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 happenin 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/chance_news_10.11.html) cities an article by Marry 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 xxx).

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, they 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 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 do 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 a 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

In general, 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. If you presented your extravation scores as in Table 2.1, no one could argue that you were trying to deceive them, or that you have not given sufficient information. The problem is that no one would be able to read anything useful from the 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.

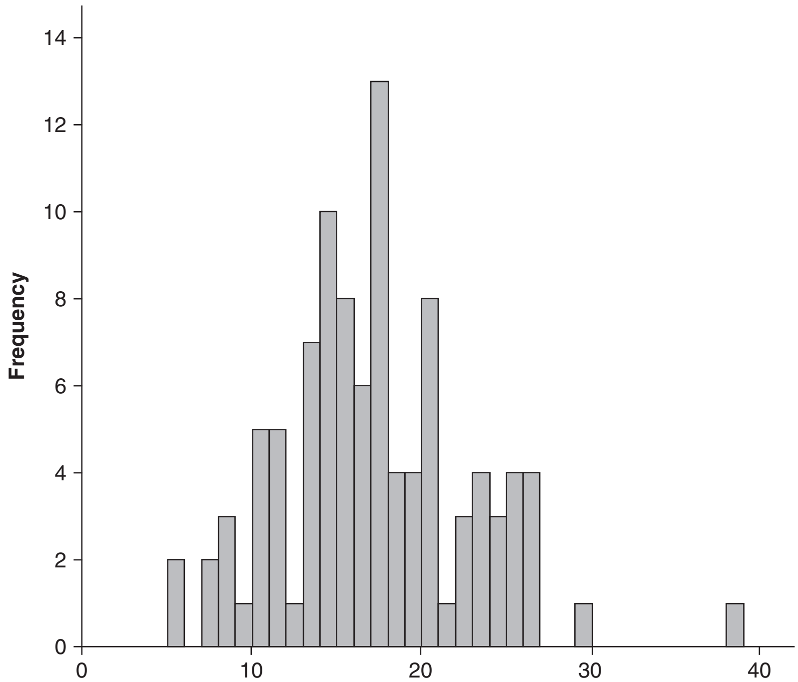


Figure 2.1 Histogram with bin size 1

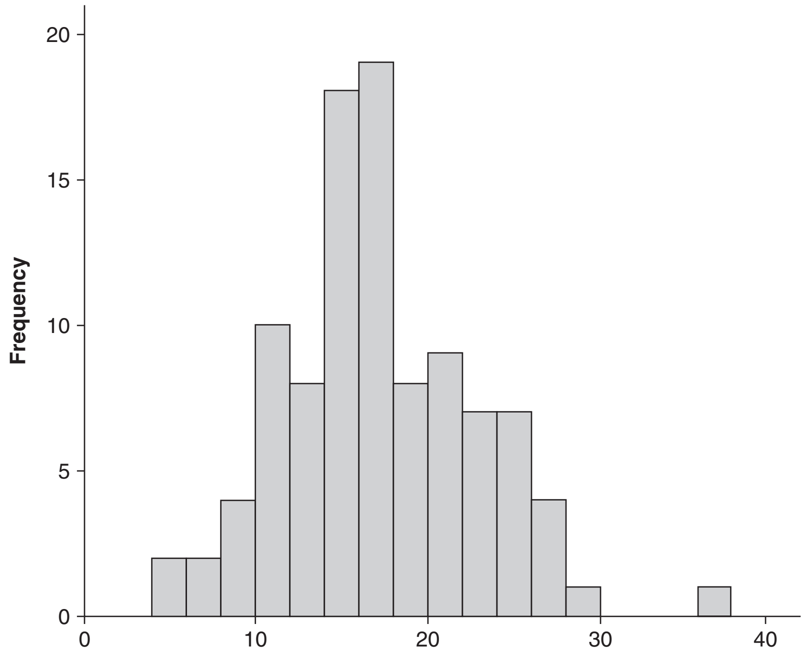


Figure 2.2 Histogram with bin size 1

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.

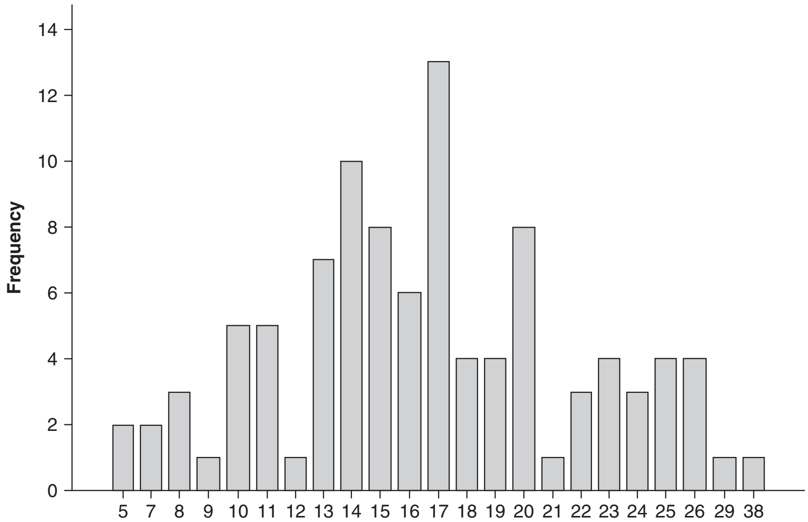


Figure 2.3 Bar chart

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 see 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. Sometimes we choose to draw a histogram by joining together all of the tops of the bars. It is useful to illustrate the shape of the distribution and show that it doesn't matter (for the shape) how many bins you have. It is usually not worth doing this with real data because the curve is too lumpy.

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.

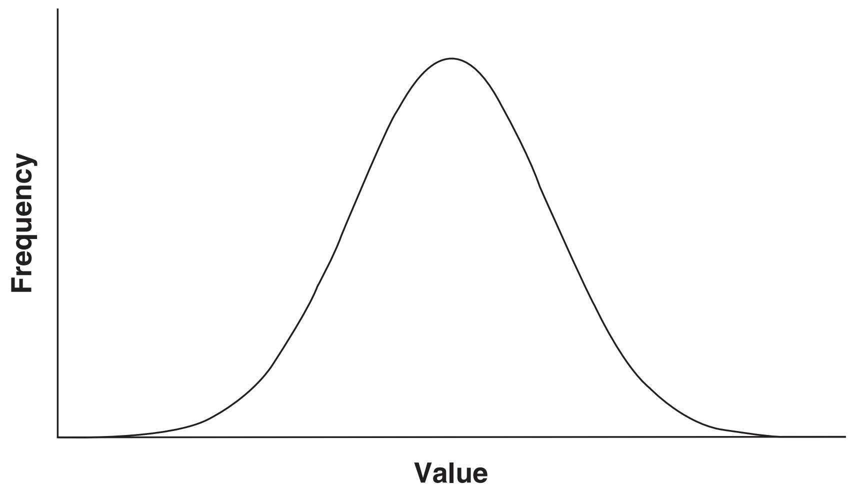


Figure 2.4 Histogram showing the shape of a normal distribution

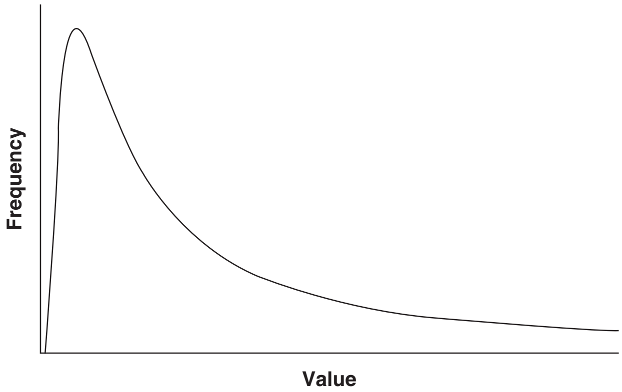


Figure 2.5 Histogram showing positively skewed distribution

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 awkward 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.

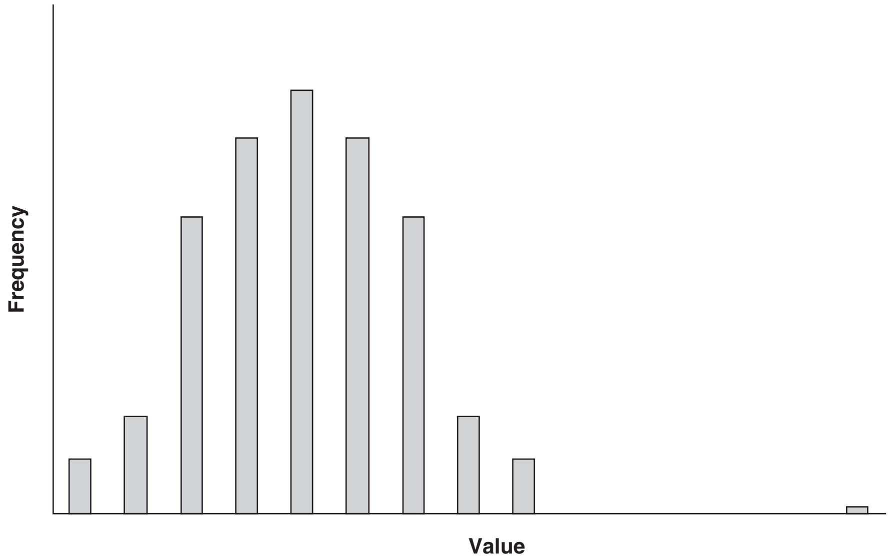


Figure 2.6 Histogram showing normal distribution with an outlier

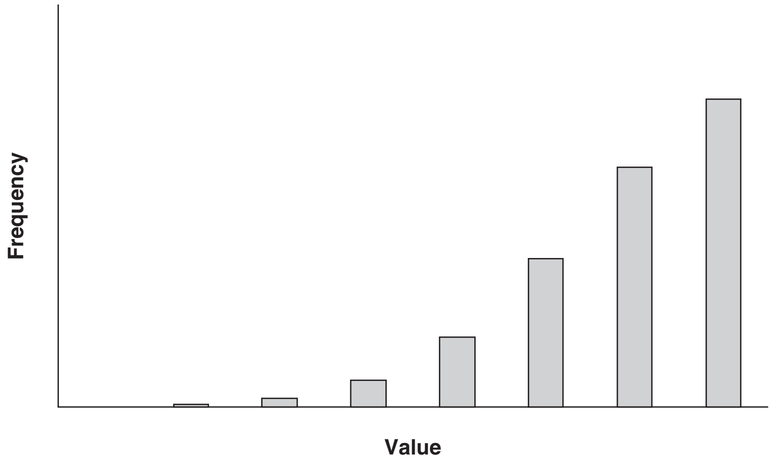


Figure 2.7 Histogram showing negatively skewed distribution

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 Figure 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

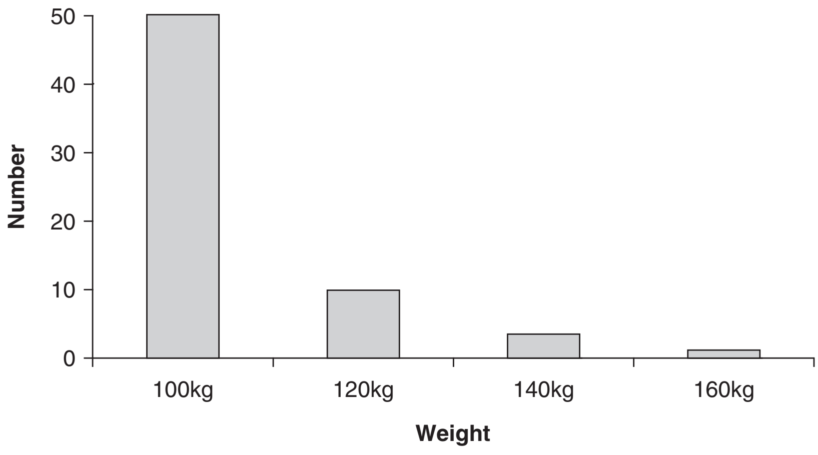


Figure 2.8 Histogram showing how many people lift different weights and illustrating a floor effect, which leads to positive skew

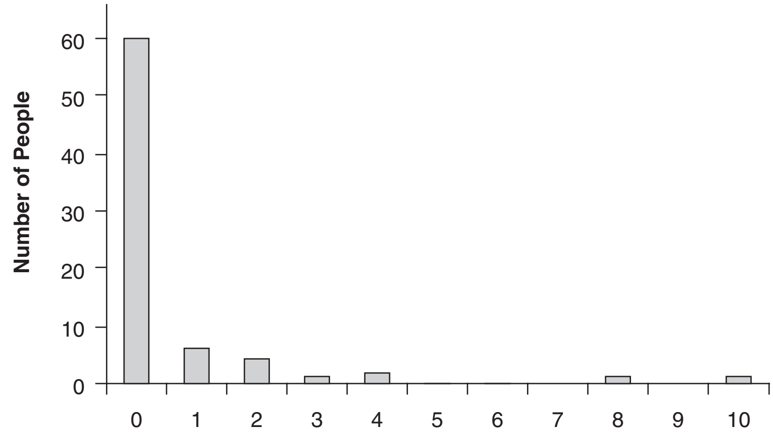


Figure 2.9 Histogram showing the number of times a group of people have been arrested

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

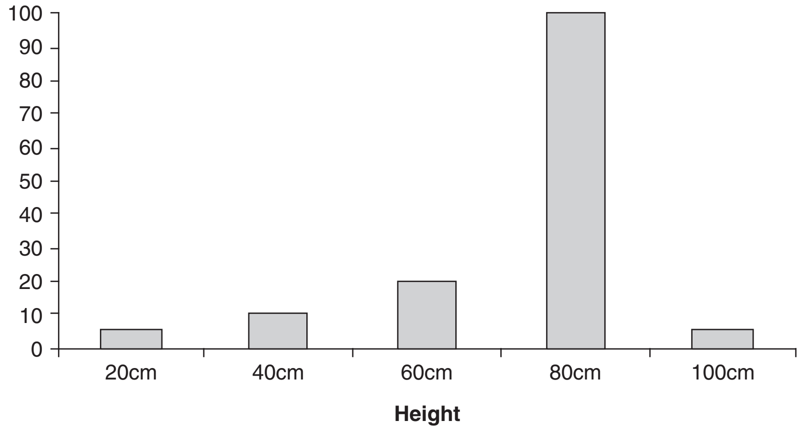


Figure 2.10 Distribution of height of barrier people can jump over, in a room with a very low ceiling. Ceiling effects cause negative skew

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 and 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).

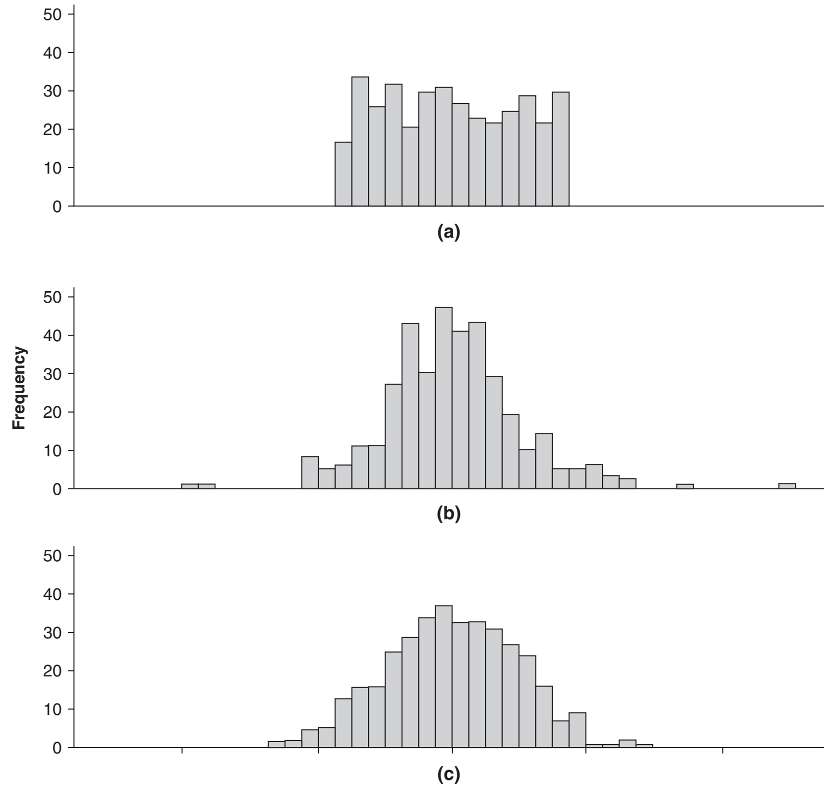


Figure 2.11 Three different distribution randomly sampled from (a) a negatively kurtosed distribution, (b) a positively kurtosed distribution, and (c) a normal distribution

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 many people in the tails. You might say that the top one were negatively kurtosed, because there were 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)

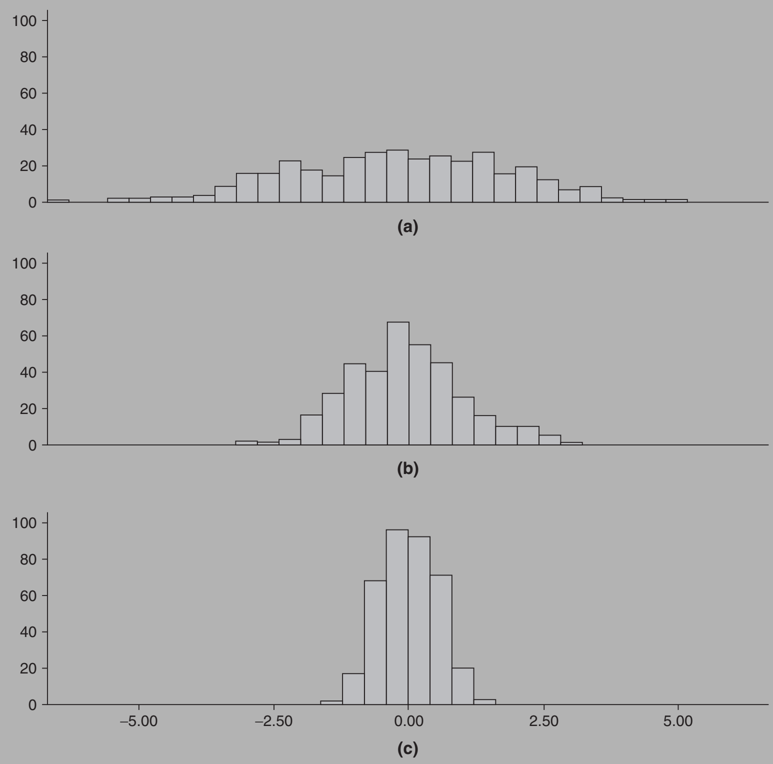


Figure 2.12 Histogram (b) shows a normal distribution, but what about (a) and (c)?

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 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' 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 individuals scores. We can write this in equation form as follows:

$$\bar{x} = \frac{\sum x}{N}$$

Whatever it is we have measured and want to know the mean of, whether it is temperature, IQ, or severity of depression, or the time it took to solve a problem, we will just refer to as x. (We could call it anything we like, such as Derek, but it is easier if we just call it x because that is what statisticians call it.)

The mean of the variable x, is written as $\bar{x}$, and often pronounced 'x-bar.'

On the top of the fraction after the equals sign is the Greak capital letter Σ, or sigma. You will come across this symbol quite often as you study research methods and statistics; it means 'add up' or 'take the sum of'. Σx therefore means 'add up all the scores in x'.

On the bottom of the fraction, we have the letter N. Again we will come across N quite often. N always refers to the number of people in the sample.

The table below shows the variable *x*. We are going to find the mean of *x*.

x	5	6	7	8	9	10

To solve this problem we need to take the equation, and substitute each of the letters and symbols in it, with a number. Then we can work the answer out. We will take this one step at a time. If you think this is easy, that is good, because this is almost the hardest thing we are going to have to do. (And you thought statistics was a difficult subject!)

1. Write down the equation. This is always a good idea as it tells everyone you know what you're doing. $\bar{x} = \frac{\sum x}{N}$
2. Σx means 'the sum of x.' That means add up all of the values in x. We will replace the Σx in the equation, with the sum. $\bar{x} = \frac{5+6+7+8+9+10}{N}$
3. *N* means the number of individuals. We find this by counting the scores, and we find there are 6 of them. $\bar{x} = \frac{5+6+7+8+9+10}{6}$
4. Now we will work out the top row: $5+6+7+8+9+10 = 45$. $\bar{x} = \frac{45}{6}$
5. Now we work out the fraction that we are left with. $\bar{x} = 7.5$

Assumptions

We often need to make assumptions in everyday life. We assume that everyone else will drive on the left-hand side of the road (or the right-hand side, depending on which country you are in). We assume that when we press a light switch, the light will come on. We assume that if we do enough work and understand enough then we will pass our exams. It's a bit disturbing if our assumptions are wrong. If we find that passing exams is not related to the amount of work we do, but that to pass an exam we actually need to have an uncle who works in the exams office, then our assumption is wrong. Similarly, if we assume everyone will drive on the left, but actually some people drive on the right, again our assumptions are wrong. In both these cases, the things that we say will be wrong. 'Work hard, and you'll pass,' is no longer correct. Assumptions in statistics are much the same as this. For our statistics to be correct, we need to make some assumptions. If these assumptions are wrong (statisticians usually say *violated*), then some of the things we say (the results of our statistical analysis) will be wrong.

However, while our assumptions will usually be broadly correct, they will never be *exactly* correct. People sometimes drive on the wrong side of the road (when parking, for example) and we manage not to crash into them (most of the time). In the same way, if our assumptions about data are wrong, but not too wrong, we need to be aware that our statistics will not be perfectly correct, but as long as the assumptions are not violated to any great extent, we will be OK.

When we calculate and interpret the mean, we are required to make two assumptions about our data.

1. The distribution is symmetrical. This means that there is not much skew, and no outliers on one side. (You can still *calculate* the mean if the distribution is not symmetrical, but *interpreting* it will be difficult because the result will give you a misleading value.)
2. The data are measured at the interval or ratio level. We saw earlier that data can be measured at a number of levels. It would not be sensible to calculate the mean colour of shoes. If we observed that half of the people were wearing blue shoes, and half of the people were wearing yellow shoes, it would make no sense to say that the mean (average) shoe colour was green.

COMMON MISTAKES: GARBAGE IN, GARBAGE OUT

When you use a computer program, it will not check that you are asking for sensible statistics. If you ask for the mean gender, the mean town people live in, and the mean shoe colour, it will give them.

Just don't put them into your report.

Median

The second most common measure of central tendency is the **median**. It is the middle score in a set of scores. The median is used when the mean is not valid, which might be because the data are not symmetrically or normally distributed, or because the data are measured at an ordinal level.

To obtain the median the scores should be placed in ascending order of size, from the smallest to the largest score. When there is an odd number of scores in the distribution, halve the number and take the next whole number up. This is the median. For example if there are 29 scores, the median is the 15th score. If there are an even number of scores, the median is the mean of the two middle scores.

We are going to find the median of the variable x below:

x:	1	14	3	7	5	4	3

The first thing to do is to rearrange the scores, in order, from smallest to largest. The rearranged data are presented below.

x:	1	3	3	4	5	7	14

We have seven data points. If we calculate $7 \div 2 = 3.5$ and we therefore take the next number above 3.5, which is 4. The fourth item is the median, and has the value 4.

Mode

The final measure of central tendency, rarely reported in research, is the **mode**. It is the most frequent score in the distribution or the most common observation among a group of scores. The mode is the best measure of central tendency for categorical data (although it's not even very useful for that).

In the dataset that we analysed in the previous example the number 3 was the only number to appear twice, and hence the mode is 3. In a frequency distribution the mode is very easy to see because it is the highest point of the distribution.

The problem with the mode is that it doesn't tell you very much. For example, imagine that we have a group of 20 females and 5 males. What is the mode? The mode is female. But that just tells us that there were more females than males. It doesn't take much more space to say that there were 20 females and 5 males.

Comparison of mean, median and mode

When deciding which of the three measures of central tendency to use, you should take into account the distribution of the scores. When the distribution is unimodal (i.e. has one mode) and symmetrical, then the mode, median and mean will have very similar values.

The mean is often the best average, for a couple of reasons. First, unlike the median, it uses all of the information available. Every number in the dataset has some influence on the mean. The mean also has useful distributional properties (don't worry about what this means, we will cover it later), which the median does not have. The downfall of the mean is that it is affected by skew and by outliers.

Consider the following example. Five psychology students and five biology students were questioned to see how many lectures in research methods and statistics they attended. The results are shown below:

Psychology	Biology
17	20
19	20
19	20
17	20
18	1

Test Yourself 3

Which group of students attended the most lectures?

The mean number of lectures attended by the psychology students was 18. The mean number of lectures attended by the biology students was 16.2. Thus it appears at first glance as if the psychology students attend more lectures. However, when we look more closely at the data, it is clear that there is an outlier who is skewing the data. If you were to think that a psychology student would be likely to attend about 18 lectures, you would be right. If you were to think that a biology student would be likely to attend about 16 lectures, you would be wrong.

If instead we use the median, we find that the median for the psychology students is 18, and the median for the biology students is 20. If you thought that the psychology student would be likely attend 18 lectures, you would be right. If you thought that a biology student would be likely to attend about 20 lectures, you would be right 4 times out of 5. This shows that when the distribution is skewed, the median is a more representative value of central tendency.

When a distribution is skewed, the skewed values have the effect of 'pulling' the mean away from the true value. In a normal (or any symmetrical, unimodal) distribution, the mean, median and mode are all the same. In a skewed distribution, they are not the same.

A survey was carried out in which 100 students were asked how many books they consulted in their statistics module. Table 2.3 shows that 6 students are honest enough to say they never used a book, and that 3 students are making the unbelievable claim that they used 9 books. If we calculate the mode, median and mean, we find that the mode is equal to 1. More students have read 1 book than any other number of books. However, the median is slightly higher, at 2, and the mean is higher again, at 2.88. The histogram in Figure 2.14 shows the frequency plot, with the mode, median and mean marked on it. This separation of the mean, median and mode in the direction of the skew is a consistent effect in a skewed distribution. If the distribution were negatively skewed we would find the effect going in the opposite direction.

Table 2.3 *Frequency table of number of books used by students studying biostatistics*

Number of Books	Frequency
0	6
1	25
2	20
3	18
4	12
5	8
6	4
7	2
8	2
9	3

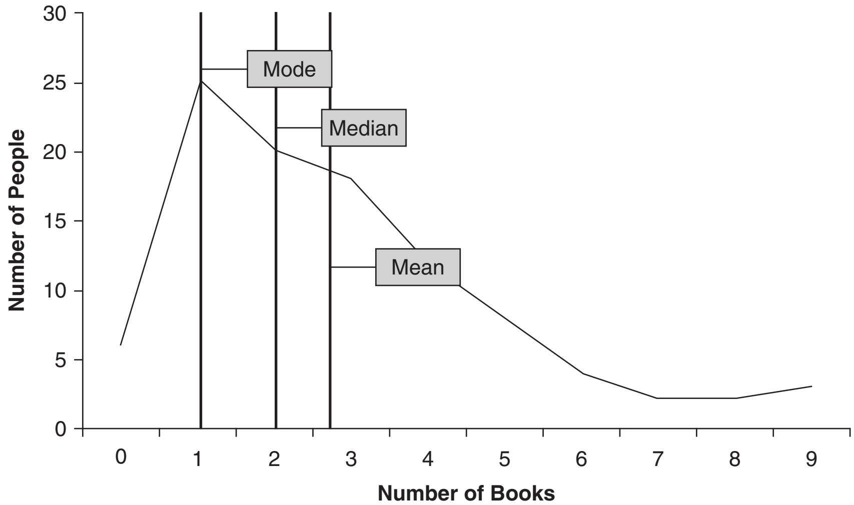


Figure 2.14 Frequency plot of number of books read and number of students. The mode, median and mean are all marked. Note that the median is to the right of the mode, and the mean is to the right of the median. This is to be expected in a skewed distribution

OPTIONAL EXTRA: LIES, DAMNED LIES AND STATISTICS (AGAIN)

One of the most common places to lie in statistics is to selectively choose the mean or median, depending on the argument that you want to present. Politicians are particularly guilty of this, for example when they talk about incomes. The distribution of income is highly skewed – some people don't earn much money, most people earn a moderate amount, and a very small number of people earn far too much. This makes the mean amount of money that people earn rather unrepresentative. If the population of the United Kingdom was 50 million people (it's not, but it makes the sums easier), and one person gets £1 billion (from somewhere) then the mean amount of money that we each have goes up, by £20. So the mean wealth of people in the UK has risen by £20, but only one persons any riches. The rest of us will see no change in our bank accounts.

However, if I am a politician, and I want to argue that you are better off, I might use the mean. An example of this was the 1992 Conservative Party election campaign, in which they said 'You'd pay £1,000 more tax a year under Labour' (Figure 2.15). What they meant was, you'd pay a mean of £ 1,000 more tax.

(Continued)

(Continued)

Figure 2.15 An example of an election campaign using the mean, in an unrepresentative fashion

If the mean is £1,000, this doesn't mean that everyone will pay £1,000. Some people (for example, Richard Branson, the bosses at Sage Publications, Andy Field) will pay a lot more, while others (for example, poverty-stricken textbook authors) may pay no more, but on average, we will pay £1,000 more. A much more representative measure would be the median.

We know that politicians know that the median is the more appropriate measure of income, so when the minimum wage was introduced into the UK, it was based on the median income, not the mean. Child poverty is also defined as being in a family that earns less than 60% of median income.

There are lies, damn lies and statistics, but only if you don't understand statistics. One thing that this book tries to do is let you see through the lies that people (and politicians) try to tell you. Read this book and be empowered.

Ordinal data

When data are measured on an ordinal scale it becomes something of a tricky issue to decide whether to use the mean or the median, or even the mode. Opinions differ between statisticians, so we have to be careful about what we decide to do.

The problem is that there is a very fuzzy line between what could definitely be called ordinal, and what could definitely be called interval. Some statisticians would argue that things like personality measures and attitude scales can only be considered to be ordinal data. The majority would argue that these can be considered to be interval data and therefore it is OK to use the mean. As we've already said (but it was an optional extra, so you probably skipped it) if you consult most journals you will find that most researchers, most of the time, will treat their data as if they are measured on an interval scale. (There is some research to back this up, but you don't really want to be reading papers about this sort of thing, do you?)

TIP: DAZED AND CONFUSED

If you can't decide whether your data are ordinal or interval then it might not be due to your lack of knowledge. Some measures can be seen as either. For example, IQ scores are commonly dealt with *as if* they are interval, but some people argue that the data are really collected on an ordinal scale.

MEASURES OF DISPERSION AND SPREAD

When describing a variable it is necessary to describe the central tendency (the mean, median or mode). However, the central tendency doesn't mean a lot without a measure of **dispersion** or spread. An experiment was conducted to examine the effects of a new antidepressant drug. An experimental group were given a dose of the drug and a control group were given a placebo (which looks like the drug but does not have the active ingredient). Later, both were scored by a psychiatrist on their level of depression. The results are shown in Table 2.4.

Table 2.4 *Mean scores in depression study*

Group	Mean score
Placebo	80
Drug	70

Test Yourself 4

Did the drug have an effect on level of depression? How much of an effect was it?

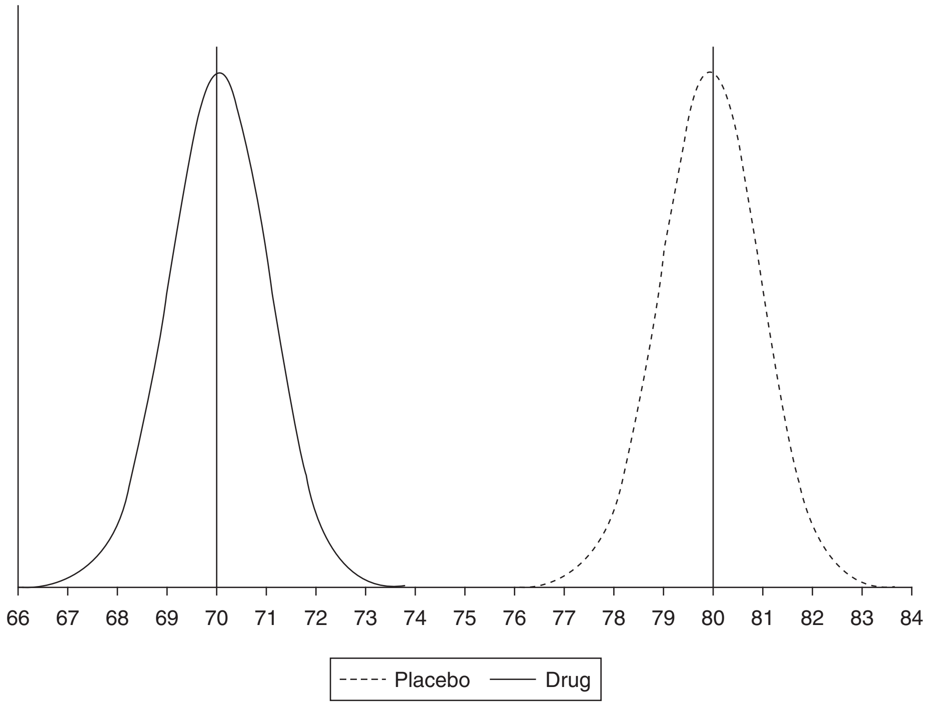


Figure 2.16 Large difference between means

Of course, you cannot give a full answer to the question in the box because we haven't given you any information about the dispersion (spread) of the scores. Have a look at Figures 2.16–2.18. In each of these figures the results are as presented from Table 2.4. The means are the same in each graph. but the graphs look very different because the dispersions of the scores in each one are very different.

When the dispersion is very small, this is a very large difference between the two groups. In Figure 2.16 there is such a small dispersion – if you knew what someone's score was, you would know which group they were in. In Figure 2.17 the difference between the means is the same, but the dispersion is larger, hence the two distributions overlap to some extent, and therefore the effect of the drug is less. In Figure 2.18 the dispersion is large, and so there is a great deal of overlap between the groups, and the drug only has a small effect.

What this shows is that it is very hard to interpret a measure of central tendency without also having a measure of dispersion. In Figure 2.16 means are different and the distributions do not overlap, so there is clearly a big difference between the groups. In Figure 2.17 the distributions overlap but the bulk of the experimental groups scored higher than the control group. In Figure 2.18 the overlap is so great that it would not be possible to predict which group an individual was in from looking at their score. Therefore, if you

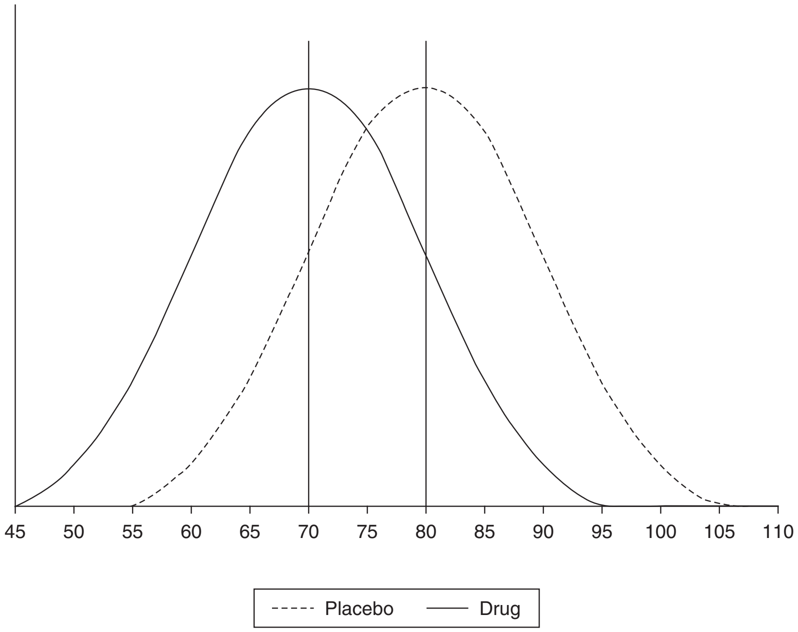


Figure 2.17 Medium difference between means

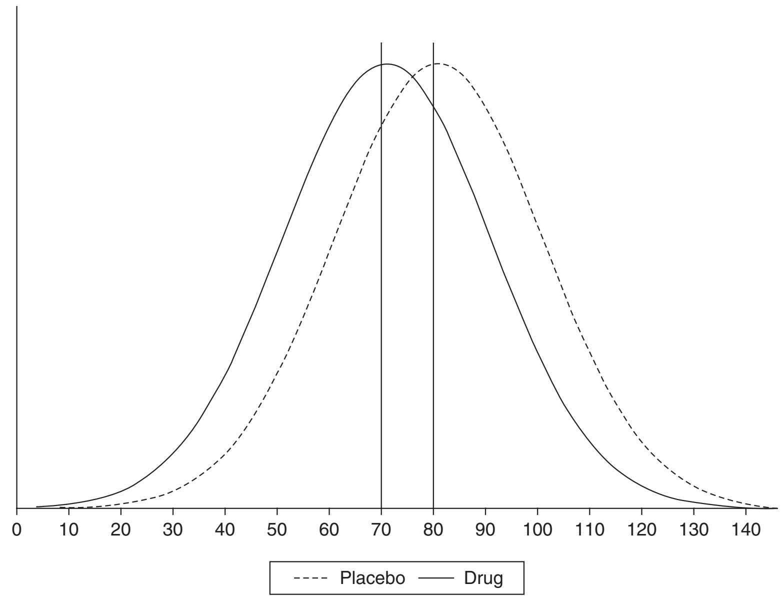


Figure 2.18 Small difference between means

only use the mean score to describe the data the results can be misleading – or, worse, meaningless. We will now look at three different measures of dispersion: the range, the inter-quartile range, and the standard deviation.

Range

The **range** is the simplest measure of dispersion. It is simply the distance between the highest score and the lowest score. It can expressed as a single number, or sometimes it is expressed as the highest and lowest scores. We will find the range of the following variable x:

x:	4	11	17	12	3	15	10	2	8

To find the range we find the lowest value (2) and the highest value (17). Sometimes the range is expressed as a single figure, calculated by subtracting the lowest value from the highest:

$$\text{Range} = 17 - 2 = 15$$

Sometimes it is expressed as the range of scores: here the range is from 2 to 17.

The range suffers from one huge problem, in that it is massively affected by any outliers that occur. If one person gets a strange score the range is very distorted and two dispersions that may actually be very similar are suddenly made to appear very different. Because of this, the range is only rarely used in psychological research. It is most commonly used to describe some aspect of a sample which does not need to be summarised with any degree of accuracy. For example ‘the ages of the participants ranged from 18 to 48’ or ‘class sizes ranged from 23 to 36’.

Inter-quartile range

The **inter-quartile range** (IQR) is used with ordinal data or with non-normal distributions. If you use the median as the measure of central tendency then you’ll probably use the inter-quartile range as a measure of dispersion.

The inter-quartile range is, you will be surprised to hear, the distance between the upper and lower quartiles. There are three quartiles in a variable – they are the three values that divide the variable into four groups. The first quartile happens one-quarter of the way up the data, which is also the 25th centile. The second quartile is the half-way point, which is the median, and is also the 50th centile. The third quartile is the three-quarter-way point, or the 75th centile.

COMMON MISTAKES

1. People often think of the quartile as being the range of the data – that is that the bottom 25% of people are in the lower quartile. This is not correct.
2. Because the median is the 2nd quartile (and the 50th centile) there is no point presenting both in a table. (A computer will do this if you ask it to.)

To find the inter-quartile range the scores are placed in rank order and counted. The half-way point is the median (as we saw previously). The IQR is the distance between the quarter and three-quarters distance points.

To show how this is calculated, look at Table 2.5. This table shows a variable with 15 values, which have been placed in order. The median is the middle point, which is the 8th value, and is equal to 16. The upper quartile is found one-quarter of the way through the data, which is the 4th point, and is equal to 6. The upper quartile is found at the three-quarter-way point, and is equal to 23. The inter-quartile range is therefore 23 – 6 = 17.

Table 2.5 *Calculation of the inter-quartile range*

Count	Value	
1	2	
2	3	
3	5	
4	6	Lower quartile = 6
5	14	
6	15	
7	16	
8	16	Midpoint – median = 16
9	16	
10	21	
11	22	
12	23	Upper quartile = 23
13	24	
14	33	
15	45	

Unlike the range, the IQR does not go to the ends of the scales, and is therefore not affected by outliers. It also is not affected by skew and kurtosis to any great extent. You might also come across the *semi-inter-quartile range*. This is the inter-quartile range divided by 2.

Standard deviation

The final measure of dispersion that we shall look at is the **standard deviation** (SD). The standard deviation is like the mean, in that it takes all of the values in the dataset into account when it is calculated, and it is also like the mean in that it needs to make some assumptions about the shape of the distribution. To calculate the standard deviation, we must assume that we have a normal distribution.

So, why do we pay the price of making those assumptions? Why not just use the inter-quartile range? The SD can be used in a wide range of further analyses, as we will see later.

The standard deviation is calculated as follows:

$$\sigma = \sqrt{\frac{\sum(x - \bar{x})^2}{N - 1}}$$

TIP

Don't be put off by the equation. Remember:
$\bar{x}$ is the mean;
Σ means 'add them all up';
σ is the standard deviation;
N is the number of cases.

Let us calculate the SD for the following variable (x):

x:	9	8	7	1	11	10	4	13	4	3	7

Just as when we looked at the mean, we will take the equation and substitute the symbols with numbers, and then we will work it out.

Table 2.6 shows some of the calulations. The main steps are as follows:

Table 2.6 *Calculating* $(\Sigma x - \bar{x})^2$

Case	**Score**	$\bar{x}$	$x - \bar{x}$	$(x - \bar{x})$
1	9	7	2	4
2	8	7	1	1
3	7	7	0	0
4	1	7	–6	36
5	11	7	4	16
6	10	7	3	9
7	4	7	–3	9
8	13	7	6	36
9	4	7	–3	9
10	3	7	-4	16
11	7	7	0	0
Totals	77			136

1. Write down the equation. The x refers to each value, $\bar{x}$ is the mean, the superscript 2 means 'square' and the Σ is the Greek letter sigma, which means 'take the sum of'.

$$\sigma = \sqrt{\frac{\sum(x - \bar{x})^2}{N - 1}}$$

2. It will help us to calculate $\sum (x–\bar{x})^2$ if we draw up a table (see Table 2.6). The score column contains the individual scores. The next column contains the mean; we looked at that earlier, but we show the workings here. The mean is 7.

$$\bar{x} = \frac{9 + 8 + 7 + 11 + 10 + 4 + 13 + 4 + 3 + 7}{11}$$

$$\bar{x} = \frac{77}{11}$$

$$\bar{x} = 7$$

3. The next stage is to calculte $x–\bar{x}$ for each person. The calculations for the first two individuals are calculated on the right, and we have filled in the rest in Table 2.6.

$9 - 7 = 2$
$8 - 7 = 1$

4. Next we need to calculate $(x–\bar{x})^2$. To do this, we square each of the values that we calculated at stage 3. Again, I have shown the first two cases on the right, and I have filled in the table 2.6.

$2^2 = 2 \times 2 = 4$
$1^2 = 1 \times 1 = 1$

5. We can add each of these values together, to find $\sum (x–\bar{x})^2$

$\sum (x–\bar{x})^2 = 4 + 1 + 0 + 36 + 16 + 9 + 9 + 36 + 9 + 16 +$

6. Calculate $N - 1$.

$\sum (x–\bar{x})^2 = 136$

$N = 11$ (that's how many rows we have in Table 2.6), so $N - 1 = 10$. This gives the bottom half of the fraction.

$$\sigma = \sqrt{\frac{136}{10}}$$

7. Now divide the top half of the fraction by the bottom half:
$136 \div 10 = 13.6$.

$$\sigma = \sqrt{13.6}$$

8. Find the square root of step 7. (You will almost certainly need a calculator to do this.)
$\sqrt{13.6} = 3.69$.
This is the standard deviation.

$\sigma = 3.69$

OPTIONAL EXTRA: IT'S (ALMOST) ALL GREEK TO ME

The symbol used for the standard deviation is a little confusing, and this arises because there are actually two different forms of the standard deviation. The first is the *sample standard deviation*, which is referred to as *s*, and is calculated using

$$s = \sqrt{\frac{\sum(x - \bar{x})^2}{n}}$$

However, the sample standard deviation suffers from a problem–it is a biased estimator of the population standard deviation. This means that if we took lots of samples from a population, and calculated the standard deviation using the above formula, we would not expect the standard average of each of those standard deviations to match the population standard deviations. The sample standard deviations would, on average, be a bit too low. (In contrast, if we did this with the mean, we would find that the mean of all the means did match the population mean – hence the mean is an *unbiased* estimator.

Instead, we use the unbiased standard deviation, or the *population standard deviation*, which is given by this section:

$$\sigma = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}}$$

Notice that it looks very similar to the previous one, except it has n–1 instead of *n*. Because this is a population estimate (we'll discuss this on page xxx) it is written with a Greek letter–the lower-case sigma,

In statistics, we never want the (biased) sample standard deviation, we always want the population standard deviation, so whenever we talk about the standard deviation, we are talking about σ, not *s*.

One more thing-if you ever happen to use Excel to calculate a standard deviation, there are two functions available. Stdev() is the standard deviation that we use in statistics, which we have called the population standard deviation and refer to as σ; and stdevp(), very confusingly, is what Excel calls the population standard deviation, but which we call the sample standard deviation, *s*.

BOXPLOTS FOR EXPLORING DATA

We have focused so for on describing data using summary statistics. However, there are also a number of graphical techniques that can be used. We'll focus on one of the most useful, the **boxplot**, or box and whisker plot.

We used Table 2.5 to calculate the inter-quartile range. We'll now use it again to draw a boxplot (see Figure 2.19)

The median in a boxplot is represented with a thick line, and the upper and lower quartiles are shown as a box around the median. The whiskers then extend from the box to the highest and the lowest points – *unless* this would mean that the length of the whisker would be more than 1.5 times the length of the box, in which case it extends to the furthest point which means it does not exceed 1.5 times the length of the box. The computer defines that point as an outlier.

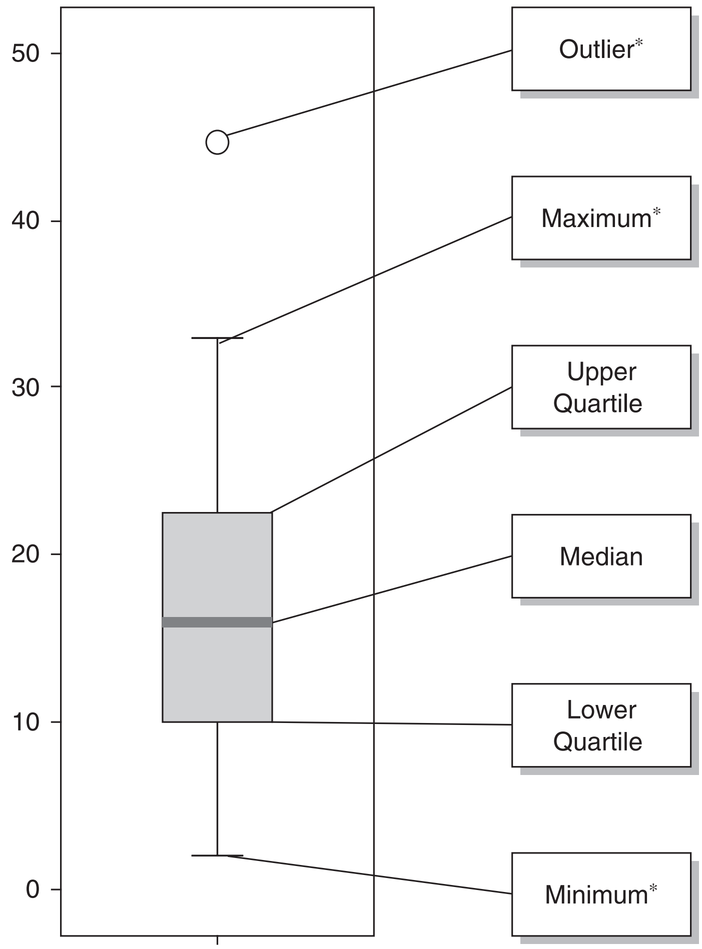


Figure 2.19 Boxplot of the data in Table 2.5

(1)

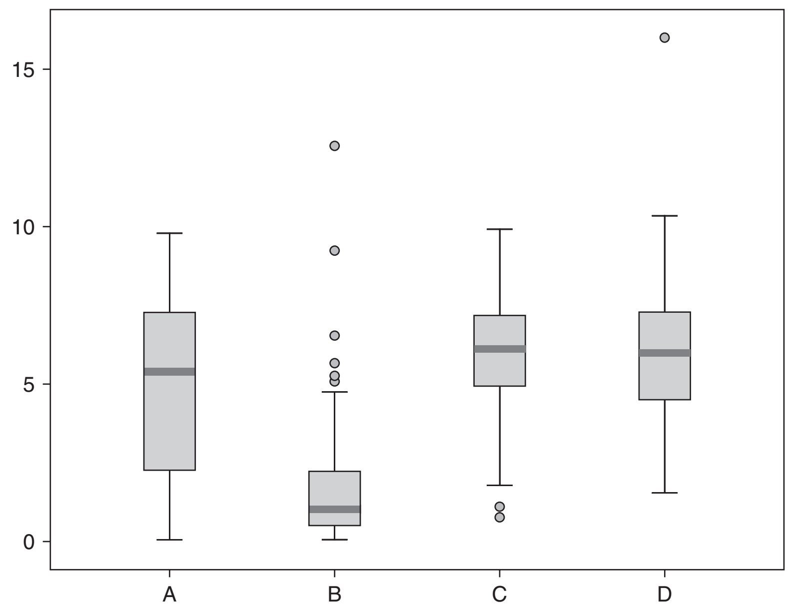
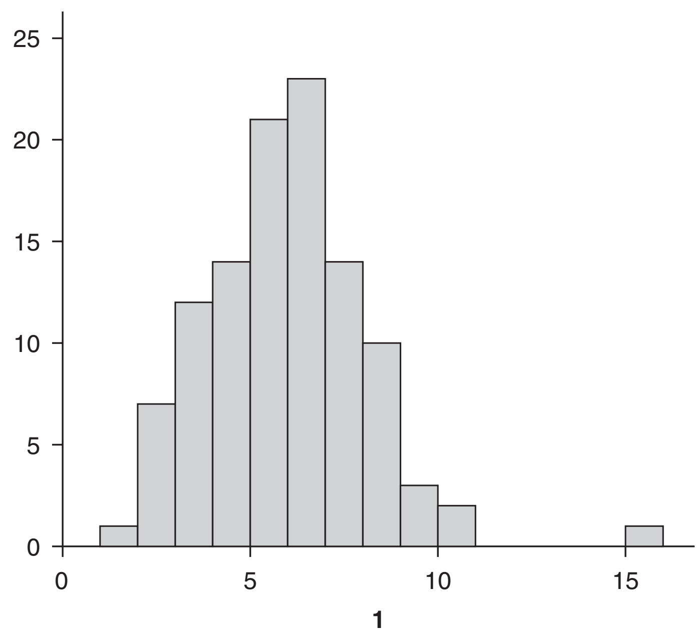


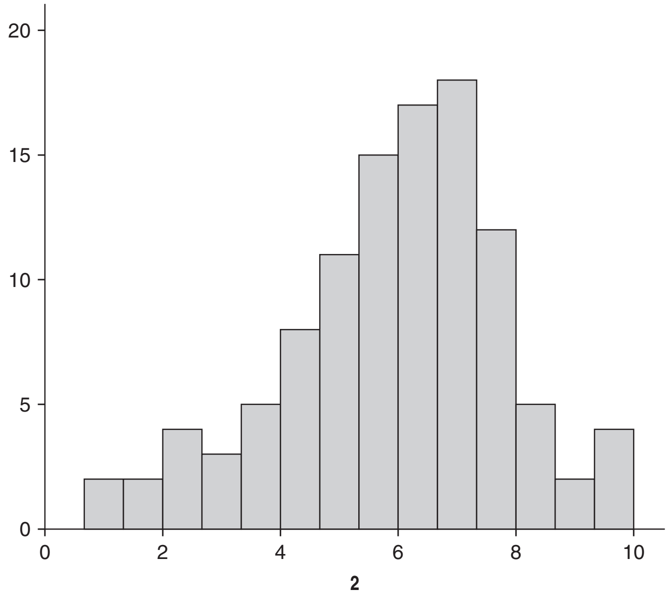
Figure 2.20 Which histogram goes with which boxplot?

(2)

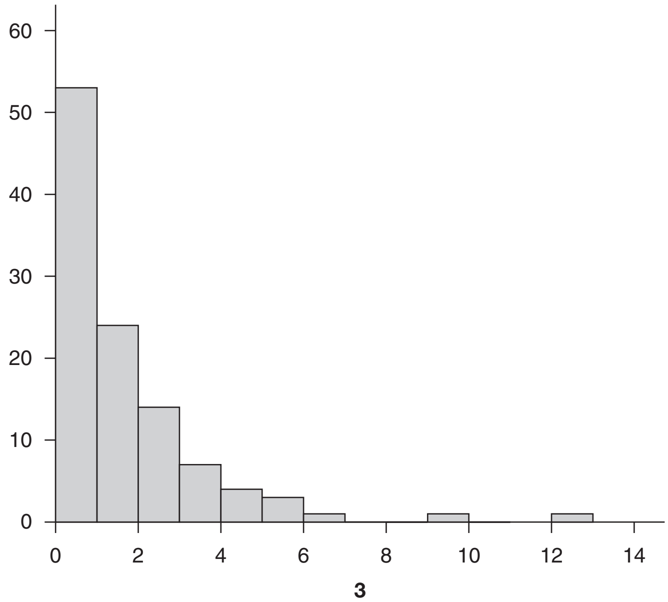


(*Continued*)

(3)



(4)



(*Continued*)

(5)

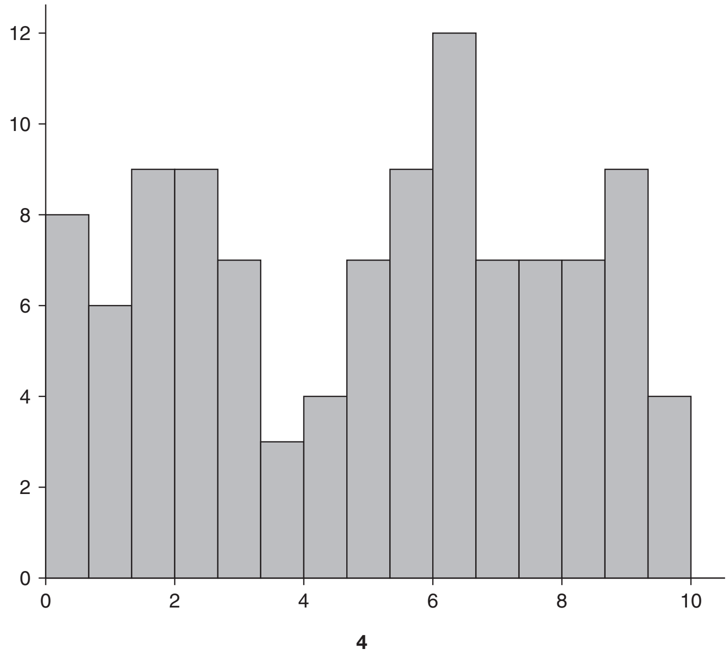


Figure 2.20 which histogram goes with which boxplot?

COMMON MISTAKES

1. If you use a computer program, such as SPSS, to draw a boxplot, you will find that it might label the outlier with something like the case number. Whilst this is useful for you – you can find the person in your dataset, it is useless for anyone reading your report – they don't have your data, after all. (It might even be worse than useless – they might think it is something important, which they don't understand.)
2. Notice that we said *the computer defines that point as an outlier*. That doesn't mean *you* have to define it as an outlier. Don't blindly follow the computer – *you* have to decide if it is an outlier.

Test Yourself 5

Look at the boxplot, showing four variables, in Figure 2.20. Below that are four histograms, showing the same variables. Except we've jumbled up the order. See if you can decided which histogram goes with each boxplot. (The sample sizes are all 100.)

Answers are at the end of the chapter.

One final point to note about boxplots, which you may have realised from this exercise – quite simply, they save space. With boxplots, we showed the distribution of four variables using one graph. If we wanted to use histograms, we would need four graphs.

SUMMARY

We've looked at some of the most common techniques that are used to describe statistics. There are a range of measures of central tendency and dispersion and the choice of technique depends on the type of measurement you have and the type of distribution you observe in the data. The principles are straightforward even if the technical terms seem complicated and the equations look off-putting. Concentrate on the principles and everything will fall into place.

Test yourself answers

Harry Potter (page xxx)

We don't know how many children were asked or what they said, but here are some data from 20 children that match the statistics the reporters gave:

1,000,000,000	3
1,000,000,000	3
1,000,000	3
10,000	3
1,000	3
500	3
300	3
20	3
3	3
3	3

Some of the children gave extreme answers, which were feasible (to watch the film a billion times would take almost 300,000 years, and that's if you didn't eat, sleep or go to the toilet). We could discard the outliers, because they are clearly not sensible (even watching it 1000 times will take over 3 months). When we do this, we have a highly skewed distribution, in which case a sensible thing to do would be to take the median, which gives 3.

Test Yourself

1. Shoe size is a measurement variable, and it is interval, not ratio. We could argue about whether it was discrete or continuous. I would say that it was continuous, because it would be possible to have size 12¼ shoes, but no one would make them.

2. Height is a continuous, ratio measure. If I am 1 metre tall, and you are 2 metres tall, you are twice as tall as me.
3. Phone number is a nominal measure – there is no meaningful order to phone numbers.
4. Degrees celsius is a continuous interval measure – 0°c is not a true zero point.
5. Position in the top 40 is an ordinal measure.
6. Number of CD sales is a discrete, ratio measure.
7. Cash earned from CD sales is a continuous, ratio measure. You could earn 0.001p, you just couldn't spend it.
8. Length of headache is a continuous, ratio measure.
9. Health rating on this scale is an ordered categorical measure.
10. Shoe colour is a nominal measure; shoes can be a number of colours, but they are not in any order.
11. Sex is a binary measure. Gender might be considered to be a continuous measure, you can be more or less masculine or feminine, but you must be male or female.
12. Number of times pecked by a duck is a ratio, discrete measure.
13. IQ is argued about. Most psychologists treat it as a continuous measure, but some people argue that it is actually ordinal. It is not ratio, because you cannot have an IQ of zero.
14. Blood pressure is a continuous measure. We could argue about whether is was ratio or not – it appears to be ratio, because it is measured in millimetres of mercury, but you cannot have a blood pressure of zero, because you would be dead. And if your blood pressure was low, around 20, say, you would be dead as well (or very soon). The zero doesn't really mean anything, so we would treat it as interval.

Test yourself 5

A4. The boxplot shows very short whiskers, relative to the length of the box. This means that the distribution is not very spread out. Histogram 4 is kurtosed; in other words, the distribution has very short tails, and does not spread out very much.

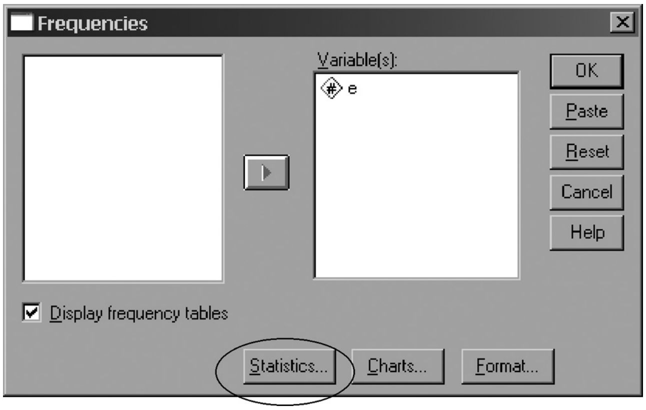
B3. In the boxplot, the median line is not in the middle of the box. This is our first clue that the data are skewed. The second clue is that the whiskers are of very different lengths, meaning that the data extend more in one direction than the other. Finally, there are some outliers at the top. All of these match the distribution shown in histogram 3.

C2. In the boxplot, the median line is in the middle of the box. The whiskers are of equal length. This indicates that the distribution is symmetrical. In addition, the whiskers are a bit longer than the length of the box. This indicates that the distribution has more points in the 'tails' than boxplot A. There are two outliers highlighted by the computer, but in a sample size of 100, this should not surprise us. The distribution shown in histogram 2 is close to a normal distribution, and this is what the boxplot shows us.

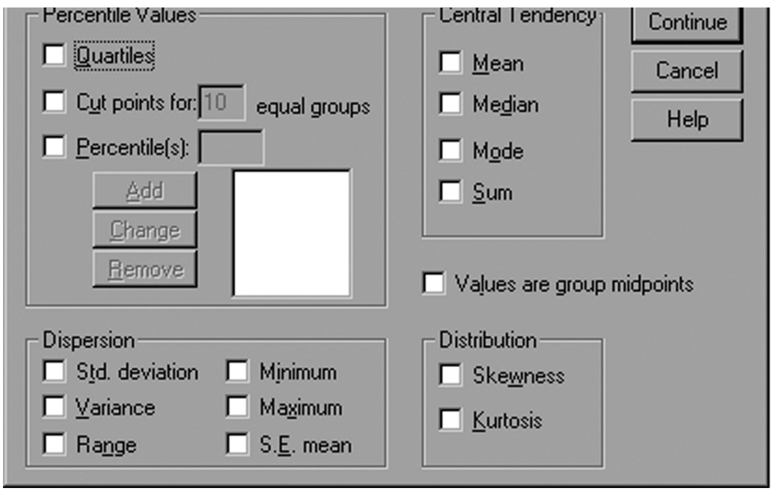
D1. Finally, boxplot D looks very similar to C, except that there is an outlier. Boxplot C showed a normal distribution, so we should expect the histogram to show a normal distribution, but also have an outlier, which is exactly what histogram 1 shows.

USING SPSS

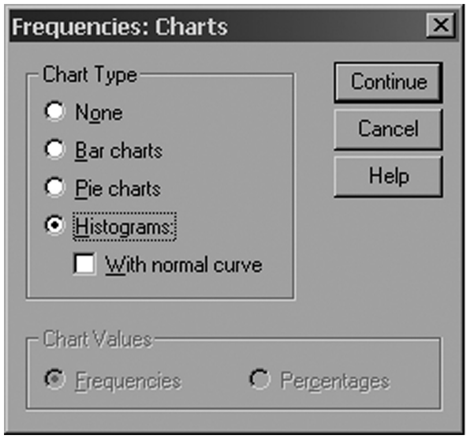
The statistics that we have shown in this chapter can all be calculated in SPSS using the Frequencies command. Select: **Analyze** ⇒ Descriptive Statistics ⇒ Frequencies:



Click on the **Statistics** button:



Choose the statistics that are best suited to your data. Click on **Continue.**
Now click on **Charts:**



Choose the type of chart you want. (You almost certainly want a histogram).

Note that the output table that is produced shows the difference between exploratory and descriptive statistics. The valid scores are given because there might be scores that are missing or that have been defined as missing values. There are two columns for percentage – percent, and valid percent. This is to help you, the data analyst, understand what is happening. In our data they are the same so don't put them both into your report.

Finally, there is the cumulative percent. Again, this might be useful to you but it probably won't be useful to the reader of your report. If it's not useful to the reader don't put it in your report.

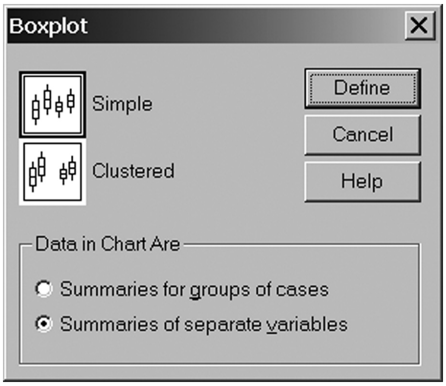
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	5	2	2.0	2.0	2.0
	7	2	2.0	2.0	4.0
	8	3	3.0	3.0	7.0
	9	1	1.0	1.0	8.0
	10	5	5.0	5.0	13.0
	11	5	5.0	5.0	18.0
	12	1	1.0	1.0	19.0
	13	7	7.0	7.0	26.0
	14	10	10.0	10.0	36.0
	15	8	8.0	8.0	44.0
	16	6	6.0	6.0	50.0
	17	13	13.0	13.0	63.0
	18	4	4.0	4.0	67.0
	19	4	4.0	4.0	71.0
	20	8	8.0	8.0	79.0
	21	1	1.0	1.0	80.0

(*Continued*)

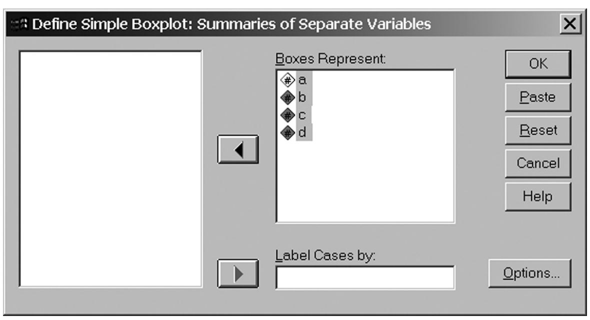
(Continued)

	Frequency	Percent	Valid Percent	Cumulative Percent
22	3	3.0	3.0	83.0
23	4	4.0	4.0	87.0
24	3	3.0	3.0	90.0
25	4	4.0	4.0	94.0
26	4	4.0	4.0	98.0
29	1	1.0	1.0	99.0
38	1	1.0	1.0	100.0
Total	100	100.0	100.0	

Boxplots are drawn through the **Graphs** command. Select **Graphs** ⇒ **Boxplot:**



Select **Simple** (this is the default). If you want to draw a boxplot for one or more variables, choose **Summaries of separate variables**. If you want to have separate boxplots for groups of variables, choose **Summaries for groups of cases.** Click on **Define**. Select the variables you want to have in your boxplot:



Click **OK**.