Correlation

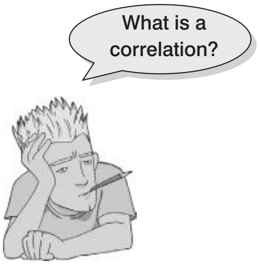
6

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

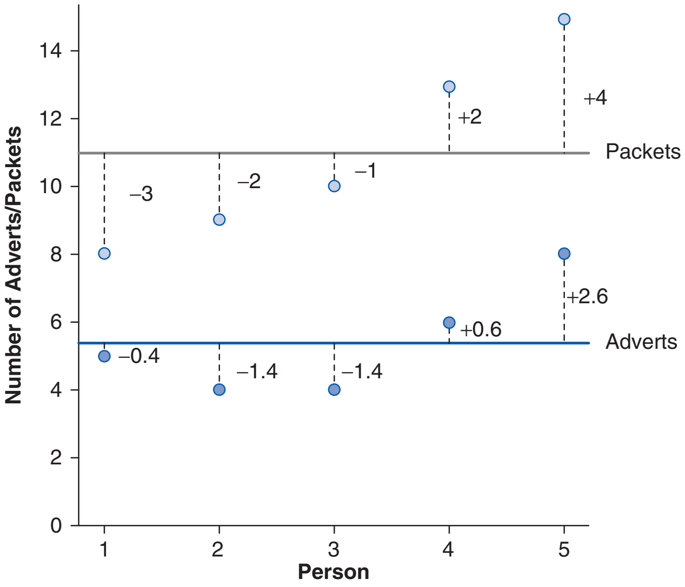


FIGURE 6.2 Graphical display of the differences between the observed data and the means of two variables

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:

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

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

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

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):

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

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 × 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.