1 Introduction: Statistics, Non-mathematical Methods and How to Use this Book

This chapter introduces statistics and the approach taken in this book. Statistics is an important subject, but many aspects of it are difficult and misconceptions common. To try to get round some of the difficulties, this book takes a non-mathematical approach to the subject, which draws on various ideas, including computer simulation, and the random choice of balls from a bucket as a metaphor for probability and other statistical concepts. This chapter also outlines the (minimal) arithmetical expertise you will need, and makes some suggestions about computer software, and how to approach statistics and this book.

▶ 1.1 Statistics

The word *statistics* comes from the same root as the word for state, which reflects the fact that statistics originally referred to the use of data by the state. The scope of statistics has now spread far wider than this, and the term itself is commonly used in three senses. Statistics on earnings, sport, or the weather, are simply lists of numbers telling us such things as how much we earn, how many goals have been scored, or how much the earth is warming up. Statisticians refer to these numbers as 'data'. Secondly, we may refer to something calculated from this data as a statistic, for example the average earnings of teachers in the UK in 2001. The third meaning of statistics is to refer to the science which helps to analyse data,[1] and draw inferences from it, often with the help of the idea of probability. This is the subject of this book.

The methods of statistics and probability are useful when you're not quite sure what's going on. When things are certain and completely predictable, you don't need statistics; whenever there are uncertainties, or things you can't predict, statistics may have a role to play. There are three, interlinked, things you can do with statistics.

The first is that you can make predictions about what will happen. For example, the following predictions have been made with the help of statistics:

- The earth is heating up and average temperatures are likely to rise by three degrees Celsius by the 2080s.[2]
- My life expectancy is another 33 years: this is a statistically based 'best guess' about how many more years I will live.[3]
- There is 1 chance in 14 000 000 of a UK national lottery ticket winning the jackpot, but being killed by an asteroid crashing into the earth is, apparently, more likely than this: according to one estimate, 750 times more likely.[4]

None of these predictions is exact. A probabilistic prediction, like the last of these, only aims to say how probable things are, not to make definite predictions; this is, however, often the best we can do. My life expectancy is an average figure – it is the average we would expect over lots of different men of my age – so there is no implication that I will survive exactly 33 years, indeed this would be most improbable. I may drop dead tomorrow, and I may live another 60 years: both are unlikely, but experts could attach a probability to them.

The second thing you can do with statistics is build a statistical 'model': this is just a description of how a situation works in probabilistic or 'averaging' terms. (The use of the word 'model' in this context may strike you as odd, I'll return to this in Section 1.4.) For example, there is very clear evidence for a link between smoking and lung cancer: other things being equal, a person smoking 20 cigarettes a day has about a 20 times greater chance of developing lung cancer than a non-smoker.[5] A more detailed model would incorporate other factors, such as age, gender, extent of passive smoking and so on, and arrive at a more realistic assessment of the risks.

Models like this could be used to predict the number of smokers who will develop lung cancer, and the predictions above are all based on models. However, that is not really the purpose of this particular model. The main value of this model is the insight it gives us into the relationship between smoking and lung cancer and the folly of smoking. This insight might be used to estimate the savings to the health service, or the increased cost of pensions, if the number of smokers were to be reduced by a given amount (see Chapter 9 for this type of model).

The third role of statistics is to answer questions about the strength of evidence and how much certainty can be attached to both predictions and models. This can be done in several ways, of which the most useful is often to express estimates as intervals. Instead of citing a prediction from an opinion poll (based on a sample of 1000 electors) that 38% of electors will vote Labour, we can use statistics to derive the less precise, but more realistic, prediction that we can be 95%[6] confident that the number of Labour voters will be somewhere between 35% and 41%. The extent of the impreci-

sion in this prediction is the width of this interval: 6%. If we wanted a more precise prediction with a narrower interval, we would need a larger sample for the opinion poll. Statistical methods can tell us what size we need (see Chapter 7).

It is difficult to overstate the importance of statistics. Economics, education, health care, business, weather forecasting, running a football club, getting elected to run a government, gambling or analysing a survey or an experiment all benefit from the thoughtful use of statistical methods. The modern view of the subatomic world is based on the theory of quantum mechanics, which cannot predict accurately what will happen to individual particles, but only what will *probably* happen or what will happen on *average*. Statistical ideas are an integral part of most academic disciplines: you need to know what statistics is about to do useful research in just about anything.

The same applies to everyday life. Whenever we're not quite sure what's going on, or exactly what the important influences or trends are, then the best approach, often the only approach, may be to use averages and probabilities: we need the statistical approach. Despite this, there are other ways of dealing with uncertainty that don't come under the umbrella of statistics: 'fuzzy logic' and the idea of 'chaos', for example. We'll look briefly at these in Chapter 4.

Statistics is not the most popular subject on the curriculum, and errors and misconceptions are common. In the next section, I'll look briefly at some of the difficulties of statistics, which leads into the rationale behind the non-mathematical approach taken in this book.

► 1.2 The difficulties of statistics

Statistics has a bad press for a variety of reasons. At an elementary level, the suspicion tends to be that results are meaningless or misleading because the data on which they are based is distorted in some way, sometimes deliberately so.[7] There are, after all, 'lies, damned lies, and statistics'.[8] When we are told that men think about sex every six seconds, or that 86% of married women committed adultery in the last year, it's natural to wonder where the information comes from, and how it could be obtained without a serious danger of distortion.

Even when data comes from a supposedly respectable source, you should be cautious. The 33-year estimate of my life expectancy is based on tables used by actuarial students in the year 2000, but they are based 'on the mortality of the male population of England and Wales in the years 1960–62'.[9] The data is based on mortality rates 40 years in the past, rather than the coming century, which is when the events responsible for my death will occur.

A different sort of difficulty arises from the statement that you are 750 times more likely to die as a result of an asteroid crashing into the earth, than to win the jackpot on the UK national lottery. There must be several thousand jackpot winners alive in the UK today; does this mean that there were 750 times this number of deaths from asteroid collisions in the last few years, and that it is far more sensible to take precautions against asteroids than it is to enter the national lottery?

It is not hard to see where the difficulties lie here.[10] Common sense and a determination to think clearly is all that is really required. At a more advanced level, however, this may not be good enough, because the statistical concepts and methods are often too convoluted and complex. For example, according to an article in the *Guardian*, 30 years ago there was a cancer cluster around the town of Aldermaston with a 'probability of chance occurrence' of 1 in 10 000 000.[11] Aldermaston was the site of a nuclear installation; the suspicion being, of course, that this was responsible for the abnormally high level of cancers in the surrounding neighbourhood. The article gives a layman's account of a common type of statistical analysis, a 'null hypothesis test'. The research raises lots of obvious questions, the most obvious being how clusters are defined, but what I want to concentrate on here is what the statistical result, a probability of 1 in 10 000 000, actually means? There are two obvious interpretations, both wrong:

1. Does it refer to the chance of getting cancer, as most people to whom I have shown the article assume? No, it does not. The probability of 1 in 10 000 000 has no relation whatsoever to the chances of contracting cancer (which is far higher than this).
2. Does it refer to the probability of the cluster being an accident and having nothing to do with the nuclear installation? This would imply that there is a 99.9999% chance of the nuclear installation being responsible for the cluster. This sounds plausible, but is again wrong.

The correct interpretation is the third, which, in my experience occurs to few, if any, statistically unsophisticated readers of the original article:

3. The probability actually refers to the chance of the cluster occurring *on the assumption that cancer cases occur at random*: that is, on the assumption that the nuclear installation had no impact. This probability differs from interpretation 2 in the same way that the probability that a successful bank robber is rich – presumably close to 100% – differs from the probability that a rich person is a successful bank robber – presumably close to 0% (Section 2.2).

Interpretations 1 and 2 both correspond to information which would be very useful to have. This is probably the reason for people assuming that one of these interpretations is correct. The correct interpretation, on the other hand, is difficult to get one's mind round, and does not tell us anything we really want to know. To the mind untutored in statistics, if we are interested in the possibility that the nuclear installation has got something to do with the cancer clusters, why mess about with probabilities based on the assumption that it hasn't? On top of that, the mathematical details of how it works are far too complex for the article to explain. These have to be taken on trust, and cannot serve to clarify the meaning of the conclusion. (We'll return to this example in Chapter 8.)

A very similar difficulty arose in the trial of Sally Clark who was accused of murdering two of her children.[12] One of the alternative explanations for the children's deaths was sudden infant death syndrome (SIDS or 'cot death'). At her original trial, a figure of 1 in 73 000 000 was quoted for the probability of having two SIDS cases in a single family. This was part of the evidence that led to her conviction, which was later overturned, but not until she had spent several years in prison.

There are three problems with this 1 in 73 000 000. First, it's wrong (see Section 5.2). Second, it's irrelevant: the defence case was the deaths were natural, but not that they were due to SIDS.[13] And third, despite its incorrectness and irrelevance, the 1 in 73 000 000 is liable to be misinterpreted as the chance that the deaths were natural, which seems to imply that the children almost certainly were murdered. This logic is completely wrong, in just the same way that the second interpretation of the Aldermaston data is wrong. In the legal context, it's known as the 'prosecutor's fallacy' (see also Exercise 6.7.3).

Unfortunately, these are not isolated examples.[14] Many statistical ideas are difficult and misconceptions are common. The idea of this book is to present a non-mathematical approach to statistics which will, I hope, make things a bit clearer. The next two sections introduce the basic elements of this approach.

▶ 1.3 Non-mathematical methods

Statistics is usually seen as a branch of mathematics. This makes non-mathematical statistics seem as sensible a concept as inedible food or illogical logic. However, I intend to prise the two domains apart by taking a particular perspective on statistics, and what some might regard as a rather restricted interpretation of mathematics. But the first question is, why

bother? Mathematical statistics seems to work OK, so what's the point in non-mathematical statistics?

There are two answers to this. The first is that mathematical statistics does not work satisfactorily, except for the experts. At anything beyond an elementary level, the conceptual and mathematical difficulties are substantial, as we have just seen. This leads to incomprehension on the part of novices trying to master the subject, and to techniques being misapplied and their results misinterpreted, even by people who think they know what they are doing.

The second advantage of the non-mathematical approach in this book is that aspects of it are sometimes superior to the conventional approach by criteria that a mathematician would appreciate: generality and the ability to solve difficult problems (for example some of the methods in Chapters 7 and 8). So what do I mean by the non-mathematical approach?

The non-mathematical[15] version of statistics presented here has *no* algebra or mathematical formulae or equations, and does not rely on mathematical proofs and computer packages doing mysterious things in mysterious ways. All arithmetical relations used are simple enough to be described in words or by means of a graph. This book makes *no* assumptions about the reader's mathematical understanding beyond elementary arithmetic and the use of simple graphs. I am assuming some familiarity with the four rules of arithmetic, fractions, decimals, percentages, negative numbers and simple graphs, but that's about it (see Section 1.5).

This is *not* the same as providing you with a gentle approach to the standard mathematical formulae of statistics. These formulae do not come into the story. They are bypassed; we get the answer without them, by using concepts and methods that are simpler and more direct than the conventional ones. For example, one problem in statistics is finding a line on a graph which gives the 'best fit' to some data (Section 9.2). There is a formula, but in this book we will use trial and error, which, when assisted by a computer, is surprisingly efficient. That way we avoid the formula, and all the hidden assumptions which you need to make but never realise you are making. This does not mean that you have to take what I say on trust. You should be able to see exactly how, and why, the non-mathematical methods work, and what the answers mean. The rationale behind the methods should be transparent. This means that the mathematical difficulties – in the sense of difficulties with equations and formulae – are eliminated. However, the other difficulties discussed above remain; problems of interpretation can then be faced without the distractions of mathematics.

You may, however, want to relate these non-mathematical methods to the formulae and concepts in other statistics books and in computer packages. To help you do this, most chapters include a section called Similar concepts:

these are described in terms of the non-mathematical methods developed in the chapter. For example, the method known as 'bootstrapping' described in Chapter 7 is a non-mathematical approach to a statistical concept known as a 'confidence interval'. The usual approach to confidence intervals develops them by means of the mathematical theory of probability: this leads to formulae for deriving confidence intervals. The Similar concepts section in Chapter 7 mentions these formulae. It does not, however, explain the use or rationale of these formulae in detail, as this would take us too far from our non-mathematical approach. The important thing is to understand what a confidence interval means, so that, for example, you know whether or not it is a sensible quantity to ask a computer package to calculate.

As the language of mathematics isn't being used to help make sense of statistics, we'll need some alternatives. One metaphor that is very useful is the idea of a bucket with some balls in it. This is introduced in the next section.

► 1.4 Bucket and ball models and computer simulation

The usual way of approaching statistics is to set up an abstract 'model' in which words are used in special ways. Probability, for example, is described in terms of 'experiments' with 'outcomes', despite the fact that most situations involving probability are not experiments, and many of the outcomes aren't really outcomes. You then consider a 'sample space', which, of course, isn't a space, write down the rules (called axioms) which probability is supposed to obey, and then use these rules to deduce more complicated rules, in other words to use the methods of mathematics to see where they lead. The final model is typically a set of mathematical equations.

My approach here is different. We start with a *physical* model – balls in a bucket – and then relate things to this as directly as possible, without using mathematics as an intermediary. Then you can see the concepts and methods of statistics in terms of something definite and visualisable. This means we can manage without any mathematics more advanced than simple arithmetic. I mean a model in the ordinary sense of the word, like a model train or a child's doll. By playing with the model you can learn about trains or people. In just the same way, by playing with the balls in a bucket, you can, for example, estimate probabilities and understand how they can be interpreted and used. Does this mean that you will need to invest in a bucket and a lot of balls before you can proceed further? Well, no, not really, although it might be helpful. It is enough to be able to *imagine* the balls in a bucket.

So why buckets and balls? When introducing ideas of probability, textbooks often illustrate concepts by references to examples about balls in urns, because they are easy to visualise and make the necessary points in a simple way. For example:

> There are two green balls and two black balls in an urn. What is the probability that a ball drawn *at random* will be black?

The phrase 'at random' here means that we choose a ball without checking its colour in such a way that all four balls are *equally likely* to be chosen. This could be achieved if all the balls were the same size, shape, weight and texture, they are thoroughly mixed beforehand and the choice is made with a blindfold on. The answer to the question, of course, is two out of four, or 1/2 or 50%, because two of the four balls are black. I will use buckets instead of urns because they are likely to be more familiar to you. An alternative image would be to think of a lottery.

We can use the model of balls in a bucket, or a lottery, for more complicated examples. A standard demonstration in lectures on simple probability is to ask members of the audience for their estimate of the probability of two members of the audience sharing the same birthday. This turns out to be more than most people think: for example in a group of 50 there is a 97% chance that there will be at least two people with the same birthday, so the lecturer can afford a substantial bet on it. Even in a group of 23 the chance is still as much as 51%. This is according to the mathematical theory of probability. How can we do it non-mathematically?

The answer is to build a model of the situation. Imagine that the lecture room is the bucket, and each member of the audience is represented by a ball. We could imagine a portrait on each ball, but as all we are interested in is birthdays, it is simpler to imagine a birthday stamped on each ball. Where do these people in the audience come from? We could imagine a second bucket, much bigger, containing balls representing people in the local community from which the members of the audience come. If we assume the audience are effectively drawn at random from this bigger bucket, we now have a model of the audience and where it comes from.

We can now run the model, or at least the important bits of it, on a computer. Imagine drawing the balls for the audience bucket from the bigger, local community, bucket. We're interested in birthdays, and it's reasonable to assume that each person's birthday is equally likely to be any of the 365 days of the year (ignoring leap years). This means we can *simulate* this process on a computer by generating 50 numbers, each chosen randomly from the range 1–365 representing each possible birthday. Computer programs (like Excel) have built-in functions for generating random numbers

like this – for behaving as though they were blindfolded and drawing balls from a bucket. If you have a computer and Excel, it would be a good idea to try it yourself to get a feel for how it works.[16]

When I did this the result was: 337 7 285 244 98 313 329 138 94 182 242 129 333 140 323 24 222 110 76 306 146 250 17 263 332 189 122 227 93 118 25 360 155 135 124 30 66 9 143 243 134 345 324 215 78 181 151 239 9 220. As you can see, there are two 'people' with the same birthday, the two 9s representing 9 January, in this 'audience'.

This, of course, is not a real audience. It's a hypothetical audience which we made by playing with our bucket and ball model. But if we now generate, say, 100 hypothetical audiences like this, we'll be in a position to say something about the real audience. To do this, a third bucket is helpful. Take one of the 100 hypothetical audiences, get a ball and put a tick on it if two people in this hypothetical audience share a birthday, otherwise put a cross on it. Repeat for each of the 100 hypothetical audiences. Put the resulting 100 balls in the third bucket.

When I generated 100 hypothetical audiences in this way, I found that 96 of the 100 balls in the third bucket were ticked. Ninety six per cent of the 100 hypothetical audiences had at least two people who shared a birthday. This suggests that about 96% of *real* audiences of 50 people will include at least two people who have the same birthday, assuming that real audiences are drawn in much the same way as the hypothetical audiences in the model. The estimate of the probability for two people sharing a birthday was 96%, not quite the 97% produced by the mathematical theory, but close enough to be useful. If I wanted a better answer, I would have to generate more 'audiences' and put more balls in the third bucket.

We don't *need* the image of buckets and balls here. We could imagine audiences of people without thinking of them as balls in a bucket. However, the idea of balls in a bucket is useful to remind us that we are dealing with a model, not with reality itself. (The 100 computer-generated audiences are imaginary, not real.) And when we come on to models which apply to a wide variety of different situations – such as the two bucket model in Chapter 5, or the pruned ball method in Chapter 6 – the bucket and ball image gives us a very useful way of describing what's going on. (The two bucket model will, in fact, simulate the birthday problem – see Section 5.4.)

The underlying metaphor may use buckets and balls, but the practical method involves 'computer simulation'. This is a very crude way of working out probabilities. All we do is run the process through on a computer lots of times and see what pattern emerges. This is, however, good enough for most purposes. Methods like this are often preferable in some ways to those obtained from the mathematical formulae: they tend to be of more general applicability and depend on fewer assumptions. Even professional

statisticians tend to use simulation methods when they are faced with a very difficult problem and cannot see a suitable mathematical formula. Very similar comments apply to the trial and error method I mentioned in Section 1.3. Like simulation, this is crude and longwinded, but very effective and it's obvious what's going on. I've called methods such as this 'crunchy methods'.[17]

► 1.5 Numbers, calculators and computers

The approach I am following in this book may be non-mathematical, but it does use numbers! I am assuming you are familiar with the four operations of addition, subtraction, multiplication and division. This includes negative numbers,[18] fractions, percentages, and the idea of squares and square roots.[19] Calculators are helpful for working out answers, but a computer and, in particular, a spreadsheet such as Microsoft Excel is far more useful. It is much more flexible, and all the numbers and calculations you enter are stored and displayed so that you can check them and change them as necessary. Whenever, possible, I would suggest using a computer for arithmetical calculations. A spreadsheet such as Excel is the most useful piece of computer software for studying statistics. There are some Excel files on the web (see Appendix C) to accompany this book, and there is a brief introduction to the use of the package in Appendix A.

For more serious work there are a range of computer packages specifically designed for statistical analysis. One such package is SPSS (Statistical Package for the Social Sciences). I have included a brief introduction to this package in Appendix B, and I have also explained how SPSS procedures relate to the topics covered in many of the chapters of this book. The purpose of this is to assist readers using SPSS for research purposes. You do not need SPSS to study statistics. To use the Excel files you will need to have Excel installed on your computer. Similarly, SPSS is a package which must be installed on your computer. If you have neither of these, you will still be able to use a program on the web – *`resample.exe`* – which works independently of any other software. This program is a computer version of what is introduced in Chapter 5 as the 'two bucket model'. It is also useful for quite a lot of the material in later chapters.

Computers are useful for studying statistics. You can key in some data, ask a question about it and get an almost immediate response, often in the form of a graph. Playing with a computer in this way is very helpful for building up intuitions about how concepts and methods work. The files on the web are designed to encourage this sort of interaction (for example you

could have a look at *reg2way.xls* for a preview of Chapter 9 – try adjusting the numbers in the green cells).

However, it is not essential to use a computer. I have written this book on the assumption that some readers will have a computer, but others won't. Despite this, some of the methods in this book depend on a computer; they would not be possible without one. In these cases, I have described what the software does in the text. If you have a computer, you can try it for yourself; otherwise you can simply follow the account in the text. At other points, the computer is not essential to the explanation. To avoid cluttering up the text, I have put details of how to use computer software in the notes which you will find at the end of the book.

I'll finish this section with a couple of reminders about numbers. You obviously don't need a spreadsheet to understand numbers, but if do have Excel available, you may find it clarifies even these simple points. It is often useful to write fractional numbers as percentages. For example, a third can be written as 1/3 or 0.3333 or 33.33%, and two-sevenths as 2/7, or 0.2857, or 28.57%. The first of these you can probably work out in your head, but you may need a calculator or computer for the second. Two-sevenths means 2 divided by 7. This is simple with a calculator. With Excel use the formula = 2/7 and you can then (if you want) format the cell (Format – Cells) as a percentage. If you do this, you should see 0.2857. . . change into the equivalent 28.57%.

If you write two-sevenths, or a third, as a decimal or a percentage, the sequence of digits goes on for ever, but Excel or your calculator will cut if off after about a dozen digits. In practice a long list of digits after the decimal point only confuses matters. You will find it much easier to see any patterns if you *round* numbers off so that you don't see too much detail. For example, two-sevenths rounded off to the nearest per cent (0 decimal places) is 29%, since 28.57% is a bit closer to 29% than to 28%. Don't forget to do this: it makes a big difference to how easy it is to spot patterns. Excel will round numbers for you (see Appendix A). Where appropriate, I have rounded off the results of calculations in the text of this book, so if you think I've got something slightly wrong, it may just be the rounding.

▶ 1.6 Suggestions for studying statistics and using this book

The non-mathematical approach adopted in this book means that you don't need to be a proficient mathematician to follow the argument. However, this is not the same as saying that you don't need to think. Many of the arguments are subtle and you will need to concentrate to follow them. To help

you, I have included a selection of exercises at the end of most chapters to give you the opportunity to practise using statistical ideas. I have also used a bold question mark at the end of a paragraph as an invitation for you to consider your answer to the question posed before reading mine. So, whenever a paragraph ends in a bold **?**, pause to think of your own answer. You may have no idea, or you may just have a rough idea, but you are more likely to understand the answer having asked yourself the question. Sometimes I will ask something so easy that it may seem patronising (for example working out the average of a small group of numbers); the point of this is to make sure you realise how easy it is. So, when reading this book, please concentrate hard on all the details, pause to ask yourself the questions posed at the end of paragraphs ending with a bold question mark, and make sure that you have a go at the exercises. Are you likely to follow this advice**?**

When I started drafting this book, my initial assumption was that you would follow this advice and read and consider every sentence, number, table and diagram, and do all the exercises. Then I gave a draft of a few chapters to a couple of friends for their comments. There were no comments on any numbers, tables, diagrams or exercises, which raised my suspicions that none of these had been studied in any detail. Then I thought of my own approach to reading mathematical arguments: I usually skip the details and try and take in the gist of the argument, and then return to the details if I want to see how to 'do' it. To be honest, unless I do return to the details, I usually end up with a slightly vague understanding: I'm not fully confident that I see just how it works, and sometimes I do get it a bit wrong. So my advice would be to find time to look at the details, pause when you see **?**, and have a go at the exercises. This advice, however, glosses over the fact that things can be understood in different ways. And some ways are more useful than others.

1.6.1 What does it mean to understand statistics?

In Section 1.4 we saw how to solve the 'birthday' problem by simulation. I explained how the probability could be worked out by simulating lots of audiences on a computer. Alternatively, I could have told you that it is possible to use a 'formula', which is equivalent to this rule: multiply 364/365 by 363/365 by 362/365 and so on all the way to 316/365, and then subtract the answer from 1. This will give you 97%, which is the probability that at least two people in an audience of 50 share a birthday. Two different ways of getting to much the same answer. Which do you prefer**?**

If you could see *why* the rule works, my guess would be that you prefer the rule. It's (probably) easier to do than the simulation, and the fact that you know why it works should mean that you have confidence that it's right,

and you should be able to adjust it for different circumstances (for example audiences of 20). You'll know exactly what the answer means, and the assumptions on which it's based.

But, it's more likely (I suspect) that you don't see why the rule works. It's just an arbitrary set of instructions: you don't know what the rationale is, so you can't adjust it, or check the assumptions on which it's based. It's a pretty useless form of understanding. In these circumstances, I hope you would prefer the simulation, because you should be able to understand *how* this works, so it should be obvious *why* it gives you the answer. This is based on the conviction that it's important to understand as much as possible of how statistical methods work and how conclusions should be interpreted. With a computer, it's often easy to get the answer, but without a deeper appreciation of how and why methods work, it is very easy to misinterpret what's going on, and very difficult to spot any snags or adapt the method to a new situation. The Aldermaston example in Section 1.2 should have alerted you to just how slippery some statistical ideas can be. The non-mathematical methods in this book are designed to be understood in this deeper sense. You should be able appreciate the rationale behind the method, as well as just seeing how to do it.

Despite this, you might feel that the simulation method is, in some sense, cheating. You just build a model of lots of audiences and look at the pattern. There is some truth in this view. If you stick to simulation, you will miss some of the insights provided by the mathematical theory of probability. But using the rule above, without any of the background understanding, would be no better. You need to understand why the rule works and where it comes from in order to get these insights. In this particular problem, this is not too difficult (you may be able to work it out after reading Section 5.2), but for more advanced problems, it would be out of reach for most people. On the other hand, the simulation method, and other similar methods, will get you answers in ways which are transparent enough to see what's going on. And they are often more flexible than most formula-based methods. That's why I'm focusing on them in this book.

1.6.2 Why are you reading this book? Advice for different readers

You may be reading this book because you have a general interest in statistics. You may want to know what statistics can offer, what the hazards are and so on, but have no specific purpose in mind. In this case the advice above applies, and that's about it. However, you may have a more specific reason for studying statistics. Does this apply to you**?**

There are three possibilities I can envisage. First, you may be enroled on a course in statistics. The Similar concepts section in each chapter should be helpful for relating the non-mathematical methods in this book to other methods covered in your course. Use the index. The second possibility is that

you may need to use statistics for a specific purpose. Perhaps you are doing a research project that requires some statistical analysis? Your interest in statistics is as a 'producer' of statistical information. The third possibility is that you may need to interpret research reports which include statistical jargon. In this case your interest is as a 'consumer' of statistics. For example, suppose you are reading a research paper which includes an 'analysis of variance'. If you look up 'analysis of variance' in the index, you will be referred to the Similar concepts section in Chapter 8. This will tell you that analysis of variance is similar to a method explained in Chapter 8, and also relates to some of the ideas in Chapter 9. These chapters should help you understand the basic idea of analysis of variance (although not the detail).

I have tried to cater for producers and consumers in the text and the exercises at the end of each chapter. However, even if you see yourself primarily as a consumer, getting stuck in and doing some analysis yourself may be the best way of getting to grips with what it means. Conversely, even if you see yourself as a producer, interpretation is obviously important. One way in which both producers and consumers of statistics are likely to differ from the general student of statistics is that their interests are likely to be more focused. If you want to do some statistical modelling (Chapter 9), you may want to dive in here and ignore the earlier chapters. As you will see in the next section, I have tried, as far as possible, to help you do this.

1.6.3 The organisation of this book

Like most books, this book is designed to be read in the order in which it appears: Chapter 1 before Chapter 2, Chapter 2 before Chapter 3 and so on. I have, however, tried to keep individual chapters as self-contained as possible, so that readers who are interested in a particular chapter can start by reading that chapter. Dipping into later chapters before reading the preceding chapters is much easier with the approach taken in this book than the more conventional approach. The chapter on confidence intervals (Chapter 7) does not, for example, build on the chapter on probability distributions (part of Chapter 5), as it would in the conventional approach. Obviously, you will miss a few specific points, and some of the background philosophy of the book, by dipping into it like this, but you should be able to see the gist of the argument. (I have included cross-references to earlier sections to help readers who may not have read these sections.)

But which chapter is relevant to your problems? I'll finish this chapter by previewing each of the remaining chapters. You will also find a slightly longer preview at the start of each chapter. (Any terms you don't understand should become clearer after you've read the chapter in question.)

Chapter 2 defines probability via the metaphor of balls in buckets, and the ignoramus who cannot distinguish one ball from the next. Statistics is viewed

as a way of treating life as a lottery. This chapter also looks at the use of samples.

Chapter 3 introduces diagrams and summary statistics: bar charts, histograms, scatter plots, averages, measures of spread and correlation. If you've got some data from a survey, or something similar, this chapter will show the main ways you can break the data down to show patterns and relationships.

Chapter 4 builds on the ideas introduced in Chapters 2 and 3, and analyses the key features of the statistical approach to life, and its strengths and weaknesses. It also gives a very brief review of related and alternative ideas, for example fuzzy logic and the idea of chaos.

Chapter 5 explains how probabilities can be estimated and interpreted by means of thought experiments with bucket models and computer simulation. It also looks at the Poisson and normal distributions.

Chapter 6 considers the problem of using evidence to decide what can be inferred about the actual world. Three approaches are introduced: the bucket and ball 33equivalent of Bayes theorem, null hypothesis testing, and confidence intervals.

Chapter 7 explains confidence intervals in more detail. The idea is to assess, in probabilistic terms, the size of the error when using a sample to guess what the general picture is like.

Chapter 8 explains how probabilities can be calculated to see how plausible a null hypothesis is.

Chapter 9 looks at regression modelling: a way of making predictions from a sample of data, and understanding the relationships between the variables in the data.

Chapter 10 concerns strategies for empirical research: the differences between surveys and experiments, and the practicalities of collecting data. It also (in Section 10.3) makes some suggestions about what to do if you meet some concepts or techniques which you have not seen before, and which are not covered in this book.

The website at www.palgrave.com/studyguides/wood has some data files and interactive Excel spreadsheets which you can download. All the computer files mentioned are on this website. (You will recognise a computer file as a word ending in one of: *.exe .htm .txt .xla .xls*.)