17

Associations

A Correlation

1 Introduction

It is exceptional for social research to produce descriptions or measurements of single variables. Most social researchers are interested in the relationship between events, and try to ascertain the existence of such relationships, their strength and their direction. For this reason, the analysis of data, or at least a large part of it, tends to concentrate on the relationships between variables.

There are many ways of evaluating the type, direction and strength of such relationships. Some measures include two variables, others contain more than two. Whereas some relate to nominal data, others refer to ordinal or interval/ratio data. For instance, such measures provide a useful tool for examining the relationship between a high education level of women and liberal social attitudes; between socioeconomic status of parents and scholastic achievement of their children; between poverty and criminality; and between high technology and rates of divorce.

Measures of correlation are employed to explore three points, namely:

1 *presence or absence of correlation*, that is, whether or not there is a correlation between the variables in question;
2 *direction of correlation*, that is, if there is a correlation, whether it is positive or negative; and
3 *strength of correlation*, that is, whether an existing correlation is strong or weak.

Existence, direction and strength of correlation are demonstrated in the coefficient of correlation. A zero correlation indicates that there is no correlation between the variables. The sign in front of the coefficient indicates whether the variables change in the same direction (*positive correlation*) or in opposite directions (*negative correlation*), except for nominal measures, where the sign has no meaning, in which case coefficients describe only the strength of the

relationship (a high or a low association) between the variables of the study. The value of the coefficient shows the strength of the association, with values close to zero meaning a weak correlation and those close to 1 a strong correlation, as we shall see later. A correlation of +1 is just as strong as one of −1; it is the direction that is different.

2 Overview of relevant options

The selection of the appropriate measure of association is based on a number of factors. Of these, the type of distribution (being continuous or discrete), the structure and characteristics of the distribution and the level of measurement of the data are the most significant. In addition, the availability of computers and relevant computer programs have a relative effect on the choice of measures.

The options are many; Table 17.1 displays the popular measures, at least in the area of social sciences, grouped according to level of measurement. In this section only one measure for each level of measurement will be discussed. These are Yule's Q, ϕ, Spearman's rho and Pearson's r.

Table 17.1 Association tests based on level of measurement

Level	*Association tests*
Nominal	Yule's Q, Lambda test, Contingency coefficient (C), Tschurprow's T, Cramer's V and ϕ coefficient
Ordinal	Spearman's rank-order correlation, Tau-α, Gamma coefficient, Sommer's d and Tau-β
Interval/ratio	Pearson's product-moment correlation

The degree of difficulty as well as the relevance of such measures to social scientists in general and to sociologists in particular vary to some extent. In this section we shall present an overview of how correlational measures are used, computed and interpreted. More procedures and more specialised and complicated tests can be found in relevant readings. Although the list covers only the minimum of measures related to association it does present a test for each possible level of measurement, offering a good start for an analysis of association between variables.

3 Nominal measures of association

In nominal measurement, data are classified in categories by means of numbers or other symbols. For this reason, nominal data cannot be analysed with statistical techniques that employ higher levels of measurement. The methods used are based on the differences that occur between certain values (e.g. expected and observed values) or on predictions made about one variable,

derived from available knowledge about the other.

The computation and interpretation of these measures are fairly straightforward. What is important to remember is that their coefficients range from 0 to 1, with 0 being the lowest level of their value. Negative values have no meaning. The closer the values are to 1, the stronger is the relationship between the variables, and the closer they are to 0, the weaker is the relationship. Three nominal measures of association will be considered in this section: Yule's Q, the ϕ coefficient and Cramer's V.

a Yule's Q

Yule's Q is a popular measure of association for nominal data and a method which is very easy to compute. Named after a famous nineteenth century statistician (Quetelet), this measure rests on the principle that *if values are set in a four-cell table, the cross-products of the internal diagonal cells will be equal when no relationship exists between the two variables* (Eckhardt and Ermann, 1977: 134). This principle is reflected in the formula given below:

$$Q = \frac{AD - BC}{AD + BC} \tag{17.1}$$

A, B, C and D refer to the cells of the relevant table. The computation of Yule's Q is very simple and involves the following steps:

Step 1: First we set up a four-cell table with its cells clearly marked using letters from A to D as shown in Table 12 in Example A.

Step 2: Substitute the values in the formula and compute Q.

Example A: A study of attitudes to feminism including 60 males and 60 females carried out in a country town of NSW produced the data presented in Table 12. How can the relationship between attitudes and gender be described?

Table 12 Attitudes to feminism by gender

Attitudes	*Females*	*Males*	*Total*
Positive	45 *A*	10 *B*	55
Negative	15 *C*	50 *D*	65
Total	60	60	120

Employing the relevant formula we find:

$$Q = \frac{AD - BC}{AD + BC} = \frac{(45 \times 50) - (10 \times 15)}{(45 \times 50) + (10 \times 15)} = \frac{2100}{2400} = 0.875$$

What does 0.875 mean? The value of Q is compared with H_o, which proposes that there is no difference, that is, no expected association between the variables. If Q is low, H_o is accepted; if Q is high, H_o is rejected. In our example, given that the value of Q is high, H_o is rejected, which means that there is a strong association between the variables.

b Phi (ϕ) coefficient

Yule's Q, although simple in logic and computation, has not been included in any of the major computer-assisted statistical packages and is therefore not used as much as other nominal measures. Researchers tend to use other measures instead. Of these measures, two that are relatively powerful and also popular are ϕ and Cramer's V. They are also available in most popular computer programs. Both will be discussed in the next chapter because they are employed also in conjunction with chi-square tests. We shall introduce them briefly next. We begin with the ϕ coefficient.

The characteristics of ϕ are that, like Yule's Q, it is suitable for 2×2 tables, it is computed by means of a simple formula (Formula 17.2), it relies on the chi-square (to be introduced in the next chapter), and is interpreted the same way as Yule's Q. Its value ranges from 0 to 1; in general, if the ϕ value is close to 0, the strength of the relationship is fairly weak; if it is about 0.4 to 0.7 it is moderate; and if it is above 0.8 it is strong or very strong.

$$\phi = \sqrt{\frac{\chi^2}{N}} \tag{17.2}$$

For example, with $\chi^2 = 9.36$ and $N = 85$, ϕ will be as follows:

$$\phi = \sqrt{\frac{9.36}{85}} = \sqrt{0.1101176} = 0.33$$

This suggests that the strength of the relationship between the variables in question is fairly weak.

c Cramer's *V*

This is another measure of association between nominal variables. Its advantage over the ϕ coefficient is that it can be employed when tables are larger than 2×2. This measure (also known as Cramer's ϕ) possesses all the characteristics listed above for the ϕ coefficient, is interpreted the same way and is calculated using Formula (17.3):

$$\text{Cramer's } v = \sqrt{\frac{\chi^2}{N(k-1)}} \tag{17.3}$$

where k is the smaller value of the number of rows and columns, and N the sample size. For instance, if in a table there were 3 rows and 4 columns, k would be 3 (since it is smaller than 4); if there were 6 rows and 4 columns, k would be 4. This measure is interpreted similarly to the ϕ coefficient.

Let us now employ V in an example. If a study in which $N = 65$ and $k = 3$ produced a chi-square of 23.45, Cramer's V would be as follows:

$$\text{Cramer's } V = \sqrt{\frac{23.45}{65(3-1)}} = \sqrt{\frac{23.45}{130}} = \sqrt{0.1803846} = 0.424717$$

This indicates that there is a moderate association between the variables in question.

d Computing ϕ and V using SPSS

In SPSS, the computation of ϕ and Cramer's V is simple, and is a part of the computation of chi-square tests. Both measures are computed simultaneously, and their values are shown together with the chi-square value. Relevant instructions for computing these measures will be given in the next chapter.

4 Ordinal measures of association

Ordinal measures are characterised by their emphasis on ranking and on pairs. They derive their coefficient from ranking pairs, expecting that knowledge of the rank order of the pairs of one variable will allow some degree of prediction about the rank order of the other variable. Of the various measures employed in ordinal data, Spearman's rho is the most common.

a Spearman's rank-order correlation coefficient

This is a fairly simple, useful and very popular *ordinal* measure of association mainly of two ordinal variables. It relates two ordered sets of ranks (not magnitudes), and allows a prediction about one set from the other. This is facilitated by means of the formula:

$$r_s = 1 - \frac{6\Sigma D^2}{N(N^2-1)} \qquad \textbf{(17.4)}$$

The only factor that needs to be calculated in this formula is the sum of the squared difference of the ranks, namely ΣD^2; N is already known. To calculate this factor we proceed as follows:

Step 1
A table with the two sets of categories and rank orders is constructed.

Step 2
The ranks are subtracted and the difference is entered in a new column under D (differences).

Step 3
The differences are squared and the products entered in another column under D^2 (squared differences).

Step 4
The squared differences are summed, giving the sought value.

Step 5
The known values are placed in the formula and the coefficient is computed. Let us apply these steps in an example.

Example B: A group of psychology students ($N = 14$) is ranked from 1 to 14 according to test scores and also according to their popularity; the ranks are as follows: Student **A** 3, 4; **B** 14, 12; **C** 12, 13; **D** 2, 1; **E** 13, 14; **F** 1, 2; **G** 8, 6; **H** 10, 11; **I** 4, 3; **J** 7, 8; **K** 5, 7; **L** 11, 10; **M** 6, 5; and **N** 9, 9. Is there an association between popularity and test results? To answer these questions rho is computed as shown above. The results are given below.

Ranking of test results

Students	*Test ranking*	*Popularity ranking*	D	D^2
A	3	4	−1	1
B	14	12	2	4
C	12	13	−1	1
D	2	1	1	1
E	13	14	−1	1
F	1	2	−1	1
G	8	6	2	4
H	10	11	−1	1
I	4	3	1	1
J	7	8	−1	1
K	5	7	−2	4
L	11	10	1	1
M	6	5	1	1
N	9	9	0	0
				$\Sigma D^2 = 22$

Employing Formula (17.4) we find:

$$r_s = 1 - \frac{6 \times 22}{14(14^2 - 1)} = 1 - \frac{132}{2730} = 1 - 0.048 = 0.952$$

The result shows (1) that there is an association between popularity and test results (the coefficient is not 0); (2) that this association is very strong; and (3) that the association is positive.

Tied ranks

The procedure described in Example B contains a straightforward process of elementary calculations without major complications. One difficulty may arise when tied ranks occur, that is, when two or more subjects are given the same rank. The rule here is that these subjects receive the average rank score, and the ranks of the remaining subjects are adjusted accordingly. For instance, if five subjects were ranked and one was ranked first, three second and the other last, the actual ranks of these five respondents would be 1, 3, 3, 3 and 5. The second, third and fourth position are all allocated rank 3.

Significance of rho

Is the correlation coefficient significant? This is tested by means of a special table containing the critical values of rho. A version of that table containing the significance level for 0.05 and 0.01 (the most popular options) is given in Table 17.2. Note that this table can be used for positive and negative observations but only up to 30 observations. When N is more than 30 a different measure (t-statistic) is employed. The degree of significance is tested in the following way:

- In column 1 of Table 17.2 the value of N that corresponds to the number of cases (sample size) of the test is located. (If the table has no odd numbers, the value of N directly below the sample size is taken (for example, if N is 13, take 12). In our case N is 14.
- The appropriate column is chosen by defining the type of test (one-tailed or two-tailed test) and the significance level; then, the critical value that corresponds to N is located in the appropriate column. This is the *critical value* of rho. In our example, the critical values are 0.456 and 0.645 (for one-tailed tests) and 0.544 and 0.715 (for two-tailed tests) respectively.
- Rho is compared with the critical value: If r is equal to or greater than the critical value, rho is significant. In our example, the coefficient is 0.952. Since the absolute value of our coefficient is larger than each of the critical values, the value of r is significant.

For larger samples (over 30) and when the table of the critical values of r is not adequate, the value of t is employed to test the significance of r. t is evaluated in the context of the t-table. If the corresponding value in the table is equal to or smaller than the t-value, rho is significant at the specified level.

Evaluation

Spearman's rho is a very useful and effective measure of association. Nevertheless, it should be approached with caution when it includes (1) a large number of 'ties'; and (2) when scores are skewed.

Table 17.2 Critical values of Spearman's rho

	Significance level (one-tailed tests)		*Significance level (two-tailed tests)*	
N	*0.05*	*0.01*	*0.05*	*0.01*
5	0.900	1.000	1.000	-----
6	0.829	0.943	0.886	1.000
7	0.714	0.893	0.786	0.929
8	0.643	0.833	0.738	0.881
9	0.600	0.783	0.683	0.883
10	0.564	0.746	0.648	0.794
12	0.506	0.712	0.591	0.777
14	0.456	0.645	0.544	0.715
16	0.425	0.601	0.506	0.665
18	0.399	0.564	0.475	0.625
20	0.377	0.534	0.450	0.591
22	0.359	0.508	0.428	0.562
24	0.343	0.485	0.409	0.537
26	0.329	0.465	0.392	0.515
28	0.317	0.448	0.377	0.496
30	0.306	0.432	0.364	0.478

Source: E.G. Olds (1940), 'The 5% significance levels for sums of squares of rank differences and a correction', *Annals of Mathematical Statistics*, 20, pp. 117–18.

b Computing rho using SPSS

There are two methods of computing Spearman's rho. In both cases data must have been entered in the PC (method 1) before the methods can be used. They are shown below.

Method A

1 Choose **Statistics > Correlate > Bivariate**
2 Shift the variables to be correlated to the **Variable(s)** box
3 Click on the squares in front of **Spearman**
4 Click on **OK**

This procedure, employed using the data in Example B, will produce the results shown below. Note that the coefficient is the same as that computed manually and the rho is significant at the 0.000 level!

```
- - - SPEARMAN    CORRELATION    COEFFICIENTS - - -

TEST          .9516
              N(    14)
              Sig .000

              POPULARITY

(Coefficient / (Cases) / 2-tailed Significance)

" . " is printed if a coefficient cannot be computed
```

Method B

1. Choose **Statistics > Summarise > Crosstabs**
2. Shift one variable to the **Rows** box and the other to the **Columns** box
3. Click on **Statistics** at the bottom of the window
4. Click on the square box in front of **Correlations**
5. Click on **Continue** and then on **OK**

This will give you the values of Spearman's rho, together with ASE1, Val/ASE0 and approximate significance. Entering the data in Example B in the computer and following the steps introduced above produces the results shown below. The figures are self-explanatory. Note that information for Pearson's coefficient is also included in the output!

```
                                                Approximate
Statistic      Value    ASE1      Val/ASE0     Significance
----------     -----    -----     --------     ----------
Pearson's R    .95165   .01593    10.73151     .00000*4
Spearman       .95165   .02272    10.73151     .00000*4
Correlation

*4 VAL/ASEO is a t-value based on a normal
approximation, as is the significance

Number of Missing Observations: 0
```

5 Interval/ratio measures of association

a Pearson's product-moment correlation coefficient, *r*

This is the most common measure of association of variables scaled on an interval level. This measure considers not ranks of pairs but rather magnitudes of observations, and can be computed in three ways: (1) by means of computers, the easiest and most reliable method; (2) through the plotting method, by plotting and examining the least-squares line; and (3) by means of the following formula:

$$r = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{\{N\Sigma X^2 - (\Sigma X)^2\}\ \{N\Sigma Y^2 - (\Sigma Y)^2\}}} \qquad \textbf{(17.5)}$$

In the following discussion we shall attempt all three methods, beginning with the plotting method.

The plotting method

Computing the correlation using this method implies the construction of a scattergram, with the two variables being placed on the two axes of the graph. Each pair of scores is then plotted on the scattergram. The resulting shape and direction of the dots indicate the type, strength and direction of correlation. Widely spread dots indicating no direction, as in (b) in Figure 17.1, suggest that there is no correlation; dots congregating around a line beginning from the origin of the graph and making a 45° angle with the abscissa, as in (a) in Figure 17.1, indicate a positive correlation; dots displaying a form similar to that of (c) in the figure indicate a negative correlation. The more linear and the closer together the dots the stronger the relationship, and vice versa.

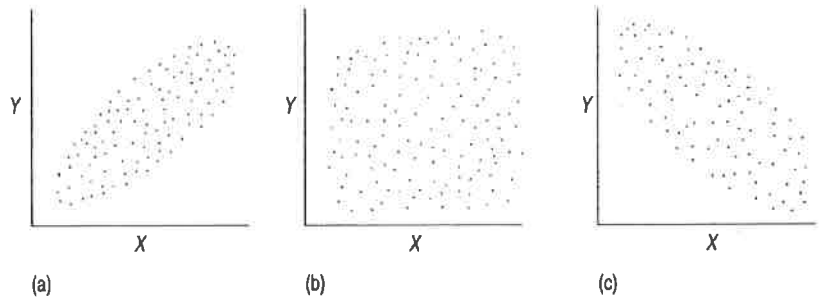

Figure 17.1 Scattergrams. (a) Strong positive relationship; (b) no relationship; (c) strong negative relationship

Plotting a scattergram using SPSS

Example C: A researcher wishes to examine the relationship between education of women and their degree of liberation. The hypothesis is that the degree of liberation is associated with their degree of education (educational status). To test this hypothesis 10 women are studied in terms of the variables in question. Education and liberation have been rated on a 15-point scale and each woman has been given a score for each of the variables according to her responses. The pair-scores obtained for education and liberation are: 2, 3; 4, 5; 5, 3; 6, 5; 7, 6; 8, 9; 9, 7; 11, 13; 13, 10; and 14, 11 respectively.

The steps to follow to obtain a scattergram are as follows:

1 Choose **Charts > Scatter**
2 Click on **Simple**
3 Click on **Define**
4 Transfer 'education' to the **Y Axis** box and 'liberation' to **X Axis** box
5 Click **OK**

These instructions will generate the output shown in Figure 17.2. The shape of the scores indicates a positive correlation. The fact that they do not form a line but are more or less spread around it suggests that the correlation is not perfect but still strong.

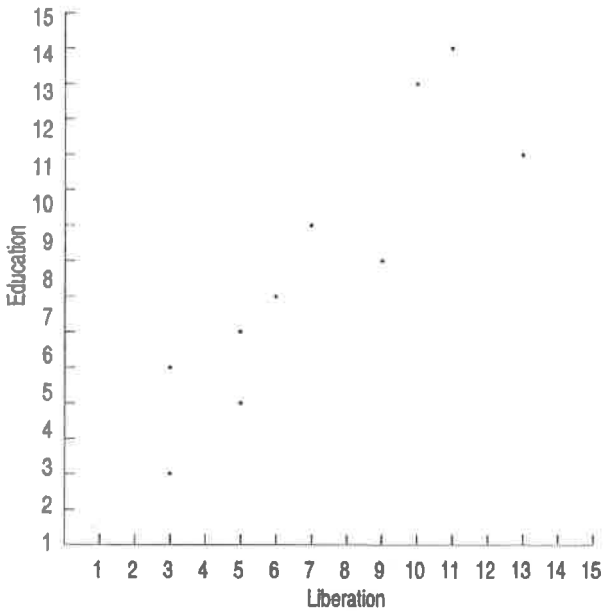

Figure 17.2 Scattergram showing a strong and positive correlation between education and liberation

Employing the formula

Computing the relationship between two variables by means of Formula (17.5) is more complicated than the plotting method; it offers, however, more information and is more accurate. To obtain the values of the elements of the formula, we first place the scores in a table, and construct columns for their squares and products. Employing the data in Example C, we obtain the results given in Table 17.3. The scores of the women have been placed in the first two columns. The squares and the products of the scores are shown in the remaining three columns.

Table 17.3 Correlation between liberalism and education of women

Education scores X	*Liberation scores* Y	X^2	Y^2	XY
2	3	4	9	6
4	5	16	25	20
5	3	25	9	15
6	5	36	25	30
7	6	49	36	42
8	9	64	81	72
9	7	81	49	63
11	13	121	169	143
13	10	169	100	130
14	11	196	121	154
$\Sigma X = 79$	$\Sigma Y = 72$	$\Sigma X^2 = 761$	$\Sigma Y^2 = 624$	$\Sigma XY = 675$

Substituting the values obtained through the table in Formula (17.5), the correlation coefficient for this example is:

$$r = \frac{10 \times 675 - 79 \times 72}{\sqrt{(10 \times 761 - 79^2)\,(10 \times 624 - 72^2)}} = \frac{1062}{\sqrt{1369 \times 1056}}$$

$$= \frac{1062}{\sqrt{1445664}} = \frac{1062}{1202.36} = 0.883$$

This coefficient is positive and high, indicating a strong and positive association between the variables. This verifies the trend depicted in the scattergram.

***Interpretation of* r**

As stated at the beginning of this chapter, a correlation coefficient indicates both the type of correlation and the strength of the relationship. A *positive correlation*, indicated by a coefficient having a positive sign, suggests that the two variables are associated in such a way that an increase/decrease in one variable is associated with an increase/decrease in the other. The common statement made by researchers in this case is 'the higher X the higher Y' or 'the lower X the lower Y'. In a more concrete example, if a positive correlation is identified the statement might be that 'the higher the degree of

dependence, the greater the chance for a person to hold a lower status in a system'; or 'the lower the class status, the lower the chance of a person attending tertiary institutions'.

A *negative correlation*, signified by a coefficient with a negative sign, suggests that there is an inverse relationship between the variables. In this case an increase in one variable is associated with a decrease in the other. The relevant statement in a study of feminism and traditionalism may be: 'the higher the degree of commitment to feminism, the lower the observance of traditional sex-differentiated household tasks'.

A *zero correlation* suggests that there is no systematic relationship between the two variables, and changes in one variable are not associated with changes in the other.

The *strength* of the relationship in a correlation is indicated by the position of the coefficient in its continuum.

The *range* of the coefficient is between –1 and +1. The correlation is generally considered to be:

- *very low* if the coefficient has a value under 0.20;
- *low* if the coefficient has a value between 0.21 and 0.40;
- *moderate* if the coefficient has a value between 0.41 and 0.70;
- *high* if the coefficient has a value between 0.71 and 0.91;
- *very high* if the coefficient is over 0.91.

This list offers a guide only to interpreting a coefficient value; it is not a rule to be followed. Nevertheless, most of the social scientists who employ *r* seem to interpret it as stated above.

***The significance of* r**

Does the identified relationship between the variables described by the correlation coefficient, *r*, correspond with the relationship that actually exists in the population, or is it a reflection of sampling or other methodological problems or procedures? In other words are the variables in the population related in the same way and to the same degree as they are found to be in the study and as they are described by the coefficient?

There are several ways of testing the significance of *r*; the simplest of all employs the critical values of *r* as summarised in relevant tables. The procedure is the same as that introduced above when we were discussing the significance of Spearman's rho. To test the significance of Pearson's *r* we compare its value with the critical value of *r* given in the table especially constructed for Pearson's *r*, as indicated by the relevant degrees of freedom and level of significance. An *r* is significant if it is equal to or greater than the relevant critical value of *r*.

b Computing Pearson's r using SPSS

There are two methods of computing Pearson's *r*. After entering the data in the computer we follow the following steps:

Method 1

1 Choose **Statistics > Correlate > Bivariate**
2 Transfer the variables to be correlated to the **Variable(s)** box
3 Click on the square in front of **Pearson**
4 Click on **OK**

Using the figures obtained in Example C, this procedure will produce the output shown below. It is clearly shown that the correlation coefficient for education and liberation is 0.8833, which is the same as the one obtained manually above. (The correlation coefficient between education and education, as well as liberation and liberation is obviously 1.) The correlation coefficient of 0.8833 is significant at the 0.001 level (P = 0.001).

```
- - - CORRELATION COEFFICIENTS - - -
             Education         Liberation

Education    1.0000            .8833
             (   10)           (  10)
             P = .             P = .001

Liberation   .8833             1.0000
             (  10)            (   10)
             P = .001          P = .

(Coefficient / (Cases) / 2-tailed Significance)

" . " is printed if a coefficient cannot be computed
```

Method 2

1 Choose **Statistics > Summarise > Crosstabs**
2 Transfer variable 1 to the **Rows** box
3 Transfer variable 2 to the **Columns** box
4 Click on **Statistics** at the bottom of the window
5 Click on the square box in front of **Correlations**
6 Click on **Continue** and then on **OK**

If we use the figures presented in Example C, the computer will display the following information. Note that the Spearman correlation is included also in the output!

```
- - Correlation coefficients - -

                                                    Approximate
Statistic      Value     ASE1      Val/ASE0       Significance

----------     -----     -----     --------       ----------

Pearson's R    .88326    .03525    5.32818        .00070
Spearman       .92075    .04401    6.67495        .00016
Correlation

Number of Missing Observations: 0
```

The output given here is somewhat different from the one produced through Method 1. Going through crosstabs produces the same figures, offers information for Pearson's *r and* Spearman's rho, gives details for ASE1 and Val/ASEO and is relatively easier to read. The important findings are, however, (a) the value of the coefficient (0.88326), which is very high (indicating a very strong, positive correlation), and (b) its significance, which is equally high.

c Coefficient of determination

It must be stressed that the coefficient of correlation measures the type and strength of linear correlations, but it does not necessarily imply a cause–effect relationship. Nevertheless, this measure can offer very useful information. If squared, for instance, the coefficient of correlation gives the coefficient of determination. This measure describes the common variance, that is, the degree of variability shared by the two variables. The symbol of this coefficient is r^2. This measure is very useful, primarily because it offers an index of predictability: it allows the researcher to make predictions about one variable if the degree of determination is known.

The coefficient of determination displays the proportion of variance in one variable that is explained by the other variable. If the coefficient is .0.81 (which is the equivalent of a coefficient of correlation of 0.9), 81 per cent of the variation is accounted for by the linear relationship with the other variable. The remaining variance cannot be explained by the other variable.

This remaining amount of variance not explained by the other variable is expressed in the *coefficient of non-determination*, and is the difference between 1 and the coefficient of determination. Obviously, the two coefficients must add up to 1.

For example, if the correlation coefficient is 0.88, the coefficient of determination is 0.77 and the coefficient of non-determination 0.23. Consequently a variable can explain 77 per cent of the variation in the values of the other. The remaining 23 per cent of variation is unique to both distributions and cannot be explained by the correlated variables.

B Regression and prediction

Regression is a method that allows social scientists to make predictions about the value of one variable (Y) if another variable (X) is known. This is an asymmetrical measure, since it allows one-direction predictions only. Predictions are made by means of the regression line, the definition of which is given by a formula containing the intercept and the slope of the line. The regression formula is as follows:

$$Y = a + bX \qquad \textbf{(17.6)}$$

where Y and X are variables, a and b constants. The constant a stands for the value of Y when X is zero and represents the Y-intercept. The constant b together with X (bX) represents the slope of the line.

The regression line is a straight line plotted on a scattergram. It is constructed by means of the method of least squares, which places the line in such a position that the squares of the vertical distances of the plots from the line are the smallest possible. The regression line is one of many but is the line of best fit. The line reduces the variance of all the distances and also passes through the mean of each variable. The position of the line on the scattergram depends on the values of a and b (which can be computed independently) and on the value of X, which can be arbitrarily defined by the researcher. Thus, to determine the value of Y, the following steps are taken:

1 b is computed by means of Formula (17.7).
2 a is computed by means of Formula (17.8).
3 These values are substituted in Formula (17.6).
4 Any two of the given values of X are substituted in Formula (17.6) where a and b are known, and the two corresponding Y-values are computed.
5 Two selected X-values and the corresponding Y-values are plotted on a scattergram.
6 The two plots are joined, which gives the regression line, thus allowing an estimation of the remaining values of Y by drawing vertical lines on the independent variable towards the regression line and horizontal lines across to the dependent variable. The point of crossing of this line is the value of the dependent variable that corresponds to the value of the independent variable that is on the foot of the line crossing the abscissa.

Computation of b is done by means of Formula (17.7)

$$b = \frac{N(\Sigma XY) - (\Sigma X)(\Sigma Y)}{N(\Sigma X^2) - (\Sigma X)^2} \tag{17.7}$$

From the computations in Table 17.3 we have $\Sigma X = 79$, $\Sigma Y = 72$, $\Sigma XY = 675$, $\Sigma X^2 = 761$, $N = 10$, $(\Sigma X)^2 = 6241$. Therefore:

$$b = \frac{10(675) - (79)(72)}{10 \times 761 - 6241} = \frac{6750 - 5688}{7610 - 6241} = \frac{1062}{1369} = 0.775$$

This means that for each unit change in X, Y changes by 0.775 units.

The value of a is obtained by means of Formula (17.8):

$$a = \frac{\Sigma Y - b(\Sigma Y)}{N} \tag{17.8}$$

Substituting our values in Formula (17.8):

$$a = \frac{72 - 0.775 \times 72}{10} = \frac{72 - 55.8}{10} = \frac{16.2}{10} = 1.62$$

This suggests that the regression line crosses the Y-line at 1.62.

The formula of the regression line now is:

$$Y = 1.62 + 0.775X$$

Taking $X_1 = 6$ and $X_2 = 11$, Y_1 and Y_2 become:

$$Y_1 = 1.62 + 0.775 \times 6 = 10.495;\ Y_2 = 1.62 + 0.775 \times 11 = 18.595$$

Knowing these values, we can now plot the values on the scattergram and after joining the two plots we can secure the regression line.

C Summary

The measures of association introduced in this chapter are very popular, very useful and effective, and also very easy to compute and interpret. They are used to measure the relationship between two variables, the strength of the relationship and the direction of the relationship. These measures are used also for nominal data and therefore in qualitative investigations.

Of these methods Pearson's r is the most popular. With regard to nominal data researchers seem to have their own personal preferences. While some do not use statistics at all, others seem to prefer ϕ and Cramer's V. The latter seems to be more popular now than before, especially with the expansion of the use of computers.

Overall, measures of association are a key tool of statistical analysis. In some areas they are used as the only statistical measure, but in most cases they are used in conjunction with measures of central tendency and dispersion, and more so with regression. In all cases, they offer useful information that assists researchers to identify correlations between variables and make relevant predictions. Unfortunately the scope of this text does not allow a more detailed discussion on regression. Information on this must be sought elsewhere.

Key concepts

Correlation
Spearman's rho
Pearson's *r*
Zero correlation
Prediction
Yule's *Q*
Gamma
Negative correlation
Regression
Sommer's *d*
Positive correlation
Coefficient of determination