UNIT 2- SEC-I

1. 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 exampleswill illustrate the situations clearly:

a) Heights and weights of persons of a certain group;

b) Sales revenue and advertising expenditure in business; and

c) 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 inbusiness. 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 onstudy increase or decrease.

In all these situations, we try to find out relation between two variables andcorrelation 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 onevariable affects the value of another variable, then variables are said to be correlated or there is correlation between these two variables.

2. TYPES OF CORRELATION

a) Positive Correlation:

Correlation between two variables is said to be positive if the values of thevariables 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 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.

b) 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 inexamination.

3. Simple, Partial and Multiple Correlation:

The distinction between simple, partial and multiple correlation is based upon the number of variables studied.

Simple Correlation: When only two variables are studied, it is a case of simple correlation. For example, when one studies relationship between the marks secured by student and the attendance of student in class, it is a problem of simple correlation.

Partial Correlation: In case of partial correlation one studies three or more variablesbut considers only two variables to be influencing each other and the effect of other influencing variables being held constant. For example, in above example ofrelationship between student marks and attendance, the other variable influencingsuch as effective teaching of teacher, use of teaching aid like computer, smart boardetc are assumed to be constant.

Multiple Correlation: When three or more variables are studied, it is a case of multiple correlation. For example, in above example if study covers the relationship between student marks, attendance of students, effectiveness of teacher, use of teaching aids etc, it is a case of multiple correlation.

4. Linear and Non-linear Correlation:

Depending upon the constancy of the ratio of change between the variables, the correlation may be Linear or Non-linear Correlation.

Linear Correlation: If the amount of change in one variable bears a constant ratio to the amount of change in the other variable, then correlation is said to be linear.

Non-linear Correlation: If the amount of change in one variable does not bear a constant ratio to the amount of change to the other variable, then correlation is said to be non-linear.

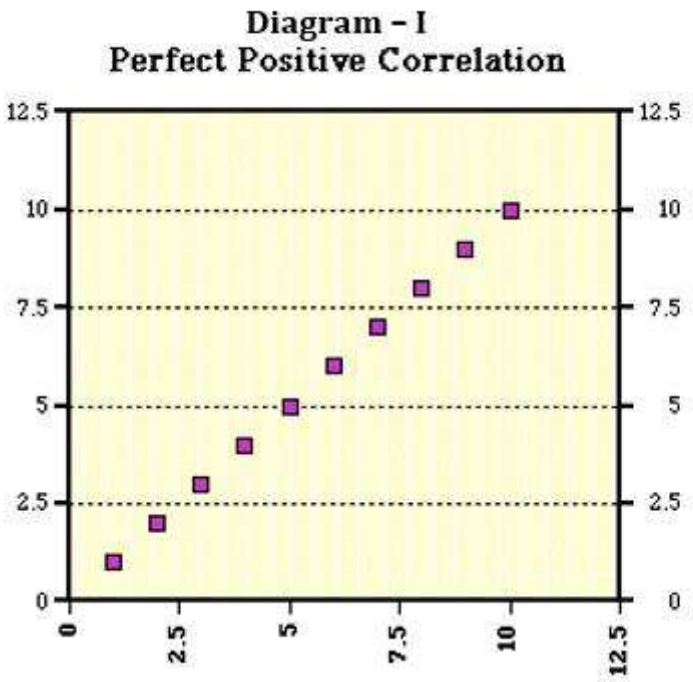
3. METHODS OF FIND CORRELATION COEFFICIENT

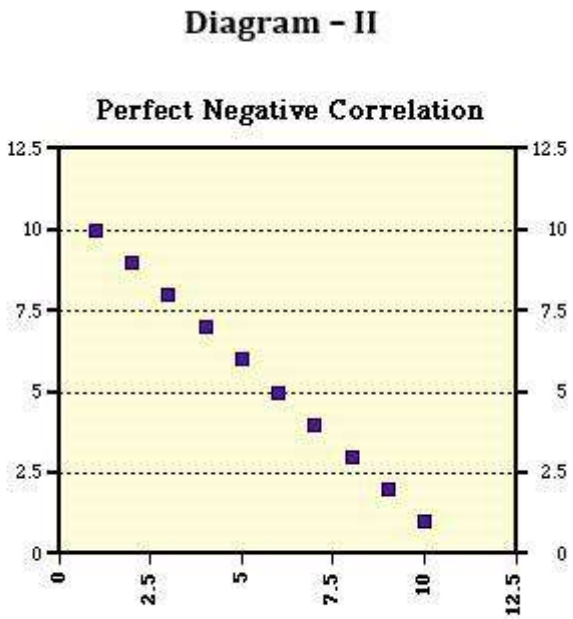
a) Scatter Diagram:

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.

In the scatter diagram





High Positive Correlation

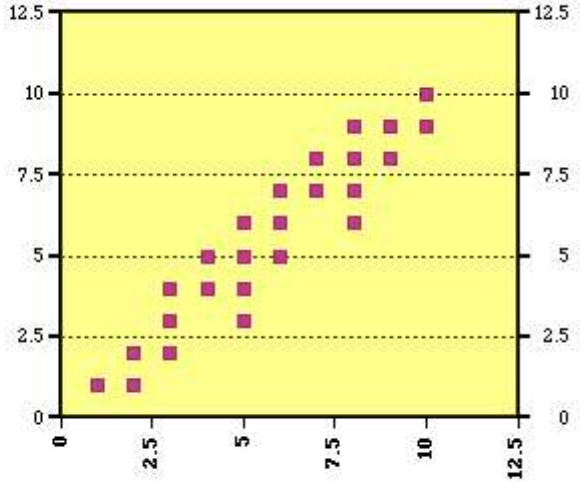


Diagram – III

High Negative Correlation

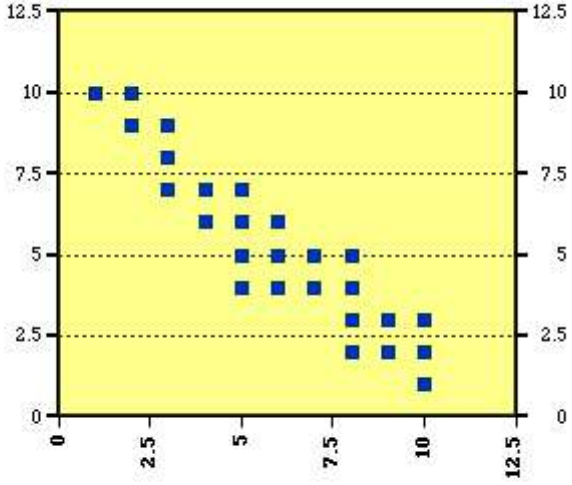


Diagram – IV

Low Postive Correlation

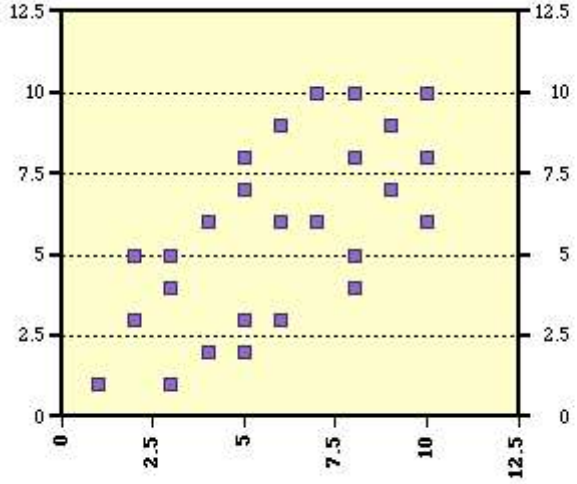


Diagram – V

Low Negative Correlation

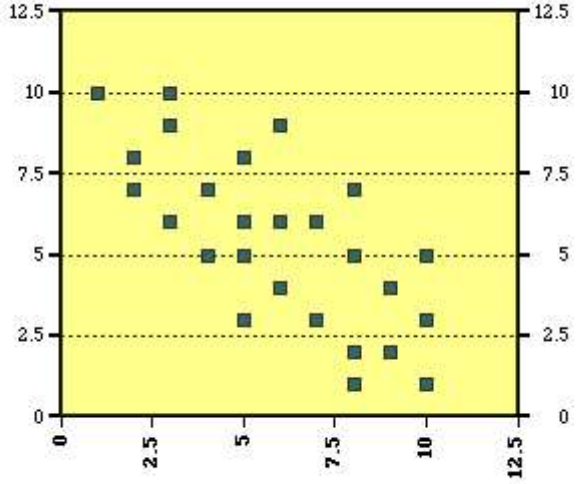


Diagram – VI

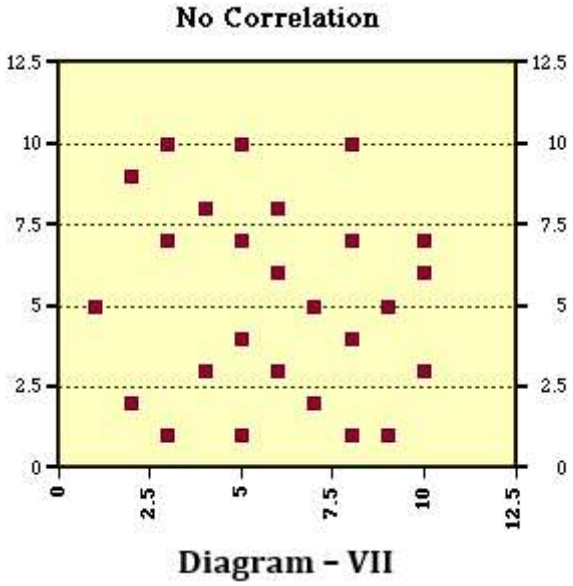


Diagram – VII

B) KARL PEARSON'S CORRELATION COFFICIENT

Scatter diagram tells us whether variables are correlated or not. But it does notindicate 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).

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

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

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:

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

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

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

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:

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

REMARK 1: Karl Pearson's correlation coefficient r is also called product moment.

REMARK 2: Karl Pearson's correlation coefficient is also denoted by $\rho(X,Y)$.

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

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

4. ASSUMPTION 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. Theyshould 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:

a) Weight and income of a person over periods of time; and
b) Rainfall and literacy in a state over periods of time.

As correlation measures the degree of linear relationship, different values ofcoefficient 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

5. PROPERTIES OF CORRELATION COFFICIENT.

Property 1: Correlation coefficient lies between −1 and +1.

Proof: We have to prove that

$$-1 \leq r(X,Y) \leq +1 \qquad (7)$$

$$r(X,Y) = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{\frac{1}{n}\Sigma (x_i - \bar{x})(y_i - \bar{y})}{\left[\frac{1}{n}\Sigma (x_i - \bar{x})^2 \cdot \frac{1}{n}\Sigma (y_i - \bar{y})^2\right]^{1/2}},$$

$$\therefore \quad r^2(X,Y) = \frac{(\Sigma a_i b_i)^2}{(\Sigma a_i^2)(\Sigma b_i^2)}, \text{ where } \begin{pmatrix} a_i = x_i - \bar{x} \\ b_i = y_i - \bar{y} \end{pmatrix} \qquad \ldots(*)$$

We have the Schwartz inequality which states that if $a_i, b_i; i = 1, 2, \ldots, n$ are real quantities then

$$\left(\sum_{i=1}^{n} a_i b_i\right)^2 \leq \left(\sum_{i=1}^{n} a_i^2\right)\left(\sum_{i=1}^{n} b_i^2\right)$$

the sign of equality holding if and only if

$$\frac{a_1}{b_1} = \frac{a_2}{b_2} = \ldots = \frac{a_n}{b_n}$$

Using Schwartz inequality, we get from (*)

$$r^2(X,Y) \leq 1 \text{ i.e., } |r(X,Y)| \leq 1 \Rightarrow -1 \leq r(X,Y) \leq 1$$

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

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

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

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

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

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

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

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

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

$$Cov(x, y) = hkCov(u, v)$$

and

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

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

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

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

Similarly,

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

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

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

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

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

i.e. correlation coefficient between X and Y is same as correlation coefficient between U and V Thus, correlation 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. If X and Y are independent variables, then

$$\text{Cov}\ (X, Y) = 0$$

$$\therefore \quad r(X, Y) = \frac{\text{Cov}\ (X, Y)}{\sigma_X \sigma_Y} = 0$$

Hence two independent variables are uncorrelated.

Hence two independent variables are uncorrelated.

But the converse of the theorem is not true, *i.e.*, two uncorrelated variables may not be independent as the following example illustrates :

X	-3	-2	-1	1	2	3	Total $\Sigma X = 0$
Y	9	4	1	1	4	9	$\Sigma Y = 28$
XY	-27	-8	-1	1	8	27	$\Sigma XY = 0$

$$\bar{X} = \frac{1}{n}\Sigma X = 0, \text{Cov}\ (X, Y) = \frac{1}{n}\Sigma XY - \bar{X}\,\bar{Y} = 0$$

$$\therefore \quad r(X, Y) = \frac{\text{Cov}\ (X, Y)}{\sigma_X\ \sigma_Y} = 0$$

Thus in the above example, the variables X and Y ~e uncorrelated. But on carful examination we find that X and Y are not independent but they are connected by the relation $Y = X^2$. Hence two uncorrelated variables need not necessarily to be independent.

Example 1.

Calculate the correlation coefficient for the following heights (in inches) of fathers (X) and their sons (Y) :

X :	65	66	67	67	68.	69	70	72
Y :	67	68	65	68	72	72	69	71

Solution.

CALCULATIONS FOR CORRELATION COEFFICIENT

	X	Y	X^2	Y^2	XY
	65	67	4225	4489	4355
	66	68	4356	4624	4488
	67	65	4489	4225	4355
	67	68	4489	4624	4556
	68	72	4624	5184	4896
	69	72	4761	5184	4968
	70	69	4900	4761	4830
	72	71	5184	5041	5112
Total	544	552	37028	38132	37560

$$\bar{X} = \frac{1}{n}\Sigma X = \frac{544}{8} = 68, \bar{Y} = \frac{1}{n}\Sigma Y = \frac{1}{8} \times 552 = 69$$

$$r(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{\frac{1}{n}\Sigma XY - \bar{X}\,\bar{Y}}{\sqrt{\left(\frac{1}{n}\Sigma X^2 - \bar{X}^2\right)\left(\frac{1}{n}\Sigma Y^2 - \bar{Y}^2\right)}}$$

$$= \frac{\frac{1}{8} \times 37560 - 68 \times 69}{\sqrt{\left[\frac{37028}{8} - (68)^2\right]\left[\frac{38132}{8} - (69)^2\right]}}$$

$$= \frac{4695 - 4692}{\sqrt{(4628{\cdot}5 - 4624)\,(4766{\cdot}5 - 4761)}} = \frac{3}{\sqrt{4{\cdot}5 \times 5{\cdot}5}} = 0{\cdot}603$$

Short Cut method:

Define $d_X = X - A_X$ and $d_Y = Y - A_Y$

Where, A_X and A_Y are assumed mean of X and Y series respectively.

The correlation coefficient is defined as

$$r = \frac{\frac{1}{n}\sum d_X d_Y - \overline{d_X}\,\overline{d_Y}}{\left\{\sqrt{\left(\frac{1}{n}\sum d_X^2 - \overline{d_X^2}\right)\quad\left(\frac{1}{n}\sum d_Y^2 - \overline{d_Y^2}\right)}\quad\right\}},$$

$$where\ \overline{d_X} = \frac{\sum d_X}{n}\ ,\ \ \overline{d_Y} = \frac{\sum d_Y}{n}$$

Example 2:

Use short-cut method to find coefficient of correlation.

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

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

x	y	$d_x = x-14$	d_x^2	$d_y = y-7$	d_y^2	$d_x d_y$
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
$\sum x = 74$	$\sum y = 40$	$\sum d_x = 4$	$\sum d_x^2 = 72$	$\sum d_y = 5$	$\sum d_y^2 = 39$	$\sum d_x d_y = 52$

$$\overline{d_X} = \frac{\sum d_X}{n} = 0.8\ ,\ \ \overline{d_Y} = \frac{\sum d_Y}{n} = 1,$$

$$r(X,Y) = \frac{\frac{1}{n}\sum d_X d_Y - \overline{d_X}\,\overline{d_Y}}{\left\{\sqrt{\left(\frac{1}{n}\sum d_X^2 - \overline{d_X^2}\right)\quad\left(\frac{1}{n}\sum d_Y^2 - \overline{d_Y^2}\right)}\quad\right\}}$$

$$r(X,Y) = \frac{\frac{1}{5}(52) - (0.8)1}{\sqrt{\left\{\left(\frac{72}{5} - 0.64\right)\left(\frac{39}{5} - 1\right)\right\}}} = \frac{9.6}{\sqrt{(13.76)(6.8)}} = \frac{9.6}{\sqrt{93.568}}$$

= 0.99

Example 3

A computer while calculating correlation coefficient between two variables X and Y from 25 pairs of observations obtained the following results :

$n = 25, \Sigma X = 125, \Sigma X^2 = 650, \Sigma Y = 100, \Sigma Y^2 = 460, \Sigma XY = 508$

It was, however, later discovered at the time of checking that he had copied down two pairs as

X	Y
6	14
8	6

while the correct values were

X	Y
8	12
6	8

Obtain the correct value of correlation coefficient.

Solution.

Corrected $\Sigma X = 125 - 6 - 8 + 8 + 6 = 125$

Corrected $\Sigma Y = 100 - 14 - 6 + 12 + 8 = 100$

Corrected $\Sigma X^2 = 650 - 6^2 - 8^2 + 8^2 + 6^2 = 650$

Corrected $\Sigma Y^2 = 460 - 14^2 - 6^2 + 12^2 + 8^2 = 436$

Corrected $\Sigma XY = 508 - 6\times14 - 8\times6 + 8\times12 + 6\times8 = 520$

$$\bar{X} = \frac{1}{n}\Sigma X = \frac{1}{25}\times 125 = 5, \quad \bar{Y} = \frac{1}{n}\Sigma Y = \frac{1}{25}\times 100 = 4$$

$$\text{Cov}(X, Y) = \frac{1}{n}\Sigma XY - \bar{X}\bar{Y} = \frac{1}{25}\times 520 - 5\times 4 = \frac{4}{5}$$

$$\sigma_X^2 = \frac{1}{n}\Sigma X^2 - \bar{X}^2 = \frac{1}{25}\times 650 - (5)^2 = 1$$

$$\sigma_Y^2 = \frac{1}{n}\Sigma Y^2 - \bar{Y}^2 = \frac{1}{25}\times 436 - 16 = \frac{36}{25}$$

$$\therefore \quad \text{Corrected } r(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{\frac{4}{5}}{1 \times \frac{6}{5}} = \frac{2}{3} = 0.67$$

Example

From the following data, compute the coefficient of correlation between X & Y:

I $\bar{x} = 25, \bar{y} = 18$

II $\Sigma (x-\bar{x})^2 = 136$

III $\Sigma (y-\bar{y})^2 = 138$

IV $\Sigma (x-\bar{x})(y-\bar{y}) = 122$

$$r(x,y) = \frac{\Sigma (x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma (x-\bar{x})^2 \cdot \Sigma (y-\bar{y})^2}}$$

$$= \frac{122}{\sqrt{136 \times 138}} = \frac{122}{136.998} = 0.89$$

Example Coefficient of correlation between X & Y is 0.3. their covariance is 9. The variance of X is 16. Find SD of Y-series.

Solution: $r(x,y) = 0.3$

$cov(x,y) = 9$

$V(x) = 16 \Rightarrow S.D(x) = 4$

$$r(x,y) = \frac{cov(x,y)}{SD(x) \cdot SD(y)}$$

$$0.3 = \frac{9}{4 \times SD(y)} \Rightarrow SD(y) = \frac{9}{4 \times 0.3} = 7.5$$

Example

$r(x,y) = 0.5$

$\Sigma (x-\bar{x})(y-\bar{y}) = 120$

$\Sigma (x-\bar{x})^2 = 90$

$SD(y) = 8 \Rightarrow$

Find n.

$$r(x,y) = \frac{\frac{1}{n}\Sigma (x-\bar{x})(y-\bar{y})}{\sqrt{\frac{1}{n}\Sigma (x-\bar{x})^2} \times 8}$$

$$\Rightarrow \frac{\frac{1}{n} \times 120}{\sqrt{\frac{1}{n} \times 90} \times 8} = 0.5$$

$$\frac{120}{\sqrt{90} \times 8} \cdot \frac{\sqrt{n}}{n} = 0.5$$

$$\frac{(120)^2}{90 \times 64} \times \frac{n}{n^2} = 0.25$$

$$\frac{14400}{5760} \times \frac{1}{n} = 0.25$$

$$\Rightarrow 2.5 \times \frac{1}{n} = 0.25$$

$$n = \frac{2.5}{0.25}$$

$$\boxed{n = 10}$$

Example

From the data given below, compute Karl Pearson's Coeff. of Correlation:

	X-series	Y-series
No. of items =	15	15
A.M =	25	18
Square of Devs from A.M	136	138

Sum of product of deviations from the Amts of X & Y = 122

Sol:

$n = 15$

$\bar{x} = 25, \quad \bar{y} = 18$

$\Sigma (x-\bar{x})^2 = 136, \quad \Sigma (y-\bar{y})^2 = 138$

$\Sigma (x-\bar{x})(y-\bar{y}) = 122$

$$r(x,y) = \frac{\Sigma (x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma (x-\bar{x})^2 \cdot \Sigma (y-\bar{y})^2}}$$

$$= \frac{122}{\sqrt{136 \times 138}} = \frac{122}{\sqrt{18768}} = \frac{122}{137} = 0.89$$