Data Analysis

A Bayesian Tutorial

D. S. SIVIA

Rutherford Appleton Laboratory

and

St Catherine's College, Oxford

CLARENDON PRESS • OXFORD

OXFORD
UNIVERSITY PRESS
Great Clarendon Street, Oxford OX2 6DP
Oxford University Press is a department of the University of Oxford
and furthers the University's aim of excellence in research, scholarship,
and education by publishing worldwide in
Oxford New York
Athens Auckland Bangkok Bogotá Buenos Aires Calcutta
Cape Town Chennai Dar es Salaam Delhi Florence Hong Kong Istanbul
Karachi Kuala Lumpur Madrid Melbourne Mexico City Mumbai
Nairobi Paris São Paulo Singapore Taipei Tokyo Toronto Warsaw
and associated companies in Berlin Ibadan

Published in the United States
by Oxford University Press Inc., New York

First published 1996

Reprinted 1997, 1998

A catalogue record for this book is available from the British Library

Library of Congress Cataloging in Publication Data
(Data applied for)

ISBN 0 19 851889 7 (Pbk)

Printed in Great Britain by Bookcraft (Bath) Ltd.
Midsomer Norton, North Somerset

To mum and dad

1
The basics

'There are three kinds of lies: lies, damned lies and statistics.'

Mark Twain (1924) probably had politicians in mind when he reiterated Disraeli's famous remarks. Scientists, we hope, would never use data in such a selective manner to suit their own ends. But, alas, the analysis of data is often the source of some exasperation even in an academic context. On hearing comments like 'the result of this experiment was inconclusive, so we had to use statistics', we are frequently left wondering as to what strange tricks have been played on the data.

The sense of unease which many of us have towards the subject of statistics is largely a reflection of the inadequacies of the 'cook book' approach to data analysis that we are taught as undergraduates. Rather than being offered a few clear principles, we are usually presented with a maze of tests and procedures; while most seem to be intuitively reasonable individually, their interrelations are not obvious. This apparent lack of a coherent rationale leads to considerable apprehension because we have little feeling for which test to use or, more importantly, why.

Fortunately, data analysis does not have to be like this! A more unified and logical approach to the whole subject is provided by the probability formulations of Bayes and Laplace. Bayes' ideas (published in 1763) were used very successfully by Laplace (1812), but were then allegedly discredited and largely forgotten until they were rediscovered by Jeffreys (1939). In more recent times they have been expounded by Jaynes and others. This book is intended to be an introductory tutorial to the Bayesian approach, including modern developments such as maximum entropy.

1.1 Introduction: deductive logic versus plausible reasoning

Let us begin by trying to get a general feel for the nature of the problem. A schematic representation of deductive logic is shown in Fig. 1.1(a): given a cause, we can work out its consequences. The sort of reasoning used in pure mathematics is of this type: that is to say, we can derive many complicated and useful results as the logical consequence of a few well-defined axioms. Everyday games of chance also fall into this category. For example, if we are told that a fair coin is to be flipped ten times, we can calculate the chances that all ten tosses will produce heads, or that there will be nine heads and one tail, and so on.

Most scientists, however, face the reverse of the above situation: Given that

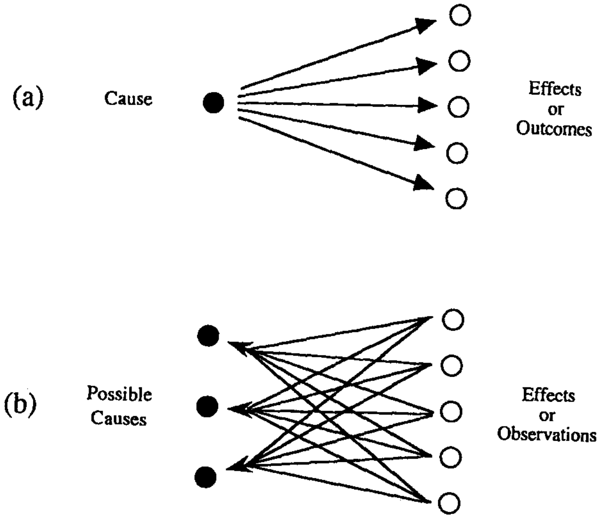

Fig. 1.1 A schematic representation of (a) deductive logic, or pure mathematics, and (b) plausible reasoning, or inductive logic.

certain effects have been observed, what is (are) the underlying cause(s)? To take a simple example, suppose that ten flips of a coin yielded seven heads: Is it a fair coin or a biased one? This type of question has to do with inductive logic, or plausible reasoning, and is illustrated schematically in Fig. 1.1(b); the greater complexity of this diagram is designed to indicate that it is a much harder problem. The most we can hope to do is to make the best inference based on the experimental data and any prior knowledge that we have available, reserving the right to revise our position if new information comes to light. Around 500 BC, Herodotus said much the same thing: 'A decision was wise, even though it led to disastrous consequences, if the evidence at hand indicated it was the best one to make; and a decision was foolish, even though it led to the happiest possible consequences, if it was unreasonable to expect those consequences.'

Even though plausible reasoning is rather open-ended, are there any general quantitative rules which apply for such inductive logic? After all, this issue is central to data analysis.

1.2 Probability: Cox and the rules for logical consistent reasoning

In 1946, Richard Cox pondered the quantitative rules necessary for logical and consistent reasoning. He started by considering how we might express our relative beliefs in the truth of various propositions. For example: (a) it will rain tomorrow; (b) King Harold died after being hit in the eye by an arrow at the

battle of Hastings in AD 1066; (c) this is a fair coin; (d) this coin is twice as likely to come up heads as tails; and so on. The minimum requirement for expressing our relative beliefs in the truth of these propositions in a consistent fashion is that we rank them in a *transitive* manner. In other words, if we believe (a) more than (b), and (b) more than (c), then we must necessarily believe (a) more than (c); if this were not so, we would continue to argue in circles. Such a transitive ranking can easily be obtained by assigning a *real* number to each of the propositions in a manner so that the larger the numerical value associated with a proposition, the more we believe it.

Cox actually took this much for granted—as being obvious—and wondered what rules these numbers had to obey in order to satisfy some simple requirements of logical consistency. He began by putting forward two axioms. The first is very straightforward: if we specify how much we believe that something is true, then we must have implicitly specified how much we believe it's false. He didn't assume any particular form for this relationship, but took it as being reasonable that one existed. The second axiom is slightly more complicated: if we first specify how much we believe that (proposition) Y is true, and then state how much we believe that X is true given that Y is true, then we must implicitly have specified how much we believe that both X and Y are true. Again, he only asserted that these quantities were related but did not specify how. To work out the actual form of the relationships, Cox used the rules of *Boolean logic*, ordinary algebra, and the constraint that if there were several different ways of using the same information then we should always arrive at the same conclusions irrespective of the particular analysis-path chosen. He found that this consistency could only be ensured if the real numbers we had attached to our beliefs in the various propositions could be *mapped* (or transformed) to another set of real *positive* numbers which obeyed the usual rules of probability theory:

$$\mathrm{prob}(X\,|\,I) + \mathrm{prob}(\overline{X}\,|\,I) = 1 \tag{1.1}$$

and

$$\mathrm{prob}(X,Y\,|\,I) = \mathrm{prob}(X\,|\,Y,I) \times \mathrm{prob}(Y\,|\,I)\,. \tag{1.2}$$

Here $\overline{X}$ denotes the proposition that X is false, the vertical bar '|' means 'given' (so that all items to the right of this conditioning symbol are taken as being true) and the comma is read as the conjunction 'and'.

Equation (1.1) is called the *sum rule*, and states that the probability that X is true plus the probability that X is false is equal to one. Technically, Cox's work only shows that the sum should be set to a constant (so that there is an overall scale in the problem); we merely choose it to be unity by convention. Equation (1.2) is called the *product rule*. It states that the probability that both X and Y are true is equal to the probability that X is true given that Y is true times the probability that Y is true (irrespective of X).

Note that we have made all the probabilities conditional on I, to denote the relevant background information at hand. Basically, there is no such thing as an absolute probability! For example, the probability that we assign to the proposition 'it will rain this afternoon' will depend on whether there are dark clouds or a clear blue sky in the morning; it will also be affected by whether or not we saw the weather forecast. Although the conditioning on I is often omitted in calculations, to reduce algebraic cluttering, we must never forget its existence. A failure to state explicitly all the relevant background information, and assumptions, is frequently the real cause of heated debates about data analysis.

1.3 Corollaries: Bayes' theorem and marginalisation

The sum and product rules of eqns (1.1) and (1.2) form the basic algebra of probability theory. Many other results can be derived from them. Amongst the most useful are two known as *Bayes' theorem* and *marginalisation*:

$$\operatorname{prob}(X \mid Y, I) = \frac{\operatorname{prob}(Y \mid X, I) \times \operatorname{prob}(X \mid I)}{\operatorname{prob}(Y \mid I)} \tag{1.3}$$

and

$$\operatorname{prob}(X \mid I) = \int_{-\infty}^{+\infty} \operatorname{prob}(X, Y \mid I)\, \mathrm{d}Y. \tag{1.4}$$

Bayes' theorem, or eqn (1.3), follows directly from the product rule. To see this, let's rewrite eqn (1.2) with X and Y *transposed* (or interchanged):

$$\operatorname{prob}(Y, X \mid I) = \operatorname{prob}(Y \mid X, I) \times \operatorname{prob}(X \mid I).$$

Since the probability of both 'Y and X' being true must be logically the same as that of 'X and Y' being true, so that $\operatorname{prob}(Y, X \mid I) = \operatorname{prob}(X, Y \mid I)$, the right-hand side of the above can be equated to that of eqn (1.2); hence, we obtain eqn (1.3). It is invaluable because it enables us to turn things around with respect to the conditioning symbol: it relates $\operatorname{prob}(X \mid Y, I)$ to $\operatorname{prob}(Y \mid X, I)$. The importance of this property to data analysis becomes apparent if we replace X and Y by *hypothesis* and *data*:

$$\operatorname{prob}(\mathit{hypothesis} \mid \mathit{data}, I) \propto \operatorname{prob}(\mathit{data} \mid \mathit{hypothesis}, I) \times \operatorname{prob}(\mathit{hypothesis} \mid I).$$

The power of Bayes' theorem lies in the fact that it relates the quantity of interest, the probability that the hypothesis is true given the data, to the term that we have a better chance of being able to assign, the probability that we would have observed the measured data if the hypothesis was true.

The various terms in Bayes' theorem have formal names. The quantity on the far right, $\operatorname{prob}(\mathit{hypothesis} \mid I)$, is called the *prior* probability; it represents our state of knowledge (or ignorance) about the truth of the hypothesis before we have analysed the current data. This is modified by the experimental measurements through the *likelihood function*, or $\operatorname{prob}(\mathit{data} \mid \mathit{hypothesis}, I)$, and yields the *posterior* probability, $\operatorname{prob}(\mathit{hypothesis} \mid \mathit{data}, I)$, representing our state of knowledge about the truth of the hypothesis in the light of the data. In a sense, Bayes' theorem encapsulates the process of learning. We should note, however, that

the equality of eqn (1.3) has been replaced with a proportionality, because the term $\operatorname{prob}(\mathit{data} \mid I)$ has been omitted. This is fine for many data analysis problems, such as those involving *parameter estimation*, since the missing denominator is simply a normalisation constant (not depending explicitly on the hypothesis). In some situations, such as *model selection*, this term plays a crucial role. For that reason, it is sometimes given the special name of *evidence*.

The marginalisation equation, (1.4), should seem a little peculiar: up to now, Y has stood for a given proposition, so how can we integrate over it? Before we answer that question, let us first consider the marginalisation equation for our standard X and Y propositions. It would take the form

$$\operatorname{prob}(X \mid I) = \operatorname{prob}(X, Y \mid I) + \operatorname{prob}(X, \overline{Y} \mid I). \tag{1.5}$$

This can be derived by expanding $\operatorname{prob}(X, Y \mid I)$ with the product rule of eqn (1.2):

$$\operatorname{prob}(X, Y \mid I) = \operatorname{prob}(Y, X \mid I) = \operatorname{prob}(Y \mid X, I) \times \operatorname{prob}(X \mid I),$$

and adding to the left- and right-hand sides, respectively, a similar expression for $\operatorname{prob}(X, \overline{Y} \mid I)$, to give

$$\operatorname{prob}(X, Y \mid I) + \operatorname{prob}(X, \overline{Y} \mid I) = [\operatorname{prob}(Y \mid X, I) + \operatorname{prob}(\overline{Y} \mid X, I)] \times \operatorname{prob}(X \mid I).$$

Since eqn (1.1) ensures that the quantity in square brackets on the right is equal to unity, we obtain eqn (1.5). Stated verbally, eqn (1.5) says that the probability that X is true, irrespective of whether or not Y is true, is equal to the sum of the probability that both X and Y are true and the probability that X is true and Y is false.

Now suppose that instead of having a proposition Y, and its negative counterpart $\overline{Y}$, we have a whole set of alternative possibilities: $Y_1, Y_2, \ldots, Y_M = \{Y_k\}$. For example, let's imagine that there are M (say five) candidates in a presidential election; then Y_1 could be the proposition that the first candidate will win, Y_2 the proposition that the second candidate will win, and so on. The probability that X is true, for example that unemployment will be lower in a year's time, irrespective of whoever becomes president, is given by

$$\operatorname{prob}(X \mid I) = \sum_{k=1}^{M} \operatorname{prob}(X, Y_k \mid I). \tag{1.6}$$

This is just a generalisation of eqn (1.5), and can easily be derived in an analogous manner as long as

$$\sum_{k=1}^{M} \operatorname{prob}(Y_k \mid X, I) = 1. \tag{1.7}$$

This *normalisation* requirement is satisfied if the $\{Y_k\}$ form a *mutually exclusive* and *exhaustive* set of possibilities. That is to say, if one of the Y_k's is true then all the others must be false, but one of them has to be true.

The actual form of the marginalisation equation in eqn (1.4) applies when we go to the *continuum limit.* For example, when we consider an arbitrarily large number of propositions about the range in which (say) the Hubble constant H_0 might lie. As long as we choose the intervals in a contiguous fashion, and cover a big enough range of values for H_0, we will have a mutually exclusive and exhaustive set of possibilities. Equation (1.4) is then just a generalisation of eqn (1.6), with $M \to \infty$, where we have used the usual shorthand notation of *calculus.* In this context, Y now represents the numerical value of a parameter of interest (such as H_0) and the integrand $\text{prob}(X,Y\,|\,I)$ is technically a *probability density function* rather than a probability. Strictly speaking, therefore, we should denote it by a different symbol, such as $\text{pdf}(X,Y\,|\,I)$, where

$$\text{pdf}(X,Y=y\,|\,I)=\lim_{\delta y\to 0}\frac{[\text{prob}(X,y\leqslant Y<y+\delta y\,|\,I)]}{\delta y}, \tag{1.8}$$

and the probability that the value of Y lies in a finite range between y_1 and y_2 (and X is also true) is given by

$$\text{prob}(X,y_1\leqslant Y<y_2\,|\,I)=\int_{y_1}^{y_2}\text{pdf}(X,Y\,|\,I)\,\mathrm{d}Y. \tag{1.9}$$

Since 'pdf' is also a common abbreviation for *probability distribution function*, which can pertain to a discrete set of possibilities, we will simply use 'prob' for anything related to probabilities; this has the advantage of preserving a uniformity of notation between the continuous and discrete cases. Thus, in the continuum limit, the normalisation condition of eqn (1.7) takes the form

$$\int_{-\infty}^{+\infty}\text{prob}(Y\,|\,X,I)\,\mathrm{d}Y=1. \tag{1.10}$$

Marginalisation is a very powerful device in data analysis because it enables us to deal with *nuisance parameters*; that is, quantities which necessarily enter the analysis but are of no intrinsic interest. The unwanted background signal present in many experimental measurements, and instrumental parameters which are difficult to calibrate, are examples of nuisance parameters. Before going on to see how the rules of probability can be used to address data analysis problems, let's take a brief look at the history of the subject.

1.4 Some history: Bayes and Laplace versus orthodox statistics

About three hundred years ago, people started to give serious thought to the question of how to reason in situations in which it is not possible to argue with certainty. James Bernoulli (1713) was perhaps the first to articulate the problem, perceiving the difference between the deductive logic applicable to games of chance and the inductive logic required for everyday life. The open question for him was how the mechanics of the former might help to tackle the inference problems of the latter.

Reverend Thomas Bayes is credited with providing an answer to Bernoulli's question, in a paper published posthumously by a friend (1763). The present-day form of the theorem which bears his name is actually due to Laplace (1812). Not only did Laplace rediscover Bayes' theorem for himself, in far more clarity than did Bayes, but he also put it to good use in solving problems in celestial mechanics, medical statistics and, by some accounts, even jurisprudence. Despite Laplace's numerous successes, his development of probability theory was rejected by mathematicians who took over the subject from the mid-nineteenth century.

The problem was not really one of substance but of concept. To the pioneers such as the Bernoullis, Bayes and Laplace, a probability represented a *degree-of-belief* or plausibility: how much they though that something was true, based on the evidence at hand. To the nineteenth-century scholars, however, this seemed too vague and subjective an idea to be the basis of a rigorous mathematical theory. So they redefined probability as the *long-run relative frequency* with which an event occurred, given (infinitely) many repeated (experimental) trials. Since frequencies can be measured, probability was now seen as an objective tool for dealing with *random* phenomena.

Although the frequency definition appears to be more objective, its range of validity is also far more limited. For example, Laplace used (his) probability theory to estimate the mass of Saturn, given orbital data that were available to him from various astronomical observatories. In essence, he computed the posterior pdf for the mass M, given the data and all the relevant background information I (such as a knowledge of the laws of classical mechanics): $\text{prob}(M\,|\,\{data\},I)$; this is shown schematically in Fig. 1.2. To Laplace, the (shaded) area under the posterior pdf curve between m_1 and m_2 was a measure of how much he believed that the mass of Saturn lay in the range $m_1\leqslant M<m_2$. As such, the position of the maximum of the posterior pdf represents a best estimate of the mass; its width, or spread, about this optimal value gives an indication of the uncertainty in the estimate. Laplace stated that '...it is a bet of 11,000 to 1 that the error of this result is not 1/100th of its value'. He would have won the bet, as another 150 years' accumulation of data has changed the

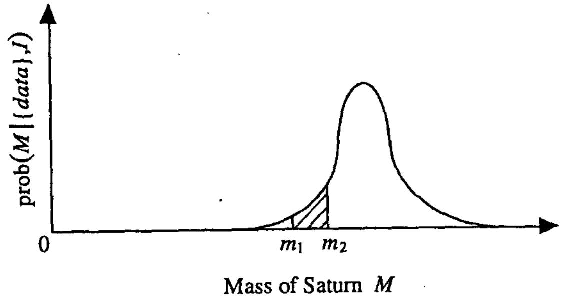

Fig. 1.2 A schematic illustration of the result of Laplace's probability analysis of the mass of Saturn.

estimate by only 0.63%! According to the frequency definition, however, we are not permitted to use probability theory to tackle this problem. This is because the mass of Saturn is a constant and not a *random variable*; therefore, it has no frequency distribution and so probability theory cannot be used.

If the pdf of Fig. 1.2 had to be interpreted in terms of the frequency definition, we would have to imagine a large *ensemble* of universes in which everything remains constant apart from the mass of Saturn. As this scenario appears quite far-fetched, we might be inclined to think of Fig. 1.2 in terms of the distribution of the measurements of the mass in many repetitions of the experiment. Although we are at liberty to think about a problem in any way that facilitates its solution, or our understanding of it, having to seek a frequency interpretation for every data analysis problem seems rather perverse. For example, what do we mean by the 'measurement of the mass' when the data consist of orbital periods? Besides, why should we have to think about many repetitions of an experiment that never happened? What we really want to do is to make the best inference of the mass given the (few) data that we actually have; this is precisely the Bayes and Laplace view of probability.

Faced with the realisation that the frequency definition of probability theory did not permit most real-life scientific problems to be addressed, a new subject was invented—*statistics*! To estimate the mass of Saturn, for example, one has to relate the mass to the data through some function called the statistic; since the data are subject to 'random' noise, the statistic becomes the random variable to which the rules of probability theory can be applied. But now the question arises: How should we choose the statistic? The frequentist approach does not yield a natural way of doing this and has, therefore, led to the development of several alternative schools of *orthodox* or *conventional* statistics. The masters, such as Fisher, Neyman and Pearson, have provided a variety of different principles, which has merely resulted in a plethora of tests and procedures without any clear underlying rationale. This lack of unifying principles is, perhaps, at the heart of the shortcomings of the cook-book approach to statistics that we are taught today.

The frequency definition of probability merely gives the impression of a more objective theory. In reality it just makes life more complicated by hiding the difficulties under the rug, only for them to resurface in a less obvious guise. Indeed, it is not even clear that the concept of 'randomness' central to orthodox statistics is any better-defined than the idea of 'uncertainty' inherent in Bayesian probability theory. For example, we might think that the numbers generated by a call to a function like RAN on a computer constitute a random process: the frequency of the numbers will be distributed uniformly between 0 and 1, and their sequential order will appear haphazard. The illusory nature of this randomness would become obvious, however, if we knew the *algorithm* and the *seed* for the function RAN (for then we could predict the sequence of numbers output by the computer). At this juncture, some might argue that, in contrast to our simple illustration above, *chaotic* and *quantum* systems provide examples of physical situations which are intrinsically random. In fact, chaos theory merely underlines the point that we are trying to make: the apparent randomness in the long-term behaviour of a classical system arises because we do not, or cannot, know its initial conditions well enough; the actual temporal evolution is entirely deterministic, and obeys Newton's Second Law of Motion. The quantum case is more difficult to address, since its interpretation (as opposed to its technical success) is still an open question for many people today (Jaynes 1990; Selleri 1990). The sub-atomic world not withstanding, it seems that 'randomness' represents our inability to predict things which, in turn, reflects our lack of knowledge about the system of interest. This is again consistent with the Bayes and Laplace view of probability, rather than the asserted physical objectivity of the frequentist approach.

To emphasize this last point, that a probability represents a state of knowledge rather than a physically real entity, consider the following example of Jaynes (1989). We are told that a dark bag contains five red balls and seven green ones. If this bag is shaken well, and a ball selected at 'random', then most of us would agree that the probability of drawing a red ball is 5/12 and the probability of drawing a green one is 7/12. If the ball is not returned to the bag, then it seems reasonable that the probability of obtaining a red or green ball on the second draw will depend on the outcome of the first (because there will be one less red or green ball left in the bag). Now suppose that we are not told the outcome of the first draw, but are given the result of the second one: Does the probability of the first draw being red or green change with the knowledge of the second? Initially, many of us would be inclined to say 'no': at the time of the first draw, there were still five red balls and seven green ones in the bag; so, the probabilities for red and green should still be 5/12 and 7/12 irrespective of the outcome of the second draw. The error in this argument becomes obvious if we consider the extreme example of a bag containing only one red and one green ball. Although the second draw cannot affect the first in a physical sense, a knowledge of the second result does influence what we can infer about the outcome of the first one: if the second ball was green, then the first one must have been red; and vice versa. Thus (conditional) probabilities represent *logical* connections rather than *causal* ones.

The concerns about the subjectivity of the Bayesian view of probability are understandable, and the aims of the orthodox statisticians to create a more objective theory quite laudable. Unfortunately, the frequentist approach does not achieve this goal: neither does the concept of randomness appear very rigorous, or fundamental, under scrutiny and nor does the arbitrariness of the choice of the statistic make it seem objective. In fact, the presumed shortcomings of the Bayesian approach merely reflect a confusion between subjectivity and the difficult technical question of how probabilities should be assigned. The popular argument goes that if a probability represents a degree-of belief, then it must be subjective, because my belief could be different from yours. The Bayesian view is that a probability does indeed represent how much we believe that something is true, but that this belief should be based on all the relevant information available. While this makes the assignment of probabilities an open-ended question, because the information at my disposal may not be the same as that accessible to you, it is not the same as subjectivity. It simply means that probabilities are always conditional, and this conditioning must be stated explicitly. As Jaynes has pointed out, objectivity demands only that two people having the same information should assign the same probability; this principle has played a key role in the modern development of the (objective) Bayesian approach.

In 1946, Richard Cox tried to get away from the controversy of the Bayesian versus frequentist view of probability. He decided to look at the question of plausible reasoning afresh, from the perspective of logical consistency. Somewhat to his surprise, he found that the only rules which met his requirements were those of probability theory. Although the sum and product rules of probability are straightforward to prove for frequencies (with the aid of a *Venn diagram*), Cox's work shows that their range of validity goes much further. Rather than being restricted to just frequencies, probability theory constitutes the basic calculus for logical and consistent plausible reasoning; for us, that means scientific inference (which is the purpose of data analysis). So, Laplace was right all along!

1.5 An outline of the book

The aim of this book is to show how probability theory can be used directly to address data analysis problems in a straightforward manner. We will start, in Chapter 2, with the simplest type of examples: namely, those involving the estimation of the value of a single parameter. They serve as a good first encounter with Bayes' theorem in action, and allow for a useful discussion about *error-bars* and *confidence intervals*. The examples are extended to two, and then several, parameters in Chapter 3, enabling us to introduce the additional concepts of *correlation* and marginalisation. In Chapter 4, we will see how the same principles used for parameter estimation can be applied to the problem of model selection.

Although Cox's work shows that plausibilities, represented by real numbers, should be manipulated according to the rules of probability theory, it does not tell us how to assign them in the first place. We turn to this basic question of assigning probabilities in Chapter 5, where we will meet the important principle of *maximum entropy* (MaxEnt). It may seem peculiar that we leave so fundamental a question to such a late stage, but it is intentional. People often have the impression that Bayesian analysis relies heavily on the use of clever probability assignments and is, therefore, not generally applicable if these are not available. Our aim is to show that even when armed only with a knowledge of pdfs familiar from high school (*binomial*, *Poisson* and *Gaussian*), and *naïveté* (a *flat*, or *uniform*, pdf), probability theory still provides a powerful tool for obtaining useful results for many data analysis problems. Indeed, we will find that most conventional statistical procedures implicitly assume such elementary assignments. Of course we can only do better by thinking more deeply about the most appropriate pdf for any given problem, but the point is that it is not usually crucial in practice.

In Chapter 6, we consider *non-parametric estimation*; that is to say, problems in which we know so little about the object of interest that we are unable to describe it adequately in terms of a few parameters. Here we will encounter MaxEnt once again, but in the slightly different guise of *image processing*. In the last chapter, we focus our attention on the subject of *experimental design*. This concerns the question of 'what are the best data to collect', in contrast to most of this book, which deals with 'what is the optimal way of analysing these data' (assuming that we have them already). This reciprocal question can also be addressed by probability theory, and is of great importance because the benefits of good experimental (or instrumental) design can far outweigh the rewards of the sophisticated analysis of poorer data.

Most of the examples used in this book involve continuous pdfs, since this is usually the nature of parameters in real life. As mentioned in Section 1.3, this can simply be considered as the limiting case of an arbitrarily large number of discrete propositions. Jaynes correctly warns us, however, that difficulties can sometimes arise if we are not careful in carrying out the limiting procedure explicitly; this is often the underlying cause of so-called paradoxes of probability theory. Such problems also occur in other mathematical calculations, which are quite unrelated to probability. To illustrate this point, consider the evaluation of the following *double integral* taken from a standard maths textbook for science undergraduates (Stephenson 1961):

$$A = \iint_R \frac{x-y}{(x+y)^3}\,\mathrm{d}x\,\mathrm{d}y,$$

where the integration is over the square region R defined by $0 \leqslant x \leqslant 1$ and $0 \leqslant y \leqslant 1$. Depending on whether we chose to integrate first with respect to x, or with y, we would tend to write (respectively)

$$A = \int_0^1 \mathrm{d}y \int_0^1 \frac{x-y}{(x+y)^3}\,\mathrm{d}x \qquad \text{or} \qquad A = \int_0^1 \mathrm{d}x \int_0^1 \frac{x-y}{(x+y)^3}\,\mathrm{d}y,$$

and proceed to do the integrals. Contrary to our expectations, we would find that the two alternatives give different results: $+1/2$ and $-1/2$. Stephenson resolves this apparent paradox by pointing out that the function to be integrated is badly behaved at the origin and so we should not expect to be able to interchange the order of integration. Although there is truth in the argument, it is only a partial answer. The correct way to deal with the problem is to consider the limiting procedure:

$$A = \lim_{\beta\to 0} \int_\beta^1 \mathrm{d}y \int_\beta^1 \frac{x-y}{(x+y)^3}\,\mathrm{d}x \qquad \text{or} \qquad A = \lim_{\beta\to 0} \int_\beta^1 \mathrm{d}x \int_\beta^1 \frac{x-y}{(x+y)^3}\,\mathrm{d}y,$$

where we carry out the integrals with a lower limit of β, and then examine the result as β goes to zero at the end. Now we will find that both calculations give the same answer of nought. This result can be confirmed, after some thought, from symmetry arguments (as the integrand is antisymmetric with respect to the diagonal $x = y$).

Despite Jaynes' insightful advice, and the above example, we will throw caution to the wind and take a somewhat cavalier approach to technical formalities in this book. This is simply because most of the concerns boil down to ones of good mathematical style in general, rather than difficulties with probability theory itself. As such, we prefer to sacrifice some rigour for increasing clarity in this introductory text. The intention is that the novice should come away with enough know-how and confidence to tackle many real data analysis problems and, perhaps, the inspiration to study the subject in greater depth. Above all, we hope that the reader will be able to share Laplace's view that: '*Probability theory is nothing but common sense reduced to calculation*'.

2

Parameter estimation I

Parameter estimation is a common data analysis problem. Like Laplace, for example, we may be interested in knowing the mass of Saturn; or, like Millikan, the charge of the electron. In the simplest case, we are only concerned with the value of a single parameter; such elementary examples are the focus of this chapter. They serve as a good introduction to the use of Bayes' theorem and allow for a discussion of error-bars and confidence intervals.

2.1 Example 1: is this a fair coin?

Let us begin with the analysis of data from a simple coin-tossing experiment. Suppose I told you that I had been to Las Vegas for my holidays, and had come across a very strange coin in one of the casinos; given that I had observed 4 heads in 11 flips, do you think it was a *fair* coin? By fair, we mean that we would be prepared to lay an even 50:50 bet on the outcome of a flip being a head or a tail. In ascribing the property of fairness (solely) to the coin we are, of course, assuming that the coin-tosser was not skilled enough to be able to control the initial conditions of the flip (such as the angular and linear velocities). If we decide that the coin was fair, the question which follows naturally is how sure are we that this was so; if it was not fair, how unfair do we think it was?

A sensible way of formulating this problem is to consider a large number of contiguous propositions, or hypotheses, about the range in which the *bias-weighting* of the coin might lie. If we denote the bias-weighting by H, then $H = 0$ and $H = 1$ can represent a coin which produces a tail or a head on every flip, respectively. There is a continuum of possibilities for the value of H between these limits, with $H = 1/2$ indicating a fair coin. The propositions could then be, for example: (a) $0.00 \leqslant H < 0.01$; (b) $0.01 \leqslant H < 0.02$; (c) $0.02 \leqslant H < 0.03$; and so on. Our state of knowledge about the fairness, or the degree of unfairness, of the coin is then completely summarised by specifying how much we believe these various propositions to be true. If we assign a high probability to one (or a closely grouped few) of these propositions, compared to the others, then this would indicate that we were confident in our estimate of the bias-weighting. If there was no such strong distinction, then it would reflect a high level of ignorance about the nature of the coin.

In the light of the data, and the above discussion, our inference about the fairness of this coin is summarised by the conditional pdf: prob$(H \mid \{data\}, I)$. This is, of course, shorthand for the limiting case of a continuum of propositions for the value of H; that is to say, the probability that H lies in an infinitesimally narrow range between h and $h+\delta h$ is given by $\text{prob}(H=h\,|\,\{data\},I)\,\mathrm{d}H$. To estimate this posterior pdf, we need to use Bayes' theorem (eqn 1.3); it relates the pdf of interest to two others, which are easier to assign:

$$\text{prob}(H\,|\,\{data\},I) \propto \text{prob}(\{data\}\,|\,H,I) \times \text{prob}(H\,|\,I). \qquad (2.1)$$

Note that we have omitted the denominator $\text{prob}(\{data\}\,|\,I)$, as it does not involve the bias-weighting explicitly, and replaced the equality by a proportionality. If required, we can evaluate the missing constant subsequently from the normalisation condition of eqn (1.10):

$$\int_0^1 \text{prob}(H\,|\,\{data\},I)\,\mathrm{d}H = 1. \qquad (2.2)$$

The prior pdf, $\text{prob}(H\,|\,I)$, on the far right-hand side of eqn (2.1), represents what we know about the coin given only the information I that we are dealing with a 'strange coin from Las Vegas'. Since casinos can be rather dubious places, we should keep a very open mind about the nature of the coin; a simple probability assignment which reflects this is a uniform, or flat, pdf:

$$\text{prob}(H\,|\,I) = \begin{cases} 1 & 0 \leqslant H \leqslant 1, \\ 0 & \text{otherwise}. \end{cases} \qquad (2.3)$$

This prior state of knowledge, or ignorance, is modified by the data through the likelihood function: $\text{prob}(\{data\}\,|\,H,I)$. It is a measure of the chance that we would have obtained the data that we actually observed, if the value of the bias-weighting was given (as known). If, in the conditioning information I, we assume that the flips of the coin were independent events, so that the outcome of one did not influence that of another, then the probability of obtaining the data 'R heads in N tosses' is given by the binormal distribution:

$$\text{prob}(\{data\}\,|\,H,I) \propto H^R(1-H)^{N-R}. \qquad (2.4)$$

We leave a formal derivation of this pdf to Chapter 5, but point out that eqn (2.4) seems reasonable because H is the chance of obtaining a head on any flip, and there were R of them, and $1-H$ is the corresponding probability for a tail, of which there were $N-R$. For simplicity, an equality has again been replaced by a proportionality; this is permissible since the omitted terms contain factors of only R and N, rather than the quantity of interest H.

According to eqn (2.1), the product of eqns (2.3) and (2.4) yields the posterior pdf that we require; it represents our state of knowledge about the nature of the coin in the light of the data. To get a feel for this result, it is instructive to see how this pdf evolves as we obtain more and more data pertaining to the coin. This is done with the aid of data generated in a computer simulation, and the results of their analyses are shown in Fig. 2.1. The panel in the top left-hand corner shows the posterior pdf for H given no data; it is, of course, the same as the prior of eqn (2.3). It indicates that we have no more reason to believe that the coin is fair than we have to think that it is double-headed, double-tailed or of any other intermediate bias-weighting.

Suppose that the coin is flipped once and it comes up heads: What can we now say about the value of H? The resulting posterior pdf, shown in the second panel of Fig. 2.1, goes to zero for $H=0$ and rises linearly to having the greatest value at $H=1$. Note that we have not normalised the pdf according to eqn (2.2), but have scaled it so that its greatest value is unity. Based purely on this single datum, it is most probable that the (strange) coin has two heads; after all, we don't have any empirical evidence that it even has a tail yet. Although $H=1$ is in some ways our 'best' estimate so far, the posterior pdf indicates that this value is not much more probable than many others. The only thing we can really be sure about is that the coin is not double-tailed; hence, the posterior pdf is zero at $H=0$.

If the coin is flipped for a second time and again comes up heads, the posterior pdf becomes slightly more peaked towards $H=1$ (proportional to H^2); this is plotted in the third panel of Fig. 2.1. Since we still haven't seen a tail, our inclinations following the first datum are just reinforced. As soon as a tail is obtained, however, the posterior pdf for $H=1$ also drops to zero; we are then sure that the coin is not double-headed either. The next panel shows the resultant pdf, given that the third flip produced a tail (proportional to $(1-H)\times H^2$). If the fourth flip also comes up tails, then the maximum of the posterior pdf is at $H=0.5$. It then becomes most probable that the coin is fair, but there is still a large degree of uncertainty in this estimate; this is indicated graphically in panel 5 of Fig. 2.1.

The remainder of Fig. 2.1 shows how the posterior pdf for H evolves as the number of data analysed becomes larger and larger. We see that the position of the maximum wobbles around, but that the amount by which it does so decreases with the increasing amount of data. The width of the posterior pdf also becomes narrower with more data, indicating that we are becoming increasingly confident in our estimate of the bias-weighting. For the coin in this example, the best estimate of H eventually converges to 0.25. This was, of course, the value chosen to simulate the flips; they can be thought of as coming from a tetrahedral coin with a head on one face and tails on the other three!

Different priors

Most people tend to be happy with the binomial distribution for the likelihood function, but worry about the prior pdf. The uniform assignment of eqn (2.3) was chosen mostly for its simplicity; it is just a naïve way of encoding a lot of initial ignorance about the coin. How would our inference about the fairness of the coin have changed if we had chosen a different prior?

To address this question, let's repeat the analysis of the data above with two alternative prior pdfs; the results are shown in Fig. 2.2. The solid line is for the uniform pdf used in Fig. 2.1, and is included for ease of comparison. One of the alternative priors is peaked around $H=0.5$ and reflects our background information that most coins are fair (even in Las Vegas); it is plotted with a dashed line. It has a width which is broad enough to comfortably accommodate the

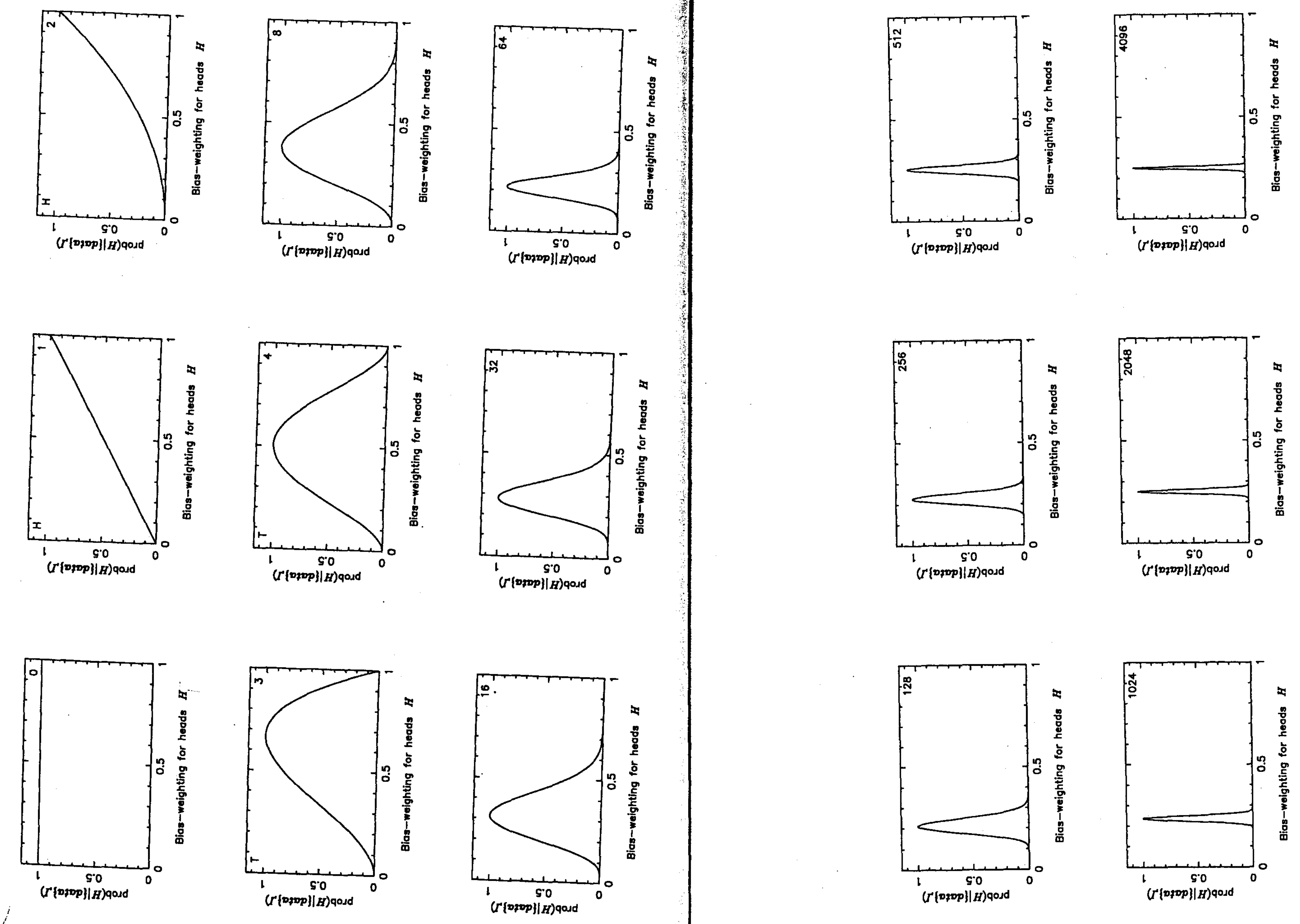

Fig. 2.1 The evolution of the posterior pdf for the bias-weighting of a coin, prob(H|{$data$}, I), as the number of data available increases. The figure in the top right-hand corner of each panel shows the number of data analysed; in the early panels, the H or T in the top left-hand corner shows whether the result of the (last) flip was a head or a tail.

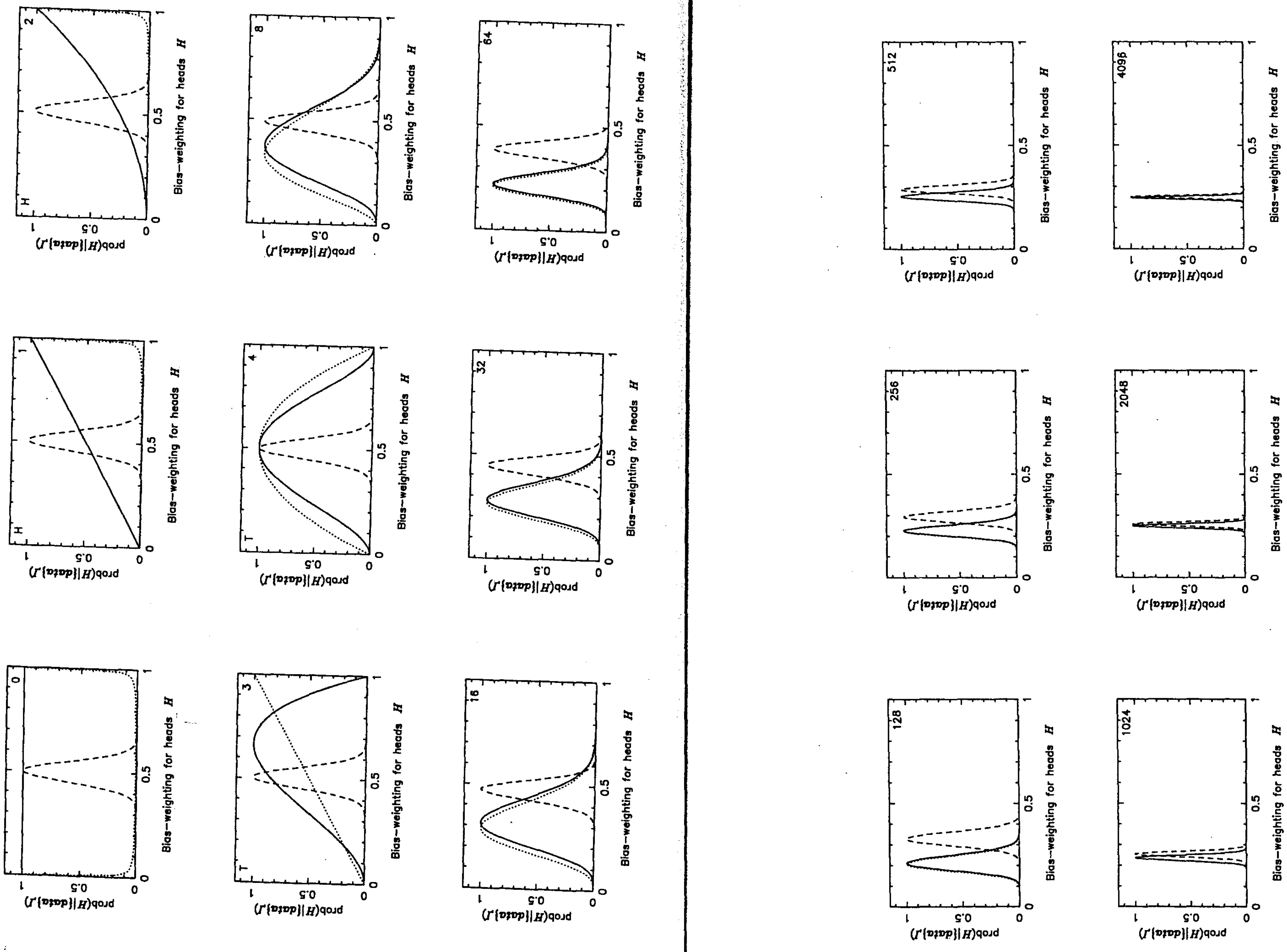

Fig. 2.2 The effect of different priors, prob$(H \mid I)$, on the posterior pdf for the bias-weighting of a coin. The solid line is the same as in Fig. 2.1, and is included for ease of comparison. The cases for two alternative priors, reflecting slightly different assumptions in the conditioning information I, are shown with dashed and dotted lines.

possibility that H might be as low as 0.35, or as high as 0.65, but expresses considerable doubt about greater levels of biased behaviour. The second alternative, shown with a dotted line, is very sharply peaked at $H = 0$ and $H = 1$; it indicates that we would expect a 'strange coin from a casino' to be heavily biased, one way or another.

Figure 2.2 shows how the posterior pdfs for the different priors evolve, as more and more data become available; they have again been scaled so that the greatest value for each is equal to unity. We find that when there are few data, the resulting posteriors are different in detail; as the number of data increases, they all become more sharply peaked and converge to the same answer. This seems quite reasonable. The outcome of only a few flips tells us little about the fairness of the coin. Our state of knowledge after the analysis of these data is, therefore, strongly dependent on what we knew or assumed before the results; hence, the posteriors are somewhat different. As the empirical evidence grows, we are eventually led to the same conclusions irrespective of our initial beliefs; the posterior pdf is then dominated by the likelihood function, and the choice of the prior becomes largely irrelevant.

Two further curious points may be noted: (i) it takes quite a lot of flips to be able to estimate the bias-weighting with some degree of confidence (about a thousand to pin it down between 0.2 and 0.3); (ii) the posterior pdfs for the solid and dotted lines converge quite quickly, but the dashed case takes much longer. The answer to the first observation is that it just does, but the number of flips required will depend on the actual bias-weighting of the coin. If the coin was tossed ten times and came up tails on every occasion, we would rapidly conclude that it was biased; on the other hand, the result of 45 heads and 55 tails in 100 flips would still leave us somewhat uncertain as to whether the coin was fair. With regard to the second point, both the flat (solid) and the spiky (dotted) priors encode a large degree of ignorance about the nature of the coin. Despite the peculiar shape of the latter, it is fairly flat for most values of H; the strange-looking spikes at $H = 0$ and $H = 1$ disappear as soon as a head and a tail have been observed. The 'fair-minded' prior (dashed line), however, claims to be moderately well-informed about the character of the coin regardless of the data. It, therefore, takes much more to be convinced that the coin is not fair. Even though the prior probability for $H = 0.5$ was about a million times greater than that of $H = 0.25$, a thousand flips were enough to drag it (kicking and screaming, perhaps) to the conclusion that the value of H was less than 0.3 (but more than 0.2).

Sequential or one-step data analysis?

If we have a set of data $\{D_k\}$ comprising the outcome of N flips of a coin ($k = 1, 2, \ldots, N$), then Bayes' theorem tells us that the posterior pdf for H is given by

$$\text{prob}(H \mid \{D_k\}, I) \propto \text{prob}(\{D_k\} \mid H, I) \times \text{prob}(H \mid I). \tag{2.5}$$

This is a one-step process in that we consider the data collectively, as a whole. As an alternative, we could also think of analysing the data sequentially (as they arrive). That is to say, we start by computing the posterior pdf based on the first datum D_1, prob$(H \mid D_1, I)$, and use it as the prior for the analysis of the second datum D_2; the new posterior could then be used as the prior for the third datum, and so on. If we continue this procedure, will we obtain the same result as the one-step approach?

To simplify matters, let us consider the case of just two data. The posterior pdf for H, based on both, is merely a special case of eqn (2.5) for $N = 2$:

$$\text{prob}(H \mid D_2, D_1, I) \propto \text{prob}(D_2, D_1 \mid H, I) \times \text{prob}(H \mid I). \tag{2.6}$$

Equally well, we could use Bayes' theorem to express the posterior in terms of pdfs conditional on D_1 throughout (like the background information I):

$$\text{prob}(H \mid D_2, D_1, I) \propto \text{prob}(D_2 \mid H, D_1, I) \times \text{prob}(H \mid D_1, I). \tag{2.7}$$

This shows that the prior in eqn (2.6) can certainly be replaced by the posterior pdf based on the first datum, but the other term doesn't quite look like the likelihood function for the second datum. Implicit in the assignment of the binomial pdf of eqn (2.4), however, was the assumption (subsumed in I) that the data were *independent*. This means that, given the value of H, the result of one flip does not influence what we can infer about the outcome of another: mathematically, we write this as

$$\text{prob}(D_2 \mid H, D_1, I) = \text{prob}(D_2 \mid H, I).$$

Substituting this in eqn (2.7) yields the desired relationship, showing that the two data can either be analysed together or one after the other. This argument can be extended to the third datum D_3, giving

$$\text{prob}(H \mid D_3, D_2, D_1, I) \propto \text{prob}(D_3 \mid H, I) \times \text{prob}(H \mid D_2, D_1, I),$$

and repeated until all the data have been included. We should not really be surprised that both the one-step and sequential methods of analysis give the same answer; after all, the requirements of such consistency was what led Cox to the rules of probability theory in the first place.

When one first learns about Bayes' theorem, and sees how the data modify the prior through the likelihood function, there is occasionally a temptation to use the resulting posterior pdf as the prior for a re-analysis of the same data. It would be erroneous to do this, and the results quite misleading. In order to justify any data analysis procedure we must be able to relate the pdf of interest, to others used in its calculation, through the sum and product rules of probability (or their corollaries). If we cannot do this then the analysis will be suspect, at best, and open to logical inconsistencies. For the case of this proposed 'bootstrapping' we would, in fact, be trying to relate the posterior pdf to itself; this can only be done by an equality, and not through the likelihood function. If we persist in our folly, and keep repeating it, the resulting posterior pdfs will become sharper and sharper; we will just fool ourselves into thinking that the quantity of interest can be estimated far more accurately than is warranted by the data.

2.2 Reliabilities: best estimates, error-bars and confidence intervals

We have seen how the posterior pdf encodes our inference about the value of a parameter, given the data and the relevant background information. Often, however, we wish to summarise this with just two numbers: the best estimate and a measure of its reliability. Since the probability (density) associated with any particular value of the parameter is a measure of how much we believe that it lies in the neighbourhood of that point, our best estimate is given by the maximum of the posterior pdf. If we denote the quantity of interest by X, with a posterior pdf $P = \mathrm{prob}(X \mid \{data\}, I)$, then the best estimate of its value X_0 is given by the condition

$$\left.\frac{\mathrm{d}P}{\mathrm{d}X}\right|_{X_0} = 0. \tag{2.8}$$

Strictly speaking, we should also check the sign of the second *derivative* to ensure that X_0 represents a maximum rather than a minimum (or a point of inflexion):

$$\left.\frac{\mathrm{d}^2P}{\mathrm{d}X^2}\right|_{X_0} < 0. \tag{2.9}$$

In writing the derivatives of P with respect to X we are, of course, assuming that X is a continuous parameter. If it could only take discrete values, our best estimate would still be that which gave the greatest posterior probability; it's just that we couldn't then use the calculus notation of eqns (2.8) and (2.9), because the *gradients* are not defined (since the increment in X cannot be infinitesimally small).

To obtain a measure of the reliability of this best estimate, we need to look at the width or spread of the posterior pdf about X_0. When considering the behaviour of any function in the neighbourhood of a particular point, it is often helpful to carry out a *Taylor series* expansion; this is simply a standard tool for (locally) approximating a complicated function by a low-order *polynomial.* Rather than dealing directly with the posterior pdf P, which is a 'peaky' and positive function, it is better to work with its *logarithm* L:

$$L = \log_e[\mathrm{prob}(X \mid \{data\}, I)], \tag{2.10}$$

since this varies much more slowly with X. Expanding L about the point $X = X_0$, we have

$$L = L(X_0) + \frac{1}{2}\left.\frac{\mathrm{d}^2L}{\mathrm{d}X^2}\right|_{X_0}(X - X_0)^2 + \ldots, \tag{2.11}$$

where the best estimate of X is given by the condition

$$\left.\frac{\mathrm{d}L}{\mathrm{d}X}\right|_{X_0} = 0, \tag{2.12}$$

which is equivalent to eqn (2.8) because L is a *monotonic* function of P.

The first term in the Taylor series, $L(X_0)$, is a constant and tells us nothing about the shape of the posterior pdf. The linear term, which would be proportional to $X - X_0$, is missing because we are expanding about the maximum (as indicated by eqn 2.12). The *quadratic* term is, therefore, the dominant factor determining the width of the posterior pdf and plays a central role in the reliability analysis. Ignoring all the higher-order contributions, the exponential of eqn (2.11) yields

$$\mathrm{prob}(X \mid \{data\}, I) \approx A \exp\left(\frac{1}{2}\left.\frac{\mathrm{d}^2L}{\mathrm{d}X^2}\right|_{X_0}(X - X_0)^2\right), \tag{2.13}$$

where A is a normalisation constant. Although this expression looks a little weird, what we have really done is to approximate the posterior pdf by the ubiquitous Gaussian distribution. Also known as the *normal* distribution, it is usually written as

$$\mathrm{prob}(x \mid \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \tag{2.14}$$

and is plotted in Fig. 2.3 (with the vertical axis multiplied by $\sigma\sqrt{2\pi}$); this function is symmetric about the maximum, at $x = \mu$, and has a width which is proportional to σ. Comparing the exponents of eqns (2.13) and (2.14), we are reassured to find that the posterior pdf for X has a maximum at $X = X_0$; its

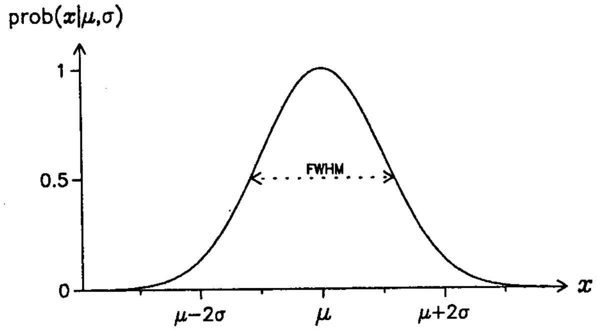

Fig. 2.3 The Gaussian, or normal, distribution. It is symmetric with respect to the maximum, at $x = \mu$, and has a full width at half maximum (FWHM) of about 2.35σ. Note that, as plotted, it is not normalised: the vertical axis has been multiplied by $\sigma\sqrt{2\pi}$, so that the maximum is equal to one.