Chapter 19

Analysing frequency data: chi-square

chapter 19

Nominal data, categories and frequency counts

There are many occasions when we want to examine the effects of an independent variable on the dependent variable when the data are nominal: the numbers indicate the category the subject belongs to rather than a position on an ordinal or interval scale. An experimenter interested in hair length of female students might categorise hair length into two categories: long (on or below the shoulder) and short (above the shoulder). Female students could then be sampled to see whether there is a preference for long or short hair on campus. Note that the data collected from the students is neither a score nor a rating. The researcher is collecting frequency data, that is adding up the number of participants in each category. If 100 female students were randomly sampled and 62 had long hair and 38 short hair can we conclude that there is a significant preference for long hair? The statistic examined in this chapter, chi-square (χ^2), allows us to analyse frequency data to answer such questions. We are not limited in the number of (independent) categories we choose, which makes this a very useful statistic, particularly when we are undertaking questionnaires or surveys. If we wanted to compare liberals and conservatives on, say, a proposed piece of new taxation legislation we could ask a number of liberals and conservatives whether they are for or against the legislation. Here we have four categories: liberals-for, liberals-against, conservatives-for and conservatives-against, with their respective frequency counts. If we included the category 'don't know' for each political group we would increase our categories to six.

Introduction to χ^2

The simplest way to view the χ^2 statistic is as the square of the z statistic:

$$\chi^2 = z^2 = \frac{(X - \mu)^2}{\sigma^2}$$

χ^2 is the square of the deviation of a score from its population mean divided by the population variance, where the population is normally distributed.

Just as we saw that the F statistic in its simplest case is t^2 and therefore never negative we also find that χ^2, also a squared value, is always positive. Like F we are only interested in the high values of the χ^2 distribution but it is always a two-tailed test in that a large positive z score or a large negative z score both square to a large positive χ^2.

In most cases we are testing samples rather than individual scores and this is where χ^2 turns out to be so useful in data analysis. If we select mutually independent samples from which to obtain X then it turns out that the sum of the individual χ^2s is also a χ^2:

$$\chi^2 = \sum z^2 = \sum\left(\frac{(X-\mu)^2}{\sigma^2}\right)$$

This means that we can find a χ^2 for each sample and the sum of the χ^2s will also be a χ^2. This allows us to compare samples against the sampling distribution of χ^2. However, the shape of the χ^2 distribution depends on the number of χ^2s that are summed, so we must take into account the degrees of freedom of the samples (the number of samples minus one). If we have four categories then the degrees of freedom for χ^2 is $c - 1 = 3$, where c is the number of categories.

In the hair length example there are two categories ($c = 2$). The two samples are mutually independent as a student cannot be in both categories. Imagine that we tested 100 women students ($N = 100$). If there was no preference for hair length then we would expect to find half the students with long hair (probability, $p_1 = 0.5$, where 'long hair' is Category 1) and half the students with short hair (probability, $p_2 = 0.5$, where 'short hair' is Category 2). Thus, when the null hypothesis is true we would expect Np_1 students ($100 \times 0.5 = 50$) to have long hair and Np_2 (50 as well) to have short hair. Are the figures of 62 and 38 significantly different from the 50 we would expect in each category under the null hypothesis? This is where χ^2 comes in. The following formula turns out to approximate the χ^2 distribution when the null hypothesis is true.

$$\chi^2 = \sum\left(\frac{(X-Np)^2}{Np}\right) \text{ with } c - 1 \text{ degrees of freedom}$$

where X is the observed frequency count in a category and Np is the frequency count we would expect when the null hypothesis is true.

This is not exactly a χ^2 distribution but the approximation is very good as long as we make sure that Np is at least 5, that is the expected frequency of each category under the null hypothesis must be at least 5. This formula provides us with a distribution to compare our actual values to in order to test the significance of our differences between frequency counts.

There are two categories in the hair length experiment, so we can work out a χ^2 using the new formula.

$$\chi^2 = \frac{(X_1 - Np_1)^2}{Np_1} + \frac{(X_2 - Np_2)^2}{Np_2} = \frac{(62 - 50)^2}{50} + \frac{(38 - 50)^2}{50}$$

$$= \frac{144}{50} + \frac{144}{50} = 5.76$$

with $df = c - 1 = 2 - 1 = 1$ degree of freedom.

If we look up the tables for the χ^2 distribution (Table A.7 in the Appendix) the critical value for $\chi^2 = 3.84$ with $df = 1$ and $p = 0.05$. As the calculated value is greater than the table value we can conclude that there is a significant preference for long hair by the female students on campus.

The more usual way to express the above formula for χ^2 is to rename X as the observed frequency (O) and Np as the expected frequency (E) so the χ^2 formula that we use is:

$$\chi^2 = \sum\left(\frac{(O - E)^2}{E}\right) \text{ with } df = c - 1$$

Chi-square (χ^2) as a 'goodness of fit' test

In many cases we wish to examine whether a pattern of frequencies significantly differs from an expected pattern of frequencies. Usually the expected frequencies are those found when the null hypothesis is true but they do not have to be, we can compare the observed frequencies with any pattern of expected frequencies we wish to choose. This is why the test is called a 'goodness of fit' test: we can use it to decide if a set of observed frequencies are a good fit for a particular pattern of expected frequencies.

A worked example

An experimenter set out to test whether there is a difference in colour preference for cars. One hundred participants were given four pictures of cars, identical but for the colour, and asked to state their preference. The colours presented were red, blue, black and white.

If there was no preference then we would expect each colour to be chosen equally, so we would expect the probability of each category being chosen to be 1/4 or $p = 0.25$ when the null hypothesis is true. With a total (N) of 100, we would expect each category to be chosen by Np of them, 100×0.25, which is 25. On performing the experiment, the researcher finds 48 participants choose the red car, 15 the blue, 10 the black and 27 the white. Do these observed frequencies differ significantly from the expected frequencies?

We compare the pattern of observed frequencies with that of the expected frequencies by calculating χ^2.

$$\chi^2 = \sum\left(\frac{(O - E)^2}{E}\right)$$

$$= \frac{(48 - 25)^2}{25} + \frac{(15 - 25)^2}{25} + \frac{(10 - 25)^2}{25} + \frac{(27 - 25)^2}{25} = 34.32$$

with $df = c - 1 = 4 - 1 = 3$.

From Table A.7, $\chi^2 = 11.34$, $df = 3$, $p = 0.01$. As our calculated value of χ^2 is greater than the table value we can reject the null hypothesis. There is a significant difference ($p < 0.01$) between the observed and expected frequencies; the four colours are not equally preferred.

Testing the 'goodness of fit' to the normal distribution

In most cases we will compare observed frequencies with those found under the null hypothesis but there is one case in particular where we might choose another set of expected frequencies. We are often making the assumption with parametric tests that the sample or samples come from normally distributed populations. There might be occasions when we actually want to check this out. This is where the χ^2 goodness of fit test can be used.

Two hundred people were tested on a complex hand–eye co-ordination test and the number of errors each participant made was measured. The scores range from 22 to 69. The sample has a mean of $\overline{X} = 46.86$ and a standard deviation of $s = 6$. Does this sample differ significantly from the normal distribution?

First we choose the categories to adopt. The more categories we choose the more sensitive the test but we end up with fewer scores in each category. With a range from 22 to 69 categories of size 5 will result in 10 categories. These are shown in the first column of the table below. The boundaries of the categories are chosen at 0.5, half the smallest possible difference between the scores. (The minimum possible difference between the scores is 1, one error.) This is done so no two categories overlap. If I had taken 25 as a category boundary then a score of 25 could go into both the 20–25 and the 25–30 category but with 25.5 as a boundary it only goes into the 20.5–25.5 category and not the 25.5–30.5. It also means that there are no gaps between the categories, they cover the whole range. The next thing to do is to allocate the 200 scores to their correct categories. These are our observed frequencies and they are shown in the second column of the table.

We now need to work out the expected frequencies. To do this we convert the category boundaries to *z* scores using the *z* formula. Unfortunately we do not have the population mean and standard deviation which we need to work out a *z* score so we estimate them using the sample values, $\overline{X}$ and *s*.

$$\text{Estimated } z = \frac{X - \overline{X}}{s} = \frac{X - 46.86}{6}$$

For the first category, scores of 20.5 and 25.5 convert to *z* scores of -4.39 and -3.56. We do this for all the category boundaries. These results are shown in the third column of the table.

If we look these figures up in the standard normal distribution table (Appendix A.1) we can find the probabilities associated with each score. These probabilities are shown in the fourth column. (Recall that the probability of a *z* score less than -4 is so small as to be taken as zero.) The difference in the probability between the category boundaries will tell us the probability of finding a score in this category when the distribution is normal. These are shown in the fifth column. (It is a little difficult finding the probability of the category surrounding the mean as one *z* score is positive and one negative. We simply take the difference of each from 0.5 and add the results.)

Multiplying the probability of finding a score in a category when the distribution is normal (*p*) by the number of participants ($N = 200$) will give us the expected frequency in each category. These are shown in the sixth column.

Category	*Observed*	*z score*		*Diff.*	*Expected*	χ^2
boundary	*frequency*		*Probability*	*in prob.*	*frequency*	
20.5	1	−4.39	0.0000	0.0002	0.04	0.1317
25.5		−3.56	0.0002			
25.5	2	−3.56	0.0002	0.0030	0.60	
30.5		−2.73	0.0032			
30.5	2	−2.73	0.0032	0.0262	5.24	
35.5		−1.89	0.0294			
35.5	26	−1.89	0.0294	0.1152	23.04	0.3803
40.5		−1.06	0.1446			
40.5	55	−1.06	0.1446	0.2644	52.88	0.0850
45.5		−0.23	0.4090			
45.5	60	−0.23	0.4090	0.3201	64.02	0.2524
50.5		0.61	0.2709			
50.5	34	0.61	0.2709	0.1960	39.20	0.6898
55.5		1.44	0.0749			
55.5	16	1.44	0.0749	0.0633	12.66	1.6823
60.5		2.27	0.0116			
60.5	3	2.27	0.0116	0.0107	2.14	
65.5		3.11	0.0009			
65.5	1	3.11	0.0009	0.0009	0.18	
70.5		3.94	0.0000			

We are nearly ready to calculate χ^2, however, there are categories with expected frequencies less than 5 and we must not allow this for the test to be valid. What we can do to overcome this is to combine categories. If we combine the top three categories to make one new one and also do the same with the bottom three categories we end up with six categories all with expected frequencies greater than 5. The new category 20.5–40.5 has an observed frequency of 5 and an expected frequency of 5.88. the new category 55.5–70.5 has an observed frequency of 20 and expected frequency of 14.98. Finally,

$$\chi^2 = \sum\left(\frac{(O - E)^2}{E}\right)$$

$$= 0.1317 + 0.3803 + 0.0850 + 0.2524 + 0.6898 + 1.6823$$

$$\chi^2 = 3.2215$$

The degrees of freedom are one less than the number of categories so are $6 - 1 = 5$. However, in this case we did not know the population mean and standard deviation and used our sample to estimate them. In doing this we 'used up' a degree of freedom on each estimation, so we take our degrees of freedom as 3. From tables χ^2 is 7.82, $df = 3$, $p = 0.05$. We can conclude that as the calculated value is *less than the table value* we have *not found* a significant difference between the distribution of our scores and a normal distribution.

Chi-square (χ^2) as a test of independence

The χ^2 test of independence operates in the same way as the goodness of fit test in that it compares observed with expected frequencies, but in the test of independence we are comparing two or more patterns of frequencies to see if they are different from each other (independent or not). If we sampled conservatives and liberals on new taxation legislation then we could see if the pattern of frequencies 'for' and 'against' was different for the conservatives compared to the liberals using the χ^2 test.

A worked example

A researcher wanted to test the difference of opinion between conservatives and liberals on some new taxation legislation. In a survey, 120 people were identified as conservatives and 80 as liberals. A question on the survey asked whether the respondent agreed with the new taxation legislation ('for'), disagreed with it ('against'), or had no opinion or did not know about it ('don't know'). The results, the observed values, are shown in the table below.

Observed frequencies	*For*	*Against*	*Don't know*	*Row totals*
Conservatives	78	30	12	120
Liberals	18	50	12	80
Column totals	96	80	24	200

Notice that, with different numbers of conservatives and liberals, we would not expect the same numbers in the various categories even under the null hypothesis. As there are more conservatives than liberals the 12 conservatives in the 'don't know' category are 12/120 or 10 per cent of their group whereas the 12 liberals in the same category are 12/80 or 15 per cent of their group. Relatively more liberals gave this answer than conservatives. What we would expect, when there is no difference between the groups in their pattern of responses, is that there is the same proportion of each group total in each category. We can work out the expected values, when the null hypothesis is true, by the following formula.

$$\text{The expected value of a cell} = \frac{\text{row total} \times \text{column total}}{\text{overall total}}$$

A cell is a category, so we have six cells, $c = 6$. Let us take the first cell (conservatives-for) as an example. If there was no difference between the two political groups in terms of the proportion answering 'for' then the 96 people who actually responded 'for' should be divided into conservative and liberal in proportion to their relative number. Out of the 200 people the proportion of conservatives is 120/200. So, of the 96 people answering 'for' we would expect the following to be the number of conservatives if there is no difference between the groups:

$$E = \frac{96 \times 120}{200} = 57.6$$

We can do this for all the cells to produce the expected values.

Expected frequencies	*For*	*Against*	*Don't know*	*Row totals*
Conservatives	57.6	48.0	14.4	120
Liberals	38.4	32.0	9.6	80
Column totals	96	80	24	200

We now work out χ^2 using the usual formula.

$$\chi^2 = \sum\left(\frac{(O-E)^2}{E}\right) = \frac{(78-57.6)^2}{57.6} + \frac{(30-48.0)^2}{48.0} + \frac{(12-14.4)^2}{14.4} +$$

$$\frac{(18-38.4)^2}{38.4} + \frac{(50-32.0)^2}{32.0} + \frac{(12-9.6)^2}{9.6}$$

$$\chi^2 = 7.23 + 6.75 + 0.4 + 10.84 + 10.13 + 0.6 = 35.95$$

To decide whether this is significant we must compare it to the appropriate χ^2 distribution. We have to be careful here, the degrees of freedom is <u>not</u> the number of categories minus one, $c - 1$. This is because we are interested in comparing the rows (the two political groups) on pattern of results across the columns (the different opinions). This is a difference between the goodness of fit and test of independence. Here, we have 2 rows, $R = 2$, and two columns, $C = 3$. For the test of independence the degrees of freedom is:

$$df = (R-1)(C-1)$$

In our example $df = (2 - 1)(3 - 1) = 2$. From tables $\chi^2 = 9.21$, $df = 2$, $p = 0.01$. As our calculated value is greater than the table value we can reject the null hypothesis at the $p = 0.01$ level of significance. There is a significant difference in the patterns of responses of the conservatives and liberals to the taxation legislation.

We must make sure that the expected frequencies are 5 or larger for the χ^2 distribution to be appropriate. In this case there was not a problem. If the ‘don’t know’ responses had been too few for an expected frequency of 5 then we could leave out the ‘don’t know’ category and compare just the ‘for’ and ‘against’ for a valid test, or collect more data to make the frequencies larger.

The chi-square distribution

Being a squared value or a sums of squares χ^2 will always be greater than zero. However, the shape of the distribution will alter with changes in the degrees of freedom. Under the null hypothesis we would expect the sums of squares to be around zero but random variation will mean that they will not always be exactly zero when the null hypothesis is true. If we sum a number of positive values, each a little bigger than zero, the sum will gradually get larger the more numbers we add. The more degrees of freedom there are,

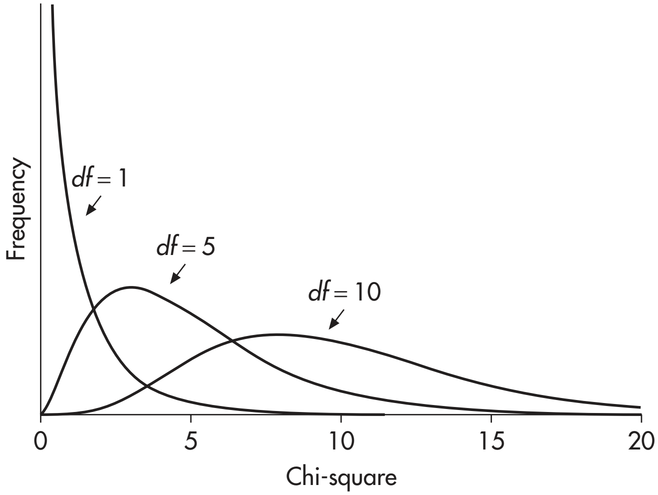

FIGURE 19.1 The chi-square distribution

then the more sums of squares we have and the larger these sums of squares become.

When *df* = 1 we expect, under the null hypothesis, most results to be close to zero with little difference between the observed and expected values (see Figure 19.1). Consider what the standard normal distribution would look like if we squared the values. Now when we increase the degrees of freedom we are adding together a set of independent χ^2s each with *df* = 1. Taking *df* = 5, for example, we have a sum of five independent χ^2s. Whilst each individual χ^2 will pile up close to zero, when added together their sum will pile up further along the scale (see Figure 19.1). As we increase the degrees of freedom the mean of the distribution moves up the scale. Whilst the distribution is very asymmetrical when the degrees of freedom are small, it becomes more symmetrical as *df* gets larger (see *df* = 10 in Figure 19.1). When the degrees of freedom get as large as 30 and above the distribution approximates the normal distribution. As a result of this tables of the χ^2 distribution usually only go up to *df* = 30, as beyond that we can use the tables of the normal distribution (Table A.7 in the Appendix).

The assumptions of the χ^2 test

In order that we compare our calculated value of χ^2 with the appropriate distribution we must make certain assumptions when performing a χ^2 test.

As with most distributions we must have randomly sampled from the population otherwise a biased sample will affect the resultant statistic. For χ^2 it is crucial that we have mutually independent categories. Essentially we must check that a subject could not possibly contribute to the frequency of more than one cell.

The chi-square distribution is 'continuous', meaning that there are no breaks in it, the curve is continuous. However, the values we calculate in the χ^2 test are not from a continuous scale but a discrete one. This is because observed frequencies vary in discrete units. We can observe a frequency of 10 or 11 but not 10.4 or 10.6. With degrees of freedom greater than 1 and with expected frequencies of at least 5 (and preferably 10) this is not a problem as the difference between the statistic and the true sampling distribution is so small. This is why large cell frequencies are encouraged. For example, the difference between 100 and 101 is small. It is a step of 1/100 or 1 per cent of the original frequency. However the step from 5 to 6 is 1/5 or 20 per cent, so is a large jump. Furthermore, because we are limited by the size of these steps (we cannot step in smaller units than whole numbers) any difference between observed and expected frequencies (even as small as 1) will appear large when we have small cell frequencies and χ^2 will tend to be significant (and possibly a Type I error).

To compensate for this problem when *df* = 1 we can apply the Yates' correction for discontinuity. This adjusts the χ^2 formula in the following manner.

$$\text{Corrected } \chi^2 = \sum \frac{(|O - E| - 0.5)^2}{E}$$

The lines either side of the $O - E$ refer to the absolute value, meaning that if the difference is negative we ignore the minus sign and treat it as positive. Thus, the χ^2 for every cell is reduced by 0.5 before it is squared. This will result in a smaller calculated value of χ^2 and will reduce the risk of a Type I error. However, the Yates' correction does tend to overcompensate for discontinuity and may result in a more conservative decision than necessary. As a simple rule, if a result is still significant with the correction or still nonsignificant without it, then we can be confident in our decision. It is only when a significant result becomes nonsignificant with the correction that a problem arises. In this case we should be cautious in making inferences from such a finding. As with any result that is 'bubbling under' (close but not quite significant) we should consider resolving the ambiguity by increasing the sample size or exploring the question further.

Details on how to calculate the chi-square statistic using the SPSS computer statistical package can be found in Chapter 14 of Hinton *et al*. (2004).

Chapter 20

Linear correlation and regression

Introduction

Do the students who spend the most time studying achieve the highest marks in examinations and do those who spend the least time studying get the lowest marks? What we are asking here is whether the variable *study time* correlates with the variable *examination performance*. If we found that this was the case then we would say that there is positive correlation between the variables, that is, as a score on one variable increases so the corresponding score on the other variable does the same. Sometimes we find a correlation between two variables where as one goes up the other goes down. This is termed a negative correlation. We are likely to find a negative correlation between smoking and health as the more a person smokes the less healthy that person tends to be.

If we find that two variables do correlate then we can use this information to predict the value of a score on one variable by using the corresponding score on the other variable. In this chapter we shall be looking at how we can produce a regression equation to allow us to do this. If we do not find a relationship between two variables we say that they are uncorrelated and a change in one cannot be used to predict a change in the other.

As an example we shall use the following data, giving the results of ten first year university students, showing how much time they spent studying (on average per week throughout the year) along with their end of year examination mark (out of 100). Do these data show a correlation?

Student	*Study time*	*Examination mark*
1	40	58
2	43	73
3	18	56
4	10	47
5	25	58
6	33	54
7	27	45
8	17	32
9	30	68
10	47	69

There appears to be a positive correlation when we look at these results by eye but a clearer way to show this is to produce a scatterplot, that is a graph of the data, where the axes are the two variables. Figure 20.1 provides a scatterplot of these results.

Note that the points are not randomly scattered about the graph (which we would expect if there was not a correlation) but generally fall within a band, indicating a correlation. (To illustrate this, imagine cutting out a piece of paper to cover up all, or most of, the points in the graph. We can do this,

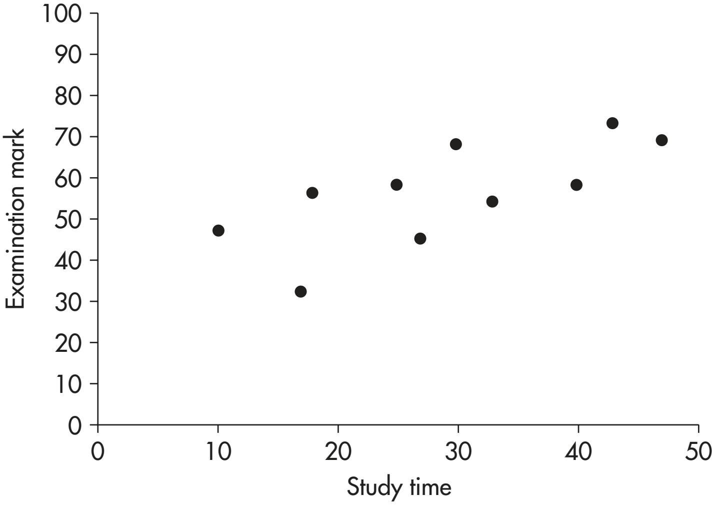

FIGURE 20.1 Scatterplot of study time by examination performance

in this case, with a fairly narrow strip of paper.) When this occurs we argue that, but for random errors, the scores would have fallen along a line, the regression line, and in our analysis we can calculate which line would 'best fit' the data. In many cases, but not all, we assume that the line of best fit is a straight line. When we make this assumption we are assuming that we have a linear correlation, and we calculate a linear regression. This is also referred to as a linear model as we are assuming that the model for the relationship between the variables is a straight line (see Chapter 23 on linear models). This is a reasonable assumption in our example as the points on the graph fall within a band that appears straight. If the pattern of points had been along a curved line there would still be a correlation but it would not be linear. In this book I am only considering linear correlation and regression.

What we need to do is to find a way to measure the strength of the correlation. If all the points lie exactly along a straight line then we have a perfect correlation. A correlation such as the one in Figure 20.1 is not perfect as the points are more widely scattered but they still fall within a fairly narrow band. This is a reasonable correlation, as we could infer that the points would lie on a straight line but for random errors. As the points become more scattered so the correlation gets weaker until we say that they are randomly scattered, and there is no correlation at all. The measurement we use to describe the degree with which the points cluster along a straight line is the Pearson correlation coefficient, *r*.

Pearson *r* correlation coefficient

In our example, as in most we examine, the two variables are measured on different interval scales. This makes it difficult to decide how well the scores on one variable correlate with the scores on the other variable. Is 30 hours per week as large a *study time* score as 60 out of 100 on *examination performance*? To overcome this problem we need to standardise the scores. We do this by finding the *z* scores of the scores on the two variables.[18] The standard scores find the position of a score relative to its mean in terms of its standard deviation. By calculating standard scores we can compare the relative position of each score on the distribution of the variable. Study time has a mean of 29 and a standard deviation of 11.42. We will call this variable *X*. Examination performance has a mean of 56 and a standard deviation of 11.80. We will call this variable *Y*. The *z* scores for each variable are shown in the table below.

Student	*Study time*	*Study time z score*	*Examination mark*	*Examination z score*	*Product of the z scores*
	X	z_X	Y	z_Y	z_Xz_Y
1	40	0.96	58	0.17	0.16
2	43	1.23	73	1.44	1.77
3	18	−0.96	56	0.00	0.00
4	10	−1.66	47	−0.76	1.26
5	25	−0.35	58	0.17	−0.06
6	33	0.35	54	−0.17	−0.06
7	27	−0.18	45	−0.93	0.17
8	17	−1.05	32	−2.03	2.13
9	30	0.09	68	1.02	0.09
10	47	1.58	69	1.10	1.74

We can now see whether the score on one variable corresponds to the same position on its distribution as the score on the second variable for each participant. Looking at the table above, the *z* scores tend to be similar for each participant: a similar size of *z* score indicates a correlation and the same sign (either both positive or both negative) indicates a positive correlation. (Had the sizes been similar but the signs different we would have been looking at a negative correlation.) How can we acknowledge this similarity mathematically? One way is to multiply the *z* scores on the two variables for each participant. When there is a correlation the size of the *z* scores will be similar, so large numbers will be multiplied by large numbers and small numbers by small numbers. With a positive correlation we will mostly multiply *z* scores of the same sign together (either both positive or both negative) to produce products that will be mostly positive. With a negative correlation we will multiply mostly *z* scores with different signs and the products will be mostly negative. Thus, if we sum the products of the *z* scores ($\sum z_Xz_Y$) we should get a large positive number when there is a positive correlation and a large negative number when there is a negative correlation. If there is no correlation at all we should get some positive products and some negative products which will tend to cancel each other out and the sum ends up around zero. If there is a perfect correlation the participants will get the same *z* score on both variables. Multiplying these together is like squaring the *z* scores of one of them. The sum of N squared *z* scores always equals N

(try it!) so a perfect positive correlation will result in the sum of the product of the z scores equalling N. When there is a perfect negative correlation the sum will be $-N$. In our example, $\sum z_X z_Y = 7.2$, so it is a positive correlation (above 0) but not perfect as $N = 10$ (we have 10 participants).

Finally, if we divide the sum of the products of the z scores by N we produce a statistic that equals 1 when there is a perfect positive correlation, −1 when there is a perfect negative correlation and 0 when there is no correlation at all. This statistic is called the Pearson correlation coefficient r.

$$r = \frac{\sum z_X z_Y}{N}$$

A positive correlation is shown by an r greater than zero and a negative correlation by r less than zero. The strength of the correlation is shown by how close r is to 1 (or −1 if the correlation is negative). In our example $r = 0.72$, which is a high positive correlation as it is much closer to 1 than 0. We will see whether it is significant in a moment.

The importance of r is that, as well as telling us the strength and direction of a correlation, it also provides us with a formula for predicting the scores on one variable by using the scores of the other variable. If we plotted the z scores of the two variables on a scatterplot we would find that r is the slope of the regression line (the straight line that best represents the linear relationship between the variables, the 'line of best fit'), the line we assume the z scores would fall along but for random error. If we write the formula for the line on the graph that best fits the z scores it is $z_Y = rz_X$. Thus, given any z score on one variable we can use this formula, now we know r, to predict what the z score would be on the other variable if the scores fell along a straight line. This is all very well but we are not actually interested in z scores! We need to get back to the original scores.

A convenient way to work out r

We do not need to work out z scores to find r. We can use an alternative formula that is identical to that above but involves only the original scores.

$$\text{Pearson's } r = \frac{SP}{\sqrt{SS_X \times SS_Y}}$$

SP is called the sums of products and gives a measures of how the scores of the two variables vary together:

$$SP = \sum(X - \bar{X})(Y - \bar{Y}) = \sum XY - \frac{(\sum X)(\sum Y)}{N}$$

SS_X is the sums of squares of the scores of the first variable, labelled *X* (in our example *study time*). This gives a measure of how these scores vary on their own:

$$SS_X = \sum(X - \bar{X})^2 = \sum X^2 - \frac{(\sum X)^2}{N}$$

SS_Y is the sums of squares of the scores of the second variable, labelled *Y* (in our example *examination performance*). This gives a measure of how these scores vary on their own:

$$SS_Y = \sum(Y - \bar{Y})^2 = \sum Y^2 - \frac{(\sum Y)^2}{N}$$

We can see that *SP* will be large if each *X* score is the same distance from its mean $\bar{X}$ as each *Y* score is from its mean $\bar{Y}$. If the *X* and *Y* scores do not vary together *SP* will be small and in the case of no correlation it will become zero. The formula $\sqrt{SS_X \times SS_Y}$ gives us a measure of individual variability of the scores in the two variables. If we can explain all the individual variability of the scores by the joint variability (*SP*) then $\sqrt{SS_X \times SS_Y}$ and *SP* will be the same size and *r* will be +1 for a positive correlation and −1 for a negative correlation.

We can use our example to show the calculation:

Participant	*X*	X^2	*Y*	Y^2	*XY*
1	40	1600	58	3364	2320
2	43	1849	73	5329	3139
3	18	324	56	3136	1008
4	10	100	47	2209	470
5	25	625	58	3364	1450
6	33	1089	54	2916	1782
7	27	729	45	2025	1215
8	17	289	32	1024	544
9	30	900	68	4624	2040
10	47	2209	69	4761	3243

$N = 10 \qquad \sum X = 290 \quad \sum X^2 = 9714 \quad \sum Y = 560 \quad \sum Y^2 = 32752 \quad \sum XY = 17211$

$$SP = \sum XY - \frac{(\sum X)(\sum Y)}{N} = 17211 - \frac{290 \times 560}{10} = 971$$

$$SS_X = \sum X^2 - \frac{(\sum X)^2}{N} = 9714 - \frac{290 \times 290}{10} = 1304$$

$$SS_Y = \sum Y^2 - \frac{(\sum Y)^2}{N} = 32752 - \frac{560 \times 560}{10} = 1392$$

$$r = \frac{SP}{\sqrt{SS_X \times SS_Y}} = \frac{971}{\sqrt{1304 \times 1392}} = 0.72$$

We now have to work out the probability of finding a value of *r* as large or larger than 0.72 by chance, that is when there really is no correlation between the variables. Only then can we decide if we have found a significant correlation.

The distribution of *r*

When there is no correlation between two variables we would expect *r* to be zero. However, there will be random variation around this point. We will,

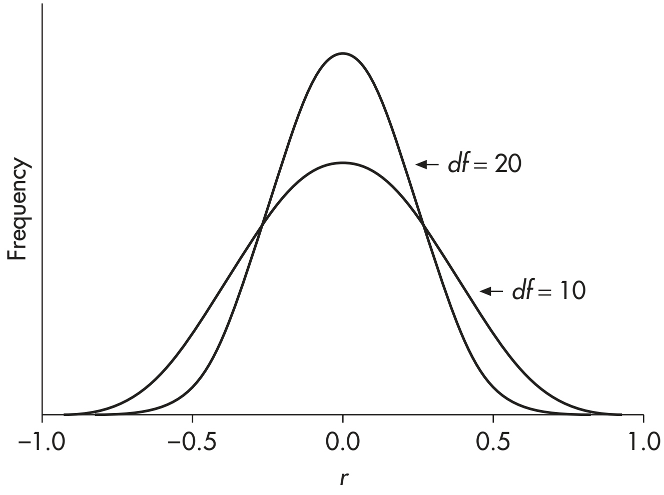

FIGURE 20.2 The distribution of Pearson's *r*

by chance, obtain values of r that deviate from zero but this will become less likely as we get closer to +1 or −1. We can see from this that the distribution of r under the null hypothesis will be symmetrical about a mean of 0, tailing off towards +1 and −1. The distribution will be flatter when there are fewer subjects and more bunched around the mean when there are more subjects. When there are more subjects the effect of individual subjects will have less influence on the correlation so there will be less chance of r deviating so far from zero.

It is not the actual number of subjects that is important when considering which distribution of r to compare our calculated value to, but the degrees of freedom. For r the degrees of freedom is $N - 2$ (and not $N - 1$) for the following reason. r is actually the slope of the 'best fit' regression line for the z scores. We need the information from at least two points to draw a specific straight line, so we have 'used up' two of our degrees of freedom in finding this line. (In other tests we use up only one degree of freedom on the sample mean.) The distribution of r is shown in Figure 20.2.

A prediction about a correlation can be one-tailed or two-tailed. A one-tailed test specifically states whether the correlation will be positive or negative, whereas a two-tailed prediction merely predicts a significant correlation. We need to take account of this in setting the significance level. In our example we are predicting a positive correlation, that examination performance increases as study time increases, so we have a one-tailed test. From the tables of r (Table A.9 in the Appendix), for a one-tailed test at

$p = 0.05$, with 8 degrees of freedom, $r = 0.5494$. As our calculated value of 0.72 is greater than the table value we can reject the null hypothesis and claim a significant correlation between the variables.

Linear regression

There are books that separate linear correlation from linear regression by putting them in different chapters. It can appear neater that way but we should not lose sight of the fact that correlation and regression are like the two sides of a coin. A linear correlation tells us how close the relationship between two variables is to a straight line. A linear regression is the straight line that best describes the linear relationship between the two variables. With a high correlation we are able to see (more or less) where the regression line occurs by drawing the scatterplot. It is not so obvious when the correlation is weak as the points might be scattered more widely than a narrow band. Yet even though we get a low correlation we can still ask: if there is a linear relationship between these variables what would that line be?

With a regression line we can predict what a score on one variable will be given a score of the other variable. We saw that r is the slope of the regression line for the z scores but this is not what we want. We would like to know the line of best fit for the actual scores so that we can predict a score on one variable from the other directly without having to go through the z scores.

We need a little algebra here, although it should not be too painful. The formula for a straight line relationship between two variables X and Y, is $Y = a + bX$, where 'a' and 'b' are constants (they always stay the same even though X and Y vary) and X and Y are the two variables. You can choose any numbers for a and b, then put any values of X you choose into the equation, work out Y, plot X and Y on a graph and the points will fall along a straight line every time. For example, if I choose, say $a = 2$ and $b = 3$ then $Y = 2 + 3X$ is a straight line. I can take any value of X, say 4, then find $Y = 2 + (3 \times 4) = 14$. I can do this for any value of X and if I plot X and Y on a graph the points will fall along a straight line. When $X = 0$ then $Y = a$ (in my example when $X = 0$, $Y = 2$), so a is the point where the straight line cuts the Y axis. The slope of the line is given by the constant b, which tells us how steeply the line rises or falls. It is like walking along a straight road going up or down hill. A slope of more than 1 is steep, as with every step we take along the X axis we are going up hill, along the Y axis, by at least the same amount and the line lies relatively close to the Y axis.

A slope of less than 1 is shallow, as with every step along the X axis we rise, along the Y axis, by less than that amount and the line lies closer to the X axis than the Y axis. Try making up a few straight line equations and plotting some points for each line on a graph with a horizonal X axis and a vertical Y axis.

We can employ the straight line formula in working out the regression line for the two variables under study. If there is a perfect correlation ($r = +1$ or -1) then the points on the scatterplot will all lie along a straight line. This is our regression line. More usually we do not get a perfect correlation and the regression line is less obvious. With the linear model we are assuming that the points would lie on a straight line but for the random variation. So we need to work out what is the most likely straight line for the data. Notice that a significant correlation gives us the confidence that there is a genuine linear relationship between the two variables. When the correlation is weak we can still work out a regression line but the linear relationship might not be genuine.

First we must decide which variable to predict (in our case we choose *examination performance*, variable Y) and which variable to use for prediction (*study time*, variable X). The first stage in the logic of regression analysis is to assume that the scores for variable X are correct and the reason why the Y scores do not fall along a straight line is due to random error. We are basing our analysis on the X scores. We express this in a formula in the following way:

$$Y = \text{Regression (on } X\text{)} + \text{Error}$$

$$Y = Y' + E$$

We are assuming that the actual Y scores are a combination of the ones along the straight line (Y') plus a deviation from that straight line due to error (E). What we want to know is what Y values we would get if they really did fall along a straight line and we could get rid of the error: what are the values of Y' where $Y' = Y - E$? The straight line that we are looking for is therefore:

$$Y' = a + bX$$

which is the regression line of Y on X without the error (E). What we now have to find are the appropriate values for a and b.

Next in the analysis we use the fact that the 'line of best fit' is the line that gives the smallest error values. We do not want a line that is nowhere

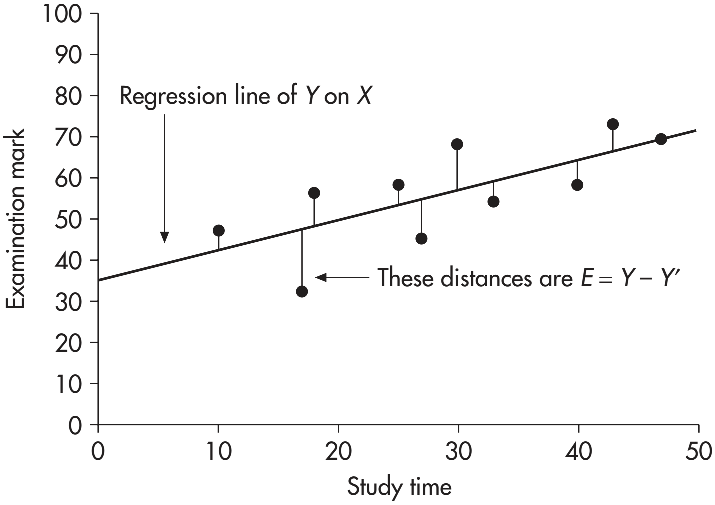

FIGURE 20.3 Finding the regression line by minimising the error values (*E*)

near the points on the scatterplot. The regression line should be the straight line that goes closest to the data points. We want to find the line that produces the smallest values for E, where $E = Y - Y'$. A mathematical way of putting it is to say that we want the line that 'minimises' E where E is the distance of an actual data point from the regression line. Figure 20.3 shows this for our example.

We work out the minimum values of E by a procedure called the least squares method of linear regression. We could add up the error $(Y - Y')$ for each subject to produce $\sum E = \sum(Y - Y')$ but some errors will be positive and some negative and so cancel each other out (as you can see from Figure 20.3), hiding the size of the error. To overcome this we square the errors so that they all become positive, to produce the sums of squares: $\sum E^2 = \sum(Y - Y')^2$. (Once again we can see the importance of 'sums of squares' at the heart of a statistical analysis.) Now we need to find when this sums of squares is at its smallest. We can replace Y' by $a + bX$ in the sums of squares to give a formula containing only X and Y, which are the values we know rather than Y' which we want to find out: $\sum(Y - a - bX)^2$. We now want to know what values of a and b would minimise this formula so that $\sum E^2 = \sum(Y - a - bX)^2$ is the smallest it can be. The way we do this is by employing a mathematical technique called differentiation. (There is not space to explain differentiation here, but for readers not familiar with it, all that is necessary to know for the logic of the current argument is that this technique exists and helps us at this point in deriving the regression line.) As a result of this, the above sums of squares is at its minimum when:

$$b = \frac{SP}{SS_X} \text{ and } a = \overline{Y} - b\overline{X}$$

where $\overline{X}$ and $\overline{Y}$ are the means of the scores of the two variables, and SP is the sums of products and SS_X the sums of squares for the scores on variable X that we worked out in the calculation of r.

All we need to do now is work out a and b to produce the regression line. For our example, looking back to the calculation of r, we have $SP = 971$, $SS_X = 1304$, $\overline{X} = 29$ and $\overline{Y} = 56$ so:

$$b = \frac{971}{1304} = 0.7446 \quad \text{and} \quad a = 56 - (0.7446 \times 29) = 34.4057$$

Finally, replacing a and b by their actual values in the formula for Y', we are able to express the regression line by the following formula:

$Y' = 34.41 + 0.74X$ (to two decimal places).

We can now use this formula to predict the values of Y (*examination performance*) from the values of X (*study time*). Below is a table of the predicted values of Y based on the regression on X.[19]

Student	*Study time* *X*	*Examination mark* *Y*	*Predicted examination mark* *Y′*
1	40	58	64.01
2	43	73	66.23
3	18	56	47.73
4	10	47	41.81
5	25	58	52.91
6	33	54	58.83
7	27	45	54.39
8	17	32	46.99
9	30	68	56.61
10	47	69	69.19

We can also use the regression line to predict other values. For example, no one studied for 35 hours per week. What examination mark would we predict for someone who did study for this time? Using the formula for Y' we get: $Y' = 34.41 + (0.74 \times 35) = 60.31$. We would expect a student who studied for 35 hours per week to get a mark of 60.31 in the examination.

r and the slope of the regression line

We have found b, the slope of the regression line, and r, the correlation coefficient, which is the slope of the z scores regression line. There is a simple relationship between the two:

$$b = r\left(\frac{\text{standard deviation of } Y}{\text{standard deviation of } X}\right)$$

b takes account of the fact that the two variables are measured on two different scales, whereas r standardises them. In our example: $b = 0.72\left(\frac{11.80}{11.42}\right) = 0.74$. So, whichever way we work out b we get the same value.[20]

Predicting *X* from *Y*

There is nothing in the logic of the regression analysis that prevents us from performing the regression the other way round, by assuming the Y values are correct and that it is the X values that deviate from a regression line due to error. The logic works in the same way to predict X from Y by the regression of X on Y. In this case we find $X' = a + bY$ (which is also a formula for a straight line), where $b = \frac{SP}{SS_Y}$ and $a = \overline{X} - b\overline{Y}$. In our example, we find $X' = 0.70Y - 10.06$. From this formula we can predict that someone who obtained a 60 in the examination studied for $(0.70 \times 60) - 10.06 = 31.94$ hours per week.

If we plot both the regression lines (Y on X, and X on Y) on the same graph we find, in our case, that they are close together (see Figure 20.4). This is because the stronger the correlation the closer the regression lines are to each other. With a perfect correlation the lines are the same. As the

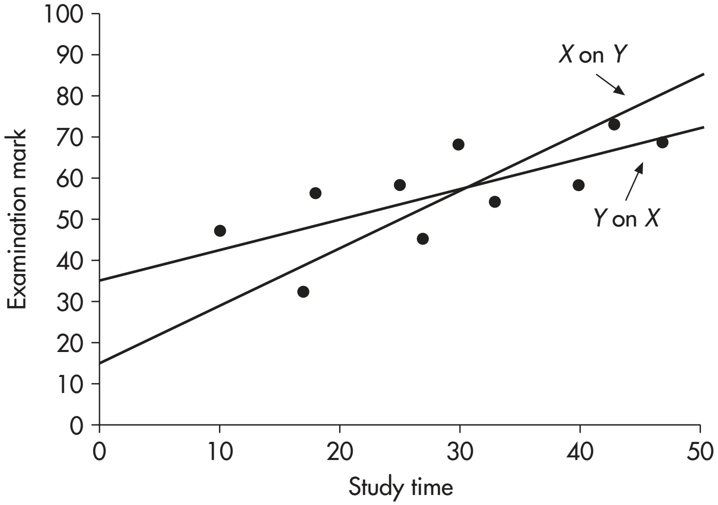

FIGURE 20.4 Regression of *Y* on *X* and the regression of *X* on *Y*

correlation gets weaker the regression lines separate until, when $r = 0$, the lines are orthogonal, that is at right angles to each other and have no predictive value as there is not a linear relationship between the variables.

The interpretation of correlation and regression

We must be careful when we interpret a significant correlation coefficient. The first point to note is that a smaller value of r is needed for significance as N increases. With a *df* of 70 for a one-tailed test, or a *df* of 100 for a two-tailed test, r is still significant (at $p = 0.05$) when it is as low as 0.2. With correlation coefficients we need to ask not just is it significant but is it big? One way of deciding the importance of the correlation is to consider how much of the variability of the scores in one variable can be explained (predicted) by the variability of the scores of the other variable. We might have a significant correlation but if it only explains a tiny amount of the variability then it may not be of much predictive worth.

Recall that Y = Regression on X + Error. We also find that the variability of the Y scores ($SS_Y = \sum(Y - \overline{Y})^2$) equals the variability due to the regression ($SS_{regression} = \sum(Y' - \overline{Y})^2$) plus the variability due to error ($SS_{error} = \sum(Y' - Y)^2$). It is reasonable to ask how much of the total variability of Y can be explained by that of the regression. We can express this as

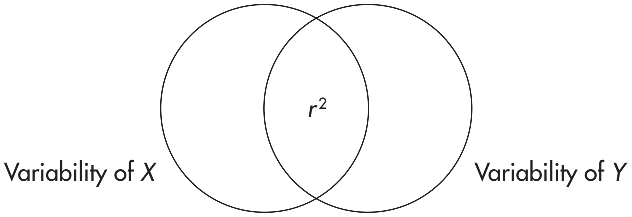

FIGURE 20.5 The coefficient of determination (r^2)

$SS_{regression}$ as a proportion of SS_Y. How much of the total Y sums of squares can be explained by the sums of squares of the regression on X? It turns out that:

$$\frac{SS_{regression}}{SS_Y} = \frac{SP^2}{SS_X SS_Y} = r^2$$

We find that the proportion of the variability in one set of scores that can be explained by the regression is actually the square of the regression coefficient, r^2, called the coefficient of determination. We can represent r^2 diagrammatically in Figure 20.5. A circle represents the total variability of the scores for one variable. The overlap of the two circles indicates the amount of variability of one variable that can be explained by the variability of the other variable, r^2.

With a perfect correlation of $r = +1$ or -1 then $r^2 = 1$ and all variability in the Y scores can be explained by the regression. The regression line is a perfect predictor of the Y scores. A high correlation, such as $r = 0.7$, yields an r^2 of 0.49 which tells us not quite half of the variability in Y can be explained by changes in X (and vice versa). With a correlation of 0.2 only 0.04 of the variability of the Y scores can be explained by the regression on X, so, in this case, despite the statistical significance we have every right to question the value of X as a predictor of Y.

Problems with correlation and regression

We must be careful to check that our data has homoscedasticity when we are undertaking a correlation. Homoscedasticity essentially means that the relationship between the two variables stays the same at all points, with the scores evenly spread along and around the regression line. Isolated points and clusters can both have a powerful influence on the correlation coefficient, and disguise the underlying relationship between the variables, particularly if we use a limited range of scores from the variables.

An example will illustrate these points. A researcher predicts that the more shop assistants smile at customers the more items are sold by the assistant. Each assistant in a store is videotaped during one day and the amount of smiling is calculated from the time an assistant greets a customer to the moment the customer decides to buy or not to buy an item. The researcher examined the correlation between the mean *smiling time* per customer for each assistant (in minutes) and the total number of *items sold* by each assistant during the day. The results for 9 assistants are shown below.

Assistant	*Smiling time* *X*	*Items sold* *Y*
1	0.4	16
2	0.8	12
3	0.8	20
4	1.2	16
5	1.4	34
6	1.8	30
7	2.2	26
8	2.6	22
9	3.0	38

When we take all 9 participants into account we find that $r = 0.69$ ($SP = 43.56$, $SS_X = 6.28$, $SS_Y = 627.56$, $df = 7$). This is significant at $p < 0.05$ (from Table A.9 in the Appendix we find that $r = 0.5822$, $p = 0.05$, $df = 7$, for a one-tailed test). However, looking at the scatterplot, Figure 20.6, we see that participant 9 is isolated from the rest. Without this participant $r = 0.52$ ($SP = 20.80$, $SS_X = 4.00$, $SS_Y = 400.00$, $df = 6$) which is no longer significant (as $r = 0.6215$, $p = 0.05$, $df = 6$, for a one-tailed test). Thus the effect of participant 9 is to make the correlation significant yet participant 9 is not typical, and so we should not take the result as practically useful despite its statistical significance. This shows how one ‘outlier’ can strongly affect the correlation.

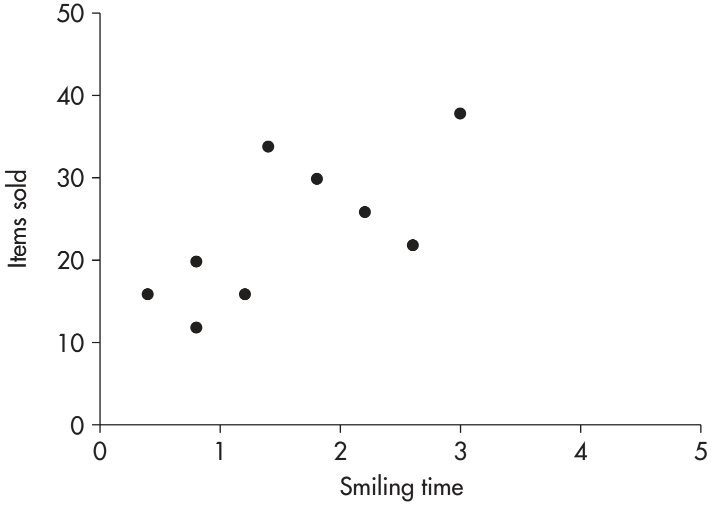

FIGURE 20.6 The scatterplot of smiling time by items sold

If we look at the scatterplot we can also see that the pattern of results is not the same for all participants: the relationship between smiling and items sold is not the same all along the regression line. If we limit our range to participants 1 to 4 we find that $r = 0$ ($SP = 0$, $SS_X = 0.32$, $SS_Y = 32.00$, $df = 2$). There is no correlation at all for these participants alone. If we now select only participants 5 to 8 we produce a correlation coefficient of $r = -1$ ($SP = -8.0$, $SS_X = 0.8$, $SS_Y = 80$, $df = 2$), which is a perfect negative correlation. These two clusters produce very different results which illustrates why we do not want a limited range in our study. The lack of homoscedasticity has resulted in a positive, zero and negative correlation dependent on which participants we select.

A similar spread of data along the regression line provides evidence that the correlation does in fact indicate a genuine underlying relationship between the variables. Isolated points, clusters and a limited range can all provide spurious correlations. We must look a little further than a statistically significant *r* when we are interpreting the meaning of a correlation.

The standard error of the estimate

We can always find a regression line for our data, regardless of the value of *r*, but just because we can calculate it does not mean that it is of theoretical significance. To be confident that our predictions are based on a genuine underling relationship we really want all the points to be close to the regression line. A way of determining how close the points are to the regression line is to calculate the standard error of the estimate, which, for the regression of Y on X, is the standard deviation of the Y scores from the regression line of Y on X. Recall that a variance is a sums of squares divided by a degrees of freedom. So the error variance, the amount by which the Y scores vary from the regression line, is $\frac{SS_{error}}{N-2}$. We find the square root to produce a standard deviation.

$$\text{Standard error of the estimate} = \sqrt{\frac{SS_{error}}{N-2}}$$

We also know from above that $SS_Y = SS_{regression} + SS_{error}$ and $r^2 = \frac{SS_{regression}}{SS_Y}$. From these two formulae we can show that $SS_{error} = (1 - r^2)SS_Y$. Replacing SS_{error} in the formula for the standard error of the estimate we get:

$$\text{Standard error of the estimate} = \sqrt{\frac{(1-r^2)SS_Y}{N-2}}$$

For the study time/examination performance example we have $r^2 = 0.52$ and $SS_Y = 1392$, so the standard error of the estimate, the standard distance of a Y score from the regression line is: $\sqrt{\frac{(1-r^2)\,SS_Y}{N-2}} = \sqrt{\frac{(1-0.52)1392}{10-2}}$ $= 9.14$.

The Spearman r_S correlation coefficient

There will be times when we wish to correlate data that is not measured on a interval scale. As long as the data are ordinal we can perform a correlation on the ranks using the Spearman r_S correlation coefficient. Each set of scores is ranked separately from lowest to highest. A Pearson's r is then calculated on the ranks. However, with ranks, as long as there are no ties, we can use a simpler formula. There will be the same ranks for both sets of scores so $SS_X = SS_Y$. If we replace SS_Y with SS_X in the formula for r we get:

$$r = \frac{SP}{\sqrt{SS_X \times SS_Y}} = \frac{SP}{SS_X}$$

It is also the case that with ranks $SP = SS_X - \frac{\sum D^2}{2}$, where D is the difference between a subject's ranks on the two variables. Furthermore, with ranks, $SS_X = \frac{N^3 - N}{12}$. Replacing SP and SS_X in the formula for r we get:

$$\text{Spearman's } r_S = 1 - \frac{6\sum D^2}{N^3 - N}$$

All we have to do for ranked data is work out r_S. We then look up the figure in the tables for r_S at the chosen level of significance (Table A.10 in the Appendix). In this case we do not use the degrees of freedom to find the correct table value of r_S but N, the number of ranks. As with all analyses on ranks we have to be careful if there are many tied ranks and should consider employing a more sensitive measure of the variable to reduce them. Alternatively, the original Pearson formula can be used.

The Spearman coefficient is useful if we are concerned that the scores on two variables appear to correlate but not linearly. As long as the two variables vary monotonically, that is as one increases the other also increases consistently or as one increases the other decreases consistently, then the r_S coefficient can be used.

A worked example

Two teachers were asked to rate the same six teenagers on the variable *how likely to do well academically at University* on a 0–20 scale, from unlikely to highly likely. The results are shown below. Is there a positive correlation between the teachers' ranking?

Teenager	*Teacher 1 ratings*	*Teacher 2 ratings*	*Teacher 1 ranks*	*Teacher 2 ranks*	*D*	*D*^{2}
1	15	8	4	3	1	1
2	12	13	3	5	–2	4
3	18	16	6	6	0	0
4	4	5	1	2	–1	1
5	8	2	2	1	1	1
6	17	10	5	4	1	1
						$\sum D^2 = 8$

The ratings for each teacher are ranked separately. From these we produce the difference scores (D), showing the difference in ranks between the teachers, and the squared difference scores (D^2). The sum of the difference scores, $\sum D^2 = 8$. There are 6 participants so $N = 6$. We now work out r_S.

$$r_S = 1 - \frac{6\sum D^2}{N^3 - N} = 1 - \frac{6 \times 8}{6^3 - 6} = 0.77$$

We have a one-tailed test as the prediction is for a positive correlation. From Table A.10 in the Appendix, $r_S = 0.829$, $p = 0.05$, $N = 6$ for a one-tailed test. The calculated value does not exceed the table value so we have not found a significant correlation in the rankings. (Notice how, with a small number of subjects, we need a high value of the coefficient for significance.)

Details on how to calculate a linear correlation and linear regression using the SPSS computer statistical package can be found in Chapter 15 of Hinton *et al.* (2004).

Chapter 21

Multiple correlation and regression

chapter 21

Introduction to multivariate analysis

Up to now we have looked at the correlation between two variables. Yet we can consider the correlation between three or more variables, say IQ, school grades, university grades and occupational performance. Dealing with many variables at the same time is referred to as multivariate analysis. In this chapter we shall be examining both correlation and regression with more than two variables as this is often an important form of analysis when we collect information about a number of factors (such as in a questionnaire or survey) and we want to investigate the relationships between them. For example, we might wish to study the relationship between housing quality, housing density, social support networks and pollution levels on health.

Partial correlation

In the previous chapter we analysed some example data to show a significant correlation between study time and examination performance. We might decide that a third variable, *intelligence*, could be influencing the correlation. If intelligence positively correlates with study time, that is, the more intelligent students spend the most time studying, and if it also positively correlates with examination performance, that is, the more intelligent students get the higher marks in the examination, then the correlation of study time and examination performance might simply be due to the third factor, intelligence. If this is the case then the relationship between study time and examination performance is not genuine, in that the reason they correlate is because they are both an outcome of *intelligence*. That is, the more intelligent students both study more and get higher marks in the examination. If we take out the effect of intelligence the relationship of study time to examination performance could disappear.

It is worth noting here that a correlation does not indicate a causal relationship. We might find that over a period of years the number of houses positively correlates with the amount of pollution in a town. It would be wrong to claim that the houses cause the pollution or that more pollution causes more houses. In this case, the correlation might arise due to a third factor *population*, which correlates with both. An increase in population (and human activity) might result in both more houses and also greater pollution. The correlation between houses and pollution is simply an outcome of a third factor rather than an important correlation in its own right.

To answer the question of the influence of intelligence on the study time/examination performance correlation we need to examine the correlation of study time and examination performance *after* removing the effects of intelligence. If the correlation disappears then we know it was due to the third factor. We do this by calculating a partial correlation. The first stage is to find out how well the factor intelligence correlates with study time and examination performance separately. To find this out we measure the students' intelligence on a standard test of intelligence. The results of this test along with the study times and examination marks are shown in the following table.

Student	*Intelligence score*	*Study time*	*Examination mark*
1	118	40	58
2	128	43	73
3	110	18	56
4	114	10	47
5	138	25	58
6	120	33	54
7	106	27	45
8	124	17	32
9	132	30	68
10	130	47	69
Mean	122	29	56
Standard deviation	9.72	11.42	11.80

Using the techniques outlined in the previous chapter we find the following correlation coefficients:

Study time and Examination performance	$r = 0.72$
Study time and Intelligence	$r = 0.37$
Examination performance and Intelligence	$r = 0.48$

The correlations indicate that intelligence is positively correlated with the other two variables so there is reason to continue the investigation.

Recall from the previous chapter that the regression allows us to predict one variable from a second. If we perform a regression of study time on intelligence this will tell us what study time scores we would predict from intelligence. Thus, the difference between the actual study time scores and those predicted by intelligence should give us the study time scores with the effects of intelligence removed. These differences are termed residuals rather than 'error' here because, whilst the difference is an 'error' in the ability of intelligence to predict study time, in this case it is what we are interested in, that is, what is left (the residual variability in the scores) after taking out the effects of intelligence on study time.

Performing a regression of study time on intelligence we get the following equation: Study time = 0.44 × Intelligence – 24.50. From this we can work out the predicted study time scores and then subtract them from the actual scores to give the residuals. The following table shows this (see Note 19).

Student	*Study time*	*Study time predicted by intelligence*	*Residual study time*
1	40	27.42	12.58
2	43	31.82	11.18
3	18	23.90	–5.90
4	10	25.66	–15.66
5	25	36.22	–11.22
6	33	28.30	4.70
7	27	22.14	4.86
8	17	30.06	–13.06
9	30	33.58	–3.58
10	47	32.70	14.30

This has removed the effect of intelligence from study time. We now need to remove it from the examination performance. We follow the same method and perform a regression of examination performance on intelligence. This gives us the regression equation: Examination performance = 0.59 × Intelligence – 15.60. We use this equation to work out the residuals for examination performance.

Student	*Examination mark*	*Examination mark predicted by intelligence*	*Residual examination mark*
1	58	54.02	3.98
2	73	59.92	13.08
3	56	48.30	6.70
4	47	51.66	−4.66
5	58	65.82	−7.82
6	54	55.20	−1.20
7	45	46.94	−1.94
8	32	57.56	−25.56
9	68	62.28	5.72
10	69	61.10	7.90

We can now correlate the residual study time scores with the residual examination marks, having removed the effects of intelligence from the two factors. The correlation of these scores yields an r of 0.665. This is called a partial correlation as it is the correlation of study time and examination performance having partialled out the effect of intelligence. In this case the size of the correlation has been reduced but it is still significant (at $p = 0.05$), so the original correlation was not entirely due to the third variable, intelligence. There is still a significant relationship between the amount of time spent studying and performance in the examination *after* we have accounted for the effects of intelligence.

We can illustrate what we have done by representing the variability of the scores of each variable by a circle. As we can see from Figure 21.1 the three circles overlap. The area $SE + SIE$ is the portion of the examination performance variability explained by study time, the area $SI + SIE$ the portion of study time explained by intelligence and $IE + SIE$ the portion of examination performance explained by intelligence. The size of these areas can be found by calculating r^2 for each correlation. When we remove the effects of intelligence we take away the intelligence circle ($I + SI + SIE + IE$) leaving $S + SE$ of the study time variability and $E + SE$ of the examination performance variability. The partial correlation of study time and examination performance, having removed the effect of intelligence, leaves us with the area SE as the residual variability of examination performance explained by the residual variability of study time.

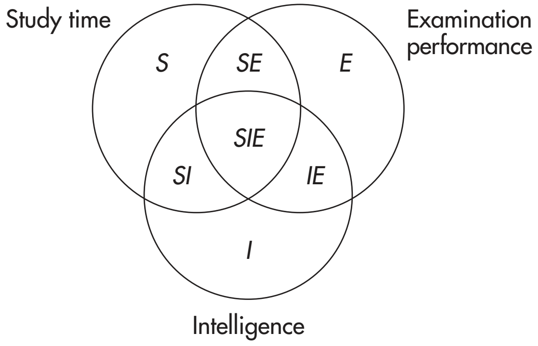

FIGURE 21.1 The variability of the scores on three variables

Fortunately, there is an easier method of calculating a partial correlation, than finding the residuals, when we know the three separate correlation coefficients. We label the variables as *1*, *2* and *3* (rather than *X* and *Y*) as it makes it easier to label additional variables. I will label examination performance as variable *1*, study time as variable *2* and intelligence as variable *3*. The correlation coefficients are labelled as r_{12} for the correlation of variables *1* and *2*, r_{13} for the correlation of variables *1* and *3* and r_{23} for the correlation of variables *2* and *3*. The partial correlation of variables *1* and *2* having removed the effects of variable *3* is termed $r_{12.3}$ and can be calculated with the following relatively simple formula.

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1 - r_{13}^2}\sqrt{1 - r_{23}^2}}$$

For our example,

$$r_{12.3} = \frac{0.72 - (0.48 \times 0.37)}{\sqrt{1 - 0.48^2}\sqrt{1 - 0.37^2}} = 0.665$$

We are not restricted to finding just the one partial correlation. We can also find $r_{13.2}$ (the correlation between examination performance and intelligence having partialled out the effect of study time) and $r_{23.1}$ (the correlation of study time and intelligence having partialled out the effect of examination performance) by using the same formula with the correlation coefficients adjusted appropriately, so for $r_{13.2}$ we would replace r_{12} with r_{13} and so on. Notice that some of these are more meaningful to work out than others. Just because the statistical reasoning provides us with the possibility of an analysis it does not mean that we decide it is worthwhile undertaking.

We can extend the analysis to partial out the effects of more than one variable from a correlation. We can remove the effects of variable *4* if we wish by the following formula:

$$r_{12.34} = \frac{r_{12.3} - r_{14.3}r_{24.3}}{\sqrt{1 - r_{14.3}^2}\sqrt{1 - r_{24.3}^2}}$$

Notice that the formula contains the partial correlations of the variables having removed variable *3*. The logic allows us to go on to remove variables *5*, *6*, etc. However, the formulae make one key assumption, that is, the variables are linearly correlated with variables *1* and *2*. We are extending the linear model to all the variables. If this assumption is not valid we will only partial out the linear components of the variables, not all their effects.

Multiple correlation

We can use partial correlations to help us calculate a multiple correlation. A multiple correlation coefficient, R, gives us a measure of how well three or more variables correlate together. We do some relabelling again here. We specify a particular variable to label as Y. This is the dependent variable and we are calculating how it correlates with the rest. It is usually the variable we wish to predict (as we shall see in multiple regression later). I shall choose *examination performance* as this is an interesting one to predict. We call the other variables *1*, *2*, *3*, etc. We have only two others so I shall call *study time* variable *1* and *intelligence* variable *2*.

R is easier to explain if we work with R^2, the coefficient of determination for the multiple correlation. We take each of the variables *1*, *2*, *3*, etc. in turn and find out what proportion of the Y variability it can explain that has not already been explained by previous variables. Adding up these portions gives us a measure of how much of the Y variability can be explained by the combination of the other variables.

The first question we must ask is how much of the variability of the Y scores (examination performance) can be explained by variable *1*, study time? This is simply the coefficient of determination of the correlation of the two variables, r_{Y1}^2. Now we ask how much of the remaining variability of Y can be explained by variable *2*, intelligence? It is not r_{Y2}^2 as some of this area has already been explained. If we look back to the interlocking circles in Figure 21.1 we see that r^2_{Y2} is the areas *SIE* and *IE*. Yet we have already explained the areas *SE* and *SIE* by r^2_{Y1}. We have already predicted the area *SIE* so we do not want to do it twice. Because intelligence and study time are correlated they both explain some of the *same* variability of examination performance (the area *SIE*). To overcome this we remove the effect of study time (variable *1*) before finding out what of the remaining variability in the examination performance can be explained by intelligence. The residual portion of examination performance after removing the effects of study time is $1 - r^2_{Y1}$ (that is, the whole area, *I*, minus that portion explained by study time, leaving $E + IE$). The amount of the area $1 - r^2_{Y1}$ explained by intelligence is the partial correlation of examination performance and intelligence having removed the effect of study time. This is $r^2_{Y2.1}$ (the area *IE*). Expressed as a portion of the residual *Y* variability this amount (*IE* as a portion of $E + IE$) is $r^2_{Y2.1}(1 - r^2_{Y1})$. In conclusion we can say that the amount of *Y* variability explained by variables *1* and *2* is:

$$R^2_{Y.12} = r^2_{Y1} + r^2_{Y2.1}(1 - r^2_{Y1})$$

(In terms of part of the examination performance circle in Figure 21.1, this is $SE + SIE$ for variable *1* plus *IE* for variable *2*.)

The multiple correlation coefficient, $R_{Y.12}$, is simply the square root of this figure. In our example, $r_{Y1} = 0.72$ and $r_{Y2.1} = 0.33$, so $R^2_{Y.12} = 0.72^2 + 0.33^2(1 - 0.72^2) = 0.57$, and the coefficient of multiple correlation, $R_{Y.12}$, is $\sqrt{0.57} = 0.75$. This tells us that more of the variability in *Y* (examination performance) can be explained by study time and intelligence ($R^2_{Y.12} = 0.57$) than by study time alone ($r^2_{Y1} = 0.52$), although not a lot more.

We can calculate a multiple correlation coefficient for any number of variables, with each new variable used to explain variability in *Y* unexplained by any previous variable. For four variables, $R^2_{Y.123} = R^2_{Y.12} + r^2_{Y3.12}(1 - R^2_{Y.12})$ where the *R*s in the formula are themselves multiple correlation coefficients. The problem is that as each additional variable is brought in, we chip away at the variability of *Y* so that *R* becomes larger. Yet as each new variable is added we increase the risk of increasing *R* by random variation rather than by genuine relationships. Therefore multiple correlations should be undertaken with caution and when a large number of variables are used as ‘predictor’ variables then a correction should be made to *R* to compensate for the increased risk of error. (Statistical computer programs such as SPSS provide an ‘Adjusted *R*’ value to correct for this – see Hinton *et al*., 2004.)

The significance of R^2

We can test the significance of a multiple correlation by using a variance ratio (F) test, comparing the estimated variance of the 'explained variability' to the estimated variance of the 'unexplained variability':

$$\frac{\dfrac{R^2}{k}}{\dfrac{1 - R^2}{N - k - 1}}$$

where N is the number of subjects and k is the number of predictor variables. Thus,

$$F = \frac{R^2(N - k - 1)}{k(1 - R^2)} \text{ with degrees of freedom } k, N - k - 1$$

In our example, with $R^2 = 0.57, N = 10, k = 2, F(2,7) = \dfrac{0.57(10 - 2 - 1)}{2(1 - 0.57)}$ $= 4.64$. From the tables of the F distribution, Table A.3 in the Appendix, $F(2,7) = 4.74$, $p = 0.05$, so the multiple correlation is not significant at $p = 0.05$. Note that if we had had the same value of R^2 but just one more participant ($N = 11$) the result would have been significant. This shows the importance of sample size when dealing with correlations.

Multiple regression

We can calculate a linear regression for more than two variables. Again we need to label one of the variables as Y because this will be the dependent variable. The other variables, the independent variables or predictor variables, will be used to predict it. Instead of having a single variable X for the linear regression we use a number of variables $X_1, X_2, \ldots, X_k$ for the regression, where k is the number of predictor variables. To work out the regression line we calculate the following linear equation:

$$Y = a + b_1X_1 + b_2X_2 + \ldots + b_kX_k$$

I shall only consider the case of two predictor variables here, the simplest case, to illustrate multiple regression. The logic is the same for more predictor variables but the calculation becomes rather complex and will not be explained in this book.

With two predictor variables we wish to solve the equation:

$$Y = a + b_1X_1 + b_2X_2$$

Recall that with just one predictor variable, $Y = a + bX$, where $b = \left(\frac{s_Y}{s_X}\right) r_{YX}$ where s_Y and s_X are the standard deviations of the scores of the two variables.[20] In the two variable case we *cannot* work out b_1 and b_2 by using $b_1 = \left(\frac{s_Y}{s_1}\right) r_{Y1}$ and $b_2 = \left(\frac{s_Y}{s_2}\right) r_{Y2}$ unless X_1 and X_2 are *not* correlated (where s_Y, s_1 and s_2 are the standard deviations of the three variables). The problem is that, as in multiple correlation, we will have some overlap in the variability of Y that the two predictor variables can explain. If we are not careful we will count this variability twice, once with X_1 and once with X_2 and our prediction will be distorted. The way to solve this problem is for the *b*s to be partial regression coefficients, that is, coefficients where the effect of one variable is partialled out when working out the b for the other. In the two predictor case:

$$b_1 = \beta_1\left(\frac{s_Y}{s_1}\right) \quad \text{and} \quad b_2 = \beta_2\left(\frac{s_Y}{s_2}\right)$$

where β_1 and β_2 are the standard partial regression coefficients:

$$\beta_1 = \frac{r_{Y1} - r_{Y2}r_{12}}{1 - r_{12}^2} \quad \text{and} \quad \beta_2 = \frac{r_{Y2} - r_{Y1}r_{12}}{1 - r_{12}^2}$$

Just as r is the slope of the line when we convert X and Y to z scores in the two variable case, β_1 and β_2 are the partial slopes of the regression of Y by the predictor variables when all the scores are converted to z scores.

To complete the linear regression we use the following formula to find a:

$$a = \overline{Y} - b_1\overline{X}_1 - b_2\overline{X}_2$$

We can illustrate the calculation by predicting examination performance (Y) using study time (X_1) and intelligence (X_2) as predictor variables. We first work out β_1 and β_2:

$$\beta_1 = \frac{0.72 - (0.48 \times 0.37)}{1 - 0.37^2} = 0.63 \qquad \beta_2 = \frac{0.48 - (0.72 \times 0.37)}{1 - 0.37^2} = 0.25$$

Next we work out b_1 and b_2 using the values for the standard deviations of the variables (found from the table on p. 285):

$$b_1 = 0.63\left(\frac{11.80}{11.42}\right) = 0.65 \qquad b_2 = 0.25\left(\frac{11.80}{9.72}\right) = 0.30$$

Finally we calculate a:

$$a = 56 - (0.65 \times 29) - (0.30 \times 122) = 0.55$$

We now have the equation for the multiple regression:[19]

$$Y' = 0.55 + 0.65X_1 + 0.30X_2$$

Replacing the symbols with the variable names gives us the formula for predicting examination performance using study time and intelligence:

Examination mark = 0.55 + 0.65 Study time + 0.30 Intelligence

From this we can predict, for example, a student with an intelligence score of 110 and who studies for 30 hours per week will obtain the following examination mark:

Examination mark = 0.55 + (0.65 × 30) + (0.30 × 110) = 53.05

Thus, on the basis of the linear multiple regression we predict that the student would get an examination mark of 53.05.

Multicollinearity

When our predictor variables are highly correlated with each other we have what is referred to as multicollinearity. This can be a problem for multiple regression. First, the predictors are explaining much the same variability in the dependent variable Y. Consider the case of two predictor variables. When the two variables are not correlated then the Y variability explained by one is different to the Y variability explained by the other but when they are correlated there is an overlap in the Y variability they explain. Second, we do not know which of the predictor variables is the more important due to the common variability explained. With many predictor variables this problem can arise quite easily. A solution to multicollinearity is to combine variables into a single variable or to leave one out if it is essentially predicting the same variability as another. As an example, imagine that you were predicting a person's height from other bodily dimensions, such as foot length, forearm length, index finger length, etc. If you had included the length of the left foot and the length of the right foot as two separate variables then you might find that these two measurements are so highly correlated that you really do not need or want both in your regression due to multicollinearity. You might decide to include only the right foot length or even the average of the two feet lengths for each person.

Calculating multiple regression

In our example we have included all the predictor variables in the regression, not surprisingly since there were only two, and this is called direct regression. When there are more predictor variables, the researcher might start by calculating the multiple regression by working out the equation using the predictor variable that correlated most highly with the dependent variable. Predictor variables are then added into the regression on the basis of the additional variance they can explain. The process is terminated when a variable no longer significantly increases R^2. This is called forward regression. An alternative is to include all the predictor variables initially but to remove variables one at a time, taking out the one that contributes the least to R^2, until removing a variable would significantly reduce R^2. At which point the regression calculation stops. This is called backward regression. Stepwise regression combines the above two methods, adding variables and taking others away at the same time. The reason why we use alternatives to the direct method is that the most predictive regression is where few variables explain lots of the variability in the dependent variable. Not only is it parsimonious, it also means that we are not including a lot of additional variables which contribute little to the prediction.

Details on how to calculate a multiple correlation and multiple regression using the SPSS computer statistical package can be found in Chapter 16 of Hinton *et al.* (2004).