UNIT 6 CORRELATION COEFFICIENT

Structure

6.1 INTRODUCTION

In Block 1, you have studied the various measures such as measures of central tendency, measures of dispersion, moments, skewness and kurtosis which analyse variables separately. But in many situations we are interested in analysing two variables together to study the relationship between them. In this unit, you will learn about the correlation, which studies the linear relationship between the two or more variables. You would be able to calculate correlation coefficient in different situations with its properties. Thus before starting this unit you are advised to go through the arithmetic mean and variance that would be helpful in understanding the concept of correlation.

In Section 6.2, the concept of correlation is discussed with examples, that describes the situations, where there would be need of correlation study. Section 6.3 describes the types of correlation. Scatter diagrams which give an idea about the existence of correlation between two variables is explained in Section 6.4. Definition of correlation coefficient and its calculation procedure are discussed in Section 6.5. In this unit, some problems are given which illustrate the computation of the correlation coefficient in different situations as well as by different methods. Some properties of correlation coefficient with their proof are also given. In Section 6.6 the properties of the correlation coefficient are described whereas the shortcut method for the calculation of the correlation coefficient is explained in Section 6.7. In Section 6.8 the method of calculation of correlation coefficient in case of bivariate frequency distribution is explored.

Objectives

After reading this unit, you would be able to

- describe the concept of correlation;
- explore the types of correlation;
- describe the scatter diagram;

- interpret the correlation from scatter diagram;
- define correlation coefficient;
- describe the properties of correlation coefficient; and
- calculate the correlation coefficient.

6.2 CONCEPT AND DEFINITION OF CORRELATION

In many practical applications, we might come across the situation where observations are available on two or more variables. The following examples will illustrate the situations clearly:

1. Heights and weights of persons of a certain group;
2. Sales revenue and advertising expenditure in business; and
3. Time spent on study and marks obtained by students in exam.

If data are available for two variables, say X and Y, it is called bivariate distribution.

Let us consider the example of sales revenue and expenditure on advertising in business. A natural question arises in mind that is there any connection between sales revenue and expenditure on advertising? Does sales revenue increase or decrease as expenditure on advertising increases or decreases?

If we see the example of time spent on study and marks obtained by students, a natural question appears whether marks increase or decrease as time spent on study increase or decrease.

In all these situations, we try to find out relation between two variables and correlation answers the question, if there is any relationship between one variable and another.

When two variables are related in such a way that change in the value of one variable affects the value of another variable, then variables are said to be correlated or there is correlation between these two variables.

Now, let us solve a little exercise.

E1) What do you mean by Correlation?

6.3 TYPES OF CORRELATION

According to the direction of change in variables there are two types of correlation

1. Positive Correlation
2. Negative Correlation

1. Positive Correlation

Correlation between two variables is said to be positive if the values of the variables deviate in the same direction i.e. if the values of one variable increase (or decrease) then the values of other variable also increase (or decrease). Some examples of positive correlation are correlation between

1. Heights and weights of group of persons;
2. House hold income and expenditure;
3. Amount of rainfall and yield of crops; and
4. Expenditure on advertising and sales revenue.

In the last example, it is observed that as the expenditure on advertising increases, sales revenue also increases. Thus, the change is in the same direction. Hence the correlation is positive.

In remaining three examples, usually value of the second variable increases (or decreases) as the value of the first variable increases (or decreases).

2. Negative Correlation

Correlation between two variables is said to be negative if the values of variables deviate in opposite direction i.e. if the values of one variable increase (or decrease) then the values of other variable decrease (or increase). Some examples of negative correlations are correlation between

1. Volume and pressure of perfect gas;
2. Price and demand of goods;
3. Literacy and poverty in a country; and
4. Time spent on watching TV and marks obtained by students in examination.

In the first example pressure decreases as the volume increases or pressure increases as the volume decreases. Thus the change is in opposite direction.

Therefore, the correlation between volume and pressure is negative.

In remaining three examples also, values of the second variable change in the opposite direction of the change in the values of first variable.

Now, let us solve a little exercise.

E2) Explore some examples of positive and negative correlations.

6.4 SCATTER DIAGRAM

Scatter diagram is a statistical tool for determining the potentiality of correlation between dependent variable and independent variable. Scatter diagram does not tell about exact relationship between two variables but it indicates whether they are correlated or not.

Let (x_i, y_i); $(i = 1,2,...,n)$ be the bivariate distribution. If the values of the dependent variable Y are plotted against corresponding values of the independent variable X in the XY plane, such diagram of dots is called scatter diagram or dot diagram. It is to be noted that scatter diagram is not suitable for large number of observations.

6.4.1 Interpretation from Scatter Diagram

In the scatter diagram

1. If dots are in the shape of a line and line rises from left bottom to the right top (Fig.1), then correlation is said to be perfect positive.

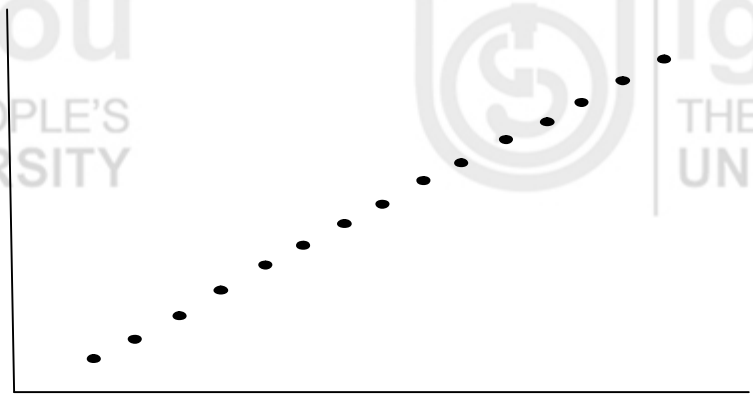

Fig. 1: Scatter diagram for perfect positive correlation

2. If dots in the scatter diagram are in the shape of a line and line moves from left top to right bottom (Fig. 2), then correlation is perfect negative.

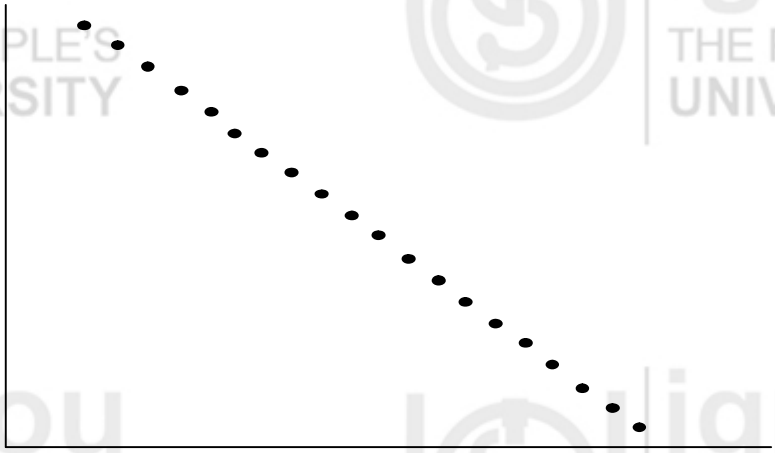

Fig. 2: Scatter diagram for perfect negative correlation

3. If dots show some trend and trend is upward rising from left bottom to right top (Fig.3) correlation is positive.

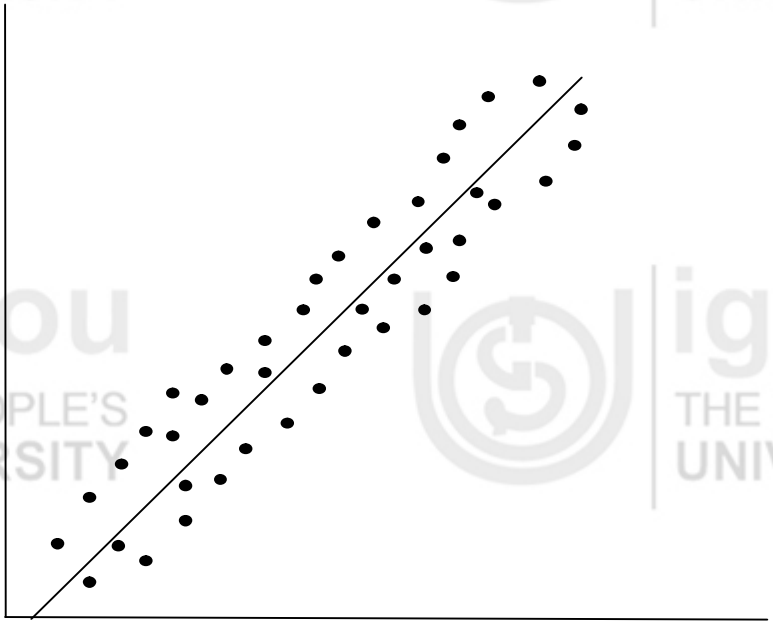

Fig. 3: Scatter diagram for positive correlation

4. If dots show some trend and trend is downward from left top to the right bottom (Fig.4) correlation is said to be negative.

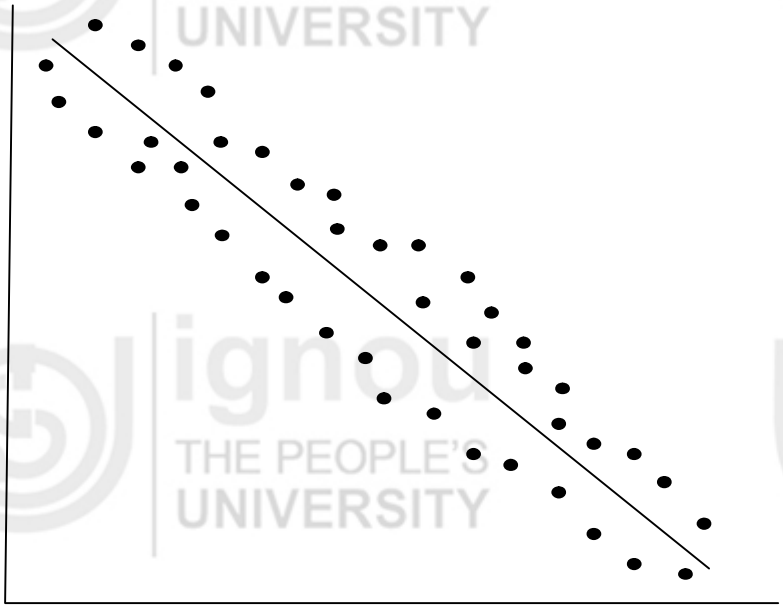

Fig. 4: Scatter diagram for negative correlation

5. If dots of scatter diagram do not show any trend (Fig. 5) there is no correlation between the variables.

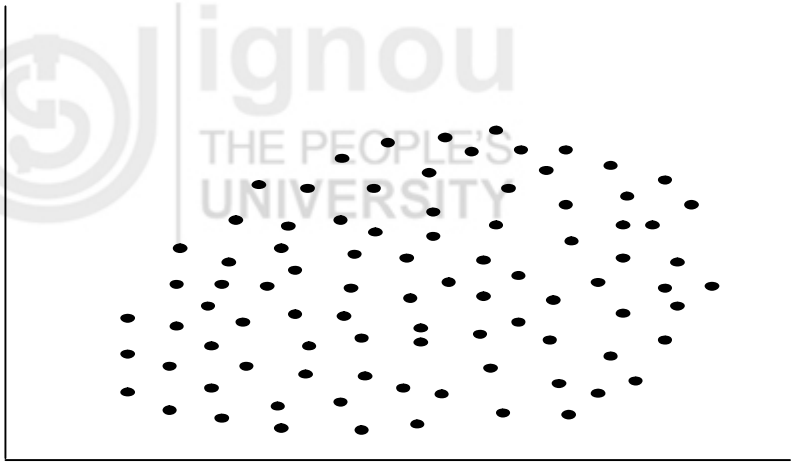

Fig. 5: Scatter diagram for uncorrelated data

6.5 COEFFICIENT OF CORRELATION

Scatter diagram tells us whether variables are correlated or not. But it does not indicate the extent of which they are correlated. Coefficient of correlation gives the exact idea of the extent of which they are correlated.

Coefficient of correlation measures the intensity or degree of linear relationship between two variables. It was given by British Biometrician Karl Pearson (1867-1936).

Note: Linear relationships can be expressed in such a way that the independent variable is multiplied by the slope coefficient, added by a constant, which determines the dependent variable. If Y is a dependent variable, X is an independent variable, b is a slope coefficient and a is constant then linear relationship is expressed as $Y = a + bX$.
In fact linear relationship is the relationship between dependent and independent variables of direct proportionality. When these variables plotted on a graph give a straight line.

If X and Y are two random variables then correlation coefficient between X and Y is denoted by r and defined as

$$r = Corr(x, y) = \frac{Cov(x, y)}{\sqrt{V(x)\,V(y)}} \qquad \ldots(1)$$

$Corr(x, y)$ is indication of correlation coefficient between two variables X and Y.

Where, $Cov(x, y)$ the covariance between X and Y which is defined as:

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

and $V(x)$ the variance of X, is defined as:

$$V(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2,$$

Note: In the above expression $\sum_{i=1}^{n}$ denotes the sum of the values for i = 1 to n; For example $\sum_{i=1}^{n} x_i$ means sum of values of X for i = 1 to n. If n = 2 i.e. $\sum_{i=1}^{2} x_i$ which is equivalent to $x_1 + x_2$. If limits are not written the summation expression i.e. $\sum x$, which indicates the sum of all values of X. You may find the discussed formulae without limits in many books. We can also write $\frac{1}{N}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ as $\frac{1}{N}\sum(x - \bar{x})(y - \bar{y})$ and both have same meaning.

Similarly,

$V(y)$ the variance of Y is defined by

$$V(y) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

where, n is number of paired observations.

Then, the correlation coefficient "r" may be defined as:

$$r = Corr(x, y) = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)\left(\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2\right)}} \quad \ldots (2)$$

Karl Pearson's correlation coefficient r is also called product moment correlation coefficient. Expression in equation (2) can be simplified in various forms. Some of them are

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)\left(\sum_{i=1}^{n}(y_i - \bar{y})^2\right)}} \quad \ldots (3)$$

or

$$r = \frac{\frac{1}{n}\sum_{i=1}^{n}x_i y_i - \bar{x}\,\bar{y}}{\sqrt{\left\{\frac{\sum_{i=1}^{n}x_i^2}{n} - \bar{x}^2\right\}\left\{\frac{\sum_{i=1}^{n}y_i^2}{n} - \bar{y}^2\right\}}} \quad \ldots (4)$$

or

$$r = \frac{\sum_{i=1}^{n}x_i y_i - n\bar{x}\,\bar{y}}{\sqrt{\left\{\sum_{i=1}^{n}x_i^2 - n\bar{x}^2\right\}\left\{\sum_{i=1}^{n}y_i^2 - n\bar{y}^2\right\}}} \quad \ldots (5)$$

or

$$r = \frac{n\sum_{i=1}^{n}x_i y_i - \sum_{i=1}^{n}x_i \sum_{i=1}^{n}y_i}{\sqrt{\left\{n\sum_{i=1}^{n}x_i^2 - \left(\sum_{i=1}^{n}x_i\right)^2\right\}\left\{n\sum_{i=1}^{n}y_i^2 - \left(\sum_{i=1}^{n}y_i\right)^2\right\}}} \quad \ldots (6)$$

6.5.1 Assumptions for Correlation Coefficient

1. **Assumption of Linearity**
   Variables being used to know correlation coefficient must be linearly related. You can see the linearity of the variables through scatter diagram.
2. **Assumption of Normality**
   Both variables under study should follow Normal distribution. They should not be skewed in either the positive or the negative direction.
3. **Assumption of Cause and Effect Relationship**
   There should be cause and effect relationship between both variables, for example, Heights and Weights of children, Demand and Supply of goods, etc. When there is no cause and effect relationship between variables then correlation coefficient should be zero. If it is non zero then correlation is termed as chance correlation or spurious correlation. For example, correlation coefficient between:

1. Weight and income of a person over periods of time; and
2. Rainfall and literacy in a state over periods of time.

Now, let us solve a little exercise.

E3) Define correlation coefficient.

6.6 PROPERTIES OF CORRELATION COEFFICIENT

Property 1: Correlation coefficient lies between -1 and $+1$.

Description: Whenever we calculate the correlation coefficient by any one of the formulae given in the Section 6.5 its value always lies between -1 and $+1$.

Proof: Consider

$$\frac{1}{n}\sum\left[\left(\frac{x-\bar{x}}{\sigma_x}\right)\pm\left(\frac{y-\bar{y}}{\sigma_y}\right)\right]^2 \geq 0$$

(Since square quantity is always greater than or equal to zero)

$$\Rightarrow \frac{1}{n}\sum\left(\frac{x-\bar{x}}{\sigma_x}\right)^2+\frac{1}{n}\sum\left(\frac{y-\bar{y}}{\sigma_y}\right)^2 \pm 2\frac{1}{n}\sum\left[\left(\frac{x-\bar{x}}{\sigma_x}\right)\left(\frac{y-\bar{y}}{\sigma_y}\right)\right] \geq 0$$

$$\Rightarrow \frac{\sigma_x^2}{\sigma_x^2}+\frac{\sigma_y^2}{\sigma_y^2}\pm\frac{2\text{Cov}(x,y)}{\sigma_x\sigma_y} \geq 0$$

Since $\frac{1}{n}\sum(x-\bar{x})^2 = \text{Variance of } X = \sigma_x^2$,

Similarly, $\frac{1}{n}\sum(y-\bar{y})^2 = \sigma_y^2$ and $\frac{1}{n}\sum(x-\bar{x})(y-\bar{y}) = \text{Cov}(x,y)$

Therefore, after putting the values

$$\Rightarrow 1+1\pm 2r \geq 0$$

$$\Rightarrow 2\pm 2r \geq 0$$

$$\Rightarrow 1\pm r \geq 0 \qquad \ldots (7)$$

If we take positive sign in equation (7) then

$$r \geq -1$$

It shows that the correlation coefficient will always be greater than or equal to -1.

If we take negative sign in equation (7) then

$$1 - r \geq 0$$

$$r \leq 1$$

It shows that correlation coefficient will always be less than or equal to $+1$. Thus

$$\Rightarrow -1 \leq r \leq 1$$

If $r = +1$, the correlation is perfect positive and if $r = -1$ correlation is perfect negative.

Property 2: Correlation coefficient is independent of change of origin and scale.

Description: Correlation coefficient is independent of change of origin and scale, which means that if a quantity is subtracted and divided by another quantity (greater than zero) from original variables , i.e. $U = \frac{X - a}{h}$ and $V = \frac{Y - b}{k}$ then correlation coefficient between new variables U and V is same as correlation coefficient between X and Y, i.e. $\text{Corr}(x, y) = \text{Corr}(u, v)$.

Proof: Suppose $u = \frac{x - a}{h}$ and $v = \frac{y - b}{k}$ then

$$x = a + h\,u \text{ and } \bar{x} = a + h\,\bar{u} \quad \ldots (8)$$

and $$y = a + k\,v \text{ and } \bar{y} = a + h\,\bar{v} \quad \ldots (9)$$

where, a, b, h and k are constants such that a > 0, b > 0, h > 0 and k > 0.

We have to prove $\text{Corr}(x, y) = \text{Corr}(u, v)$ i.e. there is no change in correlation when origin and scale are changed.

$$\text{Cov}(x, y) = \frac{1}{n}\sum (x - \bar{x})(y - \bar{y})$$

$$= \frac{1}{n}\sum (a + h\,u - a - h\,\bar{u})(b + k\,v - b - b\,\bar{v})$$

$$= \frac{1}{n} hk \sum (u - \bar{u})(v - \bar{v}),$$

$$\text{Cov}(x, y) = hk\,\text{Cov}(u, v)$$

and

$$V(x) = \frac{1}{n}\sum (x - \bar{x})^2$$

$$= \frac{1}{n}\sum (a + hu - a - h\,\bar{u})^2$$

$$= h^2 \frac{1}{n}\sum (u - \bar{u})^2$$

$$V(x) = h^2 V(u)$$

Similarly,

$$V(y) = k^2 V(v)$$

$$\text{Corr}(x, y) = \frac{\text{Cov}(x, y)}{\sqrt{V(x)V(y)}}$$

$$\text{Corr}(x, y) = \frac{hk\,\text{Cov}(u, v)}{\sqrt{h^2 V(u) k^2 V(v)}}$$

$$\text{Corr}(x, y) = \frac{\text{Cov}(u, v)}{\sqrt{V(u)V(v)}}$$

$$\text{Corr}(x, y) = \text{Corr}(u, v)$$

i.e. correlation coefficient between X and Y is same as correlation coefficient between U and V Thus, **c**orrelation coefficient is independent of change of origin and scale.

Property 3: If X and Y are two independent variables then correlation coefficient between X and Y is zero, i.e. $\text{Corr}(x, y) = 0$.

Proof: Covariance between X and Y is defined by

$$\text{Cov}(x, y) = \frac{1}{n}\sum(x - \bar{x})(y - \bar{y})$$

$$= \frac{1}{n}\sum(xy - y\bar{x} - x\bar{y} + \bar{x}\bar{y})$$

$$= \frac{1}{n}\sum xy - \bar{x}\frac{1}{n}\sum y - \bar{y}\frac{1}{n}\sum x + \bar{x}\bar{y}\frac{1}{n}\sum 1$$

$$= \frac{1}{n}\sum xy - \bar{x}\bar{y} - \bar{y}\bar{x} + \bar{x}\bar{y}$$

$$= \frac{1}{n}\sum xy - \bar{x}\bar{y}$$

$$= \bar{x}\bar{y} - \bar{x}\bar{y}$$

(if variables are independent then, $\frac{1}{n}\sum xy = \bar{x}\bar{y}$). Therefore,

$$\text{Cov}(x, y) = 0$$

Thus, correlation is

$$\text{Corr}(x, y) = \frac{\text{Cov}(x, y)}{\sqrt{V(x)V(y)}} = \frac{0}{\sqrt{V(x)V(y)}} = 0$$

As correlation measures the degree of linear relationship, different values of coefficient of correlation can be interpreted as below:

Value of correlation coefficient	Correlation is
+1	Perfect Positive Correlation
- 1	Perfect Negative Correlation
0	There is no Correlation
0 - 0.25	Weak Positive Correlation
0.75 - (+1)	Strong Positive Correlation
−0.25 - 0	Weak Negative Correlation
−0.75 - (−1)	Strong Negative Correlation

Let us discuss some problems of calculation of correlation coefficient.

Example 1: Find the correlation coefficient between advertisement expenditure and profit for the following data:

Advertisement expenditure	30	44	45	43	34	44
Profit	56	55	60	64	62	63

Solution: To find out the correlation coefficient between advertisement expenditure and profit, we have Karl Pearson's formula in many forms [(2), (3), (4), (5) and (6)] and any of them can be used. All these forms provide the same result. Let us take the form of equation (3) to solve our problem which is

$$r = Corr(x, y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)\left(\sum_{i=1}^{n}(y_i - \bar{y})^2\right)}}$$

Steps for calculation are as follow:

1. In columns 1 and 2, we take the values of variables X and Y respectively.
2. Find sum of the variables X and Y i.e.

$$\sum_{i=1}^{6} x_i = 240 \text{ and } \sum_{i=1}^{6} y_i = 360$$

3. Calculate arithmetic means of X and Y as

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{\sum_{i=1}^{6} x_i}{6} = \frac{240}{6} = 40$$

and $$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{\sum_{i=1}^{6} y_i}{6} = \frac{360}{6} = 60$$

4. In column 3, we take deviations of each observations of X from mean of X, i.e. $30 - 40 = -10$, $44 - 40 = 4$ and so on other values of the column can be obtained.
5. Similarly column 5 is prepared for variable Y i.e.

$$56 - 60 = -4,\ 55 - 60 = -5$$

and so on.

6. Column 4 is the square of column 3 and column 6 is the square of column 5.
7. Column 7 is the product of column 3 and column 5.
8. Sum of each column is obtained and written at the end of column.

To find out the correlation coefficient by above formula, we require the values of $\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$, $\sum_{i=1}^{n}(x_i-\bar{x})^2$ and $\sum_{i=1}^{n}(y_i-\bar{y})^2$ which are obtained by the following table:

x	y	$(x-\bar{x})$	$(x-\bar{x})^2$	$(y-\bar{y})$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$
30	56	−10	100	−4	16	40
44	55	4	16	−5	25	−20
45	60	5	25	0	0	0
43	64	3	9	4	16	12
34	62	−6	36	2	4	−12
44	63	4	16	3	9	12
$\sum x_i = 240$	$\sum_{i=1}^{6} y_i = 360$	$\sum_{i=1}^{6}(x_i-\bar{x}) = 0$	$\sum_{i=1}^{6}(x_i-\bar{x})^2 = 202$	$\sum_{i=1}^{6}(y_i-\bar{y}) = 0$	$\sum_{i=1}^{7}(y_i-\bar{y})^2 = 70$	$\sum_{i=1}^{6}(x_i-\bar{x})(y_i-\bar{y}) = 32$

Taking the values of $\sum_{i=1}^{6}(x_i-\bar{x})(y_i-\bar{y})$, $\sum_{i=1}^{6}(x_i-\bar{x})^2$ and $\sum_{i=1}^{6}(y_i-\bar{y})^2$ from the table and substituting in the above formula we have the correlation coefficient

$$r = Corr(x,y) = \frac{\sum_{i=1}^{6}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\left\{\sum_{i=1}^{6}(x_i-\bar{x})^2\right\}\left\{\sum_{i=1}^{6}(y_i-\bar{y})^2\right\}}}$$

$$r = Corr(x,y) = \frac{32}{\sqrt{202\times 70}} = \frac{32}{\sqrt{14140}} = \frac{32}{118.91} = 0.27$$

Hence, the correlation coefficient between expenditure on advertisement and profit is 0.27. This indicates that the correlation between expenditure on advertisement and profit is positive and we can say that as expenditure on advertisement increases (or decreases) profit increases (or decreases). Since it lies between 0.25 and 0.5 it can be considered as week positive correlation coefficient.

Example 2: Calculate Karl Pearson's coefficient of correlation between price and demand for the following data.

Price	17	18	19	20	22	24	26	28	30
Demand	40	38	35	30	28	25	22	21	20

Solution: In Example 1, we used formula given in equation (3) in which deviations were taken from mean. When means of x and y are whole number, deviations from mean makes calculation easy. Since, in Example 1, means x and y were whole number we preferred formula given in equation (3). When means are not whole numbers calculation by formula given in equation (3) becomes cumbersome and we prefer any formula given in equation (4) or (5) or (6). Since here means of x and y are not whole number, so we are preferring formula (6)

$$r = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{\left\{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right\}\left\{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2\right\}}}$$

Let us denote price by the variable X and demand by variable Y.

To find the correlation coefficient between price i.e.X and demand Y using formula given in equation (6), we need to calculate, $\sum_{i=1}^{n} x_i$, $\sum_{i=1}^{n} y_i$, $\sum_{i=1}^{n} x_i y_i$, $\sum_{i=1}^{n} x_i^2$ and $\sum_{i=1}^{n} y_i^2$ which are being obtained in the following table:

x	**y**	**x^2**	**y^2**	**xy**
17	40	289	1600	680
18	38	324	1444	684
19	35	361	1225	665
20	30	400	900	600
22	28	484	784	616
24	25	576	625	600
26	22	676	484	572
28	21	784	441	588
30	20	900	400	600
$\sum x = 204$	$\sum y = 259$	$\sum x^2 = 4794$	$\sum y^2 = 7903$	$\sum xy = 5605$

$$r = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{\left\{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right\}\left\{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2\right\}}}$$

$$r = \mathrm{Corr}(x, y)$$

$$= \frac{(9\times 5605) - (204)(259)}{\sqrt{\{(9\times 4794) - (204\times 204)\}\{(9\times 7903) - (259\times 259)\}}}$$

$$r = \mathrm{Corr}(x, y) = \frac{50445 - 52836}{\sqrt{(43146 - 41616)\times(71127 - 67081)}}$$

$$r = \mathrm{Corr}(x, y) = \frac{-2391}{\sqrt{1530\times 4046}}$$

$$r = \mathrm{Corr}(x, y) = \frac{-2391}{2488.0474}$$

$$r = \mathrm{Corr}(x, y) = -0.96$$

Note: We can use $\sum x$ instead of $\sum_{i=1}^{n} x_i$. Second expression indicates sum over x_i for i =1 to n. On the other hand first expression ($\sum x$) indicates sum over all values of X. In the current example we are using summation sign ($\sum$) without limit.

Now, let us solve the following exercises.

E4) Calculate coefficient of correlation between x and y for the following data:

x	1	2	3	4	5
y	2	4	6	8	10

E5) Find the coefficient of correlation for the following ages of husband and wife:

Husband's age	23	27	28	29	30	31
Wife's age	18	22	23	24	25	26

6.7 SHORT-CUT METHOD FOR THE CALCULATION OF CORRELATION COEFFICIENT

When values of variables are big and actual means of variables X and Y i.e. $\bar{x}$ and $\bar{y}$ are not whole number (in Example 1 mean of X and Y i.e. $\bar{x} = 40$ and $\bar{y} = 60$ were whole number) then calculation of correlation coefficient by the formula (2), (3), (4), (5) and (6) is somewhat cumbersome and we have shortcut method in which deviations are taken from assumed mean i.e. instead of actual means $\bar{x}$ and $\bar{y}$, we use assumed mean, hence $(x_i - \bar{x})$ and $(y_i - \bar{y})$ are replaced by $x_i - A_x = d_x$ and $y_i - A_y = d_y$ where A_x and A_y are assumed means of (Assumed mean may be any value of given variable of our choice) variables X and Y respectively. Formula for correlation coefficient by shortcut method is

$$r = \text{Corr}(x, y) = \frac{n\sum d_x d_y - \sum d_x \sum d_y}{\sqrt{\{n\sum d_x^2 - (\sum d_x)^2\}\{n\sum d_y^2 - (\sum d_y)^2\}}}$$

Here,

n = No. of pairs of observations,

A_x = Assumed mean of X,

A_y = Assumed mean of Y,

$\sum d_x = \sum (x - A_x)$: Sum of deviation from assumed mean A_x in X-series,

$\sum d_y = \sum (x - A_y)$: Sum of deviation from assumed mean A_y in Y-series,

$\sum d_x d_y = \sum (x - A_x)(y - A_y)$: Sum of product of deviations from assumed means A_X and A_Y in x and y series respectively,

$\sum d_x^2 = \sum (x - A_x)^2$: Sum of squares of the deviations from assumed mean A_x, in x series and

$\sum d_y^2 = \sum (y - A_y)^2$: Sum of squares of the deviations from assumed mean A_y in y series.

Note: Results from usual method and short-cut method are same.

Example 3: Calculate correlation coefficient from the following data by short-cut method:

x	10	12	14	18	20
y	5	6	7	10	12

Solution: By short-cut method correlation coefficient is obtained by

$$r = \text{Corr}(x, y) = \frac{n\sum d_x d_y - \sum d_x \sum d_y}{\sqrt{\{n\sum d_x^2 - (\sum d_x)^2\}\{n\sum d_y^2 - (\sum d_y)^2\}}}$$

$\sum d_x$, $\sum d_y$, $\sum d_x d_y$, $\sum d_x^2$ and $\sum d_y^2$ are being obtained from the following table.

Let A_x = Assumed mean of X =14 and A_y = Assumed mean of Y = 7

x	**y**	$\mathbf{d_x = x - 14}$	$\mathbf{d_x^2}$	$\mathbf{d_y = y - 7}$	$\mathbf{d_y^2}$	$\mathbf{d_x d_y}$
1	5	10 - 14 = -4	16	5–7 = –2	4	8
12	6	12–14 = –2	4	6–7 = –1	1	2
14	7	14–14 = 0	0	7–7 = 0	0	0
18	10	18–14 = 4	16	10–7 = 3	9	12
20	12	20–14 = 6	36	12–7 = 5	25	30
$\sum x = 74$	$\sum y = 40$	$\sum d_x = 4$	$\sum d_x^2 = 72$	$\sum d_y = 5$	$\sum d_y^2 = 39$	$\sum d_x d_y = 52$

Putting the required values in above formula

$$r = \frac{(5 \times 52) - (4 \times 5)}{\sqrt{\{(5 \times 72) - (4 \times 4)\}\{(5 \times 39) - (5 \times 5)\}}}$$

$$r = \frac{260 - 20}{\sqrt{\{360 - 16\}\{195 - 25\}}}$$

$$r = \frac{240}{\sqrt{\{344\}\{170\}}} = \frac{240}{241.8264} = 0.99$$

Thus, there is a very high correlation between x and y.

Now, let us solve an exercise.

E6) Find correlation coefficient between the values of X and Y from the following data by short -cut method:

x	10	20	30	40	50
y	90	85	80	60	45

6.8 Correlation Coefficient in Case of Bivariate Frequency Distribution

In bivariate frequency distribution one variable is presented in row and another in column and corresponding frequencies are given in cells (See Example 4).

If we consider two variables X and Y where, the distribution of X is given in columns and the distribution of Y is given in row. In this case we adopt the following procedure to calculate correlation coefficient.

1. Other than the given bivariate frequency distribution make three columns in the right of the table ($f_x d_x, f_x d_x^2$ and $f_x d_x d_y$) two columns in left of the table (mid value for x and class interval of variable (x) d_x), three rows in the bottom of the table $\left(f_y d_y, f_y d_y^2 \text{ and } f_y d_x d_y\right)$and two rows in the top of the table (mid value for y and d_y). Where f_x is the sum of all frequencies for the given x value i.e. $f_x = \sum_y f_{xy}$ and f_y is the sum of all frequencies for the given y values i.e. $f_y = \sum_x f_{xy}$
2. Find the mid value x and $d_x = x_i - A_x$ i.e. deviation from assumed mean A_x or step deviation i.e. $d_x = (x_i - A_x)/h$ where h is such that $(x_i - A_x)/h$ is a whole number.
3. Apply step (2) for variable Y also.
4. Find $f_x d_x$ by multiplying d_x by respective frequency f_x and get $\sum_{i=1}^{N} f_x d_x$.
5. Find $f_y d_y$ by multiplying d_y by respective frequency f_y and get $\sum_{i=1}^{N} f_y d_y$.
6. Find $f_y d_y^2$ by multiplying d_y^2 by respective frequency f_y and get $\sum_{i=1}^{N} f_y d_y^2$.
7. Find $f_x d_x^2$ by multiplying d_x^2 by respective frequency f_x and get $\sum_{i=1}^{N} f_x d_x^2$.
8. Multiply respective d_x and d_y for each cell frequency and put the figures in left hand upper corner of each cell.

9. Find $f_{xy}d_xd_y$ by multiplying f_{xy} with d_xd_y and put the figures in right hand lower corner of each cell and we apply the following formula:

$$r = \frac{\sum f_{xy}d_xd_y - \frac{\sum f_x d_x \sum f_y d_y}{N}}{\sqrt{\left\{\sum f_x d_x^2 - \frac{(\sum f_x d_x)^2}{N}\right\}\left\{\sum f_y d_y^2 - \frac{(\sum f_y d_y)^2}{N}\right\}}}$$

where, $N = \sum f_x = \sum y_y$.

Example 4: Calculate the correlation coefficient between ages of husbands and ages of wives for the following bivariate frequency distribution:

Ages of Husbands	Ages of Wives					Total
	10-20	20-30	30-40	40-50	50-60	
15-25	6	3	-	-	-	9
25-35	3	16	10	-	-	29
35-45	-	10	15	7	-	32
45-55	-	-	7	10	4	21
55-65	-	-	-	4	5	9
Total	9	29	32	21	9	100

Solution: Let, $d_y = (y - 35)/10$, where assumed mean $A_y = 35$ and $y = 10$.

$d_x = (x - 40)/10$, where assumed mean $A_x = 40$ and $h = 10$.

CI	10-20	20-30	30-40	40-50	50-60
MV (y)	15	25	35	45	55

CI	MV (x)	d_x / d_y	-2	-1	0	+1	+2	f_x	$f_x d_x$	$f_x d_x^2$	$f_{xy}d_xd_y$
15-25	20	-2	4 6 **24**	2 3 **6**	-	-	-	9	-18	36	30
25-35	30	-1	2 3 **6**	1 16 **16**	0 10 **0**	-	-	29	-29	29	22
35-45	40	0	-	0 10 **0**	0 15 0	0 7 **0**	-	32	0	0	0
45-55	50	+1	-	-	0 7 **0**	1 10 **10**	2 4 **8**	21	21	21	18
55-65	60	+2	-	-	-	2 4 **8**	4 5 **20**	9	18	36	28
		f_y	9	29	32	21	9	N = 100	$\sum f_x d_x$ = - 8	$\sum f_x d_x^2$ = 122	$\sum f_{xy}d_xd_y$ = 98
		$f_y d_y$	-18	-29	0	21	18	$\sum f_y d_y$ = -8			
		$f_y d_y^2$	36	29	0	21	36	$\sum f_y d_y^2$ = 122			
		$f_{xy}d_xd_y$	30	22	0	18	28	$\sum f_{xy} d_x d_v$ = 98			

$$r = \frac{98 - \frac{(-8 \times -8)}{100}}{\sqrt{\left\{122 - \frac{(-8)^2}{100}\right\}\left\{122 - \frac{(-8)^2}{100}\right\}}}$$

$$r = \frac{98 - 0.64}{\sqrt{\{122 - 0.64\}\{122 - 0.64\}}} = 0.802$$

6.8 SUMMARY

In this unit, we have discussed:

1. Concept of correlation;
2. Types of correlation;
3. The scatter diagrams of different correlations;
4. Calculation of Karl Pearson's coefficient of correlation;
5. Short-cut method of calculation of correlation coefficient;
6. Properties of correlation coefficient; and
7. Calculation of correlation coefficient for bi-variate data.

6.9 SOLUTIONS / ANSWERS

E1) When two variables are related in such a way that change in the value of one variable affects the value of another variable, then variables are said to be correlated or there is correlation between these two variables.

E2) Positive correlation: Correlation between

(i) Sales and profit

(ii) Family Income and year of education

Negative correlation: Correlation between

(i) No. of days students absent in class and score in exam

(ii) Time spent in office and time spent with family

E3) **C**oefficient of correlation measures the intensity or degree of linear relationship between two variables. It is denoted by r. Formula for the calculation of correlation coefficient is

$$r = Corr(x, y) = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)\left(\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2\right)}}$$

E4) We have some calculation in the following table:

x	y	$(x-\bar{x})$	$(x-\bar{x})^2$	$(y-\bar{y})$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$
1	2	-2	4	-4	16	8
2	4	-1	1	-2	4	2
3	6	0	0	0	0	0
4	8	1	1	2	4	2
5	10	2	4	4	16	8
15	30	0	10	0	40	20
15	30	0	10	0	40	20

Here $\sum x = 15$

$\Rightarrow \bar{x} = \sum x / n = 15/5 = 3$

and

$\sum y = 30$

$\Rightarrow \bar{y} = \sum y / n = 30/5 = 6$

From the calculation table, we observe that

$\sum (x-\bar{x})^2 = 10,\ \sum (y-\bar{y})^2 = 40$ and $\sum (x-\bar{x})(y-\bar{y}) = 20$

Substituting these values in the formula

$$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\left(\sum_{i=1}^{n}(x_i-\bar{x})^2\right)\left(\sum_{i=1}^{n}(y_i-\bar{y})^2\right)}}$$

$$r = \text{Corr}(x,y) = \frac{20}{\sqrt{10\times 40}} = \frac{20}{20} = 1$$

Hence, there is perfect positive correlation between X and Y.

E5) Let us denote the husband's age as X and wife's age by Y

x	y	x^2	y^2	xy
23	18	529	324	414
27	22	729	484	594
28	23	784	529	644
29	24	841	576	696
30	25	900	625	750
31	26	961	676	806
$\sum x = 168$	$\sum y = 138$	$\sum x^2 = 4744$	$\sum y^2 = 3214$	$\sum xy = 3904$

Here,

$\sum x = 168,\ \sum y = 138, \sum x^2 = 4744,$

$\sum y^2 = 3214\ \sum xy = 3904$

We use the formula

$$r = Corr(x, y) = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{\left\{n\sum x^2 - (\sum x)^2\right\}\left\{n\sum y^2 - (\sum y)^2\right\}}}$$

$$r = \frac{(6\times 3904) - (168)(138)}{\sqrt{\{(6\times 4744) - (168\times 168)\}\{(6\times 3214) - (138\times 138)\}}}$$

$$r = \frac{23424 - 23184}{\sqrt{\{28464 - 28224\}\{19284 - 19044\}}}$$

$$r = \frac{23424 - 23184}{\sqrt{240\times 240}} = 1$$

Hence there is perfect positive correlation between X and Y.

E6) By short-cut method correlation coefficient is obtained by

$$r = Corr(x, y) = \frac{n\sum d_x d_y - \sum d_x \sum d_y}{\sqrt{\{n\sum d_x^2 - (\sum d_x)^2\}\{n\sum d_y^2 - (\sum d_y)^2\}}}$$

$\sum d_x$, $\sum d_y$, $\sum d_x d_y$, $\sum d_x^2$ and $\sum d_y^2$ are being obtained through the following table.

Let A_x = assumed mean of X = 30 and A_y = Assumed mean of Y = 70

x	**y**	$\mathbf{d_x = x-30}$	$\mathbf{d_x^2}$	$\mathbf{d_y = y-70}$	$\mathbf{d_y^2}$	$\mathbf{d_x d_y}$
10	90	-20	400	20	400	-400
20	85	-10	100	15	225	-150
30	80	0	0	10	100	0
40	60	10	100	-10	100	-100
50	45	20	400	-25	625	-500
$\sum x = 150$	$\sum y = 360$	$\sum d_x = 0$	$\sum d_x^2 = 1000$	$\sum d_y = 10$	$\sum d_y^2 = 1450$	$\sum d_x d_y = -1150$

Putting the required values in above formula

$$r = \frac{(5\times -1150) - (0\times 10)}{\sqrt{\{(5\times 1000) - (0)\}\{(5\times 1450) - (10\times 10)\}}}$$

$$r = \frac{-5750}{\sqrt{\{5000\}\{7250 - 100\}}}$$

$$r = \frac{-5750}{5979.1304} = -0.96.$$