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, techniques 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 tempts 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) ...

Table 1.1 *Stages in astronomy experimentation*

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

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-needing 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, 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*

Year	Individual(s)	Key words	Events
∼1340	William of Ockham, or Occam	**Occam's Razor**	'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.
1654	Pascal, Fermat	**odds, probability theory**	Gombaud, Chevalier de Mere & Mitton pose questions on gambling odds to Pascal in ∼1654. Seven letters exchanged between Blaise Pascal & Pierre de Fermat are the genesis of probability theory.
1657	Huygens	**probability**	First publication on probability, 14 problems (+ solutions) in gambling, based on the Pascal–Fermat correspondence; the only publication on the subject for 50 years.
1662	Graunt	**descriptive statistics, life tables, survival analysis**	Publication of Graunt's *Observations on the Bills of Mortality*; first known collection and analysis of data for statistical purposes; start of actuarial risk analysis.
1692	Huygens, Arbuthnot	**probability**	*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.
1665–1676	Newton, Leibniz	**calculus**	Newton & Leibniz independently discover calculus; their dispute runs for decades. Probability theory can proceed.
1687	Newton	**binomial distribution**	In the monumental *Principia*, Newton changes the direction of physics and mathematics forever; the book includes the binomial probability distribution.