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 remark. 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 (1983, 2003) 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 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

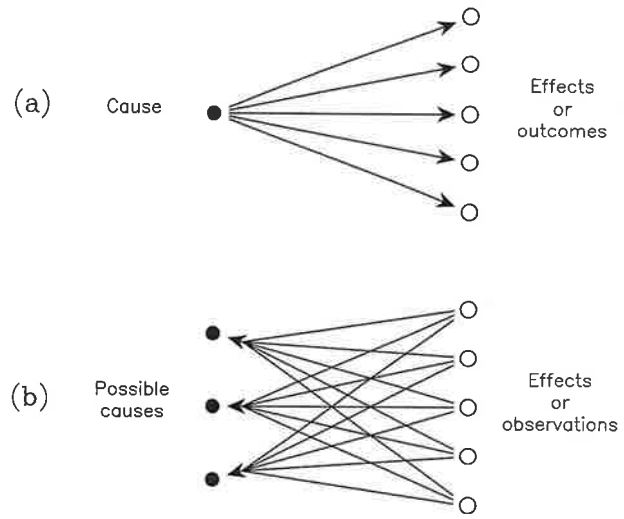

Fig. 1.1 A schematic representation of (a) deductive logic, or pure mathematics, and (b) plausible reasoning, or inductive logic.

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 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 by being hit in the eye with an arrow at the battle of Hastings in 1066 AD; (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 making two assertions. 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 assertion 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 obeyed the usual rules of probability theory:

$$\text{prob}(X|I) + \text{prob}\left(\overline{X}|I\right) = 1 \tag{1.1}$$

and

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

with $0 = \text{prob}(\text{false})$ and $1 = \text{prob}(\text{true})$ defining certainty; a proof is given in Appendix B. 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'.

We have made the probabilities conditional on I, to denote the relevant background information at hand, because there is no such thing as an absolute probability. For example, the probability 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.

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. 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). We can change the description of probability, as when we multiply by 100 to obtain percentages, but we are not allowed to change the content of eqns (1.1) and (1.2). Probability calculus uses the unique scale in which the rules take the form of un-adorned addition and multiplication.

1.3 Corollaries: Bayes' theorem and marginalization

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 *marginalization*:

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

and

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

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):

$$\text{prob}(Y,X|I) = \text{prob}(Y|X,I)\times\text{prob}(X|I)\,.$$

Since the statement that 'Y and X are both true' is the same as 'X and Y are both true', so that $\text{prob}(Y,X|I) = \text{prob}(X,Y|I)$, 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 $\text{prob}(X|Y,I)$ to $\text{prob}(Y|X,I)$. The importance of this property to data analysis becomes apparent if we replace X and Y by *hypothesis* and *data*:

$$\text{prob}(\mathit{hypothesis}|\mathit{data},I) \propto \text{prob}(\mathit{data}|\mathit{hypothesis},I)\times\text{prob}(\mathit{hypothesis}|I)\,.$$

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 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, $\text{prob}(\mathit{hypothesis}|I)$, 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 $\text{prob}(\mathit{data}|\mathit{hypothesis},I)$, and yields the *posterior* probability, $\text{prob}(\mathit{hypothesis}|\mathit{data},I)$, 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 $\text{prob}(\mathit{data}|I)$ has been omitted. This is fine for many data analysis problems, such as those involving *parameter estimation*, since the missing denominator is simply a normalization constant (not depending explicitly on the hypothesis). In some situations, like *model selection*, this term plays a crucial rôle. For that reason, it is sometimes given the special name of *evidence*. This crisp single word captures the significance of the entity, as opposed to older names, such as *prior predictive* and *marginal likelihood*, which describe how it tends to be used or calculated. Such a central quantity ought to have a simple name, and 'evidence' has been assigned no other technical meaning (apart from as a colloquial synonym of data).

The marginalization equation, (1.4), should seem a bit 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 marginalization equation for our standard X and Y propositions. It would take the form

$$\text{prob}(X|I) = \text{prob}(X,Y|I) + \text{prob}\big(X,\overline{Y}|I\big)\,. \tag{1.5}$$

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

$$\text{prob}(X,Y|I) = \text{prob}(Y,X|I) = \text{prob}(Y|X,I)\times\text{prob}(X|I)\,,$$

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

$$\text{prob}(X,Y|I) + \text{prob}\big(X,\overline{Y}|I\big) = \big[\text{prob}(Y|X,I) + \text{prob}\big(\overline{Y}|X,I\big)\big]\times\text{prob}(X|I)\,.$$

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.

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 then given by

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

This is just a generalization of eqn (1.5), and can be derived in an analogous manner, by putting $\text{prob}(X,Y_k|I) = \text{prob}(Y_k|X,I)\times\text{prob}(X|I)$ in the right-hand side of eqn (1.6), as long as

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

This *normalization* 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 marginalization equation in (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 generalization 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 $\mathrm{pdf}(X, Y|I)$, where

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

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

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

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 normalization condition of eqn (1.7) takes the form

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

Marginalization 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, Laplace and orthodox statistics

About three hundred years ago, people started to give serious thought to the question of how to reason in situations where 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, he also put it to good use in solving problems in celestial mechanics, medical statistics and even jurisprudence. Despite Laplace's numerous successes, his development of probability theory was rejected by many soon after his death.

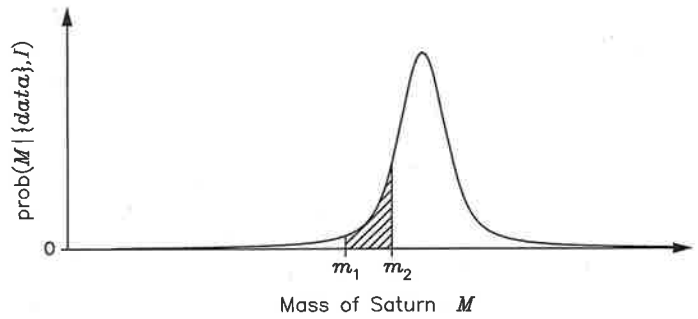

Fig. 1.2 A schematic illustration of the result of Laplace' analysis of the mass of Saturn.

The problem was not really one of substance but concept. To the pioneers such as Bernoulli, Bayes and Laplace, a probability represented a *degree-of-belief* or plausibility: how much they thought that something was true, based on the evidence at hand. To the 19th 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): $\mathrm{prob}(M|\{data\}, I)$; 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 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 realization 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, 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 students are often taught even 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 `rand` on a computer constitutes 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 `rand` (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, but we refer the interested reader to Caves *et al.* (2002) for a very insightful viewpoint. Either way, 'randomness' is what we call 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 case 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 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 aim of creating an objective theory is quite laudable. Unfortunately, the frequentist approach does not achieve this goal: neither does its 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 rôle 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. He found that the only rules which met his requirements were those of probability theory. Although the sum and product rules of probability are easy 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: 'It is remarkable that a science, which commenced with a consideration of games of chance, should be elevated to the rank of the most important subjects of human knowledge.'

1.5 Outline of 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 marginalization. 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 might do better by thinking deeply about a more 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 where 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 Chapter 7, 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'. 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.

Chapters 8 shows how one of the most widely used data analysis procedures in the physical sciences, *least-squares*, can be extended to deal with more troublesome measurements, such as those with 'outliers'.

Chapters 9 and 10 are concerned with modern numerical techniques for carrying out out Bayesian calculations, when analytical approximations are inadequate and a brute-force approach is impractical. In particular, they provide the first introductory account of the novel idea of *nested sampling*.

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 (2003) 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. All that one really needs to avoid them is basic professional care allied to common sense. Indeed, we hope that the reader will come to share Laplace's more general view that '*Probability theory is nothing but common sense reduced to calculation*'.