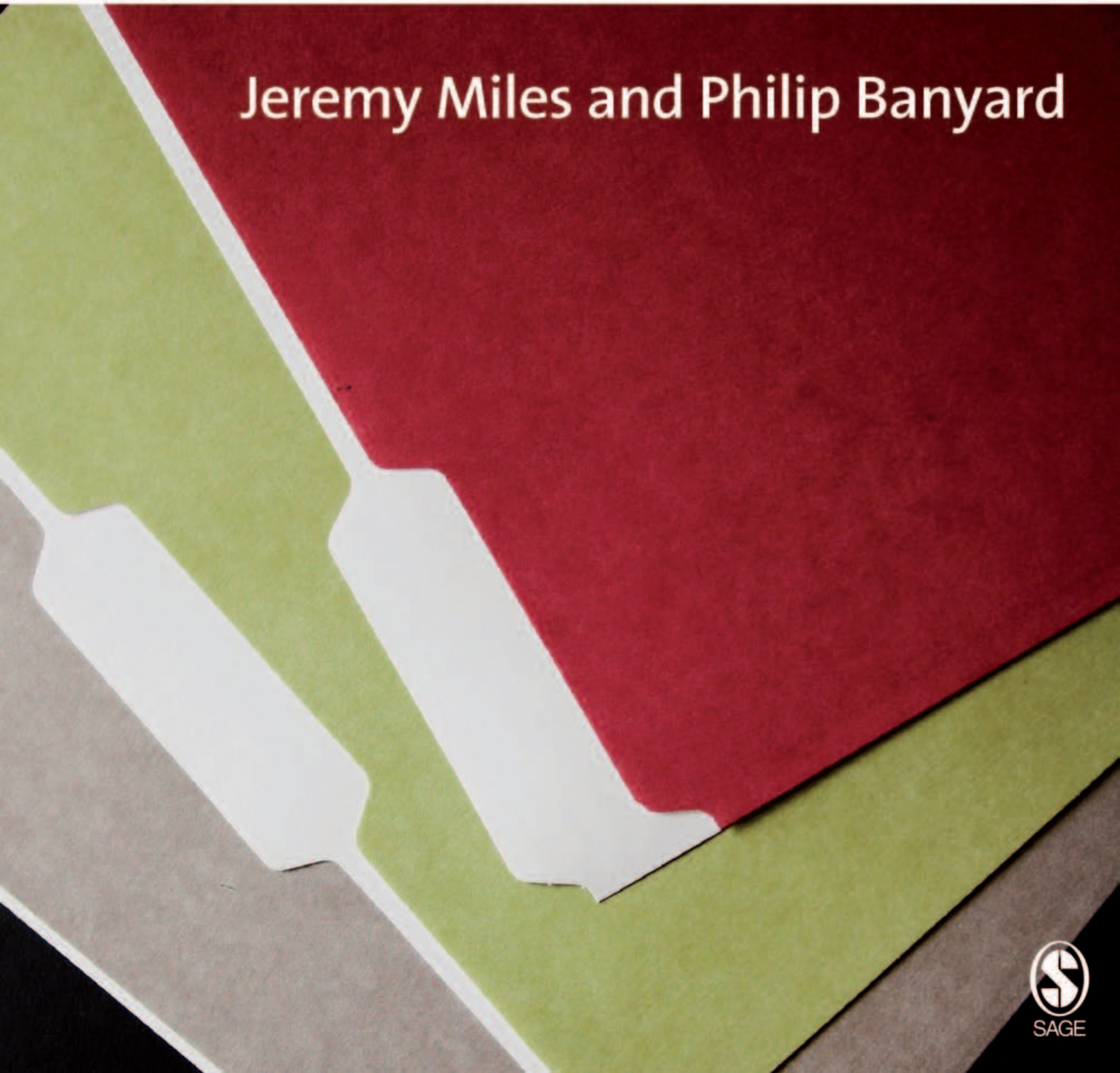
Understanding and Using Statistics in Psychology

A Practical Introduction

Jeremy Miles and Philip Banyard



SAGE

Understanding and Using Statistics in Psychology

Understanding and Using Statistics in Psychology

A Practical Introduction

or, how I came to know and love the standard error

Jeremy Miles and Philip Banyard

SAGE Publications
Los Angeles ▪ London ▪ New Delhi ▪ Singapore

First published 2007

SAGE Publications Ltd
1 Oliver's Yard
55 City Road
London EC1Y 1SP

SAGE Publications Inc.
2455 Teller Road
Thousand Oaks, California 91320

SAGE Publications India Pvt Ltd
B 1/I 1 Mohan Cooperative Industrial Area
Mathura Road, New Delhi 110 044
India

SAGE Publications Asia-Pacific Pte Ltd
33 Pekin Street #02-01
Far East Square
Singapore 048763

Library of Congress Control Number: 2006929861

British Library Cataloguing in Publication data

A catalogue record for this book is available from the British Library

ISBN 978-0-7619-4396-9
ISBN 978-0-7619-4397-6 (pbk)

Typeset by C&M Digitals (P) Ltd, Chennai, India
Printed in Great Britain by The Alden Press, Witney
Printed on paper from sustainable resources

Contents

Acknowledgements

We would like to acknowledge the support of our colleagues in the Psychology Division at Nottingham Trent University and the Health Sciences Department at the University of York who provide a positive and friendly working environment. We would also like to thank the many people who commented on the drafts and so helped in the development of this text. If we are successful in selling the film rights we promise to give them all a walk-on part. We would also like to acknowledge the support and indulgence of Michael Carmichael and Claire Reeve at Sage (and we promised not to mention the thumbscrews). They can be in the film as well if they like.

We would like to give name checks (in no particular order) to Nick Neave, (University of Northumbria at Newcastle), Peter White (Southampton University), Keith Bender (Royal Perth Hospital), Linden Ball (University of Lancaster), James Roney (University of California, Santa Barbara), Barry Wright (Selby and York Primary Care Trust), Charles Hulme (University of York), Mark Torrance (Staffordshire University) and Martin Dorahy (Queen's University, Belfast) for allowing us to use and abuse their painstakingly collected data.

Jeremy would like to thank Susanne, Alex and Daniel for allowing him to spend hours at a computer writing this text and generally ignoring them. Phil would like to thank Mark Shevlin whose reluctance to tear himself away from the bookies and collaborate with Jeremy gave him the opportunity to be the support act on this text. Phil would also like to give name checks to the darts players of the Gladstone in Carrington, the loyal supporters of Nottingham Forest and to Jeremy for writing most of this and not laughing at him when he didn't understand something.

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, even 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 set of 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 know 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 $\bar{x}$, 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 were the size of an Eddie Stobart truck), 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 $\boxed{\sqrt{}}$ and $\boxed{x^2}$ you will be fine (oh, we do use the $\boxed{\ln}$ 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 13) and approximately normally distributed (or my sample size is large).
      You need to use an independent samples t*-test (page 137) if you have two groups, or ANOVA if you have more than two (page 238).*
      (ii) My outcome measure is ordinal (see page 13) or my distribution is non-normal (and my sample size is not large).
      You need to use a Mann-Whitney U test (see page 155).
      (iii) My outcome is categorical or nominal (see pages 13 and 170–71).
      You need to use a χ^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 13) 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 113).

(ii) My outcome measure is ordinal (see Chapter 2) 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 120).

(iii) My outcome is categorical or nominal (see page 13). 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 197).

      (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 214).

   (b) One or both of my measures is either ordinal (see page 13) 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 221).

   (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 281).

   (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 274).

      (ii) My measures are categorical.
      You need to use Cohen's kappa (see page 286).

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

2

Descriptive statistics

What's in this chapter?

- Levels of measurement
- The normal distribution
- Measures of dispersion
- Measures of central tendency
- Graphical representations
- Using SPSS

KEY TERMS

binary measures
boxplot
categorical measures
ceiling effect
central tendency
continuous measures
descriptive statistics
discrete measures
dispersion
distribution
exploratory data analysis
floor effect
frequencies
histogram
inter-quartile range
interval measures
kurtosis
mean
median
mode
nominal measures
normal distribution
ordinal measures
outliers
range
ratio measures
skew
standard deviation
variable

INTRODUCTION

The purpose of descriptive statistical analysis is (you probably won't be surprised to hear) to describe the data that you have. Sometimes people distinguish between **descriptive statistics** and **exploratory data analysis**. Exploratory data analysis helps *you* to understand what is happening in your data, while descriptive statistics help you to explain to *other people* what is happening in your data. While these two are closely related, they are not quite the same thing, and the best way of looking for something is not necessarily the best way of presenting it to others.

Common Mistake

You should bear in mind that descriptive statistics do just what they say they will do – they describe the data that you have. They don't tell you anything about the data that you don't have. For example, if you carry out a study and find that the average number of times students in your study are pecked by ducks is once per year, you cannot conclude that all students are pecked by ducks once per year. This would be going beyond the information that you had.

Optional Extra: *Harry Potter* and the critics

Chance News 10.11 (http://www.dartmouth.edu/~chance/chance_news/recent_news/chance_news_10.11. html) cites an article by Mary Carmichael (*Newsweek,* 26 November 2001 p. 10), entitled 'Harry Potter: What the real critics thought':

> *'Real critics,' of course, refers to the kids who couldn't get enough of the Harry Potter movie, not the professional reviewers who panned it as 'sugary and overstuffed.'*
>
> *The article reports that: 'On average, the fifth graders Newsweek talked to wanted to see it 100,050,593 more times each. (Not counting those who said they'd see it more than 10 times, the average dropped to a reasonable three.)'*

What can you say about the data based on these summaries?

How many children do you think were asked?

What did they say?

Write down your answers now, and when you've finished reading this chapter, compare them with ours (see page 46).

LEVELS OF MEASUREMENT

Before we can begin to describe data, we need to decide what sort of data we have. This seems like a very obvious thing to say, but it is easy to make mistakes. Different sorts of data need to be summarised in different ways.

When we measure something, we are assigning numbers to individuals (where an individual is usually, but not always, a person). A measurement is usually called a **variable**. A variable is anything that can vary (or change) between individuals. There are two main kinds of variables: **categorical measures** and **continuous measures**.

Categorical measures

When we are talking about attributes, we can put each individual in a category. It is an activity that we do in our daily lives. We might categorise a person as mean or funny or male or a 4×4 owner. When we see a bird, we probably categorise it. Some people, such as the authors, will put the bird into one of four or five simple categories (for example, small brown bird, seagull, pigeon, large bird). Bird spotters, however, will categorise a bird in one of a thousand categories ('oh look, there goes a lesser spotted split-toed tufted great bustard').

These data are also called categorical, qualitative or classification variables. They come in three different kinds:

- **Binary**, where there are two possible categories (e.g. female/male, smoker/non-smoker).
- **Nominal**, where there are three or more possible categories, but there is no natural order to the categories. For example, if people are asked where they were born, they can be classified as 'England', 'Scotland', 'Wales', 'N. Ireland', or 'elsewhere'. Even though, for convenience, we may use numbers to refer to these categories, the order does not mean anything. Telephone numbers are another example of nominal categories: just because my phone number is larger than your phone number doesn't make my phone any better than yours, and if you dial my phone number with one digit wrong, you won't find someone similar to me answering the phone.
- **Ordinal**, when the categories have an order. If people are asked to rate their health as 'good', 'fairly good' or 'poor', they fall into one of three categories, but the categories are in an order.

Continuous measures

Continuous measures give you a score for each individual person. They can be classified in two ways: interval or ratio, and continuous or discrete.

Interval versus ratio

Interval measures have the same interval between each score. In other words the difference between 6 and 7 is the same as the difference between 8 and 9 – one unit. So 7 seconds comes 1 second after 6, and 9 seconds comes 1 second after 8. Blindingly obvious, you say, but this does not happen with ordinal measures even when they are presented as numbers. If we imagine the final list of people who completed a marathon, it might be that the people who came 6th and 7th crossed the line almost together and so were only half a second apart, but the people who came 8th and 9th were miles away from each other so crossed the line several minutes apart. On the final order, however, they appear as next to each other and the same distance apart as the 6th and 7th runner.

Ratio measures are a special type of interval measure. They are a true, and meaningful, zero point, whereas interval measures are not. Temperature in Fahrenheit or Celsius is an interval measure, because 0 degrees is an arbitrary point – we could have made anywhere at all zero (in fact, when Celsius devised his original scale, he made the freezing point of water 100 degrees, and boiling point 0 degrees). Zero degrees Celsius does not mean no heat, it just refers to the point we chose to start counting from. On the other hand, temperature on the kelvin scale is a ratio measure, because 0 k is the lowest possible temperature (equivalent to – 273°C, in case you were wondering). However, it is not commonly used. In psychology, ratio data are relatively rare, and we don't care very often about whether data are interval or ratio.

Discrete versus continuous

Continuous measures may (theoretically) take any value. Although people usually give their height to a full number of inches (e.g. 5 feet 10 inches), they could give a very large number of decimal places – say, 5 feet 10.23431287 inches. **Discrete measures** can usually only take whole numbers so cannot be divided any more finely. If we ask how many brothers you have, or how many times you went to the library in the last month, you have to give a whole number as the answer.

Test yourself 1

What level of measurement are the following variables?:

1. Shoe size
2. Height
3. Phone number
4. Degrees Celsius
5. Position in top 40
6. Number of CD sales
7. Cash earned from CD sales
8. Length of headache (minutes)
9. Health rating (1 = Poor, 2 = OK, 3 = Good)
10. Shoe colour (1 = Black, 2 = Brown, 3 = Blue, 4 = Other)
11. Sex (1 = Female, 2 = Male)
12. Number of times pecked by a duck
13. IQ
14. Blood pressure

Answers are given at the end of the chapter.

Luckily for us, in psychology we don't need to distinguish between discrete and continuous measures very often. In fact, as long as the numbers we are talking about are reasonably high, we can safely treat our variables as continuous.

Optional Extra: Continuous measures that might really be ordinal

There is an extra kind of data, that you might encounter, and that is continuous data which do not satisfy the interval assumption. For example, the Satisfaction With Life Scale (Diener, Emmons, Larsen & Griffin, 1985) contains five questions (e.g. 'The conditions of my life are excellent', 'So far, I have gotten [*sic*] the important things from life'), which you answer on a scale from 1 (strongly disagree) to 7 (strongly agree). Each person therefore has a score from 5 to 35. It's not quite continuous, in the strict sense of the word, but it is very close. We treat height as continuous, but people tend to give their height in whole inches, from (say) 5 feet 0 inches to 6 feet 6 inches. This has 30 divisions, the same as our scale.

Should we treat this scale as interval? If we do this, we are saying that the difference between a person who scores 5 and a person who scores 10 (5 points) is the same difference as the difference between a person who scores 30 and one who scores 35 – and by difference, we don't mean five more points, we mean the same amount more quality of life, and we are not sure what that means.

Should we treat this scale as ordinal? If we do, we are saying that a higher score just means a higher score. If one person scores 10, and another scores 20, we can just say that the person who scored 20 scored 'higher'. If a third person scores 21, we can only say that they scored higher still. We cannot say anything about the size of the gap from 10 to 20, and from 20 to 21. We can just say that it is higher. This doesn't seem very sensible either. So what are we to do?

One option is to use sophisticated (and difficult) methods that can deal with ordinal data more appropriately, and treat these data as a special kind of continuous data. These are, frankly, so difficult and frightening that we're not going to even give you a reference (they even frighten us). Anyway, these methods usually can't be used for variables that have more than (about) nine categories.

The solution, used by almost everyone, almost all of the time, is to treat the measures as if they are continuous. It isn't ideal, but it doesn't actually seem to cause any problems.

DESCRIBING DATA

We carry out a study and collect the data. We then want to describe the data that we have collected. The first thing to describe is the distribution of the data, to show the kinds of numbers that we have.

Table 2.1 shows the extraversion scores of 100 students. We could present these data just as they are. This would not be very useful, but it would be very accurate.

Table 2.1 *Extraversion scores of 100 students*

11	11	20	16	14	13	7	26	15	11
17	16	8	24	18	13	25	13	19	17
20	17	22	16	20	10	13	20	19	29
13	14	20	15	25	19	23	17	16	17
20	23	18	10	9	14	24	11	17	17
13	23	14	24	17	14	15	38	14	21
26	15	22	7	14	25	10	15	18	14
16	19	14	18	23	17	15	10	11	20
17	15	25	26	22	26	5	14	17	8
16	12	17	10	15	17	8	20	13	5

For the first time we come across the problem of summarising our data, and presenting our data accurately. Generally we will find that the more accurately we present our data, the less we summarise them, and the more space they take up. For example, in Table 2.1 we have presented our data very accurately but have failed to summarise them at all. We want a way of presenting the data that does not overwhelm the reader who is trying to see what is going on. Table 2.1 would be a very accurate way of describing the data. No one could argue that you were trying to deceive them, or that you have not given sufficient information. The problem is that hardly anyone will be able to read anything useful from that table.

One way to make more sense of the data and summarise them is to present a table of **frequencies**. This is shown in Table 2.2, and already you can start to see some patterns in

Table 2.2 *Frequency scores of information from Table 2.1*

Score	Number	Percentage
5	2	2.0
7	2	2.0
8	3	3.0
9	1	1.0
10	5	5.0
11	5	5.0
12	1	1.0
13	7	7.0
14	10	10.0
15	8	8.0
16	6	6.0
17	13	13.0
18	4	4.0
19	4	4.0
20	8	8.0
21	1	1.0
22	3	3.0
23	4	4.0
24	3	3.0
25	4	4.0
26	4	4.0
29	1	1.0
38	1	1.0

the data. For example, you can see that the high and low scores (extreme scores) have only a few individuals, whereas the middle scores (14, 15, 16, 17) have the most individuals.

You'll notice that the percentage scores are the same as the number of people. This has only happened because we had 100 people in the dataset, and usually the numbers would be different.

Charts

A chart can be a useful way to display data. Look at Figures 2.1, 2.2 and 2.3, and decide which one you think best represents the data.

Figure 2.1 shows a **histogram** with a bin size of 1, which means that there is one score represented in each bar. This chart represents exactly the information that was shown in Table 2.2.

Figure 2.2 shows a histogram with a bin size of 2, which means we have combined two sets of scores into one bar. We can see that a total of two people scored 4 or 5, and two people scored 6 or 7.

Test yourself 2

Before reading on, try to decide which of those two charts is the better.

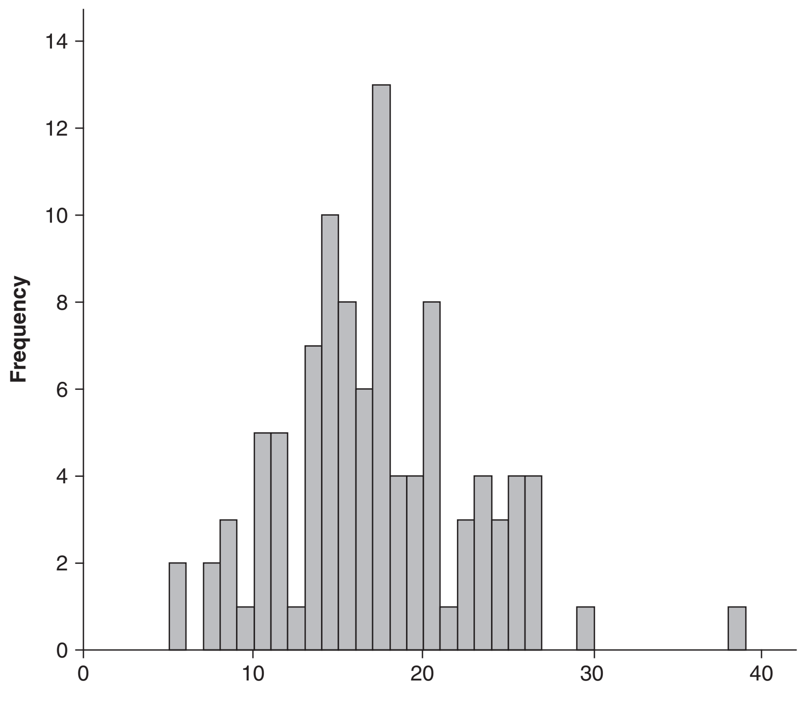


Figure 2.1 Histogram with bin size 1

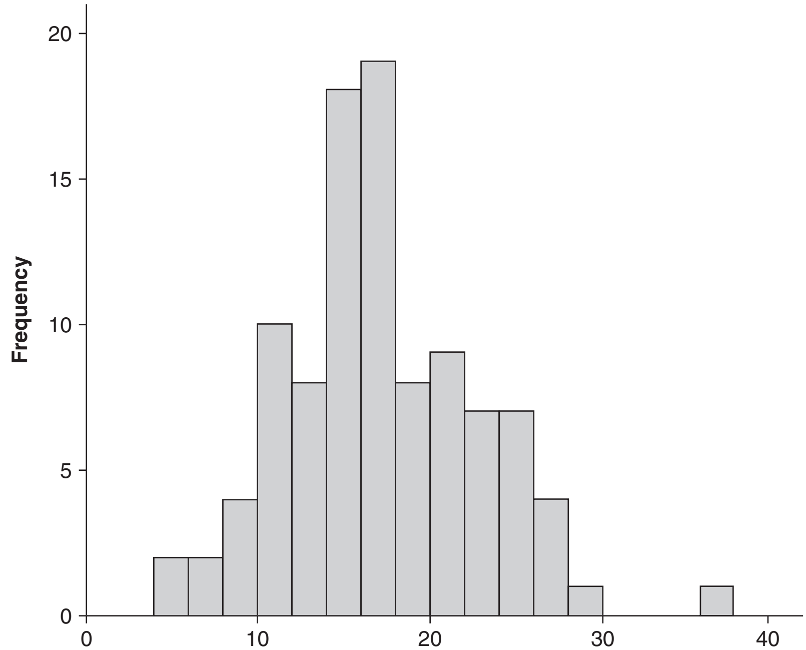


Figure 2.2 Histogram with bin size 2

When we asked you to decide which was better, we raised the issue of summarising our data, and presenting our data accurately. You can't argue with the fact that the first chart is accurate – it contains all of the information in the data. However, if we want to present our data accurately we use a table and present the numbers. A chart is used to present the pattern in our data. Using a bin size of 1 – that is, having each bar represent one point on the scale – leads to a couple of problems. First, when we have a very large number of points, we will have an awful lot of very thin stripes. Second, we are using a graph to show the pattern, and by using small bin sizes we get a very lumpy pattern, so we would rather smooth it a little by using larger bins.

A different way of presenting the data is shown in Figure 2.3. This is a bar chart.

Tip

Statisticians have developed a number of formulae to determine the best number of bins. However, the best thing to do is to draw your histogram, see what it looks like, and then if you don't like it, try a different bin size.

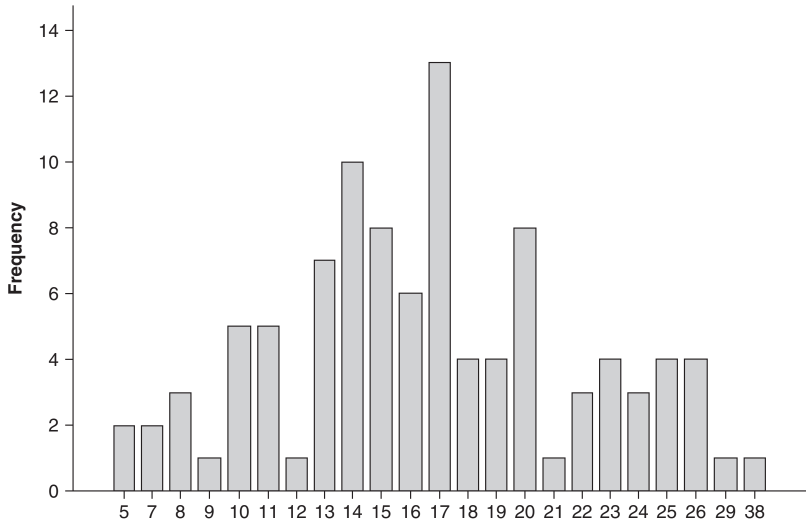


Figure 2.3 Bar chart

Common Mistakes: Draw your distribution in a bar chart

Figure 2.3 shows the same information as in Figure 2.1, except the scale along the *x*-axis (that's the horizontal one) is not spaced equally. We have treated the scale as if it were a categorical variable, not a continuous one. By doing this, we deceive the reader – it appears as if the highest score is only a little higher than the next highest score. This is not the case – it is considerably higher, as can be seen in the other charts.

Don't draw a bar chart for continuous measures.

Histograms and distributions

Histograms are very important in data analysis, because they allow us to examine the shape of the **distribution** of a variable. The shape is a pattern that forms when a histogram is plotted and is known as the distribution (if we are going to be strict about this, the distribution is defined by the formula that leads to that distribution, but thinking about it as a shape is much easier).

THE NORMAL DISTRIBUTION

One of the most commonly observed distributions is the **normal distribution** (also known as the *Gaussian distribution* even though, surprisingly, it wasn't first described by Gauss).

Optional Extra: Stigler's law of eponymy

Stigler's law of eponymy (Stigler, 1980) states that all statistical concepts which are named after someone, are named after someone who did not discover them. Gauss was not the first to mention the normal (or Gaussian) distribution – it was first used by De Moivre in 1733. Gauss first mentioned it in 1809, but claimed to have used it since 1794. It still got named after Gauss though.

The sharper-eyed amongst you will have noticed that for Stigler's law of eponymy to be correct, Stigler should not have first noted it. And, of course, he didn't.

A very large number of naturally occurring variables are normally distributed, and there are good reasons for this to be the case (we'll see why in the next chapter). A large number of statistical tests make the assumption that the data form a normal distribution. The histogram in Figure 2.4 shows a normal distribution.

A normal distribution is symmetrical and bell-shaped. It curves outwards at the top and then inwards nearer the bottom, the tails getting thinner and thinner. Figure 2.4 shows a perfect normal distribution. Your data will never form a perfect normal distribution, but as long as the distribution you have is close to a normal distribution, this probably does not matter too much (we'll be talking about this later on, when it does matter). If the distribution formed by your data is symmetrical, and approximately bell-shaped – that is, thick in the middle and thin at both ends – then you have something close to a normal distribution.

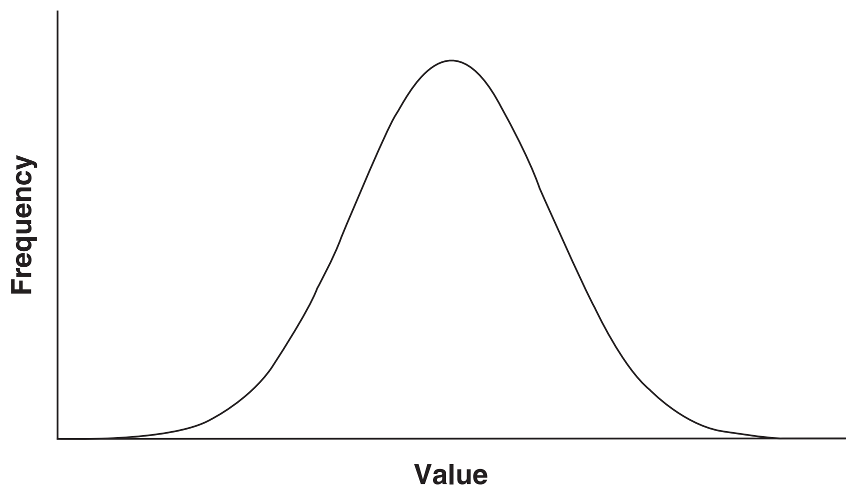


Figure 2.4 Histogram showing the shape of a normal distribution

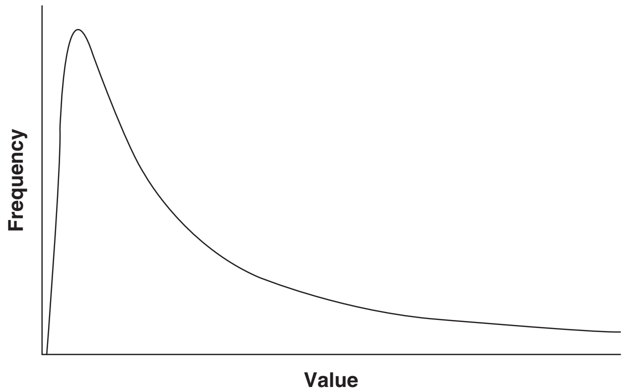


Figure 2.5 Histogram showing positively skewed distribution

Common Mistakes: What's normal?

When we talk about a normal distribution, we are using the word 'normal' as a technical term, with a special meaning. You cannot, therefore, refer to a usual distribution, a regular distribution, a standard distribution or an even distribution.

Throughout this book, we will come across some more examples of seemingly common words that have been requisitioned by statistics, and which you need to be careful with.

DEPARTING FROM NORMALITY

Some distributions are nearly normal but not quite. Look at Figures 2.5 and 2.6. Neither of these distributions is normal, but they are non-normal in quite different ways. Figure 2.5 does not have the characteristic symmetrical bell shape: it is the wrong shape. The second, on the other hand, looks to be approximately the correct shape, but has one or two pesky people on the right-hand side, who do not seem to be fitting in with the rest of the group. We will have a look at these two reasons for non-normality in turn.

Tip: Why does it matter if a distribution is normal or not?

The reason why we try and see the distributions as normal is that we have mathematical equations that can be used to draw a normal distribution. And we can use these equations in statistical tests.

A lot of tests depend on the data being from a normal distribution. That is why statisticians are often delighted to observe a normal distribution.

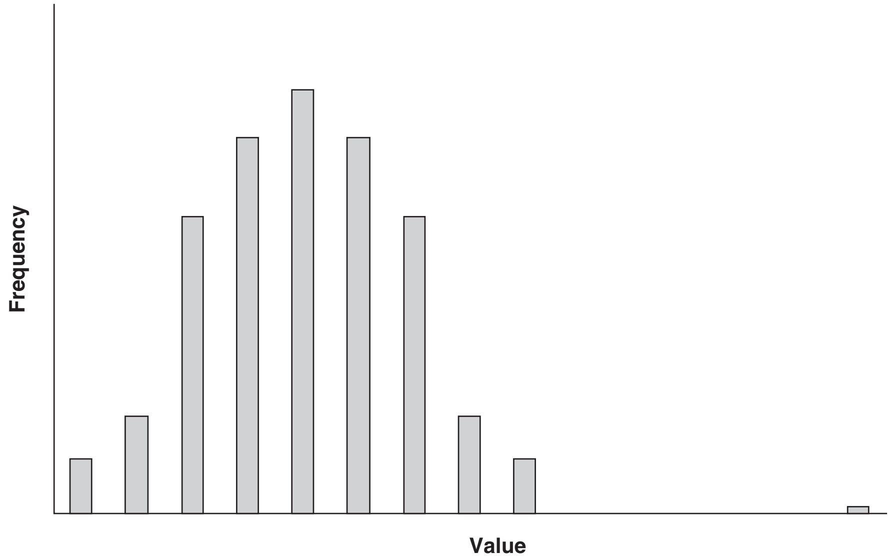


Figure 2.6 Histogram showing normal distribution with an outlier

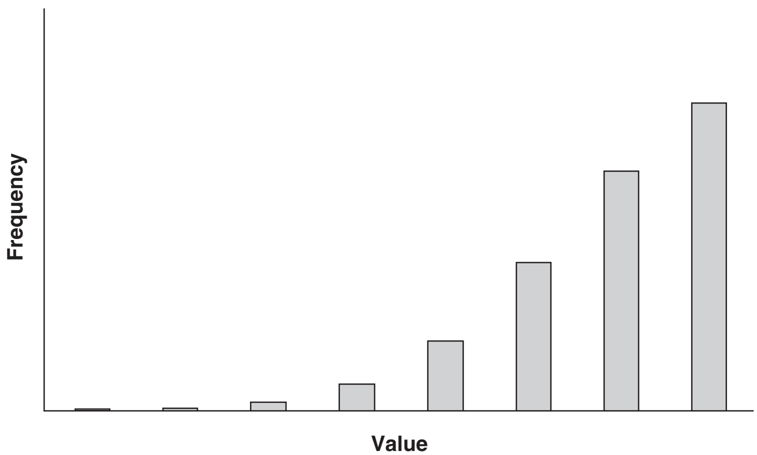


Figure 2.7 Histogram showing negatively skewed distribution

Wrong shape

If a distribution is the wrong shape, it can be the wrong shape for two reasons. First it can be the wrong shape because it is not symmetrical – this is called **skew**. Second it can be the wrong shape because it is not the characteristic bell shape – this is called **kurtosis**.

Skew

A non-symmetrical distribution is said to be *skewed*. Figures 2.5 and 2.7 both show distributions which are non-symmetrical. Figure 2.5 shows *positive* skew: this is where the curve rises rapidly and then drops off slowly. Figure 2.7 shows *negative* skew, where the curve rises slowly and then decreases rapidly. Skew, as we shall see later on, has some serious implications for some types of data analysis.

Tip: Positive and negative skew

Negative skew starts off flat, like a minus sign. Positive skew starts off going up, like part of a plus sign.

Skew often happens because of a **floor effect** or a **ceiling effect**. A floor effect occurs when only few of your subjects are strong enough to get off the floor. If you are interested in measuring how strong people are, you can give them weights to lift up. If your weights are too heavy most of the people will not get the weights off the floor, but some can lift very heavy weights, and you will find that you get a positively skewed distribution, as shown in Figure 2.8. Or if you set a maths test that is too hard then most of the class will

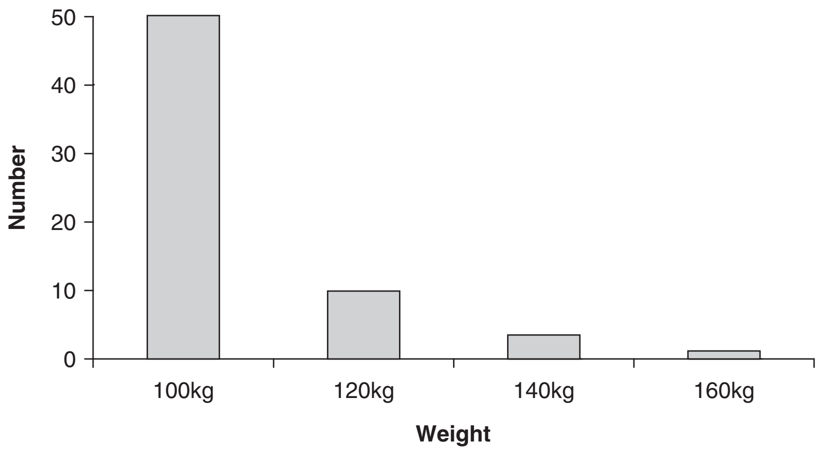


Figure 2.8 Histogram showing how many people lift different weights and illustrating a floor effect, which leads to positive skew

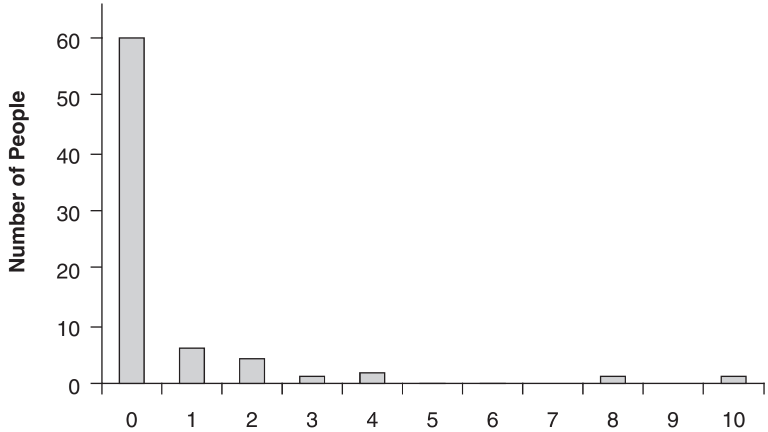


Figure 2.9 Histogram showing the number of times a group of people have been arrested

get zero and you won't find out very much about who was really bad at maths, and who was just OK.

Floor effects are common in many measures in psychology. For example, if we measure the levels of depression in a 'normal' population, we will find that most people are not very depressed, some are a little depressed and a small number are very depressed. The distribution of depression scores would look something like Figure 2.8. Often we want to measure how many times something has happened – and something cannot have happened less frequently than never. If we were interested in criminal behaviour, we could count the number of times a group of people had been arrested (Figure 2.9). Most people have never been arrested, some people have been arrested once (among them one of the authors), fewer have been arrested twice, and so on.

In a similar way, if you were carrying out a study to see how high people could jump, but found that the only room available was one that had a very low ceiling, you would find that how high people could jump will be influenced by them banging their heads on the ceiling. Figure 2.10 shows the distribution that is found in this experiment. We find that most people cannot jump over a hurdle higher than 80 cm, because they bang their heads on the ceiling. A few short people can jump over such a barrier, before they hit their head. The ceiling effect causes negative skew and a lot of headaches.

Ceiling effects are much less common in psychology, although they sometimes occur – most commonly when we are trying to ask questions to measure the range of some variable, and the questions are all too easy, or too low down the scale. For example, if you

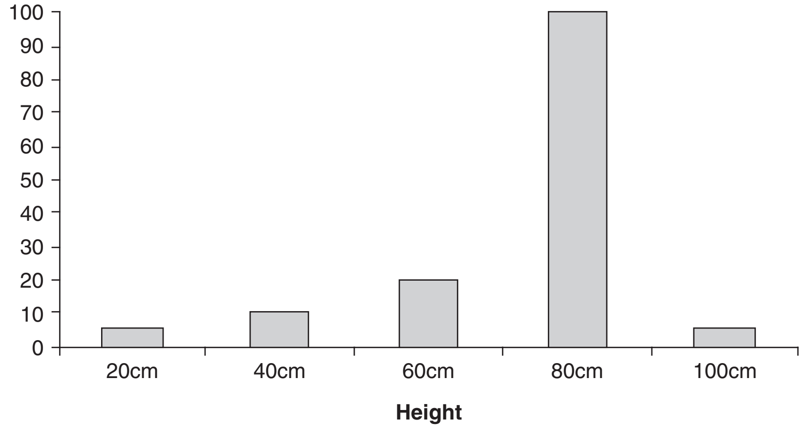


Figure 2.10 Distribution of height of barrier people can jump over, in a room with a very low ceiling. Ceiling effects cause negative skew

wanted to measure the ability of psychology students to do maths, you might give them the following test.

1. $3 + 4$
2. 2×3
3. $7 - 5$
4. $10 / 2$
5. 6×8

Hopefully they would all answer all of the questions correctly (or at least most of the students would get most of the questions right). This causes a ceiling effect, which, as before, causes a headache.

Kurtosis

Kurtosis is much trickier than skew, and luckily for us, it's usually less of a problem. We'll give a very brief overview. If you are really interested in more, see DeCarlo (1997). Kurtosis occurs when there are either too many people at the extremes of the scale, or not enough people at the extremes, and this makes the distribution non-normal. A distribution is said to be *positively kurtosed* when there are insufficient people in the tails (ends) of the scores to make the distributions normal, and *negatively kurtosed* when there are too many people, too far away, in the tails of the distribution (see Figure 2.11).

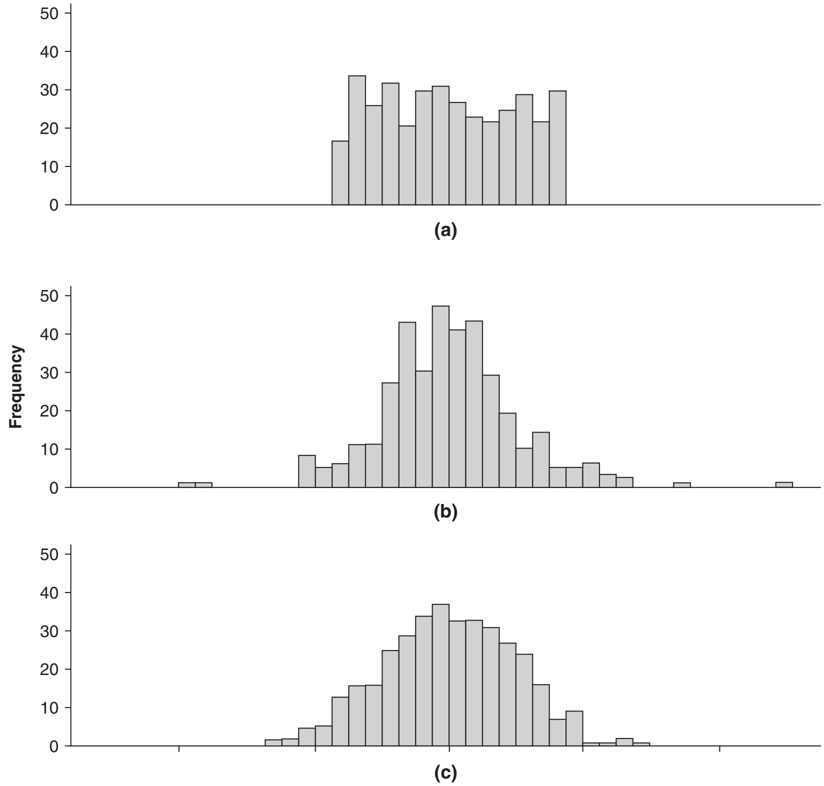


Figure 2.11 Three different distributions randomly sampled from (a) a negatively kurtosed distribution, (b) a positively kurtosed distribution, and (c) a normal distribution

Optional Extra: What's so tricky about kurtosis?

Have a look at the three distributions shown in Figure 2.12. If we told you that the middle distribution was normal, what would you say about the kurtosis of the other two? You might say that the bottom one is positively kurtosed, because there are too few people in the tails. You might say that the top one is negatively kurtosed, because there are too many people in the tails.

You'd be wrong. They are all normally distributed, but they are just spread out differently. When comparing distributions in the terms of kurtosis, it's hard to take into account the different spread, as well as the different shape.

(Continued)

(Continued)

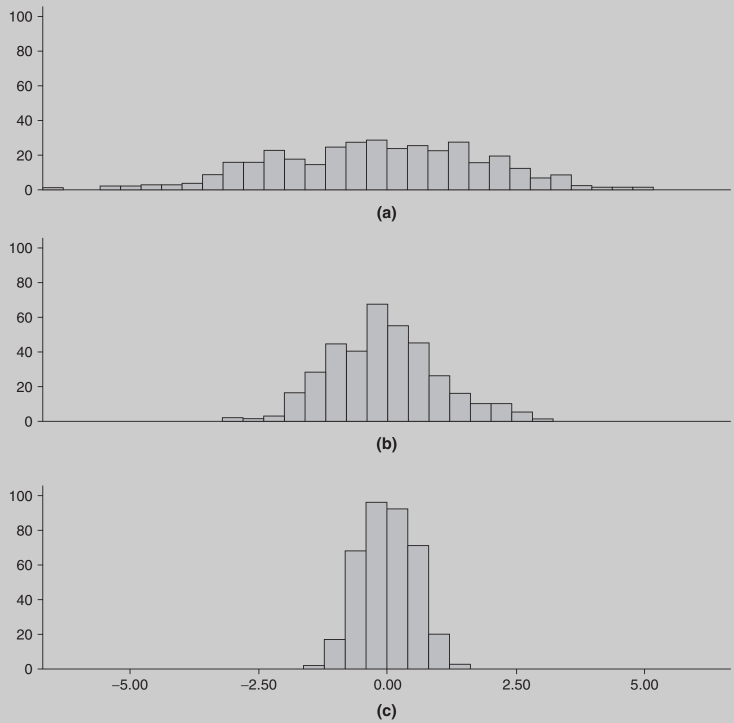


Figure 2.12 Histogram (b) shows a normal distribution, but what about (a) and (c)?

Outliers

Although your distribution is approximately normal, you may find that there are a small number of data points that lie outside the distribution. These are called **outliers**. They are usually easily spotted on a histogram such as that in Figure 2.13. The data seem to be normally distributed, but there is just one awkward person out there on the right-hand side. Outliers are easy to spot but deciding what to do with them can be much trickier. If you have an outlier such as this you should go through some checks before you decide what to do.

First, you should see if you have made an error. The most common cause of outliers is that you are using a computer to analyse your data and you have made an error while entering them. Look for numbers that should not be there. If the maximum score on a test is 10, and someone has scored 66, then you have made a mistake.

If you have checked that you did not make an error, you should now check that any measurement that you took was carried out correctly. Did a piece of equipment malfunction? If you have checked that, you should now decide whether the data point is a ‘real’

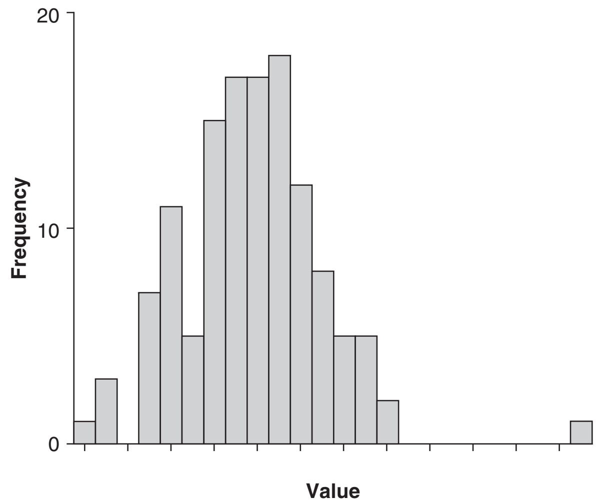


Figure 2.13 Normal distribution with a lonely outlier

data point. If it was a measure of reaction time, did the participant sneeze or yawn? Did they understand the instructions? Is there something unusual about them, which means that they should not have been measured? If any of these are the case you should try to correct the error in some way and then enter the correct value.

If you cannot eliminate any of these errors and are convinced that you have a genuine measurement, then you have a dilemma. Your first option is to eliminate the point and carry on with your analysis. If you do this you will analyse your data well, but you will not analyse *all* of your data well and, frankly, it can look a bit dodgy. If you can keep the data point then it may well have a large effect on your analysis and therefore you will analyse your data badly. Not much of a choice. (Sorry.)

MEASURES OF CENTRAL TENDENCY

Saying **central tendency** is just a posh way of saying 'average'. Average is a tricky word, because it has so many different meanings, so it is usually best to avoid it, and use a more specific word instead.

The mean

The **mean** is what we all think of as the average. Strictly speaking, it is called the *arithmetic mean* because there are other types of mean, though you are very unlikely to come across these.

As you probably know, the mean is calculated by adding up all of the scores and dividing by the number of individual scores. We can write this in equation form as follows: