1

Science and statistical data analysis

> It is remarkable that a science which began with the consideration of games of chance should have become the most important object of human knowledge.
>
> Pierre-Simon Laplace (1812)
> *Théorie Analytique des Probabilités*

Why should a scientist bother with statistics? Because science is about dealing rigorously with uncertainty, and the tools to accomplish this are statistical. Statistics and data analysis are an indispensable part of modern science.

In scientific work we look for relationships between phenomena, and try to uncover the underlying patterns or laws. But science is not just an 'armchair' activity where we can make progress by pure thought. Our ideas about the workings of the world must somehow be connected to what actually goes on in the world. Scientists perform experiments and make observations to look for new connections, test ideas, estimate quantities or identify qualities of phenomena. However, experimental data are never perfect. Statistical data analysis is the set of tools that helps scientists handle the limitations and uncertainties that always come with data. The purpose of statistical data analysis is *insight* not just *numbers*. (That's why the book is called *Scientific Inference* and not something more like *Statistics for Physics*.)

1.1 Scientific method

Broadly speaking, science is the investigation of the physical world and its phenomena by experimentation. There are different schools of thought about the philosophy of science and the scientific method, but there are some elements that almost everyone agrees are components of the scientific method.

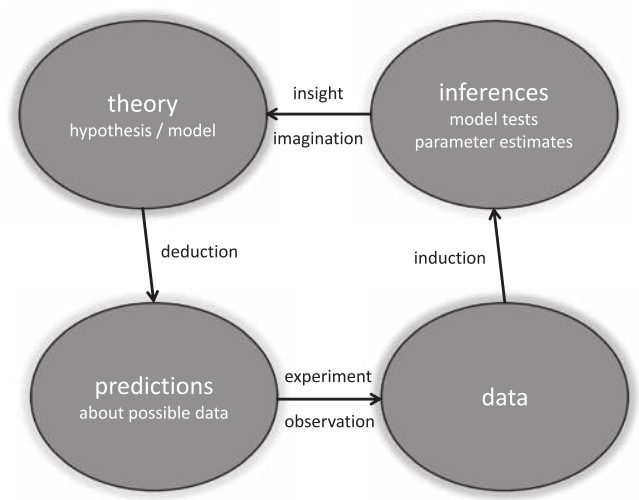

Figure 1.1 A cartoon of a simplified model of the scientific method.

Hypothesis A hypothesis or model is an explanation of a phenomenon in terms of others (usually written in terms of relations or equations), or the suggestion of a connection between phenomena.

Prediction A useful hypothesis will allow predictions to be made about the outcome of experiments or observations.

Observation The collection of experimental data in order to investigate a phenomenon.

Inference A comparison between predictions and observations that allows us to learn about the hypothesis or model.

What distinguishes science from other disciplines is the insistence that ideas be tested against what actually happens in Nature. In particular, hypotheses must make predictions that can be tested against observations. Observations that match closely the predictions of a hypothesis are considered as evidence in support of the hypothesis, but observations that differ significantly from the predictions count as evidence against the hypothesis. If a hypothesis makes no predictions about possible observations, how can we learn about it through observation?

Figure 1.1 gives a summary of a simplified scientific method. Models and hypotheses[1] can be used to make predictions about what we can observe.

[1] The terms 'hypothesis', 'model' and 'theory' have slightly different meanings but are often used interchangeably in casual discussions. A *theory* is usually a reasonably comprehensive, abstract framework (of definitions, assumptions and relations or equations) for describing generally a set of phenomena, that has been tested and found at least some degree of acceptance. Examples of scientific theories are classical mechanics, thermodynamics, germ theory, kinetic theory of gases, plate tectonics etc. A *model* is usually more specific. It might be the application of a theory to a particular situation, e.g. a classical mechanics model of the orbit of Jupiter. Some

Hypotheses may come from some more general theory, or may be more ad hoc, based on intuition or guesswork about the way some phenomenon might work. Experiments or observations of the phenomenon can be made, and the results compared with the predictions of the hypothesis. This comparison allows one to *test* the model and/or *estimate* any unknown parameters. Any mismatch between data and model predictions, or other unpredicted findings in the data, may suggest ways to revise or change the model. This process of learning about hypotheses from data is scientific inference. One may enter the cycle at any point: by proposing a model, making predictions from an existing model, collecting data on some phenomenon or using data to test a model or estimate some of its parameters. In many areas of modern science, the different aspects have become so specialised that few, if any, researchers practice all of these activities (from theory to experiment and back), but all scientists need an appreciation of the other steps in order to understand the 'big picture'. This book focuses on the induction/inference part of the chain.

1.2 Inference

The process of drawing conclusions based on what is already known is called *inference*. There are two types of reasoning process used in inference: deductive and non-deductive.

1.2.1 Deductive reasoning (from general to specific)

The first kind of reasoning is *deductive reasoning*. This starts with premises and follows the rules of logic to arrive at conclusions. The conclusions are therefore true as long as the premises are true. Philosophers say the premises entail the conclusion. Mathematics is based on deductive reasoning: we start from axioms, follow the rules of logic and arrive at theorems. (Theorems should be distinguished from theories – the former are the product of deductive reasoning; the latter are not.) For example, the two propositions 'A is true implies B is true' and 'A is true' together imply 'B is true'. This type of argument is a simple deduction known as a syllogism, which comprises a major premise and a minor premise; together they imply a conclusion:

$$\begin{aligned} \text{Major premise} &: A \Rightarrow B \text{ (read: } A \text{ is true implies } B \text{ is true)} \\ \text{Minor premise} &: A \text{ (read: } A \text{ is true)} \\ \text{Conclusion} &: B \text{ (read: } B \text{ is true).} \end{aligned}$$

Deductive reasoning leads to conclusions, or theorems, that are inescapable given the axioms. One can then use the axioms and theorems together to deduce more

authors go on to distinguish *hypotheses* as models, and their parameters, which may be speculative, as they are used in statistical inference. For now we have no need to distinguish between models and hypotheses.

theorems, and so on. A theorem[2] is something like '$A \Rightarrow B$', which simply says that the truth value of A is transferred to B, but it does not, in and of itself, assert that A or B are true. If we happen to know that A is indeed true, the theorem tells us that B must also be true. The box gives a simple proof that there is no largest prime number, a purely deductive argument that leads to an ineluctable conclusion.

Box 1.1
Deduction example – proof of no largest prime number

- Suppose there is a largest prime number; call this p_N, the Nth prime.
- Make a list of each and every prime number: $p_1 = 2$, $p_2 = 3$, $p_3 = 5$, until p_N.
- Now form a new number q from the product of the N primes in the list, and add one:

$$q = 1 + \prod_{i=1}^{N} p_i = 1 + (p_1 \times p_2 \times p_3 \times \cdots \times p_N) \tag{1.1}$$

  which is either prime or it is not.
- This new number q is larger than every prime in the list, but it is not divisible by any prime in the list – it always leaves a remainder of one.
- This means q is prime since it has no prime factors (the fundamental theorem of arithmetic says that any integer larger than 1 has a unique prime factorisation).
- But this is a contradiction. We have found a prime number q that is larger than every number in our list, in contradiction with our definition of p_N. Therefore our original assumption – that there is a largest prime, p_N – must be false.

Deduction involves reasoning from the general to the specific. If a general principle is true, we can conclude that any particular cases satisfying the general principle are true. For example:

Major premise : All monkeys like bananas
Minor premise : Zippy is a monkey
Conclusion : Zippy likes bananas.

The conclusion is unavoidable given the premises. (This type of argument is given the technical name *modus ponens* by philosophers of logic.) If some theory is true we can predict that its consequences must also be true. This applies to probabilistic as well as deterministic theories. Later on we consider flipping coins, rolling dice, and other random events. Although we cannot precisely predict the outcome of

[2] It is worth noting here that the logical implication used above, e.g. $B \Rightarrow A$, does not mean that A can be derived from B, but only that if B is true then A must also be true, or that the propositions 'B is true' and 'B and A are both true' must have the same truth value (both true, or both false).

individual events (they are random!), we can derive frequencies for the various outcomes in repeated events.

1.2.2 Inductive reasoning (from specific to general)

Inductive reasoning is a type of non-deductive reasoning. Induction is often said to describe arguments from special cases to general ones, or from effects to causes. For example, if we observe that the Sun has risen every day for many days, we can inductively reason that it will continue to do so. We cannot directly deduce that the Sun will rise tomorrow (there is no logical contradiction implied if it does not).

The basic point about the limited power of our inferences about the real world (i.e. our inductive reasoning) was made most forcefully by the Scottish philosopher David Hume (1711–1776), and is now known as the problem of induction. The philosopher and mathematician Bertrand Russell furnished us with a memorable example in his book *The Problems of Philosophy* (Russell, 1997, ch. 4):

> imagine a chicken that gets fed by the farmer every day and so, quite understandably, imagines that this will always be the case . . . until the farmer wrings its neck! The chicken never expected that to happen; how could it? – given it had no experience of such an event and the uniformity of its previous experience had been so great as to lead it to assume the pattern it had always observed (chicken gets fed every day) was universally true. But the chicken was wrong.[3]

You can see that inductive reasoning does not have the same power as deductive reasoning: a conclusion arrived at by deductive reasoning is necessarily true if the premises are true, whereas a conclusion arrived at by inductive reasoning is not *necessarily* true, it is based on incomplete information. We cannot deduce (prove) that the Sun will rise tomorrow, but nevertheless we do have confidence that it will. We might say that deductive reasoning concerns statements that are either true or false, whereas inductive reasoning concerns statements whose truth value is unknown, about which we are better to speak in terms of 'degree of belief' or 'confidence'. Let's see an example:

Major premise : All monkeys we have studied like grapes
Minor premise : Zippy is a monkey
Conclusion : Zippy likes grapes.

The conclusion is not unavoidable, other conclusions are allowed. There is no logical contradiction in concluding

Conclusion : Zippy does not like grapes.

[3] By permission of Oxford University Press.

But the premises do give us some information. It seems plausible, even probable, that Zippy likes grapes.

1.2.3 Abductive reasoning (inference to the best explanation)

There is another kind of non-deductive inference, called *abduction*, or *inference to the best explanation*. For our purposes, it does not matter whether abduction is a particular type of induction, or another kind of non-deductive inference alongside induction. Let's go straight to an example:

Premise : Nelly likes bananas
Premise : The banana left near to Nelly has been eaten
Conclusion : Nelly ate the banana.

Again the conclusion is not unavoidable, other conclusions are valid. Perhaps someone else ate the banana. But the original conclusion seems to be in some sense the simplest of those allowed. This kind of reasoning, from observed data to an explanation, is used all the time in science.

Induction and abduction are closely related. When we make an inductive inference from the limited observed data ('the monkeys in our sample like grapes') to unobserved data ('Zippy likes grapes') it is as if we implicitly passed through a theory ('all monkeys like grapes') and then deduced the conclusion from this.

1.3 Scientific inference

Scientific work employs all the above forms of reasoning. We use deductive reasoning to go from general theories to specific predictions about the data we could observe, and non-deductive reasoning to go from our limited data to general conclusions about unobserved cases or theories.

Imagine A is the theory of classical mechanics and B is the predicted path of a rocket deduced from the theory and the details of the launch. Now, we make some observations and find the rocket did indeed follow the predicted path B (as well as we can determine). Can we conclude that A is true? We may infer A, but not deductively. Other conclusions are possible. In fact, the observational confirmation of one prediction (or even a thousand) does not *prove* the theory in the same sense as a deductive proof. A different theory may make indistinguishable predictions in all of the cases considered to date, but differ in its predictions for other (e.g. future) observations.

Experimental and observational science is all about inductive reasoning, going from a finite number of observations or results to a general conclusion about

unobserved cases (induction), or a theory that explains them (abduction). In recent years, there has been a lot of interest in showing that inductive reasoning can be formalised in a manner similar to deductive reasoning, so long as one allows for the uncertainty in the data and therefore in the conclusions (Jeffreys, 1961; Jaynes, 2003).

You might still have reservations about the need for statistical reasoning. After all, the great experimental physicist Ernest Rutherford is supposed to have said

> If your experiment needs statistics, you ought to have done a better experiment![4]

Rutherford probably didn't say this, or didn't mean for it to be taken at face value. Nevertheless, statistician Bradley Efron, about a hundred years later, contrasted this simplistic view with the challenges of modern science (Efron, 2005):

> Rutherford lived in a rich man's world of scientific experimentation, where nature generously provided boatloads of data, enough for the law of large numbers to squelch any noise. Nature has gotten more tight-fisted with modern physicists. They are asking harder questions, ones where the data is thin on the ground, and where efficient inference becomes a necessity. In short, they have started playing in our ball park.

But it is not just scientists who use (or should use) statistical data analysis. Any time you have to draw conclusions from data you will make use of these skills. This is true for particle physics as well as journalism, and whether the data form part of your research or come from a medical test you were given you need to be able to understand and interpret them properly, making inferences using methods built on the same basic principles.

1.4 Data analysis in a nutshell

The analysis of data[5] can be broken into different modes that are employed either individually or in combination; the outcome of one mode of analysis may inform the application of other modes.

Data reduction This is the process of converting *raw* data into something more useful or meaningful to the experimenter: for example, converting the voltage changes in a particle detector (e.g. a proportional counter) into the records of the times and energies of individual particle detections. In turn, these may be further reduced into an energy spectrum for a specific type of particle.

[4] The earliest reference to this phrase I can find is Bailey (1967, ch. 2, p. 23).
[5] 'Data' is the plural of 'datum' and means 'items of information', although it has now become acceptable to use 'data' as a singular mass noun rather like 'information'.

Exploratory data analysis (EDA) is an approach to data analysis that uses quantitative and graphical methods in an attempt to reveal new and interesting patterns in the data. One does not test a particular hypothesis, but instead 'plays around with the data', searching for patterns suggestive of new hypotheses.

Inferential data analysis Sometimes known as 'confirmational data analysis'. We can divide this into two main tasks: model checking and parameter estimation. The former is the process of choosing which of a set of models provides the most convincing explanation of the data; the latter is the process of estimating values of a model's unknown parameters.

Exploratory data analysis is all about summarising the data in ways that might provide clues about their nature, and inferential data analysis is about making reasonable and justified inferences based on the data and some set of hypotheses.

1.5 Random samples

Our data about the real world are almost always incomplete, affected by random errors, or both. Let's say we wanted to find the answer to some important question: does the UK population prefer red or green sweets? We could survey the entire population and in principle get a complete answer, but this would normally be impractical. So we settle for a subset of the population, and assume this is representative of the population at large. Our results from the subset of people we actually survey is a *sample* and this is drawn from some *population* (of all the responses from the entire population). The sample is just one of the many possible samples that could be obtained from the same population.

But what we're interested in is the population, so we need to use what we know about the sample to infer something about the population. A small sample is easy to collect, but smaller samples are also more susceptible to random fluctuations (think of surveying just one person and extrapolating his/her answer to the entire population); a larger sample is less prone to such fluctuations but is also harder to collect. We also need to be sure to sample randomly and in an unbiased fashion – if we only sample younger people, or people in certain counties, these may not reflect the wider population. We need ways to quantify the properties of the sample, and also to quantify what we can learn about the population. This is statistics.

You may be left thinking: what's this got to do with experiments in the physical sciences? We often don't have a simple population from which we pull a random sample. Each time we perform some measurement (or series of measurements) we are collecting a sample of possible data. We can think of our sample as being drawn from a population, a hypothetical population of all the possible data that could be

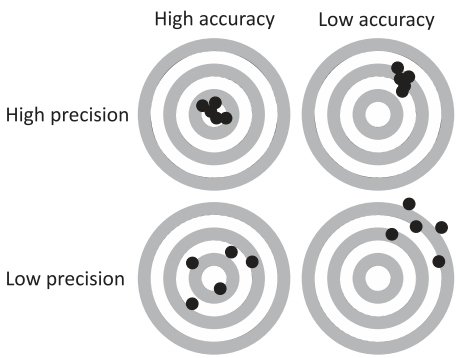

Figure 1.2 Illustration of the distinct concepts of accuracy and precision as applied to the positions of 'shot' on a target.

produced from our measurement(s). The differences between samples are due to randomness in the experiment or measurement processes.

1.5.1 Errors and uncertainty

The type of randomness described above is usually called *random error* (or measurement error) by physicists (the term *error* is used differently by statisticians[6]). Here, *error* does not mean a mistake as in the usual sense. To most scientists the 'measurement error' is an estimate of the repeatability of a measurement. If we take some data and use them to infer the speed of sound through air, what is the error on our measurement? If we repeat the entire experiment – under almost identical conditions – chances are the next measurements will be slightly different, by some unpredictable amount. As will further repeats. The 'random error' is a quantitative indication of how close repeated results will be. Data with small errors are said to have high *precision* – if we repeat the measurement the next value is likely to be very close to the previous value(s).

In addition to random errors, there is another type of error called *systematic error*. A systematic error is a bias in a measurement that leads to the values being systematically either too low or too high, and may arise from the selection of the sample under study or the calibration of the instrument used. Data with small systematic error are said to be *accurate*; if only we could reduce the random error we could get a result extremely close to the 'true' value. Figure 1.2 illustrates the difference between precision and accuracy. The experimenter usually works to reduce the impact of both random and systematic errors (by 'beating down the

[6] To a statistician, 'error' is a technical term for the discrepancy between what is observed and what is expected.

errors') in the design and execution of the experiment, but the reality is that such errors can never be completely eliminated.

It is important to distinguish between *accuracy* and *precision*. These two concepts are illustrated in Figure 1.2. Precise data are narrowly spread, whereas accurate data have values that fall (on average) around the true value. Precision is an indicator of variation within the data and accuracy is a measure of variation between the data and some 'true' value. These apply to direct measurements of simple quantities and also to more complicated estimates of derived quantities (Chapters 6 and 7).

1.6 Know your data

There are several types of data you may be confronted with. The main types are as follows.

Categorical data take on values that are not numerical but can be placed in distinct categories. For example, records of gender (male, female) and particle type (electron, pion, muon, proton etc.) are categorical data.

Ordinal data have values that can be ranked (put in order) or have a rating scale attached, but the differences between the ranks cannot be compared. An example is the Likert-type scale that you see on many surveys: 1, strongly disagree; 2, disagree; 3, neutral; 4, agree; 5, strongly agree. These have a definite order, but the difference between options 1 and 2 might not be the same as between options 3 and 4.

Discrete data have numerical values that are distinct and separate (e.g. 1, 2, 3, . . .). Examples from physics might be the number of planets around stars, or the number of particles detected in a certain time interval.

Continuous data may take on any value within a finite or infinite interval. You can count, order and measure continuous data: for example, the energy of an accelerated particle, temperature of a star, ocean depth, magnetic field strength etc.

Furthermore, data may have many dimensions.

Univariate data concern only one variable (e.g. the temperature of each star in a sample).

Bivariate data concern two variables (e.g. the temperatures and luminosity of stars in a sample). Each data point contains two values, like the coordinates of a point on a plane.

Multivariate data concern several variables (e.g. temperature, luminosity, distance etc. of stars). Each data point is a point in an N-dimensional space, or an N-dimensional vector.