Correlation

6

FIGURE 6.1
I don't have a photo from Christmas 1981, but this was taken about that time at my grandparents' house. I'm trying to play an 'E' by the looks of it, no doubt because it's in 'Take on the World'.

6.1. What will this chapter tell me? ①

When I was 8 years old, my parents bought me a guitar for Christmas. Even then, I'd desperately wanted to play the guitar for years. I could not contain my excitement at getting this gift (had it been an *electric* guitar I think I would have actually exploded with excitement). The guitar came with a 'learn to play' book and, after a little while of trying to play what was on page 1 of this book, I readied myself to unleash a riff of universe-crushing power onto the world (well, 'Skip to my Lou' actually). But, I couldn't do it. I burst into

tears and ran upstairs to hide.[1] My dad sat with me and said 'Don't worry, Andy, everything is hard to begin with, but the more you practise the easier it gets.' In his comforting words, my dad was inadvertently teaching me about the relationship, or correlation, between two variables. These two variables could be related in three ways: (1) *positively related*, meaning that the more I practised my guitar, the better a guitar player I would become (i.e., my dad was telling me the truth); (2) *not related* at all, meaning that as I practised the guitar my playing ability would remain completely constant (i.e., my dad has fathered a cretin); or (3) *negatively related*, which would mean that the more I practised my guitar the worse a guitar player I would become (i.e., my dad has fathered an indescribably strange child). This chapter looks first at how we can express the relationships between variables statistically by looking at two measures: *covariance* and the *correlation coefficient*. We then discover how to carry out and interpret correlations in **R**. The chapter ends by looking at more complex measures of relationships; in doing so it acts as a precursor to multiple regression, which we discuss in Chapter 7.

6.2. Looking at relationships ①

In Chapter 4 I stressed the importance of looking at your data graphically before running any other analysis on them. I just want to begin by reminding you that our first starting point with a correlation analysis should be to look at some scatterplots of the variables we have measured. I am not going to repeat how to get **R** to produce these graphs, but I am going to urge you (if you haven't done so already) to read section 4.5 before embarking on the rest of this chapter.

6.3. How do we measure relationships? ①

6.3.1. A detour into the murky world of covariance ①

The simplest way to look at whether two variables are associated is to look at whether they *covary*. To understand what **covariance** is, we first need to think back to the concept of variance that we met in Chapter 2. Remember that the variance of a single variable represents the average amount that the data vary from the mean. Numerically, it is described by:

$$\text{Variance}(s^2) = \frac{\sum(x_i - \bar{x})^2}{N-1} = \frac{\sum(x_i - \bar{x})(x_i - \bar{x})}{N-1} \tag{6.1}$$

The mean of the sample is represented by $\bar{x}$, x_i is the data point in question and N is the number of observations (see section 2.4.1). If we are interested in whether two variables are related, then we are interested in whether changes in one variable are met with similar changes in the other variable. Therefore, when one variable deviates from its mean we would expect the other variable to deviate from its mean in a similar way. To illustrate what I mean, imagine we took five people and subjected them to a certain number of advertisements promoting toffee sweets, and then measured how many packets of those sweets each

[1] This is not a dissimilar reaction to the one I have when publishers ask me for new editions of statistics textbooks.

Table 6.1 Adverts watched and toffee purchases

Participant:	1	2	3	4	5	Mean	s
Adverts watched	5	4	4	6	8	5.4	1.67
Packets bought	8	9	10	13	15	11.0	2.92

person bought during the next week. The data are in Table 6.1 as well as the mean and standard deviation (*s*) of each variable.

If there were a relationship between these two variables, then as one variable deviates from its mean, the other variable should deviate from its mean in the same or the directly opposite way. Figure 6.2 shows the data for each participant (light blue circles represent the number of packets bought and dark blue circles represent the number of adverts watched); the grey line is the average number of packets bought and the blue line is the average number of adverts watched. The vertical lines represent the differences (remember that these differences are called *deviations*) between the observed values and the mean of the relevant variable. The first thing to notice about Figure 6.2 is that there is a very similar pattern of deviations for both variables. For the first three participants the observed values are below the mean for both variables, for the last two people the observed values are above the mean for both variables. This pattern is indicative of a potential relationship between the two variables (because it seems that if a person's score is below the mean for one variable then their score for the other will also be below the mean).

So, how do we calculate the exact similarity between the patterns of differences of the two variables displayed in Figure 6.2? One possibility is to calculate the total amount of deviation but we would have the same problem as in the single variable case: the positive and negative deviations would cancel out (see section 2.4.1). Also, by simply adding the deviations, we would gain little insight into the relationship between the variables. Now, in the single variable case, we squared the deviations to eliminate the problem of positive and negative deviations cancelling out each other. When there are two variables, rather than squaring each deviation, we can multiply the deviation for one variable by the corresponding deviation for the second variable. If both deviations are positive or negative then this will give us a positive value (indicative of the deviations being in the same direction), but

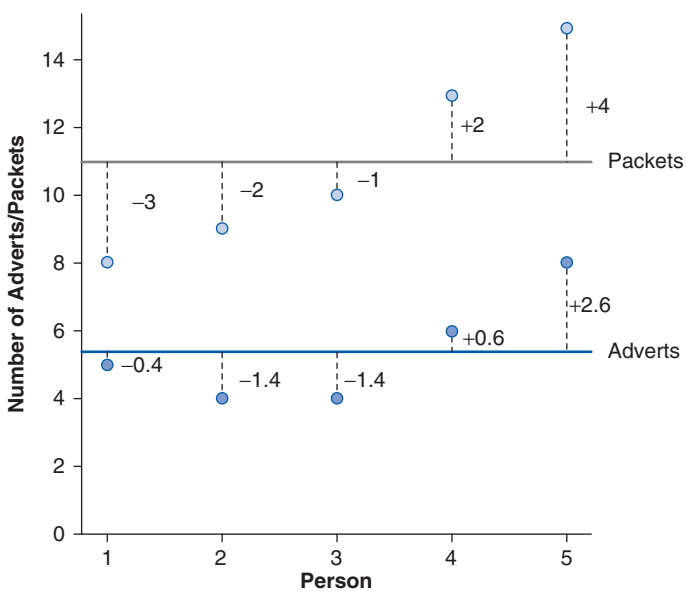

FIGURE 6.2 Graphical display of the differences between the observed data and the means of two variables

if one deviation is positive and one negative then the resulting product will be negative (indicative of the deviations being opposite in direction). When we multiply the deviations of one variable by the corresponding deviations of a second variable, we get what is known as the **cross-product deviations**. As with the variance, if we want an average value of the combined deviations for the two variables, we must divide by the number of observations (we actually divide by $N - 1$ for reasons explained in Jane Superbrain Box 2.2). This averaged sum of combined deviations is known as the **covariance**. We can write the covariance in equation form as in equation (6.2) – you will notice that the equation is the same as the equation for variance, except that instead of squaring the differences, we multiply them by the corresponding difference of the second variable:

$$\text{cov}(x,y) = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{N-1} \tag{6.2}$$

For the data in Table 6.1 and Figure 6.2 we reach the following value:

$$\begin{aligned}\text{cov}(x,y) &= \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{N-1} \\ &= \frac{(-0.4)(-3) + (-1.4)(-2) + (-1.4)(-1) + (0.6)(2) + (2.6)(4)}{4} \\ &= \frac{1.2 + 2.8 + 1.4 + 1.2 + 10.4}{4} \\ &= \frac{17}{4} \\ &= 4.25\end{aligned}$$

Calculating the covariance is a good way to assess whether two variables are related to each other. A positive covariance indicates that as one variable deviates from the mean, the other variable deviates in the same direction. On the other hand, a negative covariance indicates that as one variable deviates from the mean (e.g., increases), the other deviates from the mean in the opposite direction (e.g., decreases).

There is, however, one problem with covariance as a measure of the relationship between variables and that is that it depends upon the scales of measurement used. So, covariance is not a standardized measure. For example, if we use the data above and assume that they represented two variables measured in miles then the covariance is 4.25 (as calculated above). If we then convert these data into kilometres (by multiplying all values by 1.609) and calculate the covariance again then we should find that it increases to 11. This dependence on the scale of measurement is a problem because it means that we cannot compare covariances in an objective way – so, we cannot say whether a covariance is particularly large or small relative to another data set unless both data sets were measured in the same units.

6.3.2. Standardization and the correlation coefficient ①

To overcome the problem of dependence on the measurement scale, we need to convert the covariance into a standard set of units. This process is known as **standardization**. A very basic form of standardization would be to insist that all experiments use the same units of measurement, say metres – that way, all results could be easily compared. However, what happens if you want to measure attitudes – you'd be hard pushed to measure them

in metres. Therefore, we need a unit of measurement into which any scale of measurement can be converted. The unit of measurement we use is the *standard deviation*. We came across this measure in section 2.4.1 and saw that, like the variance, it is a measure of the average deviation from the mean. If we divide any distance from the mean by the standard deviation, it gives us that distance in standard deviation units. For example, for the data in Table 6.1, the standard deviation for the number of packets bought is approximately 3.0 (the exact value is 2.92). In Figure 6.2 we can see that the observed value for participant 1 was 3 packets less than the mean (so there was an error of –3 packets of sweets). If we divide this deviation, –3, by the standard deviation, which is approximately 3, then we get a value of –1. This tells us that the difference between participant 1's score and the mean was –1 standard deviation. So, we can express the deviation from the mean for a participant in standard units by dividing the observed deviation by the standard deviation.

It follows from this logic that if we want to express the covariance in a standard unit of measurement we can simply divide by the standard deviation. However, there are two variables and, hence, two standard deviations. Now, when we calculate the covariance we actually calculate two deviations (one for each variable) and then multiply them. Therefore, we do the same for the standard deviations: we multiply them and divide by the product of this multiplication. The standardized covariance is known as a **correlation coefficient** and is defined by equation (6.3), in which s_x is the standard deviation of the first variable and s_y is the standard deviation of the second variable (all other letters are the same as in the equation defining covariance):

$$r = \frac{\text{cov}_{xy}}{s_x s_y} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{(N-1)s_x s_y} \tag{6.3}$$

The coefficient in equation (6.3) is known as the **Pearson product-moment correlation coefficient** or **Pearson correlation coefficient** (for a really nice explanation of why it was originally called the 'product-moment' correlation, see Miles & Banyard, 2007) and was invented by Karl Pearson (see Jane Superbrain Box 6.1).[2] If we look back at Table 6.1 we see that the standard deviation for the number of adverts watched (s_x) was 1.67, and for the number of packets of crisps bought (s_y) was 2.92. If we multiply these together we get $1.67 \times 2.92 = 4.88$. Now, all we need to do is take the covariance, which we calculated a few pages ago as being 4.25, and divide by these multiplied standard deviations. This gives us $r = 4.25/4.88 = .87$.

By standardizing the covariance we end up with a value that has to lie between −1 and +1 (if you find a correlation coefficient less than −1 or more than +1 you can be sure that something has gone hideously wrong!). A coefficient of +1 indicates that the two variables are perfectly positively correlated, so as one variable increases, the other increases by a proportionate amount. Conversely, a coefficient of −1 indicates a perfect negative relationship: if one variable increases, the other decreases by a proportionate amount. A coefficient of zero indicates no linear relationship at all and so if one variable changes, the other stays the same. We also saw in section 2.6.4 that because the correlation coefficient is a standardized measure of an observed effect, it is a commonly used measure of the size of an effect and that values of ±.1 represent a small effect, ±.3 is a medium effect and ±.5 is a large effect (although I re-emphasize my caveat that these canned effect sizes are no substitute for interpreting the effect size within the context of the research literature).

[2] You will find Pearson's product-moment correlation coefficient denoted by both *r* and *R*. Typically, the uppercase form is used in the context of regression because it represents the multiple correlation coefficient; however, for some reason, when we square *r* (as in section 6.5.4.3) an upper case *R* is used. Don't ask me why – it's just to confuse me, I suspect.

JANE SUPERBRAIN 6.1

Who said statistics was dull? ①

Students often think that statistics is dull, but back in the early 1900s it was anything but dull, with various prominent figures entering into feuds on a soap opera scale. One of the most famous was between Karl Pearson and Ronald Fisher (whom we met in Chapter 2). It began when Pearson published a paper of Fisher's in his journal but made comments in his editorial that, to the casual reader, belittled Fisher's work. Two years later Pearson's group published work following on from Fisher's paper without consulting him. The antagonism persisted with Fisher turning down a job to work in Pearson's group and publishing 'improvements' on Pearson's ideas. Pearson for his part wrote in his own journal about apparent errors made by Fisher.

Another prominent statistician, Jerzy Neyman, criticized some of Fisher's most important work in a paper delivered to the Royal Statistical Society on 28 March 1935 at which Fisher was present. Fisher's discussion of the paper at that meeting directly attacked Neyman. Fisher more or less said that Neyman didn't know what he was talking about and didn't understand the background material on which his work was based. Relations soured so much that while they both worked at University College London, Neyman openly attacked many of Fisher's ideas in lectures to his students. The two feuding groups even took afternoon tea (a common practice in the British academic community of the time) in the same room but at different times! The truth behind who fuelled these feuds is, perhaps, lost in the mists of time, but Zabell (1992) makes a sterling effort to unearth it.

Basically, then, the founders of modern statistical methods were a bunch of squabbling children. Nevertheless, these three men were astonishingly gifted individuals. Fisher, in particular, was a world leader in genetics, biology and medicine as well as possibly the most original mathematical thinker ever (Barnard, 1963; Field, 2005c; Savage, 1976).

6.3.3. The significance of the correlation coefficient ③

Although we can directly interpret the size of a correlation coefficient, we have seen in Chapter 2 that scientists like to test hypotheses using probabilities. In the case of a correlation coefficient we can test the hypothesis that the correlation is different from zero (i.e., different from 'no relationship'). If we find that our observed coefficient was very unlikely to happen if there was no effect in the population, then we can gain confidence that the relationship that we have observed is statistically meaningful.

There are two ways that we can go about testing this hypothesis. The first is to use our trusty z-scores that keep cropping up in this book. As we have seen, z-scores are useful because we know the probability of a given value of z occurring, if the distribution from which it comes is normal. There is one problem with Pearson's r, which is that it is known to have a sampling distribution that is not normally distributed. This is a bit of a nuisance, but luckily, thanks to our friend Fisher, we can adjust r so that its sampling distribution *is* normal as follows (Fisher, 1921):

$$z_r = \frac{1}{2}\log_e\left(\frac{1+r}{1-r}\right) \tag{6.4}$$

The resulting z_r has a standard error of:

$$SE_{z_r} = \frac{1}{\sqrt{N-3}} \tag{6.5}$$

For our advert example, our $r = .87$ becomes 1.33 with a standard error of .71.

We can then transform this adjusted r into a z-score just as we have done for raw scores, and for skewness and kurtosis values in previous chapters. If we want a z-score that represents the size of the correlation relative to a particular value, then we simply compute a z-score using the value that we want to test against and the standard error. Normally we want to see whether the correlation is different from 0, in which case we can subtract 0 from the observed value of r and divide by the standard error (in other words, we just divide z_r by its standard error):

$$z = \frac{z_r}{SE_{z_r}} \tag{6.6}$$

For our advert data this gives us 1.33/.71 = 1.87. We can look up this value of z (1.87) in the table for the normal distribution in the Appendix and get the one-tailed probability from the column labelled 'Smaller Portion'. In this case the value is .0307. To get the two-tailed probability we simply multiply the one-tailed probability value by 2, which gives us .0614. As such the correlation is significant, $p < .05$, one-tailed, but not two-tailed.

In fact, the hypothesis that the correlation coefficient is different from 0 is usually (**R**, for example, does this) tested not using a z-score, but using a t-statistic with $N - 2$ degrees of freedom, which can be directly obtained from r:

$$t_r = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}} \tag{6.7}$$

You might wonder then why I told you about z-scores, then. Partly it was to keep the discussion framed in concepts with which you are already familiar (we don't encounter the t-test properly for a few chapters), but also it is useful background information for the next section.

6.3.4. Confidence intervals for r ③

Confidence intervals tell us something about the likely value (in this case of the correlation) in the population. To understand how confidence intervals are computed for r, we need to take advantage of what we learnt in the previous section about converting r to z_r (to make the sampling distribution normal), and using the associated standard errors. We can then construct a confidence interval in the usual way. For a 95% confidence interval we have (see section 2.5.2.1):

$$\text{lower boundary of confidence interval} = \bar{X} - (1.96 \times SE)$$

$$\text{upper boundary of confidence interval} = \bar{X} + (1.96 \quad SE)$$

In the case of our transformed correlation coefficients these equations become:

$$\text{lower boundary of confidence interval} = z_r - (1.96 \times SE_{z_r})$$

$$\text{upper boundary of confidence interval} = z_r + (1.96 \quad SE_{z_r})$$

For our advert data this gives us $1.33 - (1.96 \times .71) = -0.062$, and $1.33 + (1.96 \times .71) = 2.72$. Remember that these values are in the z_r metric and so we have to convert back to correlation coefficients using:

$$r = \frac{e^{(2z_r)} - 1}{e^{(2z_r)} + 1} \tag{6.8}$$

This gives us an upper bound of $r = .991$ and a lower bound of -0.062 (because this value is so close to zero the transformation to z has no impact).

CRAMMING SAM'S TIPS Correlation

- A crude measure of the relationship between variables is the *covariance*.
- If we standardize this value we get *Pearson's correlation coefficient, r*.
- The correlation coefficient has to lie between –1 and +1.
- A coefficient of +1 indicates a perfect positive relationship, a coefficient of –1 indicates a perfect negative relationship, and a coefficient of 0 indicates no linear relationship at all.
- The correlation coefficient is a commonly used measure of the size of an effect: values of ±.1 represent a small effect, ±.3 is a medium effect and ±.5 is a large effect. However, if you can, try to interpret the size of correlation within the context of the research you've done rather than blindly following these benchmarks.

6.3.5. A word of warning about interpretation: causality ①

Considerable caution must be taken when interpreting correlation coefficients because they give no indication of the direction of *causality*. So, in our example, although we can conclude that as the number of adverts watched increases, the number of packets of toffees bought increases also, we cannot say that watching adverts *causes* you to buy packets of toffees. This caution is for two reasons:

- **The third-variable problem**: We came across this problem in section 1.6.2. To recap, in any correlation, causality between two variables cannot be assumed because there may be other measured or unmeasured variables affecting the results. This is known as the *third-variable* problem or the *tertium quid* (see section 1.6.2 and Jane Superbrain Box 1.1).
- **Direction of causality**: Correlation coefficients say nothing about which variable causes the other to change. Even if we could ignore the third-variable problem described above, and we could assume that the two correlated variables were the only important ones, the correlation coefficient doesn't indicate in which direction causality operates. So, although it is intuitively appealing to conclude that watching adverts causes us to buy packets of toffees, there is no *statistical* reason why buying packets of toffees cannot cause us to watch more adverts. Although the latter conclusion makes less intuitive sense, the correlation coefficient does not tell us that it isn't true.