CHAPTER 1

What are statistics?

1. Antibiotics reduce the duration of viral throat infections by 1–2 days.
2. Five per cent of women aged 30–49 consult their GP each year with heavy menstrual bleeding.
3. At our health centre, 50 patients were diagnosed with angina last year.

(after Rowntree, 1981)

The above quotes may be fact or fallacy, but they are familiar examples of statistics. We use statistics every day, often without realising it. Statistics as an academic study has been defined as follows:

The science of assembling and interpreting numerical data (Bland, 2000)

The discipline concerned with the treatment of numerical data derived from groups of individuals (Armitage et al., 2001)

The term data refers to ‘items of information’, and is plural. When we use statistics to describe data, they are called descriptive statistics. All of the above three statements are descriptive.

However, as well as just describing data, statistics can be used to draw conclusions or to make predictions about what may happen in other subjects. This can apply to small groups of people or objects, or to whole populations. A population is a complete set of people or other subjects which can be studied. A sample is a smaller part of that population. For example, ‘all the smokers in the UK’ can be regarded as a population. In a study on smoking, it would be almost impossible to study every single smoker. We might therefore choose to study a smaller group of, say, 1000 smokers. These 1000 smokers would be our sample.

When we are using statistics to draw conclusions about a whole population using results from our samples, or to make predictions of what will happen, they are called
inferential statistics. Statements 1 and 2 on page 1 are examples of inferential statistics. It is important to recognise that when we use statistics in this way, we never know exactly what the true results in the population will be. For example, we shall never know how often every woman consults her GP (these data are not routinely collected in primary care at present), but we can draw a conclusion that is based on a sample of data.

A statistic is a quantity calculated from a sample, which describes a particular feature. Statistics are always estimates. The true quantities of the population (which are rarely known for certain) are called parameters.

Different types of data and information call for different types of statistics. Some of the commonest situations are described on the following pages.

Before we go any further, a word about the use of computers and formulae in statistics. There are several excellent computer software packages (as well as calculators) that can perform statistical calculations more or less automatically. Some of these packages are available free of charge, while some cost well over £1000. Each package has its own merits, and careful consideration is required before deciding which one to use. These packages can avoid the need to work laboriously through formulae, and are especially useful when one is dealing with large samples. However, care must be taken when interpreting computer outputs, as will be demonstrated later by the example in Chapter 6. Also, computers can sometimes allow one to perform statistical tests that are inappropriate. For this reason, it is vital to understand factors such as the following:

- which statistical test should be performed
- why it is being performed
- what type of data are appropriate
- how to interpret the results.

Several formulae appear on the following pages, some of which look fairly horrendous. Don't worry too much about these – you may never actually need to work them out by hand. However, you may wish to work through a few examples in order to get a 'feel' for how they work in practice. Working through the exercises in Appendix 2 will also help you. Remember, though, that the application of statistics and the interpretation of the results obtained are what really matter.
CHAPTER 2

Populations and samples

It is important to understand the difference between populations and samples. You will remember from the previous chapter that a population can be defined as every subject in a country, a town, a district or other group being studied. Imagine that you are conducting a study of post-operative infection rates in a hospital during 2014. The population for your study (called the target population) is everyone in that hospital who underwent surgery during 2014. Using this population, a sampling frame can be constructed. This is a list of every person in the population from whom your sample will be taken. Each individual in the sampling frame is usually assigned a number, which can be used in the actual sampling process.

If thousands of operations have been performed during 2014, there may not be time to look at every case history. It may therefore only be possible to look at a smaller group (e.g. 200) of these patients. This smaller group is a sample.

Remember that a statistic is a value calculated from a sample, which describes a particular feature. This means it is always an estimate of the true value.

If we take a sample of 100 patients who underwent surgery during 2014, we might find that 7 patients developed a post-operative infection. However, a different sample of 100 patients might identify 11 post-operative infections, and yet another might find 8 infections. We shall almost always find such differences between samples, and these are called sampling variations.

When undertaking a scientific study, the aim is usually to be able to generalise the results to the population as a whole. Therefore we need a sample that is representative of the population. Going back to our example of post-operative infections, it is rarely possible to collect data on everyone in a population. Methods therefore exist for collecting sufficient data to be reasonably certain that the results will be accurate and applicable to the whole population. The random sampling methods that are described in the next chapter are among those used to achieve this.

Thus we usually have to rely on a sample for a study, because it may not be practicable to collect data from everyone in the population. A sample can be used to
estimate quantities in the population as a whole, and to calculate the likely accuracy of the estimate.

Many sampling techniques exist, and these can be divided into non-random and random techniques. In random sampling (also called probability sampling), everyone in the sampling frame has an equal probability of being chosen. This approach aims to make the sample more representative of the population from which it is drawn. There are several methods of random sampling, some of which are discussed in the next chapter. Non-random sampling (also called non-probability sampling) does not have these aims, but is usually easier and more convenient to perform.

Convenience or opportunistic sampling is the crudest type of non-random sampling. This involves selecting the most convenient group available (e.g. using the first 20 colleagues you see at work). It is simple to perform, but is unlikely to result in a sample that is either representative of the population or replicable.

A commonly used non-random method of sampling is quota sampling, in which a predefined number (or quota) of people who meet certain criteria are surveyed. For example, an interviewer may be given the task of interviewing 25 women with toddlers in a town centre on a weekday morning, and the instructions may specify that seven of these women should be aged under 30 years, ten should be aged between 30 and 45 years, and eight should be aged over 45 years. While this is a convenient sampling method, it may not produce results that are representative of all women with children of toddler age. For instance, the described example will systematically exclude women who are in full-time employment.

As well as using the correct method of sampling, there are also ways of calculating a sample size that is appropriate. This is important, since increasing the sample size will tend to increase the accuracy of your estimate, while a smaller sample size will usually decrease the accuracy. Furthermore, the right sample size is essential to enable you to detect a real effect, if one exists. The appropriate sample size can be calculated using one of several formulae, according to the type of study and the type of data being collected. The basic elements of sample size calculation are discussed in Chapter 21. Sample size calculation should generally be left to a statistician or someone with a good knowledge of the requirements and procedures involved. If statistical significance is not essential, a sample size of between 50 and 100 may suffice for many purposes.
Random selection of samples is another important issue. In random sampling, everyone in the sampling frame has an equal probability of being chosen. For a sample to be truly representative of the population, a random sample should be taken. Random sampling can also help to minimise bias. Bias can be defined as an effect that produces results which are systematically different from the true values (see Chapter 24).

For example, imagine that you are conducting a study on hypertension (high blood pressure). You have 300 hypertensive patients, and want to find out what proportion have had their blood pressure checked in the past year. You might make a list of all of these patients, and decide to examine the records of the first 50 patients on the list. If most of them are found to have received blood pressure checks, are the other 250 patients likely to be similar? Furthermore, what if someone accuses you of ‘fixing’ the sample by only selecting patients who you know have received a blood pressure check? If you use a random sampling system, such doubts can be minimised.

There are many different random sampling systems, but one simple method is to use a random number table (these can be purchased or downloaded) or a computer program to produce a random number list to select the sample. Free web-based resources can generate a random number list. For example, if you want a random sample of 50 from a population of 300, you could list all 300 subjects and assign a number to each. Then use the numbers on the random number list, which match the numbers you have assigned. This produces a simple random sample. Generating 50 random numbers from 300 produces a list like the one shown in Table 3.1.

Multi-stage sampling can also be used. For example, in a study of university students in the UK, it would be difficult to obtain a complete list of all students. Even if such a list were available, the sheer number of students would be difficult to manage. To overcome this problem, multi-stage sampling could involve first selecting a simple random sample of all UK universities (first stage), and then a simple random sample of student names could be drawn from each selected university (second stage). This approach saves time, as it avoids the need to study every university. Additional stages
can be added to multi-stage sampling. For example, after randomly selecting the universities (first stage), a simple random sample of each university's faculties could be taken (second stage), and then a simple random sample of students within each faculty (third stage). Although multi-stage sampling can provide better focus and save resources, it will yield less precise results than would be obtained by taking a simple random sample from a complete list of all UK university students.

<table>
<thead>
<tr>
<th>TABLE 3.1 Random number list showing 50 random numbers</th>
</tr>
</thead>
<tbody>
<tr>
<td>8 12 14 22 24 27 33 37 49</td>
</tr>
<tr>
<td>55 67 78 79 93 95 98 104 108</td>
</tr>
<tr>
<td>113 116 125 128 133 138 143 158</td>
</tr>
<tr>
<td>163 167 169 171 173 176 184 193 203</td>
</tr>
<tr>
<td>212 218 219 221 224 225 230 232 249</td>
</tr>
<tr>
<td>264 272 273 283 285</td>
</tr>
</tbody>
</table>

Cluster sampling is similar to multi-stage sampling, except that all of the subjects in the final-stage sample are investigated. In the three-stage example just described, the randomly selected faculties would be regarded as clusters, and all students within these faculties would be studied.

It can be useful to employ stratified sampling to randomly select subjects from different strata or groups. Imagine a study designed to examine possible variations in healthcare between Asian and non-Asian patients. A random sample of patients on a list would almost certainly produce very few Asian patients, as most localities have a lower proportion of Asian residents. In such a case, we could stratify our sample by dividing patients into Asian and non-Asian subjects, and then take a random sample of the same size for each.

A less random but nevertheless useful approach is to use a systematic sampling scheme. In this method, a number is assigned to every record, and then every $n$th record is selected from a list. For example, if you want to systematically select 50 of your 300 patients with angina, the procedure would be as follows:

1. Obtain a list of all 300 patients with angina (this is your sampling frame).
2. As $300/50 = 6$, you will be taking every sixth patient.
3. Choose a number randomly between 1 and 6 as a starting point.
4. Take every sixth patient thereafter (e.g. if your starting point is 4, you will take patient numbers 4, 10, 16, 22, 28, 34, etc.).

By doing this, you are using the list rather than your own judgement to select the patients. Look at the list carefully before you start selecting. For example, choosing
every tenth patient in a list of married couples may well result in every selected person being male or every selected person being female (Donaldson & Scally, 2009).

For randomised controlled trials (see Chapter 31), random number tables can also be used to allocate patients to treatment groups. For example, the first number in the table can be allocated to the first patient, the second number to the second patient and so on. Odd numbers may be allocated to treatment group A, and even numbers to treatment group B. Other methods include subjects being randomly allocated to treatment groups by opening sealed envelopes containing details of the treatment category.
A variety of graph styles can be used to present data. The most commonly used types of graph are pie charts, bar diagrams, histograms and scattergrams.

The purpose of using a graph is to tell others about a set of data quickly, allowing them to grasp the important characteristics of the data. In other words, graphs are visual aids to rapid understanding. It is therefore important to make graphs as simple and easy to understand as possible. The use of ‘three-dimensional’ and other special effects can detract from easy and accurate understanding. Such approaches should therefore be avoided altogether, or used with great care. Also, omitting ‘0’ from a scale can make the graph misleading. Some examples of graphs follow.

The graph in Figure 4.1 is known as a pie chart, because it depicts each category as a slice of pie, with the size of each slice varying according to its proportion of the whole pie. This can be useful for comparing individual categories with the total. The pie chart in Figure 4.1 shows the distribution of different visual impairments in Swiss pathologists. It is easy to see that myopia (nearsightedness) was recorded most frequently, and that 8.6% of those with a visual impairment had hyperopia (farsightedness).

Figure 4.2 shows an example of a bar diagram. In this example, the size of each block represents the frequency recorded for the category concerned. Bar diagrams are useful for comparing one category with others. In the bar diagram shown in Figure 4.2, we can see the percentage of musculoskeletal problems recorded in Swiss pathologists. It is clear that the percentage of neck problems was nearly three times higher than hand/arm problems.

The graph shown in Figure 4.3 is called a histogram. Histograms are bar diagrams, where the areas (i.e. height and width) of the bars are proportional to the frequencies in each group. These are especially useful for frequency distributions of grouped data (e.g. age groups, grouped heights, grouped blood measurements). For example, if you use age groups of equal range (e.g. 21–30, 31–40, 41–50 years, etc.), then the width of each bar is equal, and if the 21–30 years age group has a frequency of 30, while the 31–40 years age group has a frequency of 60, then the former group is exactly half the
FIGURE 4.1 Distribution of different visual impairments among Swiss pathologists. Source: Adapted from Fritzsche et al. (2012).

FIGURE 4.2 Location of musculoskeletal problems in Swiss pathologists. Source: Adapted from Fritzsche et al. (2012).
The histogram in Figure 4.3 shows the frequency distribution of patients presenting with myelodysplastic syndrome in a given period, with the patients grouped into 5-year blocks.

An example of a scatterplot is shown in Figure 4.4. In a scatterplot, two measurements (also called variables) are each plotted on separate axes. The variable on the (horizontal) $x$-axis is usually called the independent variable, and the variable on the (vertical) $y$-axis is usually called the dependent variable. You can usually tell which variable is dependent on the other by considering which variable could have been caused by which other variable. In Figure 4.4, the weight of an adult patient can depend on (or be caused by) his or her height, whereas height cannot be dependent on (or caused by) weight. Scatterplots are discussed further in Chapter 18.
This page intentionally left blank
Suppose we ask a sample of 30 teenagers each to tell us how old they are. The list of their ages is shown in Table 5.1:

<p>| | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td>14</td>
<td>16</td>
<td>15</td>
<td>17</td>
<td>14</td>
<td>16</td>
<td>17</td>
<td>14</td>
<td>18</td>
</tr>
<tr>
<td>19</td>
<td>16</td>
<td>17</td>
<td>14</td>
<td>13</td>
<td>15</td>
<td>14</td>
<td>16</td>
<td>16</td>
<td>19</td>
</tr>
<tr>
<td>19</td>
<td>18</td>
<td>14</td>
<td>17</td>
<td>14</td>
<td>16</td>
<td>15</td>
<td>17</td>
<td>15</td>
<td>17</td>
</tr>
</tbody>
</table>

This is all very well, but when the data are presented in this way, it is difficult to make sense of them quickly. For example, how many of the teenagers are old enough to drive? How many of them are old enough to purchase alcohol legally? Are there more 15-year-olds than 16-year-olds in this group? From the listing shown, it is difficult to answer these questions. Individual ages need to be picked out and counted up every time we want to answer such a question.

A summary of the data can make things easier. What if we count up how often each individual age is recorded, and write this down? Then we can look at the count each time we need to know something about these ages. In Table 5.2, the ages are sorted into numerical order, and the number of times each age is recorded is written at the side.

It is now easy to see how often each age occurs. We can quickly tell that 11 teenagers are old enough to drive (the legal age is 17 years in the UK), 5 can legally purchase alcohol (the legal age is 18 years in the UK) and there are more 16-year-olds \( n = 6 \) than 15-year-olds \( n = 5 \).

The number of times that something occurs is known as its frequency. For example, the frequency of 14-year-olds in our sample is 7, and the frequency of 18-year-olds
is 2. Table 5.2 shows the ages and their frequencies, and is called a frequency distribution. It shows how the ages are distributed in this sample.

**TABLE 5.2** Frequency distribution of age

<table>
<thead>
<tr>
<th>Age (years)</th>
<th>Number of times recorded</th>
</tr>
</thead>
<tbody>
<tr>
<td>13</td>
<td>1</td>
</tr>
<tr>
<td>14</td>
<td>7</td>
</tr>
<tr>
<td>15</td>
<td>5</td>
</tr>
<tr>
<td>16</td>
<td>6</td>
</tr>
<tr>
<td>17</td>
<td>6</td>
</tr>
<tr>
<td>18</td>
<td>2</td>
</tr>
<tr>
<td>19</td>
<td>3</td>
</tr>
</tbody>
</table>

In the frequency distribution in Table 5.3, the frequencies are added up and percentages added (in this example, the percentages are rounded up to the nearest whole percent). This is a common way of presenting a frequency distribution.

The percentages indicate the proportion of times that a particular age is recorded. Proportions can be expressed as decimals or multiplied by 100 and expressed as percentages.

For example, if 15 out of 30 teenagers are aged 18 years, then the proportion is 0.50 (15/30 = 0.50) or the percentage is 50% (0.50 × 100 = 50).

**Note that in statistics, we normally use the symbol ‘/’ for division, instead of ‘÷’:**

If 20 of the teenagers are aged 18 years, then the proportion is 0.67 (20/30 = 0.666 or 0.67 to two decimal places) or 67% (0.67 × 100 = 67).

**TABLE 5.3** Frequency distribution of age, also showing totals and percentages

<table>
<thead>
<tr>
<th>Age</th>
<th>Frequency</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>13</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>14</td>
<td>7</td>
<td>23</td>
</tr>
<tr>
<td>15</td>
<td>5</td>
<td>17</td>
</tr>
<tr>
<td>16</td>
<td>6</td>
<td>20</td>
</tr>
<tr>
<td>17</td>
<td>6</td>
<td>20</td>
</tr>
<tr>
<td>18</td>
<td>2</td>
<td>7</td>
</tr>
<tr>
<td>19</td>
<td>3</td>
<td>10</td>
</tr>
<tr>
<td>Total</td>
<td>30</td>
<td>100</td>
</tr>
</tbody>
</table>
In Table 5.3, 5 of the 30 teenagers are aged 15 years. The proportion is 0.17 ($\frac{5}{30} = 0.1666$ or $0.17$ to two decimal places), and the percentage is 17% ($0.17 \times 100 = 17$).

In these calculations, we have sometimes rounded numbers to two decimal places. For example, if we use a calculator to work out $\frac{20}{30}$, it will probably display ‘0.6666666’ – it has displayed seven numbers after the decimal point. This is called ‘displaying to seven decimal places’. To show this as three decimal places, we round the third digit after the decimal point up to the nearest whole number. Thus when displaying 0.6666666 to three decimal places, 0.667 is nearer to the real value than is 0.666. In other words, if the last digit is 5 or more, we round up to the next whole number. If we want to round 1.222 to two decimal places, 1.22 is nearer to the true value than 1.23. So if the last digit is 4 or less, we round down to the nearest number.

Proportions or percentages are more useful than frequencies when we want to compare numbers of events in two or more groups of unequal size. For example, suppose that we want to compare the number of industrial accidents in the work forces of two different companies. In company A, there have been 37 accidents among a total of 267 workers. In company B, 45 accidents have occurred among a total of 385 workers. At which company are workers more likely to have an accident? On the face of it, company B has experienced more accidents, but it also employs more workers. Unless you are very good at mental arithmetic, it is difficult to answer the question. Let us work it out using proportions:

- company A had 37 accidents among 267 workers – the proportion of accidents is 0.139 ($\frac{37}{267}$)
- company B had 45 accidents among 385 workers – the proportion of accidents is 0.117 ($\frac{45}{385}$).

Therefore even though company A’s workforce had fewer accidents, it is statistically the more dangerous place to work, as it had a higher proportion of accidents. When we use proportions to describe the number of events, they can be called rates. In this example, therefore, the accident rate in company A is 0.139 (or 13.9%) and that in company B is 0.117 (or 11.7%).
At this stage, it is worth mentioning the need to recognise different types of data. For example, we could ask people to give us information about how old they are in one of two ways. We could ask them to tell us how old they are in whole years (i.e. their age last birthday). Alternatively, we could ask them to tell us to which of several specified age bands they belong (e.g. 20–24, 25–29, 30–34 years, etc.). Although these two methods tell us about the age of the respondents, hopefully you can see that the two types of data are not the same!

Data can be classified as either **categorical** or **numerical**.

**CATEGORICAL DATA**

This refers to data that are arranged into separate categories. Categorical data are also called **qualitative** data.

If there are only two possible categories (e.g. yes/no, female or male), the data are said to be **dichotomous**. If there are more possible categories (e.g. a range of several age groups or ethnic minority groups), the data may be described as **nominal**.

Categories can sometimes be placed in order. In this case they are called **ordinal data**. For example, a questionnaire may ask respondents how happy they are with the quality of catering in hospital, the choices may be very happy, quite happy, unhappy or very unhappy. Other examples of ordinal data include positions in hospital league tables, and tumour stages. Because the data are arranged both in categories **and** in order, ordinal data provide more information than categories alone.

**NUMERICAL DATA**

For this type of data, numbers are used instead of categories. Numerical data are also called **quantitative** data.
There are three levels (scales) of numerical data. These are presented in order according to how much information they contain.

**In discrete** data, all values are clearly separate from each other. Although numbers are used, they can only have a certain range of values. For example, age last birthday is usually given as a whole number (e.g. 22 or 35, rather than 22.45 or 35.6, etc.). Other examples of discrete data include the number of operations performed in 1 year, or the number of newly diagnosed asthma cases in 1 month. It is usually acceptable to analyse discrete data as if they were continuous. For example, it is reasonable to calculate the mean number (see Chapter 7) of total knee replacement operations that are performed in a year.

The next two scales are regarded as **continuous** – each value can have any number of values in between, depending on the accuracy of measurement (for example, there can be many smaller values in between a height of 2 m and a height of 3 m, e.g. 2.2 or 2.23 or 2.23978675). Continuous data can also be converted into categorical or discrete data. For example, a list of heights can be converted into grouped categories, and temperature values in degrees centigrade (measured to one or more decimal places) can each be converted to the nearest whole degree centigrade.

**In interval** data, values are separated by **equally spaced** intervals (e.g. weight, height, minutes, degrees centigrade). Thus the difference (or interval) between 5 kg and 10 kg, for example, is exactly the same as that between 20 kg and 25 kg. As interval data allow us to tell the precise interval between any one value and another, they give more information than discrete data. Interval data can also be converted into categorical or discrete data. For example, a list of temperature measurements in degrees centigrade can be placed in ordered categories or grouped into dichotomous categories of ‘afebrile’ (oral temperature below 37°C) or ‘febrile’ (oral temperature of 37°C or more).

**Ratio** data are similar to interval scales, but refer to the ratio of two measurements and also have a true zero. Thus weight in kilograms is an example of ratio data (20 kg is twice as heavy as 10 kg, and it is theoretically possible for something to weigh 0 kg). However, degrees centigrade cannot be considered to be a ratio scale (20°C is not, in any meaningful way, twice as warm as 10°C, and the degrees centigrade scale extends below 0°C). Ratio data are also interval data.

Sometimes people get different types of data confused – with alarming results. The following is a real example (although the numbers have been changed to guarantee anonymity). As part of a study, a researcher asks a group of 70 pregnant women to state which of a range of age groups they belong to. These are entered into a table as shown in Table 6.1.
The researcher wants to enter the data into a computerised analysis program, and to ensure ease of data entry, he decides to give each group a numerical title (so that, when entering the data, he can simply press ‘3’ for someone who is in the ’22–26’ years age group, for example). Unfortunately, he does not notice that the program assumes that the numerical titles represent continuous data. It therefore treats the age groups as if they were actual ages, rather than categories. Being busy with other matters, the researcher does not notice this in the program’s data analysis output. In his report, he states that the mean age of the pregnant women is 4.03 years! Of course, the most frequently recorded age group (27–31 years), also called the mode (see Chapter 7), is the correct measure for these data. Treating categorical data as if they were continuous can thus produce very misleading results and is therefore dangerous. Clearly, great care needs to be taken to ensure that data are collected and analysed correctly.
This page intentionally left blank
Means, medians and modes are methods of measuring the central tendency of a group of values – that is, the tendency for values in a group to gather around a central or ‘average’ value which is typical of the group.

**MEAN**

It can be very useful to summarise a group of numerical values by finding their average value. The mean gives a rough idea of the size of the values that you are dealing with, without having to look at every one of them. The mean (or to use its proper name, the arithmetic mean) is another term for the average.

Consider the HbA₁c (the percentage of glycosylated haemoglobin circulating in the blood) values for patients with diabetes, shown in the frequency distribution in Figure 7.1. It also shows the median and mode, which are discussed later in this chapter.

The formula for calculating the mean is:

\[ \frac{\sum x}{n} \]

*Add up (\( \sum \)) all of the values (\( x \)) and divide by the number of values observed (\( n \)).*

To calculate a mean:

1. add up every value in your group (call this result A)
2. count how many values are observed in the group (call this result B)
3. divide result A by result B.
22  BASIC STATISTICS AND EPIDEMIOLOGY

In the example in Figure 7.1:

1. the sum of all of the HbA1c values listed = 180.6
2. the number of values observed = 27
3. $180.6 / 27 = 6.69$ (or 6.7 if we use one decimal place).

The mean is usually represented by $\bar{x}$ (called x-bar) for samples, and $\mu$ (called mu) for populations.

Remember that, when writing the mean, it is good practice to refer to the unit measured. In this case, it is a HbA1c value of 6.7%.

Note that many calculators will work out the mean in a single process, without having to go through the steps outlined here.

The mean can be misleading if there are any extreme values in a group of numbers. For example, the mean of the group 1, 2, 3, 2, 4, 5, 19 is 5.1. The value 19 is an extreme value, as it is far higher than any of the other numbers in the group. Since only one of the values in the group is actually 5.1 or greater, the mean is not representative of the group. In this case, the median may provide a better representation.

<table>
<thead>
<tr>
<th>%</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.0</td>
<td>1</td>
</tr>
<tr>
<td>4.3</td>
<td>1</td>
</tr>
<tr>
<td>4.4</td>
<td>1</td>
</tr>
<tr>
<td>4.5</td>
<td>1</td>
</tr>
<tr>
<td>4.7</td>
<td>1</td>
</tr>
<tr>
<td>4.9</td>
<td>2</td>
</tr>
<tr>
<td>Mode</td>
<td>5.0</td>
</tr>
<tr>
<td></td>
<td>5.4</td>
</tr>
<tr>
<td></td>
<td>5.5</td>
</tr>
<tr>
<td>Median</td>
<td>5.8</td>
</tr>
<tr>
<td></td>
<td>6.0</td>
</tr>
<tr>
<td></td>
<td>6.1</td>
</tr>
<tr>
<td></td>
<td>6.2</td>
</tr>
<tr>
<td>Mean (6.69)</td>
<td>7.0</td>
</tr>
<tr>
<td></td>
<td>7.6</td>
</tr>
<tr>
<td></td>
<td>7.9</td>
</tr>
<tr>
<td></td>
<td>8.5</td>
</tr>
<tr>
<td></td>
<td>8.9</td>
</tr>
<tr>
<td></td>
<td>9.9</td>
</tr>
<tr>
<td></td>
<td>10.7</td>
</tr>
<tr>
<td></td>
<td>10.8</td>
</tr>
<tr>
<td></td>
<td>10.9</td>
</tr>
<tr>
<td></td>
<td>11.2</td>
</tr>
</tbody>
</table>

| Total | 27 |

FIGURE 7.1  Frequency distribution of HbA1c values.
MEAN, MEDIAN AND MODE

MEDIAN
This is the middle value of an ordered sample of numerical values. To calculate the median:

1. arrange all of the recorded values in order of size
2. find the middle value.

If we arrange the following numbers in numerical order, we obtain:

1, 2, 2, 3, 4, 5, 19.

The median is 3.

In this example, the median is much more representative of the group than the mean (5.1). Extreme values do not affect the median, and the median value is usually typical of the data.

If there is an even number of values, use the mean of the two middle values:

19, 24, 26, 30, 31, 34.

The median is \((26 + 30)/2 = 28\).

The median HbA1c value in Figure 7.1 is 5.8 – there are 13 values below and 13 values above it.

MODE
The mode is the value which occurs most often in a group. This can be a group of either numbers or categories.

In Figure 7.1, the HbA1c value 5.0 is recorded more often than any other value (three times in all), and so it is the mode of that group.

For example, if you want to know the most frequently used health promotion clinic (e.g. ‘smoking cessation’, ‘weight loss’, ‘well woman’, ‘well man’, etc.) at a primary care surgery, count up the attendance at each clinic over a specific period, and find the one with the highest attendance.

If there are two modes in a group of numbers, the group is described as bimodal. The mode is easy to determine, and requires no calculation. It is usually typical of the data used. Because the mode only records the most popular value, the others are not taken into account. The mode is therefore not affected by extreme values.

The mode can be used for categorical data where the mean and median are not appropriate (e.g. as in the example shown in Table 6.1).
Although the median is the middle value in a group of ordered numbers, it provides no information about the range of values, or how the values are grouped around the median. The range uses only the highest and lowest values, which may be extreme values. As we have already found when discussing the mean, extreme values may provide a misleading representation of the central tendency of the group. One approach is to effectively ignore a percentage of values at each end of the group, and to concentrate on the central area, where the majority of values are likely to lie.

**Centiles** allow us to describe the central range of a group of numbers. They are often expressed as the 25th and 75th centiles, although it is possible to calculate centiles of any value (e.g. 3rd and 97th centiles). Centiles are also referred to as **percentiles**.

The 25th centile is also called the first quartile. It is the point which separates the lower quarter of the numbers in a group, in the same way as the median separates the upper half. The 50th centile is also called the second quartile, and is equivalent to the median. The 75th centile is also called the third quartile, and is the point that separates the upper quarter of the numbers.

The **interquartile range** is the distance between the 25th and 75th centiles, and is calculated by simply subtracting the 25th centile from the 75th centile. It provides an indication of how much variation (or spread) there is between the first and third quartiles. It ignores the values below the first quartile and above the third quartile.

For example, suppose that a group of patients has the following cholesterol values (in mmol/L):

3.5, 3.5, 3.6, 3.7, 4.0, 4.1, 4.3, 4.5, 4.7, 4.8, 5.2, 5.7, 6.1, 6.3, 6.3

The 25th centile is 3.7. The 50th centile (median) is 4.5. The 75th centile is 5.7. The interquartile range is: (5.7 – 3.7) = 2.0.

This means that there is a variation of 2.0 mmol/L between the first and third quartiles, and a range of 3.5–6.3 mmol/L. A second group of patients may have an
interquartile range of 0.9 mmol/L, indicating less variation. Even if the first and last values in the second group are very extreme (e.g. 3.0 and 9.0, respectively), these will not affect the interquartile range, which concentrates on the central area of values.
We have seen that the interquartile range indicates the variation of data where the median is the measure of central tendency. Standard deviation is used where this measure is the mean. It indicates the difference between a group of values and their mean, taking all of the data into account. Although this means that it may be influenced by extreme values, the standard deviation plays an important role in many tests of statistical significance (which will be described in later chapters). The larger the standard deviation, the more the values differ from the mean, and therefore the more widely they are spread out.

For example, one small group of patients in a particular outpatient clinic may wait for a mean time of 11 minutes to be seen by a doctor, and the standard deviation from the mean for this group is 5.701. Individual waiting times vary widely – from 7 minutes up to 21 minutes. There is wide variation between these waiting times, and they are quite widely spread out from their mean. These waiting times are therefore heterogeneous or dissimilar.

On another day, another group of patients from the same clinic may also have a mean waiting time of 11 minutes, but their standard deviation is 0.707. This is much less than the first group’s standard deviation of 5.701. Looking at this group’s actual waiting times, it can be seen that they only vary from 10 to 12 minutes. Waiting times for the second group are more homogeneous – that is, the data are more similar to each other. They are less widely spread out around their mean than the first group.

Let us look at the actual waiting times recorded for each group, as shown in Table 9.1.

You can see that the data in group 1 are much more spread out than those in group 2. This difference in standard deviations can be explained by the fact that, although most patients in group 1 waited a very short time, one patient had to wait for a long time (21 minutes). Although this one ‘outlier’ waiting time is not representative of the whole group, it has a large effect on the overall results, and it strongly affects the mean and standard deviation. Several patients from group 2 actually waited longer
than group 1 patients, although the difference between the waiting times in group 2 is very slight.

**TABLE 9.1** Waiting times and standard deviation for each patient group

<table>
<thead>
<tr>
<th>Group</th>
<th>Time 1</th>
<th>Time 2</th>
<th>Time 3</th>
<th>Time 4</th>
<th>Time 5</th>
<th>Mean</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>21</td>
<td>11</td>
<td>5.701</td>
</tr>
<tr>
<td>2</td>
<td>11</td>
<td>11</td>
<td>10</td>
<td>11</td>
<td>12</td>
<td>11</td>
<td>0.707</td>
</tr>
</tbody>
</table>

Although the abbreviations **SD** or **s.d.** are used to represent standard deviation generally, s is used to represent standard deviation for **samples**, and \( \sigma \) is used to represent standard deviation for **populations**.

The most usual formula for standard deviation is as follows:

\[
\sqrt{\frac{\sum(x - \bar{x})^2}{(n-1)}}
\]

where \( x = \) individual value, \( \bar{x} = \) sample mean and \( n = \) number of values.

The equation is only suitable for a sample (or **population estimate**). This will usually be the case, since we rarely know the true population value (which in this case is the mean).

The following steps are used to work out a standard deviation.

1. Find the mean of the group.
2. Subtract this from every value in the group individually – this shows the deviation from the mean, for every value.
3. Work out the square \((x^2)\) of every deviation (that is, multiply each deviation by itself, e.g. \(5^2 = 5 \times 5 = 25\)) – this produces a squared deviation for every value.
4. Add up all of the squared deviations.
5. Add up the number of observed values, and subtract 1.
6. Divide the sum of squared deviations by this number, to produce the **sample variance**.
7. Work out the square root of the variance.

If you have to work out a standard deviation by hand, it is helpful to use a grid like the one shown in Table 9.2. We can use this to work out the standard deviation of the data for group 1 from Table 9.1.
1. We already know the mean is 11 (see previous page).
2. Subtract each time value from the mean. Note each result in the 'Deviation from the mean' column.
3. Multiply each deviation by itself, and write each result in the 'Squared deviation' column (e.g. $-4^2 = -4 \times -4 = 16$) (note that multiplying minus numbers produces positive ones).
4. Adding all of the squared deviations ($1 + 16 + 9 + 4 + 100$) gives a value of 130.
5. There are five separate values in the group. Subtract 1, and you get 4.
6. Divide the sum of squared deviations by 4, to produce the variance ($130/4 = 32.5$).
7. Use a calculator to determine the square root of the variance ($\sqrt{32.5} = 5.701$).

Of course, calculating standard deviation by hand like this is not practical if you have a large number of values. Moreover, the mean is unlikely to be a whole number as it is in the example here. Calculators and computer programs are an invaluable aid to this process, and are readily available.

Other uses of standard deviation are discussed under normal distribution (see Chapter 11).
This page intentionally left blank
Standard error (or s.e.) is another term for the standard deviation of a sampling distribution (or frequency distribution of samples), rather than just a sample. You may remember from Chapter 2 that a value found from one sample may be different to that from another sample – this is called sampling variation. For example, if we took a large number of samples of a particular size from a population and recorded the mean for each sample, we could calculate the standard deviation of all their means – this is called the standard error. Because it is based on a very large number of (theoretical) samples, it should be more precise and therefore smaller than the standard deviation.

Standard error is used in a range of applications, including hypothesis testing and the calculation of confidence intervals (which are discussed in later chapters).

The most frequently used calculations are described as follows:

**COMPARING A SAMPLE MEAN WITH A POPULATION MEAN (FOR LARGE SAMPLES)**

\[
s.e. = s / \sqrt{n}
\]

Divide the standard deviation \(s\) by the square root of the number of values \(n\) in the sample.

To calculate the standard error, follow the steps listed below.

1. Calculate the standard deviation of the sample mean.
2. Count the number of observed values.
3. Find the square root of this sum.
4. Divide the standard deviation by this number.
Using the table of HbA1c values in Figure 7.1 in Chapter 7, we can calculate the standard error as follows:

1. The standard deviation is 2.322 (not shown in Chapter 7).
2. The number of observed values = 27.
3. The square root of 27 = 5.196.
4. Divide the standard deviation (2.322) by 5.196 = 0.447.

You can see that the standard error is very much smaller than the standard deviation.

**COMPARING TWO SAMPLE MEANS (FOR LARGE SAMPLES)**

\[
\text{s.e.} = \sqrt{\frac{s_1^2 + s_2^2}{n_1 + n_2}}
\]

where: \(s_1\) = standard deviation for sample 1, \(s_2\) = standard deviation for sample 2, \(n_1\) = sample size 1 and \(n_2\) = sample size 2.

Let us work through the stages of this formula.

1. Square the first sample standard deviation (\(s_1\)).
2. Divide it by the first sample size (\(n_1\)) – note the result, and call it ‘result 1’.
3. Square the second sample standard deviation (\(s_2\)).
4. Divide it by the second sample size (\(n_2\)) – note this result, and call it ‘result 2’.
5. Add results 1 and 2.
6. Find the square root of this number – this is the standard error.

**SINGLE PROPORTION (FOR LARGE SAMPLES)**

\[
\text{s.e.} = \sqrt{\frac{p(1-p)}{n}}
\]

where \(p\) = proportion and \(n\) = sample size.

There are different formulae for calculating standard error in other situations (e.g. for comparing proportions in two independent groups, where the sample size is large), and these are covered by several other texts.

Standard error formulae for **small** samples are presented in Chapter 15.
Normal distribution

If we take a large sample of men or women, measure their heights and plot them on a frequency distribution, the distribution will almost certainly obtain a symmetrical bell-shaped pattern that looks something like the one shown in Figure 11.1.

This is known as the **normal distribution** (also called the Gaussian distribution). The least frequently recorded heights lie at the two extremes of the curve. It can be seen that very few women are extremely short or extremely tall. An outline of the normal distribution curve is drawn around the frequency distribution, and is a reasonably good fit to the shape of the distribution. With a larger sample size, the pattern of the frequency distribution will usually follow this shape more closely.

![Figure 11.1](image_url)  
**Figure 11.1** Distribution of a sample of values of women’s heights.

In practice, many biological measurements follow this pattern, making it possible to use the normal distribution to describe many features of a population.

It must be emphasised that some measurements do not follow the symmetrical
shape of the normal distribution, and can be \textbf{positively skewed} or \textbf{negatively skewed}. For example, more of the populations of developed Western countries are becoming obese. If a large sample of such a population's weights was to be plotted on a graph similar to that in Figure 11.1, there would be an excess of heavier weights which might form a similar shape to the 'negatively skewed' example in Figure 11.2. The distribution will therefore not fit the symmetrical pattern of the normal distribution. You can tell whether the skew is positive or negative by looking at the shape of the plotted data, as shown in Figure 11.2.

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{positively_negatively_skewed_data.png}
\caption{Examples of positive and negative skew.}
\end{figure}

Furthermore, the shape may be symmetrical but different to the normal distribution. The normal distribution is shown in Figure 11.3. You can see that it is split into two equal and identically shaped halves by the mean. The standard deviation indicates the size of the spread of the data. It can also help us to determine how likely it is that a given value will be observed in the population being studied. We know this because the proportion of the population that is covered by any number of standard deviations can be calculated.

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{normal_distribution.png}
\caption{The normal distribution.}
\end{figure}
For example:

- **68.27%** of all values lie within plus or minus (±) one standard deviation (either one standard deviation below the mean or one standard deviation above it)
- **95.45%** of all values lie within ± two standard deviations of the mean
- **99.73%** of all values lie within ± three standard deviations of the mean.

It is useful to know that **95%** of all values lie within **1.96 standard deviations**, and **99%** of all values lie within **2.58 standard deviations**.

The proportions of values **below** and **above** a specified value (e.g. the mean) can be calculated, and are known as **tails**. We shall discuss these in Chapter 14.

It is possible to calculate the probability that a value in any particular range will occur. The normal distribution is useful in a number of applications, including confidence intervals (see Chapter 12) and hypothesis testing (see Chapter 14).

As well as the normal distribution, a number of other distributions are important, including the following:

- **the t-distribution** – for small samples (usually below 30) (see Chapter 15 on t-tests)
- **the binomial distribution** – for dichotomous data (e.g. result can only be 0 or 1; yes or no)
- **the Poisson distribution** – for rare events that occur randomly in a large population.

The t- and binomial distributions resemble the normal distribution when large samples are used.
This interval has been constructed from the random sample data using a procedure such that, if we took many such samples and constructed a confidence interval for each, then 95% of the varying intervals would contain the population mean (a fixed, but unknown value).

Although we can calculate a sample mean, we never know exactly where the population mean is. Confidence intervals are used to estimate how far away the population mean is likely to be, with a given degree of certainty. This technique is called estimation, and the term ‘confidence interval’ is often abbreviated to c.i. or CI. Conventionally, 95% confidence intervals are used, although they can be calculated for 99% or any other value.

Figure 12.1 shows diastolic blood pressure measurements taken from a sample of 92 patients with diabetes. The mean diastolic blood pressure is 82.696 mmHg, with a standard error of 1.116. A 95% confidence interval will indicate a range above and below 82.696 mmHg in which the population mean will lie, with a 95% degree of certainty. In other words, a ‘95% confidence interval’ is the interval which will include the true population value in 95% of cases.

The formula for calculating a 95% confidence interval for a sample mean (large samples) is:

$$\bar{x} \pm (1.96 \times \text{s.e.})$$

where \(\bar{x}\) = sample mean and s.e. = standard error.

This formula is suitable for samples of around 30 or larger, where data are on the interval or ratio scale, and are normally distributed.

Note that numbers in this section are calculated to three decimal places.

To calculate a 95% confidence interval (large samples), follow the steps listed next.
1. Calculate the sample mean, the standard deviation and hence the standard error (s.e.).

2. Multiply the s.e. by 1.96, and note this result (call it result 1).

3. Add result 1 to the sample mean, and note this sum (call it sum a).

4. Take result 1 away from the sample mean, and note this sum (call it sum b).

5. The confidence interval is written as:

   \[ \text{95% c.i.} = (\text{sample mean}) ((\text{sum a}) \rightarrow (\text{sum b})). \]

Let us work through this using the diastolic blood pressure readings in Figure 12.1.

1. The sample mean is 82.696; the standard error (s.e.) is 1.116 (remember that the standard error is calculated as \(10.701/\sqrt{92}\).

2. s.e. \(\times\) 1.96 = 1.116 \(\times\) 1.96 = 2.187.

3. 82.696 + 2.187 = 84.883.

4. 82.696−2.187 = 80.509.

5. 95% c.i. is 82.696 (80.509 \rightarrow 84.883).

In the example, although the sample mean is 82.696, there is a 95% degree of certainty that the population mean lies between 80.509 and 84.883. In this case, the range is not particularly wide, indicating that the population mean is unlikely to be far away.
It should therefore be reasonably representative of patients with diabetes, so long as the sample was randomly selected. Increasing the sample size will usually result in a narrower confidence interval.

To calculate a 99% confidence interval, use 2.58 instead of 1.96 (this is the number of standard deviations which contain 99% of all the values of the normal distribution). Although a 99% confidence interval will give greater certainty, the intervals will be wider.

In the example here, we have calculated a confidence interval for a single mean, based on a fairly large sample. Confidence intervals can be calculated for other circumstances, some of which are listed as follows:

- 95% c.i. for difference between two sample means – **large samples**:

  \[(\bar{x}_1 - \bar{x}_2) \pm (1.96 \times \text{s.e.})\]

  *(see s.e. formula for comparing two sample means (large samples) in Chapter 10)*

- 95% c.i. for a single proportion \((p)\) – **large samples**:

  \[p \pm (1.96 \times \text{s.e.})\]

  *(see s.e. formula for single proportion (large samples) in Chapter 10).*

There are different formulae for calculating confidence intervals and standard error in other situations (e.g. for comparing proportions in two independent groups, where the sample size is large), and these are covered by several other texts.

- For **small samples**:

  \[\bar{x} \pm t \times \text{s.e.}\]

  *(also see Chapter 15 on t-tests).*
This page intentionally left blank
Probability is a mathematical technique for predicting outcomes. It predicts how likely it is that specific events will occur.

Probability is measured on a scale from 0 to 1.0 as shown in Figure 13.1.

For example, when one tosses a coin, there is a 50% chance of obtaining a head. Note that probabilities are usually expressed in decimal format – 50% becomes 0.5, 10% becomes 0.1 and 5% becomes 0.05. The probability of obtaining a head when a coin is tossed is therefore 0.5.

A probability can never be more than 1.0, nor can it be negative.

There is a range of methods for calculating probability for different situations.

FIGURE 13.1 The scale of probability.
TO CALCULATE THE PROBABILITY \((P)\) OF A SINGLE EVENT \((A)\) HAPPENING

For example, to find the probability of throwing a six on a die:

\[
P(A) = \frac{\text{the number of possible equally likely outcomes}}{\text{the number of possible events}}
\]

\[
P(A) = \frac{\text{the number of sixes on the die}}{\text{the number of sides on the die}}
\]

\[
P(A) = \frac{1}{6} = 0.1667 \text{ (or 16.67%)}
\]

TO CALCULATE THE PROBABILITY OF EVENT \((A)\) AND EVENT \((B)\) HAPPENING (INDEPENDENT EVENTS)

For example, if you have two identical packs of cards (pack A and pack B), what is the probability of drawing the ace of spades from both packs?

Formula: \(P(A) \times P(B)\)

\[
P(\text{pack A}) = \frac{1}{52} = 0.0192
\]

\[
P(\text{pack B}) = \frac{1}{52} = 0.0192
\]

\[
P(A) \times P(B) = 0.0192 \times 0.0192 = 0.00037
\]

This is called the rule of multiplication.

In the example, events A and B are independent of each other. This means that one event happens regardless of the other, and its outcome is not related to the other.

Sometimes probabilities are conditional, which means that one probability relies on another having already happened.

TO CALCULATE THE PROBABILITY OF EVENT \((A)\) AND EVENT \((B)\) HAPPENING (CONDITIONAL EVENTS)

What is the probability of drawing the ace of spades and the queen of clubs consecutively from a single pack of cards?

Formula: \(P(A) \times P(B | A)\)

where \((B | A)\) means

[B given that A has happened]
We already know that the probability of drawing the ace of spades from a pack of 52 cards is \( \frac{1}{52} = 0.0192 \), so \( P(A) = 0.0192 \).

The chances of now drawing the queen of clubs are a little higher, because one less card is left in the pack, so the probability \( P(B \mid A) \) is now \( \frac{1}{51} = 0.0196 \).

\[
P(A) \times P(B \mid A) = \left( \frac{1}{52} \right) \times \left( \frac{1}{51} \right) = 0.0192 \times 0.0196 = 0.0004
\]

Probabilities can be mutually exclusive. This means that one event prevents another event from happening. For example, throwing a die once will result in either a one, or a two, or a three, or a four, or a five, or a six – but only one number can be obtained. Therefore throwing a five rules out any other number. In such cases, the rule of addition is used.

**TO CALCULATE THE PROBABILITY OF EITHER EVENT (A) OR EVENT (B) HAPPENING (WHERE THE EVENTS ARE MUTUALLY EXCLUSIVE)**

For example, what is the probability of throwing either a six or a five on a die?

\[
\text{Formula: } P(A) + P(B)
\]

\[
P(A) = 0.1667
\]

\[
P(B) = 0.1667
\]

\[
P(A) + P(B) = 0.1667 + 0.1667 = 0.333 \text{ (or 33.3%)}
\]

This is called the **rule of addition** or the **additive rule**.

**TO CALCULATE THE PROBABILITY OF EITHER EVENT (A) OR EVENT (B) HAPPENING (WHERE THE EVENTS ARE NOT MUTUALLY EXCLUSIVE)**

Suppose that a local study finds that 90% of people aged over 60 years in Epitown suffer from at least one common cold during a 1-year period, and 20% suffer from heartburn at least once. What is the probability that any person over 60 years of age will suffer from either common cold or heartburn? We shall assume that common cold and heartburn occur independently of each other.

Using the rule of addition produces a probability of 0.9 + 0.2, which is equal to 1.1. This cannot be correct, since we already know that a probability can never be more than 1.0.
In this situation, we use a different formula:

\[ P(A) + P(B) - P(\text{both}) \]

\[ P(A) = 0.9 \text{ (common cold)} \]
\[ P(B) = 0.2 \text{ (heartburn)} \]
\[ P(\text{both}) = 0.9 \times 0.2 = 0.18 \]

(since we are assuming that they are independent).

So
\[ P(A) + P(B) - P(\text{both}) = (0.9 + 0.2) - 0.18 \]
\[ = 1.1 - 0.18 \]
\[ = 0.92 \text{ (or 92%)} \]

In this example, then, there is a probability of 0.92 (or 92%) that any person aged over 60 years in Epitown will suffer from either common cold or heartburn during a 1-year period.
A hypothesis is an unproved theory that is formulated as a starting point for an investigation – for example, ‘patients who take drug A will have better outcomes than those who take drug B’ or ‘drug A is better than drug B’. The hypothesis that ‘drug A is better than drug B’ is often written as $H_1$.

For every hypothesis there is a null hypothesis. In the scenarios mentioned, the null hypothesis is that ‘the outcomes of patients taking drug A will be the same as those of patients who take drug B’. Scientific experiments tend to adopt a somewhat sceptical attitude, and normally use the null hypothesis to try to disprove the real hypothesis. The null hypothesis is often written as $H_0$.

If drug A proves to be significantly better than drug B, the null hypothesis ($H_0$) is rejected, and the alternative hypothesis ($H_1$) is accepted. Hypotheses are sometimes referred to as one-tailed or two-tailed. As described in Chapter 11, the normal distribution is split in half by the mean. The proportions of values under and above a specified value (e.g. two standard deviations more than the mean) can be calculated. These are known as tails. The term one-tailed refers to the distribution either under or above a specified value. The term two-tailed refers to the whole distribution, both under and above the specified value (e.g. either two standard deviations less or two standard deviations more). In a two-tailed hypothesis, we want to find out whether there will actually be a difference between the two treatments, but we do not state which way it will go (e.g. ‘drug A will be better or worse than drug B’). In a one-tailed hypothesis, we are interested in the direction of any difference (e.g. ‘drug A is better than drug B’). The two-tailed hypothesis is usually more appropriate.

The problem is how much better does the difference or size of effect need to be in order to reach the level of statistical significance? In practice, we assess the probability that the effect we found (or a more extreme effect) would have occurred if the null hypothesis were true. If the probability is low, it follows that the effect may be due to the effectiveness of the treatment – or possibly some other cause. In order
to make this assessment, we need to calculate a test statistic and use this to determine the probability (expressed as a $P$-value). This process is called hypothesis testing.

At this point, it is useful to go back to the idea of the normal distribution and standard deviations. Remember that, in a normal distribution, 95% of all values fall within 1.96 standard deviations and 99% of them fall within 2.58 standard deviations.

If the value of a result is more than 1.96 standard deviations of the hypothetical or population mean value, its probability of occurring is less than 5%. Remembering (from Chapter 13) that probabilities are usually expressed as decimals, its probability is written as $P < 0.05$ ($< \text{means ‘less than’}$). If the value is more than 2.58 standard deviations away from the mean, its probability of occurring (if the $H_0$ is true) is less than 1%. Its probability is therefore $P < 0.01$. Probabilities of $< 0.05$ or $< 0.01$ are generally regarded as being the thresholds of statistical significance.

For many studies, a $P$-value of less than 0.05 is regarded as significant. For other more critical studies (e.g. treatment trials), significance may only be assigned when the $P$-value is $< 0.01$.

Our test statistic for comparing a sample mean with a hypothetical mean is calculated using the following relatively simple equation:

$$(\bar{x} - \mu)/\text{s.e.}$$

where $\bar{x}$ is the sample mean, $\mu$ is the hypothetical mean presumed in the $H_0$ and s.e. is the standard error of the observed value.

This test uses the normal distribution, and is thus called the normal test. It is also called the $z$-test.

**Note:** the formula here should only be used for large samples – see Chapter 15 on $t$-tests if the sample size is small.

The equation calculates the number of standard deviations that separate the hypothetical mean from the sample mean, and expresses this as something called a $z$-score (or normal score). The $z$-score is the test statistic that is used in the normal test. The larger the $z$-score, the smaller the probability of the null hypothesis being true.

The final step is to look up this $z$-score in a normal distribution table (either one-tailed or two-tailed, depending on the hypothesis) in order to obtain a $P$-value. An example of a normal distribution table for two-tailed hypotheses is provided in Appendix 1.

We know that 95% of all values under the normal distribution are contained within 1.96 standard deviations of the mean, and 99% of values are contained within 2.58 standard deviations. If the $z$-score is more than 1.96, we instantly know that the
probability is less than 5%, and its P-value will therefore be $< 0.05$. If the z-score is **more than 2.58**, the probability is less than 1%, and its P-value will therefore be $< 0.01$.

The steps for the first equation on page 46 $(\bar{x} - \mu)/s.e.$ are as follows:

1. Calculate the sample mean and standard error.
2. Subtract the hypothetical mean from the sample mean (ignore any minus values, since we are only interested in the difference between the two means).
3. Divide the result by the standard error to produce a z-score.
4. Look down each column of the normal distribution table in Appendix 1 to find your z-score, and then read across to obtain the P-value (e.g. for a z-score of 0.37, the P-value is 0.7114).

Many statistical computer programs produce P-values automatically, and it is possible that you will never actually need to calculate one.

Using the table of diastolic blood pressure readings in Chapter 12, we calculate a P-value as follows:

Suppose the **population** mean diastolic blood pressure in patients with diabetes is believed to be 84 mmHg.

1. The sample mean is 82.696 and the standard error is 1.116.
2. $82.696 - 84 = 1.304$ (ignoring the minus value).
3. $1.304/1.116 = 1.17$.
4. $z = 1.17$; in a two-tailed normal distribution table, look up 1.17 in the left-hand column, and then read across to find the P-value. The P-value = 0.2420, which is not significant. The null hypothesis (in this case, that there is no difference between the sample and the population) is **not** rejected. In fact, this sample could have come from a population with a mean blood pressure of 84 mmHg.

Now imagine that the diastolic blood pressures were taken from a group of men who have hypertension, and who have received a new antihypertensive drug in a certain clinic. We shall also assume that the population mean diastolic blood pressure in hypertensive men (whose blood pressure is either controlled or kept at a safe level by conventional drugs) aged 30–45 years who attend hypertension clinics is in fact 86 mmHg $((82.696 - 86)/1.116) = 3.304/1.116 = 2.96$.

The z-score is now 2.96. The two-tailed normal distribution table gives a P-value of 0.0031. Thus the probability of this result being obtained if the null hypothesis (that there is no difference between the treatments) were true is very low. In this case, the null hypothesis will be rejected, and the alternative hypothesis (that there **is** a difference) will be accepted. It may be concluded that this drug is either highly effective, or that the result may have been influenced by another factor. Such factors could include
problems with the sampling/randomisation process, differences between groups of patients receiving the treatments (either at the start of the study or with regard to patient management during the study) or the deliberate ‘fiddling’ of results.

It is worthwhile using a certain amount of common sense when interpreting $P$-values. A $P$-value of 0.6672 is certainly not significant, but a value of 0.0524 should not necessarily be dismissed just because it is slightly higher than the threshold. However, a $P$-value of 0.0524 will always be referred to and reported as non-significant.

A $P$-value of less than our chosen threshold of significance does not prove the null hypothesis to be true – it merely demonstrates insufficient evidence to reject it. There is always an element of uncertainty when using a $P$-value to decide whether or not to reject the null hypothesis.

When interpreting a $P$-value, two different types of possible error should be recognised:

- **type 1 error** – rejecting a true null hypothesis, and accepting a false alternative hypothesis
- **type 2 error** – not rejecting a false null hypothesis.

It is also worth remembering that a statistically significant result is not necessarily clinically significant. For example, a reduction in the mean diastolic blood pressure from 115 mmHg to 110 mmHg in a large sample of adults may well produce a $P$-value of $< 0.05$. However, a diastolic blood pressure of 110 mmHg is still well above what is considered to be a healthy level.

Although $P$-values are routinely calculated, there is a strength of feeling that confidence intervals may be a better way of testing hypotheses, since they show an estimate of where the true value actually lies. If a confidence interval does not include the hypothetical mean, this indicates significance. When reporting results, it is good practice to quote both $P$-values and confidence intervals.

There are different formulae for calculating $z$-scores in other situations (e.g. differences between proportions), and these are covered by several other texts.
CHAPTER 15

The t-tests

The previous methods of calculating confidence intervals and performing hypothesis testing are only suitable if the sample size is large. However, in some circumstances only small samples are available. For these purposes, a ‘small’ sample is usually considered to be 30 or less.

A different distribution – the \( t \)-distribution (also known as Student’s \( t \)-distribution, after WS Gossett, whose pseudonym was ‘Student’) – is used if the sample size is small. The \( t \)-distribution has a similarly shaped curve to the normal distribution, but is more widely spread out and flatter. The degree of spread and flatness changes according to the sample size. If the sample size is very large, the \( t \)-distribution becomes virtually identical to the normal distribution. The \( t \)-tests are therefore suitable for both large and small sample sizes.

For the use of a \( t \)-test to be valid, the data should be normally distributed. Although the test is described as ‘robust’, meaning that it can withstand moderate departures from normality, severely skewed data are unsuitable. For independent tests, the standard deviations should also be roughly equal.

If you are in doubt as to whether the degree of skewedness of your data violates these conditions, statistical methods exist to assess this (see Chapter 16). There are also methods of transforming skewed data to make them more ‘normal’. One alternative method for dealing with skew is to use a non-parametric test (see Chapter 17). For small samples, the Wilcoxon signed-rank test can be used instead of the paired \( t \)-test, and the Wilcoxon rank-sum test or Mann–Whitney \( U \)-test can be used instead of the independent \( t \)-test. These methods are covered by many more detailed texts.

The calculation of the \( t \)-statistic (\( t \)) is a little different to the calculation of \( z \). It takes the level of significance (e.g. 0.05, 0.01) into account, together with degrees of freedom (d.f.) which are based on sample size. Don’t worry too much about the theory behind degrees of freedom.

Degrees of freedom are calculated as follows:
\(n - 1\) for a one-sample test

where \(n\) = sample size

\((n_1 - 1) + (n_2 - 1)\) for an independent test

where \(n_1\) = sample size for group 1 and \(n_2\) = sample size for group 2.

The steps for performing a \(t\)-test are as follows:

1. Work out the standard error and \(t\)-statistic for the required test.
2. Calculate the appropriate d.f.
3. Using the \(t\)-distribution table (see Appendix 1), look up the d.f. value in the left-hand column.
4. Read across this row, until the nearest values to the left and right of your \(t\)-statistic can be seen.
5. Your \(P\)-value will be less than the \(P\)-value at the top of the column to the left of your \(t\)-statistic and greater than the \(P\)-value at the top of the column to its right (e.g. a \(t\)-statistic of 2.687 with 6 d.f. falls in between 2.447 and 3.143. The nearest value to its left is 2.447; the \(P\)-value at the top of this column is 0.05. The \(P\)-value for your \(t\)-statistic will therefore be less than 0.05, and is written \(P < 0.05\). If your \(t\)-statistic is 1.325 with 6 d.f., there is no column to its left, so the \(P\)-value will be greater than the column to its right, and is therefore \(> 0.2\)).

There are a number of different \(t\)-test formulae which are used in different situations, described as follows:

**ONE-SAMPLE \(T\)-TEST**

This test compares a sample mean with a population mean.

\[ t = (\bar{x} - \mu)/\text{s.e.} \]

where \(\bar{x}\) = sample mean, \(\mu\) = population mean and s.e. = standard error of sample mean.

\[ \text{d.f.} = n - 1 \]

where \(n\) = sample size.
s.e. = \frac{s}{\sqrt{n}}

where \( s \) = standard deviation of sample mean and \( n \) = sample size.

**95% Confidence intervals – one-sample t-test**

\[ \bar{x} \pm t_{0.05} \times \text{s.e.} \]

where \( t_{0.05} \) = value on \( t \)-distribution table in 0.05 column (two-tailed), corresponding to appropriate d.f.

For example, suppose that a group of 14 GP surgeries is running healthy eating groups to help patients to lose weight. At the start, each patient has their height measured and is weighed, and their body mass index (BMI) is calculated. The mean BMI is roughly the same for patients at each GP surgery. After 6 months, each patient is weighed and their BMI is recorded again. One surgery is interested to find out how successful its patients have been in losing weight, compared with the whole group. The BMI values of its patients are shown in Figure 15.1.

<table>
<thead>
<tr>
<th>BMI value</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>21</td>
<td>1</td>
</tr>
<tr>
<td>22</td>
<td>1</td>
</tr>
<tr>
<td>26</td>
<td>1</td>
</tr>
<tr>
<td>29</td>
<td>1</td>
</tr>
<tr>
<td>30</td>
<td>2</td>
</tr>
<tr>
<td>31</td>
<td>1</td>
</tr>
<tr>
<td>32</td>
<td>1</td>
</tr>
<tr>
<td>33</td>
<td>1</td>
</tr>
<tr>
<td>35</td>
<td>1</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>10</strong></td>
</tr>
</tbody>
</table>

Mean = 28.9
SD = 4.581

**FIGURE 15.1** Frequency distribution of BMI from a sample of patients in primary care.

The mean BMI for the 14 surgeries as a whole is 26.2 (this is a precisely known population value), compared with 28.9 for this surgery. It looks as if this surgery’s patients have been less successful, but has their performance been significantly different? Let us find out, by performing a one-sample \( t \)-test.

The steps are as follows:

1. Work out the standard error \( (n = 10; s = 4.581; \sqrt{10} = 3.162) : 4.581/3.162 = 1.449 \). The sample mean minus the population mean = 28.9 \(-\) 26.2 = 2.7. To work out the \( t \)-statistic: 2.7/1.449 = 1.863 (to three decimal places here).
2. Calculate the degrees of freedom (d.f.): 10 – 1 = 9.

3–5. Using the \( t \)-distribution table, look up d.f. = 9, and then read across this row. Our \( t \)-statistic is in between 1.833 and 2.262. Reading up the columns for these two values shows that the corresponding two-tailed \( P \)-value is less than 0.1 but greater than 0.05, and is therefore not significant.

The null hypothesis (in this case, that there is no difference between the BMI values in this GP surgery and the group as a whole) is not rejected.

To calculate a 95% confidence interval, the steps are as follows:

1. Note the sample mean, standard error and degrees of freedom.
2. Find the value in the two-tailed \( t \)-distribution table in the 0.05 column, corresponding to the degrees of freedom.
3. Multiply this value by the standard error, and note the result (call it \( \text{result 1} \)).
4. Add \( \text{result 1} \) to the mean, and note this sum (call it \( \text{sum a} \)).
5. Subtract \( \text{result 1} \) from the mean, and note this sum (call it \( \text{sum b} \)).
6. The confidence interval is written as:

\[
95\% \text{ c.i.} = (\text{sample mean}) \rightarrow (\text{sum a}) \rightarrow (\text{sum b}).
\]

Using the mentioned example, the steps are as follows:

1. The sample mean is 28.9, the standard error is 1.449 and there are 9 degrees of freedom.
2. In the \( t \)-distribution table in Appendix 1, find degrees of freedom = 9, and then read along the line until you come to the 0.05 column – the value is 2.262.
3. Multiply 2.262 by the standard error (2.262 \times 1.449 = 3.278) (\( \text{result 1} \)).
4. \( 28.9 + 3.278 = 32.178 \) (\( \text{sum a} \)).
5. \( 28.9 – 3.278 = 25.622 \) (\( \text{sum b} \)).
6. \( 95\% \text{ c.i.} = 28.9 \rightarrow 32.178 \).

Note that the confidence interval includes the mean of the group as a whole (26.2). This supports the null hypothesis that there is no difference between the BMI values.

**PAIRED (ALSO CALLED THE DEPENDENT) \( t \)-TEST**

This test is used to assess the difference between two paired measurements. It tests the null hypothesis that the mean of the difference is zero. In this case, data are naturally paired or matched (e.g. weight measurements from the same subjects at a 6-month interval or data relative to twins or couples).
The value that we analyse for each pair is the *difference* between the two measurements.

\[ t = \bar{x} / \text{s.e.} \]

where \( \bar{x} \) = mean of the differences and s.e. = standard error of the differences.

\[ \text{d.f.} = n - 1 \]

where \( n \) = sample size.

\[ \text{s.e.} = s / \sqrt{n} \]

where \( s \) = standard deviation of the differences and \( n \) = sample size.

**95% Confidence intervals – paired t-test**

\[ \bar{x} \pm t_{0.05} \times \text{s.e.} \]

where \( t_{0.05} \) = value on \( t \)-distribution table in 0.05 column (two-tailed), corresponding to appropriate d.f.

**INDEPENDENT (ALSO CALLED THE TWO-SAMPLE OR UNPAIRED) T-TEST**

This is used where data are collected from groups which are unrelated (or independent), such as the length at 1 year of a group of infants who were breastfed, compared with a group who were not breastfed.

\[ t = (\bar{x}_1 - \bar{x}_2) / \text{s.e. pooled} \]

where \( \bar{x}_1 \) = mean from group 1 and \( \bar{x}_2 \) = mean from group 2.

\[ \text{d.f.} = (n_1 - 1) + (n_2 - 1) \]

where \( n_1 \) = sample size for group 1 and \( n_2 \) = sample size for group 2.

\[ \text{s.e. pooled} = \text{see following.} \]
Calculating standard deviation and standard error for the independent t-test

If the standard deviations are not appreciably different, use the ‘pooled’ standard error:

$$\text{s.e. pooled} = \sqrt{\frac{s_{\text{pooled}}^2}{n_1} + \frac{s_{\text{pooled}}^2}{n_2}}$$

where $s_{\text{pooled}}$ is calculated in the formula following, $n_1 =$ sample size 1 and $n_2 =$ sample size 2.

To calculate a ‘pooled’ standard deviation:

$$s_{\text{pooled}} = \sqrt{\frac{[s_1^2(n_1 - 1)] + [s_2^2(n_2 - 1)]}{(n_1 + n_2) - 2}}$$

where $s_1 =$ standard deviation 1, $s_2 =$ standard deviation 2, $n_1 =$ sample size 1 and $n_2 =$ sample size 2.

If the standard deviations and/or sample sizes are appreciably different, it is advisable to consult a statistician or someone with advanced statistical skills.

95% Confidence intervals – independent t-test

$$\bar{x} + t_{0.05} \times \text{s.e. pooled}$$

where $t_{0.05} =$ value on $t$-distribution table in 0.05 column (two-tailed), corresponding to appropriate d.f.
Once your data have been collected, it is natural to want to get on with the job of analysis. Before going any further however, it is absolutely essential to check the data thoroughly. The process of data checking can be tiresome, but ignoring it may lead to your drawing the wrong conclusions. This chapter provides a brief overview of some of the issues to be considered.

**PREVENTION IS BETTER THAN CURE**

The best advice is to think carefully about your data before you get anywhere near the analysis stage – both in the planning of a study and during data collection. Piloting your data collection instrument before the study begins can help to minimise ‘bugs’ and possible misunderstandings by both participants and researchers. Thinking in advance about potential biases and problems can help to reduce data errors – this can save considerable time and distress later on (see also Chapter 28 on Questionnaires). Data entry can be tedious and requires concentration and alertness, so taking care to enter data accurately can also pay dividends. You could check a sample of data as you go along, or even ask someone else to check some or all of what has been entered before doing your analysis (discussed further below).

When reviewing data, it is helpful to ask yourself the following questions:

**HAVE THE DATA BEEN ENTERED ACCURATELY?**

This includes checking that the data have been accurately recorded and entered. It is helpful to ‘eyeball’ the data to check that the data set looks right. Typing errors are common and easily made, so it is good practice to check that data have been entered correctly by comparing the records used for data entry with what appears in your database. Some software packages allow automated checking of data validity. Of
course, any changes to your data should only be made where an actual mistake has been identified.

**ARE THERE ANY MISSING DATA?**

In the real world, it is difficult to achieve 100% completeness. For example, some participants may refuse to answer certain questions or fail to complete all fields in a questionnaire, others may leave a study for various reasons or some practices may not keep records of every variable you want to collect data for.

Where data are incomplete, you need to decide how best to act. It is possible (though unlikely) that you could attempt to go back to collect the missing data; this will be impossible if your subjects were anonymous, but may be feasible under some circumstances and if time allows. If the missing data have arisen from errors in data entry, this should be straightforward to correct as described previously. However, if this cannot be done, you could continue with one of the following options:

- Analyse what you have anyway. This may be acceptable if relatively small amounts of data are missing, but large quantities of missing data could seriously undermine the reliability of your results. Be aware that sample size would be affected. You would need to report the fact that data were missing, and discuss how this may affect your results.

- Exclude the incomplete variable(s) from your analysis. If the variables concerned are not very important to your analysis and not central to answering a research question, this may be a viable option. Otherwise, the missing data may present a major problem that could ruin your whole study. If this is the case, you could consider estimating missing values – see the next point.

- Estimate the missing values. This may be possible using various techniques such as dummy variables or applying mean values or other methods that estimate or impute missing values. These should always be used with care, preferably with the help of a statistician, and are not covered by this basic guide. When estimated values have been substituted for missing data, it is a good idea to carry out separate analyses on the variable both with missing data and with substituted data – in this case, both results should be reported, with discussion on the differences between results, where this is relevant.

If you draft a report using incomplete data and add further entries later, it is important to check the draft against your final analysis – you otherwise risk inconsistencies and errors in your report (Smeeton & Goda, 2003).
ARE THERE ANY OUTLIERS?
These are values that are either extremely high or extremely low. We have already seen in Chapters 7 and 9 that such extreme values can lead to misleading results, so your data set should be carefully checked for outliers. These may arise from errors in data provided by study participants or from data entry mistakes – or an extreme value could be real. Sometimes, an outlier can affect more than one variable. For example, if an extreme weight value was recorded, and weight will be used to calculate a separate BMI variable.

To check for outliers, you can either look at the range of a variable (the lowest and highest values) or produce a graph showing all the values, such as a histogram, for checking the range, or scatterplots for looking at the relationship between two values, e.g. weight and height, to see if they appear to be consistent with each other.

Some outliers are obviously erroneous (e.g. a human age of 240 years is definitely wrong), while others could actually be genuine (e.g. a male weight of 240 kg is very heavy, but possible). In either case, it is advisable to go back to the original data and check whether the outlier appears real. If you are sure that an error has been made and can identify the correct value, you can amend it. Careful judgement must be used when doing this, however, as it would clearly be inappropriate and unethical to delete or change a value just because it seemed wrong.

If you are unsure about whether to delete an outlier, you could (as with missing data, discussed earlier) carry out two analyses – one with the outlier left in and another with it deleted, to see how large an effect this has on your results. If the results are very different, you should consider employing more advanced statistical methods such as transformation and non-parametric tests to deal with this (Petrie & Sabin, 2009). In the latter case, it is advisable to seek expert statistical advice.

ARE THE DATA NORMALLY DISTRIBUTED?
Statistical tests make ‘assumptions’ (or have requirements) about the kind of data they can be used with, and often require that the data are normally distributed. For example, we have seen an assumption for using t-tests is that the data are normally distributed.

If data are normally distributed, we can use parametric statistical tests (such as a t-test) to analyse the data (note – there may still be unsatisfied assumptions that invalidate them, e.g. unequal variances in an independent samples t-test). For data that are not normally distributed, there are various broadly comparable techniques called non-parametric tests – for example, the Wilcoxon signed-rank test is a non-parametric equivalent of the paired t-test. There is some more detail on this in the next chapter.

The problem is that non-parametric tests are less likely to show statistical significance when there is a real difference – the risk of a type 2 error is usually greater with a non-parametric test, so technically they tend to be less powerful. Also, parametric
tests and their non-parametric equivalents do not always test the same hypothesis (e.g. paired $t$-tests test for equal means, while Wilcoxon signed-rank tests for equal medians). It is therefore always better to use parametric tests if possible.

As part of the process of screening data before carrying out analysis, we can use tests such as the Kolmogorov–Smirnov or the Shapiro–Wilks to find out whether the data are normally distributed.

When we have data that are not normally distributed, we can try transforming the data – this is done in an attempt to ‘normalise’ them (i.e. transform them into normally distributed data), so that we can use a parametric test. A commonly used transformation is the logarithmic/log$_{10}$, though there is a range of others that can be used. Transformation may or may not succeed in normalising the data. If we transform a variable, and it is then identified as ‘normally distributed’, we can more safely use a parametric test to analyse it. If the transformation does not normalise the data, we should use an appropriate non-parametric test instead.

Although you can visually inspect the data, for example, by using a histogram (Petrie & Sabin, 2009), to check whether it resembles the symmetrical bell-shaped pattern described in Chapter 11, normality is often checked using one of two just previously mentioned tests:

- **Kolmogorov–Smirnov** – for large samples (e.g. 50 or more)
- **Shapiro–Wilks** – best for sample sizes of less than 50.

When using these tests, the null hypothesis is that the distribution is normally distributed.

This means that:

- if $P < 0.05$, we reject the null hypothesis and conclude that the data are not normally distributed
- if $P \geq 0.05$, the data are not significantly non-normal, so may be assumed normally distributed.

The Q–Q plot (abbreviation of ‘quantile–quantile’ plot, produced by some programs) can also be used to check normality. If the data are normally distributed, the dots should fall along the straight line on the plot.

If our statistical testing will involve comparing groups, then the data for each group should be checked for normality. Some care needs to be taken when using these tests of normality, as they can be unreliable under certain circumstances. It is therefore advisable to also use Q–Q plots when interpreting them (Field, 2013).

Transformation and normality tests are performed using computer programs, and instructions for carrying them out differ between various packages.
Let’s now look at two examples of normality tests with abbreviated outputs produced by SPSS Statistics software.

First, we are going to check the normality of systolic blood pressure readings from a group of patients, which have been entered onto a database. The following output is produced:

<table>
<thead>
<tr>
<th>Tests of normality</th>
<th>Kolmogorov–Smirnov</th>
<th>Shapiro–Wilk</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Statistic</td>
<td>df</td>
</tr>
<tr>
<td>Systolic blood pressure (mmHg)</td>
<td>.123</td>
<td>39</td>
</tr>
</tbody>
</table>

A total of 39 systolic readings are recorded (a ‘small’ sample size), so we will use the Shapiro–Wilk test.

We can see that the \( P \)-value (shown as ‘Sig.’ in the table) is 0.021 – this is < 0.05, so indicates that systolic blood pressure is \textbf{not} normally distributed.

Looking at the Q–Q plot shown in Figure 16.1, we can see that the dots are \textbf{not} arranged along the straight line, which confirms that systolic blood pressure is not normally distributed. Although they may seem to be quite close to the line, the data
swing a few points under the line (the lowest observed values), then several over, then
do another run under. For normality, the points should be close to the line.

For the second example, we will use the database of Warwick-Edinburgh Mental Well-being Scale (WEMWBS) scores for mental well-being used in Chapter 22 on effect size. A total of 60 scores are recorded for patients receiving the ‘new therapy’ and the output looks like this:

Tests of normality

<table>
<thead>
<tr>
<th></th>
<th>Kolmogorov–Smirnov</th>
<th>Shapiro–Wilk</th>
</tr>
</thead>
<tbody>
<tr>
<td>Statistic</td>
<td>df</td>
<td>Sig.</td>
</tr>
<tr>
<td>New Therapy WEMWBS</td>
<td>.085</td>
<td>60</td>
</tr>
</tbody>
</table>

A total of 60 WEMWBS scores are recorded (a ‘large’ sample size), so this time we will use the Kolmogorov–Smirnov test.

This time, we can see that the $P$-value is 0.200 – this is > 0.05, so there is no reason to reject the assumption that these scores are normally distributed.

Looking at the Q–Q plot shown in Figure 16.2, we can see that the dots are generally arranged along the straight line (much more closely than in the previous example), which suggests that these baseline WEMWBS scores are normally distributed.

**FIGURE 16.2** Normal Q–Q plot of baseline.
THE CENTRAL LIMIT THEOREM

So far in this chapter, we have discussed the importance of checking your data for normality using formal methods (such as the Kolmogorov–Smirnov or Shapiro–Wilk tests and Q–Q plots) as a prerequisite for using parametric statistical techniques.

Having read this, it may surprise you to know that something called the central limit theorem says that as long as your sample size is reasonably large (say 20 or more), you can probably proceed as if the distribution is normal.

The central limit theorem was first proposed over 200 years ago, and is about what happens to the sample mean ‘in the limit’ when the sample size heads off towards infinity. It has since been shown that it is practically true for means of fairly small samples (in practice, sometimes even < 10) if the shape of the distribution for the individual measurements is not normal but is in some sense ‘reasonable’. In practice therefore, even if a test of normality such as Shapiro–Wilk shows that the data for a sample of say 20 or 30 are significantly non-normal, the sampling distribution of the mean will be very close to normal, and hence most standard parametric tests will be valid, unless the distribution is very skewed. This will, however, not be true of tests that do not relate to the mean.

This does not suggest that you should ignore parametric assumptions and testing for normality, but the central limit theorem should be borne in mind if, for example, you only have a small data set and wish to use a parametric analysis technique.
People often ask about the difference between parametric and non-parametric tests. We introduced the concept of parameters early in the book – these are measures of a population, rather than of a sample. Used in this context, the term refers to the ‘population’ of the normal distribution. Parametric tests are performed if a normal distribution can be assumed. Remember that the $t$-tests also require an underlying normal distribution.

However, if the data are clearly not normally distributed, non-parametric tests can be used. These are also known as distribution-free tests, and they include the following:

- Wilcoxon signed-rank test – replaces the paired $t$-test
- Mann–Whitney $U$-test or Wilcoxon rank-sum test – replaces the independent $t$-test
- Chi-squared ($\chi^2$) test – for categorical data
- Spearman’s (Spearman’s $\rho$ or rho) rank correlation coefficient – replaces Pearson’s product moment correlation coefficient
- Kendall’s $\tau$ (or tau) – alternative to Spearman’s rho (bullet point 4)
- Kruskal–Wallis test – replaces one-way analysis of variance (ANOVA).

The Chi-squared test is described in Chapter 20. The other tests are covered by several other statistical textbooks (see Further reading).
This page intentionally left blank
Correlation and linear regression

Various statistical methods exist for investigating association between variables. In the next two chapters, we will be looking at the Chi-squared ($\chi^2$) test (used for investigating the presence of an association between categorical variables), as well as briefly outlining multiple regression, logistic regression, and analysis of variance (ANOVA). This chapter, however, concentrates on methods for assessing possible association, mainly between continuous variables.

**CORRELATION**

Correlation assesses the strength of association between variables (usually interval or ratio), and linear regression allows us to use one variable to predict another.

Let’s have a look at how we can put this into practice. A rheumatologist measures and records the bone mineral density (BMD) in a group of women. She has a hypothesis that BMD decreases with age, and decides to use correlation and linear regression to explore this.

Correlation is measured using a *correlation coefficient* ($r$), which can take any value between $-1$ and $+1$. If $r = +1$, there is a **perfect positive correlation**; if $r = -1$ there is a **perfect negative correlation**; a value of $r = 0$ represents **no linear correlation** (we will discuss what is meant by linear in a moment). It follows that if $r$ is more than 0 but less than $+1$, there is **imperfect positive correlation**, and if $r$ is more than $-1$ but less than 0, there is **imperfect negative correlation**. If we plot the age and BMD data on a scatterplot (see page 11 for more information on scatterplots), the shape that the dots form will give us a clue about the relationship between age and BMD. If, for example, there is a **perfect positive correlation**, BMD increases with age ($r = +1$), and the scatterplot will appear as shown in Figure 18.1.

Each dot on the graph represents an individual’s age (shown on the $x$-axis) and their BMD value (on the $y$-axis). Note that age is the independent variable (since our hypothesis is that BMD depends on age, age is independent of any influence from
BMD), and is thus placed on the $x$-axis. This makes BMD the dependent variable, which is placed on the $y$-axis. You can see that the dots form a straight line, showing a **linear relationship** between the two variables.

**FIGURE 18.1** Scatterplot showing a perfect positive correlation between age and BMD.

If, on the other hand, there is **perfect negative correlation**, BMD decreases with age, $r = -1$, and the scatterplot will look as it does in Figure 18.2.

**FIGURE 18.2** Scatterplot showing a perfect negative correlation between age and BMD.

If there is no linear correlation at all between age and BMD ($r = 0$), the scatterplot may resemble that in Figure 18.3. In this figure, you can see that there is no discernible linear relationship between age and BMD.
There might be an imperfect correlation (either positive or negative). Figure 18.4 shows an imperfect positive correlation, where BMD increases with age, but where \( r \) is somewhere between 0 and +1 (quite close to +1, in fact). A fairly strong and clear linear relationship can be seen, but the dots do not lie in a straight line, as in perfect correlation.

There could also be an imperfect negative correlation, as can be seen in Figure 18.5. In this case, \( r \) would be quite close to \(-1\).
Finally, there may be a **non-linear relationship**, one example of which is shown in Figure 18.6.

Correlation is often calculated using the Pearson's product moment correlation coefficient (commonly known as **Pearson's \( \rho \)**), the formula for which is:

\[
 r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \sum (y - \bar{y})^2}}
\]

where: \( x = \text{individual exposure} \quad \bar{x} = \text{mean exposure} \)
\( y = \text{individual outcome} \quad \bar{y} = \text{mean outcome} \)
Do not worry if this equation looks complicated! All of the calculations can easily be done by computer, so you should never need to work this out by hand. It is important, however, that you understand some of the theory behind this process, and know how to interpret the computer outputs that are generated.

This formula should only be used when:

- there is no clear **non-linear** relationship between the variables
- only one value is recorded for each patient (e.g. observations are independent, **not** paired – *see* paired *t*-test in Chapter 15).

Coming back to our example, let's use Pearson's product moment correlation coefficient to find the strength of association between age and BMD. The data collected by our consultant rheumatologist are shown in Table 18.1.

**TABLE 18.1** Age and BMD data for 10 patients

<table>
<thead>
<tr>
<th>Age</th>
<th>BMD</th>
</tr>
</thead>
<tbody>
<tr>
<td>46</td>
<td>1.112</td>
</tr>
<tr>
<td>49</td>
<td>0.916</td>
</tr>
<tr>
<td>52</td>
<td>0.989</td>
</tr>
<tr>
<td>56</td>
<td>0.823</td>
</tr>
<tr>
<td>58</td>
<td>0.715</td>
</tr>
<tr>
<td>60</td>
<td>0.817</td>
</tr>
<tr>
<td>64</td>
<td>0.834</td>
</tr>
<tr>
<td>68</td>
<td>0.726</td>
</tr>
<tr>
<td>75</td>
<td>0.654</td>
</tr>
<tr>
<td>79</td>
<td>0.612</td>
</tr>
</tbody>
</table>

In real life we would hope to use a much larger sample than 10 patients, but we will just regard this as an example to illustrate the techniques we are studying. First of all, let us plot the data (Figure 18.7).
Note that each axis in the figure has been broken using two parallel lines, to denote that the scale does not begin at 0. The shape of the dots on the scatterplot shows an imperfect negative correlation between age and BMD (compare this with Figure 18.5). BMD does indeed appear to generally decrease with age. This alone, however, is not enough to demonstrate a correlation and test significance – for this we need to calculate Pearson’s product moment correlation coefficient.

When the data are analysed using a computer program (in this case, SPSS Statistics version 22.0 (IBM Corporation, 2013)), the following output is produced (other computer programs may produce different looking outputs, but the results are equivalent):

<table>
<thead>
<tr>
<th>Correlations</th>
<th>Age</th>
<th>BMD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>−.891**</td>
</tr>
<tr>
<td>Age</td>
<td>Sig. (2-tailed)</td>
<td>.001</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Age Pearson correlation</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Sig. (2-tailed) correlation</td>
<td>.001</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>10</td>
</tr>
</tbody>
</table>

**Correlation is significant at the 0.01 level**
Although the output does not specifically include the symbol ‘r’, it shows that the Pearson correlation (coefficient) is –0.891, and that the two-tailed significance (P-value) is 0.001. Don’t worry about the fact that both of these figures seem to be shown twice (one for age and BMD, the other for BMD and age) on the output. The correlation coefficient (r) is –0.891, which is more than –1, but less than 0. The figure of –0.891 is quite close to the maximum value of –1, and therefore indicates a relatively strong correlation between age and BMD.

When assessing the strength of an association using r, 0 to 0.19 is regarded as very weak, 0.2 to 0.39 weak, 0.40 to 0.59 moderate, 0.6 to 0.79 strong and 0.8 to 1 very strong (Swinscow & Campbell, 2002). These values can be plus or minus. These labels are useful, though somewhat arbitrary. Our value of –0.891 would therefore be regarded as ‘very strong’.

The figure ‘Sig.’ represents the P-value of 0.001, indicating a significant correlation. We can therefore conclude that there is a significant negative correlation between age and BMD in women, and can accept the consultant rheumatologist’s hypothesis that BMD decreases as age increases.

We can also calculate $r^2$ – this indicates how much variation in one variable can be explained by the other. If we square r, we get $-0.891 \times -0.891 = 0.794$. This means that age is responsible for 0.794 (or 79.4%) of the total variation in BMD. This does not mean, however, that age causes the variation in BMD (the subject of causality is discussed in Chapter 26).

If Pearson’s product moment correlation coefficient cannot be used (e.g. if there is no clear linear relationship – see previously listed criteria), it might be appropriate to employ Spearman’s rank correlation coefficient. This is the non-parametric version of Pearson’s product moment correlation coefficient, and is also called Spearman’s ρ or Spearman’s rho. It can be used when any of the following apply:

- there is a small sample size
- there is no clear linear relationship between the variables
- one or both variables are ordinal.

We have a small sample size in our age and BMD study, so in this case it is also appropriate to use the Spearman’s rank correlation coefficient. When calculated, the following SPSS output is produced:
Correlations

<table>
<thead>
<tr>
<th></th>
<th>Age</th>
<th>BMD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spearman’s rho</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age Correlation coefficient</td>
<td>1.000</td>
<td>-0.867**</td>
</tr>
<tr>
<td>Sig. (2-tailed)</td>
<td>.</td>
<td>.001</td>
</tr>
<tr>
<td>N</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>BMD Correlation coefficient</td>
<td>-.867**</td>
<td>1.000</td>
</tr>
<tr>
<td>Sig. (2-tailed)</td>
<td>.001</td>
<td>.</td>
</tr>
<tr>
<td>N</td>
<td>10</td>
<td>10</td>
</tr>
</tbody>
</table>

**Correlation is significant at the 0.01 level (2-tailed)**

This shows that although the correlation coefficient is slightly smaller than when using Pearson’s product moment correlation coefficient (−0.867 compared to −0.891), the result is still significant.

Kendall’s τ (also called Kendall’s tau) can be used as an alternative to Spearman’s rank correlation coefficient. This is covered in other texts – see Further reading.

So we have demonstrated the presence of a strong (and statistically significant) correlation between age and BMD in women.

LINEAR REGRESSION

As mentioned at the start of the chapter, we can also use linear regression to predict the value of BMD for any specific age. This is achieved by calculating a straight line that best fits the association. This line is called the linear regression line. The line describes how much the value of one variable changes when the other variable increases or decreases.

Linear regression should only be used when all of the following assumptions apply:

- the observations are independent
- an imperfect linear relationship exists between x and y
- the value of y is normally distributed, for any value of x
- the size of the scatter of the points around the line is the same throughout the length of the line.

(In practice, the last two assumptions are difficult to determine; a statistician should be consulted if there is any doubt.)

The formula for the regression line is:

\[
y = a + bx
\]
These letters represent the following:

\( y \) = the variable on the y-axis  
\( x \) = the variable on the x-axis  
\( a \) = the intercept or constant (the value of \( y \) when \( x = 0 \))  
\( b \) = the gradient of the line (the amount that \( y \) increases when \( x \) is increased by one unit).

In fact, \( a \) and \( b \) are also known as the regression coefficients, and \( b \) is sometimes labelled \( B \) or \( \beta \).

Going back to our example, we already know that \( y = \text{BMD} \) and \( x = \text{AGE} \). So the equation \( y = a + bx \) effectively says that: \( \text{BMD} = a + (b \times \text{AGE}) \)

All we need to know now are the regression coefficients, \( a \) and \( b \).

We will use a computer program for our calculations. When a linear regression is performed using SPSS, a fairly lengthy output is produced, including the following table:

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized coefficients</th>
<th>Standardized coefficients</th>
<th>t</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>Std. error</td>
<td>Beta</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>(Constant)</td>
<td>1.588</td>
<td>.140</td>
<td>11.331</td>
</tr>
<tr>
<td></td>
<td>Age</td>
<td>-.013</td>
<td>.002</td>
<td>-5.559</td>
</tr>
</tbody>
</table>

The coefficients have not actually been labelled as ‘\( a \)’ and ‘\( b \)’ in the table, so arrows indicating the regression coefficients, \( a \) and \( b \) have been added for clarity. There is no need for us to deal with the items of information that have been covered over in grey, though other textbooks discuss these in detail. We shall concentrate on the column labelled ‘B’. As mentioned earlier, \( a \) is also known as the ‘constant’, which is shown in the table as 1.588. The other coefficient, \( b \) (labelled ‘Age’), has a value of –0.013.

Our equation can now be completed:

\[ y = a + bx \]

i.e. \( \text{BMD} = a + (b \times \text{AGE}) \)

i.e. \( \text{BMD} = 1.588 + (-0.013 \times \text{AGE}) \)

i.e. \( \text{BMD} = 1.588 - (0.013 \times \text{AGE}) \)
Imagine that we would like to predict the expected BMD for a woman aged 50. This could be calculated by inserting ‘50’ for the age value:

\[
\text{BMD} = 1.588 + (-0.013 \times 50) \\
\text{i.e. BMD} = 1.588 + -0.65 \\
\text{i.e. BMD} = 1.588 - 0.65 \\
\text{BMD} = 0.938
\]

So an average woman aged 50 would have a predicted BMD of 0.938. We can easily do the same for a woman aged 60:

\[
\text{BMD} = 1.588 + (-0.013 \times 60) \\
\text{i.e. BMD} = 1.588 + -0.78 \\
\text{i.e. BMD} = 1.588 - 0.78 \\
\text{BMD} = 0.808
\]

An average woman aged 60 would therefore have a predicted BMD of 0.808. And again for a woman aged 70:

\[
\text{BMD} = 1.588 + (-0.013 \times 70) \\
\text{i.e. BMD} = 1.588 + -0.91 \\
\text{i.e. BMD} = 1.588 - 0.91 \\
\text{BMD} = 0.678
\]

**FIGURE 18.8** Scatterplot for age and BMD data, showing regression line.
An average woman aged 70 would therefore have a predicted BMD of 0.678. The predicted BMD values for these ages are summarised in Table 18.2.

**TABLE 18.2** Predicted BMD values for ages 50, 60 and 70

<table>
<thead>
<tr>
<th>Age</th>
<th>Predicted BMD</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>0.938</td>
</tr>
<tr>
<td>60</td>
<td>0.808</td>
</tr>
<tr>
<td>70</td>
<td>0.678</td>
</tr>
</tbody>
</table>

Going back to our scatterplot, we can plot the three predicted BMD values, and join them up to show the regression line (Figure 18.8).

You may have noticed that the line does not actually go through any of the observed points.

For each of the three predicted values, dotted lines have been drawn upwards from age, then across from BMD value. An ‘×’ is marked where each intersects. A line has then been drawn through the three ×s, to form the **regression line**.

Linear regression is therefore a useful technique, which allows us to use one value to predict another.
CHAPTER 19

Analysis of variance and some other types of regression

When looking at $z$- and $t$-tests in earlier chapters, we were limited to comparing only one mean value with another. However, it is often useful to examine differences between more than two means. For example, we may want to examine whether the weight of infants at 1 year of age is influenced by any of six types of milk they have received since birth. The object of the study is to find out which type of milk will produce the greatest weight gain. If each type of milk is called formula 1, 2, 3, 4 and 5, and breast milk, a total of 15 comparisons of mean weight are possible:

<table>
<thead>
<tr>
<th>TABLE 19.1 Possible combinations for five different types of formula milk compared with breast milk</th>
</tr>
</thead>
<tbody>
<tr>
<td>Formula 1; formula 2</td>
</tr>
<tr>
<td>Formula 1; formula 3</td>
</tr>
<tr>
<td>Formula 1; formula 4</td>
</tr>
</tbody>
</table>

Performing a separate $z$- or $t$-test for every possible combination would therefore require 15 separate tests. Besides being very time-consuming, such repeated testing is likely to produce statistically significant results which are misleading. A $P$-value of 0.05 or less would be expected from 5% (1 in 20) of all tests performed when there are no real differences (Kirkwood, 1988). This probability is increased if repeated tests are performed. With 15 tests, the chance of getting at least one wrong conclusion is therefore more than 50%. In other words, we have a greater risk of making a type 1 error – rejecting at least one true null hypothesis, and accepting a false alternative hypothesis.
A technique called **analysis of variance** or **ANOVA** allows several groups to be compared in one **single** statistical test, and indicates whether any significant differences exist between them.

The mentioned example compares mean weight at 1 year of age with the six groups (type of milk used). In other words, the numerical outcome variable (weight) is being compared to **one** categorical exposure group (type of milk). In this situation, **one-way ANOVA** can be used.

Where **two or more** categorical exposure groups need to be included (e.g. type of milk and ethnic group), then **two-way ANOVA** needs to be used. Details of two-way and other types of ANOVA are not covered by this basic guide, but are discussed in other texts – see Further reading. This chapter will therefore concentrate on **one-way ANOVA**.

The calculation of one-way ANOVA is normally carried out using a computer program. It assumes that data in each group are normally distributed, with equal standard deviations. This can be checked using techniques such as **Levene’s test** (it can often be carried out by programs at the same time as one-way ANOVA). If the assumptions are not met, the **non-parametric** version of one-way ANOVA – the **Kruskal–Wallis test** – should be used instead.

One-way ANOVA compares the variance (this is the square of the standard deviation – see Chapter 9) of the means **between the groups** with the variance of the subjects **within the groups**, and uses the F-test to check for differences between these two variances. The null hypothesis is that variances between the groups are due to chance, and hence the outcome is **not** influenced by differences between the exposure categories. A **P-value** of < 0.05 would suggest that the outcome **is** influenced by differences between the exposure categories.

In the example, a **P-value** of < 0.05 would indicate that weight at 1 year of age was significantly influenced by the type of milk used. Unfortunately, ANOVA does not tell us **which** type of milk produced the greatest weight gain – we would need to go back to the data and check the mean weight achieved for each group.

Let’s try using one-way ANOVA with the help of a computer program. We have an electronic database containing the BMI (body mass index – a measurement of obesity) values of 433 patients living in a town which is made up of five localities – A, B, C, D and E – with differing levels of social deprivation. We are interested in finding out whether BMI levels are influenced by which locality people live in. In other words, whether people living in the most deprived localities are more likely to be obese.

The computer program is likely to require you to define which variable is **dependent** and which is **fixed**. In this case, we are hypothesising that people’s BMI may be influenced by (or depend on) the locality in which they reside, so BMI is the **dependent variable**, and locality is ‘fixed’ – this is called the **factor**.

When a one-way ANOVA is performed using SPSS, an output is produced
including the tables following. These have been edited for simplicity. There is no need for us to deal with any items of information that have been covered in grey, though other textbooks discuss these in detail.

<table>
<thead>
<tr>
<th>Locality</th>
<th>Mean</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>26.1</td>
<td>71</td>
</tr>
<tr>
<td>B</td>
<td>26.9</td>
<td>88</td>
</tr>
<tr>
<td>C</td>
<td>29.6</td>
<td>68</td>
</tr>
<tr>
<td>D</td>
<td>29.8</td>
<td>115</td>
</tr>
<tr>
<td>E</td>
<td>30.5</td>
<td>91</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>433</td>
</tr>
</tbody>
</table>

This output table shows mean BMI values for each locality, along with a count (frequency) for each. It appears that people in the least deprived locality (A) have the lowest mean BMI value, while those having the highest BMI values reside in the most deprived locality (E). What we do not yet know, of course, is whether this effect is significant.

As mentioned previously, Levene’s test can be used to test the assumptions of one-way ANOVA.

<table>
<thead>
<tr>
<th>F</th>
<th>df1</th>
<th>df2</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>.409</td>
<td>4</td>
<td>428</td>
<td>.802</td>
</tr>
</tbody>
</table>

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

In the output shown, we can ignore the information in the grey cells, and concentrate on the significance (‘Sig.’) column. This shows a non-significant $P$-value (0.802). There is no evidence that variances across the groups (and hence standard deviations) are unequal – it is therefore appropriate to use one-way ANOVA. If this $P$-value were significant (< 0.05), the Kruskal–Wallis test should be used instead.
The following is an example computer output for a one-way ANOVA.

Tests of between-subjects effects

<table>
<thead>
<tr>
<th>Source</th>
<th>Type III sum of squares</th>
<th>df</th>
<th>Mean square</th>
<th>F</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corrected model</td>
<td>1253.230</td>
<td>4</td>
<td>313.308</td>
<td>12.291</td>
<td>.000</td>
</tr>
<tr>
<td>Intercept</td>
<td>341257.428</td>
<td>1</td>
<td>341257.428</td>
<td>13387.065</td>
<td>.000</td>
</tr>
<tr>
<td>Locality</td>
<td>1253.230</td>
<td>4</td>
<td>313.308</td>
<td>12.291</td>
<td>.000</td>
</tr>
<tr>
<td>Error</td>
<td>10910.396</td>
<td>428</td>
<td>25.492</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>369275.000</td>
<td>433</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected total</td>
<td>12163.626</td>
<td>432</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

We only need to focus on the F-statistic (F) and significance (Sig.) for **locality**. The F-statistic is 12.291, and there is a significant *P*-value of 0.000, or < 0.0001. This *P*-value suggests that (in this town), BMI is influenced by which locality people reside in.

There are considerable similarities between ANOVA and multiple regression (mentioned briefly following), and the two techniques generally give equivalent results (Kirkwood & Sterne, 2003).

**OTHER TYPES OF REGRESSION**

**Multiple regression**

In the linear regression example earlier, we used only one ‘exposure’ variable: age. It is also possible to examine the effect of more than one exposure, using **multiple regression**. For example, we could look at the effects of three continuous variables: age, BMD and height.

It is possible that height is a factor that could influence the value of BMD, as well as age. We could call this a **confounding** factor (discussed further in Chapter 24). Multiple regression could tell us whether age and BMD are still related, even when height is taken into account. If this is so, we can assume that height is not acting as a confounding factor.

The assumptions for multiple regression are the same as for linear regression. Three or more continuous variables can be used, and it is also possible to include categorical variables (e.g. ethnic group or sex). It is best, however, to keep the number of variables fairly small.

This technique goes beyond ‘basic’ statistical methods, and is covered in other texts – see Further reading.
Logistic regression
This is a technique that uses dichotomous variables (e.g. yes/no, present/absent, male/female) to predict the probability of an outcome.

For example, a total of 303 alcohol-abusing men were studied, to ascertain whether diagnosis of liver cirrhosis could be made on the basis of clinical symptoms alone, without the need to perform a surgical liver biopsy (Hamberg *et al.*, 1996). Six symptoms were studied: facial telangiectasia, vascular spiders, white nails, abdominal wall veins, fatness and peripheral oedema. In this case, the ‘dichotomous variables’ were the symptoms (because patients either have or do not have a particular symptom) and the dichotomous ‘outcome’ was liver cirrhosis. *Logistic regression* was used to predict the likelihood that a person having any combination of the symptoms actually had liver cirrhosis. A concise explanation of the logistic regression analysis used in this study was subsequently published in *Bandolier* (Freeman, 1997), and is available online at: www.medicine.ox.ac.uk/bandolier/band37/b37-5.html Results of the analysis were used to predict that people who experienced all six symptoms had a 97% chance of having cirrhosis, whereas there was a 20% chance in those who only had white nails and fatness.

Further details on logistic regression can be found in other texts – see Further reading.
This page intentionally left blank
So far we have looked at hypothesis tests for continuous variables, from which summary statistics such as means and medians can be calculated. However, when we have only categorical data, means and medians cannot be obtained. For example, it is not possible to calculate the mean of a group of colours.

The Chi-squared test (Chi is pronounced ‘ki’, as in ‘kind’ and is normally written as $\chi^2$) overcomes this problem, allowing hypothesis testing for categorical data. For example, we may wish to determine whether passive smokers are more likely to develop circulatory disease than those who are not exposed to smoke. In this example, passive smoking is the exposure and circulatory disease is the outcome. The Chi-squared test is a non-parametric test (see Chapter 17).

A good way to start examining the data is to present them in an $r \times c$ table (row $\times$ column; also known as a cross-classification or contingency table). Data are presented in cells, arranged in rows (horizontal) and columns (vertical). These often appear in the form of a $2 \times 2$ table (so called because it shows two exposures and two outcomes). An example of a $2 \times 2$ table is shown in Table 20.1.

<table>
<thead>
<tr>
<th>Exposure taken place?</th>
<th>Outcome present?</th>
<th>Yes</th>
<th>No</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>$a$</td>
<td>$b$</td>
<td></td>
<td>$a+b$</td>
</tr>
<tr>
<td>No</td>
<td>$c$</td>
<td>$d$</td>
<td></td>
<td>$c+d$</td>
</tr>
<tr>
<td>Total</td>
<td>$a+c$</td>
<td>$b+d$</td>
<td></td>
<td>$a+b+c+d$</td>
</tr>
</tbody>
</table>

If there are more than two categories of either exposure or outcome, then the number of columns or rows is increased, and the table is called a $2 \times n$ table. More categories can be used if required, in an $r \times c$ (row $\times$ column) table. The test statistic is calculated by taking the frequencies that are actually observed ($O$) and then working out...
the frequencies which would be expected \((E)\) if the null hypothesis was true. The hypothesis \((H_1)\) will be that there is an association between the variables, and the null hypothesis \((H_0)\) will be that there is no association between the variables.

The expected frequencies are calculated as follows:

\[
\frac{\text{row total} \times \text{column total}}{\text{grand total}}
\]

The expected frequency for each cell can be calculated using a 2 × 2 table as follows:

- cell a: \[\frac{(a + b) \times (a + c)}{\text{total}}\]
- cell b: \[\frac{(a + b) \times (b + d)}{\text{total}}\]
- cell c: \[\frac{(a + c) \times (c + d)}{\text{total}}\]
- cell d: \[\frac{(b + d) \times (c + d)}{\text{total}}\]

These are then compared using this formula, to produce the \(\chi^2\) statistic:

\[
\chi^2 = \sum \frac{(O - E)^2}{E}
\]

where \(O\) = observed frequencies and \(E\) = expected frequencies. Degrees of freedom (d.f.) are calculated using the following formula:

\[
d.f. = (r - 1) \times (c - 1)
\]

where \(r\) = number of rows and \(c\) = number of columns.

The greater the difference between the observed and expected frequencies, the less likely it is that the null hypothesis is true.

The Chi-squared test only works when frequencies are used in the cells. Data such as proportions, means or physical measurements are not valid. This test is used to detect an association between data in rows and data in columns, but it does not indicate the strength of any association. The Chi-squared test is more accurate when large frequencies are used – all of the expected frequencies should be more than 1, and at least 80% of the expected frequencies should be more than 5. If these conditions, called the assumptions of the test, are not met, the Chi-squared test is not valid and therefore cannot be used. If the Chi-squared test is not valid and a 2 × 2 table is being used, Fisher’s exact test can sometimes be utilised (the formula for this test is not covered in this basic guide, but many computer programs will automatically calculate it if sufficiently small expected frequencies are detected within a 2 × 2 table). If there are more than two rows and/or columns, it may be possible to regroup the data so as to create fewer columns. Doing this will increase the cell frequencies, which may then be large enough to meet the requirements. For example, if you have four age
CHI-SQUARED TEST

It might be reasonable to combine these to produce two age groups (0–14 and 15–28 years). However, regrouping data into fewer categories is a compromise, as the precision that is allowed by having so many categories will be reduced.

If the test is being carried out to detect an association between paired data where there are only two possible outcomes (e.g. the outcome is either success or failure and two different regimes are tried on the same individuals or on matched pairs), then McNemar’s Chi-squared test should be used. This is not covered in this basic guide.

Let us look at an example using some real data, as shown in Table 20.2. A study asks whether Asians with diabetes receive worse treatment in primary care than non-Asians with diabetes. This is important, since Asians are more likely to develop diabetes than non-Asians. A number of variables are studied, including whether patients with diabetes have received a HbA1c test within the previous year (we mentioned HbA1c in Chapter 7), as this is a valuable indicator of how successfully diabetes is being controlled. Having the test performed regularly is important, and is therefore a valid indicator of healthcare quality in diabetes. We can calculate that 64.6% (128/198) of Asians received the check, compared with 74.7% (430/576) of non-Asians. As such we know that a lower proportion of Asian patients was checked, but is there a significant association between ethnicity and receiving the check? Our null hypothesis is that there is no association between ethnicity and receiving a HbA1c check.

**TABLE 20.2** Frequencies for HbA1c testing by ethnic group. Adapted from Stewart and Rao (2000)

<table>
<thead>
<tr>
<th>Ethnicity of patient</th>
<th>HbA1c test done?</th>
<th>Yes</th>
<th>No</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asian</td>
<td></td>
<td>128</td>
<td>70</td>
<td>198</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(a)</td>
<td>(b)</td>
<td>(a + b)</td>
</tr>
<tr>
<td>Non-Asian</td>
<td></td>
<td>430</td>
<td>146</td>
<td>576</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(c)</td>
<td>(d)</td>
<td>(c + d)</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>558</td>
<td>216</td>
<td>774</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(a + c)</td>
<td>(b + d)</td>
<td>(a + b + c + d)</td>
</tr>
</tbody>
</table>

The frequencies for Asian/non-Asian patients with diabetes are assembled in a $2 \times 2$ table and tabulated against the frequencies in each group of patients who have/have not received the HbA1c test, as shown in Table 20.2.

To calculate $\chi^2$, use the following steps.

1. Work out the degrees of freedom (d.f.).
2. Work out the expected frequencies in each of cells $a$, $b$, $c$ and $d$ – **or more if it is a larger table**.
3. For each cell, subtract the expected frequency from the observed frequency \((O - E)\).
4. For each cell, square the result \((O - E)^2\).
5. For each cell, divide this number by the expected frequency \([\frac{(O - E)^2}{E}]\).
6. Add up the results for each cell – this gives you the \(\chi^2\) statistic.
7. Using the \(\chi^2\) distribution table in Appendix 1, look up the d.f. value in the left-hand column.
8. Read across this row until the nearest values to the left and right of your \(\chi^2\) statistic can be seen.
9. Your \(P\)-value will be less than the \(P\)-value at the top of the column to the left of your \(\chi^2\) statistic and greater than the \(P\)-value at the top of the column to its right. (For example, a \(\chi^2\) statistic of 6.128 with 2 d.f. falls in between 5.991 and 7.824. The nearest value to its left is 5.991; the \(P\)-value at the top of this column is 0.05. The \(P\)-value for your \(\chi^2\) statistic will therefore be less than 0.05, and is written \(P < 0.05\). If your \(\chi^2\) statistic is 2.683 with 2 d.f., there is no column to its left, so the \(P\)-value will be greater than the column to its right, and is therefore \(> 0.2\)).

Using the data for the Asian diabetes study, let us work out \(\chi^2\).

1. There are two rows and two columns:

\[
(r - 1) \times (c - 1) = (2 - 1) \times (2 - 1) = 1 \times 1; \text{ so d.f. } = 1.
\]

2. Work out the expected frequencies for each cell (to two decimal places in this example):

- cell a: \([(a + b) \times (a + c)/\text{total}] = (198 \times 558)/774 = 110.484/774 = 142.74\)
- cell b: \([(a + b) \times (b + d)/\text{total}] = (198 \times 216)/774 = 42768/774 = 55.26\)
- cell c: \([(a + c) \times (c + d)/\text{total}] = (558 \times 576)/774 = 321408/774 = 415.26\)
- cell d: \([(b + d) \times (c + d)/\text{total}] = (216 \times 576)/774 = 124416/774 = 160.74.\)

Going back to the assumptions mentioned earlier in the chapter, it is clear that all of the expected frequencies are more than 1 and all are also more than 5. The Chi-squared test is therefore valid and it can be used.

3–5. It is helpful to construct a grid to aid the following calculations, as shown in Table 20.3.

6. The sum of all of the \((O - E^2/E)\) results is 7.32 – this is the \(\chi^2\) statistic.
7. On the $\chi^2$ distribution table in Appendix 1, look along the row for d.f. = 1.
8. Look along the row to find the values to the left and right of the $\chi^2$ statistic – it lies in between 6.635 and 10.827.
9. Reading up the columns for these two values shows that the corresponding $P$-value is less than 0.01 but greater than 0.001 – we can therefore write the $P$-value as $P < 0.01$.

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>128</td>
<td>142.74</td>
<td>-14.74</td>
<td>217.27</td>
</tr>
<tr>
<td>b</td>
<td>70</td>
<td>55.26</td>
<td>14.74</td>
<td>217.27</td>
</tr>
<tr>
<td>c</td>
<td>430</td>
<td>415.26</td>
<td>14.74</td>
<td>217.27</td>
</tr>
<tr>
<td>d</td>
<td>146</td>
<td>160.74</td>
<td>-14.74</td>
<td>217.27</td>
</tr>
<tr>
<td>Total</td>
<td>774</td>
<td>7.32</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Thus there is strong evidence to reject the null hypothesis, and we may conclude that there is an association between being Asian and receiving a HbA1c check. Asian patients are significantly less likely to receive a HbA1c check, and appear to receive a poorer quality of care in this respect.

The $\chi^2$ formula is made more conservative by subtracting 0.5 from the product of $(O - E)$ at stage 3. We can ignore any minus numbers in the product of $(O - E)$, and it is thus written as $(|O - E|)$. This becomes $|(O - E)| - 0.5$, and is known as **Yates’ correction** (also called a **continuity correction**). It is especially important to use this when frequencies are small. Note that Yates’ correction can only be used for $2 \times 2$ tables. If Yates’ correction is applied to the data shown, we obtain the following result, as shown in Table 20.4.

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>128</td>
<td>142.74</td>
<td>14.24</td>
<td>202.78</td>
</tr>
<tr>
<td>b</td>
<td>70</td>
<td>55.26</td>
<td>14.24</td>
<td>202.78</td>
</tr>
<tr>
<td>c</td>
<td>430</td>
<td>415.26</td>
<td>14.24</td>
<td>202.78</td>
</tr>
<tr>
<td>d</td>
<td>146</td>
<td>160.74</td>
<td>14.24</td>
<td>202.78</td>
</tr>
<tr>
<td>Total</td>
<td>774</td>
<td>6.84</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Thus $\chi^2 = 6.84$, which still gives a $P$-value of $< 0.01$. However, this is closer to the 0.01 value than the previous $\chi^2$ of 7.32. The significance is therefore slightly reduced.
Chi-squared for trend can be used to test for a statistically significant trend in exposure groups which have a meaningful order and two outcomes. For example, this could apply to age groups (. . . 45–54, 55–64, 65–74, 75+) and diagnosis of dementia (Y/N) or pain severity (mild, moderate, severe) and cessation of pain (Y/N). The calculation of Chi-squared for trend is not covered in this basic guide.