1

Decision

Statistics, the most important science in the whole world: for upon it
depends the practical application of every other science and of every art.
(Florence Nightingale)

If your experiment needs statistics, you ought to have done a better
experiment.

(Ernest Rutherford)

Science is about decision. Building instruments, collecting data, reducing data,
compiling catalogues, classifying, doing theory – all of these are tools, tech-
niques or aspects which are necessary. But we are not doing science unless
we are deciding something; only decision counts. Is this hypothesis or theory
correct? If not, why not? Are these data self-consistent or consistent with other
data? Adequate to answer the question posed? What further experiments do
they suggest?

We decide by comparing. We compare by describing properties of an object
or sample, because lists of numbers or images do not present us with immediate
results enabling us to decide anything. Is the faint smudge on an image a star
or a galaxy? We characterize its shape, crudely perhaps, by a property, say the
full-width half-maximum, the FWHM, which we compare with the FWHM
of the point-spread function. We have represented a data set, the image of the
object, by a statistic, and in so doing we reach a decision.

Statistics are there for decision and because we know a background against
which to take a decision. To this end, every measurement we make, and every
parameter or value we derive requires an error estimate, a measure of range
(expressed in terms of probability) that encompasses our belief of the true value
of the parameter. We are taught this by our masters in the course of interminable
undergrad lab experiments. Why? It is because no measured quantity or property
is of the slightest use in decision and therefore in science, unless it has a 'range quantity' attached to it.

A **statistic** is a quantity that summarizes data; it is the ultimate data-reduction. It is a property of the data and nothing else. It may be a number, a mean for example, but it does not have to be. It is a basis for using the data or experimental result to make a decision. We need to know how to treat data with a view to decision, to obtain the right **statistics** to use in drawing **statistical inference**. (It is the latter which is the branch of science; at times the term statistics is loosely used to describe both the descriptive values and the science.)

The opening quotes indicate a mixed press. Nightingale was a pioneer of applied statistics and graphical presentation. Her message is clear, but suggests the age-old confusion between statistics and data. Rutherford’s message also appears clear and uncompromising, but it can only hold in some specialized circumstances. For a start, astronomers are not always free to do better experiments. The laboratory is the big stage; the Universe is an experiment we cannot re-run. Attempting to understand astrophysics and cosmology from one freeze-frame in the spacetime continuum requires some reconsideration of the classical scientific method. This scientific method of **repetition** of experimentally reproduced results does not apply. Thus, the first issue for astronomers: we cannot always re-roll the dice, and anyway, repetition implies similar conditions. We are never at the same spacetime coordinates.

There is thus need for a certain rigour in our methodology. The inability to re-roll dice has led and still leads astronomers into some of the greatest errors of inference. It becomes tempting to the point of irresistibility to use the data on which a hypothesis was proposed to verify that hypothesis.

**Example** *The Black Cloud* (Hoyle, 1958). The Black Cloud appears to be heading for the Earth. The scientific team suggests that this proves the cloud has intelligence. Not so, says the dissenting team member. Why? A golf ball lands on a golf course which contains $10^7$ blades of grass; it stops on one blade; the chances are 1 in $10^7$ of this event occurring by chance. This is not so amazing – the ball had to land somewhere. It would only be amazing if the experiment were re-run to test the newly formulated hypothesis (e.g. the blade being of special attractive character; the golfer of unusual skill) and the event was repeated. However, the importance of deciding if the Black Cloud knew about the Earth cannot await the next event or the sequence of events, and except the rush to judgement in which initial data, hypothesis and test data are combined; so in many instances in astronomy and cosmology.
The most obvious area in which this offence is committed is in claims of physical association of objects of small angular separation on the sky; or similarly, claims of alignment of objects in close proximity on the sky. Most such claims are bogus because they use the object grouping in which the association or the alignment was originally noted in subsequent tests of significance. The original data may be used to formulate a hypothesis only; testing must await examination of fresh and unbiased data which do not include the original data. It is essential to divorce hypothesis–formulation data from hypothesis–test data. There is no set of tests which can cope with a-posteriori statistics, or will ever be able to do so.

A second difference for astronomers stems from the first – the remoteness of our objects and the inability to re-run our experiments precisely means that we do not necessarily know the underlying distributions of the variables measured. The essence of classical statistical analysis is (i) the formulation of hypothesis, (ii) the gathering of hypothesis–test data via experiment, and (iii) the construction of a test-statistic. But making a decision on the basis of the test-statistic may demand that the sampling distribution of the statistic be known before a decision can be made. How else could we decide if the value we got was normal or abnormal? It may well be the case that no one, physicist, sociologist, botanist, ever does know these underlying distributions exactly; but astronomers are worse off than most because of our necessarily small samples and our inability to control experiments, leading to poor definitions of the underlying distributions.

Astronomers cannot avoid statistics and there are at least the following reasons for this unfortunate situation.

(i) Error (range) assignment – ours, and the errors assigned by others: what do they mean?
(ii) How can data be used best? Or at all?
(iii) Correlation, testing the hypothesis, model fitting; how do we proceed?
(iv) Incomplete samples, samples from an experiment which cannot be re-run, upper limits; how can we use these to best advantage?
(v) Others describe their data and conclusions in statistical terms. We need some self-defence.
(vi) But above all, we must decide. The decision process cannot be done without some methodology, no matter how good the experiment. Rutherford may not have known when he was using statistics.

This is not a book about statistics, the values or the science. It is about how to get results in astronomy, using statistics, data analysis and statistical inference. Consider first how we do science in order to see at what point ‘statistics’ enter(s) the process.
1.1 How is science done?

In simplest terms, each experiment goes round a loop which can be characterized by six stages:

1. Observe: with an observing or data-gathering programme, record or collect the data.
2. Reduce: clean up the data to remove experimental effects, i.e. flat field it, calibrate it.
3. Analyse: obtain the numbers from the clean data – intensities, positions. Produce from these summary descriptors of the data which enable comparison or modelling – descriptors that lead to reaching the decision which governed the design of the experiment; and which are statistics.
4. Conclude: carry through a process to reach a decision. Test the hypothesis; correlate; model, etc.
5. Reflect: what has been learnt? Is the decision plausible? Is it unexpected? At which experimental stage must re-entry be made to check? What is required to confirm this unexpected result? Or – what was inadequate in the experimental design? How should the next version be defined? Is an extended or new hypothesis suggested? Far too little time is spent here; perhaps the pressure of observing application deadlines and/or the perceived need to publish get the better of us.
6. Experiment design: if the hypothesis is important enough; if the data warrant it; if previous experimental experience suggests it is possible; if technical advances make it feasible – then the next experiment needs to be designed. This may (and usually does) take the form of thinking out an observing proposal, writing and submitting it. It may take the form of re-design of an instrument on a current telescope. It may take the form of a proposal to build a new instrument. It may take the form of designing a new telescope or space mission, a process which, in itself, may occupy much of a research career. The latest such projects involve multi-nation collaborations on scales of billions of dollars. The timescales from initial plans to realization may range to 40 years (e.g. the James Webb Space telescope; the Square Kilometre Array).

And so back to stage 1.

This process is a loop and ‘experiments’ may begin at different points. For instance, we disbelieve someone else’s conclusions based on their published data set. We enter at point (3) or even (4); and we may then go around the data-gathering cycle ourselves as a result. Or we enter at (5), looking at an old result in the light of new and complementary ones from other fields – and proceed to (6) and back to (1) . . .
1.1 How is science done?

Table 1.1  Stages in astronomy experimentation

<table>
<thead>
<tr>
<th>Stage</th>
<th>How</th>
<th>Examples</th>
<th>Considerations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observe</td>
<td>In person? Remotely? Depends on facility</td>
<td>Experiment design: calibration integration time Stats</td>
<td>What is wanted? Number of objects</td>
</tr>
<tr>
<td>Reduce</td>
<td>Algorithms</td>
<td>Flat field Flux calibration</td>
<td>Stats</td>
</tr>
<tr>
<td>Analyse</td>
<td>Parameter estimation, hypothesis testing</td>
<td>Intensity measurements Positions T Stats</td>
<td>Frequentist, Bayesian?</td>
</tr>
<tr>
<td></td>
<td>T Stats</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Conclude</td>
<td>Hypothesis testing</td>
<td>Correlation tests Distribution tests T Stats</td>
<td>Believable, repeatable, understandable?</td>
</tr>
<tr>
<td></td>
<td>T Stats</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reflect</td>
<td>Carefully; far too little time is invested here</td>
<td>Mission achieved? A better way? ‘We need more data’? T Stats</td>
<td>The next observations</td>
</tr>
<tr>
<td>Design</td>
<td>Hone the mission; build science case</td>
<td>New observations/instrument/telescope/space mission</td>
<td>Feasibility – cost, team design, experience, human resources; simulations, predictions Stats</td>
</tr>
</tbody>
</table>

Stats

Of course it could be argued that (6) should start the process, but we need some knowledge base before we start designing.

All too often we use (3) to set up the tests at (4). This carries the charge of mingling hypothesis and data, as in the Black Cloud example.

Table 1.1 summarizes the process. Points in Table 1.1 at which recourse to statistics or to statistical inference is important have been indicated by Stats; a T appears when the issue applies to theorists as well as to experimentalists. Few are the regions in which we can ignore statistics and statistical inference. Experiment design needs to consider from the start what statistic or summarized data form is required to achieve the desired outcome. There are then checks throughout the experiment, and finally there is analysis in which the measured statistics are used in inference. Applied statistics in the guise of forecasting is increasingly used in astronomy instrument/survey/experiment design.
1.2 Probability; probability distributions

The concept of probability is crucial in decision processes, and there is a commonly accepted relationship between probability and statistics. In a world in which our statistics are derived from finite amounts of data, we need probabilities as a basis for inference. For example, limited data yields us only a partial idea of the point-spread function, such as the FWHM; we can only assign probabilities to the range of point-spread functions roughly matching this parameter.

We all have an inbuilt sense of probability. We know, for example, that the height of adults is anything from, say, 1.5 to 2.5 m. We know this from the totality of the population, all adults. But we know what a tall person is – and it is not necessarily somebody who is 2.5 m tall. The distribution is not flat; it peaks at around 1.7 m. The distribution of the heights of all adults, normalized to have an area of 1.0, is the measured probability density function, often called the probability distribution. (We meet them in a more rigorous context in Chapter 2.) The tails contain little area; and it is the tails that give us the decision: we probably call somebody tall when they are taller than 75 per cent of us.

We have made a decision based on a statistic, by relating that statistic to a probability distribution; we have decided that the person in question was tall. Note also what we did – observe, reduce, analyse, conclude, probably all in one glance. We did not do this rigorously in making a quantitative assessment of just how tall, which would have required a detailed knowledge of the distribution of height and a quantitative measurement. And reflect? Context of our observation? Why did we wish to register/decide that the person was tall? What next as a result? How was this person selected from the population? The brain has not only done the five steps but has also set the result into an extensive context; and this in processing the single glance.

The probability distribution in our minds – the heights of adults – is unlikely to have a mathematical description; it is one determined by counting enough of the population (probably subconsciously) so that it is well defined. There are distributions for which mathematical description is very precise, such as the Poisson and Gaussian (Normal) distributions, and there are many cases in which we have good reason to believe that these must represent the underlying probability distributions well.

This is also an example of a ‘ruling-out’; here we ruled out the hypothesis that the person is of ‘ordinary’ height. There is a different type of statistical inference, the ‘ruling-in’ process, in which we compute the probability of getting a given result, and if it is ‘probable’, we accept the original hypothesis.
1.3 Bolt-on statistics?

It is also an example of ‘counting’ to find the probabilities, the frequency distribution. There are other ways of assigning probabilities, including opinion and states of knowledge; and, in fact, there are instances in which we are moderately comfortable with the paradoxical notion of assigning probabilities to unique events. It is essential that our view of statistics and statistical inference be broad enough to take such probability concepts on board.

1.3 Bolt-on statistics?

With regard to statistics and probability, in many of the conversations we have had with users of the first edition of this book, we found that ‘statistics’ is often seen as a bolt-on addition to scientific analysis, a technological feature rather like dentistry; necessary, somewhat unpleasant, but a solved piece of technology. In the aftermath of the global financial crisis beginning in 2008, the role of quantitative finance was widely discussed. One of the failures that was identified was the failure of statistical models of risk – failures that had consequences costing trillions of dollars. Why the contradiction? Surely if statistics and probability were that routine, things could not have gone wrong quite so badly?

The answer is that there is a very wide range of degree of certainty associated with the application of statistics. An early distinction was drawn by Knight (1921), who was curious about why some businesses made huge profits and some only modest ones. His suggestion was that the run-of-the-mill firms dealt in risk, whereas the very successful ones (with an obvious selection effect in operation) dealt in uncertainty. What did Knight mean by these terms, especially ‘uncertainty’, which we often use interchangeably with words like random, stochastic or probability?

Take a concept, implicit in our usual undergraduate lab statistics training, in which we think we know the mean and standard deviation of our normally distributed observable. To put the implication at its starkest, this means that we know every single observation that we will ever make; only the order is unknown. This is melodramatic phrasing, but it expresses the extraordinary power of the assumption that we know a probability distribution. Often, when we start out in statistics, we have an uneasy feeling that we are getting something for nothing. In fact, there is a high price to pay in the scope of the assumptions we make, either openly or unknowingly. This illustrates what Knight meant by mere risk, in a business context: risk involves only known probabilities. A casino is an example. Unless the roulette wheels are improperly engineered, the management of a casino can predict its profits, as long as customers keep
coming through the doors with the same amount of money in their pockets. The probabilities are known, and the casino management knows exactly how to set its margins to attain a given return.

Of course, not even a casino operates in this ideal environment. Taleb (2010) gives the example of the single biggest loss experienced by a casino of which he had apparently intimate knowledge: in a show put on to entertain idle patrons, a performing tiger ate its trainer, with consequent eye-watering claims for trauma, loss of earnings to the bereaved family, and so on. This is what Knight meant by uncertainty and Taleb by his term ‘black swans’ – not only may the probabilities not be known, they may not even have been considered.

Returning to the more familiar ground of astronomy, what do we learn for the application of statistics to our subject? There is exactly the same continuum between risk and uncertainty, reflected in the robustness of the assumptions we make in order to pursue our statistical analyses. Do we know the parameters of the distributions we assume? Probably not, but we can estimate them from the data. How well we do this depends on how much data we have. If we have a lot, we wonder if it is ‘all the same’, or whether the underlying parameters are actually varying within the data set. Indeed, the very form of the distributions we assume is an issue. Gaussian? To the extent that the central limit theorem (Section 2.4.2.3) holds, perhaps. More realistically, we need a range of distributions, each with its own prior probability and parameters . . . and so on, up the hierarchy of complexity towards greater uncertainty.

As you embark on this little handbook, remember that the statistics you will encounter represent a model of the world, in the same messy, complicated, intuition-need sense as the astronomy to which you may think you can ‘bolt it on’. Making the measurement is the easy part, understanding the error is the hard part; but as you will see if you persist with us, there is a framework (formally, the framework of Bayesian inference, aided by the concept of hyperparameters) that allows us to bound our ignorance and control its consequences – if we are fortunate. If we are not, of course, we may be eaten by a tiger.

1.4 Probability and statistics in inference: an overview of this book

Statistics are combinations of the data that do not depend on any unknown parameters. The average is a common example. When we calculate the average of a set of data, we expect that it will bear some relation to the true, underlying mean of the distribution from which our data were drawn. In the classical tradition, the variance of the variable is the key measure of the error of the approximation and the statistic is used to take it into account.

This is the classical view of the world and is the view that underlies much of the statistical analysis of the world as it is known, or at least is thought to be. We will develop these ideas and see how they can be extended to make the best use of the data that we have.

In Chapter 6 we will explore the Bayesian approach. The Bayesians view the world in a different way: they attach different probabilities to various events, and then make distributions of them to answer the question we are asking. In the classical approach, we developed a model of the world, and then used it to answer questions about it.

Chapter 2 introduces the Bayesian approach to the problem of making inferences about the world, given the data that we have. We will see that this approach is fundamentally different from the classical approach, and that it leads to different conclusions about the world.

1 ... but not in the sense that we usually think of it.
1.4 Probability and statistics in inference: an overview

tradition, we calculate the sampling distribution of the average, the probabilities of the various values it may assume as we (hypothetically) repeat our experiment many times. We then know the *probability* that some range around our single measurement will contain the true mean. This is information that we can use to take decisions.

This is precisely the utility of statistics – they are laboriously discovered combinations of observations which converge, for large sample sizes, to some underlying parameter we want to know (say, the mean). Useful statistics are actually rather few in number.

We meet the issues of probability distributions, statistics, the relation between these, and the role of random-number analyses in Chapters 2 and 3. The long development of these concepts is outlined in Table 1.2, a sketch of the timeline of the development of probability and statistics. Origins of statistical inference can be traced back to Aristotle (384–322 BC) who developed a logic framework and stated a version of Occam’s Razor. For a fascinating historical study of statistics and probability, see the erudite books by Anders Hald (1990, 1998).

In Chapter 2 we also meet a radically different way of making inferences – the *Bayesian approach*, totally distinct in its logic from the ‘classical’ or ‘frequentist’ approach just discussed. The Bayesian approach focuses on the probabilities right away, without the intermediate step of statistics. In the Bayesian tradition, we invert the reasoning just described. The data, we say, are unique and known; it is the mean that is unknown, that should have probability attached to it. Without using statistics, we instead calculate the probability of various values of the mean, given the data we have. This also allows us to make decisions. In fact, as we shall see, this approach comes a great deal closer to answering the questions that scientists actually ask. This drastic change in approach came painfully and relatively recently – see Table 1.2. From Chapter 2 on, we invoke both methodologies to greater or lesser extent; we explain why in context.

Chapter 4 *Correlation and association* provides our first look at a practical area of statistics, namely correlations, searches for them in data sets as well as tests of their significance. This area of statistics might well be the one which most readily refutes the charge that statistics as a science has not discovered anything.\(^1\) The original regression lines of Francis Galton (‘regression to mediocrity’) played a major role in genetics, while subsequently the germ theory of disease (John Snow) and the expansion of the Universe (Edwin Hubble) both emerged from correlation analyses.

---

\(^1\) ... but serves only as the lamp-post serves the drunken man: for support rather than for illumination (Andrew Lang, nineteenth-century poet and philosopher).
Table 1.2  A brief history of probability and statistics

<table>
<thead>
<tr>
<th>Year</th>
<th>Individual(s)</th>
<th>Key words</th>
<th>Events</th>
</tr>
</thead>
<tbody>
<tr>
<td>1340</td>
<td>William of Ockham, or Occam</td>
<td>Occam’s Razor</td>
<td>‘It is useless to do with more what can be done with less.’ Ockham, an ordained Franciscan, was excommunicated for his views on separation of church and state, amongst other things. In addition to the application of the principle in statistics and data modelling, Hawking (1988) attributes the discovery of quantum mechanics to it.</td>
</tr>
<tr>
<td>1654</td>
<td>Pascal, Fermat</td>
<td>odds, probability theory</td>
<td>Gombaud, Chevalier de Mere &amp; Mitton pose questions on gambling odds to Pascal in ~1654. Seven letters exchanged between Blaise Pascal &amp; Pierre de Fermat are the genesis of probability theory.</td>
</tr>
<tr>
<td>1657</td>
<td>Huygens</td>
<td>probability</td>
<td>First publication on probability, 14 problems (+ solutions) in gambling, based on the Pascal–Fermat correspondence; the only publication on the subject for 50 years.</td>
</tr>
<tr>
<td>1662</td>
<td>Graunt</td>
<td>descriptive statistics, life tables, survival analysis</td>
<td>Publication of Graunt’s <em>Observations on the Bills of Mortality</em>; first known collection and analysis of data for statistical purposes; start of actuarial risk analysis.</td>
</tr>
<tr>
<td>1692</td>
<td>Huygens, Arbuthnot</td>
<td>probability</td>
<td>Of the Laws of Chance, or, a method of Calculation of the Hazards of Game . . . ; Arbuthnot’s translation of Huygens’ work becomes the first English publication on probability.</td>
</tr>
<tr>
<td>1665–1676</td>
<td>Newton, Leibniz</td>
<td>calculus</td>
<td>Newton &amp; Leibniz independently discover calculus; their dispute runs for decades. Probability theory can proceed.</td>
</tr>
<tr>
<td>1687</td>
<td>Newton</td>
<td>binomial distribution</td>
<td>In the monumental <em>Principia</em>, Newton changes the direction of physics and mathematics forever; the book includes the binomial probability distribution.</td>
</tr>
<tr>
<td>Year</td>
<td>Individual(s)</td>
<td>Key words</td>
<td>Events</td>
</tr>
<tr>
<td>------</td>
<td>---------------</td>
<td>-----------</td>
<td>--------</td>
</tr>
<tr>
<td>1693</td>
<td>Halley</td>
<td>risk analysis, actuarial analysis</td>
<td>Halley publishes <em>An Estimate of the Degrees of Mortality of Mankind.</em></td>
</tr>
<tr>
<td>1711</td>
<td>de Moivre</td>
<td>statistical independence</td>
<td>Concept of statistical independence introduced, expressed in ratios of products of numbers of wins and losses.</td>
</tr>
<tr>
<td>1713</td>
<td>J. Bernoulli</td>
<td>binomial distribution, law of large numbers</td>
<td>N. Bernoulli publishes posthumously his uncle’s <em>Ars Conjectandi</em> proving the binomial distribution of Newton, developing issues of probability, and introducing the (weak) law of large numbers.</td>
</tr>
<tr>
<td>1730</td>
<td>de Moivre</td>
<td>central limit theorem</td>
<td>The central limit theorem in the special case of the binomial distribution.</td>
</tr>
<tr>
<td>1733</td>
<td>de Moivre</td>
<td>Normal distribution</td>
<td>The Normal distribution shown to be an approximation of the binomial distribution.</td>
</tr>
<tr>
<td>1749, 1791</td>
<td>Achenwall, Sinclair</td>
<td>‘Statistik’ = statistics</td>
<td>The study of social data (Graunt, Halley and many others) suggests to Achenwall that dealing with natural ‘states’ of society should be referred to as <em>Statistik</em>. Sinclair’s 21-volume <em>Statistical Account of Scotland</em> (1791) establishes the term.</td>
</tr>
<tr>
<td>1749</td>
<td>Mayer</td>
<td>combining observations</td>
<td>Prior to 1750, with the exception of Tycho Brahe, most observers believed that combining observations led to divergence from the best estimate. Mayer shows that the reverse is the case.</td>
</tr>
<tr>
<td>1756</td>
<td>Bayes</td>
<td>systematic errors</td>
<td>In a letter from Bayes: ‘The more observations you make with an imperfect instrument the more it seems to be that the error in your conclusion will be proportional to the imperfection of the instrument...’</td>
</tr>
</tbody>
</table>

(continuation)


<table>
<thead>
<tr>
<th>Year</th>
<th>Individual(s)</th>
<th>Key words</th>
<th>Events</th>
</tr>
</thead>
<tbody>
<tr>
<td>1757, 1760</td>
<td>Boscovich, Maire</td>
<td>combining observations</td>
<td>In 1757 they publish a synopsis of ideas on combining observations; a full description of the method appears in 1760.</td>
</tr>
<tr>
<td>1763</td>
<td>Bayes</td>
<td>Bayes’ theorem</td>
<td>Bayes’ theorem is presented to the Royal Society posthumously by Richard Price: ‘I now send you an essay which I have found among the papers of our deceased friend Mr Bayes, and which, in my opinion, has great merit…’ The theorem is finally accepted by the great Laplace in 1781.</td>
</tr>
<tr>
<td>1787</td>
<td>Laplace</td>
<td>combining observations</td>
<td>While publishing proof of the stability of the solar system, Laplace improves on the method of Mayer to combine observations.</td>
</tr>
<tr>
<td>1805</td>
<td>Legendre</td>
<td>least squares</td>
<td>In his treatise on comet orbits, Legendre develops the method of least squares.</td>
</tr>
<tr>
<td>1809–1810</td>
<td>Gauss, Laplace</td>
<td>Normal distribution</td>
<td>77 years after de Moivre, Gauss shows that observational errors are expected to have a Normal (Gaussian) distribution. Laplace provides a much better derivation.</td>
</tr>
<tr>
<td>1812</td>
<td>Laplace</td>
<td>probability theory</td>
<td>Laplace publishes his landmark <em>Théorie Analytique des Probabilités</em>.</td>
</tr>
<tr>
<td>1835</td>
<td>Quetelet</td>
<td>Normal distribution</td>
<td>Quetelet, astronomer, statistician, social scientist, publishes a statistical study of human properties showing the ubiquitous nature of the Normal distribution; he is a major influence in promoting the use of statistical studies in social and astronomical contexts.</td>
</tr>
<tr>
<td>1837</td>
<td>Poisson</td>
<td>Poisson distribution, law of large numbers</td>
<td>Poisson publishes <em>Recherches sur la Probabilité des Jugements</em>... in which he introduces the distribution now bearing his name, and coins the phrase ‘Law of Large Numbers’.</td>
</tr>
<tr>
<td>Year</td>
<td>Individual(s)</td>
<td>Key words</td>
<td>Events</td>
</tr>
<tr>
<td>----------</td>
<td>---------------</td>
<td>--------------------</td>
<td>---------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>1885–9</td>
<td>Galton</td>
<td>regression,</td>
<td>The amazing Galton, cousin of Darwin, African explorer, statisticians, psychologist, biologist, criminologist (fingerprints), meteorologist, the original behavioural geneticist, and publisher of 350 books and papers: he introduces regression plots (the heights of sons against their fathers), correlation plots and explores the bivariate Gaussian to describe these.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>correlation,</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>bivariate Gaussian</td>
<td></td>
</tr>
<tr>
<td>1893</td>
<td>Pearson</td>
<td>standard deviation</td>
<td>Pearson introduces the term, already known as error of mean square or mean error, and it becomes the accepted way of describing distribution spread.</td>
</tr>
<tr>
<td>1897</td>
<td>Pearson</td>
<td>product moment</td>
<td>Pearson introduces the standard (frequentist) method of describing the strength of a correlation.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>coefficient</td>
<td></td>
</tr>
<tr>
<td>1900</td>
<td>Pearson</td>
<td>chi-square test</td>
<td>Pearson’s polemical paper developing chi-square demonstrates inter alia that a long run of bad luck on the roulette wheel at Monte Carlo could not have been due to chance.</td>
</tr>
<tr>
<td>1904</td>
<td>Spearman</td>
<td>rank</td>
<td>Spearman develops a non-parametric way of testing for correlation even when data are not on ordinal scales.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>correlation</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>coefficient</td>
<td></td>
</tr>
<tr>
<td>1908</td>
<td>Gosset</td>
<td>‘Student’s’ t</td>
<td>Gosset derives the t distribution to test for differences between means of observed distributions.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>distribution</td>
<td></td>
</tr>
<tr>
<td>1913</td>
<td>Eddington</td>
<td>Eddington bias</td>
<td>Distortion of number counts from measurement error, introduced in a two-page paper in <em>Monthly Notices of the Royal Astronomical Society</em>.</td>
</tr>
<tr>
<td>1920</td>
<td>Malmquist</td>
<td>Malquist bias</td>
<td>Malmquist discovers the luminosity-distance correlation from a survey of fixed sensitivity, the bias that will plague astronomers for the next century.</td>
</tr>
<tr>
<td>Year</td>
<td>Individual(s)</td>
<td>Key words</td>
<td>Events</td>
</tr>
<tr>
<td>------</td>
<td>---------------</td>
<td>-----------------------------------</td>
<td>----------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>1922</td>
<td>Fisher</td>
<td>maximum likelihood</td>
<td>Fisher develops the concept and consequences of maximum likelihood.</td>
</tr>
<tr>
<td>1925</td>
<td>Fisher</td>
<td>ANOVA and much more</td>
<td>Fisher publishes the massively influential <em>Statistical Methods for Research Workers</em>, which includes the new analysis of variance and describes for the first time the full panoply of t test, chi square, the F test, randomized design, and significance testing. What we now call Fisher Information was also introduced in this year.</td>
</tr>
<tr>
<td>1933</td>
<td>Neyman and Pearson</td>
<td>the null hypothesis</td>
<td>Neyman and Pearson introduce hypothesis testing, the null hypothesis, and Type I and Type II errors.</td>
</tr>
<tr>
<td>1933</td>
<td>Hoteling</td>
<td>PCA</td>
<td>Principal component analysis, the right way to search for dependencies in multivariate data, is developed and published.</td>
</tr>
<tr>
<td>1933</td>
<td>Kolmogorov</td>
<td>Kolmogorov axioms, Kolmogorov–Smirnov test</td>
<td>Kolmogorov presents his probability axioms, a basis for a self-consistent theory of probability; and the Kolmogorov–Smirnov non-parametric test to search for significant difference between distributions.</td>
</tr>
<tr>
<td>1935</td>
<td>Fisher</td>
<td>‘The Design of Experiments’</td>
<td>The standard for decades for the integrated design and statistical analysis of experiments.</td>
</tr>
<tr>
<td>1939</td>
<td>Jeffreys</td>
<td>theory of probability</td>
<td>The first attempt to develop a fundamental theory of scientific inference based on Bayesian statistics. His ideas were well ahead of their time.</td>
</tr>
<tr>
<td>1953</td>
<td>Neyman, Scott, Shane</td>
<td>clustering of galaxies</td>
<td>A key paper identifying ‘contagion’ in the clustering of galaxies; Neyman had been working on the statistics of epidemics and so noticed the mathematical similarities between clusters of disease outbreaks and the clustering of galaxies.</td>
</tr>
<tr>
<td>Year</td>
<td>Individual(s)</td>
<td>Key words</td>
<td>Events</td>
</tr>
<tr>
<td>------</td>
<td>--------------</td>
<td>-----------</td>
<td>--------</td>
</tr>
<tr>
<td>1953</td>
<td>Metropolis <em>et al.</em></td>
<td><strong>the Metropolis algorithm</strong></td>
<td><em>Equations of State Calculations by Fast Computing Machines</em> introduces an idea which will come of age several decades later and make Bayesian inference possible for real problems.</td>
</tr>
<tr>
<td>~1955</td>
<td>Cooley, Tukey</td>
<td>FFT</td>
<td>Cooley and Tookey derive the Fast Fourier Transform algorithm, all-pervasive in image compression, CMB cosmology, aperture synthesis, etc.</td>
</tr>
<tr>
<td>~1960</td>
<td>Fisher–Pearson vs. Jaynes–Jeffreys</td>
<td>Bayes’ revival</td>
<td>The powerful and productive rivals Pearson and Fisher, while disagreeing violently and publicly on more than one issue, were frequentist in approach and so influential in the first half of the 20th century that Bayesian methods were all but forgotten. The rise of computing power and more importantly of mathematicians such as Jaynes and Jeffreys, who thought deeply about methodology, began a sea change. 1965–80 was a period of intense debate amongst statisticians who divided themselves into frequentists and Bayesians. The outcome is the current ascendancy of Bayesian methods, with frequentist methods still in use in many traditional areas of sample testing such as polling and medical drug research where simple yes-no answers are requested; and in areas where no model exists such as non-parametric testing. Modern approaches combine the methodologies.</td>
</tr>
<tr>
<td>1957</td>
<td>Scheuer</td>
<td>$p(D)$</td>
<td>The 'probability of Deflection', confusion limit analysis, developed for interferometric radio surveys, has application in all domains in which more than one signal contributes within the resolution of the instrument.</td>
</tr>
</tbody>
</table>

*(cont.)*
Chapter 5, *Hypothesis testing* is basically frequentist; here we describe the classical tests based on statistics. Chapter 5 describes both parametric and non-parametric (or ‘distribution-free’) tests. The latter is a methodology of doing frequentist-type hypothesis testing *without knowing what the underlying probability distribution actually is*, i.e. without having such a distribution which is characterized by parameters. This is of particular importance for astronomers for a number of reasons, as we have mentioned.

So why this plethora of approaches: parametric versus non-parametric/ frequentist versus Bayesian? Part of this is the very different ways in which we encounter data, ours or other people’s, or perhaps statistics – data descriptors – presented to us in lieu of data. For instance, it is important to recognize that data may be in a number of forms depending on what *measurement scale* is used; see Table 1.3. When it does come to us in forms other than those immediately recognizable, i.e. on numerical scales (magnitudes, redshifts, etc.), there are still formally valid ways of carrying out statistical inference, as we shall see. However, our options may be limited and, in particular, Bayesian techniques may not be applicable.
### Table 1.3 Measurement scales

<table>
<thead>
<tr>
<th>Scale type</th>
<th>Also called</th>
<th>Description: example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nominal/categorical</td>
<td>Binned</td>
<td>gender: male / female</td>
</tr>
<tr>
<td>Ordinal/ranking</td>
<td>Ordered</td>
<td>army ranks: private, corporal, sergeant, . .</td>
</tr>
<tr>
<td>Ratio</td>
<td>Numerical/</td>
<td>uniformly calibrated scale with zero point:</td>
</tr>
<tr>
<td></td>
<td>measures</td>
<td>temperature in kelvin</td>
</tr>
<tr>
<td>Interval</td>
<td>Numerical/</td>
<td>uniformly calibrated scale: time, whose</td>
</tr>
<tr>
<td></td>
<td>measures</td>
<td>beginning (and end) we do not know, but</td>
</tr>
<tr>
<td></td>
<td></td>
<td>which we arbitrarily ‘zero-point’ in many</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ways to give it ratio scales, e.g. Gregorian</td>
</tr>
<tr>
<td></td>
<td></td>
<td>calendar, Julian Day, UTC, GMT</td>
</tr>
</tbody>
</table>

Chapter 6 describes data modelling and parameter estimation, and is a mixture of frequentist and Bayesian techniques. Chapter 7 describes more advanced Bayesian techniques, including model selection, Markov chain Monte Carlo analysis and hyperparameters.

Chapters 8, 9, 10 and 11 have a strong astronomical orientation. Chapter 8 is about detection, catalogues, surveying, luminosity functions, incomplete samples and confusion (object-blending). Chapter 9 deals with 1D statistics, i.e. spectral scans, or time-sequenced observations, the sampling and analysis of such data. Chapter 10 is termed Statistics of large-scale structure, as it deals, in the main, with the 2D distribution of objects on the sky. Chapter 11 is a descriptive epilogue, considering the integral role of statistics in current astrophysics and cosmology by way of examples in galaxy distribution, weak gravitational lensing and cosmic microwave background (CMB) studies.

### 1.5 How to use this book

This is not a textbook of statistical theory, a guide to numerical analysis, or a review of published work. It is a practical manual, which assumes that proofs, numerical methods and citation lists can easily be found elsewhere. This book sets out to tell it from an astronomer’s perspective, and our main objective is to help in gaining familiarity with the broad concepts of statistics and probability, to understand their usefulness and to feel confident in applying them. Work through the examples and exercises; they are drawn from our experience and have been chosen to clarify the text. They vary in difficulty, from one-page calculations to mini-projects. Some need data; these may be simulated. If
preferred, example data sets are available on the book’s website – as are the solutions to the exercises. Aim to become confident in the use of Monte Carlo simulations to check any calculations, and to try out ideas. Remember, in this subject we can do useful and revealing experiments – in the computer. Do not be ashamed to let simulations guide your mathematical intuition!

For further details on statistical methods and justification of theory, there is no substitute for a full textbook. None of our topics is arcane and they will be found in the index of many books on probability and statistics. We have found several particularly helpful and these are described in Appendix A. We need to mention one indispensable reference work here, though: *Numerical Recipes* (Press *et al.*, 2007), which points the way on numerous statistical issues, provides the means for numerically solving most problems, and contains comprehensive, humorous and wise advice.

There is little algebra in this book; it would have greatly lengthened and cluttered the presentation to have worked through details. Likewise, we have not explained how various integrals were done or eigenvalues found. These things can be done by computers; packages such as the superb MATHEMATICA, used for many of the calculations in this book, can deal swiftly with more mathematical technology than any of us know. Using these packages frees us all up to think about the problem to hand, rather than searching in vain for missing minus signs or delving into handbooks for integrals which never seem to be there in quite the needed form.

We have not attempted exhaustive referencing. Rather, we have given enough key references to provide entry-points to the literature. Online bibliographic databases provide excellent cross-referencing, showing who has cited a paper and whom it cites; it is the work of minutes to collect a comprehensive reading-list on any topic. The lecture notes for many excellent university courses are now on the Web; a well-phrased search may well yield useful material to help with whatever is puzzling you.

Finally, use this book as you need it. It can be read from front to back, or dipped into. Of course, no interesting topic is self-contained, but we hope that the cross-referencing will connect all the technology needed to explore a particular topic.

### Exercises

1.1 **Discovery.** At first sight, discovery of a new phenomenon may not read as an experiment as described in Section 1.1. But it is. Describe the discovery of pulsars (*Hewish et al.*, 1968) in terms of the six experimental stages.
1.2 Significance. The significance of a certain conclusion depends very strongly on whether the most luminous known quasar is included in the data set. The object is legitimately in the data set in terms of pre-stated selection criteria. Is the conclusion robust? Believable?
2

Probability

God does not play dice with the Universe.
(Albert Einstein

Whether He does or not, the concepts of probability are important in astronomy for two reasons.

1. Astronomical measurements are subject to random measurement error, perhaps more so than most physical sciences because of our inability to re-run experiments and our perpetual wish to observe at the extreme limit of instrumental capability. We have to express these errors as precisely and usefully as we can. Thus, when we say ‘an interval of $10^{-6}$ units, centred on the measured mass of the Moon, has a 95 per cent chance of containing the true value’, it is a much more quantitative statement than ‘the mass of the Moon is $1 \pm 10^{-6}$ units’. The second statement really only means anything because of some unspoken assumption about the distribution of errors. Knowing the error distribution allows us to assign a probability, or measure of confidence, to the answer.

2. The inability to do experiments on our subject matter leads us to draw conclusions by contrasting properties of controlled samples. These samples are often small and subject to uncertainty in the same way that a Gallup poll is subject to ‘sampling error’. In astronomy we draw conclusions such as: ‘the distributions of luminosity in X-ray-selected Type I and Type II objects differ at the 95 per cent level of significance.’ Very often the strength of this conclusion is dominated by the number of objects in the sample and is virtually unaffected by observational error.

---

\[1\] The first documented statement to this effect, rather more elegantly phrased, appears in a 1926 letter to his friend Max Born (AEA 8-180). Einstein made several public and private comments of a similar nature.
2.1 What is probability?

This chapter begins with a discussion of what probability is, and proceeds to introduce the concepts of conditionality and independence, providing a basis for the consequent discussion of Bayes' theorem, with prior and posterior probabilities. Only at this point is it safe to consider the concept of probability distributions; some common probability distributions are compared and contrasted. We then introduce Monte Carlo generators, thereby providing tools with which to examine probabilities, probability distributions and relations between sample size and uncertainty, inter alia. This all sets the stage for the following chapter, dealing with statistics themselves, the penultimate product of data reduction – if conclusions/discoveries are considered as the ultimate product. The issues of expectation and errors, dependent on the distributions and statistics, are discussed in the final section of the following chapter.

2.1 What is probability?


The study of probability began with the analysis of games of chance involving cards or dice. Because of this background we often think of probabilities as a limiting case of a frequency. Many textbook problems are still about dice, hands of cards, or coloured balls drawn from urns; in these cases it seems obvious to take the probabilities of certain events according to the ratio

\[
\text{number of favourable events/total number of events}
\]

and the probability of throwing a six with one roll of the dice is ‘obviously’ 1/6.

This probability derives from what Laplace called the ‘principle of indifference’, which in effect tells us to assign equal probabilities to events unless we have any information distinguishing them. In effect we have done the following calculation:

\[
\text{probability of one spot} = x \\
\text{probability of two spots} = x \\
\text{probability of three spots} = x
\]

and so on; this is the principle of indifference step. Further, we believe that we have identified all the cases; with the convention that the probability of a
certain event (anything between one and six spots) is unity, we have

\[ 6x = 1. \]

This calculation, apparently trivial as it is, shows a vitally important feature; we cannot usefully define probability by this kind of ratio. We have had to assume that each face of the die is equally probable to start with – thus, the definition of probability becomes circular.

If we can identify equally likely cases, then calculating probabilities amounts simply to enumerating cases – not always easy, but straightforward in principle. However, identifying equally likely cases requires much more thought.

Many interesting and useful calculations can be done using the principle of indifference, either directly or by exploiting its applicability to aspects of the problem. For example, we may know that a die is biased, the faces are not ‘equally likely’. However, given some details of, say, the mass distribution of the die, we may be able to calculate the probabilities of the faces using an assumption that the initial direction of the throw is isotropic – in which case the principle of indifference applies to throw-directions.

Sometimes we estimate probabilities from data. The probability of our precious observing run being clouded out is estimated by

\[ \text{number of cloudy nights last year} / 365 \]

but two issues arise. One is the limited data – we suspect that 10 years’ worth of data would give a different, more accurate result. The second issue is simply the identification of the ‘equally likely’ cases. Not all nights are equally likely to be cloudy, some student of these matters tells us; it is much more likely to be cloudy in winter. What is ‘winter’, then? A set of nights equally likely to be cloudy?

We can only estimate the probabilities correctly once we have identified the equally likely cases, and this identification is the subjective, intuitive step that is built into our reasoning about data from apparently malevolent instrumentation in an uncertain world.

It is common to define probabilities as empirical statements about frequencies, in the limit of large numbers of cases – our 10 years’ worth of data; but, as we have seen, this definition must be circular because selecting the data depends on knowing which cases are equally likely. Defining probabilities in this way is sometimes called ‘frequentist’.

So, what is probability? The notion we adopt for the present is that probability is a numerical formalization of our degree or intensity of belief. In everyday speech we often refer to the probability of unique events, showers of rain or election results. In the desiccated example of throwing dice, \( x \) measures the
2.1 What is probability?

The strength of our belief that any face will turn up. Provided that the die is not loaded, this belief is 1/6, the same for each face.

Ascribing an apparently subjective meaning to probability in this way needs careful justification. After all, one person’s degree of belief is another person’s certainty, depending on what is known. We can only reason as best we can with the information we have; if our probabilities turn out to be wrong, the deficiency is in what we know, not the definition of probability. We just need to be sure that two people with the same information will arrive at the same probabilities. It turns out that this constraint, properly expressed, is enough to develop a theory of probability which is mathematically identical to the one often interpreted in frequentist terms.

A useful set of properties of probability can be deduced by formalizing the ‘measure of belief’ idea. The argument is originally due to Cox (1946) and goes as follows: if A, B and C are three events and we wish to have some measure of how strongly we think each is likely to happen, then, for consistent reasoning, we should at least apply the rule if A is more likely than B, and B is more likely than C, then A is more likely than C. Remarkably, this is sufficient to put constraints on the probability function which are identical to the Kolmogorov axioms of probability, proposed some years before Cox’s paper:

- any random event A has a probability \( \text{prob}(A) \) between 0 and 1;
- the sure event has \( \text{prob}(A) = 1 \);
- If A and B are exclusive events, then \( \text{prob}(A \text{ or } B) = \text{prob}(A) + \text{prob}(B) \).

The Kolmogorov axioms are a sufficient foundation for the entire development of mathematical probability theory, by which we mean the apparatus for manipulating probabilities once we have assigned them.

Example Before 1987, four naked-eye supernovae had been recorded in 10 centuries. What, before 1987, was the probability of a bright supernova happening in the twentieth century?

There are three possible answers.

1. Probability is meaningless in this context. Supernovae are physically determined events and, when they are going to happen, can, in principle, be accurately calculated. They are not random events.

   From this God’s-eye viewpoint, probability is indeed meaningless; events are either certain or forbidden. ‘God does not play dice...’.

2. From a frequentist point of view our best estimate of the probability is 4/10, although it is obviously not very well determined.
This assumes that supernovae were equally likely to be reported throughout 10 centuries, which may well not be true. Eventually some degree of belief about detection efficiency will have to be made explicit in this kind of assignment.

(3) We could try an a-priori assignment. In principle we might know the stellar mass function, the fate and lifetime as a function of mass, and the stellar birth rate. We would also need a detection efficiency. From this we could calculate the mean number of supernovae expected in 1987, and we would put some error bars around this number to reflect the fact that there will be variation caused by factors we do not know about – metallicity, perhaps, or location behind a dust cloud, and so on.

The belief-measure structure is more complicated in this detailed model but it is still there. The model deals in populations, not individual stars, and assumes that certain groups of stars can be identified which are equally likely to explode at a certain time.

Suppose now that we sight supernova 1987A. Is the probability of there being a supernova later in the twentieth century affected by this event?

Approach (1) would say no – one supernova does not affect another. Approach (2), in which the probability simply reflects what we know, would revise the probability upward to 5/10. Approach (3) might need to adjust some aspects of its models in the light of fresh data; predicted probabilities would change.

Probabilities reflect what we know – they are not things with an existence all of their own. Even if we could define ‘random events’ (approach 1), we should not regard the probabilities as being properties of supernovae.

### 2.2 Conditionality and independence

Two events $A$ and $B$ are said to be independent if the probability of one is unaffected by what we may know about the other. In this case, it follows (not trivially!) from the Kolmogorov axioms that

$$\text{prob}(A \text{ and } B) = \text{prob}(A)\text{prob}(B).$$

(2.1)

Sometimes independence does not hold, so that we would also like to know the conditional probability: the probability of $A$, given that we know $B$. The definition is

$$\text{prob}(A \mid B) = \frac{\text{prob}(A \text{ and } B)}{\text{prob}(B)}.$$  

(2.2)
2.2 Conditionality and independence

If $A$ and $B$ are independent, knowing that $B$ has happened should not affect our beliefs about the probability of $A$. Hence, $\text{prob}(A \mid B) = \text{prob}(A)$ and the definition reduces to $\text{prob}(A \text{ and } B) = \text{prob}(A)\text{prob}(B)$ again.

If there are several possibilities for event $B$ (label them $B_1, B_2, \ldots$) then we have that

$$\text{prob}(A) = \sum_i \text{prob}(A \mid B_i)\text{prob}(B_i). \quad (2.3)$$

$A$ might be a cosmological parameter of interest, while the $B$s are not of interest. They might be instrumental parameters, for example. Knowing the probabilities $\text{prob}(B_i)$ we can get rid of these ‘nuisance parameters’ by a summation (or integration). This is called marginalization.

Example Take the familiar case in astronomy where some ‘remarkable’ event is observed, for example two quasars of very different redshifts close together on the sky. The temptation is to calculate an a-priori probability, based on surface densities, of two specified objects being so close. However, the probability of the two quasars being close together is conditional on having noticed this fact in the first place. Thus, the probability of the full event is simply $\text{prob}(A \mid A) = 1$, consistent with how we should expect to measure our belief in something that we already know. We can say nothing further, although we might be able to formulate a hypothesis to carry out an experiment.

Consider now the very different case in which we wish to know the probability of finding two objects of different types, say galaxy and quasar, within a specified angular distance $r$ of each other. To be specific, we plan to search some fixed solid angle $\Omega$. The surface densities in question are $\Sigma_G$ and $\Sigma_Q$. On finding a galaxy, we will search around it for a quasar. We need

$$\text{prob}(G \text{ in field and } Q \text{ within } r) = \text{prob}(Q \text{ within } r \mid G \text{ in field})\text{prob}(G \text{ in field}).$$

This assumes that the probabilities are independent, obviously what we would like to test. A suitable model for the probabilities is the Poisson distribution, and in the interesting case where the probabilities are small we have

$$\text{prob}(G \text{ in field}) = \Sigma_G \Omega$$

and

$$\text{prob}(Q \text{ within } r) = \pi r^2 \Sigma_Q.$$
The answer we require is therefore
\[ \text{prob}(G \text{ in field and Q within } r) = \zeta G \zeta Q \Omega \pi r^2. \]

This is symmetrical in the quasar and galaxy surface densities as we would expect; it should not matter whether we searched first for a galaxy or for a quasar. Note the strong dependence on the search area that is specified before the experiment; if there is obscurity about this, then the probabilities are not well determined.

As an extension of this example, it is possible to calculate the probability of finding triples of objects aligned to some small tolerance (Edmunds & George, 1985). If the objects are all the same, the probability of a linear triple depends on the cube of the surface density and search area.

### 2.3 ... and Bayes’ theorem

Bayes’ theorem is a simple equality, derived by equating \( \text{prob}(A \text{ and } B) \) with \( \text{prob}(B \text{ and } A) \). This gives the ‘theorem’:

\[ \text{prob}(B \mid A) = \frac{\text{prob}(A \mid B) \text{prob}(B)}{\text{prob}(A)}. \quad (2.4) \]

In this, the denominator is a normalizing factor. The theorem is particularly useful when interpreted as a rule for induction; the data, the event \( A \), are regarded as succeeding \( B \), the state of belief preceding the experiment. Thus \( \text{prob}(B) \) is the prior probability which will be modified by experience. This experience is expressed by the likelihood \( \text{prob}(A \mid B) \). Finally \( \text{prob}(B \mid A) \) is the posterior probability, the state of belief after the data have been analysed.

Bayes’ theorem by itself is a perfectly innocent identity, a mathematical truism. It acquires its force from its interpretation. To see what this force is, we return to the familiar and simple problem of drawing those coloured balls from urns. It is clear, even automatic, what to calculate; if there are \( M \) red balls and \( N \) white balls, the probability of drawing three red balls and two white ones is ...
As a series of brilliant scientists realized, and as a series of brilliant scientists did not, this is generally not the problem we face. As scientists, we more often have a datum (three red balls, two white ones) and we are trying to infer something about the contents of the urn. This is sometimes called the problem of ‘inverse probability’. (An exact analogy is the inverse problem of a sample of objects ‘drawn’ or obtained in a sky survey to a fixed sensitivity, and attempting to infer from the sample how these types of object are distributed throughout the Universe.) How does Bayes’ theorem help? We interpret it to be saying

$$\text{prob(contents of urn | data)} \propto \text{prob(data | the contents of the urn)}$$

and of course we can calculate the right-hand side, given some assumptions.

The urn example illustrates the principles involved; these are far more interesting than coloured balls.

**Example** There are $N$ red balls and $M$ white balls in an urn; we know the total $N + M = 10$, say. We draw $T = 3$ times (putting the balls back after drawing them) and get $R = 2$ red balls. How many red balls are there in the urn?

Our model (hypothesis) is that the probability of a red ball is

$$\frac{N}{N + M}.$$ 

We assume that the balls are not stratified, arranged in pairs, or anything else ‘peculiar’. The probability of getting $R$ red balls, the likelihood, is

$$\binom{T}{R} \left( \frac{N}{N + M} \right)^R \left( \frac{M}{N + M} \right)^{T-R}.$$ 

This is the number of permutations of the $R$ red balls amongst the $T$ draws, multiplied by the probability that $R$ balls will be red and $T - R$ will not be red. (This is a binomial distribution; see Section 2.4.2.1.)

Thus we have the probability (data, given the model) part of the right-hand side of Bayes’ theorem. We also need probability (model), or the prior. We assume that the only uncertain bit of the model is $N$, which to start with we take as being uniformly likely between zero and $N + M$. Without bothering with the details at the moment, we plot up the left-hand side of Bayes’ theorem (the posterior probability) as a function of $N$ – see Figure 2.1. For a draw of, say, three red balls in five tries, the posterior probability peaks at
6; for 30 out of 50, the peak is still at 6 but other possibilities are much less likely.

This seems unsurprising and in accord with common sense – but notice that we are speaking now of the probability of there being 1, 2, 3, \ldots red balls in a unique urn that is the subject of our experiment. We are describing our state of belief about the contents of the urn, given what we know (the data, and our prior information).

The key point of this example is that we have succeeded in answering our scientific question: we have made an inference about the contents of the urn, and can make probabilistic statements about this inference. For example, the probability of the urn containing three or fewer red balls is 11 per cent. We are assigning probabilities to these statements to \( N \) because we are using probability to reflect our degree of certainty. Our concern, as experimental scientists, is with what we can infer about the world from what we know.

*Bayes' theorem allows us to make inferences from data, rather than compute the data we would get if we happened to know all the relevant information about our problem.*

This may seem academic; but suppose we had data from two populations and wanted to know if the means were different. Many chapters of statistics textbooks answer the opposite question for us: given populations with two different means, what data would you get? The combination of interpreting probability as a consistent measure of belief, plus Bayes’ theorem, allows us to...
answer the question we wish to pose: given the data, what are the probabilities of the parameters contained in our statistical model?

Another very significant point about this example is the use of prior information; again, we assigned probabilities to \( N \) to reflect what we know. Notice that although the word ‘prior’ suggests ‘before the experiment’, it really means ‘what we know apart from the data’. Sometimes this can have a dramatic, even disconcerting effect on our inferences; see Section 2.5. Sometimes we even need a ‘probability of a probability’:

**Example** Return to the question of supernova rate per century and consider how to estimate this; call this \( \rho \). Our data are four supernova in 10 centuries. Our prior on \( \rho \), expressing our total ignorance, is uniform between 0 and 1; we have no preconceptions or information about \( \rho \). A suitable model for \( \text{prob}(\text{data} \mid \rho) \) is the binomial distribution (Section 2.4.2.1), because in any century we either get a supernova or we do not (neglecting here the possibility of two supernovae in a century). Our posterior probability is then

\[
\text{prob}(\rho \mid \text{data}) \propto \binom{10}{4} \rho^4 (1 - \rho)^6 \times \text{prior on } \rho.
\]

We follow Bayes and Laplace in taking the prior to be uniform in the range 0 to 1. Then, to normalize the posterior probability properly, we need

\[
\int_0^1 \text{prob}(\rho \mid \text{data})d\rho = 1,
\]

resulting in the normalizing constant

\[
\int_0^1 \binom{10}{4} \rho^4 (1 - \rho)^6 d\rho,
\]

which happens to be

\[
\frac{\Gamma(10)\Gamma(4)}{\Gamma(14)} = B[5, 7],
\]

where \( \Gamma \) is the Gamma function and \( B \) is the (tabulated) beta function. In general, for \( n \) supernovae in \( m \) centuries, the distribution is

\[
\text{prob}(\rho \mid \text{data}) = \frac{\rho^n (1 - \rho)^{m-n}}{B[n+1, m-n+1]}.
\]

Our distribution \((n = 4, m = 10)\) peaks – unsurprisingly – at 4/10, as shown in Figure 2.2.
As the sample size increases, the distribution becomes narrower so that the peak posterior probability is more and more closely defined by the ratio of successes (supernovae, in our example) to sample size. This result is sometimes called the law of large numbers, expressing, as it does, the frequentist idea of a large number of repetitions resulting in a converging estimate of probability.

The key step in this example is ascribing a probability distribution to $\rho$, itself a probability. This makes no sense in a frequentist approach, nor indeed in any interpretation of probabilities as objective. Even if we are prepared to leap this metaphysical hurdle, in very many cases the assignment of a prior probability is much more difficult than in this example. Indeed, the assignment of priors in the current example is very simple. For a long time the objection to Bayesian methods focused on the Bayes/Laplace uniform prior.

Both Jeffreys (1961) and Jaynes (1968) discuss the prior on $\rho$, arguing that in many cases a uniform prior is far too agnostic. By intricate arguments, they arrive at other possibilities:

$$\text{prob}(\rho) = \frac{1}{\rho(1-\rho)}$$

and the ‘Haldane prior’

$$\text{prob}(\rho) = \frac{1}{\sqrt{\rho(1-\rho)}}.$$

These are intended to reflect the fact that in most experiments, we are expecting, with good reason, a yes or no answer.
Assigning priors when our knowledge is rather vague can be quite difficult, and there has been a long debate about this. Some ‘obvious’ priors (like the one we might use for location, simply uniform from $-\infty$ to $\infty$) are not normalizable and can sometimes get us into trouble. Out of the enormous literature on this subject, try Lee (2004) for an introduction, and Jaynes’ writings for some fascinating arguments. One of the ways of determining a prior is the maximum entropy principle; we will see an example of such a prior later (Section 7.9). A common prior for a scale factor $\sigma$ is Jeffreys’ prior, uniform in $\log \sigma$.

**Example** The use of Bayes’ theorem as a method of induction can be neatly illustrated by our supernova example. For simplicity, imagine that we establish our posterior distribution at the end of the nineteenth century, so that it is $\rho^6(1 - \rho)^6/B[5, 7]$, as shown earlier. At this stage, our data are four supernovae in 10 centuries. Reviewing the situation at the end of the twentieth century, we take this as our prior. The available new data consist of one supernova, so that the likelihood is simply the probability of observing exactly one event of probability $\rho$, namely $\rho$. The updated posterior distribution is

$$\text{prob}(\rho \mid \text{data}) = \frac{\rho^5(1 - \rho)^6}{B[6, 7]}$$

which peaks at $\rho = 5/11$, as we might expect.

In these examples, we have focused on the peak of the posterior probability distribution. This is one way amongst many of attempting to characterize the distribution by a single number. Another choice is the posterior mean, defined by

$$< \rho > = \int_0^1 \rho \text{prob}(\rho \mid \text{data})d\rho.$$  \hspace{1cm} (2.5)

If we have had $N$ successes and $M$ failures, the posterior mean is given by a famous result called Laplace’s rule of succession:

$$< \rho > = \frac{(N + 1)}{(N + M + 2)}.$$  

In our example, at the end of the nineteenth century Laplace’s rule would give $5/12$ as an estimate of the probability of a supernova during the twentieth century. This differs from the $4/10$ derived from the peak of the posterior probability, and it will do so in general.

Unless posterior distributions are very narrow, attempting to characterize them by a single number is frequently misleading. How best to characterize
Table 2.1  The common probability density functions

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Density function</th>
<th>Mean</th>
<th>Variance</th>
<th>Raison d’être</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uniform</td>
<td>( f(x; a, b) = 1/(b - a) a &lt; x &lt; b ) ( = 0, x &lt; a, x &gt; b )</td>
<td>((a + b)/2)</td>
<td>((b - a)/12)</td>
<td>In the study of rounding errors; as a tool in studies of other continuous distributions.</td>
</tr>
<tr>
<td>Binomial</td>
<td>( f(x; p, q) = x! \frac{n!}{x!} p^x q^{n-x} )</td>
<td>( np )</td>
<td>( npq )</td>
<td>( x ) is the number of ‘successes’ in an experiment with two possible outcomes, one (‘success’) of probability ( p ), and the other (‘failure’) of probability ( q = 1 - p ). Becomes a Normal distribution as ( \infty ). The limit for the binomial distribution as ( p \ll 1 ), setting ( \mu = np ). It is the ‘count-rate’ distribution, e.g. take a star from which an average of ( \mu ) photons are received per ( \Delta t ) (out of a total of ( n ) emitted; hence ( p \ll 1 )); the probability of receiving ( x ) photons in ( \Delta t ) is ( f(x; \mu) ). Tends to the Normal distribution as ( \mu \rightarrow \infty ).</td>
</tr>
<tr>
<td>Poisson</td>
<td>( f(x; \mu) = e^{-x} \frac{\mu^x}{x!} )</td>
<td>( \mu )</td>
<td>( \mu )</td>
<td>The essential distribution; see text. The central limit theorem ensures that the majority of ‘scattered things’ are dispersed according to ( f(x; \mu, \sigma) ). Vital in the comparison of samples, model testing; characterizes the dispersion of observed samples from the expected dispersion, because if ( x_i ) is a sample of ( v ) variables normally and independently distributed with means ( \mu_i ) and variances ( \sigma_i^2 ), then ( \chi^2 = \sum_{i=1}^{N} (x_i - \mu_i)^2 / \sigma_i^2 ) obeys ( f(\chi^2; v) ). Invariably tabulated and used in integral form. Tends to Normal distribution as ( v \rightarrow \infty ). For comparison of means, Normally distributed populations; if ( n_i x_i )'s are taken from a Normal population ( (\mu, \sigma) ), and if ( x_i ) and ( \sigma_i ) are determined, then ( t = \sqrt{n} (x_i - \mu_i) / \sigma_i ) is distributed as ( f(t; v) ) where ‘degrees of freedom’ ( v = n - 1 ). Statistic ( t ) can also be formulated to compare means for samples from Normal populations with the same ( \sigma ), different ( \mu ). Tends to Normal as ( v \rightarrow \infty ).</td>
</tr>
</tbody>
</table>

\[ f(t; v) = \frac{\Gamma[(v + 1)/2]}{\sqrt{\pi v} \Gamma(v/2)} \left( 1 + \frac{t^2}{v} \right)^{-v/2} \] 

Legend: The better-known variate functions
(\( f(v) \) is the correct \( \Gamma \).
\( f(v) = \frac{\Gamma(v/2)}{\sqrt{\pi v} \Gamma(v/2)} \left( 1 + \frac{t^2}{v} \right)^{-v/2} \) for \( v > 2 \). 
(i) \( f(v) \) is the probability distribution of \( \chi^2 \). 
(ii) \( f(v) \) is the pdf of the Student's t-distribution. 
(iii) \( f(v) \) is the cdf of the chi-squared distribution. 
(iv) \( f(v) \) is the cdf of the Student's t-distribution. 

We deal with samples; in this sort of probability distribution, probabilities for numbers usually seen to be integers. The sum \( n \) of probabilities is unity and a probability is intuitively defined.
the distribution depends on what is to be done with the answer, which in turn depends on having a carefully posed question in the first place.

2.4 Probability distributions

2.4.1 Concept

We have referred several times to probability distributions. The basic idea is intuitive; here is a little more detail.

Consider the boring experiment in which we toss four ‘fair’ coins. The probability of no heads is \((1/2)^4\); of one head \(4 \times (1/2)^4\); of two heads \(6 \times (1/2)^4\), etc. The sum of the possibilities for getting no heads to four heads is readily seen to be 1.0. If \(x\) is the number of heads \((0, 1, 2, 3, 4)\), we have a set of probabilities \(\text{prob}(x) = (1/16, 1/4, 3/8, 1/4, 1/16)\); we have a probability distribution, describing the expectation of occurrence of event \(x\). This probability distribution is discrete; there is a discrete set of outcomes and so a discrete set of probabilities for those outcomes.

In this sort of case we have a mapping between the outcomes of the experiment and a set of integers. Sometimes the set of outcomes maps onto real numbers instead, the set of outcomes no longer containing discrete elements. We deal with this by the contrivance of ‘discretizing’ the range of real numbers into little ranges within which we assume that the probability does not change. Thus, if \(x\) is the real number that indexes outcomes, we associate with it a probability density \(f(x)\); the probability that we will get a number ‘near’ \(x\), say within a tiny range \(\delta x\), is \(\text{prob}(x) \delta x\). We loosely refer to probability ‘distributions’, whether or not we are dealing with discrete outcomes.

Formally: if \(x\) is a continuous random variable, then \(f(x)\) is its probability density function, commonly termed probability distribution, when

(i) \(\text{prob}[a < x < b] = \int_a^b f(x)\,dx\);
(ii) \(\int_{-\infty}^{\infty} f(x)\,dx = 1\), and
(iii) \(f(x)\) is a single-valued non-negative number for all real \(x\).

The corresponding cumulative probability distribution function is \(F(x) = \int_{-\infty}^{x} f(y)\,dy\). Probability distributions and distribution functions may be similarly defined for sets of discrete values of \(x\); and distributions may be multivariate, functions of more than one variable.

2.4.2 Some common distributions

The better-known probability density functions appear in Table 2.1 together with location (where is the ‘centre’?) and dispersion (what is the ‘spread’?)
quantifiers. These quantifiers can be given by the first two moments of the distributions (Section 3.1):

$$
\mu_1(\text{mean}) = \mu = \int_{-\infty}^{\infty} x f(x) \, dx \quad (2.6)
$$

$$
\mu_2(\text{variance}) = \sigma^2 = \int_{-\infty}^{\infty} (x - \mu_1)^2 f(x) \, dx \quad (2.7)
$$

The square root of the variance, $\sigma$, is known as the standard deviation. Table 2.1 gives some indication of how or where each distribution arises. Three of them are of great importance, the binomial, Poisson, and Gaussian or Normal.

### 2.4.2.1 Binomial distribution

There are two outcomes – ‘success’ or ‘failure’. This common distribution gives the chance of $n$ successes in $N$ trials, where the probability of a success at each trial is the same, namely $\rho$, and successive trials are independent. This probability is then

$$
\text{prob}(n) = \binom{N}{n} \rho^n (1 - \rho)^{N-n}. \quad (2.8)
$$

The leading term, the combinatorial coefficient, gives the number of distinct ways of choosing $n$ items out of $N$:

$$
\binom{N}{n} = \frac{N!}{n!(N - n)!}. \quad (2.9)
$$

This coefficient can be derived in the following way. There are $N!$ equivalent ways of arranging the $N$ trials. However, there are $n!$ permutations of the successes, and $(N - n)!$ permutations of the failures, which correspond to the same result – namely, exactly $n$ successes, arrangement unspecified. Since we require not just $n$ successes (probability $p^n$) but exactly $n$ successes, we need exactly $N - n$ failures, probability $(1 - p)^{(N-n)}$ as well. The binomial distribution follows from this argument. The binomial distribution has a mean value given by

$$
\sum_{n=0}^{N} n \text{prob}(n) = Np
$$

and a variance or mean square value of

$$
\sum_{n=0}^{N} (n - Np)^2 \text{prob}(n) = Np(1 - p).
$$
Example Suppose we know, from a sample of 100 galaxy clusters selected by automatic pattern-recognition techniques, that 10 contain a dominant central galaxy. We plan to check a different sample of 30 clusters, now selected by X-ray emission. How many of these clusters do we expect to have a dominant central galaxy? If we assume that the 10 per cent probability holds for the X-ray sample, then the chance of getting \( n \) dominant central galaxies is

\[
\text{prob}(n) = \binom{30}{n} 0.1^n 0.9^{30-n}.
\]

For example, the chance of getting 10 is about 1 per cent; if we found this many, we would be suspicious that the X-ray cluster population differed from the general population.

Suppose we made these observations and did find 10 centrally dominated clusters. What can we do with this information?

![Figure 2.3](image.png)

Figure 2.3 The posterior probability distribution for the observation that 10/30 X-ray-selected clusters are centrally dominated. The dark black line uses a uniform prior distribution for this fraction; the dashed line uses the prior derived from an assumed previous sample in which 10 out of 100 clusters had dominant central members. The light curve shows the distribution for this earlier sample.

The Bayesian thing to do is a calculation that parallels the supernova example. Assuming that the X-ray galaxies are a homogeneous set, we can deduce the probability distribution for the fraction of these galaxies that have a dominant central galaxy. A relevant prior would be the results for the original larger survey. Figure 2.3 shows the results, making clear that the
data are not really sufficient to alter our prior very much. For example, there is only a 10 per cent chance that the centrally dominant fraction exceeds even 0.2; and, indeed, Figure 2.3 shows that the possibility of it being as high as 33 per cent is completely negligible. Our X-ray clusters differ markedly from the general population.

The binomial distribution is the parent of two other famous distributions, the Poisson and the Gaussian.

### 2.4.2.2 Poisson distribution

The Poisson distribution derives from the binomial in the limiting case of very rare (independent) events and a large number of trials, so that although \( p \to 0, Np \to \) a finite value. Calling the finite mean value \( \mu_1 = \mu \), the Poisson distribution is

\[
\text{prob}(n) = \frac{\mu^n}{n!} e^{-\mu}.
\]  

(2.10)

The variance of the Poisson distribution, \( \mu_2 \), is also \( \mu \).

The Poisson distribution is encountered in many walks of life, and despite our lifelong exposure to it, the everyday results from it are endlessly misinterpreted. There are two popular areas in which it produces daily newspaper headlines: crime rates and apparent spatial coincidences, both due to what is commonly termed Poisson clumping. This is a completely misleading term as the distribution itself cannot produce clumps. It is the brain, highly geared to detecting patterns, which finds the clumps, as in the following two examples.

**Example** Consider a homicide rate of 7 per 100 000 inhabitants per year (from Wikipedia, as for New York 2006; note that 48 cities of population 250 000 or greater in the USA had a higher rate). The low probability implies that the Poisson distribution applies. A population of 8 165 000 yields 572 homicides in the year, an average of 11 per week. What is the probability according to Poisson of 44 homicides in a week? The answer is about 1.5 per cent, i.e. roughly every two years there will be as many as 44 homicides in a week. You can be sure that the newspapers will headline ‘Crime Wave’ on such occasions – followed by political posturing and subsequent polarization of ‘we need more police and tougher laws’ versus ‘we need to improve education/opportunities/social conditions’. There will be irresistible demands for action. There is no doubt merit in taking action;
but the fact remains that random numbers drawn from Poisson distributions do not provide a good basis for decision.

The second example considers 2D Poisson distributions, as in the scattering of seeds at random in a field. Divide the field into little square cells: the probability of a seed falling into any one is small so that Poisson applies.

Example Suppose we place dots in two dimensions as in Figure 2.4:

![Diagram](image1.png)

(a) ![Diagram](image2.png)

(b)

Figure 2.4 Poisson in two dimensions. In which figure are the dots randomly placed?

which diagram of the two in Figure 2.4 has the dots randomly distributed? Both have about 100 dots in a square of unity in size. Figure 2.4(b) looks the better bet, but in fact is not random at all. The dots were first placed on an $X - Y$ grid and then shifted at random slightly away from the grid on a scale of about 0.1 of the grid interval. Figure 2.4(a) is truly random, and paradoxically this is the one in which we can ‘clearly’ see clumps (‘Poisson clumping’), voids, and even a semi-circle.

Here is trouble. The real-life encounter is the situation in which an interested party is trying to claim that, e.g., the occurrence of a disease is associated with the positions of power stations. The effect is the examination of two distributions (disease cases and power stations) overlaid, both of which will certainly exhibit some randomness with a dose of Poisson clumping, and some non-randomness. It is difficult to establish true probability or true evidence in the face of these effects, especially when we add in the effect of a-posteriori reasoning (Chapter 1). Moreover, a correlation does not imply a causal connection; see Chapter 4.
We mentioned sky distributions in Chapter 1, and, in particular, examining whether one type of extragalactic object is associated with another type. The difficulty of rigour is again severe, for the same reasons – two distributions, both with Poisson clumping, both probably non-random to start with, and a-posteriori statistics once again.

Finally, the Poisson distribution may play its biggest role in the lives of astronomers via the photons with which we measure emission from our chosen objects.

**Example** Poisson statistics govern the number of photons arriving during an integration. The probability of a photon arriving in a fixed interval of time is (often) small, at least at wavelengths shorter than the infrared (IR). The arrivals of successive photons are independent (apart from small correlations arising because photons obey Bose–Einstein statistics, negligible for our purposes). Thus, the conditions necessary for the Poisson distribution are met. Hence, if the integration over time $t$ of photons arriving at a rate $\lambda$ has a mean of $\mu = \lambda t$ photons, then the fluctuation on this number will be $\sigma = \sqrt{\mu}$. (In practice we usually only know the number of photons in a single exposure, rather than the mean number; obviously we can then only estimate the $\mu$. This is the subject of an exercise in the next chapter.) There are the following limiting cases:

1. Suppose we are detecting our objects with no effective background either from the sky or from our instrumentation. The photons we receive are solely from the objects measured. This idealized situation means that with $\mu = \lambda t$, the scatter on $\mu$ is (Poisson) $\sigma = \sqrt{\lambda t}$. If we `integrate' more by simply waiting for more photons, the *photon-limited* case,

   $\sigma \propto \sqrt{t}$, while signal $\propto t$.

   Thus, *signal/noise* $\propto \sqrt{t}$.

2. Now suppose our object is barely visible against the sky background, as in charge-coupled device (CCD) imaging of very faint objects in the optical regime. Our signal is still $\mu = \lambda t$, but our noise is $\sigma = \sqrt{\lambda_{\text{sky}} t}$. The net result is

   $$S/N \propto \frac{\lambda t}{\sqrt{\lambda_{\text{sky}} t}}.$$ 

   We again get

   $$S/N \propto \sqrt{t},$$
but note how much harder we have to work! The sky emission $\lambda_{\text{sky}} t$ will be much higher than the $\lambda t$ (the sky-limited case) so that to achieve similar $S/N$ to the simple photon-limited case, we shall have to integrate much longer. Moreover, the fact that integration goes as $\sqrt{t}$ implies that there is an effective limit to how good an observation can be. For example, if an $S/N$ of 2 is achieved in 2 hours, to get this $S/N$ to 4 will require four times as long, 8 hours, a whole night of precious telescope time.

3. Now suppose we have extremely bright photon-limited objects, for which we require very short exposures only, because there is vast signal, e.g. CCD observations of bright stars. Then, with huge signal and short exposures, the readout noise of the device, fixed, time-independent and placed on top of the signal, may dominate the uncertainty. Calling this fixed error $\sigma_{\text{readout}}$,

$$S/N \propto \frac{\lambda t}{\sigma_{\text{readout}}}, \text{ or } \propto t$$

for CCD of readout noise $\sigma_{\text{readout}}$.

4. Lastly, at the long-wavelength end of the spectrum, sub-millimetre and radio wavelengths, the flood of (relatively feeble) photons from objects is so enormous that the Poisson situation of rare events no longer applies. We are dealing with the receiver-limited case; $S/N$ is governed by receiver sensitivity. The physical situation is quite different, but curiously, the result is a familiar one. We now have such a quantity of photons from the object – call this flux of photons $S$ – that it is the receiver noise (assuming this to be thermal or roughly equivalent to thermal) which requires the integration:

$$S/N \propto \frac{S}{\sigma_{\text{rec}}/\sqrt{t}}, \text{ or } \propto \sqrt{t}$$

for a receiver of thermal noise $\sigma_{\text{rec}}$.

### 2.4.2.3 Gaussian (Normal) distribution

Both the binomial and the Poisson distributions tend to the Gaussian distribution (Figure 2.5), large $N$ in the case of the binomial, large $\mu$ in the case of the Poisson. The (univariate) Gaussian (Normal) distribution is

$$\text{prob}(x) = \frac{1}{\sigma \sqrt{2\pi}} \exp \left[-\frac{(x - \mu)^2}{2\sigma^2}\right] \quad (2.11)$$
from which it is easy to show that the mean is $\mu$ and the variance is $\sigma^2$ (Section 3.1). For the binomial when the sample size is very large, the discrete distribution tends to a continuous probability density

$$\text{prob}(n) = \frac{1}{\sigma \sqrt{2\pi}} \exp \left[ -\frac{(n - \mu)^2}{2\sigma^2} \right]$$

in which the mean $\mu = Np$ and variance $\sigma^2 = Np(1 - p)$ are still given by the parent formulae for the binomial distribution. Here is an instance of the discrete changing to the continuous distribution: in this approximation we can treat $n$ as a continuous variable (because $n$ changes by one unit at a time, being an integer, and so the fractional change $1/n$ is small). The Poisson distribution performs similarly – as $\mu$ is increased, the Poisson distribution becomes more symmetrical and when $\mu \geq 20$, the distribution is virtually indistinguishable from a Gaussian with mean $\mu$ and $\sigma = \sqrt{\mu}$.

**Example** Opinion-sampling, and ‘...expected to be correct within 2.1 per cent 19 times out of 20’. The phrase has become commonplace when presenting results from opinion polls. Where does it come from? *Sampling theory* is a subject in its own right – there are countless books. In opinion-polling, a representative sample of the population is asked the relevant question. (This statement is doubly loaded: how to choose the relevant sample and how to choose the question?) Suppose that total population
opinion favours option A over option B by a ratio of 3:1, i.e. 75 per cent of the population favour option A, and \( p = 0.75 \). Then suppose that 4000 representative people were asked for their opinion; \( N = 4000 \). How accurate would our result be? The result is given by the binomial distribution, but as \( pN = 3000.0 \), far above the value of \( \sim 20 \) required for the binomial distribution to approximate the Gaussian distribution, we can use the Gaussian distribution. For this we get \( \mu = 3000.0 \) and \( \sigma = 27.4 \). We know that 95 per cent of the area of this distribution lies between \( \pm 1.96\sigma = \pm 53.7 \), and as a percentage of the 4000, this is 2.1 per cent. 0.95 is 19/20, i.e. if we repeated the survey, 19 times out of 20 we should get a result within 2.1 per cent of the true proportion.

Pollsters rarely sample more than 5000 people; thus, they generally quote accuracies of this order. The reason for this limited sampling (for any total size of population) is (a) sampling is costly, and (b) the result is unlikely to improve in practice. The ‘2.1 per cent’ is only the sampling error, and at this level, other uncontrollable errors come to be of similar or even greater magnitude. These factors include (a) sampling bias, the difficulty of choosing a representative sample when you cannot ask all members of the population (the whole art of sampling design), and (b) voter fickleness, if, e.g., opinion polls are to do with elections. This comes in (at least) two forms – (a) sample members may not give their true opinions (particularly if those opinions are at extreme ends of the political spectrum), and (b) late factors, e.g. a telling debate exchange or a world event, may change voter opinions at the last moment.

The true importance of the Gaussian distribution and its dominant position in experimental science stems from the central limit theorem. A non-rigorous statement of this is as follows.

Form averages \( M_n \) from repeatedly drawing \( n \) samples from a population \( x_i \) with finite mean \( \mu \), variance \( \sigma^2 \). Then the distribution of

\[
\left[ \frac{(M_n - \mu)}{\sigma/\sqrt{n}} \right] \rightarrow \text{Gaussian distribution}
\]

with mean 0, variance 1, as \( n \rightarrow \infty \).

This is a remarkable theorem. What it says is that provided certain conditions are met – and they are in many situations – a little bit of averaging
Figure 2.6  An indication of the power of the central limit theorem. The panels show successive amounts of ‘integration’: in (a), a single value has been drawn; in (b), 200 values have been taken from an average of two values; (c), 200 values from an average of four; (d), 200 values from an average of 16.

will produce a Gaussian distribution of results no matter what the shape of the distribution from which the sample is drawn. Even eyeball integration counts. It means that errors on averaged samples will always look ‘Gaussian’. The reliance on Gaussian distributions, made valid by the unsung hero of statistical theory and, indeed, experimentation, the central limit theorem, shapes our entire view of experimentation. It is this theorem which leads us to describe our errors in the universal language of sigmas, and, indeed, to argue our results in terms of sigmas as well, which we explicitly or implicitly recognize as describing our place within or at the extremities of the Gaussian distribution.

Figure 2.6 demonstrates the compelling power of the central limit theorem. Here we have brutally truncated an exponential, clearly an extremely non-Gaussian distribution. The histogram obtained in drawing 200 random samples from the distribution (see Figure 2.6) follows it closely. When 200 values resulting from averaging just four values have been drawn, the distribution is already becoming Gaussian-like. When 200 values have been taken, it is virtually Gaussian.

Before leaving the discussion of Gaussian distributions it is important to emphasize that although the Gaussian is a useful and convenient model, it is not the final word. It is a model for many phenomena, but it is not the only model. Thus, astronomers have a 10σ, casting into doubt whether their observations support their hypothesis. This is a very different point of view from that of physicists – the latter will accept a value L (integrated over all mass) for a given object if the distribution is of the form

\[ N(> L) = K^\gamma \left( \frac{L}{\bar{M}} \right)^{\gamma - 1} \]

This is a scale-free invariant index of mass. If this index is constant, the distribution of number of events is also scale-free.

This distribution is well known from a variety of phenomena, from the distribution of stars in a galaxy to the sizes of rocks in a river bed. It is a power-law distribution, and it is one of the most important distributions in nature. It is also one of the most common distributions in nature. It is a scale-free distribution, and it is one of the most important distributions in nature. It is a power-law distribution, and it is one of the most important distributions in nature. It is a scale-free distribution, and it is one of the most important distributions in nature.
already becoming symmetrical; by the time 200 values of 16 averages have been taken, it is virtually Gaussian.

Before leaving the central limit miracle and Gaussian distributions, it is important to emphasize how tight the tails of the Gaussian distribution are (Table B.2). The range ±2σ encompasses 95.45 per cent of the area. Thus, the infamous 2σ result has a less than 5 per cent chance of occurring by chance. But we scoff—because the error estimates are difficult to make, and observers are optimistic. Things upset the distribution; there are outlying points. Thus, astronomers feel it necessary to quote results in the range 3σ to even 10σ, casting inevitable doubt on the belief in their own error estimates. In fact, experimentalists are aware of another key feature of the central limit theorem; the convergence to a Gaussian happens fastest at the centre of the distribution, but the wings may converge much more slowly to a Gaussian form. Interesting results (the 10σ ones) of course acquire their probabilistic interpretation from knowing the shape of the tails to high accuracy. In practice it is virtually impossible to be certain that an error distribution is truly Gaussian because the amount of data required to check the tails is, well, astronomical.

### 2.4.2.4 Power-law distribution

A very different type of distribution figures prominently in the life of astronomers—the power-law distribution. Take \( N \) as the number of objects or events that have a measured property (say luminosity) either greater than a value \( L \) (integral form) or within the bin \( dL \) centred on \( L \). With \( \gamma \) as the power-law exponent,

\[
N(>L) = K L^{\gamma+1}, \text{ integral form, } N > L, \text{ or }
\]

\[
dN = (\gamma + 1) K L^{\gamma} dL, \text{ differential form, } dN \text{ objects in } dL.
\]

(2.12) (2.13)

This is a scale-free or scale-independent distribution, because if \( f(x) = x^\gamma \), then \( f(ax) = a^\gamma \cdot x^\gamma = \text{const} \cdot x^\gamma = \text{const} \cdot f(x) \), the definition of scale independence.

This distribution is so different that intuition gained from our experience with Gaussians and progenitors binomial and Poisson becomes dangerous. As it stands, the mean and variance are both infinite, but in real life, this is not the case; something always physically limits each end. Nevertheless, steep power laws can appear in astronomy extending over decades, and when confronted with such distributions, astronomers have made serious mistakes. Trying to
describe results for objects selected from power-law distributions in terms of means and sigmas becomes hugely misleading at best, meaningless at worst. The slopes of the power laws are generally steep and inverse, that is to say there are many more objects of small \( L \) than large. This leads to very strong biases when objects are drawn from such a parent population, particularly if relatively large measurement errors are involved.

Power laws arise in many walks of life (see, e.g. Ball, 2004): they describe the distribution of fluctuations in the economic market, growth rates of firms, the distribution of salaries, and the size distributions of, e.g., avalanches, earthquakes and forest fires. The important thing is criticality. For instance, like the onset of an avalanche to adding sand at the apex of a sand-pile to a point where it suddenly becomes unstable. At criticality, there is no prescription as to whether a small region will slide and stop; or whether the entire side of the pile will be collectively triggered to break away and collapse. (These experiments have been done.) Earthquakes, stock-market fluctuations and forest-fires can be seen to follow similar patterns. So do many phenomena in the physical sciences, critical exponents for fluids, interconnectivity sub-networks on the Internet (the system is scale free), etc. The main feature of the power law is that the exponent is negative; there are many more small things than large – many little landslips go nowhere while very rare are the catastrophic collapses of the pile.

We meet these probability distributions in astronomy daily – the Salpeter Mass Function, magnitude or source counts (surface densities of objects on the sky), luminosity functions, the primordial fluctuation spectrum and more. There are always more faint objects than bright, more low-mass or low-luminosity objects than high; so that the exponent \( \gamma \) is invariably negative. The power law in pure form does not, of course, obey the formal definition of a probability distribution (Section 2.4.1). However, there are those physical limits which generally set upper and lower bounds.

**Example** Consider the counts of objects from one 15-arcmin-square region of the UK Schmidt Telescope survey, R-band, centre selected at random (but at high Galactic latitude), and with magnitudes measured off the plates by the CCD system of the SuperCosmos Sky Survey. Figure 2.7 shows the image and the results.

Power laws (fitted via maximum likelihood; see Section 6.1) describe the data well. Most objects in the total count are stars. The objects classified as galaxies are seen to have a significantly steeper slope, as expected for a distant population distributed approximately uniformly about us.
Figure 2.7 (a) A 15-arcmin square of sky from the R-band UKSTU sky survey at RA 22h, Dec $-18^\circ$. The scanning process recognizes about 750 images in the area. (b) The number–magnitude count (a ‘source count’ at other wavelengths) for all objects (dots) and for objects classified as galaxies (triangles). The data are plotted in numbers of objects in the area in equal bins of 0.4 mag, a ‘differential count’. Power laws have been fitted to the data. As magnitude is an inverse logarithmic scale, $m_1 - m_2 = -2.5 \log(L_1/L_2)$ where $L$ is luminosity, the power-law index is positive; of course, it would be negative if the plot were in terms of apparent luminosities.

Be aware of the many pitfalls of the power law. This distribution has no saving grace via approximations to familiar well-bounded distributions. We have mentioned selection bias. Add these issues: (a) is this power law an integral or differential distribution? This is one common way of getting the index wrong by unity. (b) Is the binning on a uniform or a log scale? If a differential distribution is binned via a uniform $\Delta \log L$ scale, instead of via $\Delta L$, the slope is reduced by unity (Figure 2.7). (c) There is no characteristic scale or spread for such a distribution, although in practice the physical limits always manage to provide high and low end-stops. Do not rely on these to make power laws tractable in terms of our normal usage of means and standard deviations. See the final example in Section 2.6.

For example, given a fixed range of a power law between, say, $a$ and $b$, the mean (from the first moment) can be calculated:

$$\mu = \left(\frac{\nu+1}{\nu+2}\right) \left[ \frac{b^{\nu+2} - a^{\nu+2}}{b^{\nu+1} - a^{\nu+1}} \right].$$

(2.14)

The expression breaks down for $\nu = -1$ or $-2$; results can readily be derived for these special cases. This mean in the interval $a$ to $b$ is useful in showing how skewed the distribution over the interval is, and also in providing suitable abscissa points for plotting a power law. The variance about this mean can also
be calculated (from the second moment) in a straightforward way, resulting in an even messier expression. This variance is highly misleading if used to describe the asymmetric ‘spread’ of the distribution.

2.5 Bayesian inferences with probability

It is in the following chapter that we describe statistics and their relation to probability distributions. This is a further step on the road to making statistical inferences with probability via the frequentist/classical route. We are not quite ready to do this yet; but as we have seen, with Bayesian methods we bypass these steps and make inferences from calculated probabilities, paying little regard to the name of the distribution we have calculated (if, indeed, it has a name). So what, then, is the relation between Bayesian inference and the probability distributions we have been describing?

One such connection is in estimating the parameters of assumed probability distributions, i.e. we are assuming a model for our data and wish to find out how this model is characterized. In this we are essentially data modelling; see Chapters 6 and 7. We have a probability distribution \( f(\text{data} \mid \alpha) \) in mind and we wish to know the parameter vector \( \alpha \). The Bayesian route is clear: compute the posterior distribution of \( \alpha \), as we have shown in several examples in this chapter.

**Example** Suppose we have \( N \) data \( X_i \), drawn from a Gaussian of known variance \( \sigma \) but unknown \( \mu \). The parameter we want is \( \mu \). To proceed, we need a prior on \( \mu \); we take the so-called ‘diffuse’ prior, where

\[
\text{prob}(\mu) = \text{constant}
\]

over some wide range of \( \mu \), the range defined by our knowledge of the problem. Of course we might have more precise information available. From Bayes, the posterior distribution follows at once:

\[
f(\mu \mid \text{data}) \propto \exp \left[ -\frac{\sum_{i=1}^{N}(X_i - \mu)^2}{2\sigma^2} \right]
\]

and with some simplification (including the absorbing of terms not depending on \( \mu \) into the ‘\( \propto \)’)

\[
f(\mu \mid \text{data}) \propto \exp \left[ -\frac{1}{N} \sum_{i=1}^{N}(X_i - \mu)^2 \right]
\]

so that the average of the data is distributed around \( \mu \), with variance \( \sigma^2/N \).
This method is related to the classical technique of maximum likelihood. If the prior is ‘diffuse’, as in the example, then the posterior probability is proportional to the likelihood term \( f(\text{data} \mid a) \). Maximum likelihood picks out the mode of the posterior, the value of \( a \) which maximizes the likelihood. This amounts to characterizing the posterior by one number, an approach which is often useful because of powerful theorems on maximum likelihood. We consider this in more detail in Section 6.1; some exercises at the end of this chapter illustrate the procedure.

Following the discussion of probability distributions, we are now able to consider some more detailed Bayesian problems, with surprising outcomes.

**Example** Suppose that we make an observation with a telescope at a randomly selected position in the sky. Our model of the data (an event labelled \( D \), consisting of the single measured flux density \( f \)) is that it is distributed in a Gaussian way (Section 2.4.2.3) about the true flux density \( S \) with a variance (Section 2.4.2) \( \sigma^2 \). The extensive body of source counts also tells us the a-priori distribution of \( S \); for the purposes of this example, we approximate this information by the simple prior for a static Euclidean universe

\[
\text{prob}(S) = K S^{-5/2}
\]

describing our prior state of knowledge. \( K \) normalizes the counts to unity; there is presumed to be one source in the beam at some flux-density level. The probability of observing \( f \) when the true value is \( S \) we take to be

\[
\exp \left[ -\frac{1}{2\sigma^2} (f - S)^2 \right].
\]

Bayes’ theorem then tells us that

\[
\text{prob}(S \mid D) = K' \exp \left[ -\frac{1}{2\sigma^2} (f - S)^2 \right] S^{-5/2},
\]

with the normalizations condensed into the single parameter \( K' \). If we were able to obtain \( n \) independent flux measurements \( f_i \), then the result would be

\[
\text{prob}(S \mid D) = K'' \exp \left[ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (f_i - S)^2 \right] S^{-5/2}.
\]

Suppose, for specific example, that the source counts were known to extend from 1 to 100 units, the noise level was \( \sigma = 1 \), and the data were 2, 1.3, 3, 1.5, 2 and 1.8. In Figure 2.8 are the posterior probabilities for the first two, then four, then six measurements. The increase in data gradually overwhelms
the prior but the prior affects conclusions markedly (as it should) when there are few measurements.

Figure 2.8 Measurement of flux density given a power-law prior (source count) and a Gaussian error distribution. The posterior probability distribution for flux density is plotted for two, four and then six of the measurements listed in the text; the form of the curve approaches Gaussian as numbers increase.

If, subsequently, we looked at a survey plate of the region we had observed, and found that the emission was from some category of object (say, a quasar) with different source counts, our prior would change and so would the posterior probability. In turn, our idea of the most probable flux density would also change.

In this example, the prior seems to be well determined. However, in some cases we wish to estimate quantities where the argument is not so straightforward. What would we take as the prior in the previous example if we were making the first ever measurements at a new wavelength?

2.6 Monte Carlo generators

We need to come clean about random numbers at this point - we have already used random-number (Monte Carlo) generation (Figure 2.6). In fact there are frequent occasions in probability calculations, hypothesis testing and model-fitting (see Chapters 5 and 7) when it is essential to have recourse to a set of numbers distributed perhaps how we guess the data might be. We may wish to test a test to see if it works as advertised; we might need to test efficiency of tests; we might wish to determine how many iterations we require; or we might even want to test that our code is working. We need random numbers.

either uniform or known frequency - much of what is open to direct commonly known.

It is important to random-number. One key issue is can the random number be repeated? (Or, in other words, is it possible to repeat it without the need to re-create it for each new set of conditions?)

Moreover, if the number of random numbers is large, the random number may be tempted to repeat itself. This is a problem if one wishes to use the random number in the calculation of a variable which is known to be a function of the random number, such as dr/dx is used to define a cumulative distribution function, $F(x) = \frac{1}{2} \int_{-\infty}^{x} f(x') dx'$ from which one can calculate the probability that $X$ is less than some value $x$. Here $x$ is the inverse, $x = \text{inverse, } x =$.

The very simple relationship $f(x)dx$ If this is $r$ and perform the curve is generated.
either uniformly distributed, or drawn randomly from a parent population of
known frequency distribution. If we introduce them here, then nothing is sacred
– much of what we say from here on, and of what we have said to this point, is
open to direct simulation by the reader through random-number generation –
commonly known as ‘Monte Carlo’.

It is important not to compromise any such analyses with bad random data. 

_Numerical Recipes_ (Press et al., 2007) presents a number of methods for 
random-number generation, from single expressions to powerful routines. A
key issue is _cycle length_; how long is it before the pseudo-random cycle is
repeated? (Or, how many random numbers do you need?) At some level,
therefore, it is necessary to understand the characteristics of the generator.
Moreover, it is essential to follow the prescribed implementation precisely. It
may be tempting to try some ‘extra randomizing’, for example by combining
routines or by modifying seeds. Be very scared of any such process. Finally it
is easy to forget that the routines generate _pseudo-random numbers_. Run them
again from the same starting point and you will get the same set of numbers.

How do we draw a set of random numbers following a _given_ frequency
distribution? Suppose, following the cautions above, that we have a
way of producing random deviates that are uniformly distributed over the
range 0.0–1.0, in, say, the variable \( \alpha \); and we have a functional form for our
frequency distribution \( \frac{dn}{dx} = f(x) \). We need a transformation \( x = x(\alpha) \) to
distort the uniformity of \( \alpha \) to follow \( f(x) \). But we know that

\[
\frac{dn}{dx} = \frac{dn}{d\alpha} \frac{d\alpha}{dx} = \frac{d\alpha}{dx} 
\]  \hspace{1cm} (2.15)

as \( \frac{dn}{d\alpha} \) is uniform by assumption; thus

\[
\alpha(x) = \int_{x'}^{x} f(x) dx, \hspace{1cm} (2.16)
\]

from whence the required transformation \( x = x(\alpha) \), the _inverse_ of
Equation (2.16), i.e. solve (2.16) for \( x \).

**Example** Thus, e.g., the example in Section 6.1: the source-count random
distribution is \( f(x) dx = -1.5 x^{-2.5} dx \), a ‘Euclidean’ differential source
count. Here \( d\alpha = -1.5 x^{-2.5} dx \), \( \alpha = x^{-1.5} \), and the transformation is
the inverse, \( x = \alpha^{-1/1.5} \).

The very same procedure works if we do _not_ have a functional form for
\( f(x) dx \). If this is a histogram, we need simply to calculate the integral version,
and perform the inverse function operation as above. Simple interpolation of a
curve is generally necessary.
Example Figure 2.9 shows an example of choosing uniformly distributed random numbers and transforming them to follow the frequency distribution prescribed by a given histogram.

![Histograms and plots](image)

Figure 2.9 An example of generating a Monte Carlo distribution following a known histogram. (a) The step-ladder histogram, with points from 2000 trials, produced by (i) integrating the function (b) and (ii) transforming the axes to produce the inverse function of the integrated distribution (c). The points with $\sqrt{N}$ error bars in the left diagram are from drawing 2000 uniformly distributed random numbers and transforming them according to the right diagram.

How do we draw numbers obeying some arbitrary distribution? The prescription above is all very well, and works when integration of the function can be done; it cannot in many cases, the Gaussian being an obvious one. There is another method, the rejection method, of generating random numbers to a prescription, and it can be coded in just a few lines. Details are in Lyons (1986) and Press et al. (2007).

The algorithm assumes that we need random numbers distributed according to some awkward function $f$, but which is ‘covered’ by a function $g$ from which we can get random numbers. This means that $f \leq Mg$ always, for some positive constant $M$. The steps of the algorithm are simple:

(i) Draw a random number $X_i$ from $g$.
(ii) Draw a random number $U_i$, uniformly distributed between 0 and 1.
(iii) If $f(X_i)/Mg(X_i) \geq U_i$, accept $X_i$.
(iv) Otherwise reject $X_i$ and repeat.
A nice proof of why this works is in MacKay (2003). While simple, the method will become very inefficient in high dimensions — but it is at least a simple route to getting samples from a multivariate distribution.

As a Monte Carlo product to which we shall turn a number of times during the course of this, let us invent a very simple universe. We will add complexity to it as we go; and it will serve to illustrate a number of specific issues.

**Example** Figure 2.10 shows the opening stages of creating our own little universe. It is in Euclidean space; its extent is from 0 (the centre) to 1.0 (the edge; we told you it was a model). We fill it uniformly with $10^6$ objects; to do so, we simply need $10^6$ values of $(r, \theta, \phi)$, with $0 < r < 1$, $0 < \phi < 2\pi$ and $0 < \theta < \pi$ where these are the conventional spherical coordinates. These must be chosen such that each volume element $r^2 \sin \theta \, dr \, d\theta \, d\phi$ is uniformly populated. As volume is $\propto r^3$, we need an inverse function to convert our uniform random numbers (0.0 to 1.0) to random distances $R$ via the inverse function, i.e. $R = R^{1/3}$ where $R$ is the random number selected uniformly over the range 0 to 1. We assign every object the same luminosity of $L = 10$ units. We suppose our universe is clean and non-relativistic, so that the flux received from each object is $L/r^2$.

![Figure 2.10](image)

(a) Looking straight down the $Y$-axis of our 3D toy universe with $R_{\text{max}} = 1.0$ and uniformly distributed objects plotted in $(x, y, z)$; this is a thin slice $\pm \Delta y$ about $y = 0$; (b) assigning all objects the same luminosity $L$, pretending that we are at the centre and observing with a telescope of sensitivity $0.1L/R_{\text{max}}^2$, here is a typical patch of sky. The sizes of the circles are proportional to the flux measured by our telescope. Note the similar sorts of pattern to those seen in Figure 2.4, the ‘Poisson clumping’.
We shall return to random numbers – in Chapters 6 and 7 in particular where we discover how to use them to do difficult integration via Monte Carlo techniques (Section 7.6), and how to do some important tests such as bootstrapping and jackknife (Section 6.6). In the interim, we urge you to gain familiarity with random number generation: it is mandatory for examining what data can tell you via probabilities and statistical inference. Besides, you may find it quite entertaining; see the following exercises.

Exercises

2.1 A warm-up on coin-tossing. This is not an astronomical problem but does provide a warm-up exercise on probability and random number generation. Every computer has a way of producing a random number between 0 and 1. Use this to simulate a simple coin-tossing game where player A gets a point for heads, player B a point for tails. Guess how often in a game of \( N \) tosses the lead will change; if A is in the lead at toss \( k \), when was the previous change of lead most likely to be? And by how much is a player typically in the lead? Try to back up these guesses with calculations. For many more game-based illustrations of probability, see Haigh (1999).

2.2 Efficient choosing. Imagine you are on a 10-night observing run with a colleague, in settled weather. You have an agreement that one of the nights, of your choosing, will be for your exclusive use. Show that, if you wait for five nights and then choose the first night that is better than any of the five, you have a 25 per cent chance of getting the best night of the 10. For a somewhat harder challenge, show that the optimum length of the ‘training sample’ is a fraction \( 1/e \) of the total. Check your results with random-number generation.

2.3 Bayesian inference. Consider the proverbial bad penny, for which prior information has indicated that there is a probability of 0.99 that it is unbiased (‘OK’); or a probability of 0.01 that it is double-headed (‘worse’). What is the (Bayesian) posterior probability, given this information, of obtaining seven heads in a row? In such a circumstance, how might we consider the fairness of the coin? Or of the experimenter who provided us with the prior information? What are the odds on the penny being fair?

2.4 Laplace’s rule and priors. Laplace’s rule (Section 2.3) \( \hat{\theta} = (N + 1)/(N + M + 2) \) depends on our prior for \( \theta \). If we have our
success and no failures, consider what the rule implies, and discuss why this is odd. How is the rule changed for alternative priors, for example Haldane's?

2.5 Bayesian reasoning in an everyday situation. The probability of a certain medical test being positive is 90 per cent, if the patient has disease D. If your doctor tells you the test is positive, what are your chances of having the disease? If your doctor also tells you that 1 per cent of the population have the disease, and that the test will record a false positive 10 per cent of the time, use Bayes' theorem to calculate the chance of having D if the test is positive. Simulate the experiment via Monte Carlo.

2.6 Inverse $\chi^2$ statistic. For a Gaussian of known mean (say, zero), show that the posterior distribution for the variance is inverse $\chi^2$. Use the 'Jeffreys' prior' for the variance: $\text{prob}(\sigma) = 1/\sigma$. Comment on the differences between this result and the one obtained by using a uniform prior on $\sigma$.

2.7 Maximum likelihood and the Poisson distribution. Suppose we have data which obey a Poisson distribution with parameter $\mu$, and in successive identical intervals we observe $n_1, n_2 \ldots$ events. Form the likelihood function by taking the product of the distributions for each $n_i$, and differentiate to find the maximum likelihood estimate of $\mu$. Is it what you expect?

2.8 Maximum likelihood and the exponential distribution. Suppose we have data $X_1, X_2 \ldots$ from the distribution $1/2a \exp(- |x|/a)$. Compute the posterior distribution of $a$ for a uniform prior, and Jeffreys' prior $\text{prob}(a) \propto 1/a$. Do the differences seem reasonable? Which prior would you choose? If $a$ were known, but the location $\mu$ was to be found, what would be the maximum likelihood estimate?

2.9 Birth control. Imagine a society where boys and girls were (biologically) equally likely to be born, but families cease producing children after the birth of the first boy. Are there more males than females in the population? Attack the problem in three ways: pure thought, by a Monte Carlo simulation, and by an analytic calculation.

2.10 Univariate random numbers. Work out the inverses of the integral functions required to generate (a) $f(x) = 2x^3$, (b) a power law, representative of luminosity functions, $f(x) = x^{-\nu}$. Use these results to produce random experiments following these probabilities by drawing 1000 random samples uniformly distributed between 0 and 1; verify by comparison with the given functions.

2.11 Make your own toy universe. Set up a toy universe as in the example. With a single luminosity of, say, $L = 10$ units and the telescope sensitive
to fluxes down to $0.1L/r_{\text{max}}^{2}$, calculate the number density per unit area as a function of received flux, and plot this source count. Now assign luminosities at random to the objects, following a power law in luminosity with slope of $-3$. ‘Resurvey’ the sky and calculate the source count again. Understand the results. This example is a prelude to Chapter 8.

Lies, damn lies, and statistics.

In embarking on the Gaussian approach, the use of statistics were no longer out of favour. There were indirect and conflated links to hypothesis testing and procedure alternatives to the philosophical work associated with this, presenting the possibility to make clear the

Statistics are defined as a branch of a statistical science. Given $X_1, X_2, \ldots$, some data, the aim is to find a single value or the average value. The finite amounts of information can try to distinguish the population alone, by upper limits, we will denote this

The summary is: (a) location and (b)