If we turn our attention to independent designs, a similar argument can be applied. We know that different participants participate in different experimental conditions and that these participants will differ in many respects (their IQ, attention span, etc.). Although we know that these confounding variables contribute to the variation between conditions, we need to make sure that these variables contribute to the unsystematic variation and *not* the systematic variation. The way to ensure that confounding variables are unlikely to contribute systematically to the variation between experimental conditions is to randomly allocate participants to a particular experimental condition. This should ensure that these confounding variables are evenly distributed across conditions.

A good example is the effects of alcohol on personality. You might give one group of people 5 pints of beer, and keep a second group sober, and then count how many fights each person gets into. The effect that alcohol has on people can be very variable because of different tolerance levels: teetotal people can become very drunk on a small amount, while alcoholics need to consume vast quantities before the alcohol affects them. Now, if you allocated a bunch of teetotal participants to the condition that consumed alcohol, then you might find no difference between them and the sober group (because the teetotal participants are all unconscious after the first glass and so can't become involved in any fights). As such, the person's prior experiences with alcohol will create systematic variation that cannot be dissociated from the effect of the experimental manipulation. The best way to reduce this eventuality is to randomly allocate participants to conditions.

SELF-TEST Why is randomization important?

1.7. Analysing data ①

The final stage of the research process is to analyse the data you have collected. When the data are quantitative this involves both looking at your data graphically to see what the general trends in the data are, and also fitting statistical models to the data.

1.7.1. Frequency distributions ①

Once you've collected some data a very useful thing to do is to plot a graph of how many times each score occurs. This is known as a **frequency distribution**, or **histogram**, which is a graph plotting values of observations on the horizontal axis, with a bar showing how many times each value occurred in the data set. Frequency distributions can be very useful for assessing properties of the distribution of scores. We will find out how to create these types of charts in Chapter 4.

Frequency distributions come in many different shapes and sizes. It is quite important, therefore, to have some general descriptions for common types of distributions. In an ideal world our data would be distributed symmetrically around the centre of all scores. As such, if we drew a vertical line through the centre of the distribution then it should look the same on both sides. This is known as a **normal distribution** and is characterized by the bell-shaped curve with which you might already be familiar. This shape basically implies that the majority of scores lie around the centre of the distribution (so the largest bars on the histogram are all around the central value).

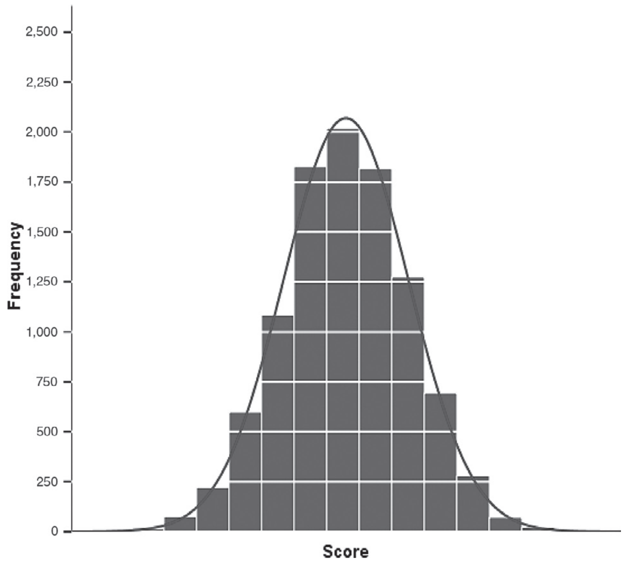
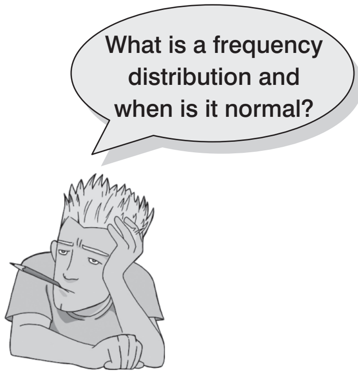


FIGURE 1.3 A 'normal' distribution (the curve shows the idealized shape)



Also, as we get further away from the centre the bars get smaller, implying that as scores start to deviate from the centre their frequency is decreasing. As we move still further away from the centre our scores become very infrequent (the bars are very short). Many naturally occurring things have this shape of distribution. For example, most men in the UK are about 175 cm tall,[12] some are a bit taller or shorter but most cluster around this value. There will be very few men who are really tall (i.e. above 205 cm) or really short (i.e. under 145 cm). An example of a normal distribution is shown in Figure 1.3.

There are two main ways in which a distribution can deviate from normal: (1) lack of symmetry (called **skew**) and (2) pointyness (called **kurtosis**). Skewed distributions are not symmetrical and instead the most frequent scores (the tall bars on the graph) are clustered at one end of the scale. So, the typical pattern is a cluster of frequent scores at one end of the scale and the frequency of scores tailing off towards the other end of the scale. A skewed distribution can be either *positively skewed* (the frequent scores are clustered at the lower end and the tail points towards the higher or more positive scores) or *negatively skewed* (the frequent scores are clustered at the higher end and the tail points towards the lower or more negative scores). Figure 1.4 shows examples of these distributions.

Distributions also vary in their kurtosis. Kurtosis, despite sounding like some kind of exotic disease, refers to the degree to which scores cluster at the ends of the distribution (known as the *tails*) and how pointy a distribution is (but there are other factors that can affect how pointy the distribution looks – see Jane Superbrain Box 2.3). A distribution with *positive kurtosis* has many scores in the tails (a so-called heavy-tailed distribution) and is pointy. This is known as a **leptokurtic** distribution. In contrast, a distribution with *negative kurtosis* is relatively thin in the tails (has light tails) and tends to be flatter than normal. This distribution is called **platykurtic**. Ideally, we want our data to be normally distributed (i.e. not too skewed, and not too many or too few scores at the extremes!). For everything there is to know about kurtosis read DeCarlo (1997).

In a normal distribution the values of skew and kurtosis are 0 (i.e. the tails of the distribution are as they should be). If a distribution has values of skew or kurtosis above or below 0 then this indicates a deviation from normal: Figure 1.5 shows distributions with kurtosis values of +1 (left panel) and –4 (right panel).

[12] I am exactly 180 cm tall. In my home country this makes me smugly above average. However, I'm writing this in The Netherlands where the average male height is 185 cm (a massive 10 cm higher than the UK), and where I feel like a bit of a dwarf.

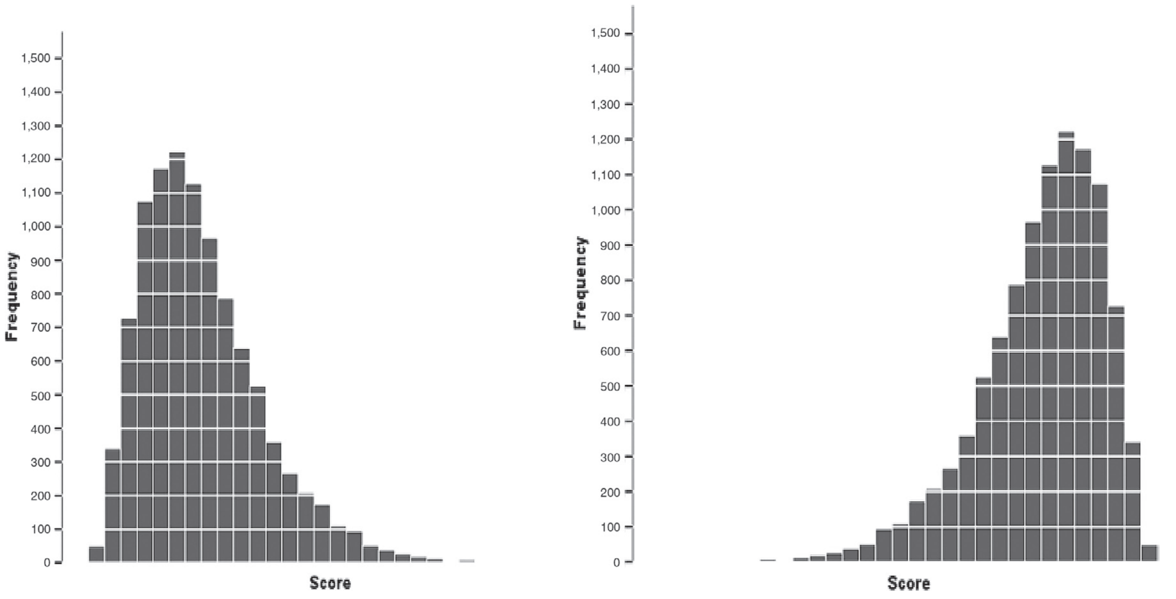


FIGURE 1.4 A positively (left figure) and negatively (right figure) skewed distribution

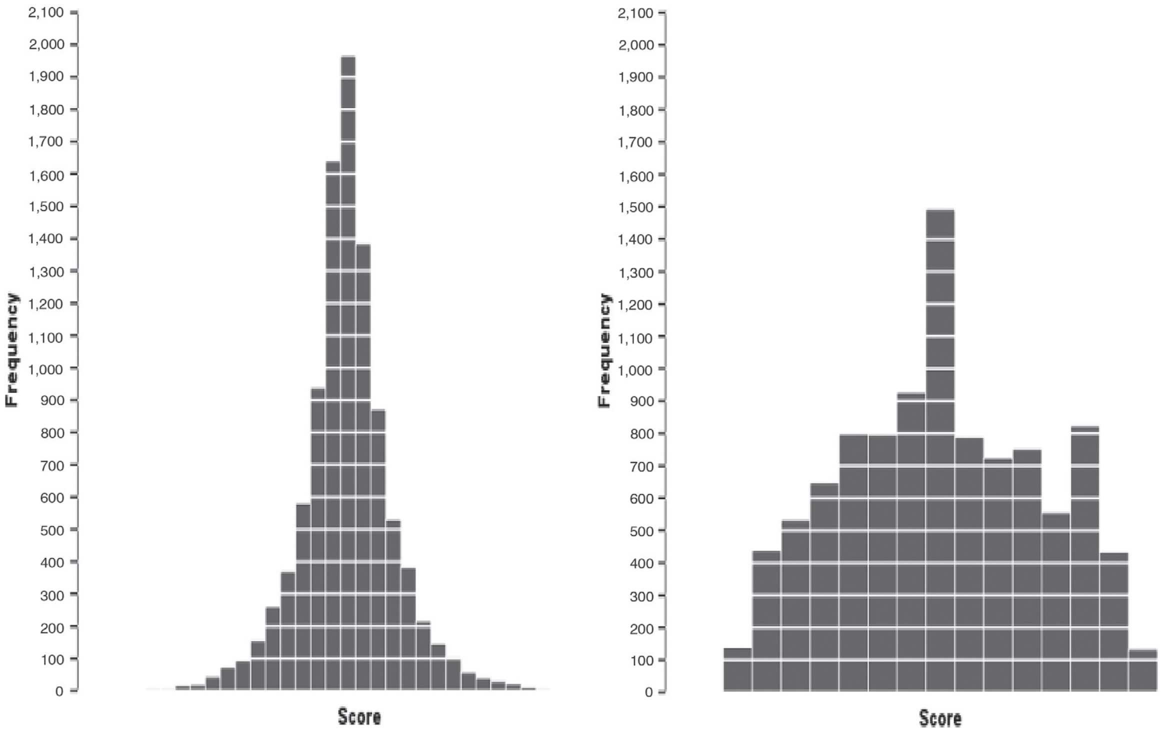


FIGURE 1.5 Distributions with positive kurtosis (leptokurtic, left figure) and negative kurtosis (platykurtic, right figure)

1.7.2. The centre of a distribution ①

We can also calculate where the centre of a frequency distribution lies (known as the **central tendency**). There are three measures commonly used: the mean, the mode and the median.

1.7.2.1. The mode ①

The **mode** is simply the score that occurs most frequently in the data set. This is easy to spot in a frequency distribution because it will be the tallest bar! To calculate the mode, simply place the data in ascending order (to make life easier), count how many times each score occurs, and the score that occurs the most is the mode! One problem with the mode is that it can often take on several values. For example, Figure 1.6 shows an example of a distribution with two modes (there are two bars that are the highest), which is said to be **bimodal**. It's also possible to find data sets with more than two modes (**multimodal**). Also, if the frequencies of certain scores are very similar, then the mode can be influenced by only a small number of cases.

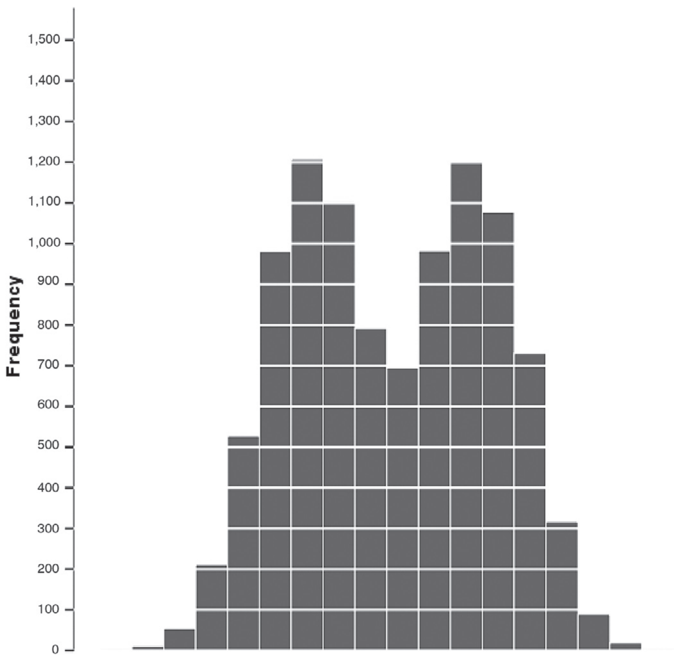


FIGURE 1.6 A bimodal distribution

1.7.2.2. The median ①

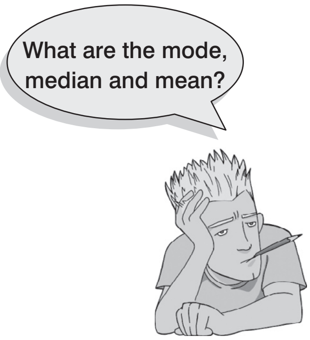
Another way to quantify the centre of a distribution is to look for the middle score when scores are ranked in order of magnitude. This is called the **median**. For example, Facebook is a popular social networking website, in which users can sign up to be 'friends' of other users. Imagine we looked at the number of friends that a selection (actually, some of my friends) of 11 Facebook users had. Number of friends: 108, 103, 252, 121, 93, 57, 40, 53, 22, 116, 98.

To calculate the median, we first arrange these scores into ascending order: 22, 40, 53, 57, 93, 98, 103, 108, 116, 121, 252.

Next, we find the position of the middle score by counting the number of scores we have collected (n), adding 1 to this value, and then dividing by 2. With 11 scores, this gives us $(n + 1)/2 = (11 + 1)/2 = 12/2 = 6$. Then, we find the score that is positioned at the location we have just calculated. So, in this example we find the sixth score:

22, 40, 53, 57, 93, 98, 103, 108, 116, 121, 252

Median



This works very nicely when we have an odd number of scores (as in this example) but when we have an even number of scores there won't be a middle value. Let's imagine that we decided that because the highest score was so big (more than twice as large as the next biggest number), we would ignore it. (For one thing, this person is far too popular and we hate them.) We have only 10 scores now. As before, we should rank-order these scores: 22, 40, 53, 57, 93, 98, 103, 108, 116, 121. We then calculate the position of the middle score, but this time it is $(n + 1)/2 = 11/2 = 5.5$. This means that the median is halfway between the fifth and sixth scores. To get the median we add these two scores and divide by 2. In this example, the fifth score in the ordered list was 93 and the sixth score was 98. We add these together (93 + 98 = 191) and then divide this value by 2 (191/2 = 95.5). The median number of friends was, therefore, 95.5.

The median is relatively unaffected by extreme scores at either end of the distribution: the median changed only from 98 to 95.5 when we removed the extreme score of 252. The median is also relatively unaffected by skewed distributions and can be used with ordinal, interval and ratio data (it cannot, however, be used with nominal data because these data have no numerical order).

1.7.2.3. The mean ①

The **mean** is the measure of central tendency that you are most likely to have heard of because it is simply the average score and the media are full of average scores.[13] To calculate the mean we simply add up all of the scores and then divide by the total number of scores we have. We can write this in equation form as:

$$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{1.1}$$

This may look complicated, but the top half of the equation simply means 'add up all of the scores' (the x_i just means 'the score of a particular person'; we could replace the letter i with each person's name instead), and the bottom bit means divide this total by the number of scores you have got (n). Let's calculate the mean for the Facebook data. First, we first add up all of the scores:

$$\begin{aligned}\sum_{i=1}^{n} x_i &= 22 + 40 + 53 + 57 + 93 + 98 + 103 + 108 + 116 + 121 + 252 \\ &= 1063\end{aligned}$$

We then divide by the number of scores (in this case 11):

$$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{1063}{11} = 96.64$$

The mean is 96.64 friends, which is not a value we observed in our actual data (it would be ridiculous to talk of having 0.64 of a friend). In this sense the mean is a statistical model – more on this in the next chapter.

[13] I'm writing this on 15 February 2008, and to prove my point the BBC website is running a headline about how PayPal estimates that Britons will spend an average of £71.25 each on Valentine's Day gifts, but uSwitch.com said that the average spend would be £22.69!

SELF-TEST Compute the mean but excluding the score of 252.

If you calculate the mean without our extremely popular person (i.e. excluding the value 252), the mean drops to 81.1 friends. One disadvantage of the mean is that it can be influenced by extreme scores. In this case, the person with 252 friends on Facebook increased the mean by about 15 friends! Compare this difference with that of the median. Remember that the median hardly changed if we included or excluded 252, which illustrates how the median is less affected by extreme scores than the mean. While we're being negative about the mean, it is also affected by skewed distributions and can be used only with interval or ratio data.

If the mean is so lousy then why do we use it all of the time? One very important reason is that it uses every score (the mode and median ignore most of the scores in a data set). Also, the mean tends to be stable in different samples.

1.7.3. The dispersion in a distribution ①

It can also be interesting to try to quantify the spread, or dispersion, of scores in the data. The easiest way to look at dispersion is to take the largest score and subtract from it the smallest score. This is known as the **range** of scores. For our Facebook friends data, if we order these scores we get 22, 40, 53, 57, 93, 98, 103, 108, 116, 121, 252. The highest score is 252 and the lowest is 22; therefore, the range is 252 – 22 = 230. One problem with the range is that because it uses only the highest and lowest score it is affected dramatically by extreme scores.

SELF-TEST Compute the range but excluding the score of 252.

If you have done the self-test task you'll see that without the extreme score the range drops dramatically from 230 to 99 – less than half the size!

One way around this problem is to calculate the range when we exclude values at the extremes of the distribution. One convention is to cut off the top and bottom 25% of scores and calculate the range of the middle 50% of scores – known as the **interquartile range**. Let's do this with the Facebook data. First we need to calculate what are called **quartiles**. Quartiles are the three values that split the sorted data into four equal parts. First we calculate the median, which is also called the **second quartile**, which splits our data into two equal parts. We already know that the median for these data is 98. The **lower quartile** is the median of the lower half of the data and the **upper quartile** is the median of the upper half of the data. One rule of thumb is that the median is not included in the two halves when they are split (this is convenient if you have an odd number of values), but you can include it (although which half you put it in is another question). Figure 1.7 shows how we would calculate these values for the Facebook data. Like the median, the upper and lower quartile need not be values that actually appear in the data (like the median, if each half of the data had an even number of values in it then the upper and lower quartiles would be the average

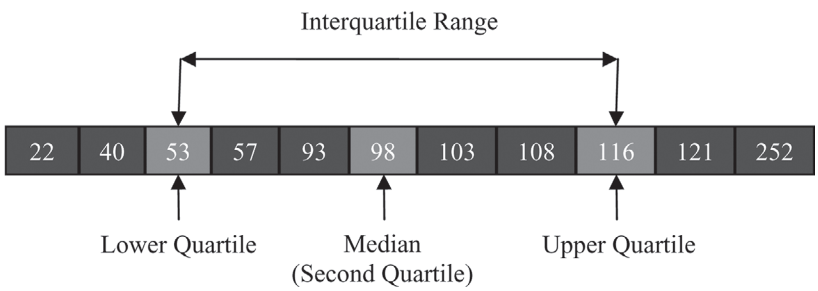


FIGURE 1.7 Calculating quartiles and the interquartile range

of two values in the data set). Once we have worked out the values of the quartiles, we can calculate the interquartile range, which is the difference between the upper and lower quartile. For the Facebook data this value would be 116–53 = 63. The advantage of the interquartile range is that it isn't affected by extreme scores at either end of the distribution. However, the problem with it is that you lose a lot of data (half of it in fact!).

SELF-TEST Twenty-one heavy smokers were put on a treadmill at the fastest setting. The time in seconds was measured until they fell off from exhaustion: 18, 16, 18, 24, 23, 22, 22, 23, 26, 29, 32, 34, 34, 36, 36, 43, 42, 49, 46, 46, 57

Compute the mode, median, mean, upper and lower quartiles, range and interquartile range

1.7.4. Using a frequency distribution to go beyond the data ①

Another way to think about frequency distributions is not in terms of how often scores actually occurred, but how likely it is that a score would occur (i.e. probability). The word 'probability' induces suicidal ideation in most people (myself included) so it seems fitting that we use an example about throwing ourselves off a cliff. Beachy Head is a large, windy cliff on the Sussex coast (not far from where I live) that has something of a reputation for attracting suicidal people, who seem to like throwing themselves off it (and after several months of rewriting this book I find my thoughts drawn towards that peaceful chalky cliff top more and more often). Figure 1.8 shows a frequency distribution of some completely made up data of the number of suicides at Beachy Head in a year by people of different ages (although I made these data up, they are roughly based on general suicide statistics such as those in Williams, 2001). There were 172 suicides in total and you can see that the suicides were most frequently aged between about 30 and 35 (the highest bar). The graph also tells us that, for example, very few people aged above 70 committed suicide at Beachy Head.

I said earlier that we could think of frequency distributions in terms of probability. To explain this, imagine that someone asked you 'how likely is it that a 70 year old committed suicide at Beach Head?' What would your answer be? The chances are that if you looked at the frequency distribution you might respond 'not very likely' because you can see that only 3 people out of the 172 suicides were aged around 70. What about if someone asked you 'how likely is it that a 30 year old committed suicide?' Again, by looking at the graph, you might say 'it's actually quite likely' because 33 out of the 172 suicides were by people aged around 30 (that's more than 1 in every 5 people who committed suicide). So based

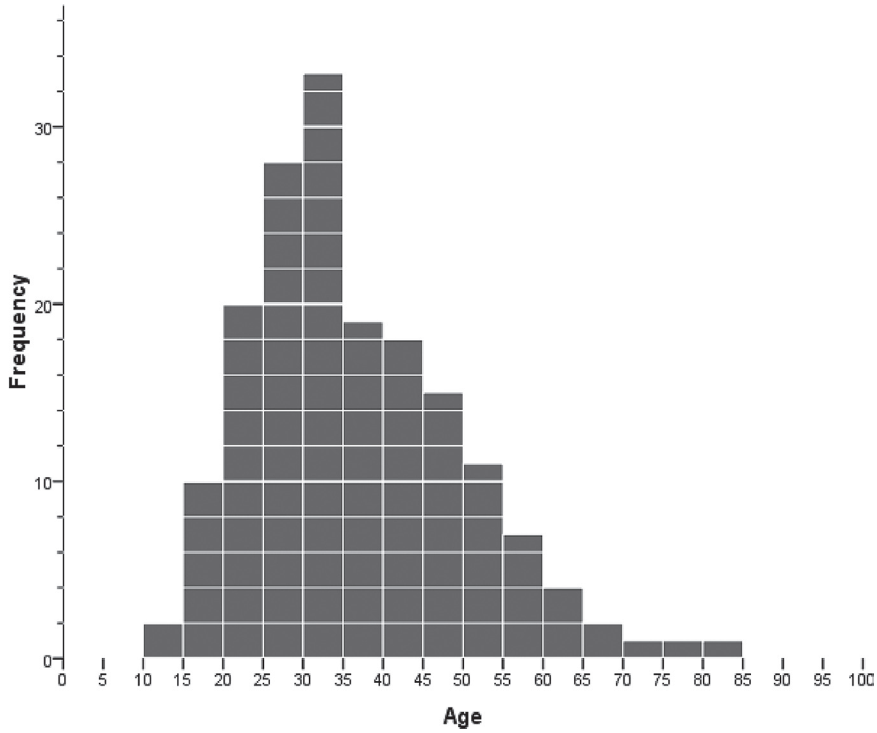
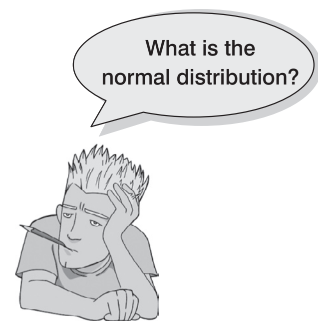


FIGURE 1.8 Frequency distribution showing the number of suicides at Beachy Head in a year by age

on the frequencies of different scores it should start to become clear that we could use this information to estimate the probability that a particular score will occur. We could ask, based on our data, 'what's the probability of a suicide victim being aged 16–20?' A probability value can range from 0 (there's no chance whatsoever of the event happening) to 1 (the event will definitely happen). So, for example, when I talk to my publishers I tell them there's a probability of 1 that I will have completed the revisions to this book by April 2008. However, when I talk to anyone else, I might, more realistically, tell them that there's a .10 probability of me finishing the revisions on time (or put another way, a 10% chance, or 1 in 10 chance that I'll complete the book in time). In reality, the probability of my meeting the deadline is 0 (not a chance in hell) because I never manage to meet publisher's deadlines! If probabilities don't make sense to you then just ignore the decimal point and think of them as percentages instead (i.e. .10 probability that something will happen = 10% chance that something will happen).

I've talked in vague terms about how frequency distributions can be used to get a rough idea of the probability of a score occurring. However, we can be precise. For any distribution of scores we could, in theory, calculate the probability of obtaining a score of a certain size – it would be incredibly tedious and complex to do it, but we could. To spare our sanity, statisticians have identified several common distributions. For each one they have worked out mathematical formulae that specify idealized versions of these distributions (they are specified in terms of a curved line). These idealized distributions are known as **probability distributions** and from these distributions it is possible to calculate the probability of getting particular scores based on the frequencies with which a particular score occurs in a distribution with these common shapes. One of these 'common' distributions is the normal distribution, which I've already mentioned in section 1.7.1. Statisticians have calculated the probability of certain scores occurring in a normal distribution with a mean of 0 and a standard deviation of 1. Therefore, if we have any data that are shaped like a normal distribution, then if the mean and standard deviation



are 0 and 1 respectively we can use the tables of probabilities for the normal distribution to see how likely it is that a particular score will occur in the data (I've produced such a table in the Appendix to this book).

The obvious problem is that not all of the data we collect will have a mean of 0 and standard deviation of 1. For example, we might have a data set that has a mean of 567 and a standard deviation of 52.98. Luckily any data set can be converted into a data set that has a mean of 0 and a standard deviation of 1. First, to centre the data around zero, we take each score and subtract from it the mean of all. Then, we divide the resulting score by the standard deviation to ensure the data have a standard deviation of 1. The resulting scores are known as **z-scores** and in equation form, the conversion that I've just described is:

$$z = \frac{X - \overline{X}}{s} \tag{1.2}$$

The table of probability values that have been calculated for the standard normal distribution is shown in the Appendix. Why is this table important? Well, if we look at our suicide data, we can answer the question 'What's the probability that someone who threw themselves off of Beachy Head was 70 or older?' First we convert 70 into a z-score. Say, the mean of the suicide scores was 36, and the standard deviation 13; then 70 will become $(70 - 36)/13 = 2.62$. We then look up this value in the column labelled 'Smaller Portion' (i.e. the area above the value 2.62). You should find that the probability is .0044, or put another way, only a 0.44% chance that a suicide victim would be 70 years old or more. By looking at the column labelled 'Bigger Portion' we can also see the probability that a suicide victim was aged 70 or less. This probability is .9956, or put another way, there's a 99.56% chance that a suicide victim was less than 70 years old.

Hopefully you can see from these examples that the normal distribution and z-scores allow us to go a first step beyond our data in that from a set of scores we can calculate the probability that a particular score will occur. So, we can see whether scores of a certain size are likely or unlikely to occur in a distribution of a particular kind. You'll see just how useful this is in due course, but it is worth mentioning at this stage that certain z-scores are particularly important. This is because their value cuts off certain important percentages of the distribution. The first important value of z is 1.96 because this cuts off the top 2.5% of the distribution, and its counterpart at the opposite end (–1.96) cuts off the bottom 2.5% of the distribution. As such, taken together, this value cuts of 5% of scores, or put another way, 95% of z-scores lie between –1.96 and 1.96. The other two important benchmarks are ±2.58 and ±3.29, which cut off 1% and 0.1% of scores respectively. Put another way, 99% of z-scores lie between –2.58 and 2.58, and 99.9% of them lie between –3.29 and 3.29. Remember these values because they'll crop up time and time again.

SELF-TEST Assuming the same mean and standard deviation for the Beachy Head example above, what's the probability that someone who threw themselves off Beachy Head was 30 or younger?

1.7.5. Fitting statistical models to the data ①

Having looked at your data (and there is a lot more information on different ways to do this in Chapter 4), the next step is to fit a statistical model to the data. I should really just