Chapter 11
Related variables

So far in the course, just one attribute (age, temperature, weight, and so on) has normally been measured on a random sample from some population. In this chapter we explore situations where more than one attribute is measured, and interest centres on how the attributes vary together (for example, height and weight). We learn how to quantify any perceived association between variables, and how to test a hypothesis that there is, in fact, no association between them. A new probability model, the bivariate normal distribution, is introduced.

Like the previous chapter, Chapter 11 is concerned with ideas and techniques for data consisting of pairs of variables; that is, with data in the form \((X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)\). Chapter 10 concentrated on regression analysis, and a key idea there was that one of the variables involved was treated as the explanatory variable, and the other as the response variable. In this chapter, the two variables are not distinguished in that way. We shall not be concerned with trying to explain how measurements on the variable \(Y\) change in response to changes in the variable \(X\), but instead we shall treat the two variables on an equal footing.

For instance, Figure 11.1 is a scatter plot of data on the heights (in cm) and weights (in kg) of 30 eleven-year-old girls. Rather than asking questions about how a girl’s weight depends on her height, in this chapter we shall ask how the weights and heights of girls vary together. What do we mean when we say that the two variables, height and weight, are related? How can we measure how closely related they are?

In situations of this sort, it is often useful to treat each pair of observations on the two random variables as one observation on a kind of two-dimensional random variable—a bivariate random variable. The resulting data are often called bivariate data. Section 11.1 describes in more detail what is meant by bivariate data, as well as exploring the idea of what it means for two variables to be related. Section 11.2 develops ways of attaching a numerical measure to the strength of the relationship between two random variables, and for using such a measure in a hypothesis test of whether the variables really are related at all. Sections 11.3 and 11.4 are concerned with data on pairs of variables, each of which is discrete and can take only a small number of values. An example to which we shall return comes from a study of risk factors for heart disease. A number of individuals were given a score on a four-point scale to indicate the amount that they snored at night. The second variable involved took only two values; it was a Yes-No measure of whether each individual had heart disease. The question of interest was whether heart disease and snoring frequency are related. Such data are very often presented in a table called a

Data provided by A.T. Graham, The Open University.

![Figure 11.1 Scatter plot of the heights and weights of 30 eleven-year-old girls in a Bradford school](image-url)
contingency table. Methods for testing for relationships between variables of this sort are rather different from those covered in Section 11.2 for continuous data and other forms of discrete data; one of the tests turns out to be a form of the chi-squared goodness-of-fit test that you met in Chapter 9. Finally, Section 11.5 returns to continuous data and presents a probability model for bivariate data—the bivariate normal distribution.

11.1 Bivariate data

11.1.1 Scatter plots and relationships

Let us begin with some examples of bivariate data.

Example 11.1 Systolic and diastolic blood pressures

The data in Table 11.1 come from a study of the effect of a drug, captopril, on blood pressure in human patients who had moderate essential hypertension (moderately raised blood pressure). The pressure of the blood inside the body varies as the heart beats, and a blood pressure measurement generally produces two values: the systolic pressure, which is the maximum pressure as the heart contracts, and the diastolic pressure, which is the minimum pressure. The data in Table 11.1 are readings taken on the fifteen patients before they were given the drug, captopril.

| Table 11.1 Blood pressure measurements for 15 patients before treatment with captopril (mm Hg) |
|----------------------------------|---------------------------------|-------------------|
| Patient number | Systolic blood pressure (mm Hg) | Diastolic blood pressure (mm Hg) |
| 1               | 210                             | 130                            |
| 2               | 169                             | 122                            |
| 3               | 187                             | 124                            |
| 4               | 160                             | 104                            |
| 5               | 167                             | 112                            |
| 6               | 176                             | 101                            |
| 7               | 185                             | 121                            |
| 8               | 206                             | 124                            |
| 9               | 173                             | 115                            |
| 10              | 146                             | 102                            |
| 11              | 174                             | 98                             |
| 12              | 201                             | 119                            |
| 13              | 198                             | 106                            |
| 14              | 148                             | 107                            |
| 15              | 154                             | 100                            |

In Figure 11.2 the data are plotted on a scatter plot. The question of interest is: how do these two measurements vary together in patients with this condition? From the scatter plot, it appears that there is something of a tendency for patients who have high systolic blood pressures to have high diastolic pressures as well. The pattern of points on the scatter-plot slopes upwards from left to right, as it did in several of the scatter plots you met in Chapter 10. In fact, it would be possible to analyse the data using the regression methods discussed in Chapter 10. But there is a problem here.


Figure 11.2 Systolic and diastolic blood pressures for 15 patients
Which of the two variables would you choose to be the explanatory variable, and which the response variable? There seems to be no clear answer. We are neither investigating how diastolic blood pressure changes in response to changes in systolic blood pressure, nor the other way round. In Figure 11.2, systolic pressure is plotted on the horizontal axis, but in the context of this investigation this was an arbitrary choice. It would have been perfectly feasible to plot the data the other way round. We cannot use the regression methods of Chapter 10, because those methods do not treat the two variables on an equal footing.

The data in Example 11.1 provide evidence that, for people with moderate essential hypertension at least, people with relatively high systolic blood pressure tend to have relatively high diastolic pressure as well. People with low systolic pressure tend to have low diastolic pressure. And, importantly, these statements work the other way round too. People with high diastolic pressure tend to have high systolic pressure. Another way to think of this is as follows. Suppose you choose at random a patient with moderate essential hypertension. On the basis of the data in Table 11.1, you could say something about what his or her diastolic blood pressure might be. Without performing any calculations, you would probably find it surprising if their diastolic blood pressure fell a long way outside the range from about 95 to about 130 mm Hg.

However, suppose you were now told that this person's systolic blood pressure was 200 mm Hg. On the range of values of systolic pressure represented in Figure 11.2, this is a high value. One might expect, then, that this person's diastolic pressure would be relatively high too. You might think it quite unlikely, for instance, to find that their diastolic pressure was as low as 100. In other words, knowing their systolic pressure has provided information about their diastolic pressure. Similarly, knowing their diastolic pressure would tell you something about their systolic pressure, if you did not know it already.

In intuitive terms, this is what it means for two random variables to be \textit{related}. Knowing the value of one of the variables tells you something about the value of the other variable.

When two variables are related, it is often possible to describe in simple terms the manner of the relationship. In Example 11.1, one might say that the variables are \textit{positively} related, because the two variables tend to be both high at the same time or both low at the same time. The pattern of points on the scatter plot slopes upwards from left to right.

\textbf{Example 11.2 Socio-economic data on US states}

Figure 11.3 shows a scatter plot of data on 47 states of the USA, taken from a study of crime rates and their determinants. The variable on the horizontal axis is a measure of the educational level of residents of each state: it is the mean number of years' schooling of the population aged 25 and over. The variable on the vertical axis is a measure of the inequality of income distribution in the state: it is the percentage of families who earned below one-half of the overall median income.

Again, the aim here is not to describe or investigate how one of the variables changes in response to changes in the other. It is to describe how the variables vary together, or in other words how they are related. The variables clearly

\textit{We are not doing regression analysis here; but if we were, the regression line would have a positive slope.}

Elements of Statistics

Figure 11.3 Educational level and income inequality are related. States with relatively high average educational level tend to have relatively low income inequality, on the measure used here; and states with low educational level have high income inequality. Therefore, knowing something about the educational level of a state tells you something about its income inequality, and vice versa. The relationship is different from that in Example 11.1 though. In Example 11.1, a high value of one variable was associated with a high value of the other and a low value of one variable was associated with a low value of the other. In this example the association works the other way round: low values of one variable go with high values of the other. The pattern of points on the scatter plot slopes downwards from left to right. Figure 11.3 shows a negative relationship between the variables.

Variables can be related in other ways. Figure 11.4 is a scatter plot of two economic variables, the percentage of the UK workforce that is unemployed and the percentage change in wage rates, for the years 1861 to 1913.

Here, the two variables are negatively related because there is a tendency for high values of one of them to go with low values of the other. The pattern of points on the scatter plot slopes downwards from left to right. The difference between this scatter plot and Figure 11.3 is that, in Figure 11.3, the points slope downwards in a more or less linear (straight-line) way, but in Figure 11.4 the points show a clearly curved pattern. However, they are still negatively related.

Figure 11.5 shows a rather different pattern.

Figure 11.5  Population density and fire service expenditure

Here, each data point corresponds to a non-metropolitan county of England. The two variables are the density of population (hectares per person) and the expenditure (in £) per head of population on the fire service. The scatter plot shows a fairly clear pattern, but it suggests a curve. Counties with low population density and high population density both spend relatively large amounts on the fire service, while points with medium population density spend relatively small amounts. Therefore, knowing the population density of a county tells you something about its expenditure on the fire service and vice versa. The two variables are related, but in this case it is not very appropriate to describe the relationship either as positive or as negative.

Exercise 11.1

For each of the scatter plots in Figure 11.6, state whether the variables involved are related, and if they are, say whether the relationship is positive, negative or neither of these.

An important feature of all the examples we have looked at so far is that in each case both the variables involved are, or can be thought of as, random variables. Bivariate data are data giving values of pairs of random
variables—hence the name. There is a contrast here with the data used in regression. In many regression situations, the explanatory variable is not random, though the response variable is. You may recall the data in Section 10.1 on duckweed. There, the explanatory variable is the week number and there is nothing random about that. The response, the number of duckweed fronds, is a random variable. Since only one of the variables is random, it would be inappropriate to ask the question: how do week number and number of duckweed fronds vary together? In this section and Section 11.2, all the data that we shall look at are bivariate data.

It is possible to be more formal (though for much of this chapter we shall not need to be) about the idea of a relationship in bivariate data. Denote by $X$ and $Y$ the two random variables involved. In Example 11.1, $X$ could be a randomly chosen patient’s systolic blood pressure and $Y$ would be his or her diastolic blood pressure. Then it would be possible to describe what is known about the value of a randomly chosen patient’s diastolic blood pressure by giving the probability distribution of $Y$. If you knew this, you could find, for instance, $P(Y = 110)$, the probability that a randomly chosen patient has a diastolic blood pressure of 110 mm Hg. Now, suppose you find out that this patient has a systolic blood pressure of 200 mm Hg. Because the two variables $X$ and $Y$ are related, $P(Y = 110)$ will no longer give an appropriate value for the probability that the patient’s diastolic pressure is 110 mm Hg. We shall denote the probability that a patient’s diastolic pressure $Y$ is 110 when we know that the patient’s systolic pressure $X$ is 200 by

$$P(Y = 110 | X = 200),$$

which is read as ‘the probability that $Y = 110$ given that $X = 200$', or ‘the probability that $Y = 110$ conditional on $X = 200'$. An expression such as $P(Y = 110 | X = 200)$ is called a conditional probability. (By contrast, the ordinary sort of probability without a $\mid$ sign in it is sometimes termed an unconditional probability.) Since $X$ and $Y$ are related, it is the case that the conditional probability that $Y = 110$ given $X = 200$ is different from the unconditional probability that $Y = 110$; that is,

$$P(Y = 110 | X = 200) \neq P(Y = 110).$$

In general, if $X$ and $Y$ are any two random variables, we define

$$P(Y = y | X = x)$$

to be the probability that the random variable $Y$ takes the value $y$ when it is known that the random variable $X$ takes the value $x$.

The random variables $X$ and $Y$ are not related if knowing the value of $X$ tells you nothing about the value of $Y$. That is, $X$ and $Y$ are not related if, for all values of $x$ and $y$,

$$P(Y = y | X = x) = P(Y = y).$$

The random variables $X$ and $Y$ are related if, for at least some of the possible values of $x$ and $y$,

$$P(Y = y | X = x) \neq P(Y = y),$$

(11.1)

that is, if knowing the value of $X$ can tell you something about the distribution of $Y$ that you did not already know.

Depending on the question of interest, it may well be appropriate to use regression methods on bivariate data as well as on data where one of the variables is not random.

See Example 10.8.

A convenient model for variation in blood pressure would be continuous, in which case the probability $P(Y = 110)$ might strictly be written $P(109.5 \leq Y < 110.5)$. But in this case a strict insistence on notation would obscure the simple nature of the message.
Exercise 11.2

(a) Suppose that \( X \) and \( Y \) are two random variables which take values on the integers 0, 1, 2, ..., and so on. It is known that \( P(Y = 10) = 0.3 \), and \( P(Y = 10|X = 4) = 0.4 \). Are the random variables \( X \) and \( Y \) related?

(b) Suppose that \( W \) and \( Z \) are two more random variables taking values on the integers 0, 1, 2, .... It is known that \( P(Z = 5) = 0.4 \), and also that \( P(Z = 5|W = 4) = 0.4 \). Are the random variables \( W \) and \( Z \) related?

This probability definition of what it means for two random variables to be related may seem a little unsatisfactory to you for two reasons. First, nothing has been said about how the conditional probabilities that have been defined can be estimated from data. If we do not know what the value of the probability \( P(Y = 110|X = 200) \) actually is, how do we decide whether or not it is different from \( P(Y = 110) \), in order to conclude whether \( Y \) and \( X \) are related? In order to estimate these probabilities, usually we must use a probability model for the random variables involved. One way to model bivariate data is discussed in Section 11.5 and further discussion will be deferred until then. Furthermore, methods of testing whether two random variables are related generally do not involve calculating probabilities of this sort directly.

Second, it has been emphasized that the idea of related variables involved treating the two variables on an equal footing; but here we looked at the conditional probability \( P(Y = y|X = x) \), and this expression treats \( X \) and \( Y \) differently. In fact, we could just as well have defined \( X \) and \( Y \) to be related if

\[
P(X = x|Y = y) \neq P(X = x),
\]

for some values of \( x \) and \( y \). It can be shown that the definitions given by (11.1) and (11.2) always agree.

You might suspect that this idea of a relationship between two random variables has something to do with the idea of independence of random variables that you met first in Chapter 3. You would be right. It can be shown that two random variables are related in the sense we have just discussed if they are not independent in the sense defined in Chapter 3. If two variables are independent, they are not related.

Section 11.2 of this chapter is concerned with ways of measuring how strong a relationship is in data of the sort we have been looking at. But Sections 11.3 and 11.4 are concerned with relationships between variables of a rather different sort, and we now turn briefly to introduce this kind of data.

11.1.2 Relationships in discrete data

Again let us start with an example.

Example 11.3 Snoring frequency and heart disease

The data in Table 11.2 come from a study which investigated whether snoring was related to various diseases. A large number of individuals were surveyed and classified according to the amount they snored, on the basis of reports from their spouses. A four-fold classification of the amount they snored was used. In addition, the researchers recorded whether or not each person had

certain diseases. These particular data relate to the presence or absence of heart disease. The table gives counts of the number of people who fell into various categories. The top left-hand number shows, for instance, that 24 out of the 2484 people involved were non-snorers who had heart disease. The numbers in the right-hand column are the row totals and show, for instance, that 2374 of the 2484 people involved did not have heart disease. The bottom row gives the column totals: for instance, 213 people snored nearly every night.

<table>
<thead>
<tr>
<th></th>
<th>Heart disease</th>
<th>Non-snorers</th>
<th>Occasional snorers</th>
<th>Snore nearly every night</th>
<th>Snore every night</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>24</td>
<td>35</td>
<td>21</td>
<td>30</td>
<td>110</td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>1355</td>
<td>603</td>
<td>192</td>
<td>224</td>
<td>2374</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>1379</td>
<td>638</td>
<td>213</td>
<td>254</td>
<td>2484</td>
<td></td>
</tr>
</tbody>
</table>

Though these data look very different in form from those we have looked at so far, they are similar in several respects. Each individual surveyed provided an observation on two discrete random variables. The first, $X$, can take two values: Yes or No, depending on whether the individual has heart disease. The second, $Y$, can take four values: Non-snarer, Occasional snorer, Snore nearly every night and Snore every night, depending on how often he or she snores. Thus the data set consists of 2484 values of the pair of random variables $X$ and $Y$. In 24 of the pairs, $X$ takes the value Yes and $Y$ takes the value Non-snarer. In 192 of them, $X$ takes the value No and $Y$ takes the value Snore nearly every night. Thus these are bivariate data. We can therefore ask the question: are $X$ and $Y$ independent? Formal methods for answering this question are developed in Section 11.4; but Exercise 11.3 will give an informal answer.

**Exercise 11.3**

(a) On the basis of the data in Table 11.2, what would you estimate to be the (unconditional) probability that the random variable $X$ takes the value Yes?

(b) What would you estimate to be the conditional probability that $X$ takes the value Yes given that $Y$ takes the value Snore every night?

**Hint** This is a probability conditional on $Y$ taking the value Snore every night, so it is only the people who snore every night who provide direct information about it.

Do you think $X$ and $Y$ are related?

This does not entirely answer the question of whether $X$ and $Y$ are related: we have not taken into account the possibility that the result is a 'fluke' resulting from sampling variability. But this example shows that the notions of bivariate data, and of related random variables, crop up in discrete data of this kind as well as in data of the sort that can be plotted in scatter plots.

Up until now in the course, it has always been insisted that random variables should be real-valued (that is, their values should be numbers, and not words like 'never', 'occasional', 'often', 'always'). In the case of Bernoulli trials, we have been careful to identify with one outcome the number 1, and with the other the number 0. This identification is essential when calculating means and variances, and the strict definition of a random variable requires that it should be real-valued. Here, however, it would be merely cumbersome to attach numbers (0, 1, 2, 3, say) to outcomes and since no modelling is taking place here, let us not bother to do so. Data of this kind are often called **categorical data**.

We shall now look at formal ways of measuring the strength of a relationship between variables.
11.2 Measures of association

This section will develop ways of measuring the strength of a relationship between two random variables, or the strength of the association or correlation between them as it is sometimes termed. The methods apply to continuous bivariate data, and can also be applied to discrete bivariate data of certain kinds, as long as it makes sense to plot the data on a scatter plot.

11.2.1 The Pearson correlation coefficient

In Section 11.1, you saw that bivariate data on two random variables might indicate that the two variables are related. Variables can be related positively or negatively (or in some other way); and in some cases the relationship can be reasonably represented by a straight line, whereas in others it cannot. There are other aspects to relationships between variables. Compare the three scatter plots in Figure 11.7, which again give data for 47 US states in 1959–60.

![Scatter plots](image)

**Figure 11.7** (a) Police expenditure, different years  (b) Police expenditure and community wealth  (c) Police expenditure and labour force participation

In each case, the two variables involved are positively related; knowing the value of one of them tells you something about the value of the other. But in Figure 11.7(a), knowing one of the police expenditure figures would tell you very accurately what the other would be. The points are not scattered very far from a straight line. By contrast, in Figure 11.7(c), knowing the police expenditure tells you very little about the labour force participation rate. The data are very scattered. Therefore, in Figure 11.7(a) we say that the two variables are strongly associated; in Figure 11.7(c) they are weakly associated; and Figure 11.7(b) comes somewhere between in terms of strength of association.

It is useful to have a summary measure of the strength of association between two random variables. Several such measures exist; one of the oldest, but still the most commonly used, is the **Pearson correlation coefficient**. As you will see later in this section, there are other correlation coefficients, but the Pearson coefficient is the most used, and it is what statisticians usually mean when they refer to a correlation coefficient without saying which one. This measure was developed by Sir Francis Galton (1822–1911), about whom Vandaele, W. (1978) Participation in illegitimate activities: Erlich revisited. In Blumstein, A., Cohen, J. and Nagin, D., (eds) Deterrence and incapacitation: estimating the effects of criminal sanctions on crime rates. National Academy of Sciences, Washington, DC, pp. 270–335.

This is also known, sometimes, as the Pearson product-moment correlation coefficient. Often the word 'coefficient' is omitted, and we speak simply of the 'Pearson correlation'.
you read in Chapter 10, and (principally) Karl Pearson (1857–1937), who held the first post of Professor of Statistics in Britain (at University College London). Karl Pearson was the father of Egon Pearson, whom you read about in connection with hypothesis testing in Chapter 8.

The Pearson correlation coefficient is a quantity which we shall denote by $r$. It takes values between $-1$ and $+1$. The sign of $r$ indicates whether the relationship between the two variables involved is positive or negative. The absolute value of $r$, ignoring the sign, gives a measure of the strength of association between the variables. The further $r$ is from zero, the stronger the relationship. The Pearson correlation coefficient takes the value $+1$ only if the plotted bivariate data show an exact straight-line relationship with a positive slope, and $-1$ if the data show an exact straight line with negative slope. Data where the two variables are unrelated have a correlation coefficient of 0. Some examples of scatter plots are given in Figure 11.8 (the data are artificial for the purposes of illustration). The three data sets depicted in Figure 11.7, for example, have the following Pearson correlation coefficients: (a) $r = 0.994$, (b) $r = 0.787$ and (c) $r = 0.121$. All are positive, reflecting the fact that in each case the two variables involved are positively related; and the stronger the relationship, the larger the value of the correlation coefficient. You should note that higher values of $r$ do not imply anything at all about the slope of the straight-line fit: they say something about the quality of the fit.

**Exercise 11.4**

Based on what you have seen about values of $r$ in different data contexts, what do you guess the value of the Pearson correlation coefficient might be for the data in Figure 11.3?

---

### The Pearson product-moment correlation coefficient

The formula for calculating the Pearson correlation coefficient $r$ from bivariate data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, where the means of the $x$-values and the $y$-values are $\bar{x}$ and $\bar{y}$ and their standard deviations are $s_x$ and $s_y$, is as follows.

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right)$$

$$= \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

$$= \frac{s_x s_y}{\sqrt{(n \sum x_i^2 - (\sum x_i)^2)(n \sum y_i^2 - (\sum y_i)^2)}}$$

Of the three equivalent versions, (11.5) is the most convenient, and least prone to rounding error, if you need to calculate $r$ using a calculator. The first

**Figure 11.8**

(a) $r = 1$

(b) $r = -1$

(c) $r = 0$
formulation (11.3) gives the clearest idea of how the definition actually works. Consider the expression \((x_i - \bar{x})/s_X\). It is positive for values of \(x_i\) above their mean, and negative for values below the mean. Suppose the random variables \(X\) and \(Y\) are positively related. Then they both tend to be relatively large at the same time, and relatively small at the same time. Thus \(x_i\) and \(y_i\) are likely to be both above their mean or both below their mean for any \(i\). If both are above their mean, then the two terms in brackets in (11.3) will be positive for that value of \(i\), and their product will be positive, so that this data point will contribute a positive value to the sum in (11.3). If both \(x_i\) and \(y_i\) are below their means, the two terms in brackets in (11.3) will be negative, so again their product is positive, and the data point will again contribute a positive value to the sum. Since \(X\) and \(Y\) are positively related, there will be fewer data points where one of the variables is below its mean while the other is above. These points contribute a negative value to the sum in (11.3). Therefore, if \(X\) and \(Y\) are positively related, the sum in (11.3) will be positive and thus \(r\) will be positive. If \(X\) and \(Y\) are negatively related, then negative terms will dominate in the sum in (11.3) and \(r\) will turn out to be negative. The \(n - 1\) divisor in (11.3) is there for much the same reasons as the \(n - 1\) divisor in the definition of the sample variance. Expressions like \((x_i - \bar{x})/s_X\) in (11.3) include the sample standard deviation because \(r\) is intended to measure the strength of association without regard to the scales of measurement of the two variables. Thus the value of the correlation coefficient for height and weight of adults, for example, should take the same value whether weight is measured in grams or kilograms or pounds. Changing the weights from kilograms to grams will multiply both the numerator and the denominator of \((x_i - \bar{x})/s_X\) by 1000, so the overall value of this expression will not change, and hence the value of \(r\) will not change.

Generally, one would use computer software to calculate correlation coefficients. However, to examine how the formulas work, let us calculate a couple of examples by hand. The easiest version of the formula for \(r\) to use for hand calculation is that given in (11.5).

**Example 11.1 continued**

For the data on blood pressure in Table 11.1, the necessary summary calculations are as follows (denoting systolic pressure by \(X\) and diastolic pressure by \(Y\)).

\[
\begin{align*}
n &= 15 \\
\sum x_i &= 210 + 169 + \cdots + 154 = 2654 \\
\sum x_i^2 &= 210^2 + 169^2 + \cdots + 154^2 = 475502 \\
\sum y_i &= 130 + 122 + \cdots + 100 = 1685 \\
\sum y_i^2 &= 130^2 + 122^2 + \cdots + 100^2 = 190817 \\
\sum x_iy_i &= (210 \times 130) + (169 \times 122) + \cdots + (154 \times 100) \\
&= 27300 + 20618 + \cdots + 15400 \\
&= 300137
\end{align*}
\]
Then, using (11.5),

\[
r = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{\sqrt{(n \sum x_i^2 - (\sum x_i)^2)(n \sum y_i^2 - (\sum y_i)^2)}}
\]

\[
= \frac{15 \times 300.137 - 2654 \times 1685}{\sqrt{(15 \times 475.502 - 2654^2)(15 \times 190.817 - 1685^2)}}
\]

\[
= \frac{30.065}{\sqrt{88.814 \times 23.030}}
\]

\[
= 0.6648.
\]

This result matches what was said in Section 11.1: these two variables are positively related. The value of the correlation coefficient is not particularly close to either 0 or 1, implying that the strength of association between these two variables is moderate. This matches the impression given by the scatter plot in Figure 11.2, where there is a moderate degree of scatter. 

**Exercise 11.5**

The data in Table 11.3 were obtained in a study of a new method of measuring body composition. They give the age and body fat percentage for 14 women.

<table>
<thead>
<tr>
<th>Table 11.3</th>
<th>Body fat percentage and age for 14 women</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (years)</td>
<td>Body fat (%)</td>
</tr>
<tr>
<td>23</td>
<td>27.9</td>
</tr>
<tr>
<td>39</td>
<td>31.4</td>
</tr>
<tr>
<td>41</td>
<td>25.9</td>
</tr>
<tr>
<td>49</td>
<td>25.2</td>
</tr>
<tr>
<td>50</td>
<td>31.1</td>
</tr>
<tr>
<td>53</td>
<td>34.7</td>
</tr>
<tr>
<td>53</td>
<td>42.0</td>
</tr>
<tr>
<td>54</td>
<td>29.1</td>
</tr>
<tr>
<td>56</td>
<td>32.5</td>
</tr>
<tr>
<td>57</td>
<td>30.3</td>
</tr>
<tr>
<td>58</td>
<td>33.0</td>
</tr>
<tr>
<td>58</td>
<td>33.8</td>
</tr>
<tr>
<td>60</td>
<td>41.1</td>
</tr>
<tr>
<td>61</td>
<td>34.5</td>
</tr>
</tbody>
</table>

Investigate how age and body fat percentage are related by (a) drawing a scatter plot; (b) calculating the Pearson correlation coefficient.

In Exercise 11.6 you should use your computer for the calculations.

**Exercise 11.6**

(a) The data in Table 11.4 are those for the heights and weights of 30 Bradford school children, illustrated in Figure 11.1.

Calculate the Pearson correlation coefficient for these data. Does the value of the coefficient match the impression of the strength of association given by the scatter plot?
Table 11.4  Heights and weights of 30 eleven-year-old schoolgirls from Heaton Middle School, Bradford

<table>
<thead>
<tr>
<th>Height (cm)</th>
<th>Weight (kg)</th>
<th>Height (cm)</th>
<th>Weight (kg)</th>
</tr>
</thead>
<tbody>
<tr>
<td>135</td>
<td>26</td>
<td>133</td>
<td>31</td>
</tr>
<tr>
<td>146</td>
<td>33</td>
<td>149</td>
<td>34</td>
</tr>
<tr>
<td>153</td>
<td>55</td>
<td>141</td>
<td>32</td>
</tr>
<tr>
<td>154</td>
<td>50</td>
<td>164</td>
<td>47</td>
</tr>
<tr>
<td>139</td>
<td>32</td>
<td>146</td>
<td>37</td>
</tr>
<tr>
<td>131</td>
<td>25</td>
<td>149</td>
<td>46</td>
</tr>
<tr>
<td>149</td>
<td>44</td>
<td>147</td>
<td>36</td>
</tr>
<tr>
<td>137</td>
<td>31</td>
<td>152</td>
<td>47</td>
</tr>
<tr>
<td>143</td>
<td>36</td>
<td>140</td>
<td>33</td>
</tr>
<tr>
<td>146</td>
<td>35</td>
<td>143</td>
<td>42</td>
</tr>
<tr>
<td>141</td>
<td>28</td>
<td>148</td>
<td>32</td>
</tr>
<tr>
<td>136</td>
<td>28</td>
<td>149</td>
<td>32</td>
</tr>
<tr>
<td>154</td>
<td>36</td>
<td>141</td>
<td>29</td>
</tr>
<tr>
<td>151</td>
<td>48</td>
<td>137</td>
<td>34</td>
</tr>
<tr>
<td>155</td>
<td>36</td>
<td>137</td>
<td>30</td>
</tr>
</tbody>
</table>

(b) An official investigation into differences in mortality between different occupational groups in England and Wales presented the data given in Table 11.5. They relate to male deaths in 1970–72. For each of 25 'occupational orders' (groups of occupations), the data give the 'smoking ratio', a measure of the number of cigarettes smoked on average by men in that group, and the lung cancer standardized mortality ratio (SMR), a measure of the death rate from lung cancer for men in the group. Both of these ratios are adjusted to allow for differences in the pattern of age of members of the groups, and for both, a value of 100 indicates that smoking or mortality is at the average level for England and Wales in 1970–72.

Table 11.5  Smoking ratio and SMR by occupation order

<table>
<thead>
<tr>
<th>Occupation order</th>
<th>Smoking ratio</th>
<th>Lung cancer SMR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Farmers, foresters, fishermen</td>
<td>77</td>
<td>84</td>
</tr>
<tr>
<td>2 Miners and quarrymen</td>
<td>137</td>
<td>116</td>
</tr>
<tr>
<td>3 Gas, coke and chemical makers</td>
<td>117</td>
<td>123</td>
</tr>
<tr>
<td>4 Glass and ceramics makers</td>
<td>94</td>
<td>128</td>
</tr>
<tr>
<td>5 Furnace, forge, foundry, rolling mill workers</td>
<td>116</td>
<td>155</td>
</tr>
<tr>
<td>6 Electrical and electronic workers</td>
<td>102</td>
<td>101</td>
</tr>
<tr>
<td>7 Engineering and allied trades not included elsewhere</td>
<td>111</td>
<td>118</td>
</tr>
<tr>
<td>8 Woodworkers</td>
<td>93</td>
<td>113</td>
</tr>
<tr>
<td>9 Leather workers</td>
<td>88</td>
<td>104</td>
</tr>
<tr>
<td>10 Textile workers</td>
<td>102</td>
<td>88</td>
</tr>
<tr>
<td>11 Clothing workers</td>
<td>91</td>
<td>104</td>
</tr>
<tr>
<td>12 Food, drink and tobacco workers</td>
<td>104</td>
<td>129</td>
</tr>
<tr>
<td>13 Paper and printing workers</td>
<td>107</td>
<td>86</td>
</tr>
<tr>
<td>14 Makers of other products</td>
<td>112</td>
<td>96</td>
</tr>
<tr>
<td>15 Construction workers</td>
<td>113</td>
<td>144</td>
</tr>
<tr>
<td>16 Painters and decorators</td>
<td>110</td>
<td>139</td>
</tr>
<tr>
<td>17 Drivers of stationary engines, cranes, etc</td>
<td>125</td>
<td>113</td>
</tr>
<tr>
<td>18 Labourers not included elsewhere</td>
<td>133</td>
<td>146</td>
</tr>
<tr>
<td>19 Transport and communications workers</td>
<td>115</td>
<td>128</td>
</tr>
<tr>
<td>20 Warehousemen, storekeepers, packers, bottlers</td>
<td>105</td>
<td>115</td>
</tr>
<tr>
<td>21 Clerical workers</td>
<td>87</td>
<td>79</td>
</tr>
<tr>
<td>22 Sales workers</td>
<td>91</td>
<td>85</td>
</tr>
<tr>
<td>23 Service, sport and recreation workers</td>
<td>100</td>
<td>120</td>
</tr>
<tr>
<td>24 Administrators and managers</td>
<td>76</td>
<td>60</td>
</tr>
<tr>
<td>25 Professional, technical workers, artists</td>
<td>66</td>
<td>51</td>
</tr>
</tbody>
</table>

Draw a scatter plot of these data, calculate the Pearson correlation between the two variables, and comment on the strength of association between the two variables.

The conclusions from Exercise 11.6 can be used to illustrate a very important point about correlation. In both parts, you found a reasonably strong positive association between the variables. In part (b), this means that occupational groups where men smoke a lot also experience, on average, high mortality from lung cancer. An obvious explanation for this is that smoking causes lung cancer. But you should note that the data in Table 11.5 do not prove this causation. They merely show that high values of the two variables tend to go together, without saying anything about why. The data support the hypothesis that smoking causes lung cancer; but, because the analysis treated the variables on an equal footing, they support equally well the hypothesis that lung cancer causes smoking. In part (a) of Exercise 11.6, you found that tall girls tend to be relatively heavy. Again, this correlation does not establish a causal explanation, either that being tall causes girls to be heavy or that being heavy causes girls to be tall. A more reasonable explanation is that as a girl grows, this causes increases in both her weight and her height.

In summary, to say that there is a correlation or association between two variables $X$ and $Y$ is merely to say that the values of the two variables vary 'together' in some way. There can be many different explanations of why they vary together, including the following.

- Changes in $X$ cause changes in $Y$.
- Changes in $Y$ cause changes in $X$.
- Changes in some third variable, $Z$, independently cause changes in $X$ and $Y$.
- The observed relationship between $X$ and $Y$ is just a coincidence, with no causal explanation at all.

Statisticians sometimes quote examples of pairs of variables which are correlated without there being a direct causal explanation for the relationship. For instance, there is a high positive correlation between the level of teachers' pay in the USA and the level of alcoholism; yet the alcoholism is not caused, to any great extent, by teachers who drink. In parts of Europe there is a high positive correlation between the number of nesting storks and the human birth rate; yet storks do not bring babies. Since the existence of a pattern on a scatter plot, or the value of a correlation coefficient, cannot establish which of these explanations is valid, statisticians have a slogan which you should remember.

Correlation is not causation.

Causation is established by other (generally non-statistical) routes; typically the aim is to carry out a study in such a way that the causal explanation in which one is interested, is the only plausible explanation for the results.

One should not forget the possibility of a relationship between SMR and occupation, without regard to smoking habits.


Statisticians can be rather reticent about telling you what causation is!
11.2.2 Care with correlation

In Section 11.1 you saw that there could be many different types of patterns in scatter plots. The Pearson correlation coefficient reduces any scatter plot to a single number. Clearly a lot of information can get lost in this process. Therefore it is hardly ever adequate simply to look at the value of the correlation coefficient in investigating the relationship between two variables. Some of the problems that arise were demonstrated very convincingly by the statistician Frank Anscombe, who invented one of the most famous sets of artificial data in statistics. It consists of four different sets of bivariate data: scatter plots of these data are given in Figure 11.9.

![Figure 11.9 Anscombe's data sets](image)

In each case the Pearson correlation coefficient takes the same value: $r = 0.816$. In Figure 11.9(a) this is probably a reasonable summary of the data; the two variables look reasonably strongly associated. However, the other three data sets tell a different story. In Figure 11.9(b) the variables look extremely strongly related, but it is not a straight-line relationship. For the Pearson coefficient, a ‘perfect’ correlation of 1 (or −1) is only found in a straight-line relationship. Therefore the Pearson correlation coefficient is an inadequate summary of this particular data set. The same is true for Figure 11.9(c), for a different reason. This time the data do lie, exactly, on a straight line—except for one outlying point. This outlier brings the correlation coefficient down; if it were omitted the Pearson correlation would be 1. In Figure 11.9(d) there is again a point that is a long way from all the others, but if this point were omitted, there would effectively be no correlation between the variables at all.

Anscombe, F.J. (1973) Graphs in statistical analysis. *American Statistician*, 27, 17–21. The regression lines for the regression of $Y$ on $X$ are also the same for each of the scatter plots.
The message from Anscombe’s examples is that you should always study the data before trying to interpret the value of a correlation coefficient. In fact, it is almost always best to look at a scatter plot before calculating the correlation coefficient; the scatter plot may indicate that calculating the correlation coefficient is not a sensible thing to do, or that some other analysis should be followed as well.

The Pearson correlation coefficient is most useful when the data form a more or less oval pattern on the scatter plot, as in Figure 11.9(a). It is less appropriate when the data show a curvilinear relationship. When there are points that are a long way from the rest of the data, as in Figures 11.9(c) and 11.9(d), it is often useful to calculate the correlation coefficient after omitting them from the data set.

**Exercise 11.7**

In *Chapter 7*, Table 7.7, you met a data set giving the concentration of the pollutant PCB in parts per million, and the shell thickness in millimetres, of 65 Anacapa pelican eggs.

(a) Produce a scatter plot of these data. Briefly describe the relationship between the variables. Are there any points which seem to be a long way from the general run of the data?

(b) Calculate the Pearson correlation coefficient for these data (using the data for all 65 eggs).

(c) Omit any data points which you think are a long way from the others. Recalculate the Pearson correlation coefficient. How would you describe the relationship between the variables in the light of this investigation?

One possibility for using the Pearson correlation coefficient with data whose scatter plot does not show a straightforward oval pattern is to transform the data in an appropriate manner in order to standardize the pattern. This method can sometimes be used to deal with data showing a curvilinear relationship (though other means of dealing with such data are discussed later in this section). It has other applications too, as Exercise 11.8 will indicate.

**Exercise 11.8**

In *Chapter 1*, Table 1.7, data are listed on the brain weight and body weight of 28 kinds of animal. You saw in *Chapter 1* that there were problems merely in producing a sensible plot of the data unless they were transformed first. The transformation suggested was a logarithmic transformation of both variables. Calculate the Pearson correlation coefficient for the untransformed data and for the transformed data.

The Pearson correlation for the untransformed data in Table 1.7 is so close to zero that it gives the impression that the variables are not related at all. The Pearson correlation for the transformed data (which, you might recall from *Chapter 1*, show a much more evenly-spread oval pattern on their scatter plot—see Figure 1.15) indicates that there is, in fact, a reasonably strong positive association between the variables.
The idea of transforming the data before calculating the Pearson correlation is a useful one, but it does depend on whether or not an appropriate transformation can be found. We now turn to another measure of correlation which does not depend for its usefulness on the scatter plot displaying an oval pattern.

### 11.2.3 The Spearman rank correlation coefficient

There exist several correlation coefficients which can meaningfully be applied in a wider range of situations than can the Pearson coefficient. The first of these to be developed was published in 1904 by the British psychologist Charles Spearman (1863–1945). Spearman was one of many psychologists who, by adapting and extending existing statistical methods to make them more suitable for analysing psychological data, and by developing entirely new statistical approaches, have made important advances in statistical science. Spearman had the powerful but simple idea of replacing the original data by their ranks, and measuring the strength of association of two variables by calculating the Pearson correlation coefficient with the ranks.

Let us see how this is done using a data set we have already met.

**Example 11.4 Body fat percentage and age**

In Exercise 11.5 you calculated the Pearson correlation coefficient for a data set on body fat percentage and age. The first step in calculating the Spearman correlation coefficient for these data is to find the ranks of the data on each of the variables separately. Thus, for instance, the lowest age in the data set is 23 years; this is given rank 1. The lowest body fat percentage is 25.2; this is also given rank 1. The highest age is 61 years, and since there are 14 ages in the data set, this gets rank 14; and so on. Where two data values for one of the variables are tied, they are given averaged ranks (as was done in Chapter 9). The resulting ranks are shown along with the original data in Table 11.6.

The remaining calculations for the correlation use just the ranks and ignore the original data. It is fairly evident from the table, and even more obvious from a scatter plot of the ranks (see Figure 11.10) that the ranks of these two variables are positively related. That is to say, low ranks go together, and high ranks go together. This is scarcely surprising; we already saw that the two variables are positively related, which means that low values go together and high values go together.

Calculating the Pearson correlation coefficient of the ranks in the usual way, we obtain the value 0.590. That is, the **Spearman rank correlation coefficient** for these data is 0.590. This value is not very different from the Pearson correlation of 0.507 for this data set, and the scatter plot of the ranks looks similar to the scatter plot of the original data. Thus the Spearman approach has not, in this case, told us anything we did not already know. But that is because these data are suitable for analysis using the Pearson correlation coefficient. The advantage of the Spearman rank correlation coefficient is that it can be used in other situations too.

The Spearman rank correlation coefficient is denoted \( r_s \) (S stands for Spearman). To calculate the Spearman rank correlation coefficient for a set of bivariate data, proceed as follows.

---

Charles Spearman (1863–1945)

Ranks are used in the Wilcoxon signed ranks test and the Mann–Whitney–Wilcoxon test of Chapter 9.
Calculation of the Spearman rank correlation coefficient

(i) Calculate the ranks for each of the variables separately (using averaged ranks for tied values);
(ii) find $r_S$ by calculating the Pearson correlation coefficient for the ranks.

Since the Spearman rank correlation coefficient is calculated using the same formula as that for the Pearson correlation coefficient, it also takes values between $-1$ and $+1$, with values near 0 denoting a low degree of association and values near $-1$ or $+1$ denoting strong association. However, the Spearman rank correlation coefficient uses a more general definition of strong association. Suppose a data set has a Spearman rank correlation coefficient of +1. Then the Pearson correlation coefficient for the ranks is +1, which means that the ranks have an exact linear (straight-line) positive relationship. The only way this can happen is when the ranks for the two variables are exactly the same: that is, the data point which has rank 1 on variable $X$ also has rank 1 on variable $Y$, the data point which has rank 2 on variable $X$ also has rank 2 on variable $Y$, and so on. This means that the original data points come in exactly the same order whether they are sorted in order according to the values of the variable $X$ or the values of the variable $Y$. This can happen if the original variables have an exact positive linear relationship, but it can also happen if they have an exact \textit{curvilinear} positive relationship, as shown in Figure 11.11, as long as the curve involved moves consistently upwards. Such a relationship is known as a \textit{monotonic increasing} relationship.

\begin{figure}[h]
\centering
\begin{tabular}{cc}
\includegraphics[width=0.4\textwidth]{figure1111a}&\includegraphics[width=0.4\textwidth]{figure1111b}
\end{tabular}
\caption{(a) $r_S = 1, r = 1$ (b) $r_S = 1, r \neq 1$}
\end{figure}

Similarly, a data set has a Spearman rank correlation coefficient of $-1$ if the two variables have a \textit{monotonic decreasing} relationship (see Figure 11.12).

\begin{figure}[h]
\centering
\begin{tabular}{cc}
\includegraphics[width=0.4\textwidth]{figure1112a}&\includegraphics[width=0.4\textwidth]{figure1112b}
\end{tabular}
\caption{(a) $r_S = -1, r = -1$ (b) $r_S = -1, r \neq -1$}
\end{figure}

Many textbooks give a simpler formula for $r_S$ which looks somewhat different from the Pearson formula. This formula can, however, be shown to be equivalent to the Pearson formula for data where there are no ties. If there are more than one or two ties, the simplified formula does not hold, anyway. Since you will generally be performing the calculations with your statistical software, and since all statistical packages can calculate the Pearson correlation coefficient, you are spared the details of the special Spearman formula.
Summarizing, the Pearson correlation coefficient is a measure of strength of linear (that is, straight-line) association, while the Spearman rank correlation coefficient is a measure of monotonic (that is, always moving in a consistent direction) association.

Another advantage of the Spearman rank correlation coefficient is that it requires only the ranks of the data: in some cases, only the ranks may be available. In other cases, it might happen that the ranks are more reliable than the original data. When the data on animal body and brain weights were introduced in Chapter 1, various questions were raised about their accuracy. The data involve measurements for extinct animals like dinosaurs, whose body and brain weights must have been inferred from fossils, so that they are unlikely to be accurate. However, it seems more likely that the data for these animals at least come in the correct rank order.

Exercise 11.9

(a) Calculate the value of $r_s$ for the data on body and brain weights. What do you conclude?

(b) What is the value of the Spearman rank correlation coefficient for the log-transformed data? (You should be able to answer this without actually doing any calculations. Think about what the logarithmic transformation does to the ordering of the data.)

We have seen that the Spearman rank correlation coefficient has several advantages over the Pearson correlation coefficient. You might even be wondering why anyone bothers with the Pearson correlation coefficient. In fact, there are several reasons. First, in some circumstances, what is required is specifically a measure of the strength of linear association, and the Spearman rank correlation coefficient cannot give this. Second, there are a number of important more advanced statistical techniques which build on and use the idea of the Pearson correlation coefficient. Third, very often the aim of calculating a correlation coefficient from a sample is to make inferences about the association between the two variables concerned in the population from which the sample was drawn. We shall see some ways of doing this in Subsection 11.2.4. If, in fact, the relationship between the variables is linear, the Pearson correlation often provides more powerful inferences than does the Spearman rank correlation coefficient.

11.2.4 Testing correlations

So far, in this section, we have largely ignored the fact that most of our data arise as samples from some larger population. Usually the question of interest is not what the strength of association between two variables is in the sample, but what the strength of association is in the population. Just as there is a correspondence between the sample mean and the population mean, and between the sample standard deviation and the population standard deviation, so there are population analogues of both the Pearson and the Spearman correlation coefficients, and methods exist for estimating these population analogues (in fact, just as we have been doing) and for testing hypotheses about them.

There are other measures of correlation which are beyond the scope of this course. In particular, the Kendall rank correlation coefficient measures monotonic association in rather the same way as does the Spearman rank correlation coefficient.

The population analogue of the Pearson correlation coefficient is discussed again in Section 11.5.
In this course, we shall confine ourselves to using the sample correlation coefficients for testing the null hypothesis that, in the population, there is no association at all. That is, we shall look at methods for investigating whether or not the two variables involved are really related. Let us begin with an example.

**Example 11.1 continued**

The data on blood pressures can be thought of as a random sample from the population of all potential patients with moderate essential hypertension. Can we be confident that there is really a relationship between systolic and diastolic blood pressure in this population? Let us test the null hypothesis that there is no such relationship, against the two-sided alternative that there is a relationship (positive or negative).

It seems appropriate to use the value of the (sample) correlation coefficient as a test statistic. Intuitively, the further the value of \( r \) is from zero, the more evidence we have that the two variables really are related. Under the null hypothesis of no association, we would expect \( r \) to be fairly close to zero. In fact, for samples of the size we have here (sample size 15) the null distribution of \( R \) is as shown in Figure 11.13. It is symmetrical about \( r = 0 \). Figure 11.13 also shows the observed value, \( r = 0.665 \), and gives the tail area above 0.665. Since this is a two-sided test, we are interested in both tails. From Figure 11.13, it is obvious that our observed value of \( r = 0.665 \) is well out into the upper tail of the distribution. In fact,

\[
SP(\text{obtained direction}) = SP(\text{opposite direction}) = 0.0034,
\]

so there is fairly strong evidence that there really is correlation in the population. ■

This hypothesis test is based on the assumption that the data involved come from a particular distribution called the bivariate normal distribution, which is discussed in Section 11.5. If the null hypothesis of no association is true, this amounts to assuming that the two variables involved have independent normal distributions. More generally, if the data do not show a more or less oval pattern of points on a scatter plot, the test might possibly give misleading conclusions; but you have already seen that using the Pearson correlation coefficient can give strange answers if the scatter plot shows a strange pattern.

It is just as straightforward to test the null hypothesis of no association against a one-sided alternative, that there is a positive relationship between the two variables in the population (or that there is a negative relationship in the population). In terms of the null distribution in Figure 11.13, this amounts to finding the significance probability from one tail of the distribution rather than from both. Explore the facilities available on your computer in attempting Exercise 11.10.

![Figure 11.13](image-url)
Exercise 11.10

Use the data on age and body fat percentage to test the hypothesis that there is no relationship between these two variables in the population, against the one-sided alternative that the two variables have a positive relationship.

In both Chapter 8 and Chapter 9, we have had recourse to computer software when the appropriate significance tests have been too complicated or merely too time-consuming to pursue otherwise. In fact, use of the Pearson correlation coefficient to test the hypothesis of no association in a population is not at all complicated. It can be shown that when there is no association, the sampling distribution of the Pearson correlation coefficient $R$ is most easily given in the form

$$R \sqrt{\frac{n-2}{1-R^2}} \sim t(n-2), \quad (11.6)$$

and (11.6) suggests a convenient test statistic. For the blood pressure data ($n = 15$), we have

$$r \sqrt{\frac{n-2}{1-r^2}} = 0.664773 \sqrt{\frac{15-2}{1-0.664773^2}} = 3.2085,$$

and this is at the 99.66% point of $t(13)$: hence the obtained $SP$ of 0.0034. For the data in Exercise 11.10, your computer should have provided you with the obtained $SP$ of 0.0323. In fact, for these data, $r = 0.506589$ with $n = 14$, so we should test

$$r \sqrt{\frac{n-2}{1-r^2}} = 0.506589 \sqrt{\frac{14-2}{1-0.506589^2}} = 2.0354$$

gainst $t(12)$. This is the 96.77% quantile of $t(12)$, giving the same obtained $SP$ of 0.0323.

Example 11.5 Testing correlations

For the data on 65 Anacapa pelican eggs discussed in Exercise 11.7, the value of the Pearson correlation coefficient $r$ is $-0.253$. We can test the null hypothesis of no association between PCB concentration and shell thickness, against a two-sided alternative, by calculating the test statistic $r \sqrt{\frac{n-2}{1-r^2}}$ and comparing it against a $t$-distribution with $65 - 2 = 63$ degrees of freedom. The value of this quantity is

$$-0.253 \sqrt{\frac{65-2}{1-(-0.253)^2}} = -2.076.$$ 

By looking this value up in statistical tables, or by using a computer, we find that the total $SP$ is 0.042. There is some evidence of a relationship between the two variables, but the evidence is not strong.
The testing procedure for the null hypothesis of no relationship in the population, using the Spearman rank correlation coefficient as the test statistic, works in precisely the same way, except that one looks up the \( SP \) using a different computer command. For the data on body fat percentage, we saw earlier that \( r_s = 0.590 \) (with \( n = 14 \)). In a test of the null hypothesis of no association against the one-sided alternative that there is a positive association, the obtained \( SP \) is 0.0144. This test provides reasonably strong evidence of a positive association between age and body fat percentage in women.

Unlike the null distribution of the Pearson correlation coefficient \( R \) given by (11.6), the null distribution of the Spearman rank correlation coefficient \( R_s \) is, in fact, extremely complicated, though it is achieved by the rather simple idea of assuming that all possible alignments of the two sets of scores are equally likely. The computations are certainly best left to a computer. For large samples, the null distribution of \( R_s \) is given by the approximation

\[
R_s \sqrt{\frac{n-2}{1-R_s^2}} \approx t(n-2).
\]  

(11.7)

Notice that the form of (11.7) is exactly that of (11.6) for the Pearson correlation coefficient.

**Exercise 11.11**

(a) It was stated earlier that the test for no association using the Pearson correlation coefficient was based on the assumption that the data involved followed a bivariate normal distribution; or in other words that they lay in an oval pattern on the scatter plot. When we looked at the Anacapa pelican eggs data earlier, we saw that there was some doubt about whether this was the case, and we reanalysed the data omitting one of the eggs which had a very thin shell. We found that \( r = -0.157 \) (with, now, \( n = 64 \)). Using this data set with one egg omitted, test the hypothesis that there is no association between PCB concentration and thickness of shell against a two-sided alternative. What do you conclude?

(b) For the data set on body and brain weights of animals, we found that \( r_s = 0.716 \) (with \( n = 28 \)). Use the large-sample approximation (even though the sample size is hardly large enough) for the distribution of \( R_s \) to test the hypothesis that there is no association between body weight and brain weight in animals, against the one-sided alternative that the two variables are positively related.

In Section 11.5 we shall return to these sorts of data, where some theoretical issues will be discussed. Meanwhile, in Sections 11.3 and 11.4, we turn to data in the form of contingency tables.
11.3 Exact tests for association in 2 x 2 contingency tables

Our attention now turns to relationships between random variables which take only a small number of discrete values—data of the sort that can be represented in contingency tables of the type you met in Example 11.3. In Section 11.2, we looked at measures of association for 'scatter plot' data, and then we considered methods of testing the null hypothesis of no association, using the measures of association (correlation coefficients) as test statistics. Similar concepts apply to contingency table data.

You saw in Example 11.3 that it makes sense to look for relationships or associations between the variables. There exist many different measures of association for data in contingency tables; most of these measures share some of the properties of correlation coefficients. However, in much analysis of contingency data, the emphasis is on testing for the presence of association rather than on measuring the size of the association, and the common tests of association for contingency table data do not use the common measures of association as test statistics. Therefore, in this section and the next we shall discuss only tests of association.

First, what is a contingency table? It is a table of counts showing the frequencies with which random variables take various values in a sample. The term 'contingency table' is most commonly used when the table records the values of two (or more) variables at the same time, usually (if there are two variables) with one variable corresponding to the rows of a square or rectangular table, and the other variable corresponding to the columns. The 'boxes' in the table, at the intersection of rows and columns, into which counts can fall, are usually called cells. In Table 11.2 the rows corresponded to the presence or absence of heart disease (2 rows) and the columns to the degree of snoring (4 columns). There were $2 \times 4 = 8$ cells in the table, which also gave information on marginal totals.

It is important to remember that contingency tables are tables of counts; if a table consists of percentages or proportions calculated from counts, then strictly speaking it is not a contingency table and it cannot be analysed by the methods covered here (unless it can be reverted to a table of counts).

In Example 11.3, the contingency table had two rows and four columns. The simplest two-way contingency tables have just two rows and two columns, and in this section we shall consider exact tests for these $2 \times 2$ tables. An approximate test that can be applied to larger tables as well as $2 \times 2$ tables is dealt with in Section 11.4.

In a $2 \times 2$ contingency table, the two variables involved can each take only two values. In testing for association, as usual, the aim is to investigate whether knowing the value of one of the variables tells us anything about the value of the other. Let us see how this can be done in an example.

**Example 11.6 Educational level and criminal convictions**

A study was carried out on factors related to a criminal conviction after individuals had been treated for drug abuse. Sixty people who had participated in a drug rehabilitation scheme were categorized according to years of education.
Elements of Statistics

(15 years or less, and more than 15 years) and to whether or not they were convicted for a criminal offence after treatment. The results of the study are given in Table 11.7.

Table 11.7 Educational level and criminal convictions

<table>
<thead>
<tr>
<th>Education</th>
<th>Convicted</th>
<th>No</th>
<th>Row totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>15 years or less</td>
<td>16</td>
<td>20</td>
<td>36</td>
</tr>
<tr>
<td>More than 15 years</td>
<td>6</td>
<td>18</td>
<td>24</td>
</tr>
<tr>
<td>Column totals</td>
<td>22</td>
<td>38</td>
<td>60</td>
</tr>
</tbody>
</table>

It was hypothesized that those with less formal education would be more likely to face a subsequent conviction. On the face of it, the data seem to support the hypothesis. Among those less well educated, 16/36 or about 44% were convicted of a criminal offence, while among those with more education the proportion was only 6/24 or 25%. If these data represent the general state of affairs, then knowing an individual's educational history will tell you how likely it is that this person (having graduated from a drug rehabilitation programme) will face a subsequent criminal conviction. The two random variables, amount of education and conviction, are related (and not, therefore, independent).

However, this table includes only 60 graduates of the scheme, which is a small sample from the potential population. It might well be that this small sample does not truly represent the situation in the population. We could find out by testing the null hypothesis that there is no relationship between the two variables. An appropriate alternative hypothesis is one-sided, because the researchers' theory is that more education reduces an individual's likelihood to reoffend (or, at any rate, their likelihood to be convicted).

You do not have to learn a new test procedure to test these hypotheses. In Subsection 8.5.1 you met Fisher's exact test, which was used to test for differences between two binomial proportions. Here, Fisher's exact test turns out to be the appropriate test too. With the data in Table 11.7, Fisher's exact test gives an obtained $SP$ of 0.104. There is, in fact, only very weak evidence from these data that the two variables really are related in the population.

It is worth exploring briefly one aspect of how this test works. When you carry out Fisher's exact test using your computer, it may well give $SP$ values for both one-sided and two-sided tests; and you may have noticed in your work in Chapter 8 that these values are not related in the simple kind of way that the one- and two-sided $SP$ values are for, say, a $t$-test. The reason is that the distribution of the test statistic is not symmetric. In fact, the test statistic for Fisher's exact test is simply the count in one of the cells of the table, say the count in the upper left cell. The distribution of this count under the null hypothesis is found under the assumption that the marginal totals, the totals of the counts in each row and column given in the margins of the main table in Table 11.7, remain fixed. Under this assumption, considering the data in Table 11.7, the value in the top left-hand cell could be as low as 0, if none of those with more education were convicted. (It could not be larger than 22 since the total of the first column is fixed at 22 and the other count in that column could not be negative.)

Thus the null distribution of the test statistic is concentrated on the values 0, 1, 2, ..., 22, and this distribution is shown in Figure 11.14.

![Figure 11.14](image)

The obtained $SP$ for our observed value in the top left-hand cell, 16, is found by adding together the probabilities for 16, 17, ..., 22: from Figure 11.14 the $SP$ is 0.104. To calculate the total $SP$, we should add to 0.104 the probabilities in the left tail of the null distribution that are less than the probability for 16. Thus we must include the probabilities for 0, 1, ..., 10. Because the distribution in Figure 11.14 is not symmetric, the total $SP$ is equal to 0.174.

In Exercise 11.12, the data are again presented as a $2 \times 2$ contingency table (with marginal totals included).

**Exercise 11.12**

Table 11.8 gives data from a study of 65 patients who had received sodium aurothiomalate (SA) as a treatment for rheumatoid arthritis. Of these patients, 37 had shown evidence of a toxic effect of the drug. The patients were also classified according to whether or not they had impaired sulphoxidation capacity; the researchers thought this might be linked to toxicity in some way. Use Fisher's exact test to test the null hypothesis that these two variables are not associated (against a two-sided alternative). What do you conclude? If you conclude that the variables are related, briefly describe the way in which they are related.

<table>
<thead>
<tr>
<th>Table 11.8</th>
<th>Impaired sulphoxidation capacity and evidence of toxicity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Impaired sulphoxidation</td>
<td>Toxicity</td>
</tr>
<tr>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Column totals</td>
<td>37</td>
</tr>
</tbody>
</table>

We concluded in Exercise 11.12 that patients with impaired sulphoxidation are more likely to exhibit a toxic reaction to the drug. It is important to remember that, in contingency tables just as in other kinds of data, 'correlation does not imply causation' so we cannot conclude on the basis of these data alone that the impaired sulphoxidation causes the toxic reaction, or vice versa.

Incidentally, it is worth examining briefly how Fisher's exact test can be used for two apparently completely different purposes: testing for association in a $2 \times 2$ contingency table and testing for differences between binomial proportions. In Exercise 8.12 you used Fisher's exact test to analyse data from an experiment where 100 male students and 105 female students asked for help; 71 of the males and 89 of the females were helped. The question of interest there was whether or not the proportions of males and females helped were the same.

To see how these data relate to contingency tables, look at Table 11.9 where they have been written out as a contingency table.

<table>
<thead>
<tr>
<th>Table 11.9</th>
<th>Helping behaviour</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Helped</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
</tr>
<tr>
<td>Male</td>
<td>71</td>
</tr>
<tr>
<td></td>
<td>100</td>
</tr>
<tr>
<td>Female</td>
<td>89</td>
</tr>
<tr>
<td></td>
<td>105</td>
</tr>
<tr>
<td>Column totals</td>
<td>160</td>
</tr>
<tr>
<td></td>
<td>205</td>
</tr>
</tbody>
</table>

This contingency table looks very like Tables 11.7 and 11.8. It does differ in one key respect, though. In Table 11.7, for instance, the researcher chose 60 people from the drug rehabilitation programme, and then classified them in two ways, by educational history and subsequent conviction record. If the researcher had chosen a different sample, the numbers in each cell of the table might have been different. In other words, the researcher did not fix either the row totals or the column totals in advance; they are all random variables that were observed during the study. Only the overall total of 60 people was fixed in advance. However, Table 11.9 is different. There, the row totals were fixed in advance, because the data come from an experiment where the researcher chose to observe exactly 100 occasions involving males and 105 involving females. The column totals are not fixed in advance, but are random variables depending on how many people happened to help overall.

The data appear to have the same form in Tables 11.7 and 11.9; but the random variables involved are different. Table 11.9 does not show bivariate data because two random variables were not recorded on each occasion when help might have been offered. Two pieces of information were recorded, the sex of the person involved and whether or not help was offered; but the sex was not a random variable.

However, Fisher argued that his exact testing procedure applied in both of these different situations. You will recall that the null distribution of the test statistic was calculated on the basis that both sets of marginal totals (row and column totals) remained fixed. Without going into all the details, according to Fisher, the procedure could be applied where none of the marginal totals were fixed (testing for association, as in Table 11.7) or where one set was fixed and the other was not (testing for difference in proportions, as in Table 11.9). Most statisticians nowadays agree with Fisher about the applicability of his test.
11.4 The chi-squared test for contingency tables

Fisher's exact test requires a computer for all but the smallest data sets, and the computations involved may be too much even for some statistical software if the data set is too large. It is therefore useful to have another method of testing for association. The method you will learn about is an application of the chi-squared goodness-of-fit test that was discussed in Chapter 9. There, you saw that the distribution of the test statistic was approximately chi-squared, but the approximation is good provided the data set is not too small. The same applies in relation to contingency tables.

There is no simple analogue of Fisher's exact test for contingency tables that have more than two rows and more than two columns, and exact tests for such tables are not always straightforward (though they can be carried out with some statistical software). The chi-squared test, however, applies just as well to larger contingency tables as to $2 \times 2$ tables. It is worth learning about the chi-squared test for $2 \times 2$ tables, before extending the test to larger tables.

11.4.1 The chi-squared test in 2 x 2 contingency tables

Let us begin with an example that shows how a chi-squared test for association can be carried out in a $2 \times 2$ table.

Example 11.7 Rheumatoid arthritis treatment

Table 11.8 presented some data on reactions to a treatment for rheumatoid arthritis. How can these data be analysed using a chi-squared goodness-of-fit test? In Table 11.8 we have counts of observed frequencies in four cells (together with a number of totals). If we can construct a table of four expected frequencies, we can calculate the chi-squared goodness-of-fit test statistic in exactly the same way as in Section 9.2.

The null hypothesis we are testing here states that the two variables involved, toxicity and impaired sulphoxidation, are not related. The expected frequencies should be calculated on the basis that the null hypothesis is true. If it is true, then knowing whether a patient has impaired sulphoxidation will tell us nothing about the toxic effect in a patient. Altogether, according to Table 11.8, 37 patients out of 65 showed toxic effect. That is, a proportion of $37/65$ of the patients suffered toxic effects. If the null hypothesis were true, we would expect that, of the 39 patients in all who had impaired sulphoxidation, a proportion of $37/65$ of them would show toxic effect. (If the proportion differed much from $37/65$, then knowing that a person had impaired sulphoxidation would tell you something about the toxic effect in that patient; but under the null hypothesis we have assumed this does not happen.) Therefore, the expected number of patients who showed toxic effect would be $(39 \times 37)/65$ or 22.2. Out of the 26 patients who did not have impaired sulphoxidation the expected proportion that shows toxic effect would also be $37/65$, and the expected number of patients would be $(26 \times 37)/65$ or 14.8.
These expected frequencies are shown in Table 11.10, together with the remaining two expected frequencies. Note that each expected frequency is found by multiplying the row total for its row by the column total for its column, and then dividing by the overall total, 65.

<table>
<thead>
<tr>
<th>Impaired sulphoxidation</th>
<th>Toxicity</th>
<th>Row totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>Yes</td>
<td>(39 × 37)/65 = 22.2</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>(26 × 37)/65 = 14.8</td>
</tr>
<tr>
<td>Column totals</td>
<td></td>
<td>37</td>
</tr>
</tbody>
</table>

The first point to note about this table is that the expected frequencies sum to the same row and column totals as the observed frequencies did. The second point is that the expected values, calculated on the basis that the two variables were independent, differ considerably from the observed counts. The differences between observed and expected values are tabulated in Table 11.11.

<table>
<thead>
<tr>
<th>Impaired sulphoxidation</th>
<th>Toxicity</th>
<th>(O − E) values for the rheumatoid arthritis data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yes</td>
<td>30 − 22.2 = 7.8</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>7 − 14.8 = −7.8</td>
</tr>
</tbody>
</table>

Here, it is interesting to note that the row and column sums of the (Observed − Expected) differences are all zero (because the row and column sums were the same for the observed and expected frequencies). As a consequence, the values in Table 11.11 are all the same, apart from the signs. It is now straightforward to complete the calculation of the chi-squared test statistic. It is defined as

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i},$$

where the summation is over all the cells (four of them, here). Thus the value of the chi-squared test statistic is

$$\chi^2 = \frac{(7.8)^2}{22.2} + \frac{(-7.8)^2}{14.8} + \frac{(-7.8)^2}{16.8} + \frac{(7.8)^2}{11.2} = 15.905.$$

To complete the test, we need to know how many degrees of freedom are appropriate for the chi-squared distribution involved. The answer is one. This is indicated by the fact that (apart from the signs) there is only one value in the table of differences between observed and expected values. To be more formal, there are four cells, and we have estimated two parameters from the data, namely the proportion of individuals who have impaired sulphoxidation and the proportion of individuals who show toxic effect, so the number of degrees of freedom is 4 − 1 = 2 = 1. Using a computer, the SP for an observed value of 15.905 against $\chi^2(1)$ is 0.000067. There is extremely strong evidence that these two variables are associated, just as you found using Fisher's exact test in Exercise 11.12.
In general, the chi-squared goodness-of-fit test can be applied to test for association in a $2 \times 2$ contingency table as follows.

**Testing for association in a $2 \times 2$ contingency table**

1. Calculate a table of expected frequencies by multiplying the corresponding row and column totals together, and then dividing by the overall total.

2. Calculate the chi-squared test statistic using the formula

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i},$$

where the summation is over the four cells in the table.

3. Under the null hypothesis of no association, the test statistic has an approximate chi-squared distribution with 1 degree of freedom. Calculate the $SP$ for the test and interpret your answer.

As with the chi-squared tests in Section 9.2, the question of the adequacy of the approximation in Step 3 arises. The same simple rule applies: the approximation is adequate if no expected frequency is less than 5, otherwise it may not be good enough.

In practice, most statistical computer packages will carry out this test.

Note that this test, as it is presented here, is essentially two-sided; because the differences between observed and expected frequencies are squared in calculating the test statistic, their signs are ignored, and the test treats departures from the null hypothesis in either direction on the same footing. It is possible to complicate the testing procedure somewhat to give a one-sided test; we shall not do this here.

**Exercise 11.13**

A study was carried out in which 671 tiger beetles were classified in two ways, according to their colour pattern (bright red or not bright red) and the season in which they were found (spring or summer). The data are given in Table 11.12. Use your computer to carry out a chi-squared test for association between the two variables involved, and report your conclusions.

<table>
<thead>
<tr>
<th>Table 11.12</th>
<th>Colour pattern and seasonal incidence of tiger beetles</th>
</tr>
</thead>
<tbody>
<tr>
<td>Season</td>
<td>Colour pattern</td>
</tr>
<tr>
<td></td>
<td>Bright red</td>
</tr>
<tr>
<td>Spring</td>
<td>302</td>
</tr>
<tr>
<td>Summer</td>
<td>72</td>
</tr>
<tr>
<td>Column totals</td>
<td>374</td>
</tr>
</tbody>
</table>


If any expected frequency is less than 5, then the only option is an exact test, for the idea of 'pooling' cells in a $2 \times 2$ contingency table is not a meaningful one.
### 11.4.2 The chi-squared test in larger contingency tables

In Example 11.3 you met a contingency table which appeared (merely from a consideration of the frequencies reported) to indicate an association between snoring frequency and heart disease. How can we test whether these two variables really are associated in the population from which the sample was drawn? The examples we have analysed so far all had two rows and two columns, but the table in Example 11.3 had four columns. One answer is to use the chi-squared goodness-of-fit test again. The procedure given in Subsection 11.4.1 applies in exactly the same way to tables larger than $2 \times 2$, except that the number of degrees of freedom for the chi-squared distribution is different. For contingency tables in general, it is conventional to denote the number of rows by $r$, and the number of columns by $c$. The procedure is as follows.

**Testing for association in a $r \times c$ contingency table**

1. Calculate a table of expected frequencies by multiplying the corresponding row and column totals together, and then dividing by the overall total.

2. Calculate the chi-squared test statistic using the formula

$$
\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i},
$$

where the summation is over the $rc$ cells in the table.

3. Under the null hypothesis of no association, the test statistic has an approximate chi-squared distribution with $(r - 1)(c - 1)$ degrees of freedom. Calculate the SP for the test and interpret your answer.

As in previous situations, the adequacy of the approximation may not be good enough if any of the expected frequencies is less than 5.

**Example 11.3 continued**

We shall use the chi-squared test to test for association between snoring frequency and heart disease in the data from Table 11.2. The data (observed frequencies) were as follows.

<table>
<thead>
<tr>
<th>Heart disease</th>
<th>Non-snorers</th>
<th>Occasional snorers</th>
<th>Snore nearly every night</th>
<th>Snore every night</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>24</td>
<td>35</td>
<td>21</td>
<td>30</td>
<td>110</td>
</tr>
<tr>
<td>No</td>
<td>1355</td>
<td>603</td>
<td>192</td>
<td>224</td>
<td>2374</td>
</tr>
<tr>
<td>Total</td>
<td>1379</td>
<td>638</td>
<td>213</td>
<td>254</td>
<td>2484</td>
</tr>
</tbody>
</table>
The expected frequencies can be calculated as follows.

<table>
<thead>
<tr>
<th>Heart disease</th>
<th>Non-snorers</th>
<th>Occasional snorers</th>
<th>Snore nearly every night</th>
<th>Snore every night</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>110 × 1379</td>
<td>110 × 638</td>
<td>110 × 213</td>
<td>110 × 254</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2484</td>
<td>2484</td>
<td>2484</td>
<td>2484</td>
<td></td>
</tr>
<tr>
<td></td>
<td>= 61.067</td>
<td>= 28.253</td>
<td>= 9.432</td>
<td>= 11.248</td>
<td>110</td>
</tr>
<tr>
<td>No</td>
<td>2374 × 1379</td>
<td>2374 × 638</td>
<td>2374 × 213</td>
<td>2374 × 254</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2484</td>
<td>2484</td>
<td>2484</td>
<td>2484</td>
<td></td>
</tr>
<tr>
<td></td>
<td>= 1317.933</td>
<td>= 609.747</td>
<td>= 203.568</td>
<td>= 242.752</td>
<td>2374</td>
</tr>
<tr>
<td>Total</td>
<td>1379</td>
<td>638</td>
<td>213</td>
<td>254</td>
<td>2484</td>
</tr>
</tbody>
</table>

The differences $O_i - E_i$ are as follows.

<table>
<thead>
<tr>
<th>Heart disease</th>
<th>Non-snorers</th>
<th>Occasional snorers</th>
<th>Snore nearly every night</th>
<th>Snore every night</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>-37.067</td>
<td>6.747</td>
<td>11.568</td>
<td>18.752</td>
</tr>
<tr>
<td>No</td>
<td>37.067</td>
<td>-6.747</td>
<td>-11.568</td>
<td>-18.752</td>
</tr>
<tr>
<td>Total</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Comparing the observed and expected frequencies, far more of the frequent snorers seem to have heart disease than would be expected under the null hypothesis of no association.

The value of the chi-squared test statistic is found from

$$
\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} = \frac{(-37.067)^2}{61.067} + \cdots + \frac{(-18.752)^2}{242.752} = 72.782.
$$

There are two rows and four columns in the contingency table so the number of degrees of freedom is $(2 - 1)(4 - 1) = 3$. Comparing the value of our test statistic against a chi-squared distribution with 3 degrees of freedom, the $SP$ is very close to zero. There is very strong evidence of an association between snoring and heart disease, though because 'correlation is not causation' we cannot conclude that snoring causes heart disease, or that heart disease causes snoring. 

Now try the following exercise.

**Exercise 11.14**

(a) Table 11.13 gives data on the number of failures of piston-rings in each of three legs in each of four steam-driven compressors located in the same building. The compressors have identical design and are oriented in the same way. One question of interest is whether there is an association between the leg in which a failure occurs and the compressor in which it occurs, or whether the pattern of the location of failures is the same for different compressors. Use a chi-squared test to investigate whether such an association exists, and report your conclusions.

The chi-squared test for contingency tables is a very useful method of analysis. Indeed, it is more widely applicable than you have yet seen. You will recall that, in $2 \times 2$ tables, Fisher’s exact test could be used to compare proportions as well as to test for association. The same is true of the chi-squared test. In a $2 \times 2$ table, it can be used to compare proportions in much the same way that Fisher’s test can be used. In tables with two rows and more than two columns where the column totals are fixed in advance, it can also be used to provide an overall test of the null hypothesis that several binomial proportions are all equal. In larger tables where one set of marginal totals (row or column totals) are fixed in advance, the chi-squared procedure can be used to test various different hypotheses.

However, these further uses of the chi-squared test are beyond the scope of this course. To complete our study of related variables, we now turn to a probability model for continuous bivariate data.

### 11.5 The bivariate normal distribution

In earlier parts of the course, when dealing with univariate data (the kind where only one random variable at a time is involved), it was often useful (and sometimes necessary) to define a probability model for the data. Often, hypothesis tests were performed or confidence intervals were calculated on the basis that the data involved came from a particular probability distribution, such as a normal distribution or a Poisson distribution. The same is true
of bivariate data. It was mentioned in Section 11.2 that the test for association using the Pearson correlation coefficient as the test statistic involved an assumption that the data came from a particular probability model, the bivariate normal distribution. (None of the other techniques in this chapter involves the use of an explicit probability model.) The aim of this section is to show you some of the properties of the bivariate normal distribution, which is the most important probability model for continuous bivariate data, and hence to give you a flavour of what is involved in probability models for bivariate data.

The bivariate normal distribution is a model for a pair of random variables $X$ and $Y$. The model defines in all respects how $X$ and $Y$ vary together. This involves defining the distributions of $X$ and $Y$ taken on their own, and you will not be surprised to learn that both these distributions are normal. But there is more to the bivariate normal distribution than that. It has to account for correlation between $X$ and $Y$. This is done by defining a bivariate version of the normal probability density function.

Think of how the idea of a univariate probability density function was arrived at earlier in the course. We started with histograms, where the height of each bar in the histogram above the axis represented the frequency with which a particular range of values occurred. Many histograms had approximately the 'bell' shape characteristic of the probability density function of the normal distribution: hence its usefulness as a probability model.

To define a bivariate probability density function, we need to be able to represent how common are different pairs $(x, y)$ of values of the two random variables involved. That is, the probability density function needs to be defined over an area of a plane, rather than simply on a line. Imagine taking a scatter plot, dividing the area up into squares, and setting up a bar on each of the squares whose height is proportional to the frequency of data in that square. The resulting object (Figure 11.15) is a bivariate version of the histogram, and in the same way that the smooth curve of the (univariate) normal probability density function looks like a smoothed and idealized histogram, so the bivariate normal probability density function is a smooth surface which is a smoothed and idealized version of a bivariate 'histogram' like that in Figure 11.15.

Two examples of bivariate normal probability density functions are shown in Figure 11.16. The 'contour' lines join up points of equal height, just as with contour lines on an ordinary map. In the first distribution (Figure 11.16(a)), the correlation between the two variables is zero. The 'hill' is circular in cross section. If a small sample were drawn from a population which was modelled by this distribution, a scatter plot of the sample would look like that shown in Figure 11.17; the variables are not related. In the second distribution (Figure 11.16(b)), the Pearson correlation coefficient for the two variables is high; it is 0.9. The 'hill' is a narrow ridge.
Figure 11.16 (a) Zero correlation  (b) High positive correlation

Figure 11.17  A scatter plot of observations drawn from a bivariate normal distribution with zero correlation

Exercise 11.15
Imagine that a small sample is drawn from a population following the second distribution (that with high positive correlation). Draw a rough sketch of what a scatter plot of this sample would look like.

A bivariate normal distribution for bivariate random variables \( X \) and \( Y \) is characterized by five parameters: the means of \( X \) and \( Y \) and their variances, which are denoted by \( \mu_X, \mu_Y, \sigma_X^2 \) and \( \sigma_Y^2 \) respectively, and the (Pearson) correlation between \( X \) and \( Y \), which is denoted by \( \rho \). We write

\[
(X, Y) \sim N(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho).
\]

A bivariate normal distribution has several interesting properties, and these are as follows.

**The bivariate normal distribution**

If the random variables \( X \) and \( Y \) have a bivariate normal distribution

\[
(X, Y) \sim N(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)
\]

then

1. the distribution of \( X \) is normal with mean \( \mu_X \) and variance \( \sigma_X^2 \),

\[
X \sim N(\mu_X, \sigma_X^2)
\]

and the distribution of \( Y \) is normal with mean \( \mu_Y \) and variance \( \sigma_Y^2 \);,

2. the conditional distribution of \( X \), given \( Y = y \), is normal

\[
(X | Y = y) \sim N(\mu_X + \rho \frac{\sigma_X}{\sigma_Y} (y - \mu_Y), \sigma_X^2 (1 - \rho^2))
\]

(and similarly for the conditional distribution of \( Y \), given \( X = x \).)

There is a formal definition for the correlation between random variables, involving the idea of expectation, just as there is a formal definition involving expectations for the mean and variance of a random variable. You need not worry about the details. The correlation is denoted by \( \rho \), the Greek letter rho, pronounced ‘roe’.
The unconditional distributions for \( X \) and \( Y \) are often referred to as the marginal distributions of \( X \) and \( Y \), because they correspond to the marginal totals in a contingency table (which give total frequencies for each of the variables).

The following example gives some idea of a context where the bivariate normal model might be applied.

**Example 11.8 Heights and weights of schoolgirls in Bradford**

For the sample of 30 schoolchildren whose heights and weights were plotted in Figure 11.1 and listed in Table 11.4, summary sample statistics (denoting height by \( X \) and weight by \( Y \) and using the Pearson product-moment correlation coefficient) are

\[
\bar{x} = 144.87, \quad \bar{y} = 36.17, \quad s_X = 7.61, \quad s_Y = 7.70, \quad r = 0.743.
\]

Suppose that it is decided to model the variation in heights and weights of eleven-year-old schoolgirls in the population by a bivariate normal distribution with parameters \( \mu_X = 145, \mu_Y = 36, \sigma^2_X = 58, \sigma^2_Y = 60, \rho = 0.75 \). That is, \((X, Y) \sim N(145, 36, 58, 60, 0.75)\).

Then, for instance, the proportion of eleven-year-old schoolgirls in the population who are more than 140 cm tall is given by

\[
P(X > 140) = 1 - \Phi\left( \frac{140 - 145}{\sqrt{58}} \right) = 0.744;
\]

the proportion of schoolgirls whose weight is below 30 kg is given by

\[
P(Y < 30) = \Phi\left( \frac{30 - 36}{\sqrt{60}} \right) = 0.219.
\]

Considering only those schoolgirls 150 cm tall (this is taller than average: \( x = 150 \) while \( \mu_X = 145 \)) the weight distribution has mean

\[
E(Y|X = x = 150) = \mu_Y + \rho \frac{\sigma_Y}{\sigma_X} (x - \mu_X)
\]

\[
= 36 + 0.75 \times \sqrt{\frac{60}{58}} \times (150 - 145)
\]

\[
= 36 + 3.8
\]

\[
= 39.8 \text{ kg},
\]

which is also heavier than average.

To find, say, the proportion of eleven-year-old schoolgirls who are at the same time taller than 150 cm and heavier than 40 kg requires calculation of the probability

\[
P(X > 150, Y > 40);
\]

here, the only feasible approach is to use a computer programmed to return bivariate normal probabilities. The answer is 0.181.

As you can appreciate, to calculate all but very simple probabilities from a bivariate normal distribution requires the use of a computer, or at the very least a set of statistical tables, a calculator and plenty of time. Your statistical software may do these calculations for you.
In this section we have been able to give only a flavour of the bivariate normal distribution. There are similar analogues of the normal distribution in more than two dimensions—the multivariate normal distribution—but that is well beyond the scope of this course.

**Summary**

1. Two random variables \( X \) and \( Y \) are said to be associated (or not independent) if
   \[
P(X = x | Y = y) \neq P(X = x)
   \]
   for some \( x \) and \( y \), where the probability on the left-hand side is called a conditional probability, and is read 'the probability that \( X = x \) given that \( Y = y \)' or 'the probability that \( X = x \) conditional on \( Y = y \)'.

2. A measure of linear association between two variables \( X \) and \( Y \) is given by the Pearson product-moment correlation coefficient \( r \) for the sample \((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\), where
   \[
   r = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{\sqrt{(n \sum x_i^2 - (\sum x_i)^2)(n \sum y_i^2 - (\sum y_i)^2)}}.
   \]

3. A measure of monotonic association between two variables \( X \) and \( Y \) is given by the Spearman rank correlation coefficient \( r_S \) for the sample \((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\); here, the observations in each sample are replaced by their ranks (averaged, if necessary) and the Pearson correlation is calculated for the ranked data.

4. Correlation is not causation.

5. The null hypothesis that the underlying Pearson correlation is zero may be tested using
   \[
   R \sqrt{\frac{n - 2}{1 - R^2}} \sim t(n - 2).
   \]

6. When data are arranged as counts in a \( 2 \times 2 \) contingency table, an exact test for no association may be provided by posing the problem as a test for equality of two proportions, and using Fisher's test.

7. When data are arranged as counts in a \( r \times c \) contingency table, an approximate test for no association is provided by the test statistic
   \[
   \chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} \sim \chi^2 ((r - 1)(c - 1)),
   \]
   where the sum is taken over all \( rc \) cells of the table, and the expected frequency \( E_i \) is given by
   \[
   E_i = \frac{\text{row total} \times \text{column total}}{\text{overall total}}.
   \]
   The approximation is adequate as long as every \( E_i \) is at least 5.
8. A probability model for bivariate data is given by the bivariate normal distribution

\[(X, Y) \sim N(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho),\]

where the correlation \( \rho \) is a measure of linear association between the two variables. The marginal distributions of \( X \) and \( Y \) are given by

\[X \sim N(\mu_X, \sigma_X^2), \quad Y \sim N(\mu_Y, \sigma_Y^2);\]

and the conditional distribution of \( X \), given \( Y = y \), is also normal:

\[(X|Y = y) \sim N\left(\mu_X + \rho \frac{\sigma_X}{\sigma_Y} (y - \mu_Y), \sigma_X^2 (1 - \rho^2)\right).\]

Similarly, the conditional distribution of \( Y \), given \( X = x \), is normal.

The bivariate probability

\[P(X \leq x, Y \leq y)\]

(and variations on it) cannot be calculated as a combination of univariate normal probabilities, and requires the use of a computer.