10 How to do it and What Does it Mean? The Design and Interpretation of Investigations

To use statistics you need to get some data. This final chapter looks at some of the different ways of acquiring data about the world – surveys, experiments and so on – and at a few of the practicalities of designing investigations and analysing data. We finish with some tentative suggestions for the situation in which, despite this book, you're stuck: you aren't sure which method to use, or you don't understand the results of an unfamiliar technique.

▶ 10.1 The logic of empirical research: surveys, experiments and so on

Your underlying motive for studying statistics is likely to be that you want to find out about, or make better sense of, the world. You might want a better understanding of how the world works, or you might want a prediction about a specific situation, perhaps to back up a decision to do one thing rather than another. To do any of this, you will need some empirical data. Then you can analyse this in the various ways we've met in earlier chapters: by means of graphical or numerical summaries (Chapter 3), by means of probability models (Chapter 5) or regression models (Chapter 9), or by using the information to derive confidence intervals (Chapter 7), test hypotheses (Chapter 8) or assess the probability of these hypotheses being valid (Section 6.4). But you need to start by getting the data.

How can we get information about the world? There are lots of possibilities and lots of different ways of categorising these possibilities. For my purposes here, I'll distinguish the following three categories.

10.1.1 Common sense and subjective judgements

This is rarely mentioned in lists like this, but it is important. How do you know that the probability of heads is 50%, that all 49 balls are equally likely to be drawn in the lottery, that telepathy is not possible or that a clock on an aeroplane keeps the same time as a clock on the ground? In most cases your answer would probably be along the lines of 'it's obvious that . . .'. We all have a vast store of things we take for granted. We don't bother to research these things; we just take them for granted – life is too short to check everything. However, it is sensible to be cautious. Some of these assumptions occasionally turn out to be wrong. The coin may be biased or the lottery may be fixed. Telepathy may be possible. And Einstein's theory of special relativity avoids the common-sense assumption that clocks on aeroplanes keep the same time as clocks on the ground. Sometimes, it is worth questioning and checking these common-sense assumptions. But in general we just accept them.

As the above examples show, some probability estimates depend on this sort of assumption. The business of statistics, however, is more concerned with situations where we have no obvious answer: we need to collect some data and find out. We need to undertake some empirical research.

10.1.2 Non-interventionist research: surveys and case studies

A survey is simply the process of collecting data about a group of people, organisations, countries, events or whatever the units of analysis are. The data is normally collected from a sample, but the purpose would typically be to find out about a target population, or some other context which goes beyond the sample from which the data was obtained (Sections 2.5 and 3.2). The aim is to collect information without intervening in any way, which may be more difficult than it sounds. You need to bear in mind that the very act of asking people may put ideas their head and alter what they do or what they think. The results may not be as pure as you think.

With any empirical study, there is a trade-off between the size of the sample and the depth in which each case is studied. One issue is how large the sample needs to be for statistical purposes (Section 7.4). You may, however, have decided to research a particular case, or a small number of cases, for the reasons discussed in Section 4.5. There is a continuum from the large-scale study to the detailed study of an individual case.

Regardless of where the research lies on this continuum, it is important to bear in mind the difficulties of using survey data to disentangle cause and effect relationships. A few years ago I did some research which found a negative correlation (Section 3.7.3) between the time students took to do a maths test and their score in the test. Those who took a shorter time tended to get higher scores. Can we jump from this result to the conclusion that it is the quickness of doing the test which is responsible for the higher scores, and so failing students should be encouraged to work more quickly**?**

No, of course not. The students who are more skilled will both score more highly and work more quickly, because they are more skilled. The comparison between those who took a long time to do the test and those who took a short time is not 'fair' because there are likely to be many more good students on the 'short time' side of the comparison.

A similar problem would arise with a study of the relation between spending on health and life expectancy in different countries. The countries spending more on health are likely to be the richer countries, whose citizens live in better conditions and have more to eat, and it may be these factors, not the spending on health, which are responsible for differences in life expectancy (see also Section 3.9).

One response to this difficulty is to try to adjust the results to take account of these extra variables. In Sections 9.4 and 9.7 we saw how multiple regression can be used to separate the effects of several variables, and answer questions about what would happen to one variable if a second is changed while the others remain unchanged. This approach, however, has two problems. First, it can only be used to 'control' for variables on which you have data (Section 9.7). And second, the adjustment depends on a crude model, which is unlikely to be fully realistic (see Section 9.4). A far more powerful way to disentangle cause and effect is to do an experiment.

10.1.3 Interventionist research: experiments and quasi-experiments

The essential feature of this kind of research is that you intervene in the situation and see what happens. There are two main reasons for doing this: to disentangle cause and effect, and to investigate things which would not happen without the experimenter's intervention. Let's look at some examples.

Experiment 1

Suppose you want to find out what happens when the metal sodium is put in water. You could do an experiment: try it and see what happens. You would find it reacts violently with the water to form sodium hydroxide. It would not be possible to find the answer by means of a survey of what happens naturally. Sodium does not occur naturally in its pure metallic form[143] because it is so reactive, so however diligent your search, you would never observe this reaction. As it doesn't happen naturally, you need a contrived experiment to see it happening. Also the result of the experiment is always the same, so you don't need a statistical analysis.

This is a chemistry experiment. Experiments in social sciences generally suffer from the problem that social situations are less easily predicted.

Experiment 2

As part of a research project on problem solving in arithmetic,[144] I wanted to test the hunch that many children were unsuccessful in solving problems because they ploughed in without thinking about what they were doing.

Accordingly, I got a group of children to do a selection of problems under normal (N) conditions without any specific instructions, and then some more problems with the instructions to plan in advance (P), that is, to say how they were going to do the problems before writing anything down or doing any calculations. I wanted to assess the effect of this intervention. The results showed that in a sample of 32 children, 2 were more successful under the N condition, 9 were more successful under the P condition and there was no difference for the other 21 children. (This depends on a clear definition of successful, which need not concern us here.) This experiment obviously has a statistical dimension.

The results seem to indicate that the P condition was more successful than the N condition. Can you see any problems with this conclusion**?**

There are two difficulties I want to focus on here, although you may well have other concerns. First, the difference may be due to chance. The sample does seem small. Another sample and the luck of the draw may give the advantage to the N condition. The *p* value cited (which takes account of the small sample size) was 3%, which indicates that this chance explanation is not very plausible (see Chapter 8 and Exercise 10.5.1). The second problem is that the P condition always followed the N condition. (It was not practicable to reverse the order.) This raises the possibility that the children improved with practice, or as they became used to the experimental setup, in which case the improvement had nothing to do with asking children to plan in advance. This is obviously a serious difficulty with this experiment.

Experiment 3

To get round this problem, and to get some more convincing evidence for the effectiveness of the P procedure, I organised a further experiment. This was on a much larger scale: it involved two classes of children from each of five different schools. In each school, I put all the children from both classes together in a list, and then divided them into a P group and an N group, at random (see Section 2.5). One of the two teachers was asked to teach the P class, and the other the N class, again, at random. All the pupils were then given a 'pre-test' to measure their initial problem-solving ability. The P groups were then taught by the P method, and the N groups by the N method. After two teaching sessions, all the children did a second test, the 'post-test'.

Teachers teaching the P method were asked to teach their pupils to plan their method in advance: they were given specific instructions about how to do this. This was the 'experimental' group. The N method teachers were simply asked to teach problem solving as they saw fit. The 'treatment'[145] I was interested in evaluating was the P method, and I wanted to show that it was better than all the other possible treatments. This meant that the 'comparison' group (often called the 'control' group) should be taught by the best alternative method. In practice, I did not know what this was, so I told the teachers to teach by whatever method they thought appropriate, but without discussing it with the P teachers.

What do you think of the design of this experiment**?** In what ways is it better than Experiment 2**?** What do you think the flaws are**?**

As with Experiment 2, this raises too many issues to give a complete discussion here. I will just focus on a few points, but you may well have thought of others. This experiment is not open to the same objection as Experiment 2. Each group had the same opportunities to practise and get used to the experiment. This is a much fairer comparison.

One of the key features of this experiment is the fact that the children, and the teachers, were assigned to the P or N groups *at random*. Obviously, there are many differences between individual children, and between individual teachers, but allocating them to the two groups at random is a way of ensuring that the N and P groups are broadly similar, provided that the sample sizes are large enough. 'Random assignment' is a way of controlling for a multitude of unknown factors – sometimes called 'noise' variables – and ensuring that the comparison is fair. The results would be much more convincing than for Experiment 2. Experiments like this are the only rigorous way of getting convincing evidence about causal relationships which are obscured by noise variables.

However, the results of this experiment were not as I had hoped. The mean post-test score for the P group was slightly *lower* than for the N group, after adjusting for the pre-test scores, the influence of sex and differences between the schools.[146] The *p* value was 7.5%, not below the formal limit of 5%, but low enough to suggest that the N group might have had a real advantage. The conclusion was that my hypothesis about the superiority of the P group was wrong.

Experiment 4

A vegetarian diet is often said to be healthier than a non-vegetarian diet. This is based on comparing health statistics for vegetarians and non-vegetarians. The difficulty with this, of course, is that there are likely to be other differences between vegetarians and non-vegetarians besides their diet: perhaps other aspects of their lifestyle, affluence, types of job and so on. It is almost impossible to take account of all these variables, so the obvious approach would be an experiment: get a sample of (say) 1000 people, randomly assign half to a vegetarian group, and the rest to a non-vegetarian group, and monitor the health of the two groups. This should show whether vegetarianism has got health advantages, because the randomisation means that the two groups should be roughly balanced on all variables. Is this experiment practicable**?**

No. It is impossible to manipulate people in this way and, if it was possible, it would not be ethical. If the intervention required was less ambitious, say trying out a particular dietary supplement, then the experiment may be possible. There are many questions in medicine, education and business, which could be answered by experiments involving interventions that are impossible or unethical. All we can do is use techniques like multiple regression to make allowances for variables which may interfere with the comparison (Sections 9.4 and 9.7).

We've looked at four very different experiments. The first three are investigating things which would happen rarely, if at all, without the experimenter's intervention. Experiment 4 is to investigate a common occurrence (vegetarianism) to determine its effect on health. Experiment 1 does not need statistics; the last three obviously do. The statistical methods in Chapters 6, 7, 8 and 9 can be used to analyse the results of experiments like the last three.

It is often possible to answer a number of subsidiary questions, for example in Experiment 3 we might want to know whether the results for boys were different from the results for girls. One issue would be to look at the *effect* of the sex variable on success in solving problems (do girls do better than boys or vice versa?), another would be to see if there is an *interaction* (Section 9.8) between the treatment used (N or P) and sex (perhaps boys do better with the P treatment, girls with the N?) – see Chapter 9 (especially Section 9.8) for the statistical methods for this sort of problem.

Experiment 2, a comparison of what happened before and after an intervention, would not be considered a 'proper' experiment by many social scientists. It's only a 'quasi-experiment', because it does not use randomisation to control for noise variables. Such before and after comparisons are often the best we can do in situations, like Experiment 4, where a proper experiment is not possible. We might, for example, monitor the health of people who have changed from being carnivorous to being vegetarian or vice versa.

▶ 10.2 The practicalities of doing empirical research

Some brief comments on the practical side of carrying out surveys and experiments are in order here. The first point is also relevant to assessing the credibility of someone else's research, as the sampling strategy is a serious flaw in many studies.

10.2.1 Think long and hard about how you select any samples

The first step in this thinking should always be to clarify the target population, or the wider context you want to research. This is usually easy, but not always. Take, for example, research to find the relationship between spending on health and life expectancy. Suppose you've got the current data from *all* (or almost all) the countries in the world. What is the target population or the wider context of interest here**?**

At first sight, as you've got data from all the countries, there isn't a sampling problem here. You've got the whole population, so that's it. However, the underlying motivation behind the research is likely to be to see whether increasing spending on health, in any or all the countries, is likely to lead to enhanced life expectancy. In other words, the aim is to research a hypothetical possibility which goes beyond the sample studied. The population metaphor fits uneasily here (what is the population – possible countries?), but if you end up with anything like a confidence interval (Chapter 7), you are presupposing some sort of wider context.

Having sorted out the context to which you want your results to apply, you then need to check that your sample is likely to be representative. The best approach is usually to take a random sample but, as we saw in Section 2.5, there are often problems. The sample may, for example, despite your best efforts, end up biased in the direction of people who are available and willing to answer your questions. And the response rates for many surveys – the proportion of the sample who respond – are often very low. Thirty per cent would be good; less than 10% is common. You can then try bribery (entry into a free draw with a prize) or nagging, but you are unlikely to be fully successful. The question then is whether you can assume that those who don't respond are similar to those who do. The honest answer to this is often 'no', but this is often barely mentioned in many research reports.

Sometimes the sample is a convenience sample: chosen because it is convenient and available, not because it's representative of anything in any systematic sense (the data in Table 3.1 comes from such a sample). Again, you need to think hard about possible biases. Our sample of countries above is, in effect, a convenience sample. We take the countries that exist, as a sample of all the hypothetical countries which might exist. Is it a satisfactory sample? I haven't a clue! But it's all we've got to work with.

10.2.2 Do a pilot study of the data collection, and the analysis, to iron out any difficulties

If you are planning a survey of a thousand people, try it out with, say, six people: get the data, *and analyse it*, then ask your six people for their comments. You should find out whether they can understand the questionnaire, whether you've done anything silly, whether you've got all the information you need, and whether you're collecting anything you don't need. You may be able to use the pilot analysis to assess how large a sample you are likely to want (see Section 7.4).

10.2.3 Coding the data, missing data, outliers and exploring the data

Once you've got the data, you need to key it into your computer. There are some notes on this in Appendix B.1, and two examples of data files on the web, `drink.xls` and `accquest.xls`.

You will need to decide on a coding scheme for each variable, that is, what should be keyed in for each possible answer. The main criterion is convenience: use a coding scheme that sticks in the memory. If you are coding countries, use E for England, F for France, rather than A and B. If you are intending to do any numerical analysis, then it obviously helps to use a numerical code. And if any data is missing, leave the cell blank. It's also a good idea to choose numbers that have a natural interpretation. If you have a five-point scale where one end indicates 'never' and the other 'often', then coding 'never' as 0 and 'often' as 4 means that your results will be far easier to interpret than if you had used (say) 5 for 'never'.

The best coding for yes/no questions is 1 for 'yes' and 0 for 'no'. Similarly, any other category variable which can take one of two values should be coded 1 for one value and 0 for the other. Then the average (mean) represents the proportion of one value, and the variable is coded as a dummy variable in case you should want to bring it into a multiple regression analysis (see Section 9.4 and the variable Sexn in Table 9.1).

If you have a question which asks people to tick all the boxes which apply, you should have one variable (column in your spreadsheet) for each box, and use 1 for a tick (yes). You will need to think about whether unticked boxes represent 'no', in which case they should be coded by 0, or 'no comment', in which case they should be coded by a blank.

Don't forget to check through your data to check for outliers, points outside the pattern of the rest of the data (see Section 3.4.1 for an example). If you find such cases, you then have to decide whether they are credible, or whether they should be removed. You should start the analysis by *exploring* the data, checking it see if it shows the patterns you expected and whether any interesting features catch your eye. SPSS has a procedure specifically for this purpose.[147]

▶ 10.3 Interpreting statistical results and solving problems

In Section 1.6.1 we looked at the different ways in which statistics can be understood. I hope I convinced you of the importance of understanding as much as possible of how a method works, what the answer means and the assumptions on which it rests. Be particularly careful about *p* values, or significance levels (Chapter 8). These are very easy to misunderstand. If you do misinterpret them, you may jump to quite the wrong conclusion.

What should you do if you come across statistical results you do *not* understand? Perhaps they are cited in an article you are reading, perhaps they are part of an SPSS printout for some of your own research, or perhaps you've followed some instructions and worked them out yourself with paper and pencil. What can you do if you feel you don't understand something statistical**?**

There are a range of tactics available:

- *Ignore it.* Printouts from computer packages almost always include more than you need. The regression Tool on Excel gives you the 't statistic', which is not mentioned in Chapter 9. SPSS is much worse: if you tick the boxes on many of the procedures, you will end up with dozens of coefficients with strange names. Some may be useful, most won't be, but unless you want to study mathematical statistics for the next ten years, you have no choice but to ignore most of them. Try to focus on really understanding what seems important and ignore the rest.
- *Look it up.* Try the index of this and other books. The manuals for SPSS may be particularly helpful (even if you aren't using SPSS) because these are aimed at giving a user's understanding, instead of a mathematical understanding. There is also a lot on the web which is easily found with a search engine.
- *Remember the building blocks.* Many statistical methods make use of the key ideas of *p* values (Chapter 8) and 'least squares' models (Chapter 9). The ideas of 'bootstrapping' (Chapter 7) and the 'approximate randomisation test' (Section 8.3) are not so standard, but they may also be helpful in making sense of various concepts. The randomisation test, for example, is roughly equivalent to a one-way ANOVA.
- *Try experimenting with small amounts of 'toy' data.* For example, the Excel worksheet `reg2way.xls` will allow you to experiment to see how regression lines work.[148] SPSS is less interactive, but you can still experiment with small data sets. Try to guess what answer you think SPSS will give you, and then check.
- *If all else fails, ask for help.*

If you're conducting your own research, the various chapters of this book should be helpful at different stages (see Section 1.6.3). Some of the approaches we've looked at, particularly those that depend on computer simulation, are not the standard ones. They are, however, perfectly respectable methods of analysis, backed up by academic literature.[149]

One 'problem' with the computer simulation of probabilities is that the answer may be different each time you run the simulation. Perhaps this is a good thing, because it emphasises the fact that statistics is an uncertain and variable business. However, if you do want stable results, this can always be achieved simply by running the simulation more times. If you do want to use conventional methods, the section on Similar concepts at the end of most chapters should help. You may be able to use the methods in this book as a sort of mental image to help you visualise what's going on.

Despite all this advice, you may, on occasions, get stuck. You may not be able to work out how to proceed. Exercise 10.5.1 below asks you to work out a *p* value, but does not tell you how to do it. You are expected to work this out for yourself: you have a problem to solve. Many of the other exercises in this book may also have been a problem for you. I could have explained how to do Exercise 10.5.1, but then you wouldn't have got any practice in working things out for yourself, which may be useful to you another time. One of the potential advantages of the non-mathematical methods in this book is that they are simpler, which means you should be able to understand them more thoroughly and be in a better position to work things out by yourself. At least, this is my aim.

The key to solving problems is often to see the right approach. For example, the Monty Hall problem (Exercise 5.8.2) is easy if you look at it in the right way. And it's easy to simulate the lottery (Section 5.4) once you've thought of coding a number you have chosen as 1 and the rest as 0. Unfortunately there is no foolproof way of coming up with these insights, except by suggestions such as:

- try to understand your problem in detail
- experiment with the situation and try out various things
- remember methods that worked with earlier problems.

Fortunately, many of the basic ideas in this book – the two bucket model, bootstrapping, the approximate randomisation test, even the idea of coding yes/no as 1/0 – keep on coming up. They are very general ideas, so there's a reasonable chance they will be useful in your new situation. If all else fails, you could also look at the classic work on 'how to solve it'.[150]

▶ 10.4 Similar concepts

There is far too much written on the logic and practice of research to mention here. Check the literature on research methods in your discipline. On the statistical side, there is a sophisticated mathematical theory of the (statistical) 'design of experiments'. The experiments above involve manipulating just one variable; this more sophisticated theory may be of interest if you want to know the effect of several variables, which may interact with each other. One application of this theory is in industrial quality management: the Japanese quality control expert, Taguchi, has popularised the use of fairly complex experiments to help design industrial processes. These are part of the so-called ‘Taguchi methods’.

▶ 10.5 Exercises

10.5.1 Solving a problem: a *p* value for experiment 2

The *p* value for Experiment 2 in Section 10.1.3 was 3% (see Chapter 8 for an explanation of *p* values). This was worked out ignoring the 21 children who did equally well under the P and the N treatments. The null hypothesis was that the each of the 11 remaining children had a 50% chance of doing better under the P treatment, and a 50% chance of doing better under the N treatment. The data showed that 9 of the 11 did better under the P condition. The significance level cited for this was 3%. Can you see how to work this out? (You should be able to use the same method, and the data in Table 3.1 or `drink.xls`, to test the null hypothesis that students, on average, drink the same amount on Saturday as they do on Sunday.)

10.5.2 Investigating the reasons for the riches of the runners

In Section 9.7 four hypotheses are put forward to explain, in causal terms, the tendency on the Isle of Fastmoney for people who can run faster to earn more money (Section 9.1). How would you investigate the truth of each of these hypotheses? Experiments may be worth considering, but would they be possible? Can you manipulate all the variables? And how should the college set about trying to demonstrate that academic qualifications really are useful?