Correlational analysis: Pearson's *r*

6

CHAPTER OVERVIEW

In the first five chapters we have given you the basic building blocks that you will need to understand the statistical analyses presented in the remainder of the book. It is important that you understand all the concepts presented in those chapters, and you can get a good measure of your understanding by trying the activities and multiple choice questions presented throughout the text and at the end of each chapter. If you find that there are certain things that you do not understand, it is very much worth your while going back to the relevant chapter and making sure that you have grasped each concept fully. Once you feel confident that you have mastered these concepts, you will be ready to tackle the more demanding statistical analyses presented from now on. Having a thorough understanding of these earlier concepts will smooth the way through the remainder of the book. In the first five chapters, you were introduced to the idea of looking at relationships between variables: for example, the relationship between hours spent studying and performance in examinations. Psychologists often wish to know whether there is a significant relationship or association between two variables. This is the topic of Chapter 6. You will need to have an understanding of the following:

- one- and two-tailed hypotheses (Chapter 5)
- statistical significance (Chapter 5)
- confidence intervals (Chapter 4).

In this chapter we will discuss ways in which we can analyse relationships or associations between variables. In the previous chapter we talked about the relationship between time spent studying and exam performance. The way to find out whether such a relationship exists is to take a number of students and record how many hours per unit of time (e.g. per week) they spend studying, and then later take a measure of their performance in the examinations. We would then have two sets of data (or two variables). Correlational analysis gives us a measure of the relationship between them. In the previous chapter we suggested that we are able to calculate a measure of the strength of a relationship: correlational analysis gives such a measure.

In the present chapter we will discuss the following:

- the analysis and reporting of studies using correlational analysis
- *r* – a natural effect size
- confidence limits around *r*.

6.1 Bivariate correlations

When we are considering the relationship between two variables, this is called *bivariate correlation*. If the two variables are associated, they are said to be co-related (correlated). This means they co-vary; as the scores on one variable change, scores on the other variable change in a predictable way. In other words, the two variables are not independent.

6.1.1 Drawing conclusions from correlational analyses

A correlational relationship cannot automatically be regarded as implying causation. Recall that in Chapter 1 we suggested that you cannot imply causation from correlations. That is, if a significant association exists between the two variables, this does not mean that *x* causes *y* or, alternatively, that *y* causes *x*. For instance, consider the following. It has been shown that there is a significant positive relationship between the salaries of Presbyterian ministers in Massachusetts and the price of rum in Havana. Now it is clearly inappropriate in this case to argue that one variable causes the other. Indeed, as Huff (1973), who supplied this example, observed, it is not necessary to infer causation because the more obvious explanation is that both figures are growing because of the influence of a third factor – the worldwide rise in the price level of practically everything!

Statistical analysis can show us whether two variables are correlated, but the analysis itself cannot tell us the reasons why they are correlated – we have to do this work ourselves! Let's assume that two variables, *x* and *y*, are correlated. This could be because:

- the variation in scores on *y* have been caused by the variation in scores on *x* (i.e. *x* has caused *y*)
- the variation in scores on *x* have been caused by the variation in scores on *y* (i.e. *y* has caused *x*)
- the correlation between *x* and *y* can be explained by the influence of a third variable, *z* (or even by several variables)
- the correlation between them is purely chance.

As an example of the last, on one occasion we asked our students to perform a correlational analysis on several variables. When doing this on the computer, it is very easy mistakenly to include variables that are not relevant. One of our students included 'participant number' with the other variables, in error of course. She then showed us that 'participant number' had a high positive correlation with self-esteem, one of the other variables. Now there was no real relationship between these variables. It is as well, therefore, always to bear in mind the possibility that the relationship revealed by a correlational analysis may be spurious. Francis Galton (1822–1911) was a cousin of Charles Darwin. Although Galton invented correlation, Karl Pearson (1857–1936) developed it, discovering spurious correlations (a statistical relationship only – not due to a real relationship between the two variables, as just explained). He found many instances of spurious correlations. It is up to the researcher to determine whether statistically significant correlations are meaningful and important (rather than just statistically significant) – and to rule out chance factors.

The exploration of relationships between variables may include the following steps:

1. Inspection of *scattergrams* (see below).
2. A statistical test called *Pearson's r*, which shows us the magnitude and degree of the relationship, and the likelihood of such a relationship occurring by sampling error, given the truth of the null hypothesis.
3. *Confidence limits* around the test statistic *r*, where appropriate.
4. *Interpretation of the results*.

6.1.2 Purpose of correlational analysis

The purpose, then, of performing a correlational analysis is to discover whether there is a meaningful relationship between variables, which is unlikely to have occurred by sampling error (assuming the null hypothesis to be true), and unlikely to be spurious. The null hypothesis is that there is no real relationship between the two variables. This is not the only information, however, that a correlational analysis provides. It also enables us to determine the following:

- the direction of the relationship – whether it is positive, negative or zero
- the strength or magnitude of the relationship between the two variables – the test statistic, called the *correlation coefficient*, varies from 0 (no relationship between the variables) to 1 (perfect relationship between the variables).

These two points are discussed in greater detail below.

6.1.3 Direction of the relationship

Positive

High scores on one variable (which we call *x*) tend to be associated with high scores on the other variable (which we call *y*); conversely, low scores on variable *x* tend to be associated with low scores on variable *y*.

Negative

High scores on one variable are associated with low scores on the other variable.

Zero

Zero relationships are where there is no *linear* (straight-line) relationship between the two variables. (What precisely is meant by the term 'linear relationship' will be explained later. For now, just assume that no linear relationship means no relationship between the two variables.)

Now think about the *direction* of the relationships in the examples given above.

Number of hours spent studying and performance in examinations

You would expect that the number of hours spent studying would have a positive relationship with examination performance – the more hours a student spends studying, the better the performance.

Cigarette smoking in ecstasy users

Fisk, Montgomery and Murphy (2009) found that in a sample of ecstasy users, there was a positive correlation between the number of cigarettes smoked and the number of reported adverse reactions.

6.1.4 Perfect positive relationships

We have already said that, in positive relationships, high scores on one variable are associated with high scores on the other, and vice versa. This can be seen by plotting the scores on a

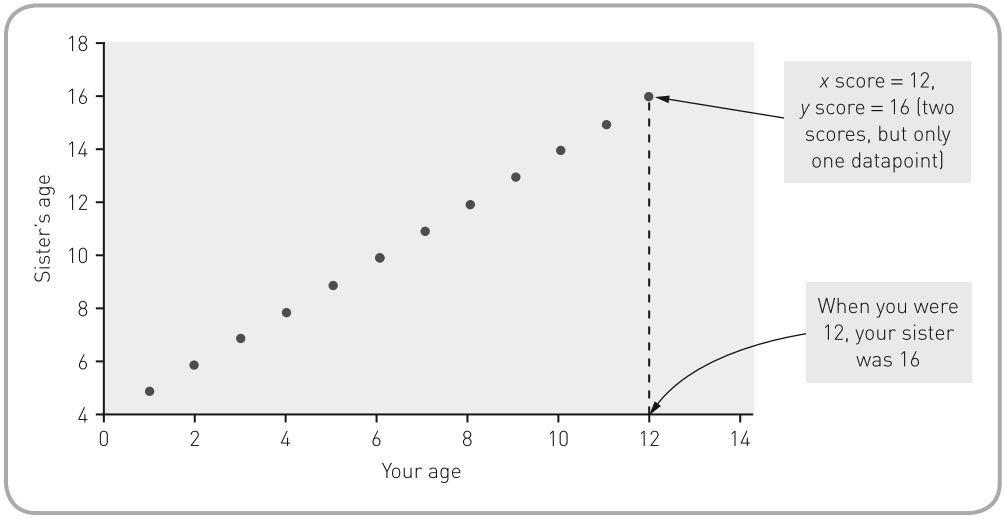


Figure 6.1 Sister's age and your age

graph called a *scattergram*, or scatterplot. When performing a bivariate correlation, we have two sets of scores. When we plot the scores on a scattergram, we assign one variable to the horizontal axis – this is always called *x*. We assign the other variable to the vertical axis – this is always called *y*. It does not matter which variable we assign to *x* or which variable to *y*.

To construct a scattergram, we take each person's score on *x* and *y*, and plot where the two meet. Each datapoint consists of two scores (*x* and *y*). You were introduced to the construction of scattergrams (using SPSS) in section 3.5. Here we go into greater detail.

A perfect positive relationship is depicted in the scattergram in Figure 6.1. A perfect relationship is where all the points on the scattergram would fall on a straight line. For instance, think of your age plotted against your sister's age. (Of course, this is an unrealistic example. No one would really want to correlate their age with their sister's age – it is just an example.) In the example below, we have assumed that your sister is four years older than you. We have allotted your sister's age to the vertical axis (*y*) and your age to the horizontal axis (*x*), and for each pair of ages we have put one point on the scattergram. It should be immediately obvious that the relationship is positive: as you grow older, so does your sister. The relationship must be perfect as well: for every year that you age, your sister ages one year as well.

An important point to note is that the above example shows that you cannot draw any inferences about *cause* when performing a correlation. After all, your age increase does not cause your sister's age to increase; neither does her growing older cause you to age!

6.1.5 Imperfect positive relationships

Imagine that we have a number of students whom we have measured on IQ and percentage marks in an exam. We want to see whether there is a relationship between IQ and exam marks. This does not mean that we are saying IQ *causes* students' exam marks; nor does it mean that the exam marks they achieved somehow had an effect on their IQ. Both high (or low) IQ and high (or low) exam marks could have been 'caused' by all sorts of factors – crammer courses, IQ practice tests, motivation, to mention just a few.

We decide to allot IQ to the vertical axis (*y*) and exam marks to the horizontal axis (*x*). Each student has two scores, an IQ score and an exam mark. However, each student contributes only one 'point' on the scattergram, as you can see in Figure 6.2.

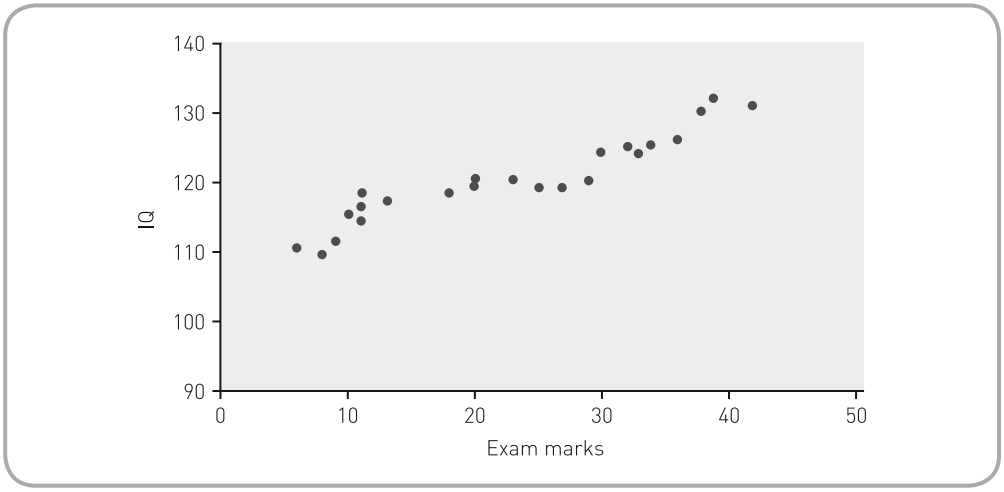


Figure 6.2 Scattergram of IQ and exam marks

You can see from this scattergram that high IQs *tend* to be associated with high exam scores, and low IQ scores *tend* to be associated with low exam scores. Of course, in this instance the correlation is not perfect. But the trend is there, and that is what is important. That is, although the dots do not fall on a straight line, this is still a positive linear relationship because they form a discernible pattern going from the bottom left-hand corner to the top right-hand corner.

Activity 6.1

Try to think of some bivariate positive relationships. Are your examples likely to be perfect relationships? Discuss your examples with others. Do you agree with each other on whether your examples are good ones?

6.1.6 Perfect negative relationships

Again, because this relationship is perfect, the points on the scattergram would fall on a straight line. Each time *x* increases by a certain amount, *y decreases* by a certain, constant, amount.

Imagine a vending machine, selling chocolate. The cost is 50p per bar. At the beginning of the day, the machine is filled with ten chocolate bars. Assuming, of course, that it works as it should (that is, no chocolate bars get stuck, the money stays in, it gives you the right change etc. . . . well, perhaps this is a bit too unrealistic but never mind), each time someone puts in 50p, the chocolate bar is ejected, and one fewer remains. This can be seen in Figure 6.3.

As you can see, with a perfect negative linear relationship the dots still fall on a straight line, but this time they go from the top left-hand corner down to the bottom right-hand corner.

6.1.7 Imperfect negative relationships

With an imperfect negative linear relationship the dots do not fall on a straight line, but they still form a discernible pattern going from the top left-hand corner down to the bottom

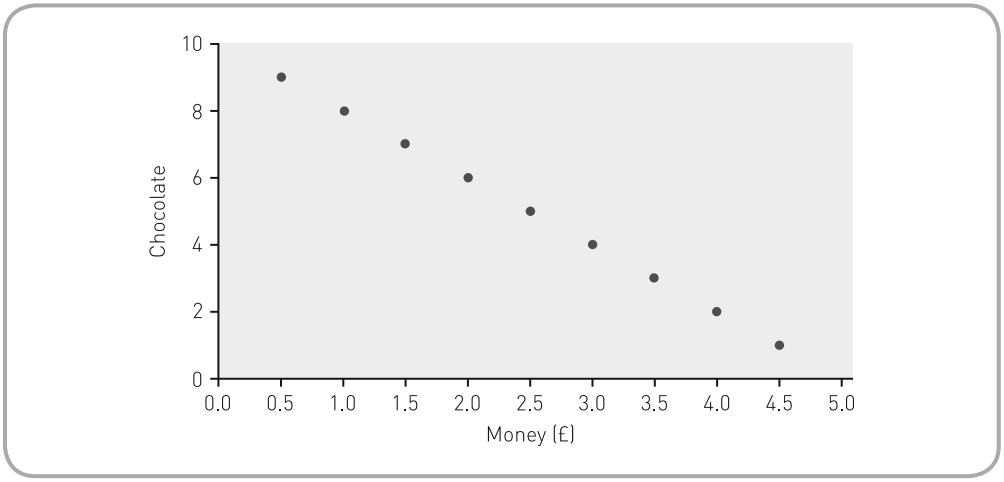


Figure 6.3 Graph of the relationship between chocolate bars in the machine and amount of money put in

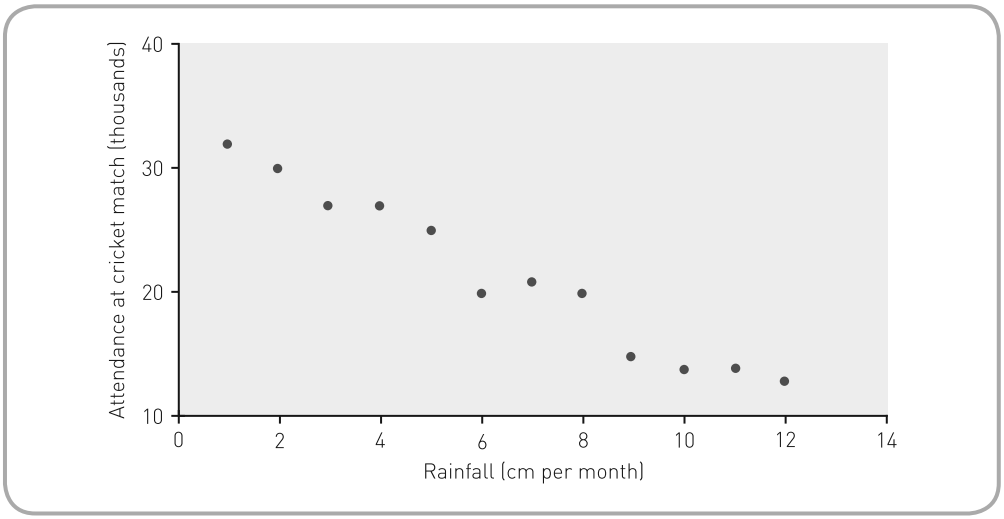


Figure 6.4 Scattergram of attendance at cricket matches and rainfall

right-hand corner. For example, suppose we had collected data on attendances at cricket matches and the amount of rainfall. The resulting scattergram might look something like Figure 6.4. Generally, the trend is for attendance at cricket matches to be lower when rainfall is higher.

6.1.8 Non-linear relationships

Note that, if a relationship is *not* statistically significant, it may not be appropriate to infer that there is *no* relationship between the two variables. This is because, as we have said before, a correlational analysis tests to see whether there is a *linear* relationship. Some relationships are not linear. An example of such a relationship is that between arousal and performance. Although we would expect a certain level of arousal to improve sports performance, too

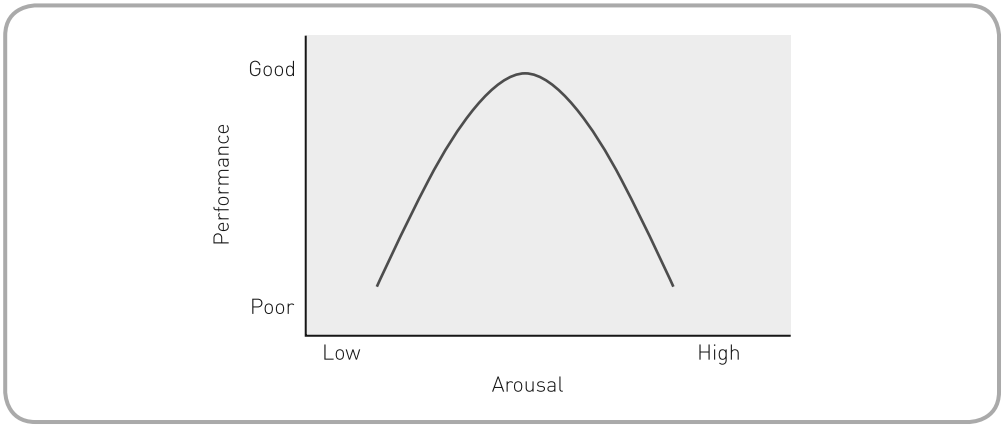


Figure 6.5 The inverted-U hypothesis (Yerkes–Dodson law, 1908)

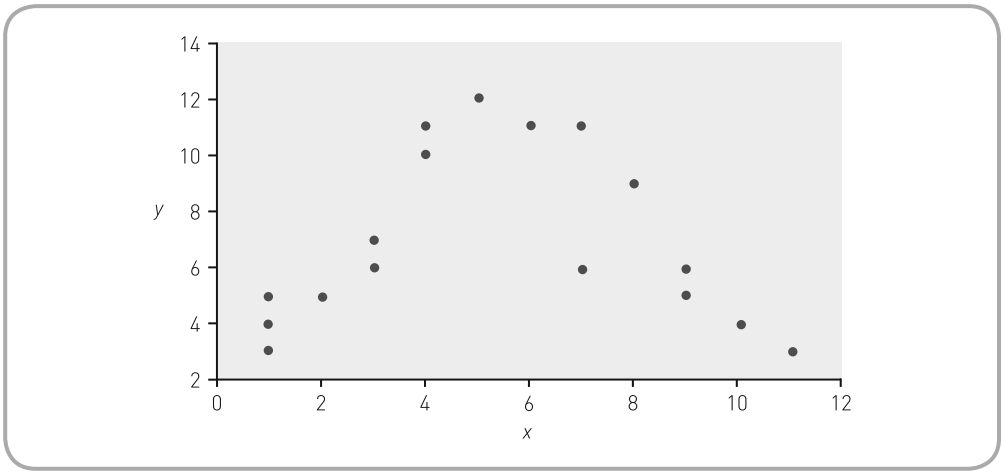
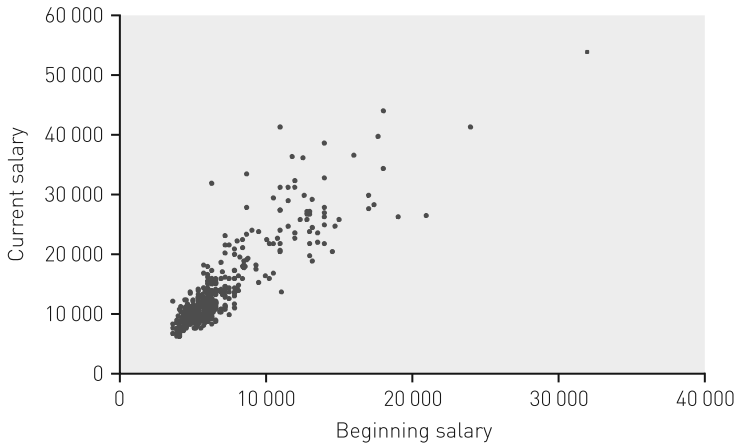


Figure 6.6 Scattergram illustrating a curvilinear relationship between *x* and *y*

much arousal could lead to a detriment in performance. Such a relation is described by the Yerkes–Dodson law (Yerkes and Dodson, 1908). This law predicts an inverted curvilinear relationship between arousal and performance. At low levels of arousal, performance (e.g. athletic performance) will be lower than if arousal was a bit higher. There is an 'optimum' level of arousal, at which performance will be highest. Beyond that, arousal actually decreases performance. This can be represented as shown in Figure 6.5.

The same relationship can be represented by the scattergram in Figure 6.6, which shows a curvilinear relationship: that is, *x* increases with *y* up to a certain extent, and then decreases with *y*. The point we are trying to make is that here there is undoubtedly a relationship between *x* and *y*, but the correlation coefficient would not be statistically significant because there is not a linear (straight-line) relationship. For this reason, you should really always look at a scattergram before you carry out your analysis, to make sure that your variables are not related in this way, because, if they are, there is not much point in using the techniques we are describing in this chapter.

Activity 6.2



Which is the most sensible conclusion? The correlation between beginning salary and current salary is:

(a) Negative
(b) Positive
(c) Zero

6.1.9 The strength or magnitude of the relationship

The strength of a linear relationship between the two variables is measured by a statistic called the *correlation coefficient*, also known as *r*, which varies from 0 to –1, and from 0 to +1. There are, in fact, several types of correlational test. The most widely used are Pearson's *r* (named after Karl Pearson, who devised the test) and Spearman's rho (Eta^2 and Cramer's V are two we mention in passing). The full name of Pearson's *r* is Pearson's product moment correlation; this is a parametric test and the one we will be discussing in this chapter. You will remember from Chapter 5, page 154, that, in order to use a parametric test, we must meet certain assumptions. The most important assumption is that data are drawn from a normally distributed population. If you have large numbers of participants, this assumption is likely to be met. If you have reason to believe that this is *not* the case, you should use the non-parametric equivalent of Pearson's *r*, which is called Spearman's rho (see Chapter 16, page 528).

In Figure 6.1 above, the relationship is represented by +1: plus because the relationship is positive, and 1 because the relationship is perfect. In Figure 6.3 above, the relationship is –1: minus because the relationship is negative, and 1 because the relationship is perfect.

Remember: +1 = perfect positive relationship
–1 = perfect negative relationship

Figure 6.7 shows you the various strengths of the correlation coefficents.

Figure 6.7 puts over the idea that –1 is *just as* strong as +1. Just because a relationship is negative does not mean that it is less important, or less strong, than a positive one. As we have said before (but repetition helps), a positive relationship simply means that high scores on *x* tend to go with high scores on *y*, and low scores on *x* tend to go with low scores on *y*, whereas a negative relationship means that high scores on *x* tend to go with low scores on *y*.

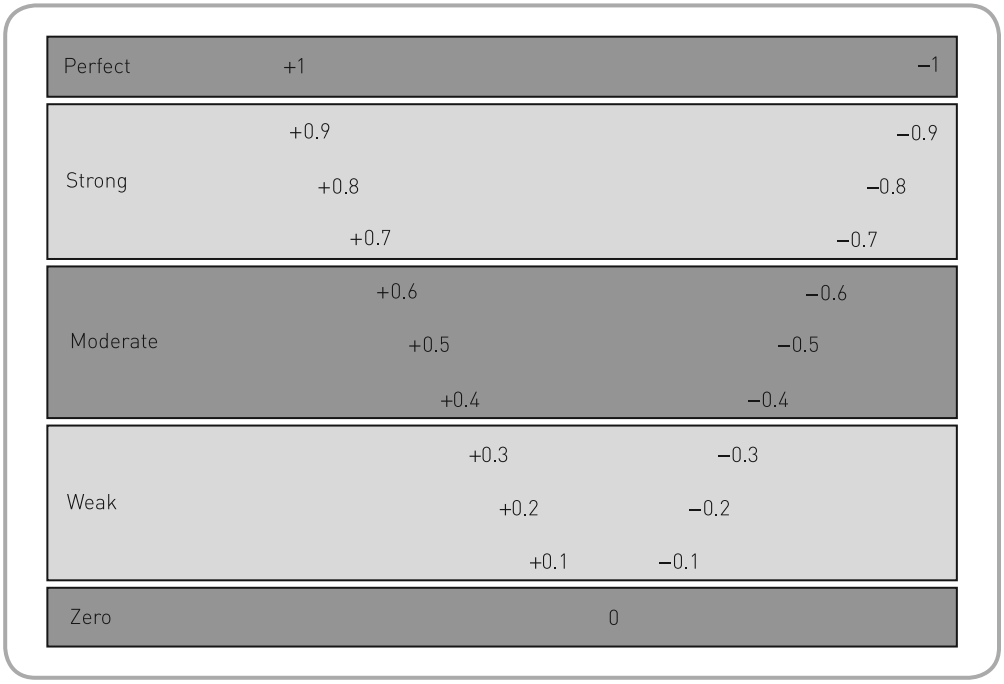


Figure 6.7 Illustration of the strength of positive and negative correlation coefficients

You can see that we have assigned verbal labels to the numbers – these are only guides. A correlation of 0.9 is a strong one. Obviously the nearer to 1 (+ or –) a correlation coefficient is, the stronger the relationship. The nearer to 0 (meaning no relationship), the weaker the correlation. Correlations of 0.4 to 0.5 are moderate. The correlation coefficient measures how closely the dots cluster together.

Activity 6.3

A correlation coefficent of +0.2 is considered:

(a) Zero
(b) Weak
(c) Moderate
(d) Strong

The scattergrams in Figures 6.8 and 6.9 give you some idea of what the correlation coefficients mean.

Walsh *et al.* (2009) used correlational analyses in order to see whether attachment anxiety and attachment avoidance were related to mindfulness. Mindfulness is a state of mind whereby a person attends to the present (rather than the past or the future). A person who is mindful tries to 'live in the present' and focus on the immediate experience. This should reduce worry and rumination.

The scattergram in Figure 6.8 shows that as attachment anxiety increases, attachment avoidance increases. The correlation is weak–moderate.

The correlation between trait anxiety and mindfulness shows that there is a moderate negative association between trait anxiety and mindfulness (see Figure 6.9).

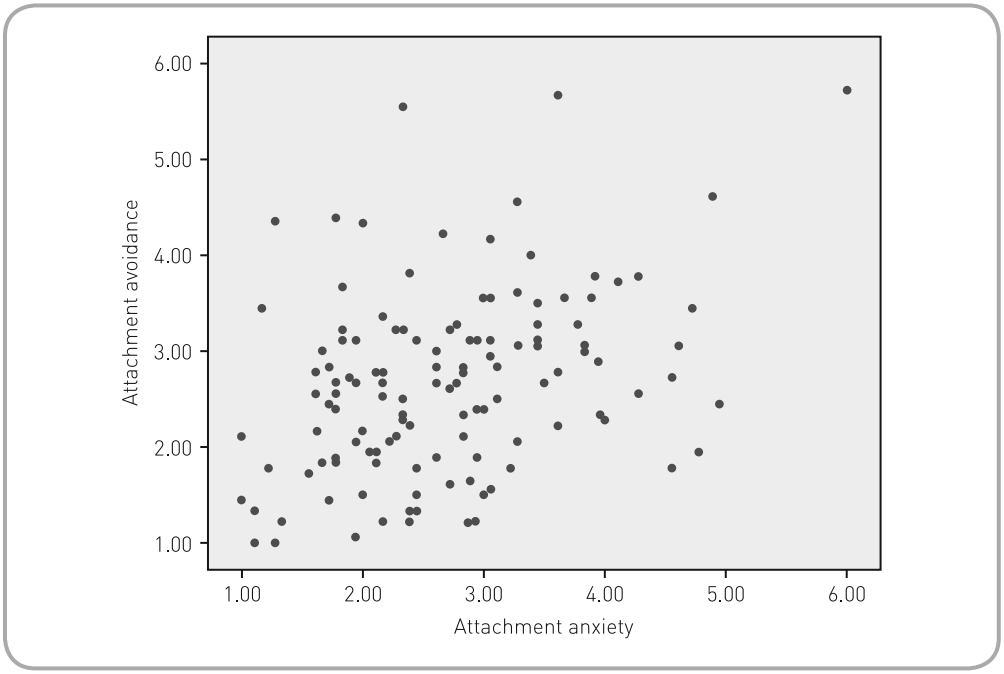


Figure 6.8 Scattergram showing the correlation between attachment anxiety and attachment avoidance ($N = 127$, $r = +0.363$, $p < 0.001$)

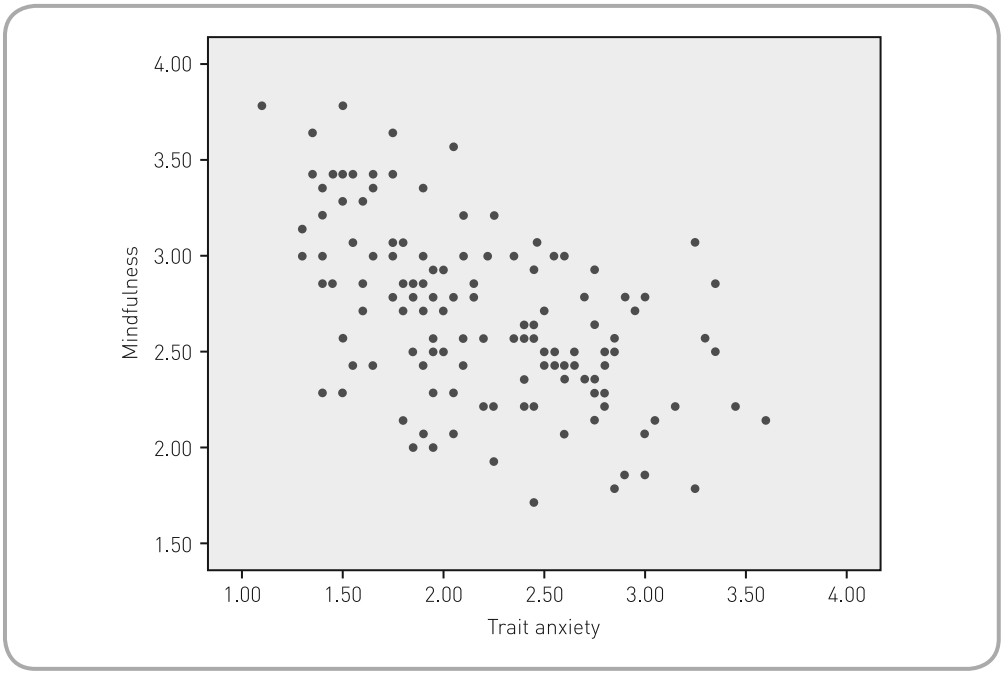


Figure 6.9 Scattergram showing the correlation between trait anxiety and mindfulness ($N = 132$, $r = -0.542$, $p < 0.001$)

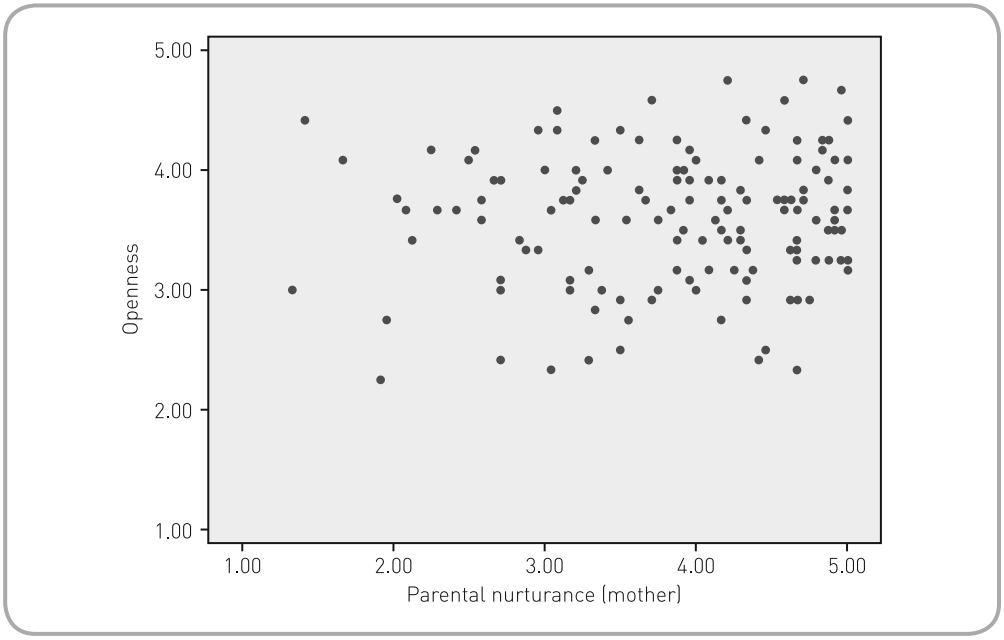


Figure 6.10 Scattergram showing the correlation between parental nurtance (mother) and openness (n = 136, r = +0.080)

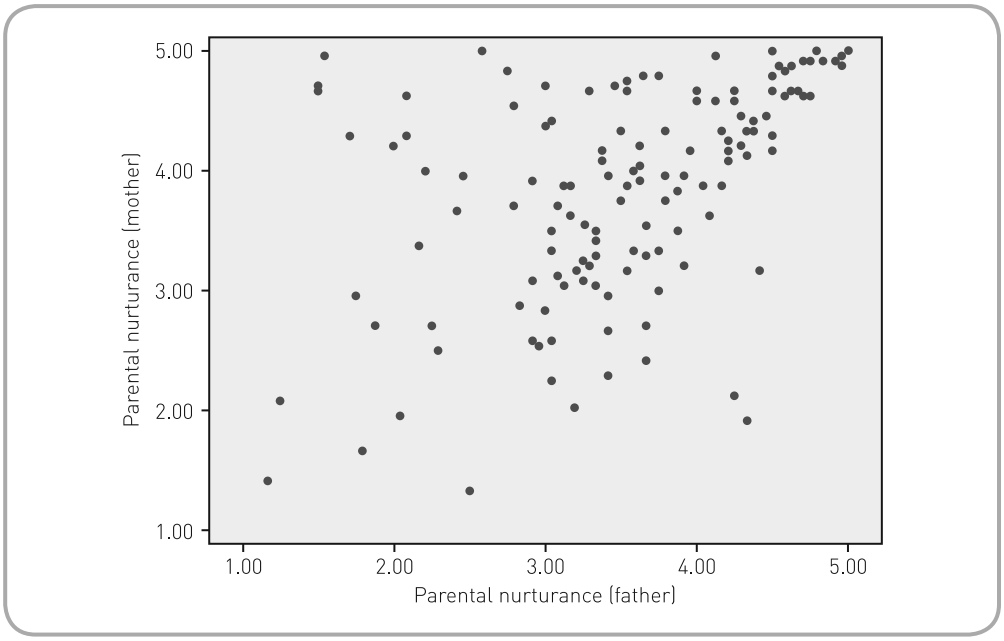


Figure 6.11 Scattergram showing the correlation between mother's nurturance and father's nurturance (n = 136; r = +0.473)

If a correlation is zero, the dots may appear to be random, and there is no discernible pattern. Thus there is no relationship between x and y.

Figure 6.10 shows that there is no association between parental nurtance (mother) and openness (n = 136; r = 0.080; p = 0.355).

Data correlating (n = 136; r = +0.473; $p < 0.001$) scores for parental nurturance (mother) and parental nurturance (father) are represented in Figure 6.11. There is a moderately strong positive relationship between these two variables.

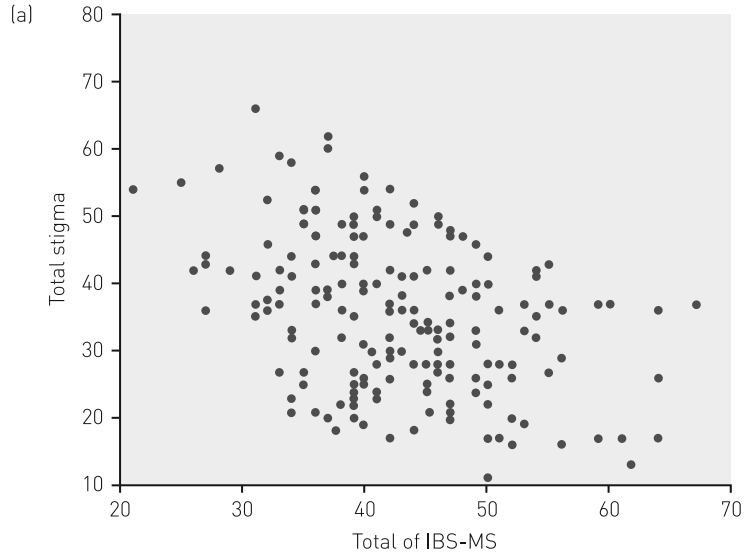
There is a moderate negative association between trait anxiety and attentional control and ($n = 136$; $r = -0.445$; $p < 0.001$) (see Figure 6.12).

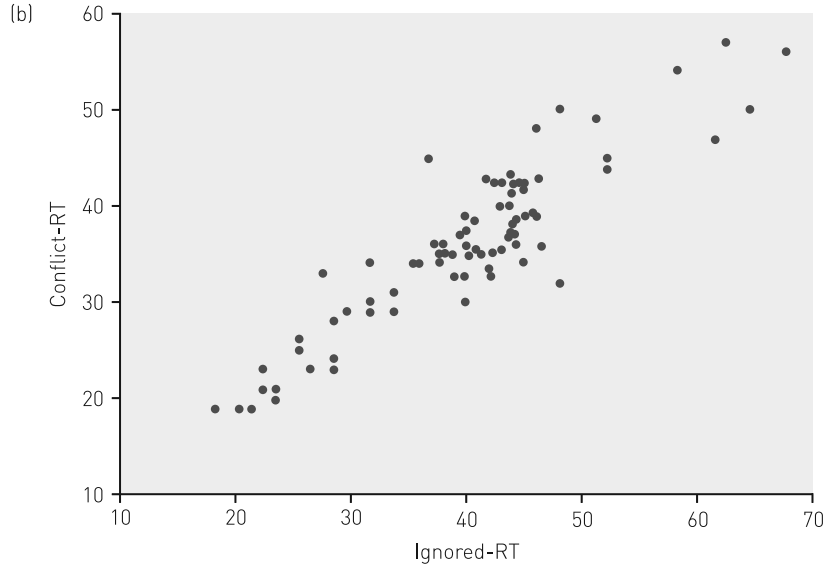
Activity 6.4

Have a look at the following scattergrams. Consider whether, just by looking at them, you can tell:

(a) The direction of the relationship (positive, negative or zero)
(b) The magnitude of the relationship (perfect, strong, weak or zero)

It is sometimes difficult to tell – which is when a test statistic like Pearson's *r* comes in handy!





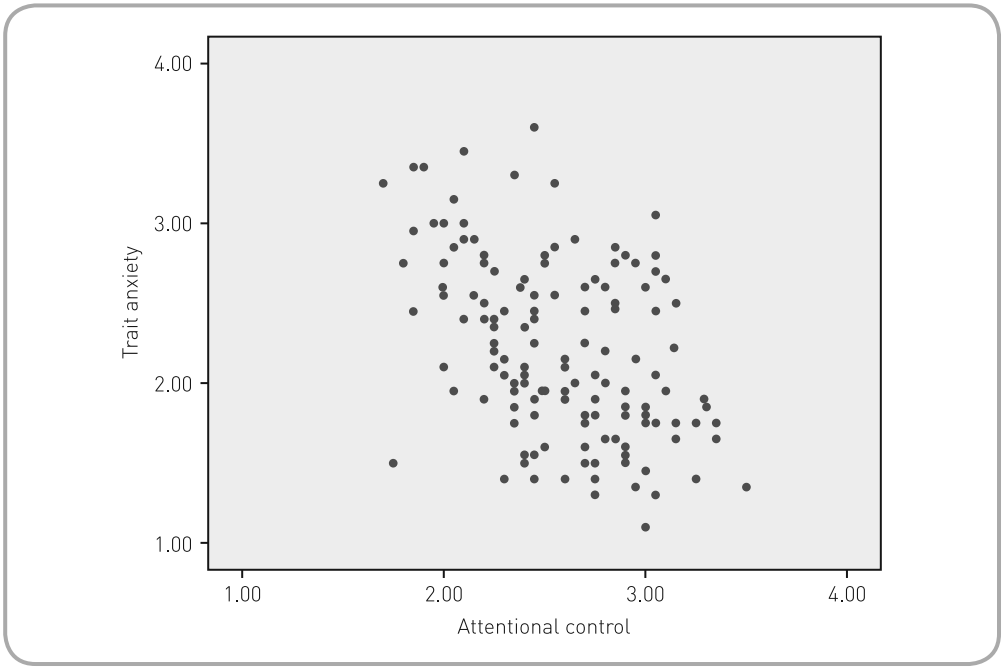


Figure 6.12 Scattergram showing the correlation between attentional control and trait anxiety ($n = 136$; $r = -0.445$)

Think about this. Does going to church stop you getting pregnant?

> There are about 118,000 teenage pregnancies a year, and half of all single parents are under 25. The UK has the highest divorce rate in Europe and the most teenage pregnancies, though other countries are coming up fast. The only reason detectable by statistics is connected to church-going. Britain's attendance figures began to drop before other countries, and everywhere as church attendance falls, divorce and single parenthood rise.
>
> (Polly Toynbee, *Radio Times*, 20–26 March 1993)

Let's look at a perfect relationship again (Figure 6.13). Imagine that this represents the relationship between the scores on two tests, Test 1 and Test 2. The fact that this is a perfect correlation means that the relative position of the participants is exactly the same for each test. In other words, if Sharmini has the top score on Test 1 (in the above example it is 23) she will also have scored the top score on Test 2 (130). Conversely, the participant who has the lowest score on Test 1 will also have the lowest score on Test 2.

Now, as we said previously, perfect relationships are rare, but the same reasoning applies with imperfect relationships. That is, in order to calculate a correlation coefficient it is necessary to relate the relative position of each participant on one variable to their relative position on the second variable.

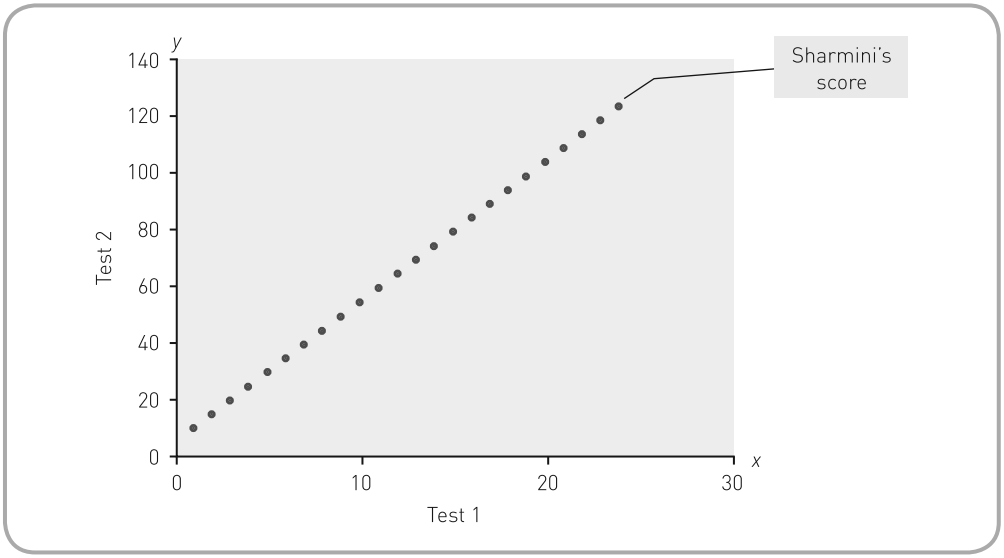


Figure 6.13 Perfect linear relationship

Example: temperature and ice-cream sales

Let's imagine that we have carried out a correlational analysis on a number of ice-cream cones bought from a van outside your college, and temperature. We ask the vendor, called Sellalot, how many ice-cream cones have been sold each day. We take the measurements over 20 days. Now we need to know whether the number of ice-cream cones sold varies along with the temperature. We would expect that, according to previous literature, ice-cream sales would increase as temperature rises. This is a one-tailed hypothesis. The data are given in Table 6.1.

Now is it quite easy to see how to plot scattergrams by hand, although when you have many scores this could be tedious. Naturally, SPSS performs this task better than we can! Instructions for how to obtain scattergrams were given to you in Chapter 3, page 69.

From the scattergram in Figure 6.14 we can see that temperature and number of ice-cream cones sold are related. It is obviously not a perfect correlation, but just by looking at the data we can see that it is positive.

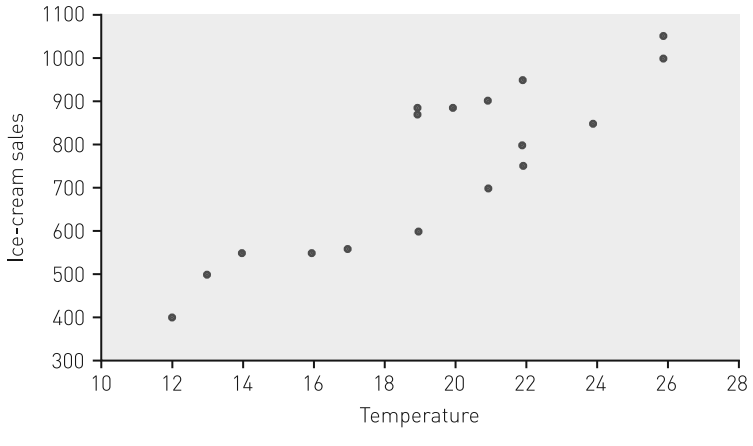


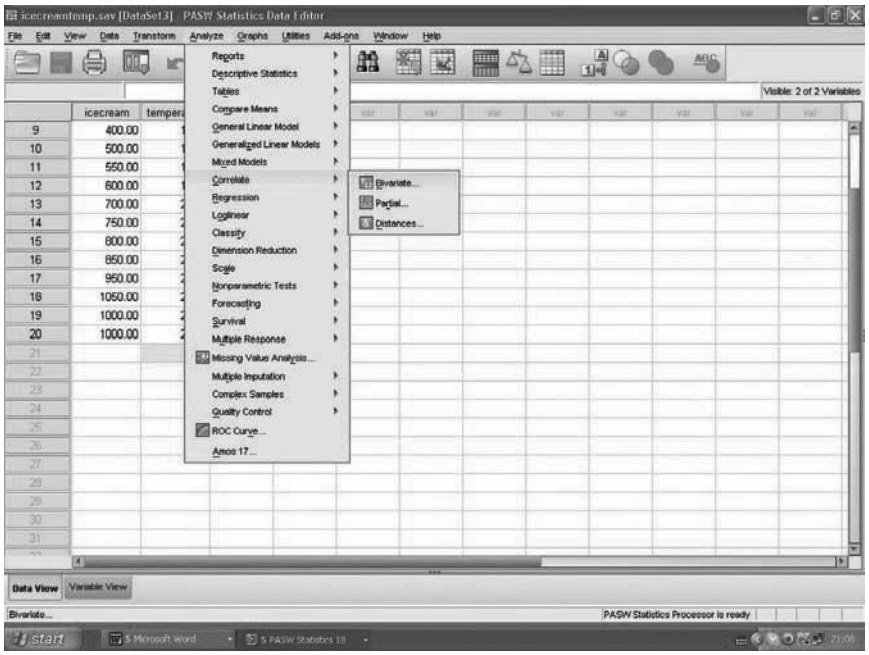
Figure 6.14 Scatterplot of the ice-cream cone data

Table 6.1 Data for the number of ice-cream comes sold on days with different temperatures

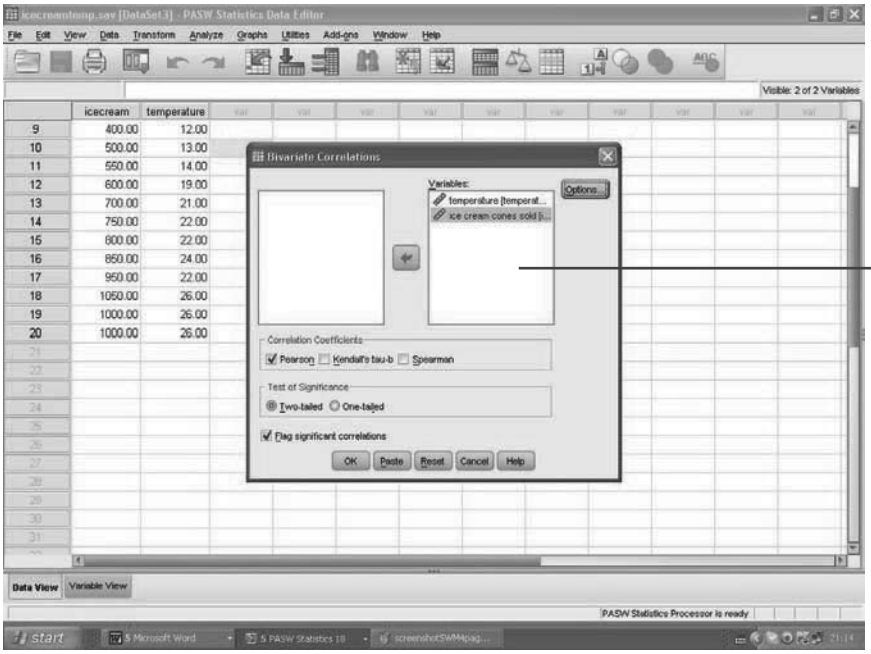
Ice-cream cones sold	Temperature	Ice-cream cones sold	Temperature
1000	26	550	14
950	22	600	19
870	19	700	21
890	20	750	22
886	19	800	22
900	21	850	24
560	17	950	22
550	16	1050	26
400	12	1000	26
500	13	1000	26

SPSS: bivariate correlations – Pearson's *r*

Now we want to know the value of the correlation coefficient and the associated probability, so again we turn to SPSS. Our data have already been entered into SPSS, so now we select *Analyze*, *Correlate*, *Bivariate*:



This brings you to the following dialogue box:



Move both variables from the left-hand side to the right-hand side. Make sure the *Pearson* and *One-tailed* options are selected. Then click on *OK*. This will obtain your results.

Let's look at the output from SPSS. The important results for your write-ups are:

- the correlation coefficient *r*; this shows us how strongly our variables relate to each other, and in what direction
- the associated probability level, letting us know the likelihood of our correlation coefficient arising by sampling error, assuming the null hypothesis to be true.

Results are given in the form of a matrix. A matrix is simply a set of numbers arranged in rows and columns. The correlation matrix is an example of a square symmetric matrix. You should find that each variable correlates perfectly with itself (otherwise something is amiss!). You will also note that results are given twice: each half of the matrix is a mirror image of itself. This means you have to look at one half (of the diagonal) only. SPSS also lets us know the number of pairs for each variable. You can see from the output below that the point where our variable ICECREAM meets our variable TEMP gives us the information we need. The first line gives us the correlation coefficient – it is usual for us to give this correct to two decimal places. The achieved significance level is given on the second line, and the third line confirms how many pairs we have in the analysis. Remember that, when SPSS gives a row of zeros, change the last zero to a 1 and use the < sign (i.e. $p < 0.001$, $n = 20$). Note that our correlation coefficient is *positive* – as temperature rises, so does the sale of ice-creams.

The correlation coefficient (r) is given in the cell where 'ice cream' meets 'temperature', i.e. $r = +0.89$

Correlations

		Ice cream	temp
Ice cream	Pearson Correlation	**1.000**	**.8931**
	Sig. (2-tailed)	.	**.000**
	N	20	**20**
temp	Pearson Correlation	.8931	1.000
	Sig. (2-tailed)	.000	.
	N	20	20

This is the achieved significance level. Remember that, when SPSS gives a row of zeros, change the last one to a '1' and use the < sign, i.e. $p < 0.001$

These results tell us that the sales of ice-cream cones are positively and strongly related to the temperature. The textual part of our analysis might therefore read as follows:

> The relationship between sales of ice-cream and outside temperature was found to be positively and strongly related ($r = +0.89$, $p < 0.001$). Thus as temperature rises, so does the sale of ice-cream.

This is all we can say at the moment, but as the chapter progresses you will see that we can add to this.

Activity 6.5

Look at the following output from SPSS:

Correlations

		Attachment anxiety	Attachment avoidance	Mindfulness	Trait Anxiety	age
Attachment anxiety	Pearson Correlation	1	.353	–.317	.310	.211
	Sig. (2-tailed)		.000	.000	.000	.019
	N	127	127	127	127	123
Attachment avoidance	Pearson Correlation	.353	1	–.247	.223	.129
	Sig. (2-tailed)	.000		.005	.012	.155
	N	127	127	127	127	123
Mindfulness	Pearson Correlation	–.317	–.247	1	–.328	–.083
	Sig. (2-tailed)	.000	.005		.000	.359
	N	127	127	127	127	123
Trait Anxiety	Pearson Correlation	.310	.223	–.328	1	.004
	Sig. (2-tailed)	.000	.012	.000		.962
	N	127	127	127	127	123
age	Pearson Correlation	.211	.129	–.083	.004	1
	Sig. (2-tailed)	.019	.155	.359	.962	
	N	123	123	123	123	123

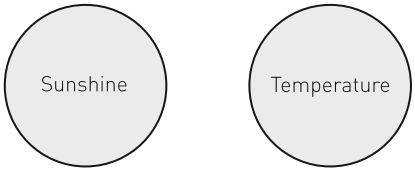
Which association is the strongest?

(a) Attachment avoidance and attachment anxiety
(b) Mindfulness and trait anxiety
(c) Mindfulness and attachment anxiety

6.1.10 Variance explanation of the correlation coefficient

The correlation coefficient (*r*) is a ratio between the covariance (variance shared by the two variables) and a measure of the separate variances.

By now you should have a good idea of what a correlation coefficient means. For instance, if we tell you that two variables are associated at 0.9, you could probably draw the scattergram pretty well. Similarly, if we tell you to draw a scattergram representing a 0.1 association, you could probably do that fairly accurately as well. But there is another way of visualising what these coefficients mean, a way that will be very useful to you later on, when we go on to regression analysis. Let's take an example of number of hours of sunshine, and temperature (this example originated from Alt, 1990). These two variables are positively associated: the more hours of sunshine, the higher the temperature. When two variables are correlated, we say that they 'share' variance. For instance, the following circles represent sunshine hours and temperature.

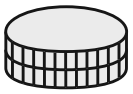


We have drawn these circles, representing sunshine and temperature, as if they are independent, but they are *not* independent. They share a lot of variance. How much variance do they share? The test statistic, a correlation coefficient, will give us the answer. We have already said that the correlation coefficient goes from 0 to +1, and 0 to –1. By *squaring* the correlation coefficient, you know how much variance, in percentage terms, the two variables share. Look at Table 6.2.

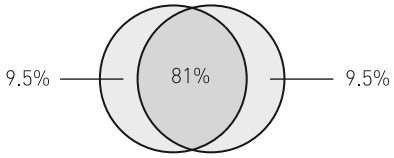
Remember, negative correlations, when squared, give a positive answer. So –0.4 squared (-0.4×-0.4) = 0.16. So 16% of the variance has been accounted for by a correlation of –0.4, just the same as if the correlation is +0.4. If you have a correlation of 0.9, you have accounted for (explained) 81% of the variance. A Venn diagram will make this clearer. If two variables were *perfectly* correlated, they would not be independent at all. The two circles for *x* and *y* would lie on top of each other, just as if you had two coins on top of each other:

Table 6.2 Table demonstrating the relationship between correlations and squared correlations

Correlation (*r*)	Correlation squared (r^2)	Variance accounted for
0.0	0.0	0.00
0.1	0.1^2	0.01 (1%)
0.2	0.2^2	0.04 (4%)
0.3	0.3^2	0.09 (9%)
0.4	0.4^2	0.16 (16%)
0.5	0.5^2	0.25 (25%)
0.6	0.6^2	0.36 (36%)
0.7	0.7^2	0.49 (49%)
0.8	0.8^2	0.64 (64%)
0.9	0.9^2	0.81 (81%)
1.0	1.0^2	1.00 (100%)



The two variables would correlate +1.00, and all the variability in the scores of one variable could be accounted for by the variability in the scores of the other variable. Take sunshine hours and temperature, which we can assume to be correlated 0.9 (81%). The two circles look like this:



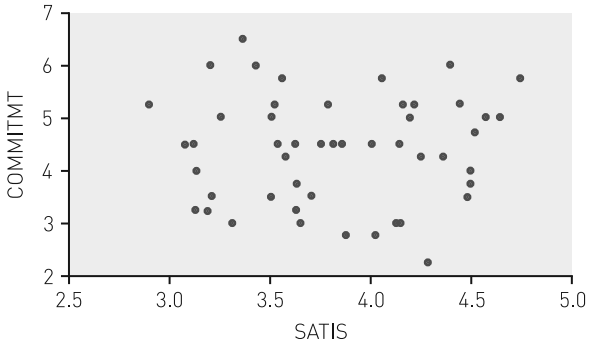
Remember, if 81% is shared variance, then 19% is not shared: it is what is known as unique variance – 9.5% is unique to sunshine, and 9.5% is unique to temperature. If the shared variance is significantly greater than the unique variances, *r* will be high. If the unique variances are significantly greater than the shared variance, *r* will be low.

$$r = \frac{\text{a measure of shared variance}}{\text{a measure of the separate variances}}$$

The shaded part (81%) is the variance they share. In other words, 81% of the variation in number of hours of sunshine can be explained by the variation in temperature. Conversely, 81% of the variation in temperature can be accounted for by reference to the variation in number of hours of sunshine – 19% is 'unexplained': that is, the variation in scores must be due to other factors as well.

Activity 6.6

Look at the scattergram below:



Which is the most sensible conclusion? The two variables show a:

(a) Moderate positive correlation
(b) Moderate negative correlation
(c) Strong negative correlation
(d) Zero correlation

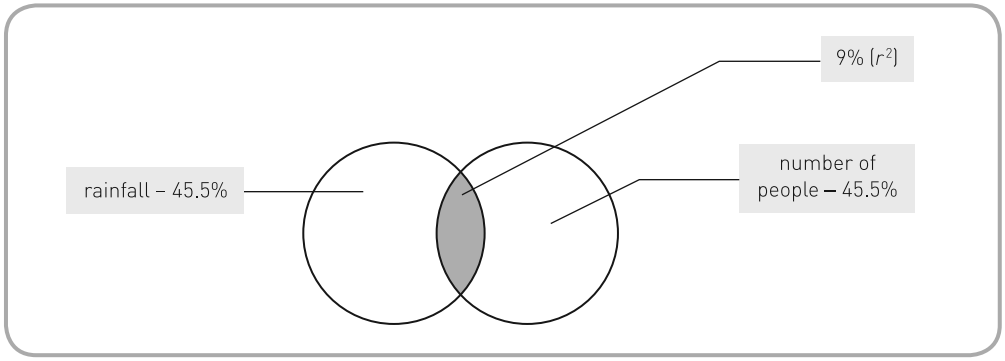


Figure 6.15 Diagram illustrating the amount of shared variance between two variables

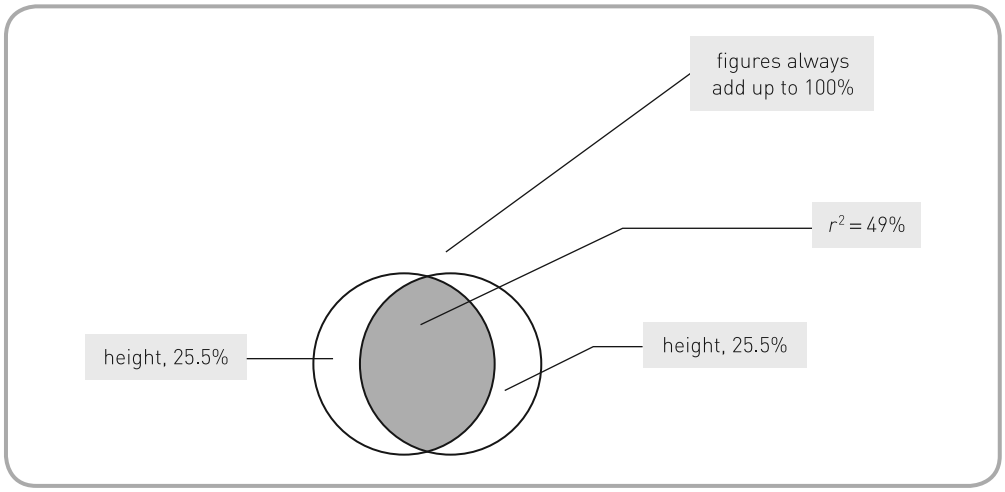


Figure 6.16 A further illustration of shared variance

Take the case of number of inches of rainfall and attendance at cricket matches. Here we would expect a negative relationship: the more rain, the fewer people attending. Assume that the relationship is −0.3. This means that 9% (−0.3 × −0.3 = +0.09) of the variance has been explained (see Figure 6.15).

As another example (see Figure 6.16), assume that we measure and weigh a class of schoolchildren, and that height and weight correlate by 0.7. How much of the variance has been accounted for? We multiply 0.7 by 0.7 = 0.49 (49%): this means that nearly half of the variation in the scores of height can be explained by the variation in weight. Conversely, nearly half of the variation in weight can be accounted for by reference to the variation in height.

This means, of course, that 51% is *unexplained*: that is, 51% is explainable by reference to other factors, perhaps age, genetics and environmental factors. A correlation coefficient can always be squared to give you the 'variance explained' (*r* squared). Similarly, if you know r^2, you can use the square-root button on your calculator to give you the correlation coefficient, *r* (although this will not tell you the direction of the relationship). You should be able to see by this that a correlation of 0.4 is *not* twice as strong as a correlation of 0.2. A correlation of 0.4 means that 16% of the variance has been explained, whereas 0.2 means that only 4% has been explained. So a correlation of 0.4 is, in fact, four times as strong as 0.2. A correlation coefficent is a good measure of effect size and can always be squared in order to see how much of the variation in scores on one variable can be explained by reference to the other variable.

Activity 6.7

When you are assessing the strength and significance of a correlation coefficient, it is important to look at:

(a) The significance level
(b) The value of the correlation coefficient
(c) Both (a) and (b)
(d) Neither (a) nor (b)

There is a (perhaps fictitious!) correlation between amount of ice-cream eaten and feelings of great happiness (+0.85). How much variation in the happiness scores can be explained by the amount of ice-cream eaten? How much variance is left unexplained?

6.1.11 Statistical significance and psychological importance

In the past, some people were more concerned about 'significance' than about the size of the correlation or the amount of variance explained. Sometimes people used to say that they had a highly significant correlation: they remembered the probability value (for instance, 0.005) but forgot the size of the correlation. The probability value means very little without reporting the *r* value. The correlation coefficient tells you how well the variables are related, and the probability value is the probability of that value occurring by sampling error.

So when you report your findings, report the correlation coefficient and think about whether *r* is meaningful in your particular study, as well as the probability value. Do not use the probability value on its own. *Remember, statistical significance does not necessarily equal psychological significance* (see Chapters 5 and 8 for further information).

Example: ice-creams and temperature revisited

Now you know about variance explained, we can adjust the textual part of our results to include it. The textual part of our analysis might now read as follows:

> The sale of ice-cream cones was strongly associated with temperature; as temperature rises, so does the sale of ice-creams. The *r* of 0.89 showed that 79% of the variation in ice-cream sales was accounted for by the variation in the temperature ($p < 0.01$). The associated probability level of 0.001 showed that such a result is highly unlikely to have arisen by sampling error alone.[1]

[1] This really means 'by sampling error, assuming the null hypothesis to be true'. Although you probably won't want to say this in your lab reports, you should always bear in mind that this is what the probability value means.