# UNIT 8 CORRELATION ANALYSIS

**Structure**



## 8.0 OBJECTIVES

After going through this unit, you will be in a position to:

- plot scatter diagram;
- measure covariance between two variables;
- compute correlation coefficient;
- compute rank correlation coefficient; and
- determine whether two variables are correlated.

## 8.1 INTRODUCTION

In the previous unit we discussed the methods of presentation of bivariate data in the form of frequency distributions. In this unit we deal with the concept of correlation which measures the strength of relationship between two variables. When we compute measures of correlation from a set of bivariate data, our interest focuses on the *degree* and *direction* of the association between the variables.

In statistical studies with several variables, there are generally two types of problems. In some problems it is of interest to study how the variables are interrelated; such problems are tackled using *correlation* techniques. For instance, an economist may be interested in studying the relationship between the stock prices of various companies; for this he may use correlation techniques.

In other problems there is a variable $y$ of basic interest and the problem is to find out what information the other variable provides on $Y$, such problems are tackled using *regression* techniques. For instance, an economist may be interested in studying what factors determine the pay of an employed person and in particular, he may be interested in exploring what role the factors such as education, experience, market demand, etc. play in determining the pay. In the above situation he may use regression techniques to set up a prediction formula for pay based on education, experience, etc.

While correlation is dealt in the present unit, regression analysis will be covered in the next unit.

## 8.2 SCATTER DIAGRAM

We first illustrate how the relationship between two variables is studied. A teacher is interested in studying the relationship between the performance in Statistics and Economics of a class of 20 students. For this he compiles the scores on these subjects of the students in the last semester examination. Some data of this type are presented in Table 8.1.

**Table 8.1**
**Scores of 20 Students in Statistics and Economics**

| *Serial Number* | *Score in* | | *Serial Number* | *Score in* | |
|---|---|---|---|---|---|
| | *Statistics* | *Economics* | | *Statistics* | *Economics* |
| 1 | 82 | 64 | 11 | 76 | 58 |
| 2 | 70 | 40 | 12 | 76 | 66 |
| 3 | 34 | 35 | 13 | 92 | 72 |
| 4 | 80 | 48 | 14 | 72 | 46 |
| 5 | 66 | 54 | 15 | 64 | 44 |
| 6 | 84 | 56 | 16 | 86 | 76 |
| 7 | 74 | 62 | 17 | 84 | 52 |
| 8 | 84 | 66 | 18 | 60 | 40 |
| 9 | 60 | 52 | 19 | 82 | 60 |
| 10 | 86 | 82 | 20 | 90 | 60 |

A representation of data of this type on a graph is a useful device which will help us to understand the nature and form of the relationship between the two variables, whether there is a discernible relationship or not and if so whether it is linear or not. For this let us denote score in Economics by $X$ and the score in Statistics by $Y$ and plot the data of Table 8.1 on the x-y plane. It does not matter which is called $X$ and which $Y$ for this purpose. Such a plot is called *Scatter Plot* or *Scatter Diagram.* For data of Table 8.1 the scatter diagram is given in Fig. 8.1.

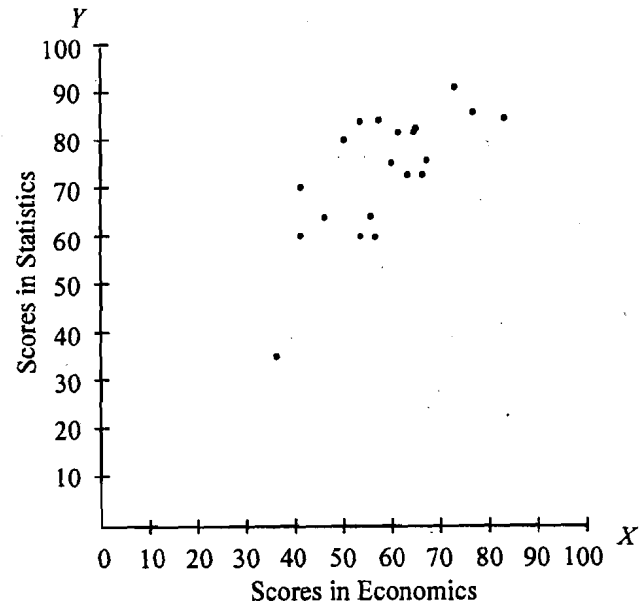


**Fig. 8.1: Scatter Diagram of Scores in Statistics and Economics**

An inspection of Table 8.1 and Fig. 8.1 shows that there is a *positive relationship* between $x$ and $y$. This means that larger values of $x$ are associated with larger values of $y$ and smaller values of $x$ with smaller values of $y$. Further, the points seem to lie scattered around both sides of a straight line. Thus it appears that a linear relationship exists between $x$ and y. However, this relationship is not *perfect* in the sense that there are deviations from such a relationship. It would indeed be useful to get a measure of the strength of this linear relationship.

## 8.3 COVARIANCE

In the case of a single variable we have learnt the concept of variance, which is defined as

$$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 \qquad \ldots(8.1)$$

In the above we use a subscript $x$ to specify that $\sigma_x^2$ represents the variance in $x$. In a similar manner we can represent $\sigma_y^2$ as the variance in $y$, and $\sigma_x$ and $\sigma_y$ as the standard deviation in $x$ and $y$ respectively.

As you know, variance measures the dispersion from mean. In the case of bivariate data we have to reach a single figure which will present the deviation in both the variables from their respective means. For this purpose we use a concept termed covariance, which is defined as follows:

$$\sigma_{xy} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) \qquad \ldots(8.2)$$

You may recall that standard deviation is always positive since it is defined as the positive square root of variance. In the case of covariance there are two terms $(X_i - \bar{X})$ and $(Y_i - \bar{Y})$ which represent the deviations in $x$ from $\bar{X}$ and $Y$ from $\bar{Y}$. Moreover, $(X_i - \bar{X})$ can be positive or negative depending on whether $x_i$ is less than or greater than $\bar{X}$. Similarly $(Y_i - \bar{Y})$ can be positive or negative. It is not necessary that whenever $(X_i - \bar{X})$ is positive $(Y_i - \bar{Y})$ will also be positive. Therefore, the product $(X_i - \bar{X})(Y_i - \bar{Y})$ can be either positive or negative. A positive value for $(X_i - \bar{X})(Y_i - \bar{Y})$ implies that whenever $X_i > \bar{X}$, we have $Y_i > \bar{Y}$. Thus a higher value of $x_i$ is associated with a relatively higher value in $y_i$. On the other hand, $(X_i - \bar{X})(Y_i - \bar{Y}) < 0$ implies that a lower value in $X_i$ is associated with a relatively higher value in $y_i$. When we sum it over all the observations and divide by the number of observations, we may obtain a negative or positive value. Therefore, covariance can assume both positive and negative values.

When covariance between $x$ and $y$ is negative ($\sigma_{xy} < 0$) we can say that the relationship could be inverse. Similarly, ($\sigma_{xy} > 0$) implies a positive relationship between $x$ and $y$. A major limitation of covariance is that it is not independent of unit of measurement. It means that if we change the unit of measurement of the variables we will get a different value for $\sigma_{xy}$.

The computation of $\sigma_{xy}$ as given in (8.2) often involves large numbers. Therefore, it is derived further as

$$\sigma_{xy} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) = \frac{1}{n}\sum_{i=1}^{n}(X_iY_i - \bar{X}Y_i - \bar{X}\bar{Y})$$

By further simplification we find that

$$\sigma_{xy} = \frac{1}{n}\sum_{i=1}^{n}X_iY_i - \frac{1}{n}\sum_{i=1}^{n}\bar{X}Y_i - \frac{1}{n}\sum_{i=1}^{n}X_i\bar{Y} + \frac{1}{n}\sum_{i=1}^{n}\bar{X}\bar{Y}$$

Since $\frac{1}{n}\sum_{i=1}^{n}\bar{X}Y_i = \frac{1}{n}\sum_{i=1}^{n}X_i\bar{Y} = \bar{X}\bar{Y}$ we have

$$\sigma_{xy} = \frac{1}{n}\sum_{i=1}^{n}X_iY_i - \bar{X}\bar{Y} \qquad \ldots(8.3)$$

## 8.4 CORRELATION COEFFICIENT

The task before us is to measure the linear relationship between $x$ and $y$. It is desirable to have this measure of strength of linear relationship independent of the scale chosen for measuring the variables. For instance, if we are measuring the relationship between height and weight, we should get the same measure whether height is measured in inches or centimetres and weight in pounds or kilograms. Similarly, if a variable is temperature, it should not matter whether it is recorded in Celsius or Fahrenheit. This can be achieved by standardising each variable, that is by considering $\frac{X-\bar{X}}{\sigma_x}$ and $\frac{Y-\bar{Y}}{\sigma_y}$ where $\bar{X}$ and $\bar{Y}$ are the means of $X$ and $Y$ respectively and $\sigma_x$ and $\sigma_y$ are standard deviations.

Let us denote these standardised variables by $u$ and $v$ respectively. Let us also use the notation $(X_i, Y_i)$ to denote the score $i^{th}$ student in Economics and Statistics respectively, $i$ ranging from 1 to $n$, the number of students, $n$ being 20 in our example. Similarly, let $(u_i, v_i)$ denote the standardised scores of $i^{th}$ student. Then recall the following formulae for mean and standard deviation:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n}\bar{X}_i; \sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n}(\bar{X}_i - \bar{X})^2;$$

$$Y = \frac{1}{n}\sum_{i=1}^{n}X_i; \sigma_y^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

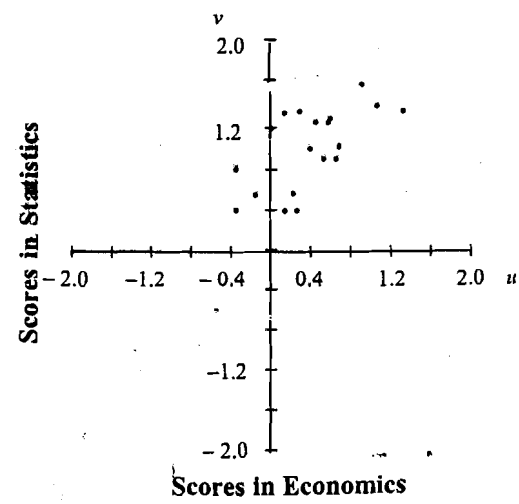


**Fig. 8.2: Scatter Diagram of Standardised Scores in Statistics and Economics**

Fig. 8.2 is the scatter diagram in terms of standardised variables $u$ and $v$. Let us observe that in this example there is a positive association between the two scores. The larger one score is, the larger the other score also is; the smaller one score is the smaller the other score is, on the whole. In view of this, most of the points are either in the *first quadrant* or in the *third quadrant*. The first quadrant represents the cases where both scores are above their respective means and third quadrant represents the cases where both scores are below their respective means. There are only a very few points in second and fourth quadrants, which represent the cases where one score is above its mean and the other is below its mean. Thus the product of the $u$, $v$ values is a suitable indicator of the strength of the relationship; this product is positive in the first and third quadrants and negative in the second and fourth. Thus the product of $u$, $v$ averaged over all the points may be considered to be suitable measure of the strength of linear relationship between $X$ and $Y$. This measure is called the *correlation coefficient* between $X$ and $Y$ and is usually denoted by $r_{xy}$ or simply by $r$, when it is clear what $x$ and $y$ in the context are. This is also called the *Pearson's Product-Moment Correlation Coefficient* to distinguish it from other types of correlation coefficients.

Thus the formula for $r$ is

$$r = \frac{1}{n}\sum_{i=1}^{n} u_i v_i \qquad \ldots(8.4)$$

If we substitute the variables $x$ and $y$ in (8.4) above

$$r = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{X_i - \bar{X}}{\sigma_x}\right)\left(\frac{Y_i - \bar{Y}}{\sigma_y}\right) = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sigma_x \sigma_y}$$

In the above expression, the term

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$$

is the *covariance* between $x$ and $y$ ($\sigma_{xy}$).

Thus the formula for correlation coefficient is

$$r = \frac{\sigma_{xy}}{\sigma_x \times \sigma_y} \qquad \ldots (8.5)$$

Incorporating the formulae for $\bar{x}, \bar{y}$, $\sigma_x$, $\sigma_y$ it becomes

$$r = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{x})(Y_i - \bar{Y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} \qquad \ldots (8.6)$$

or alternatively

$$r = \frac{n\sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{\sqrt{\left[n\sum_{i=1}^{n} X^2{}_i - \left(\sum_{i=1}^{n} X_i\right)^2\right]\left[n\sum_{i=1}^{n} Y^2{}_i - \left(\sum_{i=1}^{n} Y_i\right)^2\right]}} \qquad \ldots (8.7)$$

Let us go back to the data given in Table 8.1 and work out the value of $r$. You can use any of the formulae (8.4), (8.5), (8.6) or (8.7) to get the value of $r$. Since

all the above formulae are derived from the same concept we obtain the same value for $r$ whichever formulae we use. For the data set in Table 8.1 we have calculated it by using (8.4) and (8.7). We construct Table 8.2 for this purpose.

**Table 8.2: Calculation of Correlation Coefficient**

| *Observation No.* | $X$ | $Y$ | $X^2$ | $Y^2$ | $XY$ |
|---|---|---|---|---|---|
| 1 | 82 | 64 | 6724 | 4096 | 5248 |
| 2 | 70 | 40 | 4900 | 1600 | 2800 |
| 3 | 34 | 35 | 1156 | 1225 | 1190 |
| 4 | 80 | 48 | 6400 | 2304 | 3840 |
| 5 | 66 | 54 | 4356 | 2916 | 3564 |
| 6 | 84 | 56 | 7056 | 3136 | 4704 |
| 7 | 74 | 62 | 5476 | 3844 | 4588 |
| 8 | 84 | 66 | 7056 | 4356 | 5544 |
| 9 | 60 | 52 | 3600 | 2704 | 3120 |
| 10 | 86 | 82 | 7396 | 6724 | 7052 |
| 11 | 76 | 58 | 5776 | 3364 | 4408 |
| 12 | 76 | 66 | 5776 | 4356 | 5016 |
| 13 | 92 | 72 | 8464 | 5184 | 6624 |
| 14 | 72 | 46 | 5184 | 2116 | 3312 |
| 15 | 64 | 44 | 4096 | 1936 | 2816 |
| 16 | 86 | 76 | 7396 | 5776 | 6536 |
| 17 | 84 | 52 | 7056 | 2704 | 4368 |
| 18 | 60 | 40 | 3600 | 1600 | 2400 |
| 19 | 82 | 60 | 6724 | 3600 | 4920 |
| 20 | 90 | 60 | 8100 | 3600 | 5400 |
| **Total** | **1502** | **1133** | **116292** | **67141** | **87450** |

From Table 8.2 we note that

$$\sum_{i=1}^{20} X_i = 1502; \bar{X} = 75.1;$$

$$\sum_{i=1}^{20} Y_i = 1133; \bar{Y} = 56.65;$$

$$\sum_{i=1}^{20} X_i^2 = 116292; \sigma_x^2 = \frac{1}{20}\left[116292 - \frac{1502^2}{20}\right] = 174.59; \sigma_x = 13.21;$$

$$\sum_{i=1}^{20} Y_i^2 = 67141; \sigma_y^2 = \frac{1}{20}\left[67141 - \frac{1133^2}{20}\right] = 147.83; \sigma_y = 12.16;$$

$$\sum X_i Y_i = 87450; \sigma_{xy} = \frac{1}{20}\left[87450 - \frac{1502 \times 1133}{20}\right] = 118.09$$

Thus using formula 8.4, we have

$$r = \frac{118.09}{13.21 \times 12.16} = 0.735$$

Now let us use the formula 8.7. We have

$$r = \frac{20 \times 87450 - 1502 \times 1133}{\sqrt{(20 \times 116292 - 1502^2)(20 \times 67141 - 1133^2)}} = 0.735$$

Thus we see that both the formulae provide the same value of the correlation coefficient $r$. You can check yourself that the same value of $r$ is obtained by using the formula (8.5). For this purpose you will need values on

$\sum (X_i - \bar{X})^2$, $\sum (Y_i - \bar{Y})^2$ and $\sum (X_i - \bar{X})(Y_i - \bar{Y})$

Hence you can have five columns on

$(X_i - \bar{X})$, $(Y_i - \bar{Y})$, $(X_i - \bar{X})^2$, $(Y_i - \bar{Y})^2$ and $(X_i - \bar{X})(Y_i - \bar{Y})$ in a table and find the totals.

## 8.5 INTERPRETATION OF CORRELATION COEFFICIENT

It is a mathematical fact that the value of $r$ as defined above lies between –1 and + 1. The extreme values of –1 and + 1 are obtained only in situations where there is a *perfect linear relationship* between $X$ and $Y$. The value –1 is obtained when this relationship is perfectly negative (i.e., inverse) and + 1 when this is perfect positive (i.e., direct). The value of 0 is obtained when there is no linear relationship between $x$ and $y$.

We can make some guess work about the sign and degree of the correlation coefficient from the scatter diagram. Fig. 8.3 gives example of scatter diagrams for various values of $r$. Fig. 8.3(a) is a scatter diagram for the case $r = 0$; here there is no *linear relationship* between $x$ and $y$. Fig. 8.3(b) is also an example of scatter diagram for the case $r = 0$; here there is discernible relationship between $X$ and $Y$ but it is not of the linear type. Here, initially, $Y$ increases with $X$ but later $Y$ decreases as $X$ increases resulting in a definitive quadratic relationship. But the correlation coefficient in this case is zero. Thus the correlation coefficient is only a measure of linear relationship. This sort of scatter diagram is obtained, if we plot, for instance, body weight ($Y$) of individuals against their age ($X$). Fig. 8.3(c) is an example of a scatter diagram where there is a perfect positive linear relationship between $X$ and $Y$. We get this sort of scatter diagram if we plot, for instance, height of individuals in inches ($X$) against their heights in centimeters ($Y$); in that case $Y = 2.54X$, which is a deterministic and perfect linear relationship. Figures 8.3(d) to 8.3(k) are scatter diagrams for other values of $r$. From these scatter diagrams we get an idea of the nature of relationship and associated values of $r$.

From these it would seem that a value of 0.81 indicates a fair degree of linear relationship between scores in Statistics and Economics of these candidates. Such a quantification of relationship or association between variables is helpful for natural and social scientists to understand the phenomena they are investigating and explore these phenomena further. In an example of this sort, an educational psychologist may compute correlation coefficients between scores in various subjects and by further statistical analysis of the correlation coefficients and using psychological techniques may be able to form a theory as to what mental and other faculties are involved in making students good in various disciplines.

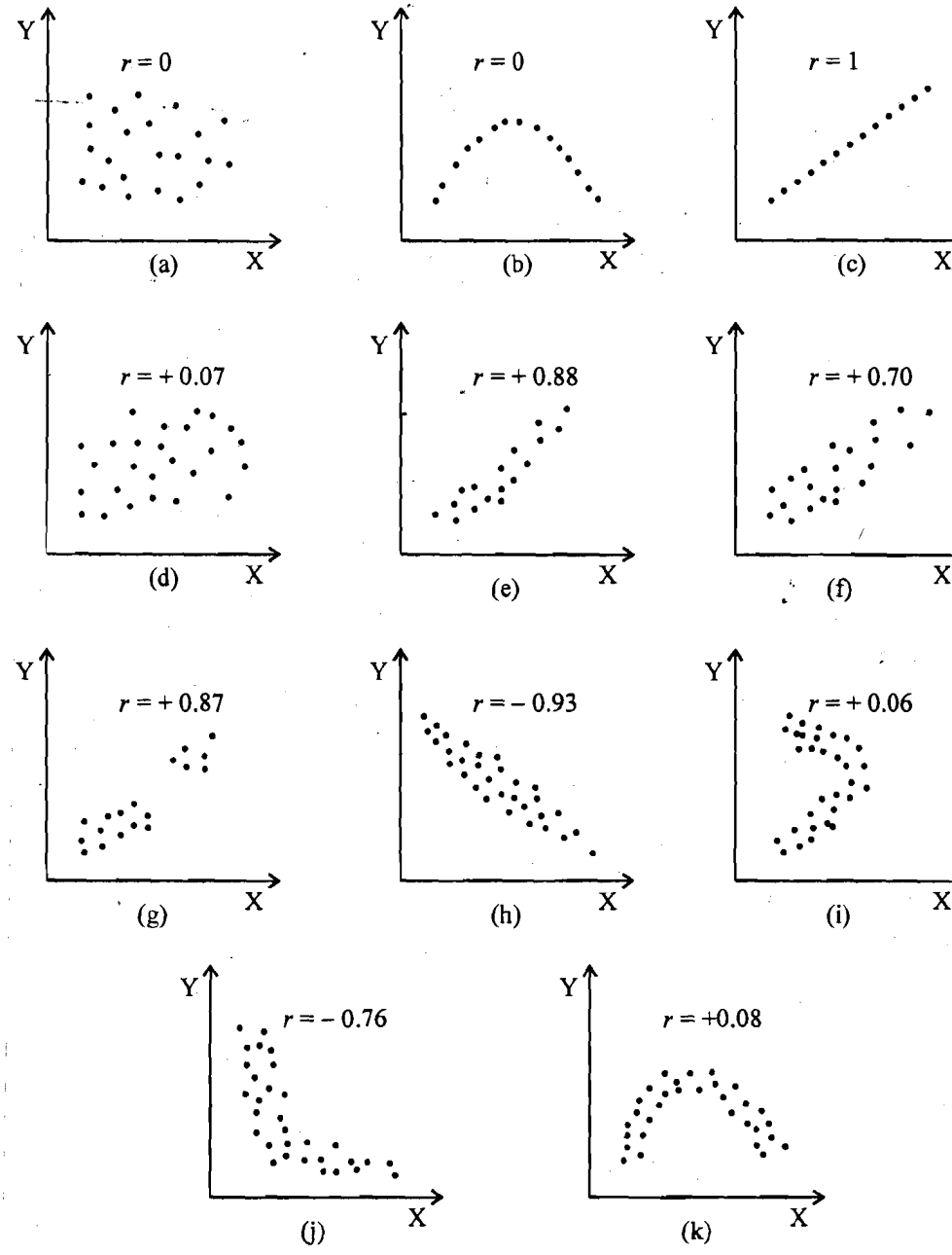


**Fig. 8.3: Scatter Plots for Various Values of Correlation Coefficient**

## Remember that

- Correlation coefficient shows the linear relationship between $X$ and $Y$. Thus, even if there is a strong non-linear relationship between $X$ and $Y$, correlation coefficient may be low.
- Correlation coefficient is independent of scale and origin. If we subtract some constant from one (or both) of the variables, correlation coefficient will remain unchanged. Similarly, if we divide one (or both) of the variables by some constant, correlation coefficient will not change.
- Correlation coefficient varies between $-1$ and $+1$. This means $r$ cannot be smaller than $-1$ and cannot be greater than $+1$.

The existence of a linear relationship between two variables is not to be interpreted to mean a cause-effect relationship between the two. For instance, if you work out the correlation between family expenditures on petrol and chocolates, you may find it to be fairly high indicating a fair degree of linear relationship. However, this

Both are luxury items and richer families can afford them and poorer ones cannot. Thus the high correlation here is caused by the high correlation of each of the variables with family income. To consider another example, suppose for each of the last twenty years, you work out the average height of an Indian and the average time per week an Indian watches television; you are likely to find a positive correlation. This does not, however, imply that watching television increases one's height or that taller people tend to watch television longer. Both these variables have an increasing trend over time and this is reflected in the high correlation. This kind of correlation between two variables is caused by the effect of a third variable on each of them rather than a direct linear cause-effect relationship between them is called *spurious correlation*.

Another aspect of the computation of correlation coefficient that we should be aware of is that the correlation coefficient like any other quantity computed from sample, varies from sample to sample and these sample fluctuations should be taken into account in making use of the computed coefficient. We do not discuss these techniques here.

Whether the presence of a linear relationship between two variables and hence a high correlation between them is genuine or spurious, such a situation is helpful to *predict* one variable from the other. We examine these prediction techniques in Unit 9.

**Check Your Progress 1**

1) Calculate $r$ from the following given results :

   $n = 10;\ \sum X = 125, \sum X^2 = 1585, \sum Y = 80, \sum Y^2 = 650, \sum XY = 1007.$

   .....

   .....

   .....

   .....

   .....

2) Calculate the coefficient of correlation for the ages of husband and wife :

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| *Age of husband* | : | 23 | 27 | 28 | 29 | 30 | 31 | 33 | 35 | 36 | 39 |
| *Age of wife* | : | 18 | 22 | 23 | 24 | 25 | 26 | 28 | 29 | 30 | 32 |

   .....

   .....

   .....

   .....

   .....

3) Specimens of similarly treated alloy steel containing various percentages of nickel are tested for toughness with the following results :

   Toughness (arbitrary units):

   47 50 52 52 54 56 58 59 60 60 62 64 65 66

   Percentage of Nickel :

   2.7 2.7 2.8 2.8 2.9 3.2 3.2 3.3 3.4 3.5 3.6 3.7 3.7 3.8

   Find the correlation coefficient between toughness and nickel content and comment on the result.

.....

.....

.....

.....

.....

.....

4) Determine the correlation coefficient between $x$ and $y$ —

| | | | | | | |
|---|---|---|---|---|---|---|
| $x$ : | 5 | 7 | 9 | 11 | 13 | 15 |
| $y$ : | 1.7 | 2.4 | 2.8 | 3.4 | 3.7 | 4.4 |

.....

.....

.....

.....

.....

.....

5) The following table gives the saving bank deposits in billions of dollars and strikes and lock-outs, in thousands, over a number of years. Compute the correlation coefficient and comment on the result.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Saving deposits : | 5.1 | 5.4 | 5.5 | 5.9 | 6.4 | 6.0 | 7.2 |
| Strikes and lock-outs : | 3.8 | 4.4 | 3.3 | 3.6 | 3.3 | 2.3 | 1.0 |

.....

.....

.....

.....

.....

.....

## 8.6 RANK CORRELATION COEFFICIENT

The Pearson's product moment correlation coefficient (or simply, the correlation coefficient) described above is suitable if both the variables involved are measurable (numerical) and the relationship between the variables is linear. However, there are situations where variables are not numerical but various items can be ranked according to the characteristics (i.e., ordinal). Sometimes even when the original variables are measurable, they are converted into ranks and a measure of association is computed. Consider for instance the situation when two examiners are asked to judge ten candidates on the basis of an oral examination. In this case, it may be difficult to assign scores to candidates, but the examiners find it reasonably easy to rank the candidates in order of merit. Before using the results, it may be advisable to find out if rankings are in reasonable concordance. For this, a measure of association between the ranks assigned by the two examiners may be computed. The Karl Pearson's correlation coefficient is not suitable in this situation. One may use the following measure called *Spearman's Rank Correlation Coefficient* for this purpose.

**Table 8.3: Ranks of 10 Candidates by two Examiners**

| S.No. | Rank given by | | Difference | |
|---|---|---|---|---|
| | Examiner I | Examiner II | $D_i$ | $D_i^2$ |
| 1 | 6.0 | 6.5 | - 0.5 | 0.25 |
| 2 | 2.0 | 3.0 | - 1.0 | 1.00 |
| 3 | 8.5 | 6.5 | 2.0 | 4.00 |
| 4 | 1.0 | 1.0 | 0.0 | 0.00 |
| 5 | 10.0 | 2.0 | 8.0 | 64.00 |
| 6 | 3.0 | 4.0 | - 1.0 | 1.00 |
| 7 | 8.5 | 9.5 | - 1.0 | 1.00 |
| 8 | 4.0 | 5.0 | - 1.0 | 1.00 |
| 9 | 5.0 | 8.0 | - 3.0 | 9.00 |
| 10 | 7.0 | 9.5 | - 2.5 | 6.25 |
| | | | $\sum D_i = 0$ | $\sum D_1^2 = 87.50$ |

Let us consider the data of Table 8.3. Here there are some ties; the tied cases are given the same rank in such a way that their total is the same as when there are no tie. For example, when there are two cases with rank 6, each is given a rank of 6.5 and there is no case with rank either 6 or 7. Similarly, if there are three cases with rank 5, then each is given a rank of 6 and there is no case with rank 5 or 7. Spearman's rank correlation coefficient, called Spearman's Rho, denoted by $\rho$, is based on the difference $D_i$ (i for i$^{th}$ observation) between the two rankings. If the two rankings completely coincide, then $D_i$ is zero for every case. The larger the value of $D_i$, the greater is the difference between the two rankings and smaller is the association. Thus the association can be measured by considering the magnitudes of $D_i$. Since the sum of $D_i$ is always zero, to find a single index on the basis of $D_i$ values, we should remove the sign of $D_i$ and consider only the magnitude. In Spearman's $\rho$, this is done by taking $D_i^2$.

However, the largeness or smallness of $\sum_{i=1}^{n} D_i^2$, where $n$ is the number of cases, will depend on $n$. Thus, in order to be able to interpret this value, we could create a ratio by dividing this sum by the largest possible value, which depends only on $n$, which is $\frac{n(n^2-1)}{6}$. However, $\frac{6\times\sum_{i=1}^{n} D_i^2}{n(n^2-1)}$ is zero for perfect association and 2 for lack of association, i. e., perfect negative association, while we would like it to be other way around. So we subtract this ratio from 1. Thus

$$\rho = 1 - \frac{6\times\sum_{i=1}^{n} D_i^2}{n(n^2-1)} \qquad \text{..... (8.8)}$$

is defined as Spearman's rank correlation.

Let us calculate the value of $\rho$ from the data given in Table 8.3.

$$\rho = 1 - \frac{6\times 87.5}{10(10^2-1)} = 1 - \frac{525}{990} = 1 - 0.53 = 0.47.$$

Like Karl Pearson's coefficient of correlation the Spearman's rank correlation has a value + 1 for perfect matching of ranks, –1 for perfect mismatching of ranks and 0 for the lack of relation between the ranks.

There are other measures of association suitable for use when the variables are of nominal, ordinal and other types. We do not discuss them here.

**Check Your Progress 2**

1) In a contest, two judges ranked eight candidates A, B, C, D, E, F, G and H in order of their preference, as shown in the following table. Find the rank correlation coefficient.

| | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| First Judge | 5 | 2 | 8 | 1 | 4 | 6 | 3 | 7 |
| Second Judge | 4 | 5 | 7 | 3 | 2 | 8 | 1 | 6 |

.....

.....

.....

.....

.....

.....

2) Compute the correlation coefficient of the following ranks of a group of students in two examinations. What conclusion do you draw from the result?

| Roll Nos. | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Rank in B.Com. Exam. | 1 | 5 | 8 | 6 | 7 | 4 | 2 | 3 | 9 | 10 |
| Rank in M. Com Exam. | 2 | 1 | 5 | 7 | 6 | 3 | 4 | 8 | 10 | 9 |

.....

.....

.....

.....

.....

.....

3) Ten competitors in a musical contest were ranked by 3 judges A, B and C in the following order :

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Ranks by A : | 1 | 6 | 5 | 10 | 3 | 2 | 4 | 9 | 7 | 8 |
| Ranks by B : | 3 | 5 | 8 | 4 | 7 | 10 | 2 | 1 | 6 | 9 |
| Ranks by C : | 6 | 4 | 9 | 8 | 1 | 2 | 3 | 10 | 5 | 7 |

Using Rank Correlation method, discuss which pair of judges has the nearest approach to common liking in music.

.....

.....

.....

.....

.....

.....

4) Ten students obtained the following marks in Mathematics and Statistics. Calculate the rank correlation coefficient.

| Student (Roll No.) | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Marks in Mathematics | 78 | 36 | 98 | 25 | 75 | 82 | 90 | 62 | 65 | 39 |
| Marks in Statistics | 84 | 51 | 91 | 60 | 68 | 62 | 86 | 58 | 53 | 47 |

.....

.....

.....

.....

.....

## 8.7 LET US SUM UP

In this unit you have learnt about scatter diagram and covariance. Also you learnt about the coefficient of correlation and the coefficient of rank correlation that will indicate the closeness of the linear association or correlation between two variables. However, correlation does not imply a cause-effect relationship.

## 8.8 KEY WORDS

**Correlation Analysis** : Refers to a measure of association between two random variables. If two random variables have been such that when one gets changed the other will do so in a related manner, they are regarded to be correlated. Variables which are independent are not correlated. The correlation coefficient is a number between –1 and + 1. It could be calculated from a number of pairs of observations which are normally referred to as points $(X, Y)$. A coefficient of 1 implies perfect positive correlation, –1 perfect negative correlation and 0 no correlation.

**Covariance** : The first product moment of two variables about their means is called covariance. The formula for the calculation of covariance is $\frac{1}{n}\sum(X_i - \bar{X})(Y_i - \bar{Y})$ or $\frac{1}{n}\left(\sum X_i Y_i - \frac{\sum X_i \sum Y_i}{n}\right)$ where $X$ and $Y$ are corresponding values of each variable and $n$ is the number of observations.

**Rank Correlation Coefficient** : There happen to be many occasions when it may not be convenient, economic or even possible to give values to variables. However, various items can be ranked. In such cases, a rank correlation coefficient may be used.

**Scatter Diagram** : A diagram showing the joint variation of two variables $X$ and $Y$. Each member is represented by a point whose coordinates, on ordinary rectangular axes, are the values of the variables. A set of $n$ observations thus provides $n$ points on the diagram and the scatter or clustering of the points exhibits the relationship between $X$ and $Y$.

## 8.9 SOME USEFUL BOOKS

Nagar, A.L. and R.K. Das, 1989 : *Basic Statistics*, Oxford University Press, Delhi.

Goon, A.M., M.K. Gupta and B. Dasgupta, 1987 : *Basic Statistics*, The World Press Pvt. Ltd., Calcutta.

## 8.10 ANSWERS/HINTS TO CHECK YOUR PROGRESS EXERCISES

**Check Your Progress 1**

1) + 0.47
2) + 0.996
3) + 0.98
4) + 0.995
5) – 0.84

**Check Your Progress 2**

1) 2/3
2) + 0.64
3) – 0.21, + 0.64, – 0.30
4) + 0.82