Contents

Chapter 11

Association Between Variables

11.4 Correlation

11.4.1 Introduction

The measures of association examined so far in this chapter are useful for describing the nature of association between two variables which are measured at no more than the nominal scale of measurement. All of these measures, ϕ, C, Cramer's xV and λ, can also be used to describe association between variables that are measured at the ordinal, interval or ratio level of measurement. But where variables are measured at these higher levels, it is preferable to employ measures of association which use the information contained in these higher levels of measurement. These higher level scales either rank values of the variable, or permit the researcher to measure the distance between values of the variable. The association between rankings of two variables, or the association of distances between values of the variable provides a more detailed idea of the nature of association between variables than do the measures examined so far in this chapter.

While there are many measures of association for variables which are measured at the ordinal or higher level of measurement, correlation is the most commonly used approach. This section shows how to calculate and interpret correlation coefficients for ordinal and interval level scales. Methods of correlation summarize the relationship between two variables in a single number called the **correlation coefficient**. The correlation coefficient is

usually given the symbol r and it ranges from -1 to +1. A correlation coefficient quite close to 0, but either positive or negative, implies little or no relationship between the two variables. A correlation coefficient close to plus 1 means a **positive relationship** between the two variables, with increases in one of the variables being associated with increases in the other variable. A correlation coefficient close to -1 indicates a **negative relationship** between two variables, with an increase in one of the variables being associated with a decrease in the other variable.

A correlation coefficient can be produced for ordinal, interval or ratio level variables, but has little meaning for variables which are measured on a scale which is no more than nominal. For ordinal scales, the correlation coefficient which is usually calculated is **Spearman's rho**. For interval or ratio level scales, the most commonly used correlation coefficient is **Pearson's r**, ordinarily referred to as simply the **correlation coefficient** or r. The latter is discussed first, with Spearman's rho being introduced in Section 11.4.6.

11.4.2 Correlation Coefficient

The correlation coefficient, r, is a summary measure that describes the extent of the statistical relationship between two interval or ratio level variables. The correlation coefficient is scaled so that it is always between -1 and +1. When r is close to 0 this means that there is little relationship between the variables and the farther away from 0 r is, in either the positive or negative direction, the greater the relationship between the two variables.

The two variables are often given the symbols X and Y. In order to illustrate how the two variables are related, the values of X and Y are pictured by drawing the **scatter diagram**, graphing combinations of the two variables. The scatter diagram is given first, and then the method of determining Pearson's r is presented. In presenting the following examples, relatively small sample sizes are given. Later, data from larger samples are given.

Scatter Diagram A scatter diagram is a diagram that shows the values of two variables X and Y, along with the way in which these two variables relate to each other. The values of variable X are given along the horizontal axis, with the values of the variable Y given on the vertical axis. For purposes of drawing a scatter diagram, and determining the correlation coefficient, it does not matter which of the two variables is the X variable, and which is Y.

Later, when the regression model is used, one of the variables is defined as an independent variable, and the other is defined as a dependent variable. In regression, the independent variable X is considered to have some effect or influence on the dependent variable Y. Correlation methods are symmetric with respect to the two variables, with no indication of causation or direction of influence being part of the statistical consideration.

A scatter diagram is given in the following example. The same example is later used to determine the correlation coefficient.

Example 11.4.1 Years of Education and Age of Entry to Labour Force

Table 11.1 gives the number of years of formal education (X) and the age of entry into the labour force (Y), for 12 males from the Regina Labour Force Survey. Both variables are measured in years, a ratio level of measurement and the highest level of measurement. All of the males are aged 30 or over, so that most of these males are likely to have completed their formal education.

Respondent Number	Years of Education, X	Age of Entry into Labour Force, Y
1	10	16
2	12	17
3	15	18
4	8	15
5	20	18
6	17	22
7	12	19
8	15	22
9	12	18
10	10	15
11	8	18
12	10	16

Table 11.1: Years of Education and Age of Entry into Labour Force for 12 Regina Males

Since most males enter the labour force soon after they leave formal schooling, a close relationship between these two variables is expected. By

looking through the table, it can be seen that those respondents who obtained more years of schooling generally entered the labour force at an older age. The mean years of schooling is $\bar{X} = 12.4$ *years and the mean age of entry into the labour force is* $\bar{Y} = 17.8$*, a difference of 5.4 years. This difference roughly reflects the age of entry into formal schooling, that is, age five or six. It can be seen though that the relationship between years of schooling and age of entry into the labour force is not perfect. Respondent 11, for example, has only 8 years of schooling but did not enter the labour force until age 18. In contrast, respondent 5 has 20 years of schooling, but entered the labour force at age 18. The scatter diagram provides a quick way of examining the relationship between* X *and* Y*.*

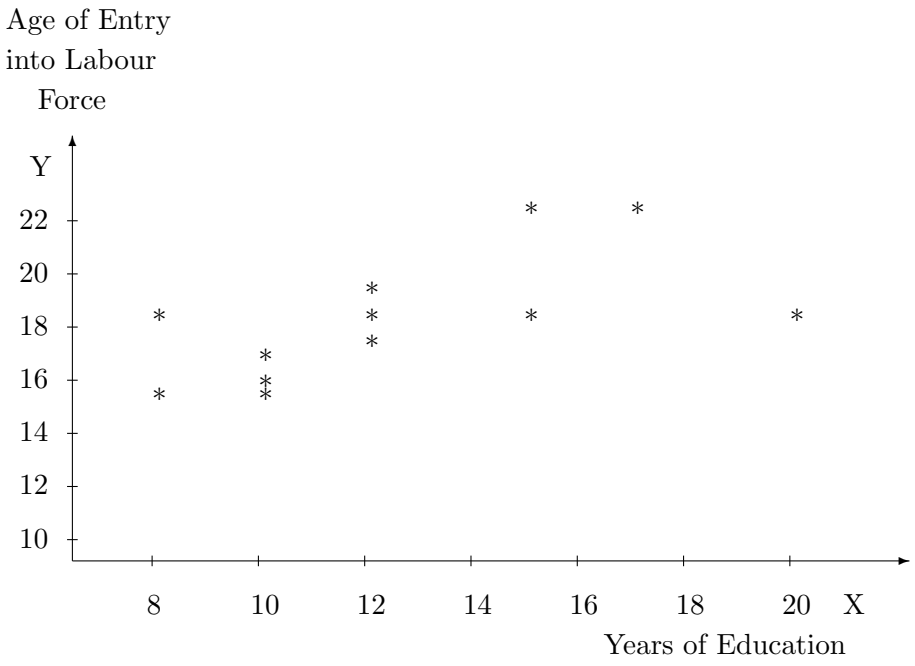


Figure 11.1: Scatter Diagram of Years of Education and Age of Entry into Labour Force

Figure 11.1 gives the scatter diagram for the data in Table 11.1. Years of education is plotted along the horizontal, and is given the symbol X. Age of entry into the labour force is plotted on the vertical axis as variable Y. Each respondent is represented by an asterik in Figure 11.1. Respondent 1 has 10 years of education and entered the labour force at age 16, and is represented by one of the two closely spaced asteriks at $X = 10$ and $Y = 16$. Respondent 2 has $X = 12$ and $Y = 17$, and the asterik representing this respondent lies to the right and slightly above the asterik for the first respondent. Each of the other respondents is similarly represented by one of the asteriks.

By examining the scatter diagram, the relationship between X and Y can be seen at a glance. It can be seen that larger values of X are associated with larger values of Y. That is, as the number of years of education increases, the age of entry into the labour force generally is greater. Respondent 5, represented by the asterik farthest to the right, with 20 years of education, but entering the labour force at age 18, is an exception. The other points generally lie along a line that goes from the lower left to the upper right of the diagram. This indicates a generally positive relationship between X and Y, so that more years of education is associated with an older age of entry into the labour force.

The scatter diagram provides a quick view of the relationship between two variables. In Table 11.1, the nature of the relationship may be a little difficult to see, given the variation in values of each of the variables. By putting all of the combinations of observed values of the variables as points on a scatter diagram, the way in which two variables are related can often be clearly pictured. The scatter diagram is a useful first approach to examining how two variables are related. Where there are many observations, the scatter diagram may be time consuming to draw, and may be difficult to read. But where there are relatively few observations, the scatter diagram is a very useful first step in examining the relationship between two variables. Most statistical software programs contain a plotting procedure which provides a scatter diagram for two variables, and it can be worthwhile to examine this scatter diagram before obtaining the value of the correlation coefficient.

When scatter diagrams are fairly similar to each other, it is difficult to compare these diagrams in order to determine which relationship is stronger. For these purposes, the correlation coefficient is used. The correlation coefficient is a measure of association that provides a single number summarizing this relationship.

11.4.3 Pearson's r

The Pearson product-moment correlation coefficient, better known as the correlation coefficient, or as r, is the most widely used correlation coefficient. Values of r for pairs of variables are commonly reported in research reports and journals as a means of summarizing the extent of the relationship between two variables. Pearson's r summarizes the relationship between two variables that have a **straight line** or **linear** relationship with each other. If the two variables have a straight line relationship in the positive direction, then r will be positive and considerably above 0. If the linear relationship is in the negative direction, so that increases in one variable, are associated with decreases in the other, then $r < 0$. The possible values of r range from -1 to +1, with values close to 0 signifying little relationship between the two variables. Exactly how different from 0 the value of r must be before providing evidence of a relationship can be determined on the basis of an hypothesis test. It will also be seen in Section 11.4.7 that the size of r can differ rather considerably depending on what type of data is being examined.

The Pearson correlation coefficient r can be defined as follows. Suppose that there are two variables X and Y, each having n values $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_n$ respectively. Let the mean of X be $\bar{X}$ and the mean of Y be $\bar{Y}$. Pearson's r is

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2}}$$

where the summation proceeds across all n possible values of X and Y. A method of computing r is presented next, with an example. Following this, there is some discussion of the meaning and interpretation of the correlation coefficient.

Calculating r. The above expression can be used to determine the correlation coefficient r. The following procedure, however, provides a more compact method of determining the correlation coefficient, as well as the regression line of Section 11.5, along with tests of hypotheses for both of these. Since all of these are often calculated at the same time, the following procedure is computationally more straightforward and less time consuming than using the above definition of r.

Beginning with the n values of the variables X and Y, sum these values to obtain $\sum X$ and $\sum Y$. Compute the squares of each of the X values, and the squares of each of the Y values. Then calculate the sums of each of

these, $\sum X^2$ and $\sum Y^2$. Further, compute the products of each pair of the X and Y values, and the sum of these, $\sum XY$.

From the values n, $\sum X$, $\sum Y$, $\sum X^2$, $\sum Y^2$ and $\sum XY$, compute:

$$S_{XX} = \sum X^2 - \frac{(\sum X)^2}{n}$$

$$S_{YY} = \sum Y^2 - \frac{(\sum Y)^2}{n}$$

$$S_{XY} = \sum XY - \frac{(\sum X)(\sum Y)}{n}$$

Once these expressions have been calculated, they can be used to obtain both the correlation coefficient and the regression line. The correlation coefficient can be shown to be equal to

$$r = \frac{S_{XY}}{\sqrt{S_{XX} S_{YY}}}.$$

This method of determining the correlation coefficient is used to determine the correlation coefficient for the data in Table 11.1.

Example 11.4.2 Correlation Coefficient for Relationship between Years of Education and Age of Entry into the Labour Force

In Example 11.4.1, a scatter diagram showing the relationship between years of formal education and age of entry into the labour force was given. From the scatter diagram, it could be seen that the two variables were

positively related, with respondents having more years of education generally entering the labour force at an older age. The relationship was not perfect though, and the correlation coefficient provides a means of summarizing the extent of the relationship between the two variables.

X	Y	X^2	Y^2	XY
10	16	100	256	160
12	17	144	289	204
15	18	225	324	270
8	15	64	225	120
20	18	400	324	360
17	22	289	484	374
12	19	144	361	228
15	22	225	484	330
12	18	144	324	216
10	15	100	225	150
8	18	64	324	144
10	16	100	256	160
149	214	1,999	3,876	2,716

Table 11.2: Calculations for r

Table 11.2 begins the calculation of the correlation coefficient. The 12 values of X and Y are given in the first two columns. The third column contains the squares of each of the X values in the first column, and the sum of the third column is $\sum X^2 = 1,999$. The fourth column contains the sum of the squares of the Y values of the second column, and the sum of the fourth column is $\sum Y^2 = 3,876$. The final column of the table contains the products of the X values of the first column and the Y values of the second column. For example, the first entry is $10 \times 16 = 160$. The sum of the fifth column is $\sum XY = 2,716$.
From Table 11.2,

$$\sum X = 149$$
$$\sum Y = 214$$
$$\sum X^2 = 1,999$$

$$\sum Y^2 = 3,876$$

$$\sum XY = 2,716$$

These values can now be used to determine S_{XX}, S_{YY} and S_{XY} as follows.

$$\begin{aligned} S_{XX} &= \sum X^2 - \frac{(\sum X)^2}{n} \\ &= 1,999 - \frac{(149)^2}{12} \\ &= 1,999 - \frac{22,201}{12} \\ &= 1,999 - 1,850.0833 \\ &= 148.9167 \end{aligned}$$

$$\begin{aligned} S_{YY} &= \sum Y^2 - \frac{(\sum Y)^2}{n} \\ &= 3,876 - \frac{(214)^2}{12} \\ &= 3,876 - \frac{45,769}{12} \\ &= 3,876 - 3,816.3333 \\ &= 59.6667 \end{aligned}$$

$$\begin{aligned} S_{XY} &= \sum XY - \frac{(\sum X)(\sum Y)}{n} \\ &= 2,716 - \frac{(149)(214)}{12} \\ &= 2,716 - \frac{31,886}{12} \\ &= 2,716 - 2,657.1667 \\ &= 58.8333 \end{aligned}$$

Based on these values,

$$r = \frac{S_{XY}}{\sqrt{S_{XX}S_{YY}}}$$

$$
\begin{aligned}
&= \frac{58.8333}{\sqrt{148.9167 \times 59.6667}} \\
&= \frac{58.8333}{94.2622} \\
&= 0.6241
\end{aligned}
$$

The correlation coefficient between years of formal education and age of entry into the labour force is $r = 0.624$. This indicates a relatively large positive relationship between the two variables. A perfect positive relationship would yield a correlation of 1 and no relationship at all between X and Y would give a correlation coefficient of 0. The relationship here is then a relatively large one, above 0.5, but considerably less than a perfect association between the two variables.

Note that several decimals have been carried through much of the calculation. In order to obtain a reasonably accurate estimate of the correlation coefficient, it is best to carry 3 or 4 significant figures through the calculations. Then the correlation coefficient can be rounded at the end of the calculations.

Interpretation of r. The above example shows how r can be calculated. If you have not examined correlation coefficients before, it may be difficult to know what is a large and what is a small correlation. The discussion that follows is intended to assist in interpreting the value of r.

The possible values for the correlation coefficient r are shown in figure 11.2. It can be seen there that the possible values range from -1 to +1. A correlation coefficient close to -1 indicates that the values of X and Y are strongly negatively associated. That is, for larger values of X, it is generally found that the values of Y are smaller. Alternatively stated, if r is close to -1, as X increases, Y generally decreases.

When r is close to 0, either on the positive or the negative side, then there is little or no association between X and Y. Exactly how different from 0 the correlation coefficient must be, before there is considered to be association, depends on the type of data being examined, and on the strength of association. Later in this section there is a test of significance for the correlation coefficient.

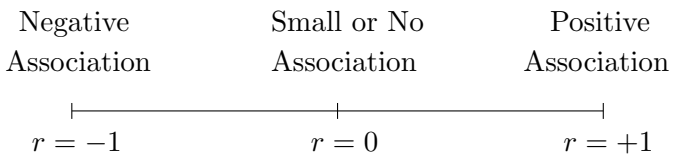


Figure 11.2: Scale for Correlation Coefficient r

When the correlation coefficient is above 0, then this provides evidence of a positive relationship between X and Y. That is, if $r > 0$, larger values of X are associated with larger values of Y. If r is close to 1, this indicates a large positive relationship between the two variables.

Figure 11.3 gives scatter diagrams for six different sets of data, along with the correlation coefficient that corresponds to each of the scatter diagrams. By examining these diagrams, it should be possible to obtain some idea of the nature of association and the size of the correlation coefficient that corresponds to each of these. Each of these is discussed briefly here.

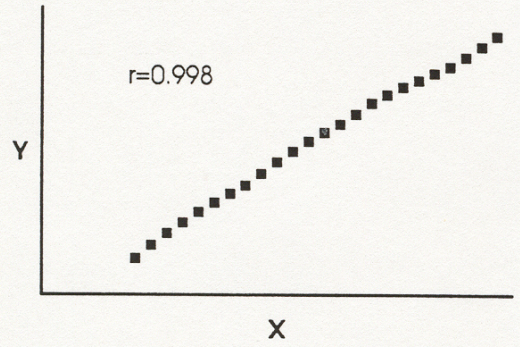
Diagram 1 of Figure 11.3 shows almost a perfect positive relationship. The correlation coefficient of $r = 0.998$ is almost equal to 1. By looking at the scatter of points, it can be seen that all of the points are very close to being along a straight line. Successively larger values of X are associated with larger values of Y, and these points lie close to a line that goes from the bottom left to the top right of the diagram.

Diagram 2 shows a large negative relationship, with $r = -0.817$. As X increases, Y generally decreases, although there are a considerable number of exceptions. A researcher examining this scatter of points would ordinarily conclude that X and Y have a fairly strong negative association with each other.

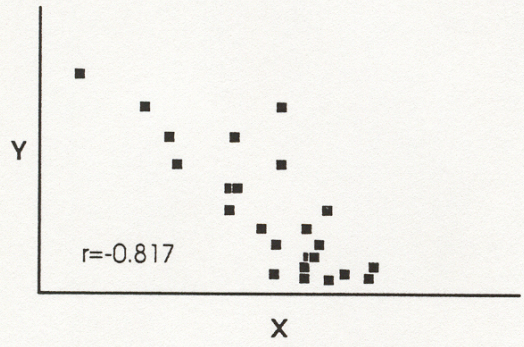
Diagram 3, with $r = 0.487$, indicates a generally positive association between X and Y, but the relationship is far from being a perfect association. In particular, for many of the values of X it appears that Y values may be either low or high. But as X increases, Y generally increases as well.

Diagram 4 shows little relationship between X and Y on the left part of the diagram. But on the right, as X increases, Y also increases. This produces a small positive association between X and Y. This is not a large relationship, but is sufficient to make $r = 0.278$.

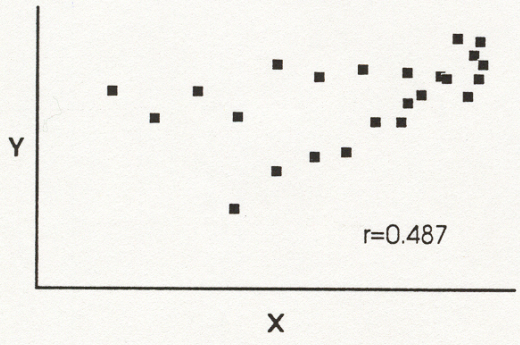
Diagram 5 provides little or no evidence of a systematic relationship between X and Y. While the scatter of points in the diagram may not be



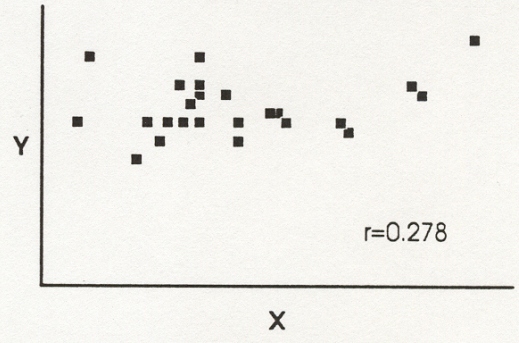
1. Large Positive Correlation



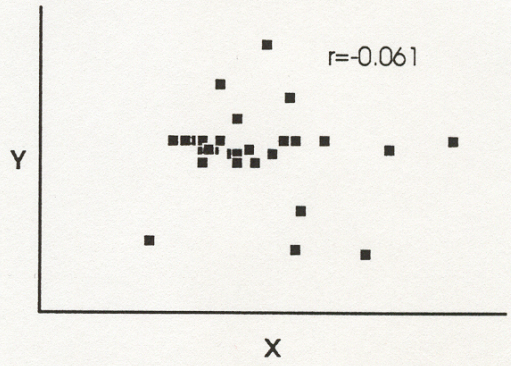
2. Large Negative Correlation



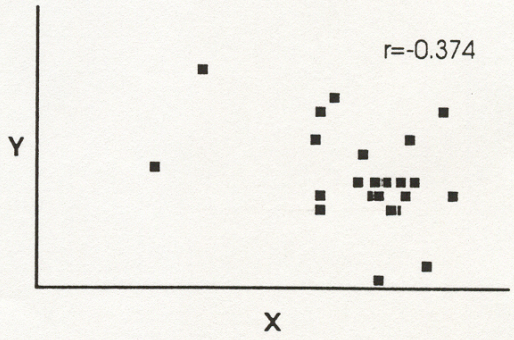
3. Modest Positive Correlation



4. Small Positive Correlation



5. Minimal Negative Correlation



6. Modest Negative Correlation

Figure 11.3: Scatter Diagram and Correlation Coefficients for Six Different Relationships

random, it is difficult to determine any positive or negative relationship between X and Y. The size of the correlation coefficient is $r = -0.061$, slightly below 0. But r is close enough to 0 that on the basis of this correlation, a researcher would ordinarily conclude that there is no relationship between X and Y.

Finally, diagram 6 shows a modest negative association between X and Y. The correlation coefficient is $r = -0.374$, evidence of a negative association between X and Y, but with nowhere near a perfectly negative association. The two points on the left, and the two or three points on the bottom right are what produces the negative association. The bulk of the points in the centre of the diagram appear to be more or less randomly scattered.

From these diagrams it should be possible to obtain some idea of the nature of association in a scatter diagram and the correlation coefficient. When examining a relationship between two variables, where there are relatively few data points, or where the data has been entered into a computer program, it is often useful to obtain the plot of the scatter diagram. The correlation coefficient summarizes the association, and this along with the association visible in the plot of the scatter diagram can give considerable information concerning the relationship between two variables.

Explanation of the Formula for r. The formulas presented above are those which are used to determine r. While the calculation is relatively straightforward, although tedious, no explanation for the formula has been given. Here an intuitive explanation of the formula is provided.

Recall that the original formula for determining the correlation coefficient r for the association between two variables X and Y is

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2}}.$$

The denominator of this formula involves the sums of the squares of the deviations of each value of X and Y about their respective means. These summations under the square root sign in the denominator are the same expressions as were used when calculating the variance and the standard deviation in Chapter 5. The expression $\sum(X_i - \bar{X})^2$ can be called the **variation in** X. This differs from the variance in that it is not divided by $n-1$. Similarly, $\sum(Y_i - \bar{Y})^2$ is the variation in Y. The denominator of the expression for r is the square root of the products of these two variations. It can

be shown mathematically that this denominator, along with the expression in the numerator scales the correlation coefficient so that it has limits of -1 and +1.

The numerator of the expression for r is

$$\sum (X_i - \bar{X})(Y_i - \bar{Y})$$

and this is called the **covariation** of X and Y. In order to understand how the covariation of X and Y behaves, Figure 11.4 may be helpful. This scatter diagram is similar to the scatter diagram of Figure 11.1, with some points shifted in order to illustrate covariation more clearly.

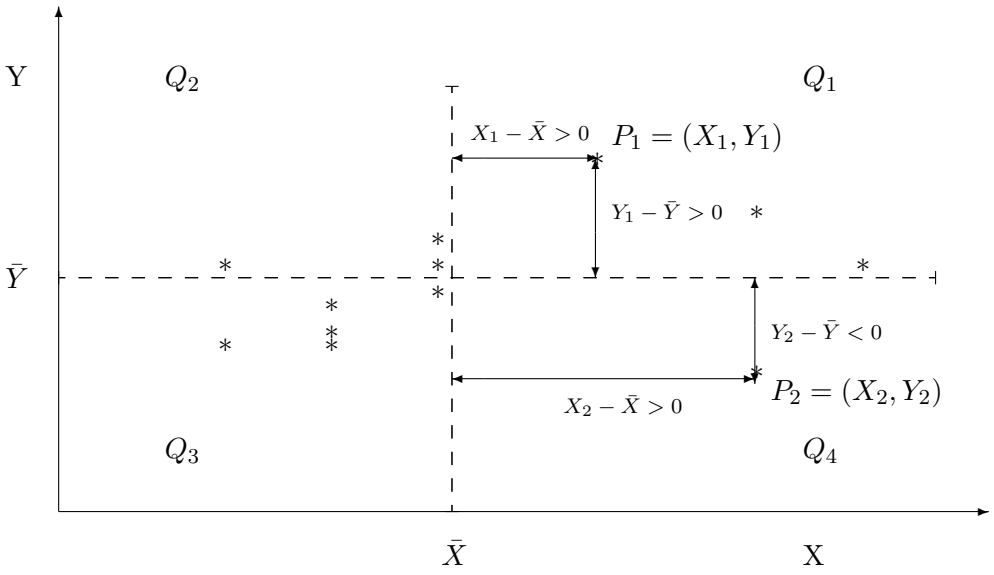


Figure 11.4: Covariation of X and Y

Covariation examines the product of the individual differences of each of X and Y about their respective means. The mean $\bar{X}$ for X is given by the dashed vertical line in the middle of the diagram. The horizontal

distance between each observation and the dashed line represents $X_i - \bar{X}$, the amount by which X_i differs from $\bar{X}$, sometimes called **the deviation about the mean**. The dashed horizontal line represents $\bar{Y}$, the mean of the observed Y values. The vertical distance between each asterik and the dashed line represents $Y_j - \bar{Y}$. If the two dashed lines are used as the new axes, then the horizontal and vertical distances of each observed value about these new axes represent the deviations about the respective mean of each variable. The products of these two deviations then represent the elements in the expression for the covariation of X and Y. These new axes also divide the diagram into four quadrants, Q_1, Q_2, Q_3 and Q_4.

First examine a point P_1 in the first quadrant. Let this point have value X_1 for the X variable, and Y_1 for the Y variable. This point has positive deviations about each mean. That is, both X and Y are greater than their respective means, so that $X_1 - \bar{X}$ and $Y_1 - \bar{Y}$ are both positive in value. For point P_1, the product $(X_1 - \bar{X})(Y_1 - \bar{Y})$ is positive, and this point contributes positively to the covariation in the numerator of r.

A different picture emerges for point P_2, with values X_2 for X and Y_2 for Y. For points in the fourth quadrant such as P_2, the horizontal deviation about the mean is positive, but the vertical deviation about the mean is negative. That is, all points in Q_2 have X values exceeding the mean $\bar{X}$, but have Y values which are less than the mean $\bar{Y}$. This means that $X_2 - \bar{X} > 0$ and $Y_2 - \bar{Y} < 0$ and the resulting product $(X_2 - \bar{X})(Y_2 - \bar{Y})$ is less than 0. All the points in Q_4 produce negative entries for the covariation in the numerator of the expression for r.

Points in each of the second and third quadrants could be analyzed in a similar manner. For the four quadrants associated with $\bar{X}$ and $\bar{Y}$ as the axes (the dashed lines), the respective entries in the expression for the covariation of X and Y are as follows.

Quadrant	Entry in Covariation
Q_1	$(X - \bar{X})(Y - \bar{Y}) > 0$
Q_2	$(X - \bar{X})(Y - \bar{Y}) < 0$
Q_3	$(X - \bar{X})(Y - \bar{Y}) > 0$
Q_4	$(X - \bar{X})(Y - \bar{Y}) < 0$

Now consider what happens when there is a scatter diagram which expresses a generally positive association, as in Figure 11.4. Again using the

mean of X and the mean of Y as the new axes, and with a positive association between X and Y, most of the points in the scatter diagram are in the first and third quadrants. While there will be some points in the second and fourth quadrants, with a positive relationship between X and Y, these will be relatively few in number. As a result, for a positive association, most of the entries in the covariation term will be positive, with only a few negative entries. This produces a positive covariation, and a correlation coefficient above 0. The denominator of the expression for r divides the covariation by terms which express the variation in X and Y. Together these produce a value for r which cannot exceed $+1$.

Also note that the greater the greater the positive association between X and Y, the more likely the points are to lie in the first and third quadrants. As the relationship between X and Y becomes less clear, the points in the scatter diagram could lie in any of the four quadrants. In this latter case, the positive and negative entries would cancel out, producing a covariation, and correlation coefficient, close to 0.

When the relationship becomes negative, there are more points in the second and fourth quadrants, and fewer in the first and third quadrants. This produces more negative and fewer positive entries in the covariation expression, and results in a value of r that is less than 0. Again the denominator of the expression for r contains terms which express the variation in X and Y, resulting in a value for r which cannot be less than -1.

What the preceding explanation shows is that the scatter diagram can be used to illustrate the nature of association between X and Y as expressed in the correlation coefficient. While the exact formula for the correlation coefficient is based on mathematical considerations, the various terms in the expression for r should make some sense. Computationally though, a more straightforward way to compute r in the manner shown in Example 11.4.2.

11.4.4 Test of Significance for r

When computing a correlation coefficient, it is also useful to test the correlation coefficient for significance. This provides the researcher with some idea of how large a correlation coefficient must be before considering it to demonstrate that there really is a relationship between two variables. It may be that two variables are related by chance, and an hypothesis test for r allows the researcher to decide whether the observed r could have emerged by chance or not.

In order to test the correlation coefficient for statistical significance, it

is necessary to define the true correlation coefficient that would be observed if all population values were obtained. This true correlation coefficient is usually given the Greek symbol ρ or *rho*. This is pronounced 'row' as in 'row of corn' in English.

The null hypothesis is that there is no relationship between the two variables X and Y. That is, if ρ is the true correlation coefficient for the two variables X and Y, when all population values are observed, then the null hypothesis is

$$H_0 : \rho = 0.$$

The alternative hypothesis could be any one of three forms, with $\rho \neq 0$, $\rho < 0$ or $\rho > 0$. If the researcher has no idea whether or how two variables are related, then the two tailed alternative hypothesis

$$H_1 : \rho \neq 0$$

would be used. If the researcher suspects, or has knowledge, that the two variables are negatively related, then

$$H_1 : \rho < 0$$

would be used. If the test is to determine whether the observed value of the statistic is enough greater than 0 to prove a positive relationship, then the null hypothesis is

$$H_1 : \rho > 0.$$

The test statistic for the hypothesis test is the sample or observed correlation coefficient r. As various samples are drawn, each of sample size n, the values of r vary from sample to sample. The sampling distribution of r is approximated by a t distribution with $n-2$ degrees for freedom. The reason why there are $n-2$ degrees of freedom will become more more apparent in Section 11.5. The standard deviation of r can be shown to be approximated by

$$\sqrt{\frac{1-r^2}{n-2}}$$

For the null hypothesis

$$H_0 : \rho = 0$$

the standardized t statistic can be written

$$t = \frac{r-\rho}{\sqrt{\frac{1-r^2}{n-2}}} = r\sqrt{\frac{n-2}{1-r^2}}$$

and there are $n-2$ degrees of freedom for this statistic. The data of Example 11.4.2 is used to illustrate how the test can be carried out.

Example 11.4.3 Test for Significance of Relationship between Years of Schooling and Age of Entry into the Labour Force

For the data in Table 11.2, $r = 0.6241$ and there were $n = 12$ observations. Let ρ be the true correlation between years of formal schooling and age of entry into the labour force for all males. The null hypothesis is that there is no relationship between these two variables and the research hypothesis is that the two variables are positively related. These hypotheses are

$$H_0 : \rho = 0$$

$$H_1 : \rho > 0$$

The test statistic is r and the standardized t statistic for r is

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

where there are $n-2$ degrees of freedom. Choosing the 0.05 level of significance, for a one tailed test, with $n-2 = 12-2 = 10$ degrees of freedom, $t = 1.813$. The region of rejection for H_0 is all t values of 1.813 or larger.

Using the data, $r = 0.6241$, so that

$$\begin{aligned} t &= r\sqrt{\frac{n-2}{1-r^2}} \\ &= 0.6241\sqrt{\frac{12-2}{1-0.6241^2}} \\ &= 0.6241\sqrt{\frac{10}{0.6104992}} \\ &= 0.6241\sqrt{16.3800381} \\ &= 0.6241 \times 4.047226 \\ &= 2.526 > 1.813 \end{aligned}$$

The correlation coefficient and the corresponding t value is in the region of rejection of H_0. The null hypothesis of no relationship between years of schooling and age of entry into the labour force can be rejected. The alternative hypothesis that there is a positive relationship between years of

schooling and age of entry into the labour force can be accepted. While the correlation coefficient is not real close to 1, it is enough greater than 0 to reject the hypothesis of no relationship. There is a probability of less than 0.05 that $r = 0.6241$*, with* $n = 12$*, could occur if there is no relationship between the two variables.*

11.4.5 Correlation and Causation

When two variables have a large positive or negative correlation with each other, there is often a tendency to regard the two variables as causally related. That is, if X and Y have a large positive correlation coefficient, then a researcher may consider this as proof that X causes Y, or that Y causes X, or that the two variables are connected in some more complex causal way. In doing this, the researcher must exercise considerable caution.

For the data in Example 11.4.2, a researcher might claim that males decide how many years of formal schooling they wish to take, and the age of entry into the labour force is a result of this. That is, after schooling is completed, these males entered the labour force. While this may be the case, and the reasonably large correlation coefficient of $r = 0.624$ supports this claim, the nature of causation may be quite different. For example, it may be that teenage males find they can get a job, or are forced by economic circumstances to leave school and look for a job. This decision may occur first, and the number of years of formal schooling that these males receive is a result of their decision to enter the labour force. The correlation coefficient by itself does not allow the researcher to decide which of these two circumstances is closer to the truth. Both explanations are reasonable ones, and the statistically significant positive correlation coefficient could be used to support either explanation.

An example of a large negative correlation, which may or may not indicate causation, follows.

Example 11.4.4 Correlation between Crude Birth Rate and Female Labour Force Participation Rate

From the 1950s through the 1970s, two socioeconomic variables showed a dramatic shift in Canada. The birth rate fell rapidly over this period, by the mid 1970s declining to almost one half the level it had been in the early to mid 1950s. At the same time the labour force participation rate of females rose very rapidly, and continued to rise through to the 1990s. From

the early 1950s to the mid 1970s, the percentage of women in the labour force doubled.

Suppose that a researcher examining these variables hypothesizes that the decline in the birth rate meant that women in Canada had less need to care for children, and as a result entered the labour force. According to the researcher, this led to an increase in the female labour force participation rate. Using the data in the first two columns of Table 11.3, draw the scatter diagram, calculate the correlation coefficient, and comment on the claim of the researcher.

The data for this example comes from the following sources. The crude birth rate is the number of births per one thousand population. It is determined by taking the annual number of births in Canada, dividing this by the mid year Canadian population and multiplied by 1000. This data comes from M. C. Urquhart and K. A. H. Buckley, **Historical Statistics of Canada**, *second edition, Ottawa, 1983, catalogue number 11-516, Series B4. The female labour force participation rate is the percentage of women aged 25 and over who are members of the labour force, as defined in Chapter 2. These data were obtained by averaging the monthly figures reported in* **Seasonally Adjusted Labour Force Statistics, January 1953-December 1972**, *Ottawa, 1973, Statistics Canada catalogue number 71-201 and* **Historical Labour Force Statistics - Actual Data, Seasonal Factors, Seasonally Adujsted Data**, *Ottawa, 1975, Statistics Canada catalogue number 71-201.*

Solution. *The scatter diagram showing the relationship between X and Y is given in Figure 11.5. By examining this figure, it appears that the two variables have a strong negative relationship with each other. For most of the years, the birth rate declines while the female labour force participation rate increases. A quick look at the table and the scatter diagram appears to support the contention of the researcher. The calculations required to determine the correlation coefficient are contained in the fourth through sixth columns of Table 11.3.*

The following summations are obtained from Table 11.3:

$$\sum X = 497.4$$

$$\sum Y = 601.6$$

$$\sum X^2 = 11,779.82$$

$$\sum Y^2 = 17,134.30$$

Year	Crude Birth Rate X	Labour Force Participation Rate Y	X^2	Y^2	XY
1953	28.1	18.5	789.61	342.25	519.85
1954	28.5	18.9	812.25	357.21	538.65
1955	28.2	19.4	795.24	376.36	547.08
1956	28.0	20.4	784.00	416.16	571.20
1957	28.2	21.8	795.24	475.24	614.76
1958	27.5	22.5	756.25	506.25	618.75
1959	27.4	23.2	750.76	538.24	635.68
1960	26.8	24.5	718.24	600.25	656.60
1961	26.1	25.5	681.21	650.25	665.55
1962	25.3	26.0	640.09	676.00	657.80
1963	24.6	26.8	605.16	718.24	659.28
1964	23.5	27.8	552.25	772.84	653.30
1965	21.3	28.6	453.69	817.96	609.18
1966	19.4	29.8	376.36	888.04	578.12
1967	18.2	30.9	331.24	954.81	562.38
1968	17.6	31.4	309.76	985.96	552.64
1969	17.6	32.3	309.76	1043.29	568.48
1970	17.5	32.9	306.25	1082.41	575.75
1971	16.8	33.8	282.24	1142.44	567.84
1972	15.9	34.4	252.81	1183.36	546.96
1973	15.5	35.7	240.25	1274.49	553.35
1974	15.4	36.5	237.16	1332.25	562.10
Total	497.4	601.6	11,779.82	17,134.30	13,015.30

Table 11.3: Calculations for Correlation between Birth Rate and Female Labour Force Participation Rate

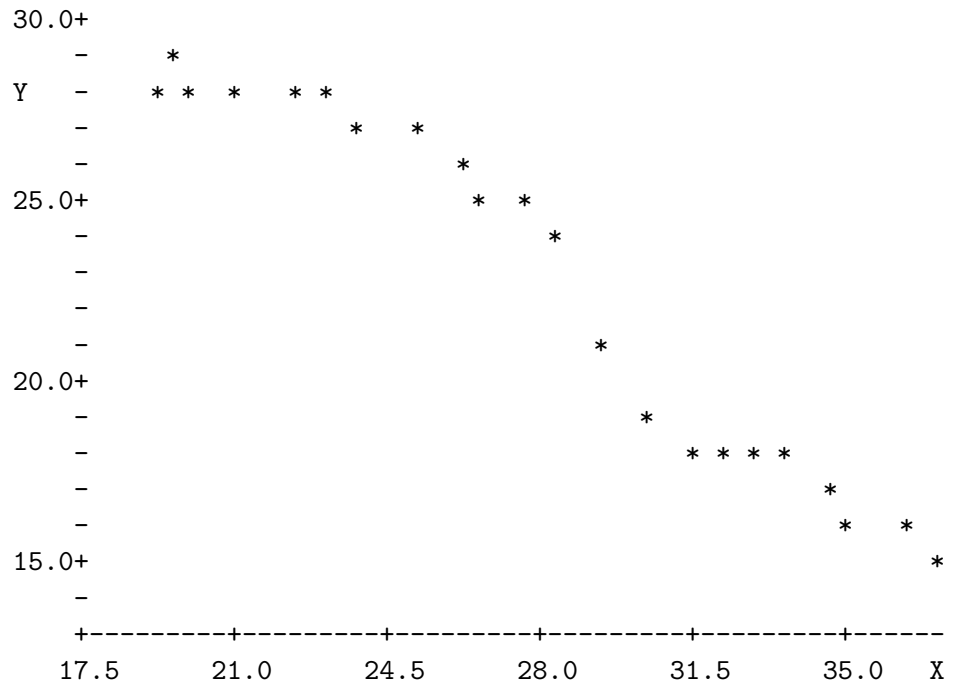


Figure 11.5: Scatter Diagram of Crude Birth Rate (X) and Female Labour Force Participation Rate, Age 25 and over, Canada, 1953-1974

$$\sum XY = 13,015.30$$

These values in are now used to determine S_{XX}, S_{YY} *and* S_{XY}.

$$\begin{aligned} S_{XX} &= \sum X^2 - \frac{(\sum X)^2}{n} \\ &= 11,779.82 - \frac{(497.4)^2}{22} \\ &= 11,779.82 - \frac{247,406.76}{22} \\ &= 11,779.82 - 11,245.76 \\ &= 534.058182 \end{aligned}$$

$$S_{YY} = \sum Y^2 - \frac{(\sum Y)^2}{n}$$

$$\begin{aligned} &= 17,134.30 - \frac{(601.6)^2}{22} \\ &= 17,134.30 - \frac{361,922.56}{22} \\ &= 17,134.30 - 16,451.02546 \\ &= 683.274545 \end{aligned}$$

$$\begin{aligned} S_{XY} &= \sum XY - \frac{(\sum X)(\sum Y)}{n} \\ &= 13,015.30 - \frac{(497.40)(601.60)}{22} \\ &= 13,015.30 - \frac{299,235.84}{22} \\ &= 13,015.30 - 13,601.62909 \\ &= -586.329091 \end{aligned}$$

Based on these values,

$$\begin{aligned} r &= \frac{S_{XY}}{\sqrt{S_{XX}S_{YY}}} \\ &= \frac{-586.329091}{\sqrt{534.058182 \times 683.274545}} \\ &= \frac{-586.329091}{604.0764532} \\ &= -0.970621 \end{aligned}$$

The correlation coefficient relating X and Y is $r = -0.971$. This appears to show a very large negative association between the crude birth rate and the female labour force participation rate. Before concluding that this is the case, the statistical significance of the correlation coefficient should be checked.

Let ρ represent the true relationship between the birth rate and the female labour force participation rate over many years. If there two variables are negatively related, as hypothesized, then $\rho < 0$. The null and research hypotheses are

$$H_0 : \rho = 0$$

$$H_1 : \rho < 0$$

The test statistic is r and the standardized t statistic for r is

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

where there are $n-2$ degrees of freedom. Choosing the 0.001 level of significance, for a one tailed test, with $n-2 = 22-2 = 20$ degrees of freedom, $t = 3.850$. The region of rejection for H_0 is all t values to the left of $t = -3.850$.

Using the data, $r = 0.971$, so that

$$\begin{aligned} t &= r\sqrt{\frac{n-2}{1-r^2}} \\ &= -0.971\sqrt{\frac{22-2}{1-0.971^2}} \\ &= -0.971\sqrt{\frac{20}{1-0.9428}} \\ &= -0.971\sqrt{349.901} \\ &= -0.971 \times 18.706 \\ &= -18.163 < -3.850 \end{aligned}$$

The correlation coefficient and the corresponding t value is in the region of rejection of H_0. The null hypothesis of no relationship between the birth rate and the female labour force participation rate can be rejected. The alternative hypothesis that there is a negative relationship between the crude birth rate and the labour force participation rate can be accepted.

The results appear to very stronly support the contention of the researcher. However, there is one serious mistake in concluding that the researcher is entirely correct. There is clearly a large negative association between the birth rate and the female labour force participation rate over these years. Of that there is little doubt, and given $r = -0.971$ when $n = 22$, the probability of obtaining a correlation this large under the assumption of no association is well under 0.001.

But the researcher has also claimed that the direction of the causation has been that the reduced birth rate **led to the increase** *in the female labour force participation rate. The correlation coefficient says nothing concerning the direction of causation and it may be that the direction of causation is the reverse of that claimed by the researcher. It may be a more*

reasonable explanation to argue that over these years, women were able to find jobs, and this led to their entry into the labour force. This may have led many women to postpone having children, or reduce the number of births. The correlation coefficient would support this explanation just as well as the first explanation hypothesized.

Another possibility is that some other variable or variables have changed over these years, simultaneously affecting both the birth rate and the female labour force participation rate. For example, there was a growth in the number of jobs and in wages over these years. In addition, there was a financial squeeze on many families. These could have led more women to search for jobs, and at the same time may have led to a reduction in the birth rate. Another explanation could be that social and attitudinal changes were occurring over this time, making childbearing a less attractive activity for women, and making labour force participation more attractive. It is likely that some combination of all of these explanations would be necessary to explain what happened over these years. The correlation coefficient supports all of these explanations, but does not really allow the researcher to decide which of these explanations is the most reasonable.

Summary. The above example show that considerable care must be taken before concluding that a correlation proves a causal connection between two variables. While a statistically significant correlation can be used to support an explanation, without further evidence the correlation coefficient itself cannot prove the claim. More evidence concerning the behaviour of related variables, and an understanding of the manner in which the two variables really are connected must also be provided.

11.4.6 Spearman's rho

The Pearson correlation coefficient just examined can be used for interval or ratio level scales. When a variable is measured at no more than the ordinal level, the researcher must decide whether to treat the ordinal scale as if it has an interval level scale, or to use a correlation coefficient designed for an ordinal level scale. There are various types of correlation coefficients which have been constructed to allow a researcher to examine the relationship between two variables each of which have at least an ordinal level of measurement. Some of these are *gamma* (the Greek letter γ), various statistics referred to as *tau* (the Greek letter τ) and Spearman's rho. The latter is the simplest to calculate without a computer, and is the only ordinal level correlation

coefficient presented in this textbook. Many other introductory textbooks present gamma and the various types of tau.

Spearman's rho is given the symbol r_s, with r used to denote that it is a correlation coefficient, and the subscript s to denote that it is named after the statistician Spearman. The true Spearman correlation coefficient is called *rho sub s*, that is, ρ_s. The Greek letter ρ is used to denote that ρ_s is the parameter, with the statistic r_s being calculated from the data.

If a scale is ordinal, it is possible to rank the different values of the variable, but the differences between these ranks may not be meaningful. In order to compute Spearman's rho for two variables whose values have been ranked, the numberical difference in the respective ranks are used. Suppose that there are two variables X and Y. For each case that is observed, the rank of the case for each of the variables X and Y is determined by ordering the values from low to high, or from high to low. For each case i, the difference in the rank on variables X and on variable Y is determined, and given the symbol D_i. These differences are squared, and then added producing a summation $\sum D_i^2$. If there are n cases, the Spearman rank correlation between X and Y is defined as

$$r_s = 1 - \frac{6 \sum D_i^2}{n(n^2 - 1)}.$$

This produces a correlation coefficient which has a maximum value of 1, indicating a perfect positive association between the ranks, and a minimum value of -1, indicating a perfect negative association between ranks. A value of 0 indicates no association between the ranks for the observed values of X and Y.

The hypothesis test for the Spearman correlation coefficient is basically the same as the test for the Pearson correlation coefficient. The standard deviation of r_s is approximated by

$$\sqrt{\frac{1 - r_s^2}{n - 2}}$$

and when n is 10 or more, r_s is approximated by a t distribution with $n-2$ degrees of freedom. When the null hypothesis is

$$H_0 : \rho_s = 0$$

the standardized t statistic can be written

$$t = r_s \sqrt{\frac{n-2}{1-r_s^2}}.$$

Example 11.4.5 Rank Correlation for Various Social Indicators among 10 Countries

Ten countries were randomly selected from the list of 129 countries in the Country Data Set of Appendix B. The countries are compared with respect to their gross national product per capita, an indicator of the average income level of each country. Researchers have found that people in countries with higher income levels tend to live longer than do people in countries with lower income levels. In addition, birth rates tend to be higher in countries with lower incomes, with birth rates declining as countries become better off economically. Table 11.4 also gives the mean life expectation at birth (LE) for the 10 countries, and the crude birth rate for each country. In columns 2 through 4 of the table, this data is contained in its interval form, as given in Appendix B. The countries are then ranked for each of the three social indicators, in columns 5 through 7. Use this data to determine the Spearman rank correlation of GNP with (i) life expectation, and (ii) the birth rate. Test to determine whether the correlation of ranks is statistically significant at the 0.05 level of significance.

Solution. *The first step in computing the Spearman rank correlation coefficient is to determine the rankings for each country with respect to each of the social indicators. GNP is ranked from high to low, so that France, with the highest GNP, is ranked first. The country with the second highest GNP is Ireland, so it is ranked 2, with Argentina having the next largest GNP, so that it is ranked third. Similarly, each of the countries is ranked in order. Sierra Leone, with a GNP per capita of only $240, comes 10th.*

Next, the countries are ranked with respect to their level of life expectation. Again, the country with the highest life expectation is ranked first. This is France, with a life expectation of 76 years. Tied for second, with life

Country	GNP	Life Expectation	CBR	Rank on: GNP	LE	CBR
Algeria	2360	65	34	4	6	4
India	340	59	31	9	9	7
Mongolia	780	62	36	8	8	2.5
El Salvador	940	64	36	7	7	2.5
Equador	1120	66	32	6	5	5.5
Malaysia	1940	74	32	5	2.5	5.5
Ireland	7750	74	18	2	2.5	9
Argentina	2520	71	21	3	4	8
France	16090	76	14	1	1	10
Sierra Leone	240	47	48	10	10	1

Table 11.4: Social Indicators for 10 Countries, Actual Value and Rank

expectation of 74 years, are Ireland and Malaysia. Where cases are tied, the ranks which would otherwise occur are averaged. With two countries tied for second place, Malayasia could be second and Ireland third, or Ireland second and Malaysia third. The ranks of 2 and 3 are tied, producing a mean rank of 2.5 for each of these countries. The next highest life expectation is 71, for Argentina, so it is ranked 4. The other countries are ranked in the same manner. Again, the poorest country, Sierra Leone, also has the lowest life expectation among the 10 countries.

For the birth rates, again the ranking is from high to low, in order to be consistent. Sierra Leone, with a CBR of 48 ranks first with respect to birth rate. Two countries are again tied for second or third, Mongolia and El Salvador, each having a birth rate of 36. These are given the average of ranks 2 and 3, or 2.5. Ranked fourth is Algeria with a birth rate of 34. Two countries are again tied for 5 and 6. Each of Ecuador and Malayysia has a birth rate of 32, so these two countries are each given a rank of 5.5. The rest of the countries are ranked in a similar manner.

For the first part of this question, the ranks, differences of ranks, the squares of these differences, and the sums of the squares are given in Table 11.5. By examining the ranks of GNP and LE, it can be seen that there is a high degree of similarity in the ranking of the 10 countries for the two social indicators. The value of r_s will allow the relationship to be summarized

in a single measure of association.

Country	Rank on: GNP	LE	Difference D_i	D_i^2
Algeria	4	6	-2	4
India	9	9	0	0
Mongolia	8	8	0	0
El Salvador	7	7	0	0
Equador	6	5	1	1
Malaysia	5	2.5	2.5	6.25
Ireland	2	2.5	-0.5	0.25
Argentina	3	4	-1	1
France	1	1	0	0
Sierra Leone	10	10	0	0
Total			0	12.50

Table 11.5: Calculations for r_s of GNP and Life Expectation

In Table 11.5, the ranks are subtracted from each other to produce the differences D_i. For example, Algeria ranks 4th on the GNP scale and 6th on the life expectancy scale, for a difference of $4-6=-2$. The difference in ranks for each of the other countries is similarly determined. Note that the sum of the difference in the ranks is 0. The final column of Table 11.5 squares these differences to produce the values D_i^2. The sum of the final column is $\sum D_i^2 = 12.50$. The value of the Spearman rank correlation coefficient is

$$\begin{aligned} r_s &= 1 - \frac{6\sum D_i^2}{n(n^2-1)} \\ &= 1 - \frac{6 \times 12.50}{10 \times (10^2-1)} \\ &= 1 - \frac{75}{10 \times 99} \\ &= 1 - \frac{75}{990} \\ &= 1 - 0.076 \\ &= 0.924 \end{aligned}$$

This confirms the suspicion that the ranks of the 10 countries on the scale of GNP and life expectation are very similar. The maximum possible correlation between ranks is +1, and +0.924 comes very close to this maximum possible association.

An hypothesis test should be conducted, and the null and alternative hypotheses would be

$$H_0 : \rho_s = 0$$

$$H_1 : \rho_s > 0$$

The null hypothesis states that there is no correlation between the ranking of the countries on the GNP and life expectancy scales. The alternative hypothesis states that the rankings are positively related. If the 0.001 level of significance, and $n - 2 = 10 - 2 = 8$ *degrees of freedom is used, the* t *value is 4.500. The region of rejection of the null hypothesis is all* t *values of 4.500 or more. The value of the standardized* t *statistic is*

$$\begin{aligned} t &= r_s\sqrt{\frac{n-2}{1-r_s^2}} \\ &= 0.924\sqrt{\frac{10-2}{1-0.924^2}} \\ &= 0.924\sqrt{\frac{8}{0.146224}} \\ &= 0.924\sqrt{54.71058} \\ &= 0.924 \times 7.39666 \\ &= 6.835 \end{aligned}$$

and this is greater than 4.500. The Spearman rank correlation coefficient is well within the region of rejection of the null hypothesis. As a result, the hypothesis of no association between the ranks of GNP and life expectancy can be rejected, and the data provides evidence of a strong positive relationship between these two variables.

For the second part of the question, the relationship between the rankings of GNP and the crude birth rate are used. These were given first in Table 11.4 and repeated in Table 11.6. In the latter table, the differences between the ranks, the squares of these differences, and the sum of these squares are given.

In Table 11.6, a quite different picture from the earlier table is presented. Many of the countries that rank highest on the GNP scale are the countries

Country	Rank on: GNP	CBR	Difference D_i	D_i^2
Algeria	4	4	0	0
India	9	7	2	4
Mongolia	8	2.5	5.5	30.25
El Salvador	7	2.5	4.5	20.25
Equador	6	5.5	0.5	0.25
Malaysia	5	5.5	-0.5	0.25
Ireland	2	9	-7	49
Argentina	3	8	-5	25
France	1	10	-9	81
Sierra Leone	10	1	9	81
Total			0	291.00

Table 11.6: Calculations for r_s of GNP and CBR

that have the lowest birth rates. For example, Sierra Leone, with the lowest GNP has the highest birth rate. Based on the comparison of these rankings, it appears that there is a fairly considerable negative relationship between the rankings of the countries on the GNP and the birth rate scales.

Compare with the earlier table, the differences D_i are much larger numerically, and the squares of these differences are even larger. The sums of the squares of the differences of the ranks in the final column is $\sum D_i^2 = 291.00$. It is this large value which produces a negative correlation coefficient here. The value of the Spearman rank correlation coefficient is

$$\begin{aligned} r_s &= 1 - \frac{6\sum D_i^2}{n(n^2-1)} \\ &= 1 - \frac{6 \times 291}{10 \times (10^2 - 1)} \\ &= 1 - \frac{1,746}{10 \times 99} \\ &= 1 - \frac{1,746}{990} \\ &= 1 - 1.764 \end{aligned}$$

$$= \quad -0.764$$

The Spearman rank correlation coefficient is a fairly large negative number, indicating a strong negative correlation between GNP and CBR. While the relationship is certainly not a perfect negative correlation, the large numerical value for r_s indicates that among these 10 countries, higher values of GNP are generally associated with lower values of the birth rate.

The null and alternative hypotheses are

$$H_0 : \rho_s = 0$$

$$H_1 : \rho_s < 0$$

The null hypothesis states that there is no correlation between the ranking of the countries on the GNP and CBR scales. The alternative hypothesis states that the rankings are negatively related. If the 0.01 level of significance, and $n - 2 = 10 - 2 = 8$ degrees of freedom is used, the t value is -2.897. The region of rejection of the null hypothesis is all t values to the left of -2.897. The value of the standardized t statistic is

$$\begin{aligned} t &= r_s\sqrt{\frac{n-2}{1-r_s^2}} \\ &= -0.764\sqrt{\frac{10-2}{1-0.764^2}} \\ &= -0.764\sqrt{\frac{8}{1-0.583141}} \\ &= -0.764\sqrt{\frac{8}{0.4168595}} \\ &= -0.764\sqrt{19.19112} \\ &= -0.764 \times 4.38077 \\ &= -3.347 < -2.897 \end{aligned}$$

and the null hypothesis of no relationship between the ranks of the countries on GNP and CBR can be rejected. At the 0.01 level of significance, there is evidence for a negative relationship between ranking of countries on the GNP and CBR scales.

Before leaving this example, it is useful to note how the size of the Spearman correlation of the ranks of the countries compare with the Pearson correlation of the original values for GNP, life expectation and the birth rate.

Correlation Between:	
GNP and LE	0.626
Rank of GNP and Rank of LE	0.924
GNP and CBR	-0.782
Rank of GNP and Rank of CBR	-0.764

Table 11.7: Pearson and Spearman Correlations

Table 11.7 gives the Pearson correlation between GNP and life expectancy as 0.626, while the Spearman correlation between the ranks of these two variables for the 10 countries as 0.924. The large positive value for the Spearman rank correlation coefficient shows that the ranks of the two variables are extremely closely related. The Pearson correlation also shows a fairly large positive association between actual values of GNP and life expectancy, but is not nearly so close to being a perfect association. In the case of the correlation between GNP and the birth rate, the association is very similar in the two cases. The rank and the Pearson correlation both produce large negative correlation coefficients, of about the same size.

Summary. The Spearman rank correlation coefficient is a very useful, and relatively easy, statistic to calculate. When at least one of the variables has no more than an ordinal level of measurement, the Spearman correlation is often used. If there are relatively few observed values of two interval or ratio level variables, and you do not want to take the time to compute the Pearson correlation coefficient, the Spearman correlation can be quickly and easily calculated. It provides a rough estimate of the size of the correlation between the two variables, often being within 0.1 of the value of the Pearson correlation coefficient.

Where there are many observed values, the Spearman correlation coefficient is not usually calculated. Even with ordinal level variables, the Pearson correlation coefficient is used when there are many values for the variable. In addition, when there are many tied values of the variable, the Spearman correlation coefficient may not be all that reliable. In that case, the Pearson correlation coefficient is often used, or measures such as gamma and tau are

calculated.

The Spearman correlation coefficient is most useful then for small data sets. It is easy to calculate, with the ranks for the two variables being easy to determine, and easy to interpret. For larger data sets, it is less useful.

11.4.7 Size of r for Various Types of Data

The size of the correlation coefficient depends on various factors, and if you have not worked extensively with correlation coefficients, then it may be difficult to determine what is a large and what is a small correlation coefficient. While a larger correlation always implies a stronger association between two variables than does a smaller correlation, many other factors also affect the size of the correlation coefficient. Survey data, with a wide variety of people being surveyed, often yield relatively low correlation coefficients. The great diversity of respondents in a large cross sectional survey of a population may mean that variables are quite strongly related, but the correlation coefficient may be no greater than 0.25 or 0.30. In contrast, data which is measured across time may yield very high correlation coefficients. Many pairs of labour force variables, for example, will yield correlation coefficients of 0.9 or more. Correlation coefficients for aggregated data, like the social indicators for 10 countries in Example 11.4.5, often lie between these two extremes.

The following examples contain Pearson correlation coefficients for several types of data. Within each type of data, a larger correlation means a stronger relationship between the variables. But it will be seen that it can be misleading to compare correlation coefficients for different types of data, and conclude that those types with higher correlation coefficients necessarily represent more important relationships between variables.

The sample size used to determine the correlation coefficient may have little effect on the size of the correlation coefficient, but it does have an effect on the level of significance of the coefficient. As the sample size increases, a correlation coefficient of any given size becomes more significant statistically. While this is generally true for all hypothesis tests, this can make it quite misleading to compare the significance of two correlation coefficients with different sample sizes.

Example 11.4.6 Times Series Data - Large Correlation Coefficients

Appendix J contains the values of several social indicators for Canada for the years 1966 through 1989. A quick glance through the values of these variables in Table J.1 in Appendix J shows that many of the variables are highly correlated with each other. The Pearson correlation coefficient for each pair of these variables is given in Table 11.8. YP is the population aged 15-24, POP is the total population, H is the annual number of homicides, SR is the suicide death rate of males aged 15-19 and DR is the divorce rate.

	Year	YP	POP	H	SR
YP	0.526				
POP	0.998	0.570			
H	0.744	0.866	0.774		
SR	0.904	0.784	0.923	0.885	
DR	0.943	0.697	0.954	0.840	0.945

Table 11.8: Pearson Correlation Coefficients for Canadian Social Indicators, 1966-1989

The correlation coefficients for each pair of variables is given in Table 11.8. For example, the correlation between the year (1966-1989) and the young population (YP) is 0.526. The correlation between year and population is 0.998, and between year and the number of homicides is 0.744. On the bottom right of Table 11.8, the correlation between the number of homicides and the suicide rate is 0.885, while it is 0.840 with the divorce rate. Finally, on the very bottom right, the correlation between the suicide rate and the divorce rate is 0.945, almost a perfect correlation. What these high correlations denote is that the two variables moved in a very similar manner over these years. For example, the divorce rate rose most of these years, and the suicide rate for young males also rose most years. Whether these two variables are causally connected is another matter. Note that in the first column, the correlation of the young male suicide rate and year is 0.904, and few people are likely to claim that these are causally related.

Also note that most of these correlation coefficients are very significant statistically. For a one tailed test with $n - 2 = 24 - 2 = 22$ *degrees of*

freedom, the t value is 3.505. Any correlation coefficient of 0.599 or more allows the researcher to reject the null hypothesis of no association, at the 0.001 level of significance. All of the correlation coefficients in the table avery significant at this level, with the exception of two near the upper left of the table. These two are significant at the 0.005 level of significance, still a very significant result statistically.

What these results show is that variables which move in the same direction over time are often very highly correlated with each other. It is not unusual to find correlation coefficients above 0.8 for two variables which are measured over time. While this correlation could be used to support a causal connection between the variables, it would be best for researchers to point out how they could be causally related. The first column of Table 11.8 shows that high correlation coefficients for time series data can be obtained without much effort, just by correlating each series with the year of the observation. But all that the high correlation means is that the variable increased over the years, and this is not an explanation of why the increase occurred.

Example 11.4.7 Cross Sectional or Survey Data - Small Correlation Coefficients

When the data from a large survey is obtained, many of the variables can be shown to have a statistically significant correlation with each other, even though the size of the correlation coefficients may seem small. If a survey is attempting to represent the population of a large city, a province, or the country as a whole, then there is great diversity among all the respondents surveyed. There are respondents of all ages, religions, ethnicity, and education, as well as a great diversity of social, economic and political views. Because of this, the relationship between variables cannot be expected to be nearly as complete as in the case of time series data. For much survey data, a correlation coefficient of 0.5 or more may be regarded as very large. Correlations no larger than 0.1 or 0.2 may be regarded by researchers as worthy of note, and demonstrating some relationship among the variables.

As an example of the size of correlation coefficients among a variety of socioeconomic and attitude variables, correlation coefficients from the 1985 Edmonton Area Study are given in Table 11.9. The Edmonton Area Study is a survey of adults in Edmonton, and provides a picture of a cross section of characteristics of all Edmonton adults. The first three variables in the table are age of respondent (AGE), years of schooling of respondent (EDUC)

and before tax household income of respondent (INCOME). The next four variables are attitude questions which ask respondents whether they disagree or agree with various explanations of unemployment. The questions asked were

> As you know, unemployment has been at a high level in this country and province for several years. Different people have different opinions about the *causes of this high unemployment.* How much do you agree or disagree with the following opinions about this?
>
> Unemployment is high because immigrants have taken up all available jobs. (IMMIGRANTS)
>
> Unemployment is high because of the introduction of widespread automation. (AUTOMATION)
>
> Unemployment is high because trade unions have priced their members out of a job. (UNIONS)
>
> World wide recession and inflation cause high unemployment. (RECESSION)
>
> Many people don't have jobs becuase they are sick or have physical handicaps. (SICK)

Each of the explanations of unemployment was given a seven point scale, with 1 being strongly agree and 7 being strongly disagree.

By examining the correlation coefficients in Table 11.9, it can be seen that none of these exceed a little over 0.3 numerically. The asteriks indicate the level of statistical significance. With a sample size of $n = 362$*, several of the correlations are significant. For example, in the first column AGE and the sickness and handicapped explanation have a correlation of 0.2488. This is a positive relationship, indication that those who are older are more in agreement with this explanation, and younger people are more in disagreement. This correlation coefficient is significant at the 0.001 level of significance.*

Education is correlated negatively with several explanations for unemployment. The largest correlations are eduation with the immigrants and the automation explanation. For immigrants, $r = -0.3118$*, meaning that "less educated respondents were more likely to agree that ... the presence of immigrants were leading to high rates of unemployment." This correlation coefficient is statistically significant at the 0.001 level, as is the correlation*

Variable	AGE	EDUC	INC	IMM	AUTO	UN	REC
EDUC	-.2808**						
INC	-.0316	.3009**					
IMM	.0371	-.3118**	-.1530*				
AUTO	.0728	-.2080**	-.1680**	.3811**			
UN	.0906	-.0617	-.0706	.2019**	.2164**		
REC	.1355*	-.0391	-.0967	-.0124	.1493*	.0730	
SICK	.2488**	-.1009	-.1095	.1000	.2621**	.0791	.1053

Number of cases: 362

1-tailed Signif: * - .01 ** - .001

Table 11.9: Correlation Coefficients for 8 Variables in Edmonton Area Study

of -0.2080 of education with the automation explanation. Again, those with more education are less likely to agree that this is a reasonable explanation of unemployment. The authors of the study note that in general, socioeconomic factors do not play an important role in explaining differences in explantion, except for the immigration and automation explanations.

Note that the correlation among the various explanations are not all that large numerically either. The largest correlation is 0.3811, and this is the correlation between the immigration and automation explanation. That is, those who agree that immigrants are responsible for unemployment also tend to believe that automation is responsible for unemployment.

(The wording of the questions and the quote were taken from H. Krahn, G. S. Lowe, T. F. Hartnagel and J. Tanner, "Explanations of Unemployment in Canada, **International Journal of Comparative Sociology**, *XXVIII, 3-4, 1987, pp. 228-236).*

The above example shows that it may be difficult to find large correlations among variables from survey data. There are too many other variables that prevent a clear observation of the relationship among these variables. In addition, for attitude and opinion variables, the responses and relationships are not so clear cut. Attitudes and many socioeconomic variables are not as clearly related as often claimed by researchers. Many attitudes are similar for people of all different socioeconomic statuses, and differences in attitudes do not always clearly relate to socioeconomic status variables at

the level researchers are able to measure these. Given these deficiencies, the correlation coefficients are often quite small, and researchers must attempt to construct explanations of relationships, even for small correlations.

Example 11.4.8 Aggregrated Data - Medium Sized Correlation Coefficients

When data have been aggregated, so that the observations refer not to individuals, but to aggregates such as cities, provinces, countries, firms, schools, or other groupings, there is a considerable averaging process. Each of these units are aggregates of individual observations, so that the great variability associated with individual cases is usually eliminated. When comparing variables for cross sections of these aggregated units, the correlation coefficients are often considerably greater than for relationships among sets of individuals. On the other hand, these are not time series data, so that the correlations do not reach the very high values that are often associated with pairs of time series variables.

As an example of the medium sized correlations, Table 11.10 contains correlation coefficients for several of the variables across the countries shown in the Country Data Set of Appendix B. Some of the variables have not been included because they show little correlation with the variables given here. The literacy rate, LIT, was left out because there were so many countries for which the literacy rate was not available. The data in Table 11.10 contain correlation coefficients for only the 123 countries where data was available for all the variables shown. The size, and statistical significance, of the Pearson correlation coefficient for each pair of the following variables is given in the table. (A more complete definition of these variables is given in Appendix B.)

IMR	Infant Mortality Rate
GNP	Per Capita Gross National Product
LE	Life Expectation in Years
CDR	Crude Death Rate
CBR	Crude Birth Rate
URB	Per Cent of Population in Urban Areas
AGE	Median Age of the Population

Note that most pairs of the variables have correlation coefficients between 0.4 and 0.9. There are no examples of variables with the almost

	IMR	GNP	LE	CDR	CBR	URB
GNP	-.3976**					
LE	-.6601**	.6364**				
CDR	.8532**	-.2148*	-.6984**			
CBR	.5602**	-.6610**	-.9030**	.4794**		
URB	-.5190**	.6054**	.7615**	-.4995**	-.6920**	
AGE	-.3276**	.7761**	.7202**	-.1177	-.8815**	.6018**

N of cases: 123

1-tailed Significance: * .01, ** .001

Table 11.10: Correlation Coefficients for Variables in Country Data Set

perfect association that can occur with time series variables. Yet the correlation coefficients are considerably greater than in Example 11.4.7 where correlations between variables obtained from individual data were obtained. Also note that most of the correlation coefficients are statistically significant at the 0.001 level. Only the correlation between the crude death rate and median age is insignificant. The correlation coefficient between the death rate and per capita GNP is -0.2148, and while this is small, it is still significant at the 0.01 level of significance.

The above examples show that there can be a considerable variety of correlation coefficients, all statistically significant, but of quite different size. A very small correlation coefficient should not necessarily be taken as evidence that there is no relationship between two variables. The statistical significance for the coefficient should always be determined, and if the correlation coefficient is significant, then some attention should be paid to the relationship. When comparing correlation coefficients from different data sets, it may be misleading to compare these coefficients if the data sets are of quite different types.

Summary. Correlation coefficients are one of the most widely used means of determining whether a relationship exists between two variables. If one of the variables has only a nominal scale of measurement, the correlation coefficient is not usually meaningful. But when the variables have at least

ordinal scales of measurement, correlation coefficients provide a useful way of summarizing the relationship between two variables. It should be remembered that the correlation coefficient is a measure of association for a straight line or linear relationship. Two variables may be related in some nonlinear manner, and the Pearson correlation coefficient does not provide a meaningful manner of describing these nonlinear relationships.

Correlation coefficients are widely used in research work, allowing the researcher to summarize the relationship between two variables in a single number. A large, or a statistically significant, correlation coefficient can be taken as evidence that the two variables may be causally related. But by itself a large or a statistically significant correlation cannot prove that two variables are causally related. There is no substitute for determining and understanding how variables are connected, and correlation coefficients cannot indicate this.

In addition to describing relationships, correlation coefficients provide the raw material for many types of multivariate statistical methods. Factor analysis, multivariate regression models, cluster analysis, and other multivariate models can all be built on sets of correlation coefficients. While these models are not examined in this textbook, they are likely to form the content of a second course in statistical methods.