

varied not just on dose of alcohol but also on their tolerance of alcohol (the systematic variation created by their past experience with alcohol cannot be separated from the effect of the experimental manipulation). The best way to reduce this eventuality is to randomly allocate participants to conditions: by doing so you minimize the risk that groups differ on variables other than the one you want to manipulate.



1.8 Analysing data



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 ([Chapter 5](#)) to see what the general trends in the data are, and also fitting statistical models to the data (all other chapters). Given that the rest of the book is dedicated to this process, we'll begin here by looking at a few fairly basic ways to look at and summarize the data you have collected.

1.8.1 Frequency distributions



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 5](#).



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 implies that the majority of scores lie around the centre of the distribution (so the largest bars on the histogram are around the central value). 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 around 175 cm tall;¹⁶ 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](#).

¹⁶ I am exactly 180 cm tall. In my home country this makes me smugly above average. However, I often visit the Netherlands, where the average male height is 185 cm (a little over 6ft, and a massive 10 cm higher than the UK), and where I feel like a bit of a dwarf.

Figure 1.3 A ‘normal’ distribution (the curve shows the idealized shape)

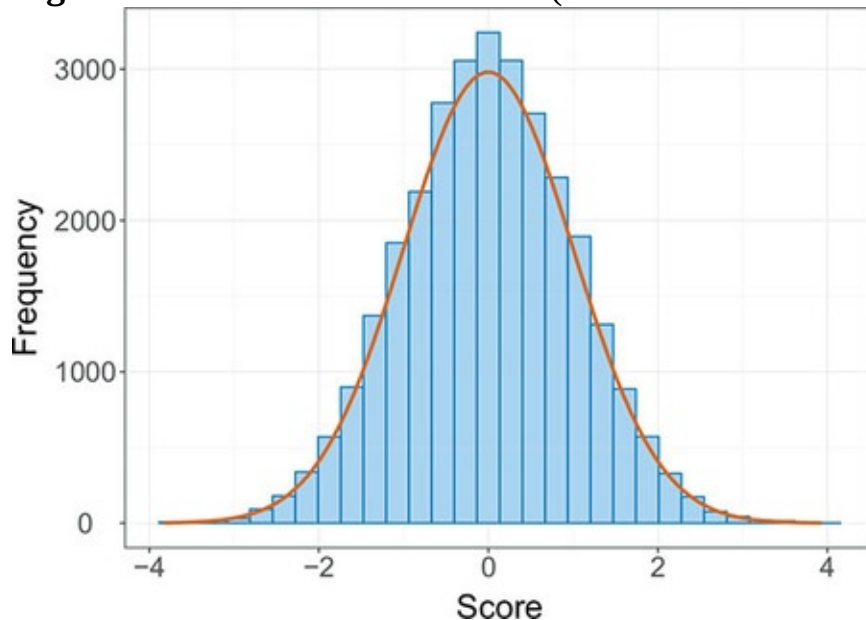
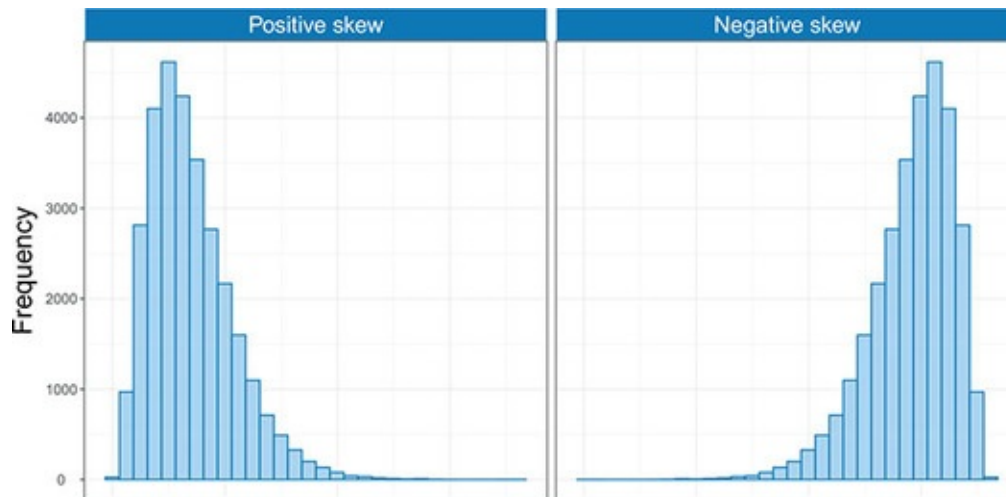


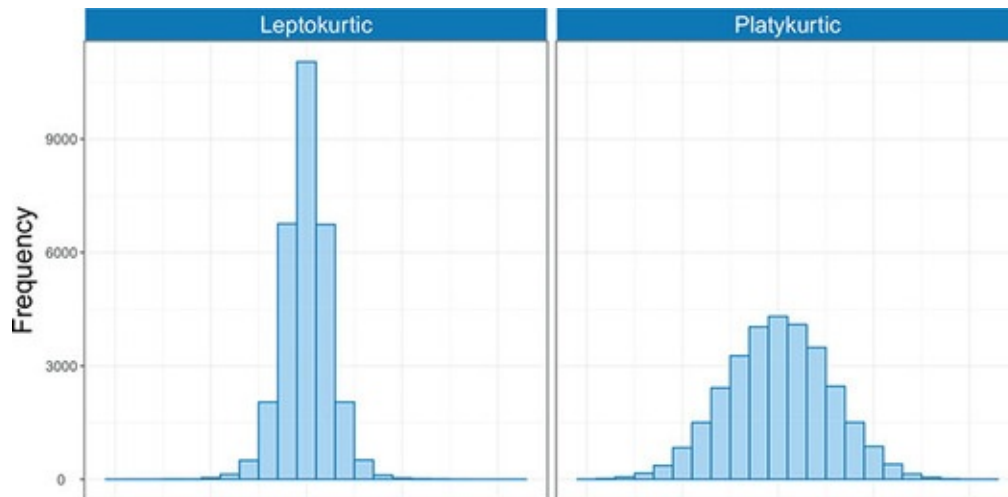
Figure 1.4 A positively (left) and negatively (right) skewed distribution



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 this tends to express itself in how pointy a distribution is (but there are other factors that can affect how pointy the distribution looks – see [Jane Superbrain Box 1.5](#)). 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).

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



In a normal distribution the values of skew and kurtosis are 0 (i.e., the tails of the distribution are as they should be).¹⁷ 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 +2.6 (left panel) and -0.09 (right panel).

¹⁷ Sometimes no kurtosis is expressed as 3 rather than 0, but SPSS uses 0 to denote no excess kurtosis.

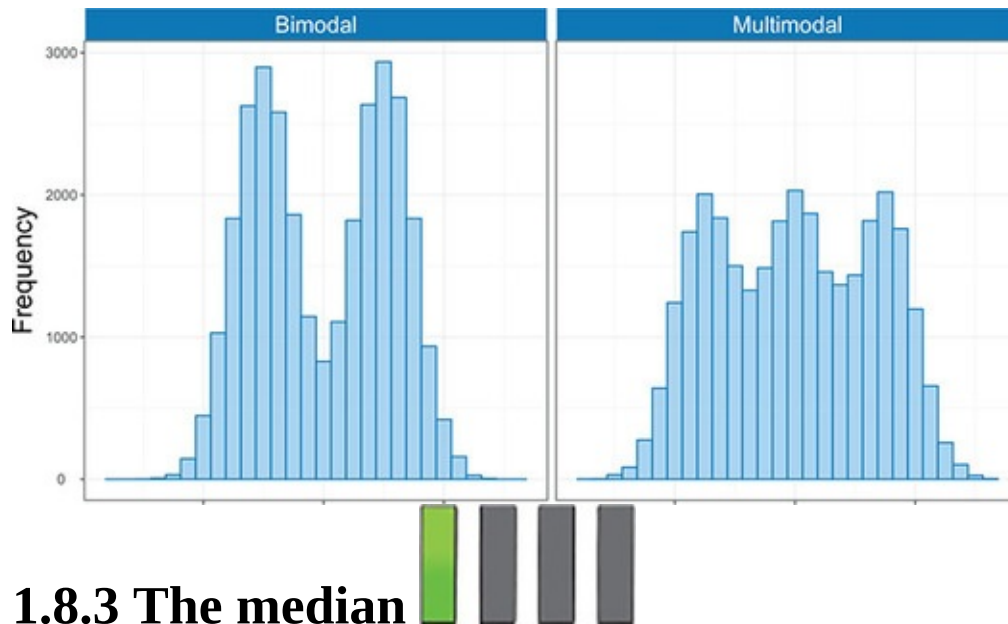
1.8.2 The mode



We can calculate where the centre of a frequency distribution lies (known as the **central tendency**) using three measures commonly used: the mean, the mode and the median. Other methods exist, but these three are the ones you're most likely to come across.

The **mode** is 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, 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 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**, and three modes (data sets with more than two modes are **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 Examples of bimodal (left) and multimodal (right) distributions



1.8.3 The median

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**. Imagine we looked at the number of friends that 11 users of the social networking website Facebook had. [Figure 1.7](#) shows the number of friends for each of the 11 users: 57, 40, 103, 234, 93, 53, 116, 98, 108, 121, 22.



To calculate the median, we first arrange these scores into ascending order: 22, 40, 53, 57, 93, 98, 103, 108, 116, 121, 234.

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 (see [Figure 1.7](#)).

This process 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 (almost 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. [Figure 1.8](#) shows this situation. As before, we 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 $(n + 1)/2 = 11/2 = 5.5$, which 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.

Figure 1.7 The median is simply the middle score when you order the data

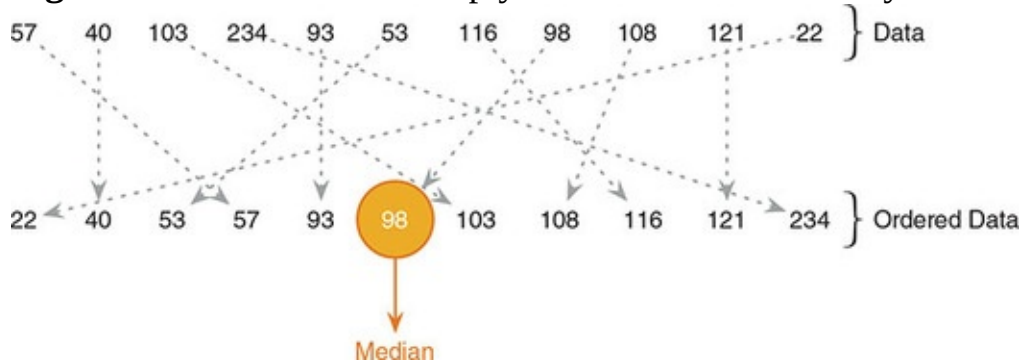
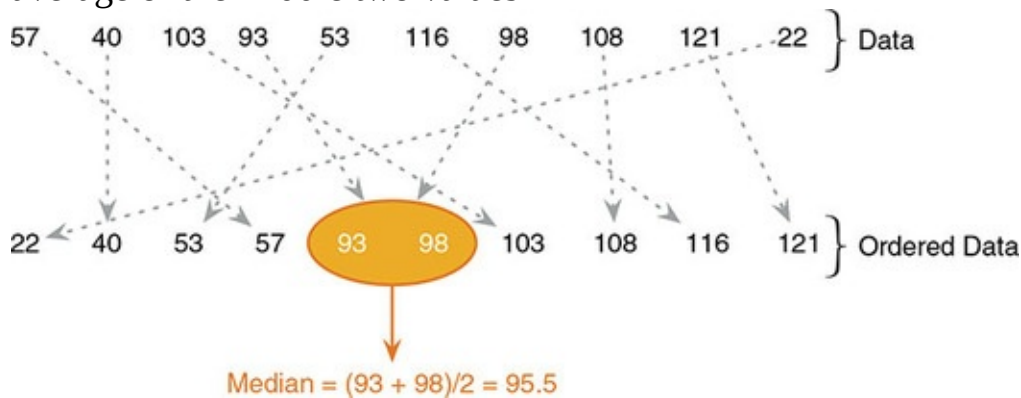


Figure 1.8 When the data contain an even number of scores, the median is the average of the middle two values



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 234. 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).



1.8.4 The mean

The **mean** is the measure of central tendency that you are most likely to have heard of because it is the average score, and the media love an average score.¹⁸ To calculate the mean we 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:

18 I wrote this on 15 February, and to prove my point, the BBC website ran a headline today about how PayPal estimates that Britons will spend an average of £71.25 each on Valentine’s Day gifts. However, uSwitch.com said that the average spend would be only £22.69. Always remember that the media is full of lies and contradictions.

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

This equation may look complicated, but the top half simply means ‘add up all of the scores’ (the x_i 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 add up all the scores:

$$\sum_{i=1}^n x_i = 22 + 40 + 53 + 57 + 93 + 98 + 103 + 108 + 116 + 121 + 234 = 1045 \tag{1.2}$$

We then divide by the number of scores (in this case 11) as in equation (1.3):

$$\bar{X} = \frac{\sum_{i=1}^n x_i}{n} = \frac{1045}{11} = 95 \tag{1.3}$$

The mean is 95 friends, which is not a value we observed in our actual data. In this sense the mean is a statistical model – more on this in the [next chapter](#).



If you calculate the mean without our most popular person (i.e., excluding the value 234), the mean drops to 81.1 friends. This reduction illustrates one disadvantage of the mean: it can be influenced by extreme scores. In this case, the person with 234 friends on Facebook increased the mean by about 14 friends; compare this difference with that of the median. Remember that the median changed very little – from 98 to 95.5 – when we excluded the score of 234, which illustrates how the median is typically 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 so often? 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 (more on that later too).

Cramming Sam’s Tips Central tendency



- The mean is the sum of all scores divided by the number of scores. The value of the mean can be influenced quite heavily by extreme scores.
- The median is the middle score when the scores are placed in ascending order. It is not as influenced by extreme scores as the mean.
- The mode is the score that occurs most frequently.



1.8.5 The dispersion in a distribution



It can also be interesting to quantify the spread, or dispersion, of scores. 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 data we saw that if we order the scores we get 22, 40, 53, 57, 93, 98, 103, 108, 116, 121, 234. The highest score is 234 and the lowest is 22; therefore, the range is $234 - 22 = 212$. One problem with the range is that because it uses only the highest and lowest score, it is affected dramatically by extreme scores.



Compute the range but excluding the score of 234.

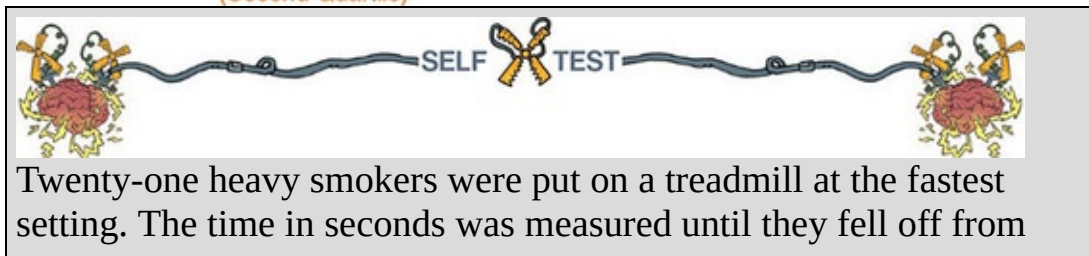
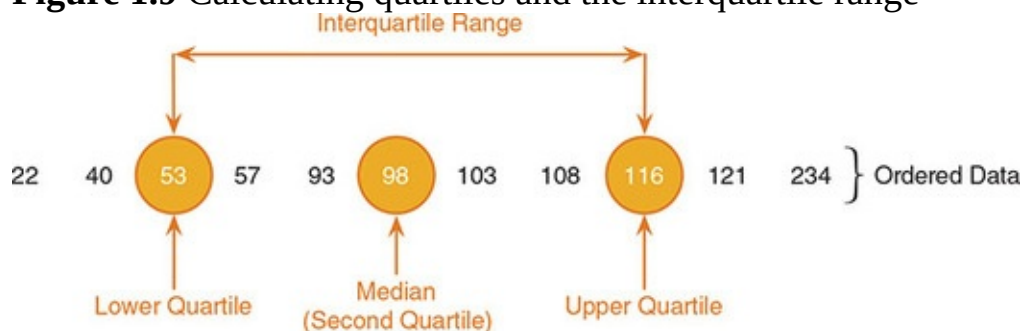
If you have done the self-test task you'll see that without the extreme score the range drops from 212 to 99 – less than half the size.

One way around this problem is to calculate the range but excluding 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. As a rule of thumb 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 1.9](#) shows how we would calculate these values for the Facebook 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 of two values in the data set (therefore, the upper and lower quartile need not be values that actually appear in the data). 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).

It's worth noting here that quartiles are special cases of things called **quantiles**. Quantiles are values that split a data set into equal portions. Quartiles are quantiles that split the data into four equal parts, but there are other quantiles such as **percentiles** (points that split the data into 100 equal parts), **noniles** (points that split the data into nine equal parts) and so on.

Figure 1.9 Calculating quartiles and the interquartile range



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.

If we want to use all the data rather than half of it, we can calculate the spread of scores by looking at how different each score is from the centre of the distribution. If we use the mean as a measure of the centre of a distribution, then we can calculate the difference between each score and the mean, which is known as the **deviance** (Eq. 1.4):

$$\text{deviance} = x_i - \bar{x} \quad (1.4)$$

If we want to know the total deviance then we could add up the deviances for each data point. In equation form, this would be:

$$\text{total deviance} = \sum_{i=1}^n (x_i - \bar{x}) \quad (1.5)$$

The sigma symbol (Σ) means ‘add up all of what comes after’, and the ‘what comes after’ in this case is the deviances. So, this equation simply means ‘add up all of the deviances’.

Let’s try this with the Facebook data. [Table 1.2](#) shows the number of friends for each person in the Facebook data, the mean, and the difference between the two. Note that because the mean is at the centre of the distribution, some of the deviations are positive (scores greater than the mean) and some are negative (scores smaller than the mean). Consequently, when we add the scores up, the total is zero. Therefore, the ‘total spread’ is nothing. This conclusion is as silly as a tapeworm thinking they can have a coffee with the Queen of England if they don a bowler hat and pretend to be human. Everyone knows that the Queen drinks tea.

To overcome this problem, we could ignore the minus signs when we add the deviations up. There’s nothing wrong with doing this, but people tend to square the deviations, which has a similar effect (because a negative number multiplied by another negative number becomes positive). The final column of [Table 1.2](#) shows these squared deviances. We can add these squared deviances up to get the **sum of squared errors, SS** (often just called the *sum of squares*); unless your scores are all exactly the same, the resulting value will be bigger than zero, indicating that there is some deviance from the mean. As an equation, we would write: equation (1.6), in which the sigma symbol means ‘add up all of the things that follow’ and what follows is the squared deviances (or *squared errors* as they’re more commonly known):

$$\text{sum of squared errors (SS)} = \sum_{i=1}^n (x_i - \bar{x})^2 \quad (1.6)$$

We can use the sum of squares as an indicator of the total dispersion, or total deviance of scores from the mean. The problem with using the total is that its size will depend on how many scores we have in the data. The sum of squares for the Facebook data is 32,246, but if we added another 11 scores that value would increase (other things being equal, it will more or less double in size). The total dispersion is a bit of a nuisance then because we can't compare it across samples that differ in size. Therefore, it can be useful to work not with the *total* dispersion, but the *average* dispersion, which is also known as the **variance**. We have seen that an average is the total of scores divided by the number of scores, therefore, the variance is simply the sum of squares divided by the number of observations (N). Actually, we normally divide the SS by the number of observations minus 1 as in equation (1.7) (the reason why is explained in the [next chapter](#) and [Jane Superbrain Box 2.2](#)):

Table 1.2 Table showing the deviations of each score from the mean

Number of Friends (x_i)	Mean (\bar{x})	Deviance ($x_i - \bar{x}$)	Deviance squared ($(x_i - \bar{x})^2$)
22	95	-73	5329
40	95	-55	3025
53	95	-42	1764
57	95	-38	1444
93	95	-2	4
98	95	3	9
103	95	8	64
108	95	13	169
116	95	21	441
121	95	26	676
234	95	139	19321
		$\sum_{i=1}^n x_i - \bar{x} = 0$	$\sum_{i=1}^n (x_i - \bar{x})^2 = 32246$

$$\text{variance}(s^2) = \frac{\text{SS}}{N-1} = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{N-1} = \frac{32,246}{10} = 3224.6 \quad (1.7)$$

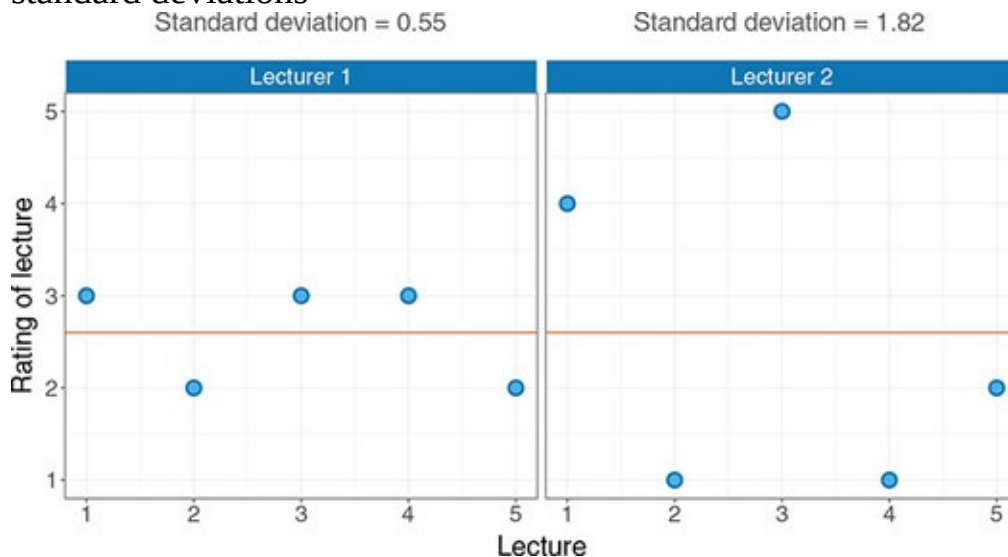
As we have seen, the variance is the average error between the mean and the observations made. There is one problem with the variance as a measure: it gives us a measure in units squared (because we squared each error in the calculation). In our example we would have to say that the average error in our data was 3224.6 friends squared. It makes very little sense to talk about friends squared, so we often take the square root of the variance (which ensures that the measure of average error is in the same units as the original measure). This measure is known as the **standard deviation** and is the square root of the variance (Eq.

1.8).

$$s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{N-1}} \quad (1.8)$$
$$= \sqrt{3224.6}$$
$$= 56.79$$

The sum of squares, variance and standard deviation are all measures of the dispersion or spread of data around the mean. A small standard deviation (relative to the value of the mean itself) indicates that the data points are close to the mean. A large standard deviation (relative to the mean) indicates that the data points are distant from the mean. A standard deviation of 0 would mean that all the scores were the same. [Figure 1.10](#) shows the overall ratings (on a 5-point scale) of two lecturers after each of five different lectures. Both lecturers had an average rating of 2.6 out of 5 across the lectures. However, the first lecturer had a standard deviation of 0.55 (relatively small compared to the mean). It should be clear from the left-hand graph that ratings for this lecturer were consistently close to the mean rating. There was a small fluctuation, but generally her lectures did not vary in popularity. Put another way, the scores are not spread too widely around the mean. The second lecturer, however, had a standard deviation of 1.82 (relatively high compared to the mean). The ratings for this second lecturer are more spread from the mean than the first: for some lectures she received very high ratings, and for others her ratings were appalling.

Figure 1.10 Graphs illustrating data that have the same mean but different standard deviations



1.8.6 Using a frequency distribution to go beyond the

data

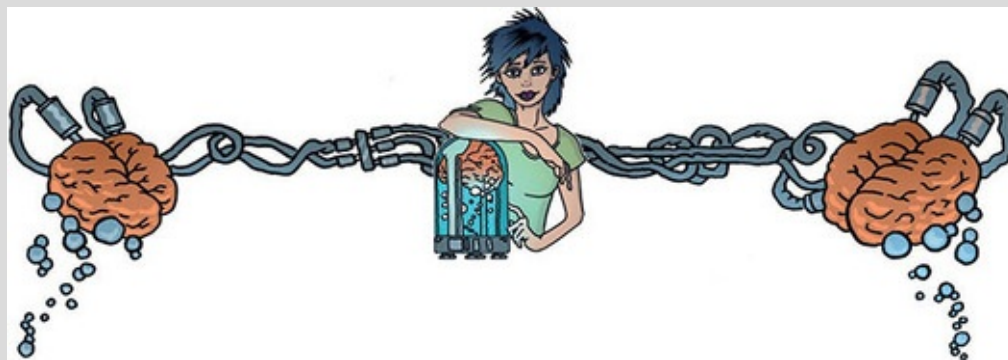


Another way to think about frequency distributions is not in terms of how often scores actually occurred, but how likely it is that a score would occur (i.e., probability). The word ‘probability’ causes most people’s brains to overheat (myself included) so it seems fitting that we use an example about throwing buckets of ice over our heads. Internet memes tend to follow the shape of a normal distribution, which we discussed a while back. A good example of this is the ice bucket challenge from 2014. You can check Wikipedia for the full story, but it all started (arguably) with golfer Chris Kennedy tipping a bucket of iced water on his head to raise awareness of the disease amyotrophic lateral sclerosis (ALS, also known as Lou Gehrig’s disease).¹⁹ The idea is that you are challenged and have 24 hours to post a video of you having a bucket of iced water poured over your head; in this video you also challenge at least three other people. If you fail to complete the challenge your forfeit is to donate to charity (in this case, ALS). In reality many people completed the challenge *and* made donations.

¹⁹ Chris Kennedy did not invent the challenge, but he’s believed to be the first to link it to ALS. There are earlier reports of people doing things with ice-cold water in the name of charity, but I’m focusing on the ALS challenge because it is the one that spread as a meme.

Jane Superbrain 1.5 The standard deviation and the shape of the

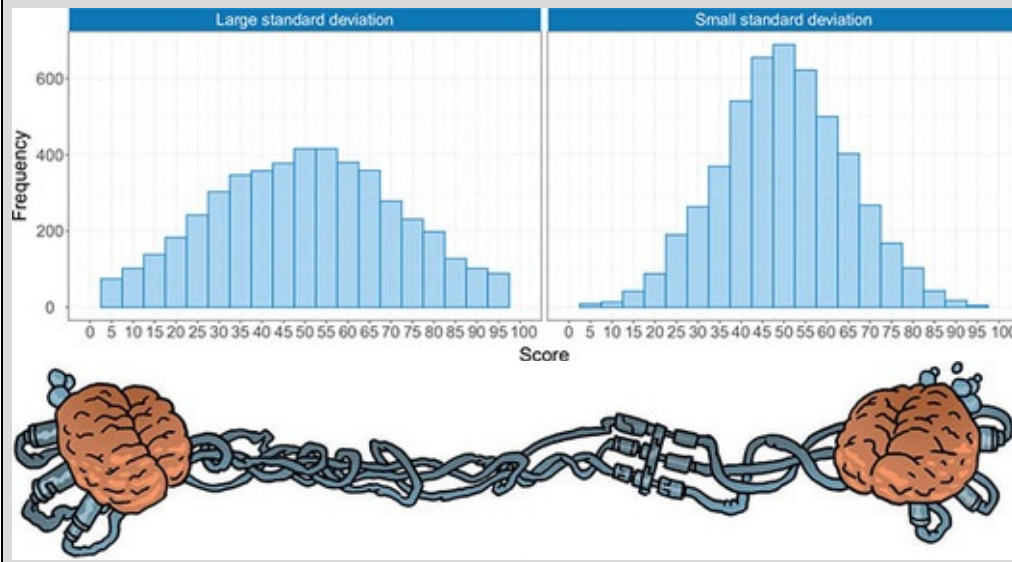
distribution 



The variance and standard deviation tell us about the shape of the distribution of scores. If the mean represents the data well then most of the scores will cluster close to the mean and the resulting standard deviation is small relative to the mean. When the mean is a worse

representation of the data, the scores cluster more widely around the mean and the standard deviation is larger. [Figure 1.11](#) shows two distributions that have the same mean (50) but different standard deviations. One has a large standard deviation relative to the mean ($SD = 25$) and this results in a flatter distribution that is more spread out, whereas the other has a small standard deviation relative to the mean ($SD = 15$) resulting in a pointier distribution in which scores close to the mean are very frequent but scores further from the mean become increasingly infrequent. The message is that as the standard deviation gets larger, the distribution gets fatter. This can make distributions look platykurtic or leptokurtic when, in fact, they are not.

Figure 1.11 Two distributions with the same mean, but large and small standard deviations

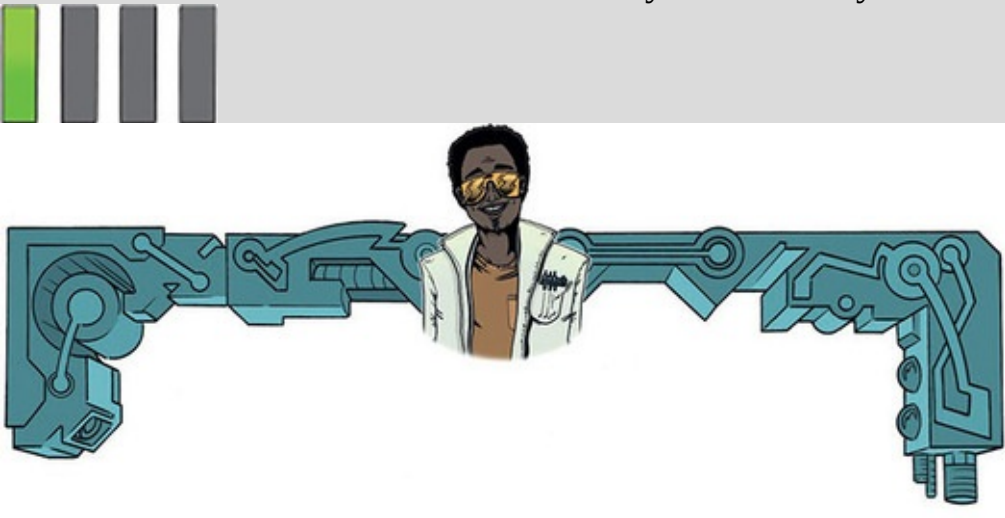


The ice bucket challenge is a good example of a meme: it ended up generating something like 2.4 million videos on Facebook and 2.3 million on YouTube. I mentioned that memes often follow a normal distribution, and [Figure 1.12](#) shows this: the insert shows the ‘interest’ score from Google Trends for the phrase ‘ice bucket challenge’ from August to September 2014.²⁰ The ‘interest’ score that Google calculates is a bit hard to unpick but essentially reflects the relative number of times that the term ‘ice bucket challenge’ was searched for on Google. It’s not the total number of searches, but the relative number. In a sense it shows the trend of the popularity of searching for ‘ice bucket challenge’. Compare the line with the perfect normal distribution in [Figure 1.3](#) – they look fairly similar, don’t they? Once it got going (about 2–3 weeks after the first video) it went viral, and popularity increased rapidly, reaching a peak at around

21 August (about 36 days after Chris Kennedy got the ball rolling). After this peak, popularity rapidly declines as people tire of the meme.

[20](#) You can generate the insert graph for yourself by going to Google Trends, entering the search term ‘ice bucket challenge’ and restricting the dates shown to August 2014 to September 2014.

Labcoat Leni's Real Research 1.1 Is Friday 13th unlucky?

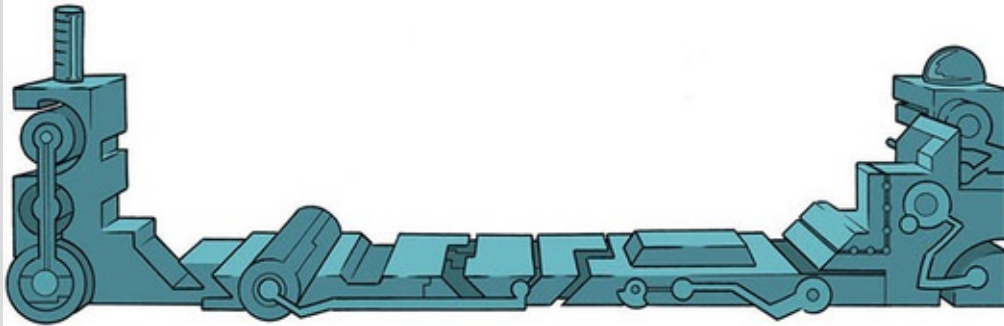


Scanlon, T. J., et al. (1993). *British Medical Journal*, 307, 1584–1586.

Many of us are superstitious, and a common superstition is that Friday the 13th is unlucky. Most of us don't literally think that someone in a hockey mask is going to kill us, but some people are wary. Scanlon and colleagues, in a tongue-in-cheek study (Scanlon, Luben, Scanlon, & Singleton, 1993), looked at accident statistics at hospitals in the south-west Thames region of the UK. They took statistics both for Friday the 13th and Friday the 6th (the week before) in different months in 1989, 1990, 1991 and 1992. They looked at both emergency admissions of accidents and poisoning, and also transport accidents.

Date	Accidents and Poisoning		Traffic Accidents	
	Friday 6th	Friday 13th	Friday 6th	Friday 13th
October 1989	4	7	9	13
July 1990	6	6	6	12
September 1991	1	5	11	14
December 1991	9	5	11	10
March 1992	9	7	3	4
November 1992	1	6	5	12

Calculate the mean, median, standard deviation and interquartile range for each type of accident and on each date. Answers are on the companion website.



Cramming Sam's Tips Dispersion



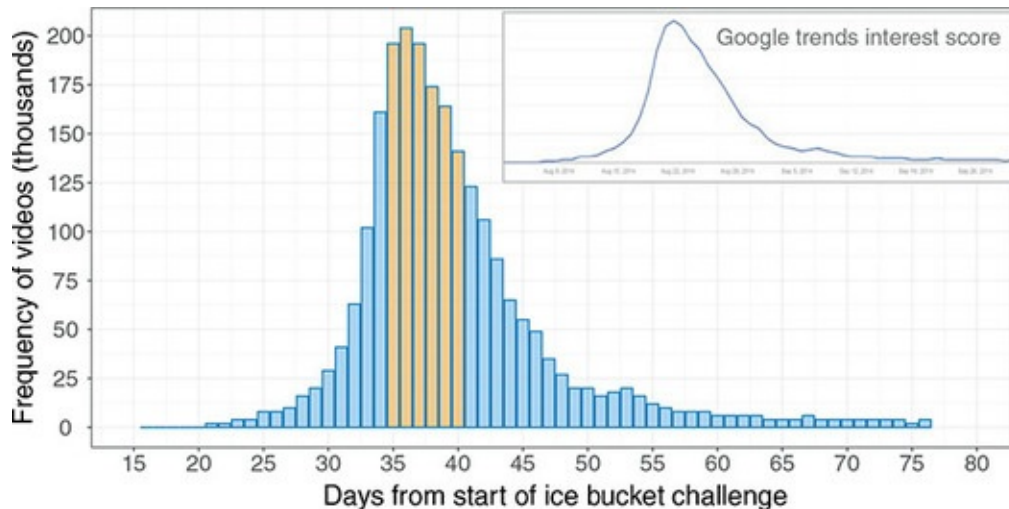
- The deviance or error is the distance of each score from the mean.
- The sum of squared errors is the total amount of error in the mean. The errors/deviances are squared before adding them up.
- The variance is the average distance of scores from the mean. It is the sum of squares divided by the number of scores. It tells us about how widely dispersed scores are around the mean.
- The standard deviation is the *square root of the variance*. It is the variance converted back to the original units of measurement of the scores used to compute it. Large standard deviations relative to the mean suggest data are widely spread around the mean, whereas small standard deviations suggest data are closely packed around the mean.
- The range is the distance between the highest and lowest score.
- The interquartile range is the range of the middle 50% of the scores.



The main histogram in [Figure 1.12](#) shows the same pattern but reflects something a bit more tangible than ‘interest scores’. It shows the number of videos posted on YouTube relating to the ice bucket challenge on each day after Chris Kennedy’s initial challenge. There were 2323 thousand in total (2.32 million) during the period shown. In a sense it shows approximately how many people took up the challenge each day.²¹ You can see that nothing much happened for 20 days, and early on relatively few people took up the challenge. By about 30 days after the initial challenge things are hotting up (well, cooling down, really) as the number of videos rapidly accelerated from 29,000 on day 30 to 196,000 on day 35. At day 36, the challenge hits its peak (204,000 videos posted) after which the decline sets in as it becomes ‘yesterday’s news’. By day 50 it’s only the type of people like me, and statistics lectures more generally, who don’t check Facebook for 50 days, who suddenly become aware of the meme and want to get in on the action to prove how down with the kids we are. It’s too late, though: people at that end of the curve are uncool, and the trendsetters who posted videos on day 25 call us lame and look at us dismissively. It’s OK though, because we can plot sick histograms like the one in [Figure 1.12](#); take that, hipster scum!

²¹ Very very approximately indeed. I have converted the Google interest data into videos posted on YouTube by using the fact that I know that 2.33 million videos were posted during this period and by making the (not unreasonable) assumption that behaviour on YouTube will have followed the same pattern over time as the Google interest score for the challenge.

Figure 1.12 Frequency distribution showing the number of ice bucket challenge videos on YouTube by day since the first video (the insert shows the actual Google Trends data on which this example is based)

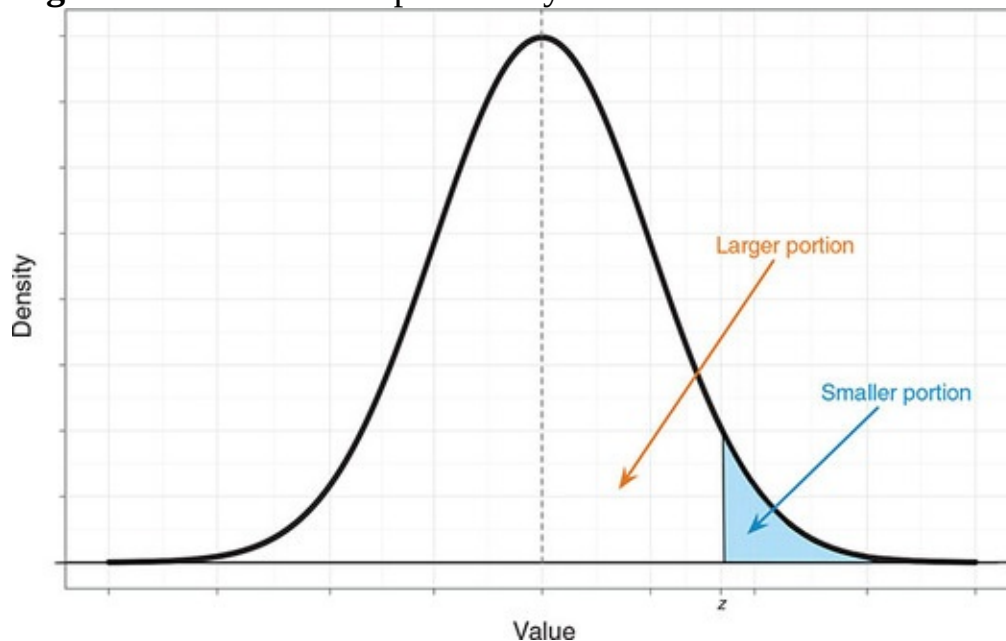


I digress. We can think of frequency distributions in terms of probability. To explain this, imagine that someone asked you ‘How likely is it that a person posted an ice bucket video after 60 days?’ What would your answer be? Remember that the height of the bars on the histogram reflects how many videos were posted. Therefore, if you looked at the frequency distribution before answering the question you might respond ‘not very likely’ because the bars are very short after 60 days (i.e., relatively few videos were posted). What if someone asked you ‘How likely is it that a video was posted 35 days after the challenge started?’ Using the histogram, you might say ‘It’s relatively likely’ because the bar is very high on day 35 (so quite a few videos were posted). Your inquisitive friend is on a roll and asks ‘How likely is it that someone posted a video 35 to 40 days after the challenge started?’ The bars representing these days are shaded orange in [Figure 1.12](#). The question about the likelihood of a video being posted 35-40 days into the challenge is really asking ‘How big is the orange area of [Figure 1.12](#) compared to the total size of all bars?’ We can find out the size of the dark blue region by adding the values of the bars ($196 + 204 + 196 + 174 + 164 + 141 = 1075$); therefore, the orange area represents 1075 thousand videos. The total size of all bars is the total number of videos posted (i.e., 2323 thousand). If the orange area represents 1075 thousand videos, and the total area represents 2323 thousand videos, then if we compare the orange area to the total area we get $1075/2323 = 0.46$. This proportion can be converted to a percentage by multiplying by 100, which gives us 46%. Therefore, our answer might be ‘It’s quite likely that someone posted a video 35-40 days into the challenge because 46% of all videos were posted during those 6 days’. A very important point here is that the size of the bars relates directly to the probability of an event occurring.

Hopefully these illustrations show that we can use the frequencies of different

scores, and the area of a frequency distribution, to estimate the probability that a particular score will occur. 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 July. However, when I talk to anyone else, I might, more realistically, tell them that there's a 0.10 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). If probabilities don't make sense to you then you're not alone; just ignore the decimal point and think of them as percentages instead (i.e., a 0.10 probability that something will happen is a 10% chance that something will happen) or read the chapter on probability in my other excellent textbook (Field, 2016).

Figure 1.13 The normal probability distribution



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 (known as **probability density functions, PDF**) that specify idealized versions of these distributions. We could draw such a function by plotting the value of the variable (x) against the probability of it occurring (y).²² The resulting curve is known as a **probability distribution**; for a normal distribution

(Section 1.8.1) it would look like [Figure 1.13](#), which has the characteristic bell shape that we saw already in [Figure 1.3](#).

22 Actually we usually plot something called the *density*, which is closely related to the probability.

A probability distribution is just like a histogram except that the lumps and bumps have been smoothed out so that we see a nice smooth curve. However, like a frequency distribution, the area under this curve tells us something about the probability of a value occurring. Just like we did in our ice bucket example, we could use the area under the curve between two values to tell us how likely it is that a score fell within a particular range. For example, the blue shaded region in [Figure 1.13](#) corresponds to the probability of a score being z or greater. The normal distribution is not the only distribution that has been precisely specified by people with enormous brains. There are many distributions that have characteristic shapes and have been specified with a probability density function. We'll encounter some of these other distributions throughout the book, for example the t -distribution, chi-square (χ^2) distribution, and F -distribution. For now, the important thing to remember is that all of these distributions have something in common: they are all defined by an equation that enables us to calculate precisely the probability of obtaining a given score.

As we have seen, distributions can have different means and standard deviations. This isn't a problem for the probability density function – it will still give us the probability of a given value occurring – but it is a problem for us because probability density functions are difficult enough to spell, let alone use to compute probabilities. Therefore, to avoid a brain meltdown we often use a normal distribution with a mean of 0 and a standard deviation of 1 as a standard. This has the advantage that we can pretend that the probability density function doesn't exist and use tabulated probabilities (as in the Appendix) instead. The obvious problem is that not all of the data we collect will have a mean of 0 and a standard deviation of 1. For example, for the ice bucket data the mean is 39.68 and the standard deviation is 7.74. However, 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 (X) and subtract from it the mean of all scores (\bar{X}). To ensure the data have a standard deviation of 1, we divide the resulting score by the standard deviation (s), which we recently encountered. The resulting scores are denoted by the letter z and are known as **z-scores**. In equation form, the conversion that I've just described is:



$$z = \frac{X - \bar{X}}{s} \quad (1.9)$$

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 ice bucket data, we can answer the question ‘What’s the probability that someone posted a video on day 60 or later?’ First, we convert 60 into a z-score. We saw that the mean was 39.68 and the standard deviation was 7.74, so our score of 60 expressed as a z-score is 2.63 (Eq. 1.10):

$$z = \frac{60 - 39.68}{7.74} = 2.63 \quad (1.10)$$

We can now use this value, rather than the original value of 60, to compute an answer to our question.

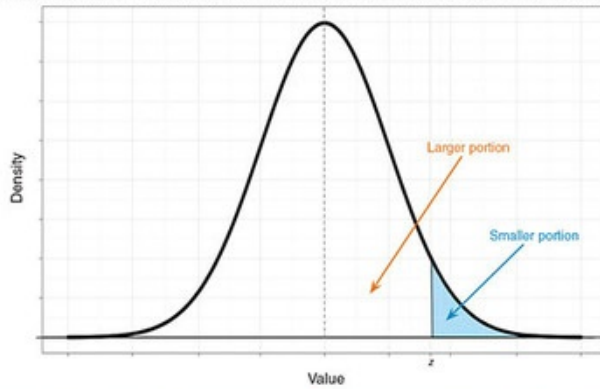
[Figure 1.14](#) shows (an edited version of) the tabulated values of the standard normal distribution from the Appendix of this book. This table gives us a list of values of z , and the density (y) for each value of z , but, most important, it splits the distribution at the value of z and tells us the size of the two areas under the curve that this division creates. For example, when z is 0, we are at the mean or centre of the distribution so it splits the area under the curve exactly in half. Consequently, both areas have a size of 0.5 (or 50%). However, any value of z that is not zero will create different sized areas, and the table tells us the size of the larger and smaller portions. For example, if we look up our z -score of 2.63, we find that the smaller portion (i.e., the area above this value, or the blue area in [Figure 1.14](#)) is 0.0043, or only 0.43%. I explained before that these areas relate to probabilities, so in this case we could say that there is only a 0.43% chance that a video was posted 60 days or more after the challenge started. By looking at the larger portion (the area below 2.63) we get 0.9957, or put another way, there’s a 99.57% chance that an ice bucket video was posted on YouTube within 60 days of the challenge starting. Note that these two proportions add up to 1 (or 100%), so the total area under the curve is 1.

Another useful thing we can do (you’ll find out just how useful in due course) is to work out limits within which a certain percentage of scores fall. With our ice bucket example, we looked at how likely it was that a video was posted between 35 and 40 days after the challenge started; we could ask a similar question such as ‘What is the range of days between which the middle 95% of videos were posted?’ To answer this question we need to use the table the opposite way

around. We know that the total area under the curve is 1 (or 100%), so to discover the limits within which 95% of scores fall we're asking 'What is the value of z that cuts off 5% of the scores?' It's not quite as simple as that because if we want the *middle* 95%, then we want to cut off scores from both ends. Given the distribution is symmetrical, if we want to cut off 5% of scores overall but we want to take some from both extremes of scores, then the percentage of scores we want to cut from each end will be $5\%/2 = 2.5\%$ (or 0.025 as a proportion). If we cut off 2.5% of scores from each end then in total we'll have cut off 5% scores, leaving us with the middle 95% (or 0.95 as a proportion) – see [Figure 1.15](#). To find out what value of z cuts off the top area of 0.025, we look down the column 'smaller portion' until we reach 0.025, we then read off the corresponding value of z . This value is 1.96 (see [Figure 1.14](#)) and because the distribution is symmetrical around zero, the value that cuts off the bottom 0.025 will be the same but a minus value (-1.96). Therefore, the middle 95% of z -scores fall between -1.96 and 1.96 . If we wanted to know the limits between which the middle 99% of scores would fall, we could do the same: now we would want to cut off 1% of scores, or 0.5% from each end. This equates to a proportion of 0.005. We look up 0.005 in the *smaller portion* part of the table and the nearest value we find is 0.00494, which equates to a z -score of 2.58 (see [Figure 1.14](#)). This tells us that 99% of z -scores lie between -2.58 and 2.58 . Similarly (have a go), you can show that 99.9% of them lie between -3.29 and 3.29 . Remember these values (1.96, 2.58 and 3.29) because they'll crop up time and time again.

Figure 1.14 Using tabulated values of the standard normal distribution

A.1. Table of the standard normal distribution

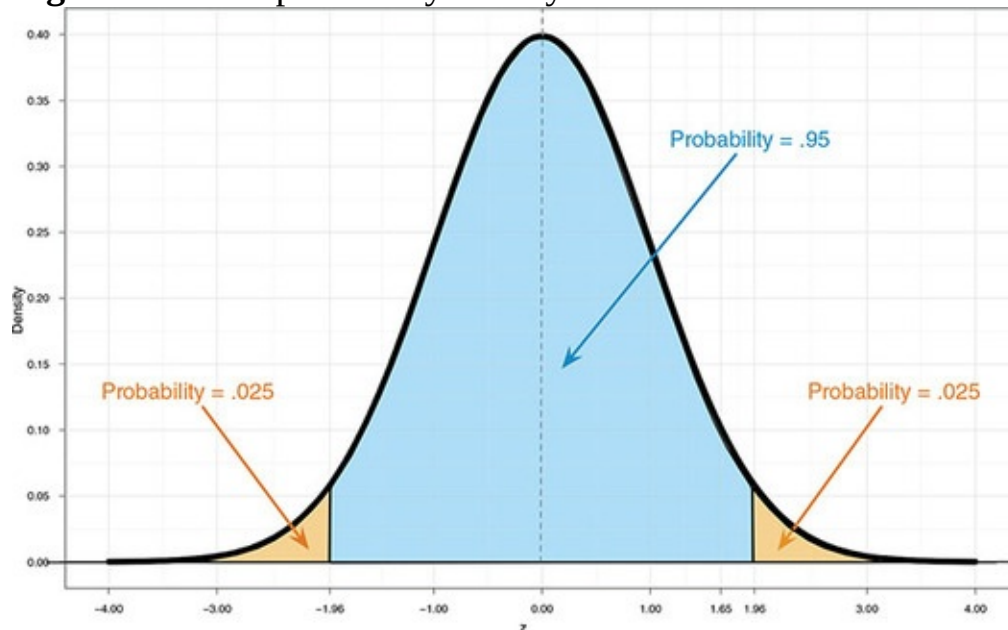


z	Larger Portion	Smaller Portion	y	z	Larger Portion	Smaller Portion	y
.00	.50000	.50000	.3989	.12	.54776	.45224	.3961
.01	.50399	.49601	.3989	.13	.55172	.44828	.3956
.02	.50798	.49202	.3989	.14	.55567	.44433	.3951
.03	.51197	.48803	.3988	.15	.55962	.44038	.3945
.04	.51595	.48405	.3986	.16	.56356	.43644	.3939

1.56	.94062	.05938	.1182	1.86	.96856	.03144	.0707
1.57	.94179	.05821	.1163	1.87	.96926	.03074	.0694
1.58	.94295	.05705	.1145	1.88	.96995	.03005	.0681
1.59	.94408	.05592	.1127	1.89	.97062	.02938	.0669
1.60	.94520	.05480	.1109	1.90	.97128	.02872	.0656
1.61	.94630	.05370	.1092	1.91	.97193	.02807	.0644
1.62	.94738	.05262	.1074	1.92	.97257	.02743	.0632
1.63	.94845	.05155	.1057	1.93	.97320	.02680	.0620
1.64	.94950	.05050	.1040	1.94	.97381	.02619	.0608
1.65	.95053	.04947	.1023	1.95	.97441	.02559	.0596
1.66	.95154	.04846	.1006	1.96	.97500	.02500	.0584
1.67	.95254	.04746	.0989	1.97	.97558	.02442	.0573
1.68	.95352	.04648	.0973	1.98	.97615	.02385	.0562

2.27	.98840	.01160	.0303	2.57	.99492	.00508	.0147
2.28	.98870	.01130	.0297	2.58	.99500	.00494	.0143
2.29	.98899	.01101	.0290	2.59	.99520	.00480	.0139
2.30	.98928	.01072	.0283	2.60	.99534	.00466	.0136
2.31	.98956	.01044	.0277	2.61	.99547	.00453	.0132
2.32	.98983	.01017	.0270	2.62	.99560	.00440	.0129
2.33	.99010	.00990	.0264	2.63	.99573	.00427	.0126

Figure 1.15 The probability density function of a normal distribution



Assuming the same mean and standard deviation for the ice bucket example above, what's the probability that someone posted a video within the first 30 days of the challenge?

Cramming Sam's Tips Distributions and z-scores



- A frequency distribution can be either a table or a chart that shows each possible score on a scale of measurement along with the number of times that score occurred in the data.
- Scores are sometimes expressed in a standard form known as z-scores.
- To transform a score into a z-score you subtract from it the mean of all scores and divide the result by the standard deviation of all scores.

- The sign of the z-score tells us whether the original score was above or below the mean; the value of the z-score tells us how far the score was from the mean in standard deviation units.



1.8.7 Fitting statistical models to the data

Having looked at your data (and there is a lot more information on different ways to do this in [Chapter 5](#)), the next step of the research process is to fit a statistical model to the data. That is to go where eagles dare, and no one should fly where eagles dare; but to become scientists we have to, so the rest of this book attempts to guide you through the various models that you can fit to the data.

1.9 Reporting data

1.9.1 Dissemination of research

Having established a theory and collected and started to summarize data, you might want to tell other people what you have found. This sharing of information is a fundamental part of being a scientist. As discoverers of knowledge, we have a duty of care to the world to present what we find in a clear and unambiguous way, and with enough information that others can challenge our conclusions. It is good practice, for example, to make your data available to others and to be open with the resources you used. Initiatives such as the Open Science Framework (<https://osf.io>) make this easy to do. Tempting as it may be to cover up the more unsavoury aspects of our results, science is about truth, openness and willingness to debate your work.

Scientists tell the world about our findings by presenting them at conferences and in articles published in scientific **journals**. A scientific journal is a collection of articles written by scientists on a vaguely similar topic. A bit like a magazine, but more tedious. These articles can describe new research, review existing research, or might put forward a new theory. Just like you have magazines such as *Modern Drummer*, which is about drumming, or *Vogue*, which is about