UNIT 16 CORRELATION : IMPORTANCE AND INTERPRETATION

Structure

16.1 INTRODUCTION

In Units 13 & 14, we have discussed those statistical measures that we use for a single variable i.e. the distribution relating to one quantitative variable. Now, we shall study the problem of describing the degree of simultaneous variation of two or more variables. The data in which we secure measures of one variable for each individual is called a univariate distribution. If we have pairs of measures on two variables of each individual, the joint presentation of the two sets of scores is called a bivariate distribution.

You may want to study the possibility of a relationship between two variables and see what kind of relationship exists. For example if you want to study the relationship between height and weight - whether the change in one will bring a change in other or not. Or if you want to find the relationship between hours of study and achievement, sex and enrolment etc., you can do so by finding correlation between them. In simple words, we can say that the statistical tool which helps us to study the relationships between two or more than two variables is called correlation. According to Tuttle, "*Correlation is an analysis of the co-variation between two or more variables*". With the change in one variable, the other related variable changes is known as correlation. You know that as the child grows, the height as well as weight increases. Similarly, if a person has more height than the other, the person is likely to have more weight than the latter and we can say height and weight are positively correlated. You will study about correlation in greater details in this unit.

16.2 OBJECTIVES

After reading this Unit, you will be able to:

- define correlation;
- define co-efficient of correlation;
- recognize various types of correlation;
- calculate the co-efficient of correlation according to nature of scores and their distribution;
- interpret the results obtained i.e. interpret the Co-efficient of correlation;
- take necessary precautions in interpreting the co-efficient of correlation; and
- discuss the importance of correlation and its coefficient.

16.3 THE CONCEPT OF CORRELATION

Let us take an example of the scores of 5 students in mathematics and physics. What pattern do you find in the data? You may notice that in general those students who score well in mathematics also get high scores in physics. Those who are average in mathematics get just average scores in physics and those who are poor in mathematics get low scores in physics. In short, there is a tendency for students to score at par on both variables. Performance on the two variables is related; in other words the two variables are related, hence co-vary.

If the change in one variable appears to be accompanied by a change in the other variable, the two variables are said to be co-related and this inter-variation is called correlation.

You must have understood from the above discussion that correlation is a statistical technique that helps us know whether there is any relationship or not between any two pairs of variables and how strong is this relationship. This relationship is perfect or not can be determined by further analysis of observations. This is a measure of simultaneous variation of variables described by an integer. This can also be called the *measure of relationship*. If you take measure of every subject (individual) of your sample on two different variables e.g. if you take the marks of every student in Urdu and Mathematics in a Class VI of your school and find out the relationship between these marks of Urdu and Mathematics, relationship which is found out is called correlation. Such a data where we study two variables at a time is called bi-variate data. In this kind of data scores of one individual in one subject could be paired with the scores of the same individual in another subject. In this case scores of Urdu can be paired with scores of Mathematics. This relationship is important as the scores of the individual may change with the scores of other or the relationship could be due to some common factor between them.

You will find that when one of these variables increases or decreases on observation, the other paired variable is also changed proportionately. However, it may not be taken as change of one variable with the manipulation of other variable. When you are using correlation, you should not assume that a change in one variable causes a change in the other. It is not a cause and effect relationship. Even a high degree of correlation does not necessarily mean that a relationship of cause and effect exists between the variables. Thus the *correlation is simply a degree of relationship between the variable of a bi-variate data*. This correlation can tell you just how much of the variation in one variable is related to the other paired or correlated variable. Like all statistical techniques correlation is only appropriate for certain kind of data. Correlation works only with the quantifiable data in which numbers are meaningful, usually in quantities of some sort. It cannot be used for purely non-quantifiable data like categorical data such as gender, socio economic status, goodness, etc. Correlation is just a co-variation and does not manifest any kind of causation of functional relationship.

16.4 CO-EFFICIENT OF CORRELATION

The degree of association or the degree of relationship between two variables is measured quantitatively in the form of an index which is termed as co-efficient of correlation.

Co-efficient of correlation is a single number that tells us to what extent the two variables are related and to what extent the variations in one variable changes with the variations in the other.

This coefficient of correlation is determined to find the relationships because it is simple to understand and convenient to express. It is a constant.

Symbol of co-efficient of correlation

The coefficient of correlation is always symbolized either by r or r (Rho). The notion 'r' is known as product moment correlation co-efficient or Karl Pearson's Coefficient of Correlation. The symbol 'r' (Rho) is known as Rank Difference Correlation Coefficient or Spearman's Rank Correlation Coefficient. Sometimes it is also written as r_{xy} which means coefficient of correlation between *x* and *y* variables.

16.4.1 Maximum Range of Values of Co-Efficient of Correlation

The measurement of correlation between two variables results in a maximum value that ranges from –1 to +1, through zero. The ±1 values denote ***perfect coefficient of correlation.***

Check Your Progress 1

Note: a) Write your answers in the space given below.

b) Compare your answers with those given at the end of the Unit.

1. What do you mean by the term correlation?

   ....

   ....

   ....

2. What is Co-efficient of Correlation?

   ....

   ....

   ....

3. What can be the possible range of Co-efficient of correlation?

   ....

   ....

   ....

16.5 TYPES OF CORRELATION

Correlation can be classified in several different ways. Three most common ways of classification can be:

(i) Positive and negative Correlation

(ii) Simple, partial and multiple Correlation

(iii) Linear and non-linear Correlation

Before we come to the classified types of correlation, many times you will come across two more terms about correlation i.e. (i) no correlation and (ii) perfect correlation. We will study about them first, before we come to types of correlation.

(i) ***No Correlation-*** when there is no relationship observed or established between two variables i.e. the change in one variable cannot establish the change of other variable in a definite pattern. It may first time increase, second time decrease and at third time may not change. In such cases, we can say no definite change or Zero correlation. Most of the times one variable changes and other remains constant. Sometimes, it may even be a chance that there is a relationship.

(ii) ***Perfect Correlation –*** When *c*hange in one variable brings similar change in the other variable under paired variate, it can be said to have perfect correlation, that is, it has correlation value = one. In such cases the variable changes in a constant proportion. Specifically, when the coefficient correlation is '+1', it is called as perfect positive correlation and when it is '-1', it is called as perfect negative correlation.

Now let us come to classified correlation.

16.5.1 Positive and Negative Correlation

As the name clearly suggests they are directional associations or correlations which means the two variables of a bi-variate data vary in the same direction or the different direction. You will understand it better with the following explanations about them one by one.

***(i) Positive Correlation*:** If both the paired variables under study vary in the same direction with any change, then, they are said to be positively correlated. It means that if one of them increases, the other related variables also increases and if one of them decreases, the other also decreases. Both of them have similar direction of change. For example, height is positively correlated to weight. Similarly, height and weight are positively correlated to age. In education, the motivation or intelligence may be positively correlated to achievement. Other examples could be reading or study habits and achievement, attendance and achievement, etc. This positive correlation can further be explained with the following example.

Example:

In a class the marks of ten students were studied in Urdu and History to know whether the marks in Urdu have some relationship with the marks obtained in History or to know the correlation between marks of two subjects. We got the marks of ten students as under:

Student	Marks in Urdu	Marks in History
1	25	35
2	20	30
3	35	45
4	28	39
5	45	50
6	36	47
7	50	65
8	18	25
9	12	20
10	40	48

Now you can see that the marks of Urdu and History achieved by all the ten students are moving in the same direction i.e. increasing/decreasing in the same order. Thus, we say achievement of students in Urdu and History is positively correlated.

(ii) Negative Correlation: If the two paired variables under study vary in opposite direction with any change, then they are said to be negatively correlated. It means that if one of them increases, the other related variable will decrease and if one of the related variable decreases, the other will increase. Both the variables have opposite direction of change. For example, the supply and the prices of commodities. You know, if the commodity is available in abundance, the prices fall and if it is in shortage, the prices will rise. The same could be the case with absenteeism and achievement. More the students absent themselves from the school, the lesser will be the achievement and vice versa. These are common examples which must have clarified the concept of negative correlation. You could give some more examples of negative correlation at your end. Negative correlation is not very common in our studies but it may appear, needing explanation in many of your studies.

16.5.2 Simple, Multiple and Partial Correlation

This classification is based on the number of variables being studied.

(i) ***Simple Correlation:*** In simple correlation only two variables are studied. For example, in most of bi-variate studies like the correlation between height and weight, correlation between the hours of study and achievement, correlation between the method of teaching and achievement of students etc., we have to find simple correlation.

(ii) ***Multiple Correlation:*** In multiple correlation three or more variables are studied simultaneously. For example, when you want to study method of teaching, sex, social-economic status and achievement and find the correlations among them simultaneously, you have to see the effect of interactions among the variables along with correlation and these are studied through multiple correlations.

(iii) ***Partial Correlation:*** Partial correlation is also used in study of correlation of more than two variables, but in this case it is supposed that we want to study only two variables which are said to be influencing each other, the effect of other influencing variables being kept constant. For example, if we want to study the effect of hours of study and achievement but the third variable intelligence is all the time influencing, then its effect has to be kept constant or partialled out.

16.5.3 Linear and Curvilinear Correlation

(i) ***Linear Correlation:*** Most of the correlations you will deal with in your research studies will be linear correlations and in common sense correlation is generally understood by the term simple linear correlation only. If the amount of change in one variable tends to bear a constant ratio to the amount of change in the other related variable, then the correlation is said to be linear correlation.

Example:

The scores obtained by students of Class VIII on two variables say intelligence and achievement are as under:

Intelligence (IQ)	*Achievement*
50	25
60	30
70	35
80	40
90	45

Now it is clear from the above data that the ratio of change between one variable to the other is the same. If such a data of the variable quantities is plotted on a graph, you will get a straight line.

(ii) Curvilinear Correlation (Non – Linear Correlation): Now it would be clear that if the amount of change in one variable does not bear a constant ratio with the amount of change in the other related variable, the correlation is said to be curvy – linear Correlation.

Example:

The following data in hours of study and the achievement scores of five students are given below :

Hours of Study	Achievement Score
1	60
2	65
3	72
4	78
5	83

We will find a non-linear relationship between the two variables. If we plot a graph of such a data, we would get a curve.

Normally, it is assumed that a linear relationship exists between the two variables under study because the techniques of correlation for curvi-linear correlation are complicated and do not bring much difference.

Check Your Progress 2

Note: a) Write your answers in the space given below.

b) Compare your answers with those given at the end of the Unit.

4. Suggest examples of the following:

a) Perfect positive correlation.

....

....

....

b) Zero correlation.

....

....

....

c) Curvilinear correlation.

....

....

....

16.6 METHODS OF COMPUTING CO-EFFICIENT OF CORRELATION (UNGROUPED DATA)

In case of ungrouped data of bivariate distribution, the following three methods are used to compute the value of co-efficient of correlation.

1. Rank Difference Co-efficient of Correlation or Spearman's Rank Order Co-efficient of Correlation.

2. Pearson's Product Moment Co-efficient of Correlation.

3. Pearson's Product Moment Co-efficient of Correlation from Scatter Diagram.

16.6.1 Spearman's Rank Difference Coefficient of Correlation (ρ)

Rank Difference Coefficient of Correlation is used only when the data is in ordinal scale and it is calculated with the help of ranks. It may be defined as the correlation between the ranks assigned to the individuals in two characters. This is the most widely used because of its easiness to compute. This is denoted by the symbol rho (ρ) and is computed as follows:

1. First of all the data being discrete is converted into Rank form. For ranking the data, we have to start ranking from the highest to the lowest. Highest score is give rank value as 1 and the next is given the rank 2 and so on. We begin with giving ranks to first set of scores followed by giving ranks to the next set of scores separately. Suppose there are two individuals getting the same score, then their positional ranks are added and an average rank is given to each individual. For example, if 6^{th} and 7^{th} highest ranked individuals are getting the same score, then their ranks are added and average is taken which is 6+7/2 = 6.5. So each of them is given a rank of 6.5. Similarly, if three individuals get the same score, then their positional ranks are added and average is taken. After completing the ranking of first set of data, the second set of data is also awarded ranks in the same manner. Although too many such ties will affect the size of correlation coefficient; but usually there are not so many ties to justify the formula that is available for correcting these ties.

2. After giving ranks to both the set of variables, find the difference between these ranks of two set of scores. This is calculated as an absolute score because the sign of these differences does not effect and is of no importance as the differences have to be squared in the next step.

3. Square each of these differences in the next column.

4. Add all these difference squares.

5. Find the Rank Order Coefficient of Correlation by using the following formula.

$$(\rho) = 1 - \frac{6\sum D^2}{N(N^2-1)}$$

Where (ρ) = Rank Order Correlation Coefficient

N = Total number of set of scores in either set or Number of pairs of scores.

D = Differences of ranks of each pair

Let us understand this with an example to illustrate it further.

Example:

Suppose you have administered a Personality test and a Mental Ability test to some students of your school and you are interested in finding the degree of association between personality and mental ability. Let us assume that you got the following scores on the two tests. These scores are taken for convenience only.

Scores of students on Personality Test and Mental Ability Test

S.No. of Student	Score on Personality	Score on Mental Ability
1	60	60
2	54	68
3	53	40
4	49	52
5	49	51
6	47	38
7	46	51
8	45	32
9	45	39
10	45	41
11	43	50
12	41	48
13	39	36
14	38	48
15	32	40
16	32	46
17	30	37

To find the rank order correlation coefficient among these two variables, let us find the ranks and form another table for computations.

Table 16.1 : Table for Computation of Rank Order Coefficient of Correlation

Student	Personality Score	Mental Ability Score	Rank Personality (R_1)	Rank Mental Ability (R_2)	D = R_1-R_2	D^2
1	60	60	1	2	1	1
2	54	68	2	1	1	1
3	53	40	3	11.5	8.5	72.25
4	49	52	4.5	3	1.5	2.25
5	49	51	4.5	4.5	0	0
6	47	38	6	14	8	64
7	46	51	7	4.5	2.5	6.25
8	45	32	9	17	8	64
9	45	39	9	13	4	16
10	45	41	9	10	1	1
11	43	50	11	6	5	25
12	41	48	12	7.5	4.5	20.25
13	39	36	13	16	3	9
14	38	48	14	7.5	6.5	42.25
15	32	40	15.5	11.5	4	16
16	32	46	15.5	9	6.5	42.25
17	30	37	17	15	2	4
						ΣD^2 = 386.5

Now in this Table:

(i) First find the ranks of Personality scores. In column P these ranks have been calculated like 60 is the highest score which has been given rank 1. 54,53 being the next scores have been given ranks 2,3 respectively whereas, 42 is repeated two times. So on averaging 4 and 5, both of these are given the rank 4.5 each. Similarly, 32 is repeated two times and is given the rank of 15.5 each on averaging 15 and 16 ranks. So all the personality scores are given the ranks from 1 to 17.

(ii) Similarly, find the ranks of Mental Ability scores. In column Q, these ranks have been calculated like, score 68 is given rank one, 60 and 52 have been given ranks 2 and 3 respectively whereas, 51 appears two times and has been given the ranks of 4.5 each. Similarly, 48 appears two times and is given rank of 7.5 each being an average of 7 and 8. Similarly 40 is also repeated and is given the rank of 11.5 each, being an average of 11 and 12. Thus, scores are awarded the ranks from 1 to 17.

(iii) Now find the difference of Ranks of each of the score pairs of Personality and Mental ability in each row. Thus, the difference of ranks of each of these ranks is column R = Column Q – Column P without sign.

(iv) In next column find the squares of the difference of ranks, i.e. S column is the square of the Difference. Column S is (column R)^2. For example, in Table you have got D^2 = 1,1,72.25, 2.25 etc. for the first four students. Hence Column S = Column R^2.

(v) Apply the values in the formula to find rank order coefficient of correlation:

$$(\rho) = 1 - \frac{6\sum D^2}{N(N^2-1)}$$

In this example

$\sum D^2$ = 386.50

N = 17

N^2 = 289

$$(\rho) = 1 - \frac{6 \times 386.5}{17\ (289 - 1)}$$

$$(\rho) = 1 - \frac{2319.0}{4896}$$

$$= 1 - 0.47 = 0.53$$

16.6.2 Pearson's Product Moment Coefficient of Correlation (r)

Pearson's Coefficient of correlation is determined when the data of variables consists of scores paired in some meaningful way. There are several ways of calculating 'r' in such cases.

(a) Deviation Score Method

In this method, coefficient of correlation is calculated on the basis of deviations of the scores from their means and is defined by the following formula.

$$r_{xy} = 1 - \frac{\sum xy}{nS_x S_y}$$

Where r_{xy} = Correlation coefficient between x and y.

x = Deviation of first variable scores from their mean ($X - \overline{X}$)

y = Deviation of second variable scores from the mean ($Y - \overline{Y}$)

n = Total number of pairs of scores

S_x = Standard Deviation of first variable scores

S_y = Standard Deviation of second Variable Scores.

Example

Suppose you have the scores of ten students in Physics and Mathematics as follows and are interested in finding the coefficient of correlation between them.

Table 16.2 : Scores of students in Physics and Mathematics

Student	**Scores in Physics**	$X = X - \overline{X}$	X^2	**Scores in Maths**	$Y = Y - \overline{Y}$	Y^2
1	37	-2.1	4.41	75	-2.6	6.76
2	41	1.9	3.61	78	0.4	0.16
3	48	8.9	79.21	88	10.4	108.16
4	32	-7.1	50.41	80	2.4	5.76
5	36	-3.1	9.61	78	0.4	0.16
6	39	-0.1	0.01	71	-6.6	43.56
7	40	0.9	0.81	75	-2.6	6.76
8	45	5.9	34.81	83	5.4	29.16
9	39	-0.1	0.01	74	-3.6	12.96
10	34	-5.1	26.01	74	-3.6	12.96
	$\sum X=391$		$\sum X^2=208.9$	$\sum Y=776$		$\sum Y^2=226.4$

$$\overline{X} = \sum X / N = 39.1 \qquad \overline{Y} = \sum Y/N = 77.6$$

$$S_x = \sqrt{\frac{\Sigma X^2}{N}} = 4.57$$

$$S_y = \sqrt{\frac{\Sigma Y^2}{N}} = 4.76$$

In this Method,

i. Find means of both the scores i.e. scores of Physics and Mathematics as calculated above are 39.1 and 77.6

ii. Then find deviation of each score from means as (x) and (y).

iii. Then find deviation squares in each case as calculated below and sum them up. These values in this example are: $\sum X^2=208.9$ and $\sum X^2=226.4$

Table 16.3 : Table for Calculation of Product Moment Coefficient of correlation by Deviation Score Method

Student	Phy (x)	Maths (y)	x Deviation	y Deviation	xy	x^2	y^2
1	37	75	-2.1	–2.6	5.46	4.41	6.76
2	41	78	1.9	+0.4	0.76	3.61	0.16
3	48	88	8.9	+10.4	92.56	79.21	108.16
4	32	80	-7.1	+2.4	-17.04	50.41	5.76
5	36	78	-3.1	+0.4	-1.24	9.61	0.16
6	39	71	-0.1	–6.6	0.66	0.01	43.56
7	40	75	0.9	–2.6	-2.34	0.81	6.76
8	45	83	5.9	+5.4	31.86	34.81	29.16
9	39	74	-0.1	–3.6	0.36	0.01	12.96
10	34	74	-5.1	–3.6	18.36	26.01	12.96
	Σx = **391** **M=39.1**	Σy = **776** **M=77.6**		**Total**	**Σxy= 129.4**	**ΣX^2= 208.9**	**ΣY^2= 226.4**

iv. Then find standard deviation of both the group of scores i.e. scores of Physics and Scores of Mathematics as calculated above are 4.57 and 4.76. These are done after finding x^2 and y^2 and their sum.

v. Then find 'x × y' and their sum as done above, Σxy=129.4.

vi. Then find coefficient of correlation.

$$r_{xy} = \frac{\Sigma xy}{nS_x S_y}$$

$$= 129.4 \ / \ (10 \times 4.57 \times 4.76)$$

$$= 129.4 \ / \ 217.53$$

$$= 0.59$$

(b) Computation of 'r' by Raw Score Method

In the raw score method if you substitute the values of deviations in raw score form and value of standard deviations in raw score form you will get the following formula for the calculation of coefficient of correlation from raw scores :

$$r = \frac{N\Sigma XY - (\Sigma X)\Sigma Y}{\sqrt{\left(N\Sigma X^2 - (\Sigma X)^2\right)\left(N\Sigma Y^2 - (\Sigma Y)^2\right)}}$$

Let us take an example for formula and find out the coefficient of correlation. In following example, raw scores and the steps for calculating 'r' by Raw score method have been provided.

Example

Table 16.4: Table for Computation of 'r' by Raw Score Method

S. No.	Score in language	Score in History (x)	X^2 (y)	Y^2	XY
1	20	12	400	144	240
2	18	16	324	256	288
3	16	10	256	100	160
4	15	14	225	196	210
5	14	12	196	144	168
6	12	10	144	100	120
7	12	9	144	81	108
8	10	2	100	4	20
9	8	7	64	49	56
10	5	8	25	64	40
	$\Sigma X=130$	$\Sigma Y=100$	$\Sigma X^2=1878$	$\Sigma Y^2=1138$	$\Sigma XY=1410$

(i) First find sum of all raw scores of language and History ΣX and ΣY Where, $\Sigma X = 130$ and $\Sigma Y = 100$ as above

(ii) Then find squares of the scores of language and History X^2 and Y^2 which are calculated in column no. 4 and 5.

(iii) Find sum of squares of raw scores of language and History as above

$\Sigma X^2 = 1878$ and $\Sigma Y^2 = 1138$

(iv) Then find multiple of raw scores of each pair (XY)

(v) Find sum of multiple of raw scores of each pair as above.

$\Sigma XY = 1410$

(vi) Now find Coefficient of Correlation by substituting each value.

$$r = \frac{N\Sigma XY-(\Sigma X)(\Sigma Y)}{\sqrt{\left(N\Sigma X^2-(\Sigma X)^2\right)\left(N\Sigma Y^2-(\Sigma Y)^2\right)}}$$

$$r = \frac{10(1410)-(130)(100)}{\sqrt{\left[10(1878)-(130)^2\right]\left[10(1138)-(100)^2\right]}}$$

$$r = \frac{14100-13000}{\sqrt{(18780-16900)(11380-10000)}}$$

$$= \frac{1100}{\sqrt{(1880)(1380)}}$$

$= 1100 \ / \ 1610.7$

$= 0.68$

(c) Standard Score or Z – Score Method

In this method, we neither use raw scores nor their deviations from the mean. In stead we use the standard scores to find the coefficient of correlation. Thus in this case first we transform both the variables into their standard scores and then find coefficient of correlation by the following formula:

$$r = \frac{\Sigma(Z_x Z_y)}{N}$$

The following steps are employed in these calculations:

(i) Find Mean of both the scores sets (X and Y).

(ii) Find deviation of each score from their respective mean which will help you in finding standard deviation of both the groups of scores X and Y.

(iii) Now find Standard Deviation (S_x and S_y) of both the sets of scores.

(iv) Then transform both the group scores into their standard score (Z scores) by the formula which you know

$$Z_x = \frac{x}{S_x} \quad \text{and} \quad Z_y = \frac{y}{S_y}$$

(v) Then find multiple of both the Z scores (Z_x) (Z_y) and sum it up

(vi) Find correlation coefficient by substituting the above value in the formula

$$r = \frac{\Sigma(Z_x Z_y)}{N}$$

Let us illustrate it with the same example as in the previous case of scores of language and scores of history.

Example

Table 16.5 : Table for the calculation of Coefficient of correlation by Standard Score of Z – Score Method

S.No.	X	Y	X	Y	Z_x	Z_y	$Z_x Z_y$
1	20	12	7	2	1.61	0.54	0.8694
2	18	16	5	6	1.15	1.62	1.8637
3	16	10	3	0	0.69	0.00	0
4	15	14	2	4	0.46	1.08	0.4968
5	14	12	1	2	0.23	0.54	0.1242

6	12	10	–1	0	–0.23	0.00	0
7	12	9	–1	–1	–0.23	–0.27	0.0621
8	10	2	–3	–8	–0.69	–2.16	1.4904
9	8	7	–5	–3	–1.15	–0.81	0.9315
10	5	8	–8	–2	–1.84	–0.54	0.9936
	$\Sigma x=130$	$\Sigma y=100$	$S_x=4.34$				$\Sigma_{ZxZy}=6.8317$
	$\overline{X} = 13$	$\overline{Y} = 10$	$S_y=3.71$				

$$r = \frac{6.8317}{10}$$

$$= 0.68$$

The above calculation is similar to the one calculated by other method. It does not change for the same set of data. Thus we can say that the *coefficient of correlation calculated from deviation method, raw score method or the standard scores method will remain unchanged* because all these methods are derived from the common formula of deviation in product moment coefficient of correlation.

16.6.3 Pearson's Product Moment Co-efficient of Correlation (Scattergram)

Scattergram as you know is based on plotting the scores of two sets of paired data on a graph or diagram in a table. By doing so, it gives us a visual picture of the bi variate data formed by the pairs of related scores. For drawing scatter gram, the unit of each variable is decided on the basis of its range and accordingly, score is represented on a graph. The advantage of this method is that it gives an 'eye check' on the number of factors that may influence the value of coefficient, which you can use when you interpret your observations. It can also give you a rough check of your coefficient of correlation, if you have already calculated by any of above three techniques. For example, if you find that scatter plot or scatter diagram corresponds to some value r = +0.72, whereas your calculated value by other means comes out to be 0.42. Then you would be aware of the discrepancy and can begin a recheck of your calculations. In computer, it is very easy to plot scatter diagrams with the help of simple statistical softwares. You may be curious to know the construction of a scatter diagram. Let us understand the construction of a scatter diagram in the following simple steps:

(1) Make a table of your set of paired scores.

(2) Decide the class interval for each set of scores and make a frequency table for each set by making tallies. These class intervals may be between 10 and 20 but 10 to 15 are most convenient classes.

    Now you have both the scores in class interval forms. It is not necessary to have exactly the same number of class interval for both the variables, nor is necessary to use the same class interval width in both variables. This selection of width and number of classes depends upon the kind of data or the scores therein.

(3) Next we tally the scores pair by pair so that two scores meet at a single point i.e. we enter our tally mark in the cell where these two class intervals meet at one single place. Then we take the next pair and again we mark a tally for the pair in the same way. Thus we make tally for each pair of scores.

(4) Inspect your own tally marks once again to see that all tallies are marked correctly.

(5) Now you add your tallies on either side of your table to get frequencies and total of each of these sums will be the same. This is at the bottom of the table for the sum of columns and on right hand side for the sum of rows frequencies. These are row A and column P in your table.

(6) Now, we write moments on both sides of these sum total columns and sum total rows. In these moments the highest is at the top of column and designated as y and is at the right side in a row which is designated as x. These are deviations from arbitrary reference points. So you get second column and row of the bottom as moments i.e. Y' and X'. These are row B and column Q of table

(7) In the next column and next row multiply each moment by their frequencies. So in next row you will get fx' and in next column on right you will get fy'. These are row C and column R of your table.

(8) Now in next column and next row multiply each of previous row and column with moment again. So in next row you will get $\Sigma fx'^2$ and in next column you will get fy'^2. These are row D and column S of your table.

(9) In next column you find sum of last two rows C and D i.e. find $\Sigma f x'$ and $\Sigma fx'^2$.

(10) Similarly, sum your column R and S i.e. find $\Sigma fx'$ and $\Sigma fx'^2$.

(11) Now take the product of the deviations from the two arbitrary reference points. Thus, it is product of moments as you have seen above that these deviations from arbitrary reference points are called moments. Here the arbitrary points are not the class interval but it is the mid point of that class interval. These products of moments are calculated by multiplying the y' with frequency of x' (total frequency).

(12) Now find the sum of these product moments i.e. x' y'.

(13) Find the coefficient of correlation by the following formula.

$$r = \frac{\Sigma x'y' - \left[(\Sigma fx')(\Sigma fy')/N\right]}{\sqrt{\left\{\Sigma fx'^2 - \left[(\Sigma fx')^2/N\right]\right\}\left\{\Sigma fy'^2 - \left[(\Sigma fy')^2/N\right]\right\}}}$$

Let us understand this scatter diagram with the help of an example.

Example

Suppose, you have scored 35 students of your class on the mental ability before teaching and examining them on their achievement after six months. You got the scores and you are interested in finding the correlation between their mental ability scores and the achievement scores with the help of scattergram. The scores of the students in Mental Ability Test and Achievement Test are as in table 16.6.

Table 16.6 : Scores of the students of a class on Mental Ability and Achievement

S.No. of the student	Mental Ability Score	Achievement Score
1.	80	61
2.	95	28
3.	94	74
4.	101	46
5.	105	44
6.	89	38
7.	106	72
8.	92	41
9.	105	49
10.	107	69
11.	111	82
12.	114	76
13.	83	39
14.	112	64
15.	91	77
16.	88	50
17.	105	55
18.	106	59
19.	105	86
20.	80	63
21.	85	31
22.	93	57
23.	85	70
24.	92	43
25.	90	70
26.	89	54
27.	85	51
28.	96	58
29.	85	63
30.	98	73
31.	101	71
32.	106	76
33.	112	76
34.	93	59
35	109	72

Table for Scattergram of Pearson's Product Moment Correlation

X-Axis – Mental Ability

Y-Axis – Achievement

		25-29	30-34	35-39	40-44	45-49	50-54	55-59	60-64	65-69	70-74	75-79	80-84	85-89	P	Q	R	S	T
															f	y'	$f\,y'$	$f\,y'^2$	$y'\,x'$
114-116												$_{24}1^{24}$			1	6	6	36	24
111-113									$_{5}1^{5}$			$_{20}1^{20}$	$_{25}1^{25}$		3	5	15	75	50
108-111											$_{12}1^{12}$				1	4	4	16	12
105-107					$_{-9}1^{-9}$	$_{-6}1^{-6}$		$_{0}2^{0}$		$_{6}1^{6}$	$_{9}1^{9}$	$_{12}1^{12}$		$_{18}1^{18}$	8	3	24	72	30
102-104															0	2	0	0	0
99-101						$_{-2}1^{-2}$					$_{3}1^{3}$				2	1	2	2	1
96-98								$_{0}1^{0}$			$_{0}1^{0}$				2	0	0	0	0
93-95		$_{6}1^{6}$						$_{0}2^{0}$			$_{-3}1^{-3}$				4	–1	–4	4	3
90-92					$_{12}2^{6}$						$_{6}1^{6}$	$_{-8}1^{-8}$			4	–2	–8	16	10
87-89				$_{12}1^{12}$			$_{6}2^{3}$								3	–3	–9	27	18
84-86			$_{20}1^{20}$				$_{4}1^{4}$		$_{-4}1^{-4}$		$_{-12}2^{-12}$				4	–4	–16	64	8
81-83				$_{20}1^{20}$											1	–5	–5	25	20
78-80									$_{-12}2^{-6}$						2	–6	–12	72	–12
A	f	1	1	2	3	2	3	5	4	1	7	4	1	1	N=35				
B	x'	–6	–5	–4	–3	–2	–1	0	1	2	3	4	5	6					
C	$f\,x'$	–6	–5	–8	–9	–4	–3	0	4	2	21	16	5	6	$\sum fx' = 33$				
D	$f\,x'^2$	36	25	32	27	8	3	0	4	4	63	64	25	36	$\sum x'^2 = 327$				
T	$x'y'$	6	20	32	3	–8	10	0	–11	6	15	48	25	18	$\sum x'y' = 164$				

Let us understand the calculations in this scattergram table through the steps we had studied.

1. Paired set of scores are there in the form of mental ability scores and achievement scores.

2. The classes of mental ability are formed from 78 to 116 with class intervals of 4 as shown on Y axis.

3. The classes of achievement scores are formed from 25 to 90 with class intervals of 5 each. They are shown on X-axis.

4. Now scores are tallied as per their value on both the axis as shown in between the Table. We enter first score pair (80, 61) in their respective column and row. Similarly next (95, 28) pair is placed as tally in the respective position and so on.

5. Recheck that they have marked tally in correct place.

6. Add the tally on either side in their respective columns and rows to get frequency as shown in row A and column P. These sums of all frequencies of A and P will be same i.e. total N. This will recheck the tallies.

7. Now mark moments as the deviation from the middle class interval. These are there in column Q and row B. The deviation of starting class is 'O' and above the class, deviations are +1, +2, +3 .... and below the class –1, –2, –3.... etc.

8. Now multiply each moment by their respective frequency on either side i.e. columns and rows to get f y' and f x' i.e. column R and row C in this table. It is product of

   P × Q = R and A × B = C

9. Now again multiply them by moment on each column and in each row to get S and D column and row respectively in this table. It is the product of R × Q = S column and

   B × C = D row. This is fx'^2.

10. Now sum all the rows and columns to get their sum (Σ) i.e. get Σf x', $\Sigma fx'^2$, Σy' and $\Sigma y'^2$.

11. Now find the product of the moments, i.e., the product of the x' and y' in T column.

12. Add the product moments which are there in column T i.e. you get Σx'y'.

13. Now find the correlation 'r' by the following formula

$$r = \frac{\Sigma x'y' - \left[(\Sigma fx')(\Sigma fy')/N\right]}{\sqrt{\left\{\Sigma fx'^2 - \left[(\Sigma fx')^2/N\right]\right\}\left\{\Sigma fy'^2 - \left[(\Sigma fy')^2/N\right]\right\}}}$$

$$r = \frac{164 - (33 \times 6 / 35)}{\sqrt{327 - (33^2/35) \times 409 - 6^2/35)}}$$

$$= \frac{164 - 5.66}{\sqrt{(327 - 31.11) \times (409 - 1.03)}}$$

$$= \frac{158.34}{\sqrt{295.89 \times 407.97}}$$

$$= \frac{158.34}{347.44} = 0.46 \text{ (Approx)}$$

Check Your Progress 3

Note: a) Write your answers in the space given below.

b) Compare your answers with those given at the end of the Unit

5. Compute correlation of the marks of 10 students in rank order method.

   Scores (Math) : 30, 37, 70, 45, 69. 55, 40, 45, 55, 63

   Scores (English) : 33, 40, 64, 39, 52, 40, 36, 44, 52, 60

   ....

   ....

   ....

6. Compute correlation of the marks of 10 students in Pearson's product moment deviation score method.

   Scores (Sc.) : 35, 37, 60, 40, 62. 55, 40, 43, 55, 50

   Scores (Soc. Sc.) : 33, 40, 61, 39, 52, 50, 36, 44, 50, 60

   ....

   ....

   ....

16.7 INTERPRETATION OF THE CO-EFFICIENT OF CORRELATION

Merely computation of correlation does not have any meaning until and unless we determine how large the coefficient need to be in order to be significant, and what does correlation tell us about the data? Coefficient of correlation is positive or negative depending upon the direction as explained earlier in positive and negative correlations. Thus coefficient of correlation will have values between +1 and –1. When the value of obtained correlation is closer to 1 or –1, then the variables are more closely related and if it is close to 0, it means that there is no relationship between these variables. The coefficient of correlation tells us the intensity of linear relationship in quantitative terms. In real life situations perfectly correlated variables are rare.

Full interpretation of correlation 'r' depends upon the circumstances, one of which is the size of the sample. All that can really be said is that when estimating the value of one variable from the value of other; the higher is the value of coefficient of correlation; the better will be the estimate. The closeness of relationship is not proportional to the value of "r". If the value of this "r" is 0.85, it does not necessarily indicate that a relationship is five times as close as one having "r" 0.17. We can only say that they have different kind of relationships. It will represent the proportion of common variation in two variables both in strength and magnitude.

Coefficients of correlation are not even additive. An average of correlation coefficients in a number of samples does not represent an average correlation in all those samples. Thus we should not add the coefficients of correlation to find correlation of two or more samples. But we should first convert them into additive measures and find correlations after that.

The coefficient of correlation is interpreted in verbal description. The rule of thumb for interpreting the size of a correlation coefficient is presented below:

Table 16.7: Correlation coefficient and its interpretation

Size of Correlation	Interpretation
±1	Perfect Positive/negative Correlation
± .90 to ± .99	Very High Positive/Negative Correlation
± .70 to ± .90	High Positive/Negative Correlation
± .50 to ± .70	Moderate Positive/Negative Correlation
± .30 to ± .50	Low Positive/Negative Correlation
± .10 to ± .30	Very low Positive/Negative Correlation
± .00 to ± .10	Markedly Low and Negligible Positive/Negative Correlation

16.8 MISINTERPRETATION OF THE COEFFICIENT OF CORRELATION

Sometimes, we misinterpret the value of coefficient of correlation and establish the cause and effect relationship, i.e. one variable causing the variation in the other variable. Actually we cannot interpret in this way unless we have sound logical base. Correlation coefficient gives us, a quantitative determination of the degree of relationship between two variables X and Y, not information as to the nature of association between the two variables.

- The degree of relationship or association is not ordinarily interpretable in direct proportions to the magnitude of the coefficient of correlations. For example, with 'r' = 0.25 we cannot say one variable has one fourth association with the other variable, or with "r" = 0.50 we can not interpret that one variable has half relationship with the other variable. This cannot be objectively considered. In general, change of 0.10 point in coefficient may have greater consequences when applied to coefficients having a high value than if the same is applied to lower values.

- The strength of relationship between two variables depends, among other things, on the nature of measurement of the two variables as well as on the kind of subjects being studied. Thus, it is not possible to speak of the correlation between two variables without taking these factors into consideration.

- Sometimes, you may find a high correlation between two variables, which may tempt you to think that they are substantially correlated. But in actual practice they may not have any relationship. You have to logically see whether such an association could exist or not. There must be some common cause for the association between the two. Mere high coefficient is insufficient to claim a relation between the two variables.

- Although a straight line graph between two variables is an indication of a straight correlation or high correlation between the two variables, but sometimes, it may not be so and in such cases it may be mere chance or it is misleading.

- The correlation coefficient is affected by the range of talent (variability) characterizing the measurements of the two variables. In general, the smaller is the range of talent in two variables; the lower will be coefficient of correlation, other things being equal. For example, in any school the correlation between the aptitude test score and marks obtained by the students may be +0.50. But if you find the coefficient of correlation of the students having high scores, the coefficient may be much lower.

- Coefficient of correlation like any other statistics is also dependent upon sampling fluctuations. It is also affected by sampling variations. Depending upon the characteristics of a particular sample, the obtained coefficient may be higher or lower than it will be in a different sample.

- Correlation works with quantifiable data in which the number is meaningful as we have already explained above. It cannot be used for purely categorical data such as gender, colours etc.

- Squaring coefficient of correlation makes it easier to understand and gives the percent of the variation in one variable that is related to the other variable and makes it easier to understand. If r = 0.5; then square becomes 0.25 and we can say there is 25% variation in one variable with respect to the other variable under study. This squared correlation coefficient (r^2) is the proportion of variance in y variable that can be accounted for by knowing X variable. One important property of variance is that it may be portioned into separate additive parts. Like total correlation of a sample variable X with variable Y where X is total human beings. So X if divided into Male and Female, the total "r" can also be divided into two parts r_1 and r_2 where $r = r_1 + r_2$.

16.9 FACTORS INFLUENCING THE SIZE OF THE CORRELATION COEFFICIENT

It will be helpful for you to know that the following factors influence the size of the coefficient of correlation and can some times lead to misinterpretation:

1. The size of "r" is very much dependent upon the variability of measured values in the correlated sample. The greater the variability, the higher will be the correlation, everything else being equal.

2. The size of "r" is altered, when an investigator selects an extreme group of subjects in order to compare these groups with respect to certain

behaviour. “r” obtained from the combined data of extreme groups would be larger than the “r” obtained from a random sample of the same group.

3. Addition or dropping the extreme cases from the group can lead to change in the size of “r”. Addition of the extreme case may increase the size of correlation, while dropping the extreme cases will lower the value or “r”.

16.10 IMPORTANCE AND USE OF CORRELATION IN EDUCATIONAL MEASUREMENT AND EVALUATION

Correlation is important in many areas of measurement and evaluation on education. It is of particular importance in the study of similarities and individual differences. With this common property of correlation in mind, correlation is used in the following contexts:

(i) For Determining Reliability:

Suppose there are two examiners examining the same class on some parameters or in some examination. The comparison of two Examiners or two observations of behaviour by the same examiner can be judged with the correlation. Correlation tells how reliable the inter observer rating of behaviour is. It tells us whether the two observers agree on the scores given or whether there is agreement between the two obtained values or scores.

It is not only the inter observer score but even the intra observations by the same examiner for two or more times can be easily compared with the help of correlation.

(ii) For Determining Validity of Scores in Predictions:

Like reliabilities, the correlation is useful in determining validity of two or more sets of scores. A high correlation is an indicator of high validity. For example, when you are interested in knowing to what extent the scores of students on aptitude test are related to their achievement or performances. If there is a high correlation between the two variables, then you can rely on any one of the tests which can be a valid measure of the variable.

(iii) Study of Individual Differences:

These relationships and the differences can be studied through the correlations of negative and positive kind. The concept of correlation is basic to the theory and practice of trait measurement. For example, the use of different types of tests like mental ability tests, aptitude tests, tests of reasoning, mathematical ability test, etc. are all used based on their property of correlation with achievement or performance, etc.

(iv) For predicting one or More Variables:

You might have estimated academic performance of the children of your school on the basis of your knowledge of their aptitude or some other test score. Even sometimes you might have predicted the scores of the students in their Board examination on the basis of their scores in previous examination or Pre-Board examination. If some one asks you the basis of your prediction, you may answer the query on the basis of your knowledge of correlation. Correlation is helpful for predicting or making estimation of one variable from our knowledge of the other variable.

(v) ***For Determining the Usefulness of Regression Line:***

Coefficient of correlation is very useful and the most frequently used measure for determining the usefulness of a regression line for the estimation purposes. It helps in proper estimations and the partial correlation of different variables that affect a particular variable.

(vi) ***Determining Prediction Validity of Tests & Measurement:***

Predictive validity is itself a correlation between a set of test scores or some other predictor with an external measure. This external measure is called the criterion. For example, to validate the intelligence tests you obtain a set of scores on a group of your school students and later find out the grade point averages (GPA) that these students get during their next examination. A correlation is then run between our two set of scores or measurements.

(vii) ***Qualifying the Relationship Between Variables:***

As correlation qualifies the relationship between the two variables, it also tells us whether the two variables are related to each other or not. It tells us about the kind of relationship, i.e., what kind of relationship exists between two variables.

(viii) ***Quantification of Degree of Relationship:***

The correlation not only qualifies for the relationship but also tells us about the quantity of relationship. The coefficient of correlation tells us the quantity of relationship. Square of coefficient of correlation (r^2) is the coefficient of determination which may even help in determining one variable with the knowledge of the other related variable. Hence, coefficient of correlation helps us in quantification of the degree of relationship between two related variables.

(ix) ***Finding Error in Prediction:***

Coefficient of correlation also gives a way of specifying the error involved in prediction of one variable from the other. The square of coefficient of correlation expresses the amount of variance commonly shared by the two variables, offering a means whereby the variability of human behaviour and performance can be explained.

16.11 LET US SUM UP

Measure of relationship of paired variables is quantified by a coefficient of correlation. The correlation coefficient is a value between – 1.0 to + 1.0. It should be noted that a high coefficient does not imply a cause-and-effect relationship, but merely quantifies a relationship that has been logically established.

The size of the correlation coefficient is affected by the homogeneity of the scores on the variables. If a relationship exists between two variables and that relationship is linear, then the scores are more heterogeneous and the greater the range of measurement, the greater is the absolute value of correlation.

The correlation coefficient is very useful in educational evaluation, standardizing tests and in making predictions.

16.12 REFERENCES AND SUGGESTED READINGS

Aggarwal, Y.P. (1990), *'Statistical Methods – Concepts, Applications and Computation'*, Sterling Publishers Pvt. Ltd., New Delhi.

Anderson, Richard L. (1952), *'Statistical Theory in Research'*, McGraw hills Book Co., New York.

Best John W and James V. Kohn (2004), *'Research in Education'*, Prentice Hall of India Pvt. Ltd., New Delhi.

Downie, N.M. (1983), *'Basic Statistical Methods'*, HarperCollins College Div., New York.

Ferguson, G.A. (1974), *'Statistical Analysis in Psychology and Education'*, McGraw Hill B.K. Co., New York.

Garrett. H.E. & Woodworth, R.S.(1969), *'Statistics in Psychology and Education'*. Vakils, Feffer & Simons Pvt. Ltd., Bombay.

Guilford, J.P. & Benjamin, F.(1973), *'Fundamental Statistics in Psychology and Education'*, McGraw Hill B.K. Co., New York.

Kerlinger, F.N.(1993), *'Foundations of Behavioural Research: Educational and Psychological Inquiry'*. Holt, Rinchart & Winston, New York.

Loether, H.L. & McTavish, D.G. (1988), *'Descriptive and Inferential Statistics: An Introduction'*, 3rd ed. Boston: Allyn & Bacon.

Sonedecor, G.W. & Cochran, W.G.(1967), *'Statistical Methods'* (6th ed.), Lowa State University Press, Ames, Lowa.

16.13 ANSWERS TO CHECK YOUR PROGRESS

1. Correlation is the association or relationship between two variables.
2. Coefficient of correlation is a number that tells us the extent to which the two given variables are related or the change in one variable is accompanied by a change in the other variable.
3. A coefficient of correlation can vary from a value of +1.00 to –1.00, through zero.
4. Self exercise (a, b, and c)
5. Self exercise
6. Self exercise