{C}^{-1} \cdot \mathbf{a} - 2\mathbf{P}^T \cdot \mathbf{a}. \end{aligned} \tag{5.15}$$

Let us pause a moment to remark on $\mathbf{C}, \mathbf{a}, \mathbf{P}$, since we will meet them many times below. $\mathbf{C}$ is an $L \times L$ matrix depending on the nonlinear parameters but not the data; $\sigma^2\mathbf{C}$ is called the covariance matrix. $\mathbf{P}$ is a column vector, and a sort of inner product of data and model. And $\mathbf{a}$ is the vector of L amplitudes to be inferred. The likelihood is completely specified by $\mathbf{C}, \mathbf{P}, d^2$, and σ^2, and we will not need to refer explicitly to the data any more.

Completing the squares inside Q^2, and using the fact that $\mathbf{C}$ is symmetric, we can rewrite (5.15) as

$$\begin{aligned} \mathfrak{prob}(D \,|\, \mathbf{a}, \omega, \sigma) &= (2\pi)^{-N/2} \sigma^{-N} \, \mathfrak{exp}\left[-\frac{Q^2}{2\sigma^2}\right] \\ Q^2 &= d^2 - \mathbf{P}^T \cdot \mathbf{C} \cdot \mathbf{P} + (\mathbf{a} - \mathbf{C} \cdot \mathbf{P})^T \cdot \mathbf{C}^{-1} \cdot (\mathbf{a} - \mathbf{C} \cdot \mathbf{P}). \end{aligned} \tag{5.16}$$

We see that the likelihood is an L-dimensional Gaussian in the amplitudes $\mathbf{a}$. Let us put a flat prior on $\mathbf{a}$, so the posterior is also Gaussian in $\mathbf{a}$. We can then easily write down the mean and (using the formula M.14 from page 71) dispersion:

$$\langle \mathbf{a} \rangle = \mathbf{C} \cdot \mathbf{P}, \quad \left\langle (\mathbf{a} - \langle \mathbf{a} \rangle)\,(\mathbf{a} - \langle \mathbf{a} \rangle)^T \right\rangle = \sigma^2 \mathbf{C}. \tag{5.17}$$

In other words, $\mathbf{a}$ has an L-dimensional Gaussian distribution with mean $\mathbf{C} \cdot \mathbf{P}$ and dispersion $\sigma^2\mathbf{C}$. The diagonal elements of the covariance matrix $\sigma^2\mathbf{C}$ give the dispersions in $\mathbf{a}$ and the off-diagonal elements indicate how correlated the components of $\mathbf{a}$ are. (Recall that $\mathbf{C}$ doesn't depend on the data.)

If we change from $\mathbf{a}$ to some linearly transformed amplitudes $\mathbf{b}$, the covariance of $\mathbf{b}$ will be related to the covariance of $\mathbf{a}$ via the transformation matrix:

$$\begin{aligned}&\left\langle (\mathbf{b} - \langle \mathbf{b} \rangle) (\mathbf{b} - \langle \mathbf{b} \rangle)^T \right\rangle = \\ &\left(\frac{\partial \mathbf{b}}{\partial \mathbf{a}}\right) \left\langle (\mathbf{a} - \langle \mathbf{a} \rangle) (\mathbf{a} - \langle \mathbf{a} \rangle)^T \right\rangle \left(\frac{\partial \mathbf{b}}{\partial \mathbf{a}}\right)^T .\end{aligned} \tag{5.18}$$

Having extracted as much as we can about the amplitudes, we now get them out of the way by marginalizing. Using (M.13) from page 71 gives us

$$\begin{aligned}\mathfrak{prob}(D \,|\, \omega, \sigma) = (2\pi)^{(L-N)/2} \sigma^{L-N} \times \\ |\det \mathbf{C}|^{1/2} \exp\left[-\frac{d^2 - \mathbf{P}^T \cdot \mathbf{C} \cdot \mathbf{P}}{2\sigma^2}\right].\end{aligned} \tag{5.19}$$

Now we have to deal with the nonlinear parameters. In general (5.19) will have to be investigated numerically, by Monte-Carlo. But if $\mathbf{P}^T \cdot \mathbf{C} \cdot \mathbf{P}$ has a sharp maximum, at $\overline{\omega}$ say, then it is reasonable to linearize about $\overline{\omega}$:

$$F_i(\omega) \simeq F_i(\overline{\omega}) + (\omega_k - \overline{\omega}_k) \left(\frac{\partial F_i(\omega)}{\partial \omega_k}\right)_{\overline{\omega}} . \tag{5.20}$$

In this approximation $(\omega_k - \overline{\omega}_k)$ behave like more amplitudes. So we can give uncertainties on ω by treating D_{kl}, where

$$D^{-1}_{kl} = \sum_{i=1}^{N} \left(\frac{\partial F_i}{\partial \omega_k}\right)_{\overline{\omega}} \left(\frac{\partial F_i}{\partial \omega_l}\right)_{\overline{\omega}} . \tag{5.21}$$

as a covariance matrix (also called a Fisher matrix in this context). If we transform from ω to some other parameters μ, and $\overline{\mu}$ corresponds to $\overline{\omega}$, the covariance matrix in terms of μ is given analogously to (5.18):

$$\begin{aligned}&\langle (\mu_k - \overline{\mu}_k)(\mu_l - \overline{\mu}_l) \rangle = \\ &\sum_{pq} \left(\frac{\partial \mu_k}{\partial \omega_p}\right) \left(\frac{\partial \mu_l}{\partial \omega_q}\right) \langle (\omega_p - \overline{\omega}_p)(\omega_q - \overline{\omega}_q) \rangle .\end{aligned} \tag{5.22}$$

This is called the formula for propagation of errors, and is much used, especially in its diagonal form

$$\langle (\mu_k - \overline{\mu}_k)^2 \rangle = \sum_{p} \left(\frac{\partial \mu_k}{\partial \omega_p}\right)^2 \langle (\omega_p - \overline{\omega}_p)^2 \rangle , \tag{5.23}$$

but we should note that it is valid only in the linearized approximation. A simple corollary is that if there are N independent least-squares estimates of some quantity, all with equal error-bars, then the error-bar of the combined estimate will be down by $\sqrt{N}$.

PROBLEM 5.2: Suppose we have several data points $d_1, d_2, \ldots, d_K$ at the same value of the independent variable t, amongst other data points at other t. In least squares, should one include the K data points individually, or combine them into a single point with a smaller error bar? **[2]**

EXAMPLE [Error propagation on square roots] In practice, people use the error propagation formula (5.23) quite freely in nonlinear situations. For example, if ω has been estimated as $\overline{\omega} \pm \Delta\omega$, then if μ is defined as $\sqrt{\omega}$, one gets from (5.23):

$$\mu = \sqrt{\overline{\omega}} \pm \frac{\Delta\omega}{2\sqrt{\overline{\omega}}}.$$

The key assumption, of course, is that the data are good enough that $\Delta\omega \ll \omega$. □

PROBLEM 5.3: The simple error propagation formula isn't always good enough. If we are going to take $\sqrt{\omega}$, then we really have a prior that requires $\omega \geq 0$. Let us set the prior to zero for $\omega < 0$ and a constant otherwise. The likelihood could be Gaussian in ω (with mean $\overline{\omega}$ and dispersion $\Delta\omega$ say). Quite possibly $\overline{\omega} < 0$, but that does not prevent one from deriving a perfectly good posterior for $\mu = \sqrt{\omega}$. In fact, one can derive a Gaussian approximation to the posterior for μ, and this has mean $\overline{\mu}$ and dispersion $\Delta\mu$ given by

$$\overline{\mu}^2 = \tfrac{1}{2}\big(\overline{\omega} + \sqrt{\overline{\omega}^2 + 2\Delta\omega^2}\big), \quad \Delta\mu^2 = \frac{\Delta\omega^2}{2\sqrt{\overline{\omega}^2 + 2\Delta\omega^2}}.$$

Derive this approximation. **[3]**

One doesn't always have an estimate for the noise dispersions σ_i. What one then does is (i) assume the σ_i equal some σ and (ii) take σ as a scale parameter and marginalize it out. Multiplying (5.19) by a $1/\sigma$ prior and integrating using (M.5) we get

$$\begin{aligned} \mathfrak{prob}(D\,|\,\omega) &= \int \mathfrak{prob}(D\,|\,\omega,\sigma)\,\sigma^{-1}\,d\sigma \\ &= \pi^{(L-N)/2}\,|\,\mathfrak{det}\,\mathbf{C}|^{1/2} \times \\ &\qquad \Gamma\big(\tfrac{1}{2}(N-L)\big)\big(d^2 - \mathbf{P}^T\cdot\mathbf{C}\cdot\mathbf{P}\big)^{(L-N)/2}. \end{aligned} \tag{5.24}$$

This is usually called a Student-t distribution.[1] We can get an estimate for σ^2 too, by taking an expectation

$$\langle\sigma^2\rangle = \frac{\int \sigma^2\,\mathfrak{prob}(D\,|\,\omega,\sigma)\,\sigma^{-1}\,d\sigma}{\int \mathfrak{prob}(D\,|\,\omega,\sigma)\,\sigma^{-1}\,d\sigma} \tag{5.25}$$

gives (adapting the calculation of 5.24)

$$\langle\sigma^2\rangle = \left(\frac{d^2 - \mathbf{P}^T\cdot\mathbf{C}\cdot\mathbf{P}}{N-L-2}\right). \tag{5.26}$$

[1] The real name of the person who derived something of this form in a somewhat different context, and published it under the name of 'Student', is known to have been W.S. Gosset, but nobody calls it the Gosset distribution.

PROBLEM 5.4: In most of this chapter we assume that there are no errors in an independent variable, i.e., in the t_i in equation (5.1). This problem is about the simplest case when there *are* errors in the independent variable.

Suppose, we have some data (x_i, y_i), where both the x_i and the y_i have errors. For simplicity we'll assume the noise dispersions in x_i and y_i are all unity. (This can always be arranged by rescaling.) We want to fit a straight line $y = mx + c$.

The likelihood function is no longer (5.6), and your job is to work out what it is. To do this, first consider the likelihood for one datum, $\mathrm{prob}(x_1, y_1 \,|\, m, c)$. This is going to depend on the noise in x and y, or the distance of (x_1, y_1) from the corresponding noiseless point on the line. But we don't know what that noiseless point is, we just know it lies somewhere on $y = mx + c$. **[4]**

So much for fitting a model by least-squares. What if we have several models, and want to find which one the data favour? For example, we might be fitting polynomials of different orders, and need to find which order is favoured. We need the global likelihoods; that is, we have to marginalize out the parameters, but with normalized priors. For very nonlinear parameter dependences we have to do that numerically, but for linear and linearizable parameters the marginalization can be done exactly.

For parameter fitting we took a flat prior on $\mathbf{a}$. Now we modify that to a Gaussian prior with a very large dispersion; it's still nearly flat for $\mathbf{a}$ values having non-negligible posterior, so our parameter estimates aren't affected. We take

$$\mathrm{prob}(\mathbf{a} \,|\, \delta) = (2\pi)^{-L/2} \delta^{-L} |\det \mathbf{C}|^{-1/2} \exp\left[-\frac{\mathbf{a}^T \cdot \mathbf{C}^{-1} \cdot \mathbf{a}}{2\delta^2}\right], \tag{5.27}$$

where δ is a new parameter assumed $\gg \sigma$. Then the exponent in $\mathrm{prob}(D \,|\, \mathbf{a}, \omega, \sigma) \times \mathrm{prob}(\mathbf{a} \,|\, \delta)$ has the same form as (5.15), but with the substitution

$$\mathbf{C}^{-1} \leftarrow \mathbf{C}^{-1}(1 + \sigma^2/\delta^2). \tag{5.28}$$

Marginalizing out the amplitudes we get, in place of (5.19):

$$\begin{aligned} \mathrm{prob}(D \,|\, \omega, \sigma, \delta) = (2\pi)^{-N/2} \delta^{-L} \sigma^{L-N} (1 + \sigma^2/\delta^2)^{-L/2} \times \\ \exp\left[-\left(\frac{d^2 - \mathbf{P}^T \cdot \mathbf{C} \cdot \mathbf{P}}{2\sigma^2} + \frac{\mathbf{P}^T \cdot \mathbf{C} \cdot \mathbf{P}}{2\delta^2}\right)\right]. \end{aligned} \tag{5.29}$$

Since we assumed $\delta \gg \sigma$, we can discard the factor of $(1 + \sigma^2/\delta^2)^{-L/2}$. Now we have to marginalize out δ; we do our usual trick (cf. equation 2.12 on page 19) of assigning a Jeffreys prior with cutoffs:

$$\mathrm{prob}(\delta) = \frac{\delta^{-1}}{\ln(\delta_{\max}/\delta_{\min})}, \text{ or (say) } \frac{\delta^{-1}}{\Lambda_\delta}. \tag{5.30}$$

We then marginalize out δ by integrating—the limits of integration should properly be $\delta_{\min}$ to $\delta_{\max}$, but we further assume that we can approximate the limits as 0 to ∞ with negligible error. This gives

$$\begin{aligned}\mathfrak{prob}(D\,|\,\omega,\sigma) &= \Lambda_\delta^{-1}(2\pi)^{-N/2}2^{L/2}\sigma^{L-N}\Gamma\left(\tfrac{1}{2}L\right)\times \\ &\left(\mathbf{P}^T\cdot\mathbf{C}\cdot\mathbf{P}\right)^{-L/2}\mathfrak{exp}\left[-\tfrac{1}{2}\sigma^{-2}\left(d^2-\mathbf{P}^T\cdot\mathbf{C}\cdot\mathbf{P}\right)\right].\end{aligned} \tag{5.31}$$

Marginalizing σ out in similar fashion gives

$$\begin{aligned}\mathfrak{prob}(D\,|\,\omega) &= \Lambda_\delta^{-1}\Lambda_\sigma^{-1}\pi^{-N/2}\times \\ &\Gamma\left(\tfrac{1}{2}L\right)\Gamma\left(\tfrac{1}{2}(N-L)\right)\left(\mathbf{P}^T\cdot\mathbf{C}\cdot\mathbf{P}\right)^{-L/2}\left(d^2-\mathbf{P}^T\cdot\mathbf{C}\cdot\mathbf{P}\right)^{(L-N)/2}.\end{aligned} \tag{5.32}$$

For model comparison, one would take the ratios of the expressions in (5.31) or (5.32) for two different models; provided both models had at least one amplitude, the Λ factors would cancel.

PROBLEM 5.5: In the following table the y-values are a polynomial in x plus Gaussian noise.

x	y	x	y	x	y	x	y
0.023	0.977	0.034	0.949	0.047	0.943	0.059	0.948
0.070	0.924	0.080	0.914	0.082	0.920	0.087	0.919
0.099	0.916	0.129	0.886	0.176	0.842	0.187	0.840
0.190	0.849	0.221	0.827	0.233	0.787	0.246	0.789
0.312	0.757	0.397	0.697	0.399	0.696	0.404	0.713
0.461	0.657	0.487	0.643	0.530	0.609	0.565	0.565
0.574	0.549	0.669	0.484	0.706	0.433	0.850	0.251
0.853	0.245	0.867	0.238	0.924	0.151	1.000	-0.016

What is the degree of the polynomial? **[4]**

Finally, we have to test the goodness of fit, that is, we have to ask whether the data at hand could plausibly have come from the model we are fitting. After all, if none of the models we are fitting is actually correct, neither the parameter estimates nor the error bars on them mean very much.

We need a statistic to measure goodness of fit. A natural choice (though still an ad hoc choice—see the discussion on page 10) is the logarithm of the likelihood. Going back to equation (5.1) and assuming the σ_i are known, we define

$$\chi^2 = \sum_{i=1}^{N}\frac{[d_i - F_i(\omega)]^2}{\sigma_i^2}, \tag{5.33}$$

and from (5.2) the likelihood is

$$\mathfrak{prob}(D \,|\, \omega, M) \propto \mathfrak{exp}\left[-\tfrac{1}{2}\chi^2\right]. \tag{5.34}$$

To use χ^2 as the goodness-of-fit statistic, we need its p-value: the probability, given ω in some model, that a random data set will fit less well than the actual data set (see page 10), or

$$1 - \mathfrak{prob}(\chi^2 < \chi^2_D \,|\, \omega), \tag{5.35}$$

where χ^2_D is what the actual data set gives. If $\mathfrak{prob}(\chi^2 < \chi^2_D \,|\, \omega)$ is very close to 1, it means the data are a very bad fit. For example, if $\mathfrak{prob}(\chi^2 < \chi^2_D \,|\, \omega) > 0.99$, the model is said to be rejected with 99% confidence.

The χ^2 test is the most-used goodness-of-fit test, and it helps that $\mathfrak{prob}(\chi^2 < \chi^2_D \,|\, \omega)$ is fairly easy to calculate, at least approximately. To start with, note that the likelihood is a density in N-dimensional data space. The peak is at $\chi = 0$, and outside of that the density falls as $e^{-\frac{1}{2}\chi^2}$. Thus, if we consider χ as a radial coordinate in N-dimensions, the density depends only on radius. We can calculate that density as a function directly of radius, or $\mathfrak{prob}(\chi^2 \,|\, \omega)$. Since the volume element in N dimensions is $\propto \chi^{N-1}\, d\chi$, and $\propto (\chi^2)^{N/2-1}\, d\chi^2$, we have

$$\mathfrak{prob}(\chi^2 \,|\, \omega) \propto (\chi^2)^{N/2-1}\, \mathfrak{exp}\left(-\tfrac{1}{2}\chi^2\right). \tag{5.36}$$

Note: by convention, $\mathfrak{prob}(\chi^2 \,|\, \omega)$ is the probability density with respect to χ^2, not χ.

Actually, (5.36) applies only if the parameters are fixed—if we fit L amplitudes (including linearizable nonlinear parameters) then we effectively remove L of the N dimensions over which the data can vary independently. In that case

$$\mathfrak{prob}(\chi^2 \,|\, \omega) \propto (\chi^2)^{\nu/2-1}\, \mathfrak{exp}\left(-\tfrac{1}{2}\chi^2\right), \quad \nu = N - L. \tag{5.37}$$

Here ν is called the number of "degrees of freedom" and χ^2/ν is sometimes called "reduced χ^2". There is no analogous prescription for adjusting for nonlinear parameters that cannot be linearized.

We can derive a reasonably good Gaussian approximation for the χ^2 distribution (5.37). Using the approximation formula (3.35) from page 27 we get

$$\mathfrak{prob}(\chi^2 \,|\, \omega) \simeq \frac{1}{\sqrt{4\pi\mu}}\, \mathfrak{exp}\left[-\frac{(\chi^2 - \mu)^2}{4\mu}\right] \tag{5.38}$$

where

$$\mu = N - L - 2. \tag{5.39}$$

The content of (5.38) is that χ^2 should be close to μ; if it more than $5\sqrt{\mu}$ away something is fishy; it is more useful to know this fact than the value of $\mathfrak{prob}(\chi^2 < \chi^2_D \,|\, \omega)$. Nevertheless, the latter can be computed, by writing

$$\mathfrak{prob}(\chi^2 < \chi^2_D \,|\, \omega) = \frac{\int_0^{\chi^2_D/2} x^{\mu/2-1} e^{-x}\, dx}{\int_0^{\infty} x^{\mu/2-1} e^{-x}\, dx} \tag{5.40}$$

and observing that the right hand side is the definition of an incomplete Γ function. Thus

$$\mathfrak{prob}(\chi^2 < \chi^2_D \,|\, \omega) = \Gamma\left(\tfrac{1}{2}\mu, \tfrac{1}{2}\chi^2_D\right) \tag{5.41}$$

One can only apply the χ^2 test if the σ_i are known. If σ is estimated by equation (5.26), that just sets χ^2 at its expectation value. Interestingly, μ in (5.39) is the same as the denominator in (5.26). To test for acceptability of a fit for the case of unknown σ, we would have to consider the set of residuals $d_i - F_i(\omega)$, and ask whether these could plausibly have been drawn from a Gaussian distribution with $m = 0$ and σ^2 given by (5.26). That's part of the material of the next chapter.

PROBLEM 5.6: What is the analogue of χ^2 if the data are counts n_i from a Poisson distribution with mean m? (The answer is known as the Cash statistic.) **[1]**

PROBLEM 5.7: [Tremaine's paradox] For a given model, we have from (5.34):

$$\mathfrak{prob}(D \,|\, \omega) \propto \mathfrak{exp}\left[-\tfrac{1}{2}\chi^2(D, \omega)\right].$$

If ω takes a flat prior, then using Bayes' theorem $\mathfrak{prob}(\omega \,|\, D)$ is also proportional to the right hand side. On the other hand, consider

$$\mathfrak{prob}(\chi^2 \,|\, \omega, D) \propto \mathfrak{exp}\left[-\frac{\left(\chi^2(D, \omega) - \mu\right)^2}{4\mu}\right].$$

Again, if ω takes a flat prior, Bayes' theorem gives $\mathfrak{prob}(\omega \,|\, D, \chi^2)$ proportional to the right hand side. Or does it?? We seem to have two very different expressions for the posterior! What is going on here? **[4]**

6. Distribution Function Fitting

In the previous chapter we considered deterministic models with noise added. In this chapter we'll consider some situations where the model itself is probabilistic.

Suppose we sample a distribution function $\mathfrak{prob}(x)$ with a non-uniform but known sampling function $\mathfrak{prob}(S\,|\,x)$. Here x need not be one-dimensional; for example, $\mathfrak{prob}(x)$ might represent the distribution of a certain type of object in the sky and $\mathfrak{prob}(S\mid x)$ the detection efficiency in different parts of the sky. Anyway, the data consist of the set $x_1, \ldots, x_N$ actually sampled. We can calculate $\mathfrak{prob}(x\,|\,S)$ using Bayes' theorem:

$$\mathfrak{prob}(x\,|\,S) = \frac{\mathfrak{prob}(S\,|\,x)\,\mathfrak{prob}(x)}{\int \mathfrak{prob}(S\,|\,x')\,\mathfrak{prob}(x')\,dx'} \tag{6.1}$$

and hence the likelihood is

$$\mathfrak{prob}(x_1, \ldots, x_N|S) = \frac{\prod_{i=1}^{N} \mathfrak{prob}(S\,|\,x_i)\,\mathfrak{prob}(x_i)}{\left[\int \mathfrak{prob}(S\,|\,x)\,\mathfrak{prob}(x)\,dx\right]^N}. \tag{6.2}$$

If we have a model $\mathfrak{prob}(x\mid \omega, M)$ for the distribution function, we can use (6.2) for parameter fitting and model comparison as usual, but these will almost always have to be done numerically.

PROBLEM 6.1: Derive $\mathfrak{prob}(x\,|\,S)$ in (6.1), but without using Bayes' theorem. Consider a sequence of events: $\mathfrak{prob}(x)$ is the probability that the next event will be at x, while $\mathfrak{prob}(x\,|\,S)$ is the probability that the next event *to be sampled* is at x. Then do a calculation sort of like in Problem 1.1. **[3]**

EXAMPLE [The lighthouse] A lighthouse is out at sea off a straight stretch of coast. As its works rotate, the lighthouse emits collimated flashes at random intervals, and hence in random directions θ. There are detectors along the shore which record the position x where each flash reaches the shore, but not the direction it came from. So the data are a set of positions $x_1, \ldots, x_N$ along the shore. The problem is to infer the position of the lighthouse.

Say the lighthouse is at distance a along the shore and b out to sea. If a flash has direction θ, and this is towards the shore, then it will reach the shore at $x = a + b\tan\theta$. We have for the likelihood:

$$\begin{aligned} \mathfrak{prob}(x\,|\,a,b) &= \frac{d\theta}{dx} = \frac{1}{2\pi}\frac{b}{(a-x)^2+b^2}, \\ \mathfrak{prob}(D\,|\,a,b) &\propto \prod_{i=1}^{N} \frac{b}{(a-x_i)^2+b^2}. \end{aligned} \tag{6.3}$$

This problem is often used by Bayesians as a warning against blindly using the sample mean. Symmetry might suggest that the mean of the x_i estimates a; in fact, since the likelihood is a Lorentzian, it doesn't have a mean. On the other hand, the posterior $\mathfrak{prob}(a, b\,|\,D)$ is perfectly well behaved, and constrains a and b more and more tightly as more data are recorded. □

PROBLEM 6.2: Now consider two lighthouses, of relative intensities w and $1-w$, at (a_1, b_1) and (a_2, b_2) respectively. Take $a_1 = 0.5$, $b_1 = 0.2$, $a_2 = 0$, $b_2 = 0.9$, and concoct some fictitious data as follows. Generate a set of x values,

$$80\% \text{ according to } x = a_1 + b_1 \tan\theta,$$
$$20\% \text{ according to } x = a_2 + b_2 \tan\theta,$$

(i.e., $w = 0.8$) with θ random and uniformly distributed in $[-\pi/2, \pi/2]$. Keep the first 200 x_i that lie in $[-1, 1]$ (the detectors extend along only part of the shoreline).

Now use the Metropolis algorithm to recover all five parameters, with 90% confidence intervals. Use a flat prior in $[-1, 1]$ for a_1 and b_1 and a flat prior in $[0, 1]$ for w, b_1, and b_2. (Don't forget the denominator in the likelihood.) **[4]**

EXAMPLE [Where the model is noisy too] Sometimes we have a situation where we can't calculate a model distribution function, but we can generate a model sample. If the model sample size could be made arbitrarily large, that would be as good as a distribution function, but practical considerations (e.g., computer power) may limit this. So we have a finite sample of some underlying distribution function, and we want to compare it with data, i.e., calculate likelihoods and posteriors. If the procedure for generating the model sample has some adjustable parameters, we can do parameter fitting in the usual way.

Since the model and data samples will not be at coincident points, we need to be able to get at the probability of the data sample points from nearby model points. Basically, we need to smooth the model somehow. The crudest way of smoothing is to bin the model and data; in effect assuming that the underlying probability distribution is constant over a bin. Suppose we do this. The bin size must be small enough for the constancy over bins approximation to be reasonable, but large enough for most data sample points to share their bin with some model points.

Let us say we have a reasonable binning, of B bins. Say the underlying probability of a (model or data sample) point being in the i-th bin is w_i. Also say there are M model points and S data sample points in all, m_i and s_i respectively in the i-th bin. The probability distribution for the bin occupancies will be a multinomial distribution (see equation 2.2 on page 15):

$$\mathrm{prob}(s_i, m_i \,|\, w_i) = M!\, S! \prod_{i=1}^{B} \frac{w_i^{m_i+s_i}}{m_i! s_i!}. \tag{6.4}$$

Since we don't know the w_i, except that $\sum_{i=1}^{B} w_i = 1$, let us marginalize them out with a flat prior. Using the identity

$$\left(\prod_{i=1}^{B} \int w_i^{n_i}\, dw_i \right) \delta\left(\textstyle\sum_{j=1}^{B} w_j - 1 \right) = \frac{1}{(N+B-1)!} \prod_{i=1}^{B} n_i! \tag{6.5}$$

we get

$$\mathrm{prob}(s_i, m_i) = \frac{M!\, S!\, (B-1)!}{(M+S+B-1)!} \prod_{i=1}^{B} \frac{(m_i+s_i)!}{m_i! s_i!}. \tag{6.6}$$

The $(B-1)!$ factor comes from normalizing the prior on the w_i.

In similar fashion, we can calculate $\mathfrak{prob}(m_i)$, and combining that with $\mathfrak{prob}(s_i, m_i)$ in (6.6) we get

$$\mathfrak{prob}(s_i \,|\, m_i) = \frac{S!\,(M+B-1)!}{(M+S+B-1)!} \prod_{i=1}^{B} \frac{(m_i+s_i)!}{m_i! s_i!}. \tag{6.7}$$

If M *can* be made arbitrarily large, we can make the bins so small that each bin contains at most one data sample point. Then (6.7) simplifies to

$$\mathfrak{prob}(s_i \,|\, m_i) \propto \prod_j (1+m_j), \tag{6.8}$$

where the product is only over boxes where $s_j = 1$. Assuming $m_j \gg 1$, this is just the formula for the continuous case. In other words, (6.7) has the correct large-m_i limit. □

PROBLEM 6.3: In the preceding example, we took M and S as fixed, and got a multinomial distribution for $\mathfrak{prob}(s_i, m_i \,|\, w_i)$. An alternative method would be to let M and S be the expectation values for the totals when m_i and s_i are drawn from the bin probabilities by a Poisson process. What would this procedure give for $\mathfrak{prob}(s_i, m_i)$? **[3]**

Apart from fitting model distribution functions to samples, and comparing different models for the same data, we need to worry about goodness of fit. We may have to judge whether a given data set could plausibly have come from a particular distribution function, or whether two discrete samples could plausibly have come from the same probability distribution. In fact, it is enough to consider the second problem, since the first is a limiting case of it.

We have to invent a statistic that tries to measure the mismatch between the two samples in question; then we calculate the distribution of the statistic for random samples drawn from the same probability distribution and check if the actual value is anomalous. We could use the likelihood itself (or a function of it) as the statistic, as we did with the χ^2 test. But if the data are one-dimensional, there is an easier way for which we don't have to know what the underlying probability distribution is.

The trick is to base the statistic on the cumulative distribution function. Suppose that from some probability distribution $\mathfrak{prob}(x)$ there are two samples: $u_1, u_2 \ldots$ of size N_u, and $v_1, v_2, \ldots$ of size N_v. We consider the difference of cumulative fractions

$$s(x) = \text{frac}\,(\{u_i\} < x) - \text{frac}\,(\{v_i\} < x).$$

Figure 6.1 illustrates. Changing the variable x in $\mathfrak{prob}(x)$ will not change s at any of the sample points; in particular, x can be chosen to make $\mathfrak{prob}(x)$ flat. Thus, any statistic

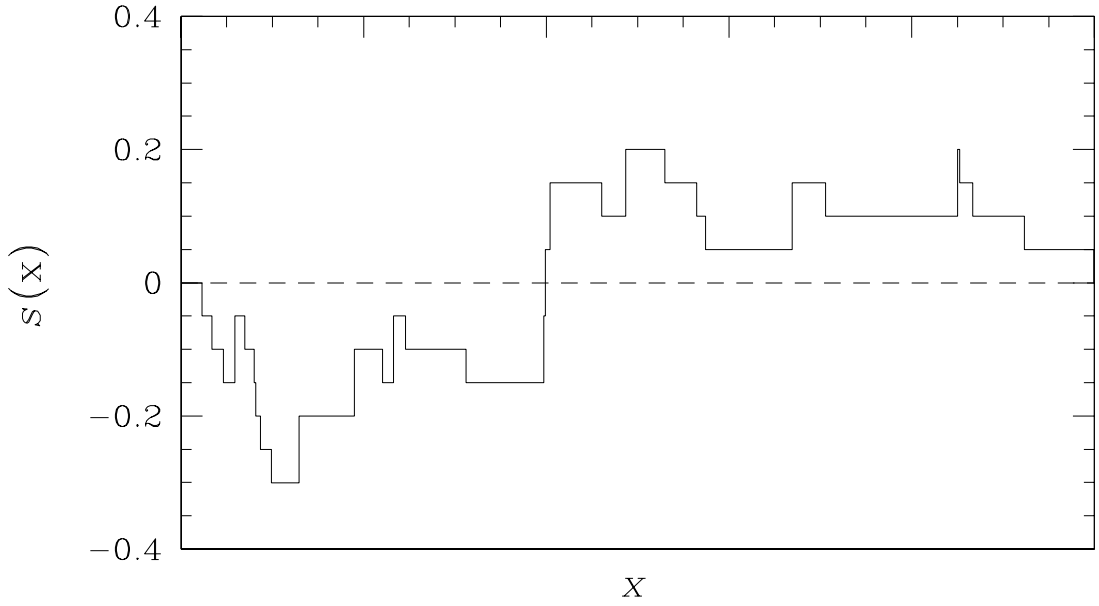

Figure 6.1: The difference of the cumulative distributions of two samples $u_1, \ldots, u_{10}$ and $v_1, \ldots, v_{20}$. The sizes of the two samples should be evident from the staircase. The x axis can be stretched or shrunk without changing any statistic evaluated from the heights of the steps.

depending only on $s(x)$ at the sample points will have its distribution independent of $\mathfrak{prob}(x)$.

There are many statistics defined from the $s(x)$ staircase. The best known is the Kolmogorov-Smirnov statistic, which is simply the maximum height or depth at any point on the staircase. Another statistic (named after Cramers, von Mises, and Smirnov in various combinations and permutations) is sum of the squares of the vertical midpoints of the steps. The distribution of these particular statistics can be calculated by pure thought, but we won't go into that. For applications, it's more important to understand how to calculate the distributions by simulating, and that is the subject of the next problem.

PROBLEM 6.4: The important thing to notice about the $s(x)$ staircase is that it is just a one-dimensional random walk constrained to return to the origin. This suggests a way of simulating the distribution of statistics based on $s(x)$.

1) Generate two samples of random numbers in $[0, 1]$, $u_1, u_2 \ldots$ of size N_u, and $v_1, v_2, \ldots$ of size N_v.
2) Sort the samples $\{u_i\}$ and $\{v_i\}$.
3) Go up along the sorted arrays of u and v; whenever the next lowest x comes from the u (v) sample, increase (decrease) $s(x)$ by $1/N_u$ ($1/N_v$). This generates the staircase.
4) Calculate the statistic from the staircase.

Iterating the process above gives the distribution of the statistic. Your job is to write a program to do this. Plot up the p-value distribution $\mathfrak{prob}(\kappa' > \kappa)$ for the Kolmogorov-Smirnov statistic κ for

$N_u = 10, N_v = 20$. Overplot the asymptotic pure-thought formula

$$\mathfrak{prob}(\kappa' > \kappa) = f(\mu\kappa), \quad f(\lambda) = 2\sum_{k=1}^{\infty}(-1)^{k-1}\,\mathfrak{exp}[-2k^2\lambda^2],$$
$$\mu \simeq \left(\hat{N} + 0.12 + 0.11/\hat{N}\right), \quad \hat{N} = \sqrt{N_u N_v/(N_u + N_v)}.$$

Then invent your own statistic and plot the p-value distribution for it. **[4]**

An important modification of the above is when you want to compare a sample against a given probability distribution function. For instance, after doing a least-squares fit you may want to test whether the residuals are consistent with a Gaussian of mean 0 and the estimated dispersion. The modification is easy: you just let u be the sample and v the distribution function with $N_v \to \infty$. (For statistics involving vertical midpoints, to be well-defined you'll have to use only the rising steps.) In some cases the distribution function in question might already have been fitted to the data by adjusting some parameters. We encountered this problem in the χ^2 test too, and in the linearizable case we solved it by reducing the number of degrees of freedom (equation 5.37 on page 43) in the distribution of χ^2. But for Kolmogorov-Smirnov and related tests, I know of no way of allowing for the effect of fitted parameters.

Another possible generalization is to several different samples: just choose a statistic which is calculated for pairs of samples and then combined in some sensible way.

Bear in mind, though, that cumulative statistics work only in one-dimension. There is a "two-dimensional Kolmogorov-Smirnov test", but it assumes that the two-dimensional distribution function involved is a product of two one-dimensional distribution functions. In general, if you are dealing with a sample in two or more dimensions then life is much more difficult.

7. Entropy

So far in this book, prior probabilities have not been very interesting, because we have not had interesting prior information to put into them. In the last two chapters, however, the interest will shift to assigning probability distributions so as to incorporate prior constraints. We will derive informative priors, ready to be multiplied by data likelihoods in the usual way. Sometimes these priors will be so informative that they start to behave almost like posteriors; this can happen when the data, though incomplete, are so accurate that one might as well combine them with prior information in one probability distribution and dispense with the likelihood.

The key player in this chapter is the entropy function

$$S = -\sum_{i=1}^{N} p_i \log p_i \tag{7.1}$$

associated with a probability distribution $p_1, \ldots, p_N$. We have seen entropy briefly already (page 12), but now we digress into some elementary information theory to see where it comes from.

DIGRESSION [Shannon's Theorem] We suppose there is a real function $S(p_1, \ldots, p_N)$ that measures the uncertainty in a probability distribution, and ask what functional form S might have. We impose three requirements.

(i) S is a continuous function of the p_i.

(ii) If all the p_i happen to be equal, then S increases with N. In other words,

$$s(N) \equiv S\left(\tfrac{1}{N}, \ldots, \tfrac{1}{N}\right) \tag{7.2}$$

is a monotonically increasing function of N. Qualitatively, this means that if there are more possibilities, we are more uncertain.

(iii) S is additive, in the following sense. Suppose we start with two possibilities with probabilities p_1 and $1 - p_1$. Then we decide to split the second possibility into two, with probabilities p_2 and $p_3 = 1 - p_1 - p_2$. We require that $S(p_1, p_2, p_3)$ should be the uncertainty from $p_1, 1 - p_1$ plus the uncertainty associated with splitting up the second possibility. In other words

$$S(p_1, p_2, p_3) = S(p_1, 1 - p_1) + (1 - p_1) S\left(\frac{p_2}{1 - p_1}, \frac{p_3}{1 - p_1}\right). \tag{7.3}$$

More generally, consider M mutually exclusive propositions with probabilities $q_1, \ldots, q_M$. Instead of giving these probabilities directly we do the following. We group the first k together and call the group probability p_1, and write p_2 for the next group of the $(k+1)$st to $(k+l)$th and so on for N groups. Then we give the conditional probabilities for propositions within

each group, which will be $q_1/p_1, \ldots, q_k/p_1$, $q_{k+1}/p_2, \ldots, q_{k+l}/p_2$ and so on. The additivity requirement is then

$$\begin{aligned} S(q_1, \ldots, q_M) = S(p_1, \ldots, p_N) + p_1 S\left(\frac{q_1}{p_1}, \ldots, \frac{q_k}{p_1}\right) \\ + p_2 S\left(\frac{q_{k+1}}{p_2}, \ldots, \frac{q_{k+l}}{p_2}\right) + \cdots \end{aligned} \tag{7.4}$$

The continuity requirement means that we only need to fix the form of S for rational values of the p_i. So let us consider (7.4) for

$$p_i = \frac{n_i}{M}, \quad q_i = \frac{1}{M}, \tag{7.5}$$

where M and the n_i are all integers. Then (7.4) becomes

$$s(M) = S(p_1, \ldots, p_N) + \sum_{i=1}^{N} p_i s(n_i), \tag{7.6}$$

and this is true for general rational p_i. Now choose $n_i = M/N$, assuming N divides M. This gives

$$s(M) = s(N) + s(M/N), \tag{7.7}$$

and among other things it immediately tells us that $s(1) = 0$. Had M and N been real variables rather than integers, we could differentiate (7.7) with respect to N, set $N = 1$ and integrate with the initial condition $s(1) = 0$ to obtain the unique solution $s(N) \propto \ln N$. As things stand, $s(N) \propto \ln N$ clearly is a solution, but we still have to prove uniqueness.

This is where the monotonicity requirement comes in. First, we note that (7.7) implies

$$s(n^k) = ks(n) \tag{7.8}$$

for any integer k, n. Now let k, l be any integers ≥ 2. Then for sufficiently large n we can find an integer m such that

$$\frac{m}{n} \leq \frac{\ln k}{\ln l} < \frac{m+1}{n}, \quad \text{or } l^m \leq k^n < l^{m+1}. \tag{7.9}$$

Since s is monotonic, we have from (7.8) and (7.9) that

$$ms(l) \leq ns(k) \leq (m+1)s(l), \quad \text{or} \quad \frac{m}{n} \leq \frac{s(k)}{s(l)} \leq \frac{m+1}{n}. \tag{7.10}$$

Comparing (7.10) and (7.9) we have

$$\left| \frac{\ln k}{\ln l} - \frac{s(k)}{s(l)} \right| \leq \frac{1}{n}. \tag{7.11}$$

Since n could be arbitrarily large, (7.11) forces $s(N) \propto \ln N$. We can just write $s(N) = \log N$ leaving the base arbitrary. Substituting in (7.6) and using (7.5), we are led uniquely to (7.1) as a measure of uncertainty. □

ASIDE It is easy enough to generalize the entropy expression (7.1) to continuous probability distributions. We have

$$S(p(x)) = -\int p(x) \log\left(\frac{p(x)}{w(x)}\right) dx. \tag{7.12}$$

Here the weight $w(x)$ is needed to keep S invariant under changes of the variable x. But in practice the continuous expression (7.12) does not see much use. In problems where a continuous probability distribution is needed, people usually discretize initially and then pass to the continuous limit late in the calculation when entropy no longer appears explicitly. □

According to the maximum entropy principle, entropy's purpose in life is to get maximized (by variation of the p_i) subject to data constraints. In general, the data constraints could have any form, and the maximization may or may not be computationally feasible. But if the data happen to consist entirely of expectation values over the probability distribution then the maximization is both easy and potentially very useful. Let us take up this case.

Consider a variable ε which can take discrete values ε_i, each having probability p_i. The number of possible values may be finite or infinite, but $\sum_i p_i = 1$. There are two known functions $f(\varepsilon)$ and $g(\varepsilon)$, and we have measured their expectation values

$$X = \sum\nolimits_i p_i f(\varepsilon_i), \quad Y = \sum\nolimits_i p_i g(\varepsilon_i). \tag{7.13}$$

[There could be one, three, or more such functions, but let's consider two.] We want to assign the probabilities p_i on the basis of our knowledge of the functional forms of f and g and their measured expectation values. We therefore maximize the entropy subject to the constraints (7.13) and also $\sum_i p_i = 1$. The equation for the maximum is

$$\begin{aligned}\frac{\partial}{\partial p_i}\Big(\sum\nolimits_i p_i \ln p_i + \lambda \sum\nolimits_i p_i \\ + x \sum\nolimits_i p_i f(\varepsilon_i) + y \sum\nolimits_i p_i g(\varepsilon_i)\Big) = 0,\end{aligned} \tag{7.14}$$

where λ, x, y are Lagrange multipliers to be set so as to satisfy the constraints. (We might as well replace $\log$ in the entropy with $\ln$, because the base of the $\log$ just amounts to a multiplicative factor in the Lagrange multipliers.) Eliminating λ to normalize the probabilities we get

$$\begin{aligned} p_i &= \frac{\exp[-x f(\varepsilon_i) - y g(\varepsilon_i)]}{Z(x,y)}, \\ Z &= \sum_i \exp[-x f(\varepsilon_i) - y g(\varepsilon_i)]. \end{aligned} \tag{7.15}$$

Here and later x and y are understood to be functions of the measured values X and Y. Z is traditionally called the partition function[1] and it is very important because lots of things can be calculated from it.

[1] The symbol Z comes from the partition function's German name Zustandsumme, or sum-over-states, which is exactly what it is.

DIGRESSION [Global entropy maximum] At this point we should make sure that the solution (7.15) really is the global maximum of the entropy, and not just any stationary point. To do this let us backtrack a bit and consider the p_i as not yet fixed. Now consider a set of non-negative numbers q_i satisfying $\sum_i q_i = 1$; these are formally like probabilities, but are not necessarily equal to the p_i. Now $\ln x \leq (x - 1)$ for any finite non-negative x, with equality if and only if $x = 1$; this implies

$$\sum_i p_i \ln\left(\frac{q_i}{p_i}\right) \leq \sum_i p_i \left(\frac{q_i}{p_i} - 1\right) = 0, \tag{7.16}$$

and hence

$$S(p_1, p_2, \ldots) \leq -\sum_i p_i \ln q_i, \tag{7.17}$$

with equality if and only if all the $q_i = p_i$. If we now choose

$$q_i = \frac{\exp[-x\, f(\varepsilon_i) - y\, g(\varepsilon_i)]}{Z(x, y)}, \tag{7.18}$$

with Z defined as in (7.15), then (7.17) becomes

$$S(p_1, p_2, \ldots) \leq \ln Z(x, y) + xX + yY. \tag{7.19}$$

But the assignment (7.15) for the p_i gives precisely this maximum value, so (7.15) must be the maximum-entropy solution. □

When calculating the partition function it is very important to be aware of any degeneracies. A degeneracy is when there are distinct values $\varepsilon_i \neq \varepsilon_j$ giving $f(\varepsilon_i) = f(\varepsilon_j)$ and $g(\varepsilon_i) = g(\varepsilon_j)$. If we disregard degeneracies, we may wrongly sum over distinct values of f and g rather than distinct values of ε.

EXAMPLE [Gaussian and Poisson distributions revisited] Suppose ε is a continuous variable. Then the partition function is

$$Z = \int w(\varepsilon)\, \exp[-x\, f(\varepsilon) - y\, g(\varepsilon)]\, d\varepsilon, \tag{7.20}$$

where $w(\varepsilon)$, which expresses the degeneracy, is called the density of states. Now suppose that the data we have are $X = \langle \varepsilon \rangle$ and $Y = \langle \varepsilon^2 \rangle$. Then

$$Z = \int w(\varepsilon)\, \exp\left[-x\varepsilon - y\varepsilon^2\right] d\varepsilon, \tag{7.21}$$

and hence

$$\mathrm{prob}(\varepsilon) \propto w(\varepsilon) \exp\left[-x\varepsilon - y\varepsilon^2\right], \tag{7.22}$$

with x and y taking appropriate values so as to give the measured $\langle \varepsilon \rangle$ and $\langle \varepsilon^2 \rangle$. The interpretation is that when all we are given about a probability distribution are its mean and variance, our most conservative inference is that it is a Gaussian.

Now consider another situation. We are given that the expectation value for the number of events in a certain time interval is m. What can we say about the probability distribution of the number of events? To work this out, divide the given time interval into N subintervals, each so small that only 0 or 1 event can occur in it. The degeneracy for n events in the full interval is the number of ways we can distribute n events among the N subintervals, or

$$ {}^{N}C_{n} \simeq \frac{N^n}{n!} \text{ for } N \gg n. \tag{7.23} $$

We then have

$$ Z(x) = \sum_{n=0}^{\infty} \frac{N^n}{n!} e^{-nx} = \exp\left[N e^{-x}\right]. \tag{7.24} $$

We solve for x (which recall is the Lagrange multiplier) by equating the mean implied by $Z(x)$ to m, getting

$$ x = -\ln \frac{m}{N}, \quad Z = e^m, \tag{7.25} $$

whence

$$ \mathrm{prob}(n \,|\, \langle n \rangle = m) = \frac{e^{-nx}}{Z(x)} = e^{-m} \frac{m^n}{n!}. \tag{7.26} $$

In other words, for an event-counting probability distribution where we know only the mean, our most conservative inference is that it is Poisson. □

PROBLEM 7.1: Using equation (7.15) we can write

$$ \langle x(\varepsilon) \rangle = -\frac{\partial \ln Z}{\partial x}. $$

The left hand side is the mean X. Give

$$ \frac{\partial^2 \ln Z}{\partial x^2} $$

some kind of interpretation. **[3]**

Let us leave partition functions now, before they completely take over this chapter, and return to least-squares, but with a new complication.

A general problem in image reconstruction is to infer an image on many picture elements (pixels) f_j from data d_i taken with a blurred and noisy camera:

$$ d_i = \textstyle\sum_j R_{ij} f_j + n_i(\sigma). \tag{7.27} $$

Here R_{ij} is a blurring function, indicating that a fraction of the light that properly belongs to the j-th pixel in practice gets detected on the i-th pixel. [I am stating the problem in optical terms, but really R_{ij} could be any linear operator.] There is noise on top of the blurring. This looks like a straightforward least-squares problem with likelihood (cf. equation 5.5 on page 36)

$$ \mathrm{prob}(D \,|\, f_1, f_2, \ldots) \propto \exp\left[-\tfrac{1}{2}\sigma^{-2} \sum_i \left(\textstyle\sum_j R_{ij} f_j - d_i\right)^2\right]. \tag{7.28} $$

The complication is that the number of data points is not much greater than the number of pixels we want to reconstruct; it may even be less. So it is important to think about what else we know. For instance, if the f_j represent an image, the values must be non-negative. What else should we put in the prior?

DIGRESSION [The monkey argument] One line of reasoning that leads to a workable prior is what's nowadays called the monkey argument. Imagine a team of monkeys randomly chucking peanuts into jars. When the peanuts land in the jars they get squashed and turned into peanut butter. There are N peanuts and M jars. When the monkeys have finished, what is the probability distribution for the amount of peanut butter in each jar? To work this out, say the i-th jar gets n_i peanuts. The distribution of the n_i follows a multinomial distribution

$$\mathrm{prob}(n_i \,|\, N, M) = M^{-N} \frac{N!}{\prod_{i=1}^{M} n_i!}. \tag{7.29}$$

If the number of peanuts is so much larger than the number of jars that $n_i \gg 1$ then we can use Stirling's approximation (formula M.7 on page 70), and

$$\ln \mathrm{prob}(n_i \,|\, N, M) = -N \ln M - \sum_{i=1}^{M} n_i \ln(n_i/N). \tag{7.30}$$

If we write f_i for n_i/N (the fraction of the total peanut butter) then (7.30) becomes

$$\ln \mathrm{prob}(f_1, f_2, \ldots \,|\, N, M) = -N \ln M - N \sum_i f_i \ln f_i. \tag{7.31}$$

We can take this story as a parable for an image processing problem: the jars correspond to pixels and the peanut butter to brightness at each pixel. It suggests the prior

$$\mathrm{prob}(f_1, f_2, \ldots) = \exp\left(-\alpha \textstyle\sum_i f_i \ln f_i\right). \tag{7.32}$$

Here α is an unknown parameter; it arose from our discretizing into peanuts. In applications α will have to be marginalized away or fixed at the value that maximizes the posterior. □

It is tempting to interpret images as probability distributions, and identify (7.31) with the entropy. But the connection, if there is one, is not so simple, because (7.31) is a measure of prior probability while (7.1) is a measure of the uncertainty associated with a probability distribution. This issue remains controversial, and until it is resolved, it seems prudent to refer to (7.31) and as a 'configurational' entropy, not to be identified too closely with information theoretic entropy.

Another reason not to identify f_j with a probability distribution is that it would leave us at a loss to interpret the following,[1] which is a prior for when f_j is allowed to be negative.

DIGRESSION [Macaulay and Buck's prior] We now consider a different form of the monkey argument. This time, we have the monkeys tossing both positive and negative peanuts, and what interests us is the difference between positive and negative kinds of peanut butter in each jar. Also,

[1] V.A. Macaulay & B. Buck, Nucl Phys A 591, 85–103 (1995).

instead of having a fixed number of peanuts and hence a multinomial distribution among the jars, we suppose that the distribution of peanuts in any jar follows a Poisson distribution with mean μ_+ for positive peanuts and μ_- for negative peanuts. (The other case could be worked out with a Poisson distribution as well.) The probability for the i-th jar to have n_{i+} positive peanuts and n_{i-} negative peanuts is

$$\mathrm{prob}(n_{i+}, n_{i-}) = e^{-(\mu_+ + \mu_-)} \frac{(\mu_+)^{n_{i+}} (\mu_-)^{n_{i-}}}{n_{i+}! n_{i-}!}. \tag{7.33}$$

Let us consider the term from one jar and suppress the subscript i. Writing $n = \frac{1}{2}(n_+ + n_-)$ and $q = \frac{1}{2}(n_+ - n_-)$ we have

$$\mathrm{prob}(q, n) = e^{-(\mu_+ + \mu_-)} \left(\frac{\mu_+}{\mu_-}\right)^q (\mu_+ \mu_-)^n \frac{1}{(n+q)!(n-q)!}. \tag{7.34}$$

Writing $\beta^2 = (\mu_+ \mu_-)$ and $\gamma = (\mu_+ / \mu_-)$ changes (7.34) to

$$\mathrm{prob}(q, n) = e^{-\beta(\sqrt{\gamma} + 1/\sqrt{\gamma})} \frac{\gamma^q \beta^{2n}}{(n+q)!(n-q)!}. \tag{7.35}$$

Since we are really interested in q, we marginalize n away by summing over it. The series is essentially a modified Bessel function:

$$\mathrm{prob}(q) = \sum_{n=|q|}^{\infty} \mathrm{prob}(q, n) = e^{-\beta(\sqrt{\gamma} + 1/\sqrt{\gamma})} \gamma^q \, I_{2q}(2\beta). \tag{7.36}$$

Using the asymptotic large-q form of I_{2q} and ignoring factors varying as $\ln q$, we have

$$\begin{aligned} \log \mathrm{prob}(q) = \sqrt{(2q)^2 + (2\beta)^2} - \beta(\sqrt{\gamma} + 1/\sqrt{\gamma}) + \\ q \ln \gamma - 2q \sinh^{-1} \frac{q}{\beta}, \end{aligned} \tag{7.37}$$

where the base of the logarithm is an arbitrary constant. Changing variables from the integer q to the real number $f = 2q\epsilon$ (where ϵ is sort of the amount of butter per peanut), and also writing $w = 2\beta\epsilon$, we have

$$\log \mathrm{prob}(f) = \sqrt{f^2 + w^2} - \tfrac{1}{2} w(\sqrt{\gamma} + 1/\sqrt{\gamma}) + \tfrac{1}{2} \ln \gamma - f \sinh^{-1} \frac{f}{w}. \tag{7.38}$$

If we further write $\overline{f} = w \sinh\left(\frac{1}{2} \ln \gamma\right) = \frac{1}{2} w(\sqrt{\gamma} - 1/\sqrt{\gamma})$ then the prior has the nice form

$$\begin{aligned} \log \mathrm{prob}(f) = (f^2 + w^2)^{\frac{1}{2}} - (\overline{f}^2 + w^2)^{\frac{1}{2}} \\ + f\left(\sinh^{-1}(\overline{f}/w) - \sinh^{-1}(f/w)\right), \end{aligned} \tag{7.39}$$

for one pixel, the full $\log \mathrm{prob}$ being a sum of such terms. Here $\overline{f}$, w, and the proportionality constant are parameters which have to determined or marginalized away. □

We now have two priors,

$$\begin{aligned}\log \mathrm{prob}(f_1, f_2, \ldots) = &\sum\nolimits_i - f_i \ln f_i \quad \text{or} \\ &\sum\nolimits_i (f_i^2 + w^2)^{\frac{1}{2}} - (\overline{f}^2 + w^2)^{\frac{1}{2}} \\ &+ f_i \Big(\sinh^{-1}(\overline{f}/w) - \sinh^{-1}(f_i/w) \Big),\end{aligned} \tag{7.40}$$

the first if f_i must be positive, the second if f_i can be negative. In either case the posterior is

$$\begin{aligned}&\mathrm{prob}(f_1, f_2, \ldots | D) \propto \exp\left[\alpha \ln \mathrm{prob}(f_1, f_2, \ldots) - \tfrac{1}{2}\chi^2\right], \\ &\chi^2 = \sigma^{-2} \sum\nolimits_i \Big(\textstyle\sum_j R_{ij} f_j - d_i \Big)^2.\end{aligned} \tag{7.41}$$

Despite their formidable appearance, these equations are quite practical, even with millions of pixels. It helps numerical work tremendously that both forms of the prior, as well as χ^2, are convex functions of the f_i. The standard technique to locate the maximum of the posterior proceeds iteratively in two stages. In the first stage, one just tries to reach a place with the correct value of χ^2; that value is the number of pixels minus the number of non-negligible singular values of R_{ij} (cf. equation 5.38 on page 43). The second stage of iterations holds χ^2 fixed while increasing $\mathrm{prob}(f_1, f_2, \ldots)$ as far as possible. This amounts to treating α as a Lagrange multiplier. Finally, one can present uncertainties in various ways. Applications are numerous and varied.[1]

[1] See B. Buck & V.A. Macaulay eds., *Maximum entropy in action* (Oxford University Press 1991) for some examples.

8. Entropy and Thermodynamics

To conclude this book we'll consider entropy in its historically original context, and see that the information theoretic ideas from the previous chapter lead to a beautiful reinterpretation of the branch of physics known as thermodynamics.[1] This chapter will try to explain all the necessary physical ideas as it goes along, so it is not essential to have seen thermodynamics before; but if you have, it will be easier to follow.

Thermodynamics is about measuring a few macroscopic properties (such as internal energy, volume) of systems that are microscopically very complex—a gram of water has $> 3 \times 10^{22}$ molecules—and predicting other macroscopic properties (such as temperature, pressure). Given some macroscopic data we assign probabilities $p_1, p_2, \ldots$ to different microstates using the principle of maximum entropy, and then use the assigned probabilities to predict other macroscopic quantities. (Microstate refers to the detailed state—including position and energy of every molecule—of the system. Naturally, we never deal with microstates explicitly.)

In general, maximizing entropy subject to arbitrary data constraints is a hopelessly difficult calculation even for quite small systems, never mind systems with 10^{22} molecules. What saves thermodynamics is that the macroscopic measurables are either expectation values, that is, of the form

$$X = \sum\nolimits_i p_i \, \mathfrak{x}(\varepsilon_i), \quad Y = \sum\nolimits_i p_i \, \mathfrak{y}(\varepsilon_i) \tag{8.1}$$

which we considered in the previous chapter, or somehow related to expectation values. For example, if $\mathfrak{x}(\varepsilon_i)$ is the energy of the microstate labelled by ε_i, then the expectation value X will be the internal energy. Any measurable that is extensive (meaning that it doubles when you clone the system) such as internal energy, volume, particle number, can play the role of X, Y. [There may be any number of X, Y variables in a problem, two is just an example.] Non-extensive measurables like temperature and pressure cannot be X, Y variables but will turn out to be related to them in an interesting way.

The key to maximizing the entropy subject to X, Y is the partition function

$$Z(x, y) = \sum_i \mathfrak{exp}[-x \, \mathfrak{x}(\varepsilon_i) - y \, \mathfrak{y}(\varepsilon_i)]. \tag{8.2}$$

The maximum-entropy probability values are given by

$$p_i = \frac{\mathfrak{exp}[-x \, \mathfrak{x}(\varepsilon_i) - y \, \mathfrak{y}(\varepsilon_i)]}{Z(x, y)} \tag{8.3}$$

[1] In standard physics usage, this chapter is about thermodynamics and statistical mechanics. However, I will use the less common terms 'classical thermodynamics' and 'statistical thermodynamics', which are a little more descriptive.

(cf. equation 7.15 on page 52). The x, y variables are Lagrange multipliers associated with the entropy maximization (originally appearing in equation 7.14) and hence ultimately functions of the measured X, Y. Later on, the x, y will turn out to represent non-extensive thermodynamic variables, but for now they are just arguments of Z.

Let us evaluate the entropy for the probability distribution (8.3), and then write $S(X, Y)$ for the maximum value.[1] We get

$$S(X, Y) = \ln Z(x, y) + xX + yY, \tag{8.4}$$

and from (8.2) and (8.3) we have

$$X = -\left(\frac{\partial \ln Z}{\partial x}\right)_y, \quad Y = -\left(\frac{\partial \ln Z}{\partial y}\right)_x. \tag{8.5}$$

Differentiating (8.4) and inserting (8.5) gives us two complementary relations

$$x = \left(\frac{\partial S}{\partial X}\right)_Y, \quad y = \left(\frac{\partial S}{\partial Y}\right)_X. \tag{8.6}$$

Equations (8.4) to (8.6) express a so-called Legendre transformation. The functions $S(X, Y)$ and $\ln Z(x, y)$ are Legendre transforms of each other. The variables x and y are conjugate to X and Y respectively in the context of Legendre transforms; correspondingly, $-X$ and $-Y$ are conjugate to x and y. What this means is that although we started by assuming X and Y as measured (independent variables) and x and y as functions of them, we equally can take x and y as measured and use the same relations to infer X and Y. In fact, we are free to choose any two out of X, Y, x, y, S, Z as the independent variables.

The partition function (8.2) and the Legendre transform relations (8.4) to (8.6) encapsulate the formal structure of statistical thermodynamics. Statistical thermodynamics is a mixed micro- and macroscopic theory: the data are all macroscopic, but to compute the partition function we need to have some microscopic knowledge; for example, if X is the internal energy then $\mathfrak{x}(\varepsilon_i)$ in equations (8.1) and (8.2) will involve the microscopic energy levels and Z will involve a sum over possible energy levels. Once the partition function is in hand, the Legendre transform relations express the macroscopic thermodynamic variables in terms of microscopic quantities, and all sorts of useful results can be derived. All predictions are of course probabilistic, since they refer to the maximum of $S(p_1, p_2, \ldots)$; but given the extremely large number of microstates (the so-called thermodynamic limit) the maximum of $S(p_1, p_2, \ldots)$ is usually extremely precise, and the macroscopic uncertainties are usually negligible.

---

[1] We are making a subtle but important semantic shift here. So far we used 'entropy' to mean $S(p_1, p_2, \ldots)$, a function of the probability distribution and hence of our state of knowledge. From now on 'entropy' will mean $S(X, Y)$ which is the *maximum* value of $S(p_1, p_2, \ldots)$, and depends only on the measured X, Y. It is, unfortunately, conventional to use the same name and symbol for both of these distinct concepts. Maybe authors secretly enjoy the endless confusion it causes.

The macroscopic subset of statistical thermodynamics is classical thermodynamics. In classical thermodynamics the partition function is never calculated explicitly, hence macroscopic variables can be related to each other but not to anything microscopic. Formally, classical thermodynamics consists just of the Legendre transform relations (8.4) to (8.6) and their consequences. A major role is played by differential formulas of the type

$$dS = x\,dX + y\,dY, \tag{8.7}$$

(which follows from equations 8.4 and 8.6). In fact, all classical thermodynamics formulas are basically identities between partial derivatives. But because the partial derivatives are physically relevant, classical thermodynamics is a powerful physical theory. And because it requires no microscopic knowledge it is the most robust branch of physics, the only branch to remain fundamentally unchanged through the 20th century.

But enough about formal structure. Let us see now how the formalism actually works, taking examples first from classical thermodynamics and then from statistical thermodynamics.

Suppose X and Y are the internal energy E and the volume V of a system. The system is in equilibrium, i.e., there is no tendency for E, V to change at the moment. Then the Legendre transform relations (8.4) and (8.6) give

$$\begin{aligned} &S(E,V) = \frac{1}{\tau}E + \frac{p}{\tau}V + \ln Z(\tau, p/\tau) \\ &\text{where} \quad 1/\tau \equiv \left(\frac{\partial S}{\partial E}\right)_V, \quad p/\tau \equiv \left(\frac{\partial S}{\partial V}\right)_E. \end{aligned} \tag{8.8}$$

Despite the hopeful choice of symbols τ, p, we have as yet no physical interpretation for these, nor any way of assigning a numerical value to S on a macroscopic basis. We do, however, have the important principle that the total entropy of a system does not change during a reversible process, though entropy may be transferred between different parts of a system. A reversible change cannot change the net information we have on the system's internal state, so S does not change. During a reversible change, a system effectively moves through a sequence of equilibrium states. Applying the principle

$$\text{reversible} \Rightarrow \text{constant total } S$$

to

$$\tau\,dS = dE + p\,dV, \tag{8.9}$$

(which is just 8.7 for the case we are considering) we can interpret τ as temperature, p as pressure, $\tau\,dS$ as heat input, and define scales for measuring temperature and entropy. The following digression explains the physical arguments leading to all this.

DIGRESSION [Heat is work and work is heat] To get macroscopic interpretations for τ, p and S, we do some thought experiments. We imagine a gas inside a cylinder with a piston, with external heat sources and sinks, and subject the gas to various reversible processes. This imaginary apparatus is just to help our intuition; the arguments and conclusions are not restricted to gases in cylinders. We reason as follows.

(1) That p must be pressure is easy to see. If we insulate the gas to heat we have $dS = 0$, and if we then compress it $dE = -p\,dV$; since dE is the differential internal energy, from energy conservation $p\,dV$ must be the differential work done by the gas, so p must be pressure.

(2) Next, we can infer that $\tau\,dS$ is differential heat input. If we supply (or remove) external heat, the input heat must be the internal energy change plus the work done by the gas. The interpretation of $\tau\,dS$ follows from (8.9). Under reversibility, the total entropy of gas and heat source/sink is still constant, but entropy moves with the heat. We may now identify equation (8.9) as a relation between heat and work; it is one form of the first law of thermodynamics.

(3) Interpreting τ as temperature is more complicated, because unlike pressure, temperature is not already defined outside thermodynamics. First we show that τ is a parameter relating to heat transfer. To do this, let us rewrite the Legendre transform (8.8) slightly: instead of taking E, V as independent variables let us take S, V as independent and write $E(S, V)$. This gives us

$$\begin{aligned} &E(S,V) = \tau S - pV + G(\tau, p) \\ &\text{where} \quad \tau = \left(\frac{\partial E}{\partial S}\right)_V, \quad p = -\left(\frac{\partial E}{\partial V}\right)_S, \end{aligned} \tag{8.10}$$

and we have introduced $G(\tau, p) = -\tau \ln Z$. The variables τ, p in (8.10) are really the same as in (8.8): τ clearly so, and p because of the partial-derivative identity

$$\left(\frac{\partial E}{\partial S}\right)_V \left(\frac{\partial S}{\partial V}\right)_E \left(\frac{\partial V}{\partial E}\right)_S = -1. \tag{8.11}$$

Equation (8.10) is a Legendre transform where $-p$ is a conjugate variable to V, and τ is conjugate to S. If two gases are separated by an insulated piston, then $p\,dV$ work will be done by the gas at higher p on the gas at lower p. Analogously, if two gases are kept at fixed volume then $\tau\,dS$ heat will flow from the one at higher τ to the one at lower τ. This shows that τ is some sort of temperature. But it still does not let us assign numbers to τ and S separately, only to $\tau\,dS$.

(4) To make further progress, we take the gas through a cycle of processes

$$\begin{array}{ccc} S_{\text{low}}, \tau_{\text{hot}} & \rightarrow & S_{\text{high}}, \tau_{\text{hot}} \\ \uparrow & & \downarrow \\ S_{\text{low}}, \tau_{\text{cold}} & \leftarrow & S_{\text{high}}, \tau_{\text{cold}} \end{array}$$

known as a Carnot cycle. The horizontal arrows indicate changes of S at constant τ, as heat is transferred into or out of the system. (To do this, the heat or source or sink must offer a slight temperature difference, but for the sake of reversibility the difference is assumed infinitesimal.) The vertical arrows indicate changes of τ at constant S. All four arrows involve changes in V. Say we start at the upper left. Along the rightward arrow, some heat (say Q_{in}) is input; meanwhile the gas expands, doing work against the cylinder's piston. During the downward

arrow, the gas is insulated and allowed to expand further, thus doing more work. Along the leftward arrow the gas outputs some heat (say Q_{out}); meanwhile the gas has to be compressed, so work is done on the gas through the piston. Finally during the upward arrow, the gas is insulated and compressed further, thus doing more work on it, until it returns to its initial state. (We cannot check directly that S has gone back to S_{low}, but since only two variables are independent here, it is enough to check that p, V are at their initial values.) We now have

$$\begin{aligned} &Q_{\text{in}} = \tau_{\text{hot}}(S_{\text{high}} - S_{\text{low}}), \quad Q_{\text{out}} = \tau_{\text{cold}}(S_{\text{high}} - S_{\text{low}}), \\ &Q_{\text{in}} > Q_{\text{out}}. \end{aligned} \tag{8.12}$$

Thus, a Carnot cycle takes heat Q_{in} from a source, gives Q_{out} to a sink, and converts the difference into net $p\,dV$ work.

(5) We see from (8.12) that in a Carnot cycle

$$\frac{Q_{\text{in}}}{Q_{\text{out}}} = \frac{\tau_{\text{hot}}}{\tau_{\text{cold}}} \tag{8.13}$$

regardless of the details. We can use this ratio of heats to define a scale for τ. In fact, the Kelvin temperature scale is defined in this way: for a Carnot cycle between 100 K and 200 K will have $Q_{\text{in}} = 2Q_{\text{out}}$, and so on.[1] The definition of 1 K is a matter of convention: it amounts to a multiplicative constant in $1/\tau$ and S but has no physical significance.

The preceding arguments show that it is possible to measure entropy macroscopically, through integrals of the type

$$\Delta S = \int \frac{dE + p\,dV}{\tau} \tag{8.14}$$

up to an additive constant. But the additive constant requires microscopic information and is not measurable within classical thermodynamics. □

So far we have used three kinds of thermodynamic variables: first S itself, then the extensive measurables (X, Y variables), and third the non-extensive conjugates (x, y variables). We can create a fourth kind by taking Legendre transforms; these are generically known as thermodynamic potentials, and are rather non-intuitive.

We have, of course, already seen $Z(x, y)$ as a Legendre transform of $S(X, Y)$. We can also define a sort of hybrid variable $Z'(x, Y)$, where the Legendre transform has been applied with respect to X but not Y:

$$S(X, Y) = \ln Z'(x, Y) + xX. \tag{8.15}$$

Analogous to (8.5) we have in this case

$$X = -\left(\frac{\partial \ln Z'}{\partial x}\right)_Y, \quad y = \left(\frac{\partial \ln Z'}{\partial Y}\right)_x. \tag{8.16}$$

[1] The number $1 - \tau_{\text{cold}}/\tau_{\text{hot}}$ is called the efficiency of a Carnot cycle, since it is the fraction of heat converted to work. See any book on thermodynamics for why this is interesting and important.

From (8.15) and (8.16) we have

$$d[\ln Z'] = -X\,dx + y\,dY. \tag{8.17}$$

$Z'(x, Y)$ is a kind of partition function, but different from $Z(x, y)$. In the sum (8.2) for $Z(x, y)$, x and y were parameters (originally Lagrange multipliers) used to enforce the required values of the extensive variables X and Y respectively. In $Z(x, Y)$, y does not appear, instead the extensive variable Y appears directly as a parameter, i.e., the sum over states ε_i is restricted to fixed Y:

$$Z'(x, Y) = \sum_{i,\ \text{fixed } Y} \exp[-x\,\varepsilon(\varepsilon_i)]. \tag{8.18}$$

We can go further and define yet another partition function via

$$S(X, Y) = \ln Z''(X, Y), \tag{8.19}$$

for which the explicit form is the weird-looking

$$Z''(X, Y) = \sum_{i,\ \text{fixed } X,Y} 1. \tag{8.20}$$

Here X, Y are constrained directly in the sum, and no Lagrange multipliers are required. This case amounts to working out all the microstates consistent with given X, Y, and assigning equal probability to all those microstates, since in this case maximum-entropy reduces to indifference.

When we need to compute a partition function, we are free to start with any of the partition-function sums (8.2), or (8.18), or (8.20), and derive the others via the Legendre transform relations (8.4), (8.15), or (8.19). In different situations, different partition functions may be easiest to compute.

We can invent further thermodynamic potentials by swapping dependent and independent variables and then taking Legendre transform. A good example is $G(\tau, p)$, introduced in (8.10) as a Legendre transform of $E(S, V)$:

$$G(\tau, p) = E - \tau S + pV. \tag{8.21}$$

G is called the Gibbs free energy, and equals $-\tau \ln Z$. Another example is the enthalpy $H(S, p)$: to derive it we start with $E(S, V)$ again and replace V by its conjugate variable while leaving S as it is, thus:

$$E(S, V) = -pV + H(S, p), \tag{8.22}$$

and $p = -(\partial E/\partial V)_S$ we have already identified (before equation 8.10) as the pressure.

The zoo of thermodynamic variables we have begun to explore admits some elegant identities. Going back to our original Legendre transform and differentiating (8.5) and (8.6) we get

$$\left(\frac{\partial X}{\partial y}\right)_x = \left(\frac{\partial Y}{\partial x}\right)_y, \quad \left(\frac{\partial y}{\partial X}\right)_Y = \left(\frac{\partial x}{\partial Y}\right)_X. \tag{8.23}$$

[If there were more than two X, Y variables we would take them in pairs.] Such identities are known as Maxwell relations, and are very useful for relating thermodynamic variables in often surprising ways.

PROBLEM 8.1: Applying the generic Maxwell relation formulas (8.23) to equation (8.21) we get:

$$\left(\frac{\partial S}{\partial p}\right)_\tau = -\left(\frac{\partial V}{\partial \tau}\right)_p, \quad \left(\frac{\partial \tau}{\partial V}\right)_S = -\left(\frac{\partial p}{\partial S}\right)_V. \tag{8.24}$$

You can derive two more identities as follows. Introduce a new thermodynamic potential $F(\tau, V)$ by a suitable Legendre transform. Relate F to the enthalpy H from (8.22). Then derive the additional Maxwell relations

$$\left(\frac{\partial S}{\partial V}\right)_\tau = \left(\frac{\partial p}{\partial \tau}\right)_V, \quad \left(\frac{\partial \tau}{\partial p}\right)_S = \left(\frac{\partial V}{\partial S}\right)_p.$$

$F(\tau, V)$ is known as the Helmholtz free energy. **[3]**

So far we have considered only equilibria and reversible processes, the latter being interpreted as sequences of equilibria. All these involve no net loss of information, and the total entropy remains constant. In non-equilibrium situations and during irreversible processes the theory is not valid, but if an irreversible process begins and ends at equilibrium states we can still relate the beginning and end states to each other.

What we mean by an irreversible process in thermodynamics is that some information about the past state of a system is lost. Thus the total entropy increases, and in place of the equality (8.7) we have

$$dS \geq x\, dX + y\, dY. \tag{8.25}$$

Equality holds for reversible changes. The most important form of (8.25) is

$$\tau\, dS \geq dE + p\, dV, \tag{8.26}$$

which replaces (8.9); it is a statement of the second law of thermodynamics.

We can calculate the entropy change in an irreversible process if we can connect the beginning and end states by an imaginary reversible process. Typically, this reversible process will introduce an imaginary external agent that makes the system do some extra work and pays the system an equivalent amount of heat (and hence entropy); the irreversible process neither does the extra work nor receives the heat, it generates the entropy by itself.

EXAMPLES [Some irreversible processes] A generic kind of irreversible process is an equalization process: two systems in different equilibria are brought together and allowed to settle into a common equilibrium.

A simple example is temperature equalization: two systems at τ_1 and $\tau_2 > \tau_1$ are brought together and allowed to exchange heat till their temperatures equalize. This would normally happen in a non-equilibrium, irreversible manner. But we can imagine an outside agent mediating the heat transfer reversibly. The agent takes an infinitesimal amount of heat δQ from the system at τ_1, and goes through a Carnot cycle to transfer heat to the system at τ_2. The heat transferred through the Carnot cycle will be $(\tau_2/\tau_1)\delta Q$, while $(1 - \tau_2/\tau_1)\delta Q$ is converted to work, with no net change of entropy. Now suppose the agent turns the work into heat and transfers it (still reversibly) to the system at τ_2. The system gains entropy

$$dS = \left(\frac{1}{\tau_2} - \frac{1}{\tau_1}\right)\delta Q > 0 \tag{8.27}$$

from the agent, but no net energy. For the irreversible process there is no agent, but the entropy increase is the same.

Another example is pressure equalization. Consider two gases at the same τ but different p_1, p_2. They are separated by a piston which is free to move, and hence the pressures get equalized. The system as a whole is insulated. Now imagine an agent that reversibly compresses the lower-pressure (say p_2) gas and lets the p_1 gas expand, both by infinitesimal volume dV. The net work done on the agent is $(p_1 - p_2)\,dV$. The agent turns this work into heat and distributes it between the gases so as to keep their temperatures equal. The system gains

$$dS = \frac{p_1 - p_2}{\tau}\,dV \tag{8.28}$$

from the agent, and no net energy. Again, for the irreversible process there is no agent, the system increases its own entropy.

A variant of the pressure-equalization example is gas expanding into a vacuum. Here (8.28) simplifies to

$$dS = \frac{p}{\tau}\,dV > 0. \tag{8.29}$$

Another variant of pressure equalization is mutual diffusion: two gases initially at the same temperature and pressure are allowed to mix. Each gas expands to the full volume, and though expansion is not into a vacuum, once again there is no $p\,dV$ work and no heat input. The entropy change is again given by (8.29), but one must add the contributions from both gases.

But what if the mutually diffusing gases are the same? Then there is no macroscopic change, but the above argument still gives us an entropy increase! This is the famous Gibbs paradox, and the source of it is a tacit assumption that the mutually diffusing gases are distinguishable. One can work around the Gibbs paradox by taking indistinguishability of gas molecules into account in different ad-hoc ways[1] but ultimately the paradox goes away only when we have a microscopic theory for indistinguishable particles, or quantum statistics, on which more later. □

[1] Gibbs's original resolution is in the closing pages of J.W. Gibbs *Elementary Principles of Statistical Mechanics* (originally 1902, reprinted by Ox Bow Press 1981). An ingenious resolution using (almost) entirely macroscopic arguments is given in Section 23 of Yu.B. Rumer & M.Sh. Ryvkin *Thermodynamics, Statistical Physics, and Kinetics* (Mir Publishers 1980).

PROBLEM 8.2: Consider two cylinders with pistons, each filled with a gas and the two connected by a valve. The pistons maintain different constant pressures p_1, p_2 as gas is forced through the valve. (The volumes of gas V_1, V_2 change of course.) The system is thermally insulated. Show that the enthalpy H of the gas remains constant as it is forced through the valve. Hence show that the entropy change is given by

$$dS = -\frac{V}{\tau}\, dp$$

integrated over an imaginary reversible process.

To integrate the entropy change we would need to know V/τ as a function of p at constant H; that depends on the gas's intermolecular forces and will not concern us here.

Constant-enthalpy expansion is very important for a different reason. It turns out that $(\partial\tau/\partial p)_H$ depends sensitively on intermolecular forces, and the expansion may be accompanied by heating or cooling, depending on the gas. With a cunning choice of gas, constant-enthalpy expansion (known as the Joule-Thompson or Joule-Kelvin effect) does most of our domestic refrigeration. **[3]**

The entropy-increase statement (8.25) implies that as an irreversible process approaches equilibrium at known values of X, Y, the entropy increases to a maximum. Rather than X, Y, however, we may prefer to take x, Y or x, y as independently known. In particular, many kinds of process (especially in chemistry) take place at given τ, p or τ, V, but not given E, V. So it is useful to Legendre-transform the inequality (8.25) to appropriate variables. If x, Y are fixed we take the differential $d[\ln Z'(x, Y)]$ from (8.17) and insert the inequality (8.25), which gives us

$$d[\ln Z'(x, Y)] \geq -X\, dx + y\, dY. \tag{8.30}$$

If x, y are fixed we instead compute the differential $d[\ln Z(x, y)]$ and insert the inequality (8.25), and that gives us

$$d[\ln Z(x, y)] \geq -X\, dx - Y\, dy. \tag{8.31}$$

The inequalities (8.30) and (8.31) mean that as a system approaches equilibrium and $S(X, Y)$ increases to a maximum, a thermodynamic potential also moves to an extremum. Conventionally, thermodynamic potentials are defined as Legendre transforms of E, rather than of S as here; that gives them an extra factor of $-\tau$ and makes them *decrease* to a minimum during an irreversible process.

PROBLEM 8.3: Derive thermodynamic potentials that are minimized by equilibria at (i) fixed τ, V, and (ii) fixed τ, p. **[2]**

We now move on from classical thermodynamics, and incorporate some microscopic considerations. In this book we will only manage the minimal theory of the so-called ideal gas, but with quantum mechanics included. 'Gas' is meant in a very general sense, it can refer to the gas of conduction electrons in a metal, or to liquid helium.

For the particles in such a gas there are a number (possibly infinity) of energy levels, say level 0 with energy ε_0, level 1 with energy ε_1, and so on. A microstate state of the gas has n_0 particles in level 0, n_1 particles in level 1, and so on. The measured thermodynamic parameters are the number of particles $N = \sum_i n_i$ and the total energy $E = \sum_i n_i \varepsilon_i$. The partition function is a sum over all possible sets $\{n_i\}$. Writing $-\mu/\tau$ and $1/\tau$ for the Lagrange multipliers corresponding to N and E respectively, we have

$$Z = \sum_{\{n_i\}} \exp\left[\textstyle\sum_i n_i(\mu - \varepsilon_i)/\tau\right]. \tag{8.32}$$

We can rearrange the double sum as

$$\prod_i \sum_{n_i} \exp[n_i(\mu - \varepsilon_i)/\tau], \tag{8.33}$$

in which case the $\{n_i\}$ are generated automatically by the product.

Quantum mechanics has two important consequences that we have to take into account when evaluating Z.

(i) Identical particles are indistinguishable, so permuting particles doesn't count as a separate state. We have already used this fact in writing (8.32)—for distinguishable particles there would have been some factorials.

(ii) Quantum particles come in two varieties: Bose-Einstein particles (or bosons) for which n_i may be any non-negative integer; and Fermi-Dirac particles (or fermions) for which n_i can only be 0 or 1. Electrons are fermions, while Helium atoms are bosons.

Working out the sum with and without the restriction on n_i, we have

$$Z = \prod_i \left(1 \pm e^{(\mu-\varepsilon_i)/\tau}\right)^{\pm 1}, \tag{8.34}$$

where the upper sign refers to the Fermi-Dirac case and the lower sign to the Bose-Einstein case. Evaluating the expectation values using (8.5) we have

$$\begin{aligned} E &= -\left(\frac{\partial \ln Z}{\partial\,[1/\tau]}\right)_{\mu/\tau} = \sum_i \frac{\varepsilon_i}{e^{(\varepsilon_i-\mu)/\tau} \pm 1}, \\ N &= \left(\frac{\partial \ln Z}{\partial\,[\mu/\tau]}\right)_{\tau} = \sum_i \frac{1}{e^{(\varepsilon_i-\mu)/\tau} \pm 1}, \end{aligned} \tag{8.35}$$

and consequently for the the occupancy of the i-th energy level we have

$$n_i = \frac{1}{e^{(\varepsilon_i-\mu)/\tau} \pm 1}. \tag{8.36}$$

If the energy levels ε_i are continuous, or spaced closely enough to be approximable as continuous, we can write continuous forms of (8.36) and (8.35):

$$n(\varepsilon) = \frac{\rho(\varepsilon)}{e^{(\varepsilon-\mu)/\tau} \pm 1}, \quad E = \int \varepsilon\, n(\varepsilon)\, d\varepsilon, \quad N = \int n(\varepsilon)\, d\varepsilon. \tag{8.37}$$

where $\rho(\varepsilon)$ is the density of states.

EXAMPLE [The classical ideal gas] In gases of our everyday experience, usually $\varepsilon \gg \mu$, in which case the partition function (8.34) simplifies to

$$\ln Z \simeq e^{\mu/\tau} \sum_i e^{-\varepsilon_i/\tau} \tag{8.38}$$

and the distinction between Bose-Einstein and Fermi-Dirac disappears; the common limiting form is called the classical or Maxwell-Boltzmann gas. If the gas is monoatomic, the density of states is expressible in terms of the momentum p as

$$\rho(\varepsilon) = \frac{4\pi}{h^3} p^2 \, dp \, dV, \quad \varepsilon = \frac{p^2}{2m} \tag{8.39}$$

where h is a physical constant (Planck's constant) and m is the particle mass. Approximating the partition function (8.38) by an integral and using (M.4) from page 70 gives us

$$\ln Z = e^{\mu/\tau} \frac{\tau^{3/2} V}{\Gamma}, \quad \Gamma = \left(\frac{h^2}{2\pi m} \right)^{3/2}. \tag{8.40}$$

Using (8.5) we have

$$\begin{aligned} N &= \left(\frac{\partial \ln Z}{\partial [\mu/\tau]} \right)_\tau = \ln Z, \quad \frac{\mu}{\tau} = \ln \left(\frac{N\Gamma}{V\tau^{3/2}} \right), \\ E &= -\left(\frac{\partial \ln Z}{\partial [1/\tau]} \right)_{\mu/\tau} = \tfrac{3}{2}\tau N, \end{aligned} \tag{8.41}$$

Substituting from the above into the Legendre transform (8.4) we have

$$S = \ln Z + \frac{E}{\tau} - \frac{\mu}{\tau} N = \left(\tfrac{5}{2} + \ln \left(\frac{\tau^{3/2} V}{N\Gamma} \right) \right) N. \tag{8.42}$$

And here we recover the well-known result that for a monoatomic ideal gas $\tau V^{2/3}$ is constant in constant-S processes.

We may now invoke the relation (8.8) for p/τ from classical thermodynamics, which gives us

$$\frac{p}{\tau} = \left(\frac{\partial S}{\partial V} \right)_{E,N} = \frac{N}{V} \tag{8.43}$$

(since $\tau = \frac{2}{3} E/N$ and thus τ is constant in the partial derivative) and we have finally

$$pV = N\tau, \tag{8.44}$$

which is the equation of state for a classical gas. □

PROBLEM 8.4: Photons in a cavity form a Bose-Einstein gas, with the difference that photons can be freely absorbed and emitted by any matter in the cavity, or its walls, so there is no constraint on N. For photons, the energy level is given by

$$\varepsilon = h\nu,$$

where ν is the frequency. The density of states has the form

$$\rho(\varepsilon)\,d\varepsilon = \frac{8\pi}{c^3}\nu^2\,d\nu\,dV$$

where c is the speed of light. Show that the energy per unit volume in photons with frequencies between ν and $\nu + d\nu$ is

$$\frac{8\pi h}{c^3}\frac{\nu^3\,d\nu}{e^{h\nu/\tau} - 1}.$$

This is known, for historical reasons, as the blackbody radiation formula.

A good example of a photon gas is the radiation left over from the Big Bang. The expansion of the universe has reduced temperature and frequencies a lot, and today the typical ν is in the microwave range and the temperature is 2.73 K. **[2]**

Appendix: Miscellaneous Formulas

This Appendix collects a number of formulas that are used in the main text, and are often useful in other places too. Most of them will not be derived here, but any book on mathematical methods for physical sciences [1] will derive most or all of them.

1. First, a reminder of the definition of e:

$$e \equiv \lim_{N\to\infty} \left(1 + \tfrac{1}{N}\right)^N. \tag{M.1}$$

2. Integrals for marginalization often involve Gamma and Beta functions which are defined as

$$\Gamma(n+1) \equiv \int_0^\infty x^n e^{-x}\, dx = n! \tag{M.2}$$

and

$$B(m,n) \equiv \int_0^1 x^m (1-x)^n\, dx = \frac{m!n!}{(m+n+1)!}. \tag{M.3}$$

Note that m, n in these two formulas need not be integers. Useful non-integer cases are

$$\Gamma\left(\tfrac{1}{2}\right) = \sqrt{\pi}, \quad \Gamma\left(\tfrac{3}{2}\right) = \tfrac{1}{2}\sqrt{\pi}. \tag{M.4}$$

Changing variables in (M.2) gives another useful integral:

$$\int_0^\infty \sigma^{-n} \exp\left(-\frac{K}{2\sigma^2}\right) \frac{d\sigma}{\sigma} = \Gamma\left(\frac{n}{2}\right)\left(\frac{K}{2}\right)^{-n/2}. \tag{M.5}$$

A useful approximation for the Gamma function is Stirling's formula (actually the leading term in an asymptotic series)

$$n! \simeq \sqrt{2\pi}\, e^{-n} n^{n+\frac{1}{2}}, \tag{M.6}$$

and its further truncation

$$\ln(n!) \simeq n \ln n - n. \tag{M.7}$$

3. Also useful for marginalization is the most beautiful of all elementary integrals

$$\int_{-\infty}^\infty e^{-x^2}\, dx = \sqrt{\pi} \tag{M.8}$$

[1] My favourites are G.B. Arfken & H.J. Weber, *Mathematical Methods for Physicists* (Academic Press 1995—first edition 1966) for detail and P. Dennery & A. Krzywicki, *Mathematics for Physicists* (Dover Publications 1996—first edition 1967) for conciseness.

and a number of integrals derived from it. These are sometimes called Gaussian integrals.

From (M.8) we have

$$\int_{-\infty}^{\infty} e^{-\frac{1}{2}\alpha x^2}\,dx = \sqrt{2\pi}\alpha^{-\frac{1}{2}}. \tag{M.9}$$

Invoking Leibnitz's rule for differentiating under the integral sign, we differentiate (M.9) with respect to α, and divide the result by (M.9) itself, to get

$$\frac{\int_{-\infty}^{\infty} x^2 e^{-\frac{1}{2}\alpha x^2}\,dx}{\int_{-\infty}^{\infty} e^{-\frac{1}{2}\alpha x^2}\,dx} = \alpha^{-1} \tag{M.10}$$

Going to two dimensions, we have

$$\begin{aligned}&\int_{-\infty}^{\infty} e^{-\frac{1}{2}(\alpha x^2+\beta y^2-2\gamma xy)}\,dx\,dy \\ &= \left(\frac{2\pi}{\beta}\right)^{\frac{1}{2}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(\alpha\beta-\gamma^2)x^2/\beta}\,dx \\ &= \frac{2\pi}{\sqrt{\alpha\beta-\gamma^2}}\end{aligned} \tag{M.11}$$

and

$$\frac{\int_{-\infty}^{\infty} x^2 \exp\left[-\frac{1}{2}(\alpha x^2+\beta y^2-2\gamma xy)\right]dx\,dy}{\int_{-\infty}^{\infty} \exp\left[-\frac{1}{2}(\alpha x^2+\beta y^2-2\gamma xy)\right]dx\,dy} = \frac{\beta}{\alpha\beta-\gamma^2}. \tag{M.12}$$

Formulas (M.11) and (M.12) look awful, but their content is simple: (M.11) shows that marginalizing out one of the variables in a 2D Gaussian leaves a 1D Gaussian, and (M.12) tells us the dispersion of the latter.

The multi-dimensional generalizations of (M.9) and (M.10) can be written concisely in matrix notation. Let $\mathbf{x}$ denote a column vector with L components each ranging from $-\infty$ to ∞ and $\mathbf{H}$ denoting a constant L × L matrix. Then $\frac{1}{2}\mathbf{x}^T\mathbf{H}\mathbf{x}$ generalizes $-\frac{1}{2}\alpha x^2$ to L dimensions. We then have the formulas[1]

$$\int \exp\left(-\tfrac{1}{2}\mathbf{x}^T\mathbf{H}\mathbf{x}\right) dx_1..dx_L = \frac{(2\pi)^{L/2}}{\sqrt{|\det \mathbf{H}|}} \tag{M.13}$$

and

$$\frac{\int \mathbf{x}\mathbf{x}^T \exp\left(-\tfrac{1}{2}\mathbf{x}^T\mathbf{H}\mathbf{x}\right) dx_1..dx_L}{\int \exp\left(-\tfrac{1}{2}\mathbf{x}^T\mathbf{H}\mathbf{x}\right) dx_1..dx_L} = \mathbf{H}^{-1}. \tag{M.14}$$

We see that $\mathbf{H}^{-1}$ behaves like σ^2 in the Gaussian. Moreover, marginalizing some of the variables in a multi-dimensional Gaussian amounts to discarding the corresponding rows and columns of $\mathbf{H}^{-1}$.

[1] See the Appendix of Sivia for a derivation.

4. Now we have a very brief summary of Fourier transforms. A function $f(x)$ and its Fourier transform, say $F(k)$, are related by

$$\begin{aligned} F(k) &\equiv \int_{-\infty}^{\infty} e^{ikx} f(x)\, dx \\ f(x) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-ikx} F(k)\, dk. \end{aligned} \tag{M.15}$$

Naturally, $f(x)$ needs to fall off fast enough at large $|x|$ for the Fourier transform to exist. Fourier transforms have many useful properties, four of which are used in chapter 3.

(i) The Fourier transform of $f'(x)$ is $(-ik)F(k)$, (easy to verify).

(ii) The Fourier transform of the convolution of two functions is the product of the Fourier transforms (the convolution theorem):

$$\int_{-\infty}^{\infty} f(y)g(x-y)\, dy = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-ikx} F(k)G(k)\, dk. \tag{M.16}$$

It is not hard to derive this, at least if we are allowed to freely change the order of integration.

(iii) The Fourier transform of a Gaussian is another Gaussian:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}x^2/\sigma^2} \Leftrightarrow F(k) = e^{-\frac{1}{2}k^2\sigma^2}. \tag{M.17}$$

The functional form stays the same, only σ is replaced by $1/\sigma$.

(iv) Although the Fourier transform of a constant is not defined, it can be given a meaning in the limit. Consider the limit of (M.17) as $\sigma \to 0$: $f(x)$ gets arbitrarily sharp and narrow but the area under it remains 1. We have

$$f(x) \to \delta(x), \quad F(k) \to 1, \tag{M.18}$$

where $\delta(x)$ (called a δ function or Dirac δ) is a curious beast with the property that for any well-behaved $h(x)$

$$\int \delta(x-a)\, h(x)\, dx = h(a) \tag{M.19}$$

if the integral goes over $x = a$, and 0 otherwise.

5. Finally, a note on the infinite product (1.5) on page 4, which has little to do with anything else in this book, but is too pretty to pass by. Consider

$$\prod_{p \text{ prime}} \left(1 - p^{-z}\right)^{-1} \tag{M.20}$$

where $z > 1$. We can binomial-expand each term in the product to get

$$\prod_{p \text{ prime}} \left(1 + p^{-z} + p^{-2z} + p^{-3z} + \cdots\right). \tag{M.21}$$

Expanding out this product, and using the fact that every natural number n has a unique prime factorization, gives

$$\sum_{n=1}^{\infty} \frac{1}{n^z}. \tag{M.22}$$

The sum (M.22) is the Riemann Zeta function, originally introduced for real z by Euler and later analytically continued to complex z by Riemann. This function has a spooky tendency to appear in any problem to do with primes. For $z = 2$ it equals[1] $\pi^2/6$.

[1] See, for example, equation (6.90) in R.L. Graham, D.E. Knuth, & O. Patashnik, *Concrete Mathematics* (Addison-Wesley 1990).

Hints and Answers

ANSWER 1.1: Adding probabilities for 1st, 2nd,... turn

$$\mathfrak{prob}(\text{Tortoise rolls 6}) = \tfrac{1}{6}\left[\tfrac{5}{6} + \left(\tfrac{5}{6}\right)^3 + \ldots\right] = \tfrac{5}{11}.$$

Alternatively

$$\mathfrak{prob}(\text{Tortoise rolls 6}) = \tfrac{5}{6}\,\mathfrak{prob}(\text{Achilles rolls 6}) = \tfrac{5}{11}.$$

ANSWER 1.2:

$$\mathfrak{prob}(65\,|\,N) = \mathfrak{prob}(1729\,|\,N) = N^{-1}, \quad \text{provided } N \geq 1729$$

and the prior on N is N^{-1}. Hence

$$\mathfrak{prob}(N\,|\,65, 1729) \propto N^{-3},\ N \geq 1729.$$

ANSWER 1.3: The instruction sets GGG and RRR give coincidence no matter what the switch settings are. Each of the remaining instruction sets (RRG RGR RGG GRR GRG GGR) gives non-coincidence for only 4 out of the 9 possible switch combinations.

ANSWER 2.1: It is the probability of having n events in the time interval $[0, \tau]$, which is $e^{-m\tau}m^n\tau^n/n!$, and then one more event in the time interval $[\tau, \tau + d\tau]$. As the Gamma function formula (M.2) shows, the result is normalized.

ANSWER 2.2: Modify the derivation of (2.16), put $K = L$, and then use the Beta function (M.3) for the integral over u.

ANSWER 2.3: Straightforward manipulations of the sums.

ANSWER 2.4: $(n_1 - n_{12})(n_2 - n_{12})/n_{12}$.

ANSWER 3.1: $\mathfrak{cf}(k) = \mathfrak{sin}(k)/k$.

ANSWER 3.2: Like equation (3.16), with $\Delta^2 = \frac{1}{2}$ in units of a step length.

ANSWER 3.3: Since 0.3% of a Gaussian lies outside 3σ, the difference between the extremes in 1024 trials will be about 6σ. In fact in this example $m = 137.5$ and $\sigma = 7$.

ANSWER 3.4: A Gaussian, times four factors of the type (3.43), times a combinatorial factor ${}^5C_2 \times 3$.

ANSWER 3.5: In the Gaussian approximation 130 is a 3σ result.

ANSWER 4.1: In polar coordinates

$$x = r\sin\theta\cos\phi, \quad y = r\sin\theta\sin\phi, \quad z = r\cos\theta$$

the volume element

$$dx\,dy\,dz = r^2\,dr\,d(\cos\theta)\,d\phi$$

so choose r according to $r^2 f(r)$ and choose ϕ and $\cos\theta$ uniformly.

For $g(r)$, start with $f(r)$ and then use rejection for the extra factor.

ANSWER 4.2: See Figure A.1.

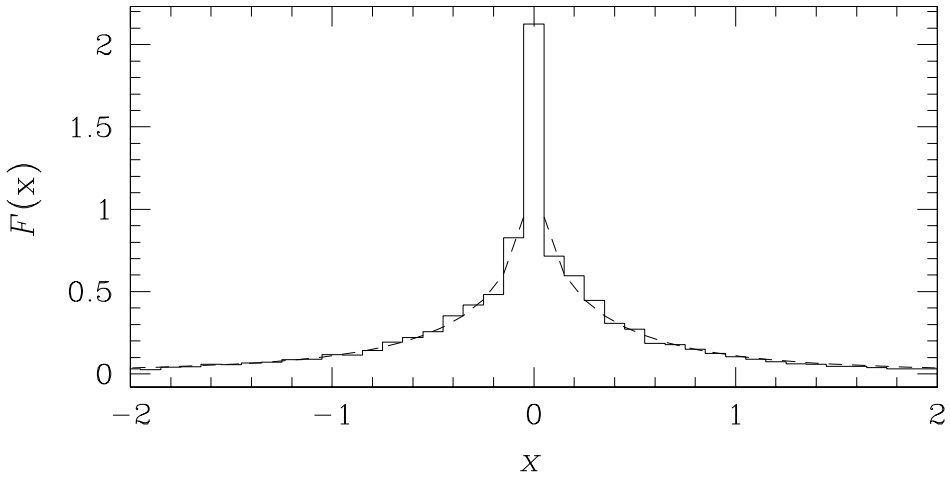

Figure A.1: This uses 40000 points. The integrable singularity at $x = 0$ is harmless.

ANSWER 5.1: We get

$$y = 1.015 - 1.047x$$

The bootstrap matrix is

$$\begin{pmatrix} 0.0017 & -0.0030 \\ -0.0030 & 0.0072 \end{pmatrix}$$

versus $\sigma^2\mathbf{C}$ (using equation 5.26 on page 40 for σ^2) of

$$\begin{pmatrix} 0.0014 & -0.0022 \\ -0.0022 & 0.0060 \end{pmatrix}$$

ANSWER 5.2: Replacing the K points with the mean and replacing σ with $\sigma/\sqrt{K}$, changes Q^2 only by a constant.

ANSWER 5.3: $\mathfrak{prob}(\mu\,|\,D) \propto (d\omega/d\mu)\ \mathfrak{prob}(\omega\,|\,D)$. Then use the Gaussian approximation formula (3.35).

ANSWER 5.4: Marginalize along the line, which gives

$$\mathfrak{prob}(x_1, y_1\,|\,m, c) \propto (1+m^2)^{-\frac{1}{2}}\ \mathfrak{exp}\left[-\frac{Q^2}{2(1+m^2)}\right]$$
$$Q^2 = (y_1 - mx_1 - c)^2$$

which is similar (5.6) but with $\sqrt{1+m^2}$ in place of σ. The argument of the exponent is the perpendicular distance from (x_1, y_1) to the line.

ANSWER 5.5: The polynomial is

$$1 - x + x^2 - x^3$$

and Figure A.2 shows data and fits. The factor

$$\Gamma\left(\tfrac{1}{2}L\right)\Gamma\left(\tfrac{1}{2}(N-L)\right)\left(\mathbf{P}^T\cdot\mathbf{C}\cdot\mathbf{P}\right)^{-L/2}\left(d^2 - \mathbf{P}^T\cdot\mathbf{C}\cdot\mathbf{P}\right)^{(L-N)/2}$$

from equation (5.32) evaluates to

$$10^{28.5},\quad 10^{36.3},\quad 10^{43.3},\quad 10^{41.0},\quad 10^{39.0},\quad 10^{37.3}$$

for polynomials of degree 1 to 6.

ANSWER 5.6: Use $-2\ln(\text{likelihood})$, or

$$2\sum\nolimits_i \left(\ln(n_i!) + m - n_i \ln m\right)$$

which reduces to χ^2 for large m.

ANSWER 5.7: Beware the denominator in Bayes' theorem!

In our first invocation of Bayes' theorem, the denominator $\mathfrak{prob}(D)$ is a well-defined constant. In our second invocation, the denominator is $\mathfrak{prob}(\chi^2)$. But since χ^2 depends on ω, marginalizing out ω to get $\mathfrak{prob}(\chi^2)$ has no meaning.

ANSWER 6.1: There could be any number of unsampled events before the next sampled event, so

$$\mathfrak{prob}(x\,|\,S) = \sum_{n=0}^{\infty}\left(\mathfrak{prob}(\overline{S})\right)^n\ \mathfrak{prob}(S, x).$$

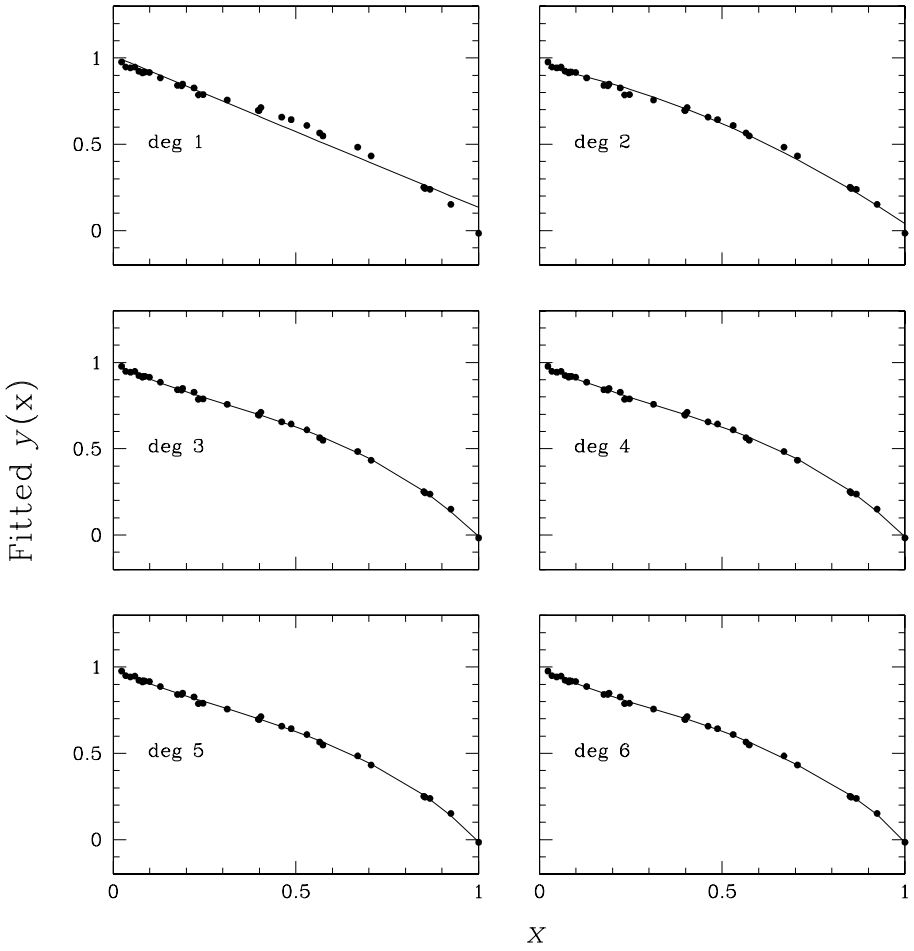

Figure A.2: Fitted polynomials of degrees 1 to 6.

ANSWER 6.2: Figure A.3 shows the data my program generated. (The binning is just for the figure, the analysis does not bin the data.) From the figure we would expect, that a_1, b_1 would be easy to infer, a_2, b_2 much harder. My Metropolis program gets the following 90% confidence intervals.

$$\begin{aligned} w &= 0.88^{+0.08}_{-0.16} \\ a_1 &= 0.48^{+0.03}_{-0.04} \quad b_1 = 0.21^{+0.06}_{-0.06} \\ a_2 &= -0.29^{+0.35}_{-0.31} \quad b_2 = 0.42^{+0.22}_{-0.09} \end{aligned}$$

ANSWER 6.3: In equation (6.6), replace $M!$ by $e^{-M}M^M$ and $S!$ by $e^{-S}S^S$.

ANSWER 6.4: See Figures A.4 and A.5.

ANSWER 7.1: The variance $\langle \mathfrak{x}^2(\varepsilon)\rangle - \langle \mathfrak{x}(\varepsilon)\rangle^2$.

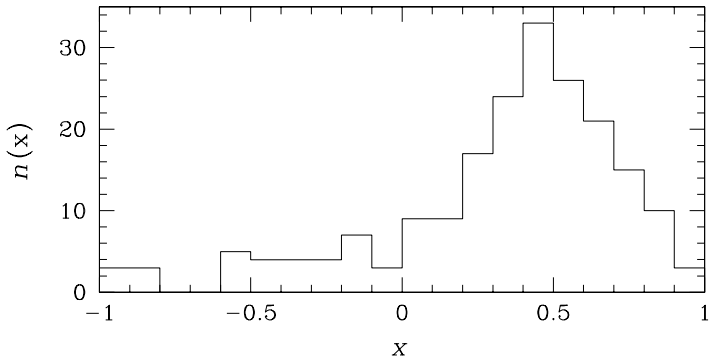

Figure A.3: Histogram of detections in the two-lighthouse problem.

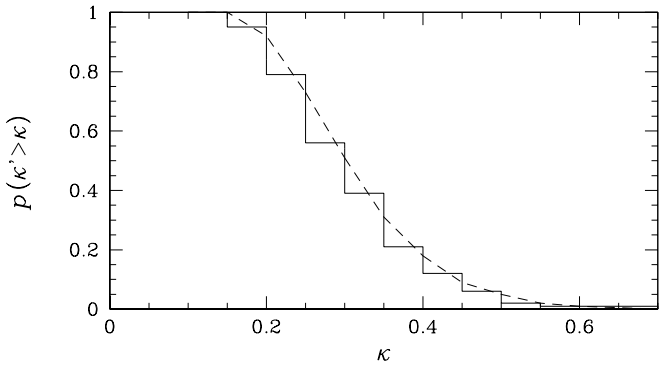

Figure A.4: Plot of p-values for the KS statistic, for $N_u = 10, N_v = 20$. The histogram is from simulations, the curve is the asymptotic formula.

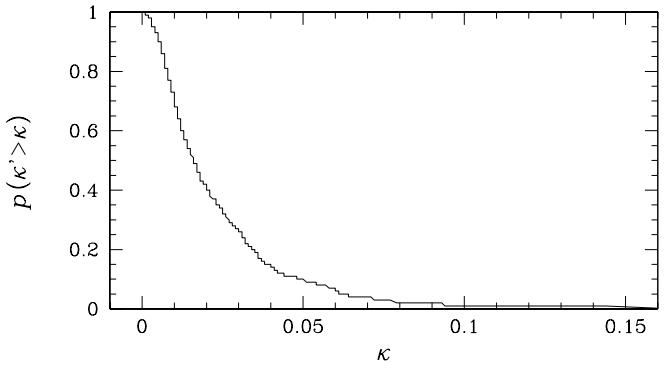

Figure A.5: Plot of p-values for the statistic $\kappa = \sum_i x_i x_{i+1}$ (suggested by A. Austin) also for $N_u = 10, N_v = 20$.

ANSWER 8.1:

$$F(\tau, V) = E(S, V) - \tau S$$
$$H(S, p) = E(S, V) + pV$$
$$H(S, p) = \tau S + Vp + F(\tau, V)$$

ANSWER 8.2: Since p_1, p_2 are constant, work done on the gas is $p_2V_2 - p_1V_1$, and hence $H = U + pV$ is constant. Using

$$dH = \tau\, dS + V\, dp$$

gives dS.

ANSWER 8.3: $F(\tau, V)$ and $G(\tau, p)$.

ANSWER 8.4: Substitute in (8.37), with $\mu = 0$ because there is no particle-number constraint.

Index