# 1

# Introduction: How to get started with statistics

## WHAT'S IN THIS CHAPTER?

- The misuse of statistics
- Computers and statistics
- How to use this text
- Which chapter do you want?

## INTRODUCTION

This chapter explains why we need statistics and why you should love them. It explains why it is important to understand statistics, which is principally so that we don't get fooled by numbers. It also provides a guide to how this book is best used. We realise that most readers will go to the chapter that best suits their immediate needs and are only reading this if it is the last book in their bag and their train has been indefinitely delayed. If you are this position then we hope it gets better soon.

## STUDYING STATISTICS IS GREAT

This heading is not an indication of madness on the part of the authors. Statistics really is great and it is a remarkable observation that when students finish their statistics courses after

much pain and gnashing of teeth they often come to this conclusion as well. It is the most useful thing you will learn on your degree. Give us a minute (or a couple or paragraphs) and we will attempt to convince you that this statement is not as deranged as it may seem.

## TIP: STATISTICS AND STATISTICS

Rather confusingly, the word 'statistics' means two things. Originally, 'statistics' were numbers. The mean of a sample, for example, is a statistic. However, the study of those statistics gave rise to an academic subject, also called 'statistics'. Hence we can say: 'Statistics are great, I love them' and 'Statistics is great, I love it'. Both sentences are grammatically correct, but have different meanings. The first is talking about numbers, the second is talking about the subject.

We learn about statistics because we want to find stuff out. We want to find out two sorts of things. First, we want to find out what our results tell us, and we can do this by using statistics to analyse data. When we analyse our data, and see what they are telling us, we find stuff out. Sometimes we shed light on a problem, sometimes we don't. Whichever we do, we make a contribution to knowledge, if that knowledge is only 'don't try to do it this way, it won't work'. If we don't do statistical analysis on our data, we will not be able to draw appropriate conclusions. In short, if we don't do statistics, we won't know what works. This text is aimed at illuminating how statistics work and what they tell us.

## TIP: DATA

'Data' is the plural of the singular term 'datum'. You should write 'data are analysed' and 'data have been entered into the computer', not 'data is …' or 'data has been …'. Be sure to point out when your lecturers make this mistake. Lecturers enjoy it when students point out this sort of simple error.

Second, we want to know about the statistics we get from other people. This is most important because we are bombarded with statistical data every day and they are often used to confuse rather than to clarify. There is a famous quote attributed to Andrew Lang: 'He uses statistics like a drunk uses a lamppost – more for support than for illumination.' We need to know when people are trying to illuminate what they have found, and when they are trying to simply support their preformed opinions.

Consider the following extract:

> The number of automatic plant shutdowns (scrams) remained at a median of zero for the second year running, with 61% of plants experiencing no scrams.
>
> *(Nuclear Europe Worldscan,* July/August 1999)

Do you know anything more about nuclear plants after reading that? It is likely that whoever wrote this was using statistics for support rather than for illumination. (Many more examples can be found in *Chance News*, at http://www.dartmouth.edu/~chance/ chance_ news/news.html.)

## THE MISUSE OF STATISTICS

Perhaps the most famous quote about statistics is commonly attributed to British Prime Minister, Benjamin Disraeli,[1] who is reported to have said:

> *There are three kinds of lies: lies, damned lies and statistics.*

Less well known is the comment attributed to another British Prime Minister, Winston Churchill, who said:

> *when I call for statistics about the rate of infant mortality, what I want is proof that fewer babies died when I was Prime Minister than when anyone else was Prime Minister. That is a political statistic.*

It is a popular view that statistics can be made to say anything you want and therefore they are all worthless. While it is clearly true that people will selectively present data to

[1] There's actually a bit of controversy about who *really* said this. Leonard Henry Courtney (1832–1918) wrote it down, in an essay in *The National Review* in 1895: 'After all, facts are facts, and although we may quote one to another with a chuckle the words of the Wise Statesman, "Lies – damned lies – and statistics," still there are some easy figures the simplest must understand, and the astutest cannot wriggle out of.' Mark Twain quoted it in his autobiography: writing: 'the remark attributed to Disraeli would often apply with justice and force: "There are three kinds of lies: lies, damned lies and statistics"'. It seems that Twain thought that Courtney was quoting Disraeli when he wrote 'the Wise Statesman', but Courtney was referring to a hypothetical wise statesman, not a specific one. Rather spoiling the whole quote, it has been suggested that the dashes are parenthetical, and Courtney was trying to say something like 'Lies (damned lies!) and statistics'. Most people haven't heard of Courtney, so they say that it was either Twain or Disraeli who said it. So that's clear then.

misrepresent what is actually happening, it is not true that statistics are therefore worthless. If we have a better understanding of where data come from and how they are being presented then we will not be fooled by the politicians, advertisers, journalists, homeopaths and assorted other charlatans who try to confuse and fool us.

**TIP**

One of the reasons why statistics sometimes appear difficult is that they are often counter–intuitive. Think about your friends, for example: half of them are below average. Or, in a wider context, if you have the view that the average car driver is ignorant and thoughtless, then by definition half of them are *even more* ignorant and thoughtless than that. Then there was the man who drowned crossing a stream with an average depth of 6 inches (attributed to W.I.E. Gates).

## IS STATISTICS HARD AND BORING?

When students find out that they have to learn about statistics as part of their course, they are often somewhat dismayed. They think that statistics is likely to be hard, and is also likely to be boring. In this text we will try and make it not quite so hard and not quite so boring, but you have to be the judge of how successful we are.

We have made this text as clear as we can and as straightforward as we can, but we have not simplified it so much that we skip over important bits. Albert Einstein wrote, 'Everything should be made as simple as possible, but not simpler', and we have tried to follow this principle.

One way to make statistics less hard is to provide a set of clear and explicit instructions, much like a cookbook. For example, if you want to make mashed potatoes, you can follow a instructions like this:

1. Wash and peel potatoes.
2. Cut larger potatoes in half.
3. Put potatoes in saucepan of hot water and boil for 20 minutes.
4. Drain the potatoes.
5. Add milk, salt, butter to saucepan.
6. Mash, with a potato masher, using an up-and-down motion.

This isn't hard. It just involves following a set of rules, and doing what they say. It isn't very interesting, and there is no room for creativity or flexibility. We don't expect you to

understand anything about why you do what you do. We do not try to explain to you anything about the potatoes, or the cooking process, we just expect you to follow the rules. If you had to follow instructions like this every time you made a meal you would find it very dull, however, and would probably just send out for a kebab.

A bigger problem would be that if something went wrong with the cooking, you would be in no state to fix it because you don't what is happening and why. The cookbook approach to statistics might get you to the right answer but you will only have a limited understanding of how you got there. The problem with this is that it is difficult to discuss the quality of your data and the strength of your conclusions. The cookbook approach is not so hard to do, but it doesn't help your understanding.

The approach in this text is to give you the cookbook recipe but also to tell you why it is done this way and what to do in a wide range of circumstances. We hope this allows you to still get to the right result fairly quickly but also to understand how you got there. Staying with the cooking analogy, we will tell you a bit about potatoes and the general process of cooking. 'Too much detail!', you might cry, but you'll thank us for it later.

## TIP

Statistics can be off-putting because of the terms and equations that appear all over the pages like a rash. Don't be put off. The equations are much more straightforward than they look, and if you can do multiplication and subtraction you should be fine. For example, the mean score is commonly written as $\bar{x}$ equation, and once you get used to this and some of the other shorthand then it will become clearer. Imagine you are in a foreign country with a language you can't speak. You don't need to know the whole language, just a few key phrases like 'two beers, please' and 'where's the toilet?'. It is the same with statistics, so just get comfortable with a few key terms and the Land of Statistics will be there for you to explore.

There is another way to deal with statistics, and that is the way that we commonly deal with technology. We open the box, connect everything up and puzzle our way through the various controls. We will only look at the instructions at the point where it either refuses to work or we have broken it. Let's face it, instructions are for wimps! We anticipate that many readers will have adopted this strategy and will be reading this book because their analysis has just gone horribly wrong. It clearly does not help to suggest that this was probably not the best strategy, but all is not lost and the last chapter, with its checklist of important points, will hopefully diagnose your problem and tell you where to go in the text to find the answer.

## COMPUTERS AND STATISTICS

Computers have made statistics much harder.

Well, they haven't really, but they have made *learning* about statistics much harder. And they have done this by making it easier to do hard things.

OK, we know that this is a statistics book, which you were expecting to be a bit tricky, at least in places. And you are reading nonsense like this before you have even got to the statistics, so let us explain. When we were students (and computers needed entire rooms to themselves), learning about statistics primarily involved learning about lots of different formulae. We were presented with formulae and we had to apply them and use them. The majority of the time that people spent doing statistics was spent working through the formulae that were given in books. This wasn't difficult, except that it was difficult to find the time to do it. Some of the statistical techniques that we will cover in this text would take hours or days to carry out. Some techniques were never used, because it was not physically possible to do them. Now we use computers. Computers have made it much easier to find the time to do statistical analysis, because they are much faster. They will do in seconds an analysis that would have taken days in the past.

Our desktop computers can now take all of the long, boring bits away from us. The parts of doing statistics that were boring (they weren't hard, remember, they just involved following a recipe to the letter) aren't there any more. What this means is that there is lots more time to spend on the parts of statistics which are not boring, but which may be a little harder. In the past, we spent a lot of time talking about how statistics were calculated, and considerably less time thinking about what they actually *meant*. Today we can spend much more time thinking about what they *mean*. Spending time thinking about what our analysis means is a good thing, because that is what we are interested in. We are not interested in statistics *per se*, we are interested in what those statistics can tell us.

### The double-edged sword that is SPSS

Throughout this book, we are going to assume that if you are using a computer, it will be running SPSS.

There is a downside to computers in that they allow us to do some very complex tasks without ever understanding what we are doing. If you put your data into SPSS you can click your way happily through the various menus until you appear to get a statistical analysis. The problem is whether you have carried out the *appropriate* analysis. SPSS is a very clever program in that it can carry out some amazing calculations, but it is not clever in terms of understanding what it is doing. It won't suddenly say to you, 'Look, are you sure you want to do a regression on these data?', because it doesn't *know* you are doing a regression and even if it did it wouldn't care. The other problem with SPSS for the happy clicker is that it generates bucketloads of output with numerous test results (Roy's largest

root is our favourite) and you need to have some idea of what you are doing so that you can understand this output.

## HOW TO USE THIS TEXT

This text introduces the basic principles of statistical analysis and works through examples of the most commonly used tests in undergraduate psychology projects. We have attempted to provide the recipe for conducting the tests and also to give the rationale behind the tests (boil the potatoes for 20 minutes, because this makes them soft). We have added some tips and asides to help you through the text and some simple tests so that you can assess your progress.

You don't have to be a mathematician to work through the examples and understand the process of statistical analysis. Although the equations might appear complex, the mathematical principles for calculation are pretty straightforward. As long as you have a calculator with all the standard keys (+, −, ×, ÷) plus $\sqrt{}$ and $x^2$ you will be fine (oh, we do use the ln button once or twice too).

At the end of each chapter we tell you how to carry out the calculations in SPSS. If you understand what you are doing before you tackle SPSS, then SPSS is very straightforward. (If you don't, you will struggle, because you won't know what you want, and you won't know what SPSS is doing.)

The final chapters of the book help you complete your research report by outlining the key features that are required in the write-up and the key issues that you need to deal with in the analysis.

Every chapter tells a story and they can be read in any order. If you have an immediate problem with, for example, regression, then you might go straight to Chapter 8 and work your way through the tests. Though having said that, it might well make most sense to start at the beginning and work your way through to the end. The reason for this is that we look at some general principles of tests in the first few chapters which then keep coming up in the later chapters. Have it your own way, though.

To help you through the chapters we have added some features to break up the story, help you take a breath and ease you over the difficult bits. You will find:

- *Tips*. These will commonly suggest shortcuts or ways to think about the material.
- *Optional extras*. There is always some debate about how much you really need to know. We have tried to write this text on a 'need to know' basis, but there are also loads of other things you might *like* to know. We have therefore put in some optional extras that have, for example, other possible tests, or fascinating (yes, really) pieces of information about statistics or statisticians.
- *Common mistakes*. There are a number of common traps that people fall into with statistics. Many of these have arisen because people have learnt how to use statistics by developing some simple 'rules of thumb' that work *most* of the time, but not *all* of the time. We have tried to point out these common traps so you don't fall into them.

- *Steps*. Where we use statistical tests we have broken them down in steps to make the process clearer and, we hope, to help you carry them out more efficiently.
- *Test yourself*. Practice makes perfect, so we have included a few questions for you to try, and just to be nice we have also included the answers. Some of the questions are a straight test of what you have learnt during the chapter, and some will help you tease out what the test is all about.
- *Using SPSS*. At the end of the chapters we have included simple instructions for doing the tests in SPSS, complete with screen dumps. The bluffing student might be tempted to go just to this section, but be careful you are sure what it all means as SPSS can confuse as well as illuminate.
- *Key terms*. We have identified the key terms in each chapter by **emboldening** them and listing them at the beginning of the chapter. Impress your tutors by learning these terms and using them liberally in any conversation with them. Understanding them is optional, of course.
- *Introductions and summaries*. At the beginning and end of each chapter we set the scene and briefly review what we have dealt with.

## WHICH CHAPTER DO YOU WANT?

We like to think that students will buy this book at the beginning of their course, read it all of the way through (possibly making notes), as they do the course, and then they will simply refer back to the parts that they need. If you are like most students, this isn't what you will do. Instead, you will pick up the book in the week before your assignment is due, try to find which part you need and then read and try to understand that part. Of course, that will be harder, so to help you we've put signposts back to other parts that you might need to know about.

In this section, we'll try to help you to understand which part of the book you need to read, based on what you need to know. Find the highest–level question that matches most closely to your question, and then answer the subsequent questions.

### TIP

Read all the questions and see which is the closest match to your question. Don't stop when you get to the first one that matches.

1. I'm looking for a difference between two (or more) groups or treatments.
   (a) I have two or more treatments, and each is applied to a different group of people (i.e. an independent samples, independent groups or between-participants design).

(i) My outcome measure is continuous (see page XXX) and approximately normally distributed (or my sample size is large).
*You need to use an independent samples t-test (page XXX) if you have two groups , or ANOVA if you have more than two (page XXXX).*

(ii) My outcome measure is ordinal (see Chapter XXX) or my distribution is non-normal (and my sample size is not large).
*You need to use a Mann-Whitney U test* (*see page XXX*).

(iii) My outcome is categorical or nominal (see page XXX).
*You need to use a $\chi^2$ (chi-square) test.*

(b) I have two or more treatments, applied to the same people (i.e. a repeated measures or within-participants design).

(i) My outcome measure is continuous (see page XXX) and the differences in the scores are approximately normally distributed (or my sample size is large).
You need to use a repeated measures *t*-test (see page XXX).

(ii) My outcome measure is ordinal (see Chapter XXX) or the differences in the scores are non-normal (and my sample size is not large).
*You need to use a Wilcoxon test (see page XXX).*

(iii) My outcome is categorical or nominal (see page XXX). If you have two possible outcomes, you can use a sign test. .

2. I'm looking for a relationship between two measures.

(a) Both of my measures are continuous and approximately normally distributed.

(i) I want to know what value of an outcome variable I would expect, given a particular value on a predictor variable.
*You need to use regression (see page XXX).*

(ii) I want to know the strength of the (linear) relationship between my two measures.
*You need to use a Pearson (product moment) correlation (see page XXX).*

(b) One or both of my measures is either ordinal (see page XXX) or is highly non-normal.

(i) I want to know the strength of the (linear) relationship between my two measures.
*You need to use a Spearman (rank) correlation (see page XXX).*

(c) At least one of my measures is categorical.
*This is the same as looking for a difference between groups. Go to Question 1.*

3. I'm looking to see if two (or more) measures of the same thing agree.

(a) I've got a multiple item questionnaire, and I want to see if the items seem to be measuring the same thing.
*You need coefficient alpha,* (*see page XXXX).*

(b) I've got two measures of the same thing, and I want to see if they are giving the same score.

(i) My measures are continuous.
*You need to use the limits of agreement measure (see page XXX).*

(ii) My measures are categorical.
*You need to use Cohen's kappa (see page XXX).*

(iii) My measures are ordinal.
*This is a difficult one. If there aren't many possibly values, you could use kappa.*

If your question isn't here, there are three possibilities:

1. We don't cover that technique, because it is too advanced for this book.
2. No technique exists that can answer your question.
3. Your question isn't analysable. (That's horribly easy to do.)

## WEBSITE

You can find further advice in the blog relating to the book, which you'll find at http://www.jeremymiles.co.uk/learningstats. If you still struggle, then send us a question, and if it's clear and interesting, we'll try to answer it (in the same place). We can't promise anything though.

## SUMMARY

Statistics great! Statistics are fun! They are interesting and not so difficult as you might think. They are also an essential component of almost information source, so if you know how they work you are ahead of the game. Enjoy.