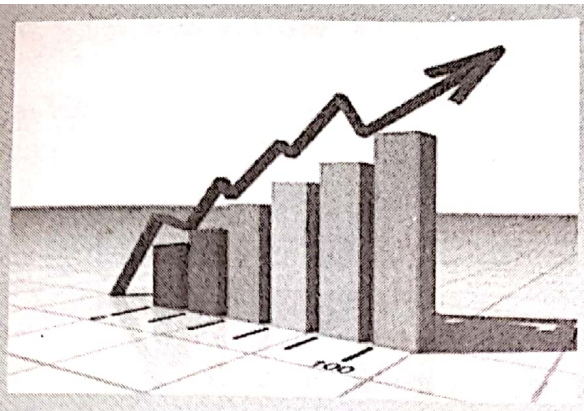


8

Correlation Analysis

8.1. MEANING AND DEFINITIONS OF CORRELATION

Meaning

Correlation, in statistics, refers to relationship between any two, or more variables viz. height and weight, rainfall and yield, price and demand, income and expenditure, wages and price index, production and employment etc. Two variables are said to be correlated if with a change in the value of one variable there arises a change in the value of another variable. On the other hand, if a change in the value of one variable does not bring any change in the value of another variable, the two variables are said to have no relation with each other. Thus, if with a change in the price of a commodity the demand for that commodity changes, we would say that the variables price, and the demand are related with each other. But there is no correlation between the heights of certain persons and the price of certain commodity because, any change in the price level of a commodity is not expected to cause any change in the height level of certain persons. In statistics, the study of correlation between any two, or more variables becomes necessary with a view to estimating the values of one variable in the light of the values of another variable. In the field of business, and economics the work of estimating the values of certain variables like cost, sales, profit, price, demand etc. becomes usual, and indispensable without which a businessman can not succeed in his business, and an economist can not draw his relevant laws and principles. But, before making such estimates, it is necessary to know first, if the two concerned variables have any relationship with each other. For this, the study of correlation between any two variables becomes necessary. However, the term correlation has been defined variously by different authors. Some of the important definitions are quoted here, as under :

Definitions

1. According to **Croxton** and **Cowden,** "When the relationship is of a quantitative nature, the appropriate statistical tool for discovering and measuring the relationship, and expressing it in a brief formula is known as correlation."
2. According to L.R. Corner, "If two, or more quantities vary in sympathy, so that movements in the one tend to be accomplished by corresponding movements in the other(s) then they are said to be correlated."
3. According to **Ya Lun Chou**, "Correlation analysis attempts to determine the 'degree of relationship' between variables".

4. From all the definitions cited above, it transpires that correlation means the inter-relation, or inter-dependence between any two, or more variables which are expressed in some quantities. As such, measures of correlation means the various methods that are applied for studying the presence, or absence, and the extent of correlation between any two, or more variables.

8.2. USES OF CORRELATION

Before going to deal with the various methods of correlation, it is necessary to know the various uses of correlation in statistical analysis which can be cited as follows :

(*i*) It is used in deriving precisely the degree, and direction of relationship between variables like price and demand, advertising expenditure and sales, rainfalls and crops yield etc.

(*ii*) It is used in developing the concept of regression, and ratio of variation which help in estimating the values of one variable for a given value of another variable.

(*iii*) It is used in reducing the range of uncertainty in the matter of prediction.

(*iv*) It is used in presenting the average relationship between any two variables through a single value of coefficient of correlation.

(*v*) In the field of economics it is used in understanding the economic behaviour, and locating the important variables on which the others depend.

(*vi*) In the field of business it is used advantageously to estimate the cost of sales, volume of sales, sales prices, and any other values on the basis of some other variables which are financially related to each other.

(*vii*) In the field of science and philosophy, also, the methods of correlation are profusely used in making progressive developments in the respective lines.

(*viii*) In the field of nature also, it is used in observing the multiplicity of the inter-related forces.

8.3. CORRELATION VS. CAUSATION

Correlation is very often misunderstood as causation *i.e.* a cause and effect relationship. But as a matter of fact, correlation implies only covariation between any two, or more variables. A very high degree of correlation obtained from the calculation does not necessarily mean that there is some cause and effect relationship between the two variables. A correlation measure may give us some value of coefficient of correlation between the variables of marks and weight but it can not be concluded there from that the 'mark' variable can be a cause, or effect of the weight variable under any circumstances. Such type of conclusion, or interpretation of cause and effect relationship is nothing but non-sense, or spurious. Therefore, before interpreting the value of correlation between any two variables as the causation, or the cause and effect relationship, care must be taken to see that the two variables are of such nature that there can exist some sort of relationship between them in reality for which one of them can be either a cause, or an effect of another.

In this connection, the following points should be taken in to consideration before interpreting the correlation as a causation.

(*i*) The correlation between any two series may be observed due to pure chance as under :

Marks	**20**	**30**	**40**	**50**	**60**	**70**	**80**
Heights	12	18	24	30	36	42	48

8.6. DIFFERENT MEASURES OF SIMPLE CORRELATION

There are different methods of studying correlation between any two, or more series. But for measuring the correlation between any two variables *i.e.* simple correlation, selection of the suitable method may be made out of the following :

(*i*) Diagrammatic method.
(*ii*) Graphic method.
(*iii*) Karl Pearson's coefficient method.
(*iv*) Spearman's coefficient method.
(*v*) Concurrent deviation method.
(*vi*) Least square method.

Each of the above methods is described at length as under :

(*i*) Diagrammatic method

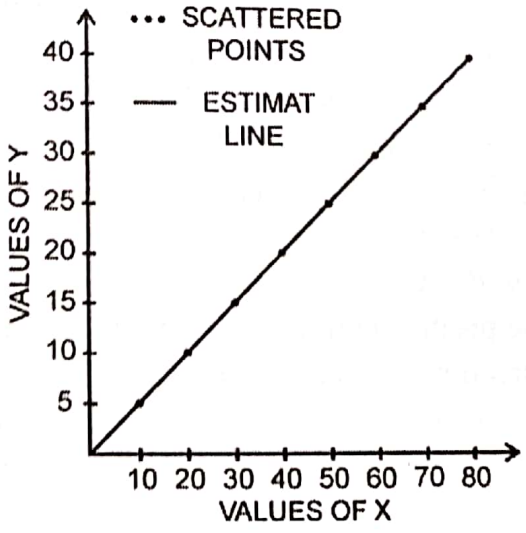
Under this method, a scatter diagram is drawn on the basis of the corresponding values of any two variables. The values of one of the variables are represented on the X axis and those of the other variable on the Y axis through natural scales on which equal subdivision represents equal values. For each of the pairs of the values of the variables, a dot is plotted on the graph paper. The dots so plotted on the graph paper give an indication of the direction of the diagram. If, the diagram appears to be upward from the left bottom to the right top, it indicates the sign of positive correlation. If the diagram appears to be downward from the left top to the right bottom, it indicates the sign of negative correlation. On the other hand, if the diagram does not show any direction *i.e.* either upward, or downward, it indicates the absence of correlation between the two variables. Further, if the diagram appears with a straight line of upward direction, it indicates the sign of perfect positive correlation. On the other hand, if the diagram appears with a straight line of downward direction, it indicates the sign of perfect negative correlation between the two series.

ILLUSTRATION 1. ***From the following pairs of data, study the correlation through a scatter diagram and draw an approximate estimating line by free-hand.***

X :	10	20	30	40	50	60	70	80
Y :	5	10	15	20	25	30	35	40

SOLUTION

Study of correlation through the scatter diagram



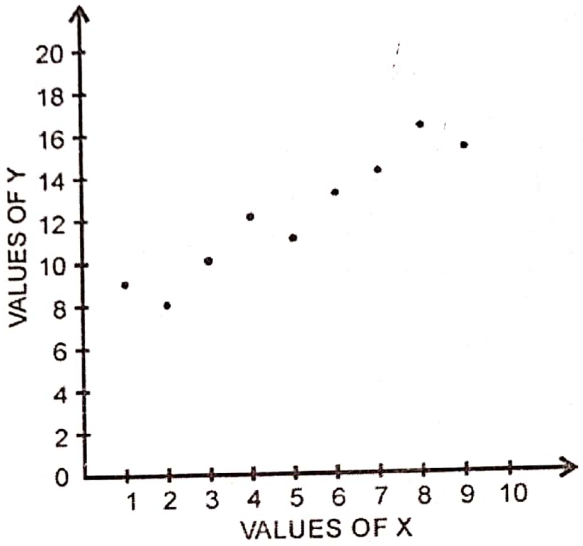
Comment. The above scatter diagram indicates that there is a perfect positive correlation between the values of the two variables X, and Y.

ILLUSTRATION 2. *From the data given below, study the correlation between the two variables by drawing a scatter diagram, and comment thereon.*

X :	9	8	7	6	5	4	3	2	1
Y :	15	16	14	13	11	12	10	8	9

SOLUTION

Study of correlation between the two variables X, and Y through the scatter diagram



Comment. From the shape of the scattered points thus exhibited in the above diagram with an upward trend rising from the lower left hand corner to the upper right hand corner of the diagram, it comes out that there is a positive correlation between the two variables X, and Y but not in a perfect manner.

(*ii*) Graphic Method

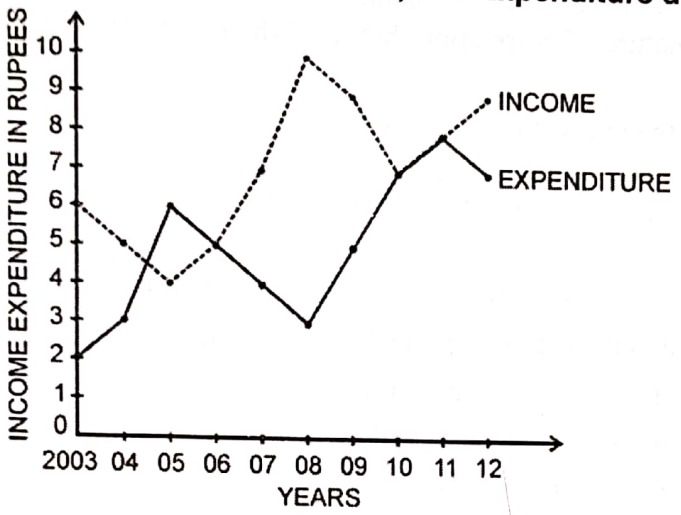
Under this method, graphs are drawn for each of the variables under study. Such graphs can be drawn either on a natural scale, or on a semi-logarithmic or ratio scale depending upon the size of the magnitude of the data. If the size of the magnitudes of the data appear to be very big, the semi-logarithmic scales are advantageously used. Further, if the minimum values of the variables are much above zero, a false base line is drawn in order to avoid the unnecessary empty spaces in the graph, and to exhibit the graphs more prominently on a large space of the paper. However, under this method, the values of all the variables are represented on the Y axis, and the values of a common reference *viz. time, place etc. are represented on the X axis i.e.* the base line. The different graphs so drawn, if move in the same direction, indicate the positive correlation between the variables. On the other hand, if the graphs move in the opposite direction *i.e.* one moves upward, and the other downward, it would indicate a negative correlation between the variables. If, the graphs move criss-cross and show erratic movements, it would indicate that either there is no correlation, or there is a very low degree of correlation between the two variables under study.

ILLUSTRATION 3. ***From the following data, study the correlation between the two variables of income, and expenditure using the graphic method :***

Year	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012
Income	6	5	4	5	7	10	9	7	8	9
Expenditure	2	3	6	5	4	3	5	7	8	7

SOLUTION

Graphic study of correlation between Income, and Expenditure during the 10 years



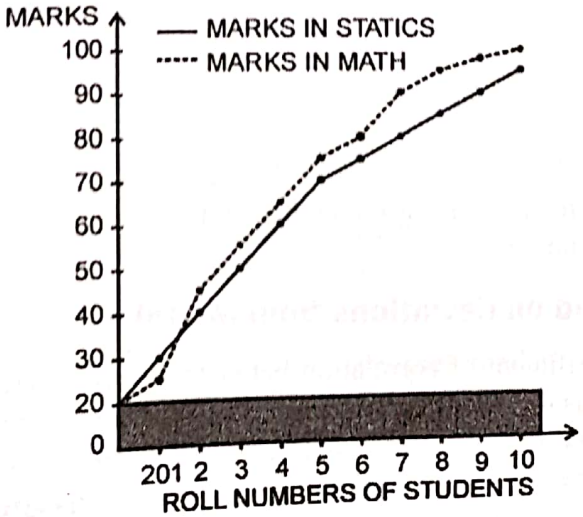
Comment. From the above graphs it appears that when the income graph is falling, the expenditure graph is rising, and when the expenditure graph is falling, the income graph is rising. It is thus, observed that the two graphs move in an opposite direction, which indicates that there is a negative correlation between the values of the income, and expenditure variables.

ILLUSTRATION 4. ***From the following data, ascertain by graphic method, if there is any correlation between the marks in Statistics and the marks in Mathematics.***

Roll No. of students :	201	202	203	204	205	206	207	208	209	210
Marks (in Statistics) :	30	40	50	60	70	75	80	85	90	95
Marks (in EQM)	25	45	55	65	75	80	90	95	98	100

SOLUTION

Graphic measurement of correlation between the marks in Statistics, and the marks in Mathematics



Comment. From the above graphs, it appears that both the graphs move in the same direction. This signifies that there is a positive correlation between the two variables of Statistics and Mathematics.

(iii) Karl Pearson's Method of Coefficient of Correlation

The British Biometrician Prof. Karl Pearson has devised several formulae of algebraic nature for measuring not only the nature of correlation, but also, the exact extent of the correlation in numerical forms.

For this, he represents the coefficient of correlation through the letter '*r*' *and asserts that the value of his 'r' must be in between* ±1. It is independent of the change of origin and scale of reference and is a pure number independent of the unit of measurement. His interpretation of the different values of '*r*' are as follows :

When, $r = 1$, it is indicative of perfect positive correlation

$r = -1$, it is indicative of perfect negative correlation

$r > 0$, but < 1 it is indicative of imperfect positive correlation

$r < 0$, but > -1 it is indicative of imperfect negative correlation

$r = 0$, it is indicative of absence of correlation

$r > 1$, or < -1 it is indicative of erroneous result

$r \geq \pm 0.90$, it is indicative of very high degree of correlation

$r \geq \pm 0.75$, but $< \pm 0.90$ it is indicative of fairly high degree of correlation

$r \geq \pm 0.50$, but $< \pm 0.75$ it is indicative of moderate degree of correlation

$r \geq \pm 0.25$, but $\leq \pm 0.50$ it is indicative of low degree of correlation

$r < \pm 0.25$, it is indicative of very low degree of correlation.

The above interpretations of the nature of correlation can be shown in a chart as under :

Chart of Correlation

Results	Degree of correlation
±1	Perfect correlation
± 0.90 or more	Very high degree of correlation
≥ ± 0.75 and < ± 0.90	Fairly high degree of correlation
≥ ± 0.50 and < ± 0.75	Moderate degree of correlation
≥ ± 0.25 and < ± 0.50	Low degree of correlation
< ± 0.25	Very low degree of correlation
0	No correlation

The different formulae for measuring the coefficient of correlation as have been devised by Prof. Pearson may be depicted as under :

1. Direct Method (based on deviations from Mean)

Under this method, coefficient of correlation between any two variables is measured on the basis of the deviations of the items obtained from their respective actual arithmetic averages. As this method is based on the product of the first moment about the Mean in the two series, it is styled as the product moment method also. Under this method, **the coefficient of correlation is defined as the ratio of**

covariance between the two variables to the product of their standard deviation *i.e.* $\frac{\Sigma xy}{N} : \sigma_x\sigma_y$.

This method is conveniently used where the values of the variables are of very big size, and their deviations from their respective Means are found to be in whole numbers.

Under this method, the fundamental formula of coefficient of correlation stands as under :

(*i*) $$\boxed{r = \frac{\Sigma xy}{N\sigma_x\sigma_y}}$$

where,

x = deviation of the first variable from its Mean.

y = deviation of the second variable from its Mean.

Σxy = total of the product of the deviations of the first, and the second variable.

N = number of pairs of the variables.

σ_x = standard deviation of the first variable.

σ_y = standard deviation of the second variable.

The above formula can be converted into the following formula for convenience in the matter of calculation :

(*ii*) $$\boxed{r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 . \Sigma y^2}}}$$

$$\because\ N\sigma_x\sigma_y = N\sqrt{\frac{\Sigma x^2}{N}} \times \sqrt{\frac{\Sigma y^2}{N}}$$

$$= N\sqrt{\frac{\Sigma x^2}{N} \cdot \frac{\Sigma y^2}{N}}$$

$$= \sqrt{\Sigma x^2 . \Sigma y^2}$$

In the above formula, calculation of the standard deviations of the two variables is dispensed with as shown in the right hand side.

STEPS :

The calculation of 'r' with the above formula involves the following steps in turn :

1. Find the arithmetic average of both the variables as they stand, or after making change of their scale, or origin as the case may be.
2. Find the deviations of the values of both the variables from their respective Means and present them as *x and y respectively.*
3. Square up the deviations of each of the variables, and present them as x^2 and y^2 respectively.
4. Find the product of each pair of the deviations, and get them totalled under Σxy.
5. Find the total of the squares of the deviations of each of the variables, and present them as Σx^2 and Σy^2 respectively.

   However, in case of the former formula, obtain the standard deviations of each of the variables, and the total number of pairs of the variables.
6. Put the respective values in the relevant formula, and get the result.
7. See that the result lies between ± 1.

ILLUSTRATION 5. *From the following statistics, find the Karl Pearson's coefficient of correlation by the direct method basing on the deviations :*

X :	8	4	10	2	6
Y :	9	11	5	8	7

SOLUTION

Calculation of the Karl Pearson's coefficient of correlation by the direct method based on the deviations

	X	$(X-\bar{X})$ x	x^2	Y	$(Y-\bar{Y})$ y	y^2	xy
	8	2	4	9	1	1	2
	4	–2	4	11	3	9	–6
	10	4	16	5	–3	9	–12
	2	–4	16	8	0	0	0
	6	0	0	7	–1	1	0
Total	30	–	40	40	–	20	–16

Arithmetic average of the first variable, or $\bar{X} = \frac{\Sigma X}{N} = \frac{30}{5} = 6$

Arithmetic average of the second variable or $\bar{Y} = \frac{\Sigma Y}{N} = \frac{40}{5} = 8$

Karl Pearson's coefficient of correlation is given by

$$r = \frac{\Sigma xy}{N\sigma_x\sigma_y}, \text{ where, } \Sigma xy = -16,$$

$$N = 5$$

We have, $$\sigma_x = \sqrt{\frac{\Sigma x^2}{N}} = \frac{40}{5} = \sqrt{8} = 2.83$$

And $$\sigma_y = \sqrt{\frac{\Sigma y^2}{N}} = \sqrt{\frac{20}{5}} = \sqrt{4} = 2$$

Substituting the respective values in the above formula we get,

$$r = \frac{-16}{5 \times 2.83 \times 2} = \frac{-16}{28.3} = -0.57 \text{ approx.}$$

Alternatively $$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 . \Sigma y^2}} = \frac{-16}{\sqrt{40 \times 20}} = \frac{-16}{\sqrt{800}} = \frac{-16}{28.3} = -0.57$$

Comment. Under both the formulae applied as above, the coefficient of correlation comes to be – 0.57 approx. which indicates that there is a moderate degree of negative correlation between the two variables.

ILLUSTRATION 6. *From the following data relating to sales and cost of sales of 10 companies, find out the Karl Pearson's coefficient of correlation by the direct method.*

Sales :	50	60	55	65	75	70	75	80	90	80
Cost of sales :	10	14	15	11	12	15	16	20	18	19

SOLUTION

Computation of the Karl Pearson's coefficient of correlation by the direct method

Sales	$(X-\bar{X})$		Cost of sales	$(Y-\bar{Y})$		xy
X	x	x^2	Y	y	y^2	
50	–20	400	10	–5	25	100
60	–10	100	14	–1	1	10
55	–15	225	15	0	0	0
65	–5	25	11	–4	16	20
75	5	25	12	–3	9	–15
70	0	0	15	0	0	0
75	5	25	16	1	1	5
80	10	100	20	5	25	50
90	20	400	18	3	9	60
80	10	100	19	4	16	40
Total 700 N = 10	–	1400	150	–	102	270

Arithmetic average of the first series, or $\bar{X} = \frac{\Sigma X}{N} = \frac{700}{10} = 70$

Arithmetic average of the second series or $\bar{X} = \frac{\Sigma Y}{N} = \frac{150}{10} = 15$

Karl Pearson's coefficient of correlation is given by

$$r = \frac{\Sigma xy}{N\sigma_x\sigma_y} = \frac{\Sigma xy}{N\sqrt{\frac{\Sigma x^2}{N} \times \frac{\Sigma y^2}{N}}} = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \Sigma y^2}}$$

Substituting the respective values in the above formula we get,

$$r = \frac{270}{\sqrt{1400 \times 102}} = \frac{270}{\sqrt{142800}} = \frac{270}{378} = 0.71 \text{ approx.}$$

The above value of 'r' indicates that there is a moderate degree of positive correlation between the two variables.

ILLUSTRATION 7. **(*On change of scale*).** ***Determine the coefficient of correlation from the following data using the direct method based on deviations given by Pearson.***

M1 :	1200	–1000	–800	–400	1200	1400	–600	–1000
M2 :	3000	1800	–2100	–3600	1200	2400	3300	–3600

SOLUTION

Determination of the Karl Pearson's coefficient of correlation by the direct method based on the deviations and change of scale

	M1/200	$(X-\bar{X})$			$M_{2/300}$	$(Y-\bar{Y})$		xy
M_1	X	x	x^2	M_2	Y	y	y^2	
1200	6	6	36	3000	10	9	81	54
–1000	–5	–5	25	1800	6	5	25 64	–25
–800	–4	–4	16	–2100	–7	–8	169	32
–400	–2	–2	4	–3600	–12	–13	9	26
1200	6	6	36	1200	4	3	49	18
1400	7	7	49	2400	8	7	100	49
–600	–3	–3	9	3300	11	10	169	–30
–1000	–5	–5	25	–3600	–12	–13		65
Total N = 8	0	–	200	–	8	–	666	189

Mean of the first series, or $\bar{X} = \frac{\Sigma X}{N} \frac{0}{8} = 0$

Mean of the second series, or $\bar{Y} = \frac{\Sigma Y}{N} \frac{8}{8} = 1$

Karl Pearson's coefficient of correlation is given by

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \Sigma y^2}}$$

Substituting the respective values in the above formula we get,

$$r = \frac{189}{\sqrt{200 \times 666}} = \frac{189}{\sqrt{133200}} = \frac{189}{365} = 0.52$$

The above result shows that there is a moderate degree of positive correlation between the two variables.

ILLUSTRATION 8. ***From the following observations, find the extent of correlation between the age, and crime by using the product moment formula of Karl Pearson.***

Age :	0–10	10–20	20–30	30–40	40–50
No. of Persons in lakhs :	80	60	50	40	30
No. of crimes :	50	45	40	35	30

Age :	50–60	60–70	70–80	80–90	90–100
No. of Persons in lakhs :	25	20	15	10	5
No. of crimes :	20	15	11	8	3

SOLUTION

Since, it is asked to study the correlation between the age variable, and the crime variable, it is first required to find the number of crimes in terms of a common denominator, say 100 lakhs.

As such, the number of crimes would be as follows :

For the age group of 0–10, out of 80 lakhs of people, the number of crimes = 50

$\therefore$ Out of 100 lakhs of people, the number of crimes = $\frac{50}{80} \times 100 = 62$ approx.

In the similar manner, the other values of the crimes would be 75, 80, 88, 100, 80, 75, 70, 80 and 60 respectively.

Computation of the Karl Pearson's correlation coefficient by the product moment formula

Age	Mid points of ages	$(X - \bar{X})/5$		No. of crimes	$(Y - \bar{Y})$		
	X	x	x^2	Y	y	y^2	xy
0–10	5	–9	81	62	–15	225	135
10–20	15	–7	49	75	–2	4	14
20–30	25	–5	25	80	3	9	–15
30–40	35	–3	9	88	11	121	–33
40–50	45	–1	1	100	23	529	–23
50–60	55	1	1	80	3	9	3
60–70	65	3	9	75	–2	4	–6
70–80	75	5	25	70	–7	49	–35
80–90	85	7	49	80	3	9	21
90–100	95	9	81	60	–17	289	–153
Total N = 10	500	–	330	770	–	1248	–92

Arithmetic average of the first series, or $\bar{X} = \frac{\Sigma X}{N} = \frac{500}{10} = 50$

Arithmetic average of the second series, or $\bar{X} = \frac{\Sigma Y}{N} = \frac{770}{10} = 77$

We have, $$\sigma = \frac{\Sigma xy}{\sqrt{x^2 \times \Sigma y^2}} = \frac{-92}{\sqrt{330 \times 1248}} = \frac{-92}{\sqrt{411840}} = \frac{-92}{642}$$

$$= -0.14 \text{ approx.}$$

The above value of the coefficient of correlation shows that the correlation between the age, and the crime is negatively very low. This implies that with an increase in the age of the people, the number of crime slightly decreases.

2. Short-cut Method (Based on Deviations from Assumed Mean)

This method is advisable when it is not possible to get the arithmetic averages of both the variables in whole or round numbers. Under this method, the deviations of values of each of the variables are taken from an assumed average. As such, the formula for computation is modified as under :

(*i*) $$r = \frac{\Sigma d_x d_y - N(\bar{X} - A_x)(\bar{Y} - A_y)}{N \sigma_x \sigma_y}$$

where, d_x = deviation from assumed average of the first, or X series

d_y = deviation from assumed average of the second, or Y series.

$\Sigma d_x d_y$ = total of the product of the pairs of the deviations from the assumed averages.

N = number of pairs of the variables

$\overline{X}$ = actual arithmetic average of the first, or X series

A_x = assumed average of the X, or first series.

$\overline{Y}$ = actual arithmetic average of the second, or Y series.

A_y = assumed average of the Y, or second series.

σ_x = standard deviation of X, or first series.

σ_y = standard deviation of Y, or second series.

For simplifying the computational work, the above formula can be modified as under :

(*ii*)
$$r = \frac{\Sigma d_x d_y - N\left(\frac{\Sigma d_x}{N}\right)\left(\frac{\Sigma d_y}{N}\right)}{N\sqrt{\frac{\Sigma d_x^2}{N} - \left(\frac{\Sigma d_x}{N}\right)^2}\cdot\sqrt{\frac{\Sigma d_y^2}{N} - \left(\frac{\Sigma d_y}{N}\right)^2}}$$

In the above formula, $(\overline{X} - A_x)$, and $(\overline{Y} - A_y)$ have been replaced by $\frac{\Sigma d_x}{N}$ and $\frac{\Sigma d_y}{N}$ respectively. This is because, the difference between the actual Mean, and the assumed Mean is equal to the average of deviations from the assumed Mean. Further σ_x and σ_y have been replaced by the formulae of standard deviation under the short cut method which is $\sigma = \sqrt{\frac{\Sigma d_x^2}{N} - \left(\frac{\Sigma d_x}{N}\right)^2}$

For further simplification of the computation work, the above formula can be reduced as follows

(*iii*)
$$r = \frac{N\Sigma d_x d_y - \Sigma d_x . \Sigma d_y}{\sqrt{N\Sigma d_x^2 - (\Sigma d_x)^2}\sqrt{N\Sigma d_y^2 - (\Sigma d_y)^2}}$$

Unless otherwise is specifically asked for, this formula should be invariably used as it makes the computation work very easy.

STEPS

The short cut method of calculating the coefficient of correlation as depicted above involves the following steps in common :

1. Find the deviations of both the series from their respective assumed averages under the heading d_x and d_y respectively.
2. Square up the assumed deviations under the heading d_x^2 and d_y^2 respectively.
3. Find the products of the assumed deviations under the heading $d_x.d_y$.
4. Get the totals of the columns d_x, d_y, d_x^2, d_y^2, and $d_x d_y$ as Σd_x, Σd_y, Σd_x^2, Σd_y^2, and $\Sigma d_x d_y$.
   In case of the first type of formula cited above, find the values of $\overline{X}$ and $\overline{Y}$ *i.e.* actual arithmetic averages of the first, and the second series, and also the standard deviations of both the series *i.e.* σ_x and σ_y.
5. Put the relevant formula, and get the result by substituting the respective values therein.

The following examples would illustrate the application of the method as under :

ILLUSTRATION 9. ***From the following data, find the coefficient of correlation between the two variables X and Y using the short-cut method.***

X :	5	10	5	11	12	4	3	2	7	6
Y :	1	6	2	8	5	1	4	6	5	2

SOLUTION

Computation of the coefficient of correlation by the short-cut method at $A_x = 7$ and $A_y = 5$

	X	$(X-A_x)$ d_x	d_x^2	Y	$(Y-A_y)$ d_y	d_y^2	$d_x d_y$
	5	–2	4	1	–4	16	8
	10	3	9	6	1	1	3
	5	–2	4	2	–3	9	6
	11	4	16	8	3	9	12
	12	5	25	5	0	0	0
	4	–3	9	1	–4	16	12
	3	–4	16	4	–1	1	4
	2	–5	25	6	1	1	–5
	7	0	0	5	0	0	0
	6	–1	1	2	–3	9	3
Total N = 10	65	–5	109	40	–10	62	43

By the short cut method

$$r = \frac{\Sigma d_x d_y - N(\bar{X} - A_x)(\bar{Y} - A_y)}{N \sigma_x \sigma_y}$$

where, $\Sigma d_x d_y = 43$, $N = 10$, $A_x = 7$, $A_y = 5$,

$$\bar{X} = \frac{\Sigma X}{N} = \frac{65}{10} = 6.5.$$

$$\bar{Y} = \frac{\Sigma Y}{N} = \frac{40}{10} = 4$$

$$\sigma_x = \sqrt{\frac{\Sigma d_x^2}{N} - \left(\frac{\Sigma d_x^2}{N}\right)} = \sqrt{\frac{109}{10} - \left(\frac{-5}{10}\right)^2} = \sqrt{10.9 - 0.25} = \sqrt{10.65} = 3.26 \text{ approx.}$$

And $$\sigma_y = \sqrt{\frac{\Sigma d_y^2}{N} - \left(\frac{\Sigma d_y^2}{N}\right)} = \sqrt{\frac{62}{10} - \left(\frac{-10}{10}\right)^2} = \sqrt{6.2 - 1} = \sqrt{5.2} = 2.28$$

Thus, substituting the respective values in the formula we get,

$$r = \frac{43 - 10(6.5 - 7)(4 - 5)}{10 \times 3.26 \times 2.28} = \frac{43 - 5}{74.33} = \frac{38}{74.33} = 0.51$$

The above result of 'r' indicates that there is a moderate degree of positive correlation between the two series.

ILLUSTRATION 10. ***From the following data, determine the Karl Pearson's coefficient of correlation taking 79 and 132 as the average for X and Y variables respectively.***

X :	61	68	79	59	69	96	89	78
Y :	108	123	136	107	112	156	137	125

SOLUTION

Since it is instructed to take certain values as the averages, it is required to use the short-cut method as under.

Computation of Karl Pearson's coefficient of correlation by the short-cut method at $A_x = 79$ and $A_y = 132$

X	$(X-A_x)$ dx	d_x^2	Y	$(Y-A_y)$ dy	d_y^2	$d_x\, d_y$
61	–18	324	108	–24	576	432
68	–11	121	123	–9	81	99
79	0	0	136	4	16	0
59	–20	400	107	–25	625	500
69	–10	100	112	–20	400	200
96	17	289	156	24	576	408
89	10	100	137	5	25	50
78	–1	1	125	--7	49	7
Total	–33	1335	N = 8	–52	2348	1696

By the short-cut method (*iii*), Karl Pearson's coefficient of correlation is given by

$$r = \frac{N\Sigma d_x d_y - \Sigma d_x . \Sigma d_y}{\sqrt{N\Sigma d_x^2 - (\Sigma d_x)^2}\sqrt{N\Sigma d_y^2 - (\Sigma d_y)^2}}$$

Substituting the respective values in the above formula we get,

$$r = \frac{8 \times 1696 - (-33 \times -52)}{(8 \times 1335) - (-33)^2 \sqrt{8 \times 2348 - (-52)^2}}$$

$$= \sqrt{\frac{13568 - 1716}{(10680 - 1089)(18784 - 2704)}}$$

$$\text{or } r = \frac{11852}{\sqrt{9591 \times 16080}}$$

$$\therefore \log r = \log 11852 - \frac{1}{2}(\log 9591 + \log 16080)$$

$$= 4.0738 - \frac{1}{2}(3.9818 + 4.2063) = 4.0738 - \frac{1}{2}(8.1881)$$

$$= 4.0738 - 4.0940 = -0.0202 \text{ or } \bar{1}.9798$$

and, r = Antilog of $\bar{1}.9798 = 0.9545$.

The above result of the coefficient of correlation shows that there is a perfect positive correlation between the two variables.

ILLUSTRATION 11. ***From the following observations, compute the Karl Pearson's coefficient of correlation by the short-cut method between the age and the success of the candidates.***

Ages :	15	16	17	18	19	20	21	22
No. of candidates appeared :	200	300	100	50	150	400	250	150
No. of successful candidates :	120	180	60	30	90	250	140	80

SOLUTION

Since, it is asked to find out the correlation between the age and the success of the candidates, it is required to find out first the number of successful candidates in terms of a common base, say 100. These will be calculated as under :

Out of 200 candidates, no. of successes = 120

So, out of 100 candidates, no. of successes = $\frac{120 \times 100}{200} = 60$

In the similar manner, the number of successful candidates per 100 for other age groups would be 60, 60, 60, 60, 60, 63, 56, 53 respectively. Now, the computation of r will proceed as follows :

Computation of Karl Pearson's coefficient of correlation by the Short-cut method at $A_x = 18$ and $A_y = 60$

Age X	$(X-A_x)$ d_x	d_x^2	No. of success Y	$(Y-A_y)$ d_y	d_y^2	$d_x d_y$
15	–3	9	60	0	0	0
16	–2	4	60	0	0	0
17	–1	1	60	0	0	0
18	0	0	60	0	0	0
19	1	1	60	0	0	0
20	2	4	63	3	9	6
21	3	9	56	–4	16	–12
22	4	16	53	–7	49	–28
Total N = 8	4	44	–	–8	74	–40

By the short-cut method

$$r = \frac{N\Sigma d_x d_y - \Sigma d_x . \Sigma d_y}{\sqrt{N\Sigma d_x^2 - (\Sigma d_x)^2}\sqrt{N\Sigma d_y^2 - (\Sigma d_y)^2}}$$

Substituting the respective values in the formula we get

$$r = \frac{8\times(-40)-(4\times-8)}{(8\times44-(-4)^2\sqrt{8\times74-(-8)^2}} = \sqrt{\frac{-320+32}{(352-16)(592-64)}} = \frac{-288}{\sqrt{336\times528}} = -0.68 \text{ approx.}$$

Comment. The above result shows that there is a moderate degree of negative correlation between the age and the success of the candidates which means that with the increase in ages, the chance of success decreases with the candidates.

3. Direct Method Based on Values of Items

This method is suitable, where the size of the given values of the variables are small, or all the values of the variables can be reduced to small size by change of their scale, or origin. Here the assumed mean is considered as zero and the formula is quite analogous to that of shortcut method. The formula of coefficient of correlation under this method stands as under

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

where,

r = Pearson's coefficient of correlation

X = given, or reduced values of the first variable

Y = given, or reduced value of the second variable, and

N = number of pairs of observations.

Note. When each of the values of a variable is divided, or multiplied by a common factor, it is a case of change of scale, when each of the values of a variable is added or subtracted by a common factor, it is a case of change of origin. Since *'r' is a pure number, it is independent both of scale and origin.*

Steps. The computation of the coefficient of correlation with the above formula will involve the following steps :

(*i*) Arrange the given data in a tabular manner, representing the first variable through X, and the second variable through Y axes.

(*ii*) If the given values are of big size, reduce them to the smallest possible size by dividing them all by a common factor. The common factor may be different for the different variables.

(*iii*) Multiply each pair of the values of X and Y, and get them totalled as ΣXY.

(*iv*) Get the totals of X and Y as ΣX and ΣY respectively.

(*v*) Square up the values of X and Y, and get them totalled as ΣX^2 and ΣY^2 respectively.

(*vi*) Square up the totals of the X and Y, and represent them as $(\Sigma X)^2$ and $(\Sigma Y)^2$ respectively.

(*vii*) Get the total of the number of pairs as N.

(*viii*) Substitute the different values in the formula given above, and find the value of *'r' through calculations.*

Another formula can also be derived from the above formula as thus,

$$r = \frac{\Sigma X\,Y - N\,\bar{X}\,\bar{Y}}{\sqrt{\Sigma X^2 - N(\bar{X})^2}.\sqrt{\Sigma Y^2 - N(\bar{Y})^2}}$$

ILLUSTRATION 12. *From the data given below, find out the coefficient of correlation between the two variables using Pearson's direct method based on values :*

Marks in English :	1	2	3	4	5
Marks in Statistics :	6	7	8	9	10

SOLUTION

Computation of the Karl Pearson's coefficient of correlation by the direct method based on values

Marks in English X	X^2	Marks in Statistics Y	Y^2	XY	
1	1	6	36	6	
2	4	7	49	14	
3	9	8	64	24	
4	16	9	81	36	
5	25	10	100	50	
$\Sigma X = 15$	$\Sigma X^2 = 55$	$\Sigma Y = 40$	$\Sigma Y^2 = 330$	$\Sigma XY = 130$	$N = 5$

Pearson's coefficient of correlation is given by

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

Substituting the respective values in the above formula we get,

$$r = \frac{5(130) - 15 \times 40}{\sqrt{5 \times 55 - (15)^2}\sqrt{5 \times 330 - (40)^2}} = \frac{650 - 600}{\sqrt{(275 - 225)(1650 - 1600)}} = \frac{50}{\sqrt{50 \times 50}} = \frac{50}{50} = +1$$

Comment. The above result of coefficient of correlation being + 1, indicates that the correlation between the two variables is perfectly positive.

ILLUSTRATION 13. (*On change of scale*) *From the following data, determine the coefficient of correlation using the Pearson's direct method based on values :*

M_1 :	75	60	45	30	15
M_2 :	150	175	200	225	250

SOLUTION

Computation of the Karl Pearson's coefficient of correlation by the direct method (based on values)

M_1	$M_1/15$ X	X^2	M_2	$M_2/25$ Y	Y^2	XY	
75	5	25	150	6	36	30	
60	4	16	175	7	49	28	
45	3	9	200	8	64	24	
30	2	4	225	9	81	18	
15	1	1	250	10	100	10	
Total	15	55	–	40	330	110	N = 5

Pearson's coefficient of correlation is given by

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

Substituting the derived values in the above formula we get,

$$r = \frac{5\times 110 - 15\times 40}{\sqrt{5\times 55 - (15)^2}\sqrt{5\times 330 - (40)^2}} = \frac{550 - 600}{\sqrt{(275-225)(1650-1600)}} = \frac{-50}{\sqrt{50\times 50}} = \frac{-50}{50} = -1$$

Comment. The above value of the coefficient of correlation being –1, indicates that there is a perfect negative correlation between the two variables.

ILLUSTRATION 14. (*On change of origin*) *From the following pairs of the data find out the Karl Pearson's coefficient of correlation.*

M_1 :	24	26	28	30	32
M_2 :	–1	–2	–3	–4	–5

SOLUTION

Calculation of the Pearson's coefficient of correlation by the direct method (based on values)

M_1	M_1-23 X	X^2	M_2	M_2+11 Y	Y^2	XY	
24	1	1	–1	10	100	10	
26	3	9	–2	9	81	27	
28	5	25	–3	8	64	40	
30	7	49	–4	7	49	49	
32	9	81	–5	6	36	54	
Total	25	165	–	40	330	180	N = 5

Pearson's coefficient of correlation is given by

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

Substituting the derived values in the above formula we get,

$$r = \frac{5(180) - (25 \times 40)}{\sqrt{(5 \times 165) - (25)^2}\sqrt{5 \times 330 - (40)^2}}$$

$$= \frac{900 - 1000}{\sqrt{(875 - 625)(1650 - 1600)}}$$

$$= \frac{-100}{\sqrt{(200 \times 50)}} = \frac{-100}{\sqrt{10000}} = \frac{-100}{100} = -1$$

Comment. The above value of 'r' indicates that the correlation between the two variables is perfectly negative.

8.7. CORRELATION OF FREQUENCY DISTRIBUTION

When the values of the two variables are frequently distributed in large numbers, computation of Karl Pearson's coefficient of correlation involves the following three steps :

(*i*) Formation of a bivariate frequency table,

(*ii*) Construction of a correlation table, and

(*iii*) Application of the frequency oriented formula.

Each of the above steps is analysed as under :

(*i*) Formation of a bivariate frequency table. For the construction of a bivariate table the following steps are to be taken up in turn :

1. Group the values of both the variables into class intervals of suitable number and magnitude. This may be same or different for the two different variables.
2. Arrange the class intervals of one of the variables in the column to the left of the table, and those of the other variable in the row at the top of the table.
3. Draw up the intersecting lines of rows, and columns, and thereby show the different cells against each of the class intervals in the table.
4. Put the tally bars in the respective cells for each of the coordinating values, and get the total of the frequencies in the respective cells, and the total of cells as well.

If the data are already in the form of a bivaribte table, it will not be necessary to form such a table any more.

The following examples will illustrate the formation of a bivariate frequency table :

ILLUSTRATION 15. ***From the following frequency distribution of marks, form a bivariate frequency table with class intervals of 20 in case of both the variables.***

Roll No.	1	2	3	4	5	6	7	8	9	10
Marks in Statistics	80	50	60	15	40	90	30	25	20	5
Marks in Accounting	70	75	65	35	20	22	88	74	30	18
Roll No.	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
Marks in Statistics	35	17	15	18	90	81	22	30	12	8
Marks in Accounting	27	15	20	25	16	7	9	8	95	85

SOLUTION

Bivariate Frequency Table

Marks in Accounting \ Marks in Statistics	0–20	20–40	40–60	60–80	80–100	Total
0–20	|| = (2)	|| = (2)	= (0)	= (0)	||= (2)	6
20–40	||| = (3)	|| = (2)	| = (1)	= (0)	| = (1)	7
40–60	= (0)	= (0)	= (0)	= (0)	= (0)	0
60–80	= (0)	| = (1)	| = (1)	1 = (|)	1 = (1)	4
80–100	|| = (2)	| = (1)	= (0)	= (0)	= (0)	3
Total	7	6	2	1	4	20

ILLUSTRATION 16. *From the following data, prepare a bivariate frequency table with class interval of 10 for English and 20 for Mathematics.*

Roll Nos.	1	2	3	4	5	6	7	8	9	10	11	12
Eng. marks	5	15	25	17	18	22	30	38	17	24	39	21
Math marks	20	30	50	10	60	80	75	90	18	20	65	60
Roll nos.	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**
Eng. marks	39	25	3	0	8	15	28	35	17	14	30	32
Math marks	21	67	81	83	93	17	82	48	45	37	77	96

SOLUTION

Bivariate Frequency Table

Marks in Math \ Marks in English	0–10	10–20	20–30	30–40	Total
0–20		||| = (3)			3
20–40	| = (1)	|| = (2)	| = (1)	| = (1)	5
40–60		| = (1)	| = (1)	| = (1)	3
60–80		| = (1)	|| = (2)	||| = (3)	6
80–100	|||= (3)		|| = (2)	|| = (2)	7
Total	4	7	6	7	24

Note. It is to be noted that a bivariate frequency table serves the function of a scatter diagram which through the shape of its frequency cells indicates the nature of correlation between the two variables. If in all the

cases, greater frequencies are noticed in the cells coordinating the larger and larger, or smaller and smaller values of both the variables, correlation is taken to be positive. On the other hand, if in all the cases, greater frequencies are noticed in the cells coordinating the larger value of one variable with the smaller value of the other variable, correlation is taken to be negative. If the association of the greater frequencies with the coordinating values in either of the above manner is not regular, the correlation is said to be absent. The following examples would illustrate the point in issue.

EXAMPLE 1

A bivariate table that indicates a positive correlation

Marks in Law / Marks in Auditing	10–20	20–30	30–40	40–50	50–60	60–70	Total
10–20	1	1					22
20–30	2	12	1				15
30–40		4	10	1			15
40–50			3	6	1		10
50–60				2	4	2	8
60–70					1	2	3
Total	3	17	14	9	6	4	53

EXAMPLE 2

A bivariate table that indicates a negative correlation

Age of wives	Age of husbands				Total
	20–30	30–40	40–50	50–60	
15–25			4	5	9
25–35		5	6		11
35–45	2	4			6
Total	2	9	10	5	26

(*ii*) Construction of a Correlation Table

After the correlated frequencies for each of the correlated values are determined through a bivariate table, the next step will be to construct a correlation table as a process of computation of the coefficient of correlation. The construction of such a table will involve the following steps :

(*i*) Put the values of the X variable in the caption or column headings, and those of Y variables in the stubs or row headings at the left of the table or vice versa.

(*ii*) Find the step deviations of the X variable against the heading d_x, and those of the Y variable under the heading d_y.

(*iii*) Put the inter-secting lines of columns and rows against each of the step deviations of both the variables, and thereby find the cells for each of the correlated step deviations.

(*iv*) Put the correlated frequencies thus obtained from the bivariate table in the left bottom corner of the respective correlated cells, and show the respective total of the frequencies in the adjacent column and row against the headings F.

(*v*) Find the products of d_x and F in the Fd_x row, and those of d_y and F in the Fd_y column and get them totalled as ΣFd_x and ΣFd_y respectively.

(*vi*) Find the products of d_x^2 and F in the Fd_x^2 row, and those of d_y^2 and F in the Fd_y^2 column, and get them totalled as ΣFd_x^2 and ΣFd_y^2 respectively.

(*vii*) Find the products of d_x, d_y and the respective frequencies of each cell, and put them in the right hand upper corner of each such cell. Get these products totalled in an intersecting cell in the column, and row of ΣFd_x d_y. The form of the table described above will appear as under :

Format of a Correlation Table

Y d_y \ X d_x		1	2	3	4				
		–2	–1	0	1	F	Fd_y	Fd_y^2	Fd_xd_y
1	–1	$\frac{f\,d_xd_y}{f}$	$\frac{f\,d_xd_y}{f}$	$\frac{f\,d_xd_y}{f}$	$\frac{fd_xd_y}{f}$	F	Fd_y	Fd_y^2	Fd_xd_y
2	0	$\frac{f\,d_xd_y}{f}$	$\frac{f\,d_xd_y}{f}$	$\frac{f\,d_xd_y}{f}$	$\frac{f\,d_xd_y}{f}$	F	Fd_y	fd_y^2	Fd_xd_y
3	1	$\frac{f\,d_xd_y}{f}$	$\frac{f\,d_xd_y}{f}$	$\frac{f\,d_xd_y}{f}$	$\frac{f\,d_xd_y}{f}$	F	Fd_y	Fd_y^2	Fd_xd_y
4	2	$\frac{f\,d_xd_y}{f}$	$\frac{f\,d_xd_y}{f}$	$\frac{f\,d_xd_y}{f}$	$\frac{f\,d_xd_y}{f}$	F	Fd_y	Fd_y^2	Fd_xd_y
	F	F	F	F	F	ΣF	ΣFd_y	ΣFd_y^2	ΣFd_xd_y
	Fd_x	Fd_x	Fd_x	Fd_x	Fd_x	ΣFd_x			
	Fd_x^2	Fd_x^2	Fd_x^2	Fd_x^2	Fd_x^2	ΣFd_x^2			
	$Fd_x\,d_y$	Fd_xd_y	Fd_xd_y	Fd_xd_y	Fd_xd_y	ΣFd_xd_y			

(*iii*) Application of the Frequency Oriented Formula

After the correlation table is constructed in the above manner, the following formula of Karl Pearson is to be applied to find out the value of the coefficient of correlation.

$$r = \frac{N\Sigma F d_x d_y - \Sigma F d_x \,.\, \Sigma F d_y}{\sqrt{N\Sigma F d_x^2 - (\Sigma F d_x)^2}\,\sqrt{N\Sigma F d_y^2 - (\Sigma F d_y)^2}}$$

The values of the relevant factors will be obtained from the correlation table, and will be substituted in the above formula. The computed value will be the value of the coefficient of correlation which will definitely lie between 1, and signify if the correlation is positive or negative. The following illustrations will make the point clear.

ILLUSTRATION 17. ***From the following bivariate table, find out the Karl Pearson's Coefficient of correlation.***

Marks in Economics	Marks in Statistics					
	0–20	20–40	40–60	60–80	80–100	Total
0–20	2	2	0	0	2	6
20–40	3	2	1	0	1	7
40–60	0	0	0	0	0	0
60–80	0	1	1	1	1	4
80–100	2	1	0	0	0	3
Total	7	6	2	1	4	20

SOLUTION

Coefficient of correlation is given by

$$r = \frac{N\Sigma F d_x d_y - \Sigma F d_x . \Sigma F d_y}{\sqrt{N\Sigma F d_x^2 - (\Sigma F d_x)^2}\sqrt{N\Sigma F d_y^2 - (\Sigma F d_y)^2}}$$

Substituting the respective values as found in the table as stated below in the above formula we get,

$$r = \frac{(20\times 2)-(-11\times -9)}{\sqrt{(20\times 51)-(-11)^2}\sqrt{(20-\times 47)-(-9)^2}}$$

$$= \frac{(20\times 2)-(-11\times -9)}{\sqrt{(20\times 51)-(-11)^2}\sqrt{2-\times 47)-(-9)^2}}$$

$$= \frac{-59}{879} = -0.07$$

Correlation Table

Mark in Economies Y	m	d_y	X: m	Marks in Statistics: 0–20	20–40	40–60	60–80	80–100				
			d_x	10	30	50	70	90				
				–2	–1	0	1	2	F	Fd_y	Fd_y^2	$Fd_x d_y$
0–20	10	–2		8 (2)	4 (2)	0 (0)	0 (0)	–8 (2)	6	–12	24	4
20–40	30	–1		6 (3)	2 (2)	0 (1)	0 (0)	–2 (1)	7	–7	7	6
40–60	50	0		0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0	0	0	0
60–80	70	1		0 (0)	–1 (1)	0 (1)	1 (1)	2 (1)	4	4	4	2
80–100	90	2		–8 (2)	–2 (1)	0 (0)	0 (0)	0 (0)	3	6	12	–10
		F		7	6	2	1	4	ΣF =20	ΣFd_y =–9	ΣFd_y^2 =47	$\Sigma Fd_x d_y$ =2

	Fd_x	–14	–6	0	1	8	$\Sigma Fd_x = -11$	
	Fd_x^2	28	6	0	1	16	$\Sigma Fd_x^2 = 51$	
	Fd_xd_y	6	3	0	1	–8	$\Sigma Fd_xd_y = 2$	

The above value of the coefficient of correlation indicates that there is a very low degree of negative correlation between the marks in Statistics and the marks in Economies.

ILLUSTRATION 18. ***From the following data arranged in a bivariate table, calculate the coefficient of correlation by the method of Karl Pearson.***

Marks in statistics	Marks in English				
	0–10	10–20	20–30	30–40	Total
0–20	0	3	0	0	3
20–40	1	2	1	1	5
40–60	0	1	1	1	3
60–80	0	1	2	3	6
80–100	3	0	2	2	7
Total	4	7	6	7	24

SOLUTION

Correlation Table

Mark in Statistics Y	m	X d_x	Marks in English 0–10	10–20	20–30	30–40	F	Fd_y	Fd_y^2	Fd_xd_y
			5	15	25	35				
		d_y	–2	–1	0	1				
0–20	10	–2	0 0	6 3	0 0	0 0	3	–6	12	6
20–40	30	–1	2 1	2 2	0 1	–1 1	5	–5	5	3
40–60	50	0	0 0	0 1	0 1	0 1	3	0	0	0
60–80	70	1	0 0	–1 1	0 2	3 3	6	6	6	2
80–100	90	2	–12 3	0 0	0 2	4 2	7	14	28	–8
		F	4	7	6	7	$\Sigma F = 24$	$\Sigma Fd_y = 9$	$\Sigma Fd_y^2 = 51$	$\Sigma Fd_xd_y = 3$
		Fd_x	–8	–7	0	7	$\Sigma Fd_x = -8$			
		Fd_x^2	16	7	0	7	$\Sigma Fd_x^2 = 30$			
		Fd_xd_y	–10	7	0	6	$\Sigma Fd_xd_y = 3$			

From the above said table it appears that $\Sigma F = 24$, $\Sigma F d_y = 9$, $\Sigma F d_y^2 = 51$, $\Sigma F d_x = -8$, $\Sigma F d_x^2 = 30$ and $\Sigma F d_x d_y = 3$.

By putting the above values in the following formula of Karl Pearson we get,

$$r = \frac{N\Sigma F d_x d_y - \Sigma F d_x . \Sigma F d_y}{\sqrt{N\Sigma F d_x^2 - (\Sigma F d_x)^2}\sqrt{N\Sigma F d_y^2 - (\Sigma F d_y)^2}}$$

$$= \frac{(24 \times 3) - (-8 \times 9)}{\sqrt{(24 \times 30) - (-8)^2}\sqrt{(24 - \times 51) - (.9)^2}}$$

$$= \frac{72 + 72}{\sqrt{(720 - 64)(1224 - 81)}}$$

$$= \frac{144}{\sqrt{(656 \times 1143)}}$$

$$= \frac{144}{\sqrt{749808}} = \frac{144}{866} = 0.17 \text{ approx.}$$

The above value of the coefficient of correlation signifies that there is a positive correlation between the two variables. However, such correlation is of very low degree as it is less than 0.25.

ILLUSTRATION 19. ***Compute the Karl Pearson's coefficient of correlation for the following bivariate frequency distribution.***

Marks in Statistics	Marks in Law						
	10–20	20–30	30–40	40–50	50–60	60–70	Total
10–20	1	1					2
20–30	2	12	1				15
30–40		4	10	1			15
40–50			3	6	1		10
50–60				2	4	2	8
60–70					1	2	3
Total	3	17	14	9	6	4	53

SOLUTION

Correlation Table

Mark in Statistics Y	m	X / m / d_x / d_y	Mark in Law 10–20 (15, −2)	20–30 (25, −1)	30–40 (35, 0)	40–50 (45, 1)	50–60 (55, 2)	60–70 (60, 3)	F	Fd_y	Fd_y^2	$Fd_x d_y$
10–20	15	−2	4 / 1	2 / 1	0 / 0	0 / 0	0 / 0	0 / 0	2	−4	8	6
20–30	25	−1	4 / 2	12 / 12	0 / 1	0 / 0	0 / 0	0 / 0	15	−15	15	16

30–40	35	0	0 0	0 4	0 10	0 1	0 0	0 0	15	0	0	0
40–50	45	1	0 0	0 0	0 3	6 6	2 1	0 0	10	10	10	8
50–60	55	2	0 0	0 0	0 0	4 2	16 4	12 2	8	16	32	32
60–70	65	3	0 0	0 0	0 0	0 0	6 1	18 2	3	9	27	24
		F	3	17	14	9	6	4	ΣF =53	ΣFd_y =16	ΣFd_y^2= 92	ΣFd_xd_y =86
		Fd_x	–6	–17	0	9	12	12	ΣFd_x =10			
		Fd_x^2	12	17	0	9	24	36	ΣFd_x^2 = 98			
		Fd_xd_y	8	14	0	10	24	30	ΣFd_xd_y = 86			

From the above said table the relevant factors are obtained as follows :

ΣF or $N = 53, \Sigma Fd_x = 10, \Sigma Fd_y = 16, \Sigma d_x^2 = 98, \Sigma Fd_y^2 = 92$ and $\Sigma Fd_x\, d_y = 86$

Karl Pearson's coefficient of correlation is given by

$$r = \frac{N\Sigma F d_x d_y - \Sigma Fd_x . \Sigma Fd_y}{\sqrt{N\Sigma F d_x^2 - (\Sigma Fd_x)^2}\sqrt{N\Sigma F d_y^2 - (\Sigma Fd_y)^2}}$$

Substituting the derived values in the above formula we have,

$$r = \frac{(53\times 86) - (10\times 16)}{\sqrt{(53\times 98) - (10^2)(53\times 92) - (16)^2}}$$

$$= \frac{4558 - 160}{\sqrt{(5194 - 100)(4876 - 256)}}$$

$$= \frac{4398}{\sqrt{5094\times 4620}} = 0.91$$

The above value of the coefficient of correlation indicates that there is a very high degree of positive correlation between the two variables. This is also indicated by the bivariate table given at the outset.

Algebraic properties of Pearson's coefficient of correlation

Prof. Karl Pearson's coefficient of correlation thus discussed above has the following algebraic properties :

1. Its value **must lie between + 1 and –1 ≤ *i.e.*–1≤ r ≤ + 1.** This property provides us with a yardstick of checking the accuracy of the calculations.
2. It is **independent of the changes of origin and scale as well.**

By change of origin we mean subtraction or addition of some constant value from/to each value of a variable. Such constants may be the same or different for the two variables X and Y. Further, by change of scale we mean dividing or multiplying each value of a variable by some constant figure and such constant figures may also be the same or different for the two variables of X and Y. This property implies that the value of the coefficient of correlation will remain the same, even if, there occurs a change of origin or a change of scale. This property helps us in simplifying the process of calculations.

3. **It is independent of the units of measurement.** This implies that even if the two variables are expressed in two different units of measurement *viz.* rain fall in inches, and yield of crops in quintals, the value of the coefficient of correlation comes out with a pure number. Thus, it does not require that the units of measurement of both the variables should be the same.
4. It is independent of the order of comparison of the two variables. Symbolically, $r_{xy} = r_{yx}$. This is because,

$$r_{xy} = \frac{\Sigma xy}{N\sigma_x\sigma_y},\ r_{yx} = \frac{\Sigma yx}{N\sigma_y\sigma_x} = \frac{\Sigma xy}{N\sigma_x\sigma_y} \therefore r_{xy} = r_{yx}.$$

5. It is the geometric mean of the two regression co-efficients *i.e.* $r = \sqrt{b_{xy} \times b_{yx}}$.

Prof.

$$b_{xy} = \frac{\sigma_x}{\sigma_y} \text{ and } b_{yx} = \frac{\sigma_y}{\sigma_x}$$

$$b_{xy} \times b_{yx} = r\frac{\sigma_x}{\sigma_y} \times r\frac{\sigma_y}{\sigma_x} = r^2$$

$\therefore$

$$r = \sqrt{b_{xy} \times b_{yx}}.$$

Assumptions of the Pearson's Coefficient of Correlation

Prof. Pearson's coefficient of correlation is based on the following assumptions :

1. Linear relationship

In devising the formulae, Prof. Pearson has assumed that there is a linear relationship between the variables which means that if the values of the two variables are plotted on a scatter diagram, it will give rise to a straight line.

2. Cause and effect relationship

Prof. Pearson has assumed that there is a cause, and effect relationship between the correlated variables which means that a change in the value of one variable is a cause for effecting a change in the value of another variable. According to him, without such relationship, correlation would carry no meaning at all.

3. Normalcy in distribution

It is assumed that the population from which the data are collected are normally distributed.

4. Multiplicity of causes

Prof. Pearson has assumed further that each of the variables under study is affected by multiplicity of causes so as to form a normal distribution. Variables like age, height, weight, price, demand, supply, yield, temperature, etc. which are usually taken to study correlation are affected by multiplicity of causes.

5. Probable error of measurement

Prof. Pearson has further assumed that there is probability of some error which may creep into the measurement of the co- efficient of correlation. But, the magnitude of such error must lie within a limit which is obtained by the following formula :

$$PE_{(r)} = 0.6745 \frac{1-r^2}{\sqrt{n}}$$

where, r = Coefficient of correlation, and n = number of pairs of the two variables.

If the constant .6745 is omitted from the above formula of probable error, we get the standard error of the coefficient of correlation.

Thus, $$SE_{(r)} = \frac{1-r^2}{\sqrt{n}}$$

The above formula of probable error helps us in interpreting the significance of the coefficient of correlation as follows :

(*i*) The correlation is taken to be almost absent, if $r < PE_{(r)}$.

(*ii*) The correlation is taken to be significant, if $r > 6 \times PE_{(r)}$.

(*iii*) The correlation is taken to be moderate, if $r > PE_{(r)}$ but < 6 times $PE_{(r)}$.

(*iv*) The limits of the correlation coefficient of the population, or $\rho_{(rho)} = r \pm PE_{(r)}$.

ILLUSTRATION 20. ***From the data given below, find out the probable error, and the standard error of the Pearson's coefficient of correlation. Also, determine the limits of the correlation co-efficient of the population.***

$$r = 0.5, \quad and \quad n = 100.$$

SOLUTION

$$PE_{(r)} = 0.6745 \times \frac{1-r^2}{\sqrt{n}} = 0.6745 \times \frac{1-(.5)^2}{\sqrt{100}} = 0.6745 \times \frac{1-0.25}{10}$$

$$= 0.6745 \times \frac{0.75}{10} = 0.05 \text{ approx.}$$

$$SE_{(r)} = \frac{1-r^2}{\sqrt{n}} = \frac{1-(.5)^2}{\sqrt{100}} = \frac{1-0.25}{10} = 0.075$$

Upper limit of the Population correlation coefficient

$$\rho_{(rho)} = r + PE_{(r)} = 0.5 + 0.05 = 0.55$$

Lower limit of the Population correlation coefficient

$$\rho_{(rho)} = r - PE_{(r)} = 0.5 - 0.05 = 0.45.$$

ILLUSTRATION 21. ***Find the probable error, and the standard error from the following data :***

$$r = 0.6 \text{ and } n = 64.$$

SOLUTION

We have, $PE_{(r)} = 0.6745 \frac{(1-r^2)}{\sqrt{n}} = 0.6745 \frac{\{1-(0.6)^2\}}{\sqrt{64}} = 0.6745 \frac{\{1-0.36\}}{8}$

$= 0.6745 \times \frac{0.64}{8} = 0.6745 \times 0.08 = 0.05396 = 0.05$ approx.

And $SE_{(r)} = \frac{1-r^2}{\sqrt{n}} = \frac{1-(0.6)^2}{\sqrt{64}} = \frac{1-0.36}{8} = \frac{0.64}{8} = 0.8.$

COEFFICIENT OF DETERMINATION

Coefficient of determination is, also, another measure of determining the correlation in the same manner as that of the Pearson's coefficient of correlation. This has been introduced by the famous Statistician, Tuttle who says that it is much easier to understand, and very much useful for interpreting the results of the coefficient of correlation between any two variables through the coefficient of determination.

According to him, coefficient of determination denoted as 'r^2' is a number that indicates the proportion of the variance in the dependent variable that is predictable from the independent variables.

In other words, **it is defined as the ratio of the explained variations to the total variations.** Thus, if $r = 0.8$, $r^2 = 0.64$, which would mean that 64 per cent of the variations in the dependent variable has been explained by the independent variable, and the rest 36 per cent of the variations is due to the other factors.

Coefficient of determination is, therefore, represented as follows :

$$\text{Co-eff. of Determination} = r^2 = \frac{\text{Explained Variation}}{\text{Total Variation}}$$

From the above formula it is to be noted that the value of the coefficient of determination shall always remain positive, and its maximum value will be 1. As such, it cannot reveal if the correlation between any two variables is positive or negative. Further, it is to be noted that the value of r^2 decreases more rapidly than the value of r and that the value of r will always be greater than the value of r^2 as shown below :

Table showing the relationship between r and r^2.

r	r^2	r	r^2
1.0	1.00	0.5	0.25
0.9	0.81	0.4	0.16
0.8	0.64	0.3	0.09
0.7	0.49	0.2	0.04
0.6	0.36	0.1	0.01

COEFFICIENT OF NON-DETERMINATION

Coefficient of non-determination explains the amount of unexplained or unaccounted for, variance between two variables or between a set of variables. It is defined by Tuttle as one minus the coefficient of determination *i.e.* $1-r^2$. Symbolically, it is represented by K^2 as belows :

$$\boxed{K^2 = 1-r^2}$$

or $$K^2 = 1 - \frac{\text{Explained Variation}}{\text{Total Variation}}$$

Thus, the coefficient of no-determination is the ratio of unexplained variation to the total variation.

CO-EFFICIENT OF ALIENATION

This is also a development over the coefficient of determination. Coefficient of alienation is a statistic that measures the lack of linear association between two variables.

This is symbolically represented by K as the square root of the coefficient of non-determination.

Thus, the coefficient of alienation, or $\boxed{K = \pm\sqrt{1-r^2}}$

$$= \pm\sqrt{1 - \frac{\text{Explained Variation}}{\text{Total Variation}}}$$

ILLUSTRATION 22. ***If r = 0.8, and N = 81, find (1) the coefficient of determination, (2 coefficient of non-determination, and (3) coefficient of alientation. Also, interpreter the results.***

SOLUTION

(*i*) Coefficient of determination = $r^2 = (0.8)^2 = 0.64$.

This result shows that 64 per cent of the change in the dependent variable is due to the change in the independent variable, and the rest 36 per cent of the change is due to the other factors. This can, also, be represented as follows :

$$\text{Coefficient of determination} = \pm\sqrt{1 - \frac{\text{Explained Variation}}{\text{Total Variation}}} = \frac{0.64}{1} = 64\%.$$

(*ii*) Coefficient of Non-determination

or $$K^2 = 1 - r^2 = 1 - 0.8^2 = 1 - 0.64 = 0.36$$

or $$K^2 = \frac{\text{Unexplained Variation}}{\text{Total Variation}} = \frac{0.36}{1} = 36\%.$$

The above result shows that 36 per cent of the change in the dependent variable is on account of the changes in the other factors.

(*iii*) Coefficient of Alienation

or $$K = \sqrt{1-r^2} = \sqrt{1-0.8^2} = \sqrt{1-0.64} = \sqrt{0.36} = 0.6$$

or $$K = \sqrt{\frac{\text{Unexplained Variation}}{\text{Total Variation}}} = \sqrt{\frac{0.36}{1}} = 0.6.$$

The above result indicates that if the effect of the other factors are alienated, the degree of correlation will be 60 per cent.

Merits and Demerits of Pearson's Method of Studying Correlation

Merits

The following are the chief points of merit that go in favour of the Karl Pearson's method of correlation :

1. This method not only indicates the presence, or absence of correlation between any two variables but also, determines the exact extent, or degree to which they are correlated.
2. Under this method, we can also ascertain the direction of the correlation *i.e.* whether the correlation between the two variables is positive, or negative.
3. This method enables us in estimating the value of a dependent variable with reference to a particular value of an independent variable through regression equations.
4. This method has a lot of algebraic properties for which the calculation of coefficient of correlation, and a host of other related factors *viz. coefficient of determination, are made easy.*

Demerits

Despite the above points of merit, this method also suffers from the following demerits :

1. It is comparatively difficult to calculate as its computation involves intricate algebraic methods of calculations.
2. It is very much affected by the values of the extreme items.
3. It is based on a large number of assumptions *viz.* linear relationship, cause and effect relationship etc. which may not always hold good.
4. It is very much likely to be misinterpreted particularly in case of homogeneous data.
5. In comparison to the other methods, it takes much time to arrive at the results.
6. It is subject to probable error which its propounded himself admits, and therefore, it is always advisable to compute its probable error while interpreting its results.

(*iv*) SPEARMAN'S RANK CORRELATION

This method is a development over Karl Pearson's method of correlation on the point that (*i*) it does not need the quantitative expression of the data, and (*ii*) it does not assume that the population under study is normally distributed.

This method was introduced by the British Psychologist **Charles Edward Spearman in 1904.** Under this method, correlation coefficient is measured on the basis of the ranks rather than the original values of the variables. A rank correlation coefficient measures the degree of similarity between two rankings and can be used to assess the significance of the relationship between them. For Rank Correlation coefficient, the values of the two variables are first converted into ranks in a particular order depicted as under :

Order of Assigning the Ranks

The ranks may be assigned to the different values either in ascending, or in descending order. In case of ascending order, the smallest rank 1 is assigned to the smallest value of a variable, and the subsequent ranks 2, 3, 4 etc. are given to the other values in order of their largeness in size. In case of descending order, the smallest rank 1 is given to the largest value of a variable and the subsequent ranks 2, 3, 4 etc. are given to the other values in order of their smallness in size. In whatever order, the ranks may be given, the same order of ranking must be followed in case of both the variables. When

two, or more values of a variable are found to be identical, each of them is to be assigned with the average of their progressive ranks. For example, if there are three 30s coming after the 4th rank, their progressive ranks would be 5, 6 and 7 respectively and their average would be $\frac{5+6+7}{3}$ *i.e.* 6. Thus, each of the three identical values will be assigned with the rank 6, and the next rank will begin with 8 to be assigned to the next value of the series.

Method of Computation

After the ranks are assigned to the values of both the variables under the two columns R_x and R_y respectively, the squares of the differences between the two corresponding ranks are found out under the next column $(R_x - R_y)^2$ or D^2. Lastly, the D^2 column is totalled, and the following formula is put to find out the values of the coefficient of correlation which also lies between ± 1.

$$r_s = 1 - \frac{6\left\{\Sigma D^2 + \frac{1}{12}(m_1^3 - m_1) + \frac{1}{12}(m_2^3 - m_2) + \frac{1}{12}(m_3^3 - m_3) + \ldots\right\}}{N^3 - N}$$

Where,

$r_{(s)}$ = Spearman's rank coefficient of correlation

ΣD^2 = Sum of the squares of the differences between the corresponding ranks of the two variables *i.e.* $(R_x - R_y)^2$.

N = number of pairs of the two variables.

m_1 = the number of items which are assigned with the first repeated rank.

m_2 = the number of items which are assigned with the second repeated rank.

m_3 = the number of items which are assigned with the third repeated rank.

Note. It is to be noted that when there is no repetition of any rank the above formula may be reduced as follows :

$$\boxed{r_{(S)} = 1 - \frac{6\Sigma D^2}{N^3 - N}}$$

ILLUSTRATION 23. ***From the following data, find out the Spearman's rank Coefficient of correlation, and comment on the result:***

Roll nos. :	1	2	3	4	5	6	7	8	9	10
Marks in commerce :	60	56	25	90	35	14	52	27	54	72
Marks in economics :	42	34	56	35	40	50	45	60	58	36

SOLUTION

Computation of the Spearman's rank coefficient of correlation

Marks in Com. X	Marks in Eco. Y	Rank in ascending order		Difference between the rank & $(R_x - R_y)$ D	Square of the differences in ranks D^2
		R_X	R_Y		
60	42	8	5	3	9
56	34	7	1	6	36

25	56	2	8	–6	36
90	35	10	2	8	64
35	40	4	4	0	0
14	50	1	7	–6	36
52	45	5	6	–1	1
27	60	3	10	–7	49
54	58	6	9	–3	9
72	36	9	3	6	36
Total	N = 10	–	–	0	$\Sigma D^2 = 276$

By putting the formula of Spearman's rank coefficient of correlation :

We have, $r_{(S)} = 1 - \frac{6\Sigma D^2}{N^3 - N} = 1 - \frac{6(276)}{10^3 - 10}$

$= 1 - \frac{1656}{990} = 1 - 1.67 = -0.67.$

Comment

From the above result of the Spearman's rank correlation it comes out that there is a moderate degree of negative correlation between the marks in Commerce, and the marks in Economics. This means that if a student secures more marks in Commerce subject his marks in Economics is likely to decrease to a moderate extent.

ILLUSTRATION 24. ***Find out the Spearman's coefficient of rank correlation from the following data relating to the ranks assigned by the two judges on a certain competition.***

Candidates :	A	B	C	D	E	F	G	H	I	J
Marks by Judge I :	26	25	38	37	41	45	60	42	53	57
Marks by Judge II :	52	25	30	35	48	77	38	43	68	64

SOLUTION

Computation of the Spearman's coefficient of rank correlation

Marks by Judge I X	Marks by Judge II Y	Ranks in descending order		Difference in ranks (d) ($R_x - R_y$)	D^2
		R_x	R_y		
26	52	9	4	5	25
25	25	10	10	0	0
38	30	7	9	–2	4
37	35	8	8	0	0
41	48	6	5	1	1
45	77	4	1	3	9
60	38	1	7	–6	36
42	43	5	6	–1	1
53	68	3	2	1	1
57	64	2	3	–1	1
Total	N = 10	–	–	0	$\Sigma D^2 = 78$

By the formula of Spearman we have,

$$r_{(S)} = 1-\frac{6\Sigma D^2}{N^3-N} = 1-\frac{6(78)}{10^3-10} = 1-\frac{468}{990}$$

$$= 1-0.47 = 0.53.$$

Comment. The above result indicates that there is a moderate degree of positive correlation between the two variables.

ILLUSTRATION 25. ***From the following data relating to the marks secured by a batch of candidates, ascertain the rank coefficient of correlation and interpreter the results.***

Candidates :	A	B	C	D	E	F	G	H	I	J
Marks in English :	50	40	50	35	37	18	30	22	15	5
Marks in Economics :	58	60	48	50	30	32	45	37	42	52
Marks in Commerce :	70	68	75	40	80	50	30	85	25	90

SOLUTION

Computation of the Rank Coefficient of Correlation between the marks in the three subjects

Marks in English X	Marks in Economics Y	Marks in Commerce Z	Ranks in ascending order			Squares of the Diff. in ranks		
			Rx	Ry	Rz	D_1^2	D_2^2	D_3^2
50	58	70	10	7	6	1	16	9
40	60	68	8	10	5	4	9	25
50	48	75	9	6	7	9	4	1
35	50	40	6	7	3	1	9	16
37	30	80	7	1	8	36	1	49
18	32	50	3	2	4	1	1	4
30	45	30	5	5	2	0	9	9
22	37	85	4	3	9	1	25	36
15	42	25	2	4	1	4	1	9
5	52	90	1	8	10	49	81	4
Total	N = 10	–	–	–		106	156	162

Rank correlation coefficient

or $$r_{(S)} = 1-\frac{6\Sigma D^2}{N^3-N}$$

Thus, $$r_{(S)(x-y)} = 1-\frac{6(106)}{10^3-10} = 1-\frac{636}{990} = 1-0.64 = 0.36$$

$$r_{(S)(x-z)} = 1-\frac{6(156)}{10^3-10} = \frac{936}{990} = 1-0.95 = 0.05$$

And $$r_{(S)(y-z)} = 1-\frac{6(162)}{10^3-10} = \frac{972}{990} = 1-0.98 = 0.02.$$

From the above results, it appears that there are positive correlation between each pair of the subjects, but the correlation between the English and Economics marks appears to be more closer.

ILLUSTRATION 26. (*On equal ranks*) ***From the following data relating to sales and net profits of a firm find the rank correlation coefficient.***

Sales in ₹ :	60	80	90	60	100	130	120	110
Profits in ₹ :	30	40	50	40	60	70	40	75

SOLUTION

Computation of the Rank Coefficient of Correlation between the Sales and Profits

Sales X	Profits Y	Ranks in desending order		$(R_x - R_y)$ D	D^2
		R_x	R_y		
60	30	7.5	8	–0.5	0.25
80	40	6	6	0.0	0.00
90	50	5	4	1.0	1.00
60	40	7.5	6	1.5	2.25
100	60	4	3	1.0	1.00
130	70	1	2	–1.0	1.00
120	40	2	6	–4.0	16.00
110	75	3	1	2.0	4.00
Total	–	–	–	0	25.50

By the relevant formula of rank correlation we have ;

$$R_{(S)} = 1 - \frac{6\left\{\Sigma D^2 + \frac{1}{12}(m_1^3 - m_1) + \frac{1}{12}(m_2^3 - m_2) \ \ldots\right\}}{N^3 - N}$$

Where, $r_{(S)}$ = the required rank correlation coefficient

ΣD^2 = Sum of the squares of the differences between the corresponding ranks *i.e.* 25.50.

m_1 = the number of items assigned with the first repeated rank 7.5 *i.e.* 2

m_2 = the number of items assigned with the second repeated rank 6 *i.e.* 3

And N = number of pairs of the ranks *i.e.* 8.

Substituting the relevant values in the above formula we have,

$$r_{(S)} = 1 - 6\left\{\frac{25.5 + \frac{1}{12}(2^3 - 2) + \frac{1}{12}(3^3 - 3)}{8^3 - 8}\right\}$$

$$= 1 - \frac{6\left\{25.5 + \frac{1}{12}(6) + \frac{1}{12}(24)\right\}}{512 - 8}$$

$$= 1 - \frac{6\{25.5 + 0.5 + 2\}}{504} = 1 - \frac{6(28)}{504}$$

$$= 1 - \frac{168}{504} = 1 - 0.33 = 0.67 \text{ approx.}$$

Thus, the rank correlation coefficient between the sales, and profits is 0.67 which signifies a positive correlation between the two variables.

ILLUSTRATION 27. ***From the following data relating to the costs and profits of a concern, find out the rank coefficient of correlation, and interpret the result.***

Year	1	2	3	4	5	6	7	8	9
Cost	50	60	65	50	55	60	60	30	40
Profit	10	20	25	15	20	30	35	5	7

SOLUTION

Calculation of the rank coefficient of correlation

Cost X	Profit Y	Ranks in ascending order R_x	Ranks in ascending order R_y	Difference between Ranks $(R_x - R_y) = D$	Square of Difference D^2
50	10	3.5 m_1	3	0.5	0.25
60	20	7 m_2	5.5 m_3	1.5	2.25
65	25	9	7	2.0	4.00
50	15	3.5 m_1	4	–0.5	0.25
55	20	5	5.5 m_3	–0.5	0.25
60	30	7 m_2	8	–1.0	1.00
60	35	7 m_2	9	–2.0	4.00
30	5	1	1	0.0	0.00
40	7	2	2	0.0	0.00
Total	–	–	–	0.0	12.00

By the relevant formula of rank coefficient of correlation we have,

$$r_{(S)} = 1 - \frac{6\left\{SD^2 + \frac{1}{12}(m_1^3 - m_1) + \frac{1}{12}(m_2^3 - m_2) + \frac{1}{12}(m_3^3 - m_3)\right\}}{12}$$

Substituting the derived values in the above formula we get,

$$r_{(S)} = 1 - \frac{6\left\{12 + \frac{(2^3 - 12)}{12} + \frac{(3^3 - 3)}{12} + \frac{(2^3 - 2)}{12}\right\}}{9^3 - 9}$$

$$= 1 - \frac{6\left\{12 + \frac{2(2^2 - 1)}{12} + \frac{3(3^2 - 1)}{12} + \frac{2(2^2 - 1)}{12}\right\}}{9(9^2 - 1)}$$

$$= 1 - \frac{6\{12 + 0.5 + 2 + 0.5\}}{720} = 1 - \frac{6 \times 15}{720}$$

$$= 1 - \frac{90}{720} = 1 - \frac{1}{8} = 1 - 0.125 = -0.875.$$

The above result of the rank coefficient of correlation shows that there is a very high degree of positive correlation between the cost and profit variables.

Special Features of Rank Correlation

From the above analysis and illustrations, the special features of Spearman's rank coefficient of correlation can be outlined as under :

1. The value of such coefficient of correlation lies between +1 and –1.
2. The sum of the differences between the corresponding ranks *i.e.* $\Sigma D = 0$.
3. It is independent of the nature of distribution from which the sample data are collected for calculation of the co- efficient.
4. It is calculated on the basis of the ranks of the individual items rather than their actual values.
5. Its result equals with the result of Karl Pearson's co-efficient of correlation unless there is repetition of any rank. This is because, Spearman's correlation is nothing more than the Pearson's coefficient of correlation between the ranks.

Merits and Demerits

Like any other method, this method of measuring the correlation has also certain merits and demerits which can be outlined as under :

Merits

1. In comparison to Karl Pearson's method, this method is much easy to understand, and simple to calculate.
2. This method can be applied to the phenomena of qualitative nature viz. honesty, beauty, efficiency etc. which can be ranked in some order.
3. This method is not affected by the extreme items.
4. This method is considered indispensable when the data are given in the form of ranks rather than their real values.
5. This method does not need the assumption that the population from which the samples are taken should be parametric, or normally distributed.

Demerits

1. This method is not suitable for frequency distributions *i.e. grouped data.*
2. This method is not suitable when the number of pairs of the variables is larger because, the work of ranking in that case becomes very much cumbersome.
3. The result obtained by this method differs from that of Pearson's method when there are repetitions of the ranks.
4. This method is not capable of further algebraic treatment like that of the Pearson's method.
5. This method is not based on the original values of the observations.

(v) CONCURRENT DEVIATION METHOD

Correlation Coefficient by concurrent deviation method indicates whether the correlation is in positive or in negative direction especially in the short-term fluctuated data. This method is a development over the rank correlation method in the sense that its process of calculation is the simplest, and shortest of all the algebraic methods discussed above. Any number of observations can be easily solved under this method.

Special Features

The special features of this method are as follows :

1. For both the variables of X and Y, deviations of each of the succeeding values from its immediately proceeding value is noted in terms of the direction of changes (*i.e.*, +, –, or 0) under two separate columns styled *dx* and *dy*.
2. Each of the pairs of deviation signs thus noted under dx and dy are multiplied in a separate column styled *dxdy*.
3. Only the positive signs *i.e.* the products of the concurrent, or similar signs in the *dxdy* column are totalled, and shown as the value of the concurrent deviations, or C.
4. The number of pairs of deviations rather than the number of pairs of observations is taken as the value of n, and for this, the number of pairs of observation is reduced by one in as much as the first pair of observations does not fall under the pairs of the deviations.
5. The following formula is applied to find both the degree and direction of the correlation which always lies between ± 1.

$$r_{(c)} = \pm\sqrt{\pm\frac{2C-n}{n}}$$

Where,

$r_{(c)}$ = coefficient of concurrent deviation.

C = number of concurrent deviations and

n = number of pairs of deviation.

The purpose of putting symbol signs inside the root is to convert the negative value of $\frac{2C-n}{n}$, if any, into the positive value by multiplying the same with the minus sign in order that the root of $\frac{2C-n}{n}$ can be found out algebraically.

Further, the purpose of putting ± signs outside the root is to show the ultimate result with its original sign by multiplying the derived result with the minus sign again. In case $\frac{2C-n}{n}$ gives a positive value, all such multiplications cited above will not be necessary.

ILLUSTRATION 28. ***Find out the coefficient of concurrent deviation from the following observations.***

X :	109	122	96	142	151	124	125	102	109	156	122
Y :	14.9	6.3	5.8	12.2	33.2	13.3	14.6	8.8	4.9	39.8	6.3

SOLUTION

Computation of the coefficient of concurrent deviations

X	Y	± Dvns. of X *dx*	± Dvns. of Y *dy*	Product of Dvns. *dxdy*
109	14.9			
122	6.3	+	–	–
96	5.8	–	–	+

142	12.2	+	+	+
151	33.2	+	+	+
124	13.3	–	–	+
125	14.6	+	+	+
102	8.8	–	–	+
109	4.9	+	–	–
156	39.8	+	+	+
122	6.3	–	–	+
Total	N = 11	$n = 10$	–	C = 8

By the formula of concurrent deviations we have,

$$r_{(c)} = \pm\sqrt{\pm\frac{2C-n}{n}}$$

Where, $r_{(c}$ = coefficient of concurrent deviations

C = number of concurrent deviations or positive signs of the products *i.e.* 8

and n = number of pairs of deviations *i.e.* N – 1 or 11 – 1 = 10.

Substituting the respective values in the formula cited above we get,

$$r_{(c)} = \sqrt{\pm\frac{2(8)-10}{10}} = \sqrt{\pm\frac{16-10}{10}} = \pm\sqrt{0.6} = 0.774 = 0.77 \text{ approx.}$$

The above result indicates that there is a high degree of positive correlation between the two variables X, and Y.

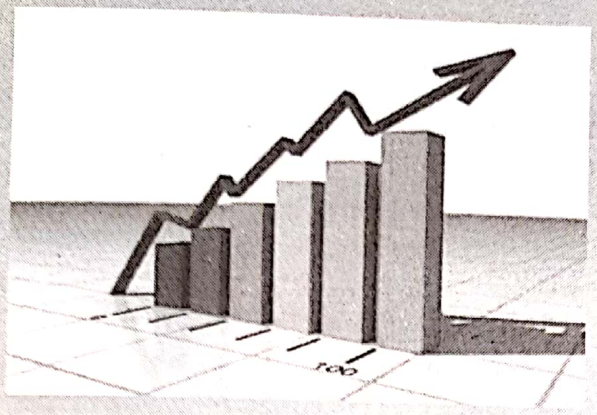
ILLUSTRATION 29. ***Determine the coefficient of concurrent deviations from the following data.***

Months :	Jan.	Feb.	Mar.	April	May	June	July	Aug.	Sept.	Oct.
Index of Calcutta price :	169	182	182	192	198	209	227	238	250	253
Index of Delhi price :	204	222	225	228	229	233	249	266	255	255

SOLUTION

Computation of the coefficient of concurrent deviations

Index of Calcutta price X	Index of Delhi price Y	± Dvns. of X variable *dx*	± Dvns. of Y variable *dy*	Product of deviations *dx dy*
169	204			
182	222	+	+	+
182	225	0	+	0
192	228	+	+	+
198	229	+	+	+
209	233	+	+	+



9

Regression Analysis

9.1. MEANING, DEFINITIONS AND CHARACTERISTICS

Meaning

Although, lexically the term 'regression' means 'going back', or 'stepping down, the **regression analysis is a statistical tool for measuring the average relationship between any two, or more closely related (positively, or negatively) variables in terms of the original units of their data.**

It is advantageously used by the statisticians in estimating the unknown values of a dependent variable say Y from the known values of an independent variable say X. This technique is extensively used as a formidable instrument in almost all the sciences *viz.,* Natural science, Physical science, and Social science. Particularly, in the fields of business and economics that come under the social science, this technique is invariably used for studying the relationship between two, or more related variables viz., Price and Demand, Demand and Supply, Production and Consumption, Expenditure on Advertisement and Volume of Sales, Cost, Volume and Profit etc.

This technique was developed by the British Biometrician Sir Francis Galton in 1877 in course of his studying the relationship between the heights of fathers and the hights of sons. He used this term' regression' for the first time in his paper 'Regression towards Mediocrity in Hereditary Stature" in which he established :

(*i*) that tall fathers will have tall sons, and short fathers will have short sons;

(*ii*) that the average height of the tall fathers' sons will be less than the average height of their fathers ;

(*iii*) that the average height of the short fathers' sons will be more than the average height of their fathers ; and

(*iv*) that the deviations of the mean height of the sons will be less than the deviations of the mean height of the fathers from the mean height of the race, or that when the fathers' height move above or below the mean, the sons' height tend to go back (regress) towards the mean.

Professor Galton studied the average relationship between the above two variables (*i.e.* the heights of the fathers and the heights of the sons) graphically, and named the line describing the relationship, the 'Line of Regression'.

Definitions

The technique of regression analysis introduced as above has been defined variously by various authors. Some of the important and meaningful definitions are reproduced here, as under :

1. In the words of Sir **Francis Galton**, the regression analysis is defined as "the law of regression that tells heavily against the full hereditary transmission of any gift...the more bountifully the parent is gifted by nature, the more rare will be his good fortune if he begets a son who is richly endowed as himself, and still more so if he has a son who is endowed yet more largely."
2. According to **Taro Yamane**, "One of the most frequently used techniques in economics and business research to find a relation between two or more variables that are related causally, is regression analysis."
3. In the words of **Ya Lun Chou,** "Regression analysis attempts to establish the nature of the relationship between variables that is to study the functional relationship between the variables and thereby provide mechanism for prediction or forecasting".

Characteristics

From the above definitions, the essential characteristics of regression analysis can be brought out as under :

(*i*) It consists of mathematical devices that are used to measure the average relationship between two, or more closely related variables.

(*ii*) It is used for estimating the unknown values of some dependent variable with reference to the known values of its related independent variables.

(*iii*) It provides a mechanism for prediction or forecast of the values of one variable in terms of the values of the other variable.

(*iv*) It consists of two lines of equation *viz.*, (*i*) equation of X on Y and (*ii*) equation of Y on X.

9.2. UTILITIES AND LIMITATIONS OF REGRESSION ANALYSIS

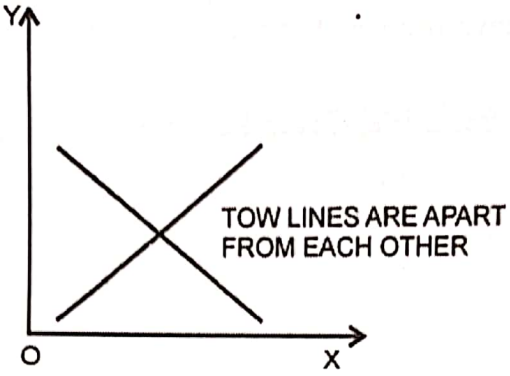
Utilities

The regression analysis as a statistical tool has a number of uses, or utilities for which it is widely used in various fields relating to almost all the natural, physical and social sciences. The specific uses, or utilities of such a technique may be outlined as under :

1. It provides a functional relationship between two or more related variables with the help of which we can easily estimate or predict the unknown values of one variable from the known values of another variable.
2. It provides a measure of errors of estimates made through the regression lines. A little scatter of the observed (actual) values around the relevant regression line indicates good estimates of the values of a variable, and less degree of errors involved therein. On the other hand, a great deal of scatter of the observed values around the relevant regression line indicates inaccurate estimates of the values of a variable and high degree of errors involved therein.
3. It provides a measure of coefficient of correlation between the two variables which can be calculated by taking the square root of the product of the two regression coefficients *i.e.*

$$r = \sqrt{b_{xy} . b_{yx}} .$$

(*v*) When there is a low degree of correlation.



The regression line of Y on X will help us in estimating the value of Y for any value of X, and the regression line of X on Y will help us in estimating the value of X for any value of Y.

Line of Regression

Thus, a line of regression is a graphic line which gives the best estimates of one variable for any given value of the other variable. Such a line can be drawn on a graph paper by any of the following two methods :

(*a*) Scatter diagram method

(*b*) Method of Least square.

These methods of drawing the lines of regression are explained as below :

(a) SCATTER DIAGRAM METHOD

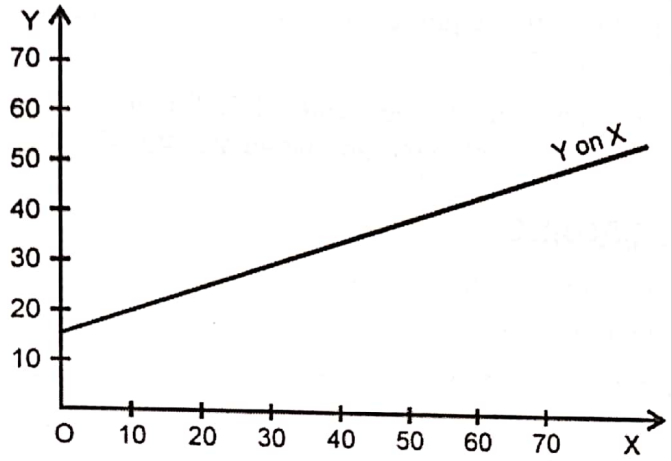
Under this method a graph paper is taken on which the independent variable say, X is represented along the horizontal axis, and the dependent variable say, Y is represented along the vertical axis. The points are then plotted on the graph paper representing the various pair of values of both the variables X and Y which give the picture of a scatter diagram with several points scattered around. After this, two free–hand straight lines are drawn across the scattered points in such a manner that sum of the deviations of the points on one side of a line is equal to sum of the deviations of the points on its other side. The line which is drawn in between such vertical deviations is represented as the regression line of Y on X and the line which is drawn in between such horizontal deviations is represented as the line of regression of X on Y. The point at which both these regression lines cut each other represents the Mean of the two variables. However, the drawal of the regression lines in such a free hand manner involves a great deal of difficulties for which a piece of thread is to be repeatedly adjusted to see that the sum of the deviations on both the sides of the thread is equal. The procedure of drawal of the regression lines under the above method is illustrated as under :

ILLUSTRATION 1. ***Using the scatter diagram method, draw the two regression lines associated with the following data both separately and jointly.***

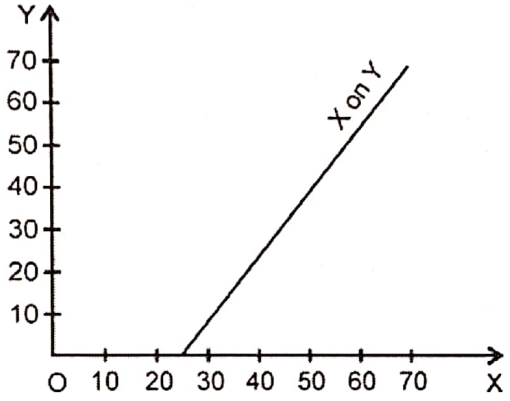
Variable X :	**40**	**50**	**60**	**40**	**45**	**50**	**70**	**50**	**55**
Variable Y :	30	30	50	35	30	40	50	40	35

SOLUTION

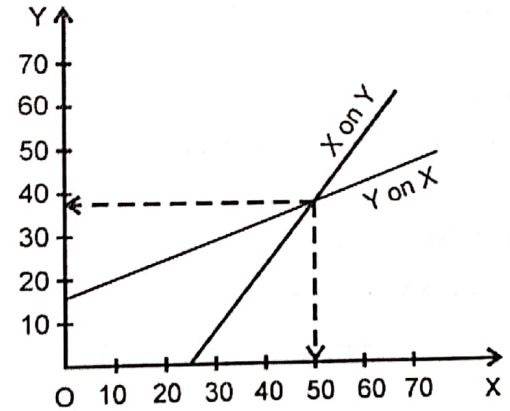
(*i*) Diagrammatic representation of the regression line of Y on X



(*ii*) Diagrammatic representation of the regression line of X on Y



(*iii*) Diagrammatic representation of both the regression lines.



From the intersection point of the two regression lines in the above graph, it must be seen that the mean value of X variable is 51 and that of the Y variable is 38 approx. This can be verified by the algebraic method of arithmetic average as follows :

$$\overline{X} = \frac{\Sigma X}{N} = \frac{460}{9} = 51 \text{ approx , and } \overline{Y} = \frac{\Sigma Y}{N} = \frac{340}{9} = 38 \text{ approx.}$$

If, we wish to estimate any value of Y for a given value of X, we can do so easily by drawing a perpendicular from the required value point of the X axis to the point at which it cuts the regression line of Y on X, and then by drawing a perpendicular therefrom to Y axis. The point at which the perpendicular touches the Y axis is the required value of Y for the given value of X. Thus, in the above graph it may be seen that when X = 70, Y = 45.

In the similar manner we can estimate the value of X for any value of Y with reference to the regression line of X and Y. Thus, in the above graph when Y = 20, X = 30.

(*b*) METHOD OF LEAST SQUARE

This method is a development over the scatter diagram method discussed above in which we face a lot of difficulties in drawing the regression lines accurately in a free–hand manner.

Under this method we are to draw **the lines of best fit as the lines of regression**. These, lines of regression are called the lines of the best fit because, with reference to these lines we can get the best estimates of the values of one variable for the specified values of the other variable. Further, this method is called as such because, under this method the sum of the squares of the deviations between the given values of a variable and its estimated values given by the concerned line of regression is the least or minimum possible.

Under this method, **the line of the best fit for Y on X (*i.e.* the regression lines of Y on X) is obtained by finding the value of Y for any two (preferably the extreme ones) values of X through the following linear equation :**

$$Y = a + bX,$$

where a and b are the two constants whose values are to be determined by solving simultaneously the following two normal equations :

$$\Sigma XY = a\,\Sigma X + b\,\Sigma X^2 \qquad \text{....(ii)}$$

where, X and Y represent the given values of the X and Y variables respectively.

Further, **the line of the best fit for X on Y (*i.e.* the regression line of X on Y) is obtained by finding the values of X for any two (preferably the extreme ones) values of Y** through the following linear equation :

$$X = a + bY,$$

where, the values of the two constants a and b are determined by solving simultaneously the following two normal equations :

$$\Sigma X = Na + b\,\Sigma Y \qquad \text{....(i)}$$

$$\Sigma XY = a\,\Sigma Y + b\,\Sigma Y^2 \qquad \text{....(ii)}$$

As pointed out earlier, in order to determine the value of Y for a given value of X, we will draw a perpendicular from the given value point of the X axis to the Y axis via the point at which it cuts the regression line of Y on X. The point at which the Y axis is touched by such perpendicular will indicate the required value of Y for the given value of X. In the similar manner, the required value of X for any given value of Y can be determined by drawing a perpendicular from the concerned point of the Y axis to the X axis via the point at which it cuts the regression line of X on Y.

The following illustration will show how regression lines are drawn and values of a variable are estimated under the method of least square explained above :

ILLUSTRATION 2. *Draw the two regression lines for the data given below using the method of least square.*

Variable X :	5	10	15	20	25
Variable Y :	20	40	30	60	50

SOLUTION

Determination of the regression lines by the method of least square

X	Y	X^2	Y^2	XY
5	20	25	400	100
10	40	100	1600	400
15	30	225	900	450
20	60	400	3600	1200
25	50	625	2500	1250
$\Sigma X = 75$	$\Sigma Y = 200$	$\Sigma X^2 = 1375$	$\Sigma Y^2 = 9,000$	$\Sigma XY = 3400$

(*i*) Regression Line of Y on X

This is given by $Y = a + bX$

where a and b are the two constants which are found by solving simultaneously the two normal equations as follows :

$$\Sigma Y = Na + b\,\Sigma X \quad \text{....}(i)$$

$$\Sigma XY = a\,\Sigma X + b\,\Sigma X^2 \quad \text{....}(ii)$$

Substituting the given values in the above equations we get,

$$200 = 5a + 75b \quad \text{....}(i)$$

$$3400 = 75a + 1375b \quad \text{....}(ii)$$

Multiplying the eqn. (*i*) by 15 under the eqn. (*iii*) and subtracting the same from the eqn. (*ii*) we get,

$$3400 = 75a + 1375b \quad \text{....}(ii)$$

$$(-)\ 3000 = 75a + 1125b \quad \text{....}(iii)$$

Thus, $400 = 250b$

$\therefore$ $$b = \frac{400}{250} = 1.6$$

Putting the above value of b in the eqn. (*i*) we get,

$$200 = 5a + 75(1.6)$$

or $$5a = 200 - 120$$

or $$a = \frac{80}{5} = 16.$$

Thus, $a = 16$, and $b = 1.6$

Putting these values in the equation $Y = a + bX$ we get

$$Y = 16 + 1.6X$$

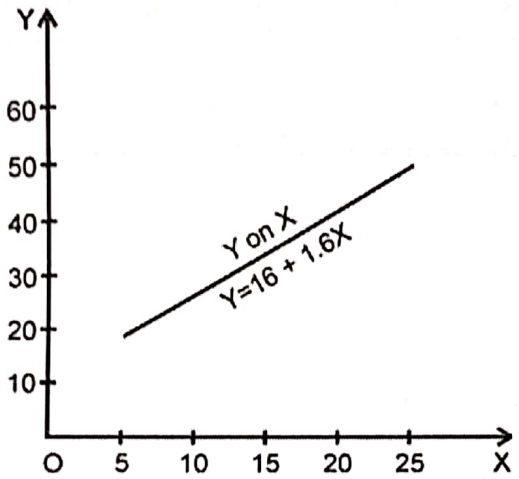
With this equation, the estimated values of Y for the two extreme values of X will be :

When $X = 5,\ Y = 16 + 1.6(5) = 24$

and when $X = 25,\ Y = 16 + 1.6(25) = 56$

With these two pairs of values *viz.* (5,24) and (25, 56) the regression line of Y on X will be drawn as under :

Regression line of Y on X



(*ii*) Regression Line of X on Y

This is given by $X = a + bY$

where a and b are the two constants whose values are determined by solving the two normal equations as follows :

$$\Sigma X = N a + b \Sigma Y \quad \text{....(i)}$$

$$\Sigma XY = a \Sigma Y + b \Sigma Y^2) \quad \text{....(ii)}$$

Substituting the given values in the above equations we get,

$$75 = 5a + 200b \quad \text{....(i)}$$

$$3400 = 200a + 9000b \quad \text{....(ii)}$$

Multiplying the eqn. (*i*) by 40 under the eqn. (*iii*) and getting the same subtracted from the eqn. (*ii*) we get,

$$3400 = 200a + 9000\ b \quad \text{...(ii)}$$

$$\underline{(-)\ 3000 = 200a + 8000b} \quad \text{...(iii)}$$

Thus, $400 = 1000b$

$\therefore$ $$b = \frac{400}{1000} = 0.4$$

Putting the above value of b in the equation (*i*) we get,

$$75 = 5a + 200(.4)$$

or $$5a = 75 - 80 = -5$$

$\therefore$ $$a = \frac{-5}{5} = -1$$

Hence, $a = -1$, and $b = 0.4$

Putting these values in the equation $X = a + bY$ we get

$$X = -1 + 0.4Y$$

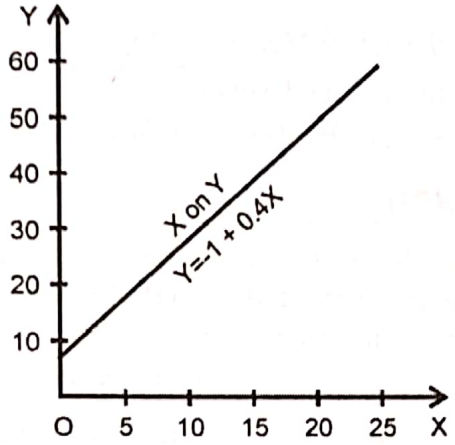
With this equation, the estimated values of X for the two extreme values of Y will be :

When $Y = 20,\ X = -1 + 0.4(20) = 7$

and When $Y = 60,\ X = -1 + 0.4(60) = 23$

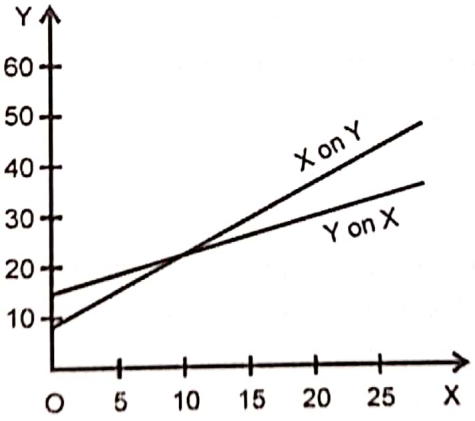
With these two pairs of values *viz.*, (20, 7) and (60, 23) the regression line of X on Y will be drawn as follows :

Regression line of X on Y



When both the regression lines of Y on X, and X on Y will be represented together, the graph will appear as under :

Regression lines of Y on X, and of X on Y



Characteristics of a Straight Line of Regression

The chief characteristics of a straight line of regression fitted by the method of least square explained as above may be noted as follows :

(*i*) It goes through the overall Mean of a variable.

(*ii*) It gives the best fit to the data because the sum of the squared deviations from the line is smaller than they would be from any other straight line.

(*iii*) The sum of the positive and negative deviations from this line is zero.

(*iv*) This line gives the best estimate of the population regression when the data represent a sample drawn from a large population.

2. ALGEBRAIC METHOD

Under this method the two regression equations are formulated to represent the two regression lines, or the lines of estimates respectively *viz.*,

(*i*) The regression line of X on Y and

(*ii*) The regression line of Y on X.

To obtain such equations we are to apply any of the following algebraic methods :

(*i*) Normal equation method.

(*ii*) Method of deviation from the actual Means ; and

(*iii*) Method of deviation from the assumed Means.

The procedure of each of the above methods is explained in detail as under :

1. NORMAL EQUATION METHOD

This method is just similar to that of the method of least square explained above except that the required values of a variable are estimated directly by the formulated equations rather than through the lines of estimates drawn on a graph paper.

Thus, under this method, the two regression equations *viz.* :

$$(i)\ Y = a + bX \text{ and } (ii)\ X = a + bY$$

are obtained on the basis of the two normal equations cited under the method of least square explained earlier. Here, the regression equation.

$$Y_e = a + bX$$

is constructed to compute the estimated values of the Y variable for any given values of the X variable.

In this formula the various factors carry the meanings as follows :

Y_e = estimated value of Y variable

a = Y-intercept at which the regression line crosses the Y- axis *i.e.* the vertical axis. Its value remains constant for any given straight line.

b = Slope of the straight line and it represents a change in the Y variable for a unit change in the X variable. Its value remains constant for any given straight line.

X = a given value of the X variable for which the value of Y is to be computed.

The above equation $Y_e = a + bX$ is to be formulated on the basis of the following two normal equations :

$$\Sigma Y = N a + b \Sigma X \quad ...(i)$$

$$\Sigma XY = a \Sigma X + b \Sigma X^2 \quad ...(ii)$$

In these formulae, Σ = total of the given values of Y variable,

ΣX = total of the given values of X variable,

ΣX^2 = total of the squares of the given values of X variable,

ΣXY = total of the product of the given values of X and Y variable, and

a and b = the two constants explained above.

The above two normal equations are to be solved simultaneously to determine the values of the two constants a and b that appear as the essential factors in the regression equation

$$Y_e = a + bX.$$

Similarly, the other equation, $X_e = a + bY$ is developed to compute the estimated values of the X variable for any given values of the Y variable.

In this formula,

X_e = estimated value of the X variable,

a = X-intercept at which the regression line crosses the X- axis *i.e.* the horizontal axis. Its value remains constant for any given straight line.

b = Slope of the straight line and it represents a change in the X variable for a unit change in the Y variable. Its value also remains constant for any given straight line, and

Y = a given value of the Y variable for which the value of X is to be computed.

The above equation $X_e = a + bY$ is to be formulated on the basis of the following two normal equations :

$$\Sigma X = N a + b \Sigma Y \quad ...(i)$$

$$\Sigma XY = a \Sigma Y + b \Sigma Y^2 \quad ...(ii)$$

In these formulae, the various factors involved carry the same meaning as explained earlier and

ΣY^2 = total of the squares of the given values of the Y variable.

The following illustration will show how the two regression equations are formed under the method of normal equation :

ILLUSTRATION 3. ***From the following data form the regression equations,***

$$Y_e = a + bX \quad \textit{and} \quad X_e = a + bY$$

Use the normal equation method :

X :	1	3	5	7	9
Y :	15	18	21	23	22

Also, estimate the value of Y when X = 4, and the value of X when Y = 24.

SOLUTION

Formulation of the regression equations

	X	Y	X^2	Y^2	XY
	1	15	1	225	15
	3	18	9	324	54
	5	21	25	441	105
	7	23	49	529	161
	9	22	81	484	198
Total	$\Sigma X = 25$	$\Sigma Y = 99$	$\Sigma X^2 = 165$	$\Sigma Y^2 = 2003$	$\Sigma XY = 533$, $N = 5$

(*i*) Regression equation of X on Y. This is given by

$$X_e = a + bY$$

To find the values of the constants a and b in the above formula, the following two normal equations are to be simultaneously solved :

$$\Sigma X = N a + b \Sigma Y \quad ...(i)$$

$$\Sigma XY = a \Sigma Y + b \Sigma Y^2 \quad ...(ii)$$

Substituting the respective values in the above formula we get,

$$25 = 5a + 99b \quad ...(i)$$

$$533 = 99a + 200\ 3b \quad ...(ii)$$

Multiplying the equation (i) by 99 and eqn. (ii) by 5 and presenting them in the form of a subtraction we get,

$$2475 = 495\ a + 9801\ b \quad ...(iii)$$

$$\underline{(-)\ 2665 = 495\ a + 10015\ b} \quad ...(iv)$$

Thus $-190 = -214b$

or $214b = 190$

$$\therefore \quad b = \frac{190}{214} = .888 \text{ approx.}$$

Putting the above values of b in the eqn. (i) we get,

$$25 = 5a + 99\ (.888)$$

or $5a = 25 - 87.912 = -62.912$

$$\therefore \quad a = \frac{-62.912}{5} = -12.5824$$

Thus, $a = -12.5824$ and $b = 0.888$.

Substituting the above values of the constants a and b, we get the regression equation of X on Y as,

$$X_e = -12.5824 + 0.888Y$$

Thus, when $Y = 24$, $X_e = -12.5824 + 0.888\ (24)$

$$= -12.5824 + 21.312$$

$$= 8.7296.$$

(*ii*) Regression equation of Y on X. This is given by

$$Y_e = a + bX$$

To find the values of the constants a and b in the above formula, the following two normal equations are to be simultaneously solved as under :

$$\Sigma Y = N a + b \Sigma X \quad ...(i)$$

$$\Sigma XY = a \Sigma X + b \Sigma X^2 \quad ...(ii)$$

Substituting the respective values in the above formula we get,

$$99 = 5a + 25b \quad ...(i)$$

$$533 = 25a + 165\ b \quad ...(ii)$$

Multiplying the equation (1) by 5 and getting the same subtracted from the equation (ii) *we get,*

$$533 = 25\ a + 165b \quad ...(ii)$$

$$\underline{(-)\ 495 = 25\ a + 125\ b} \quad ...(iii)$$

Thus, $38 = 40b$

$\therefore \quad b = \frac{38}{40} = 0.95$

Putting the above values of b in the equation (i) we get,

$99 = 5a + 25\,(0.95)$

or $5a = 99 - 23.75 = 75.25$

$\therefore \quad a = \frac{75.25}{5} = 15.05$

Thus, $a = 15.05$ and $b = 0.95$

Substituting the above values of the two constants a and b we get the regression equation of Y on X as,

$Y_e = 15.05 + 0.95X$

Thus, when $X = 4$, $Y_e = 15.05 + 0.95\,(4)$

$= 15.05 + 3.80$

$= 18.85.$

Note. It may be noted that the above normal equation method of formulating the two regression equations is very lengthy and tedious. In order to do away with such difficulties, any of the following two methods of deviation may be used advantageously.

2. METHOD OF DEVIATION FROM THE ACTUAL MEANS

Under this method, the two regression equations are developed in a modified form from the deviation figures of the two variables from their respective actual Means rather than their actual values. For this, the two regression equations are modified as under :

(*i*) Regression equation of X on Y. This is given by

$$X = \overline{X} + b_{xy}\,(Y - \overline{Y}) \quad \text{or} \quad X - \overline{X} = b_{xy}\,(Y - \overline{Y})$$

(*ii*) Regression equation of Y on X. This is given by

$$Y = \overline{Y} + b_{yx}\,(X - \overline{X}) \quad \text{or} \quad Y - \overline{Y} = b_{yx}\,(X - \overline{X})$$

In the above formulae,

X = given value of the X variable

Y = given value of the Y variable

$\overline{X}$ = arithmetic average of the X variable,

$\overline{Y}$ = arithmetic average of the Y variable,

b_{xy} = regression coefficient of X on Y *i.e.* $r\,\frac{\sigma_x}{\sigma_y}$

$\therefore \quad \boxed{b_{xy} = r\,\frac{\sigma_x}{\sigma_y}}$

where,

r = correlation coefficient,

σ_x = standard deviation of X variable

σ_y = standard deviation of Y variable

And b_{yx} = regression coefficient of Y on X *i.e* $.r\frac{\sigma_y}{\sigma_x}$

$\therefore$ $\boxed{b_{yx} = r\frac{\sigma_y}{\sigma_x}}$

Regression Coefficient

A regression coefficient is a vital factor that measures the change in the value of one variable with respect to a unit change in the value of another variable. From the above formulae, it must be observed that the regression equation of X on Y is formed on the basis of its regression coefficient *i.e.* b_{xy} , or $r\frac{\sigma_x}{\sigma_y}$ which measures the change in the value of X variable for a unit change in the value of Y variable.

Similarly, the regression equation of Y on X is formed on the basis of its regression coefficient *i.e.* b_{yx} or , $r\frac{\sigma_y}{\sigma_x}$ which measures the change in the value of Y variable for a unit change in the value of X variable.

To simplify the process of finding the above two regression coefficients, the following formulae may be substituted in place of the given ones :

(*i*) b_1 or $b_{xy} = \frac{\Sigma xy}{\Sigma y^2}$; and

(*ii*) b_2 or $b_{yx} = \frac{\Sigma xy}{\Sigma x^2}$.

In the above formulae,

b_1 or b_{xy} = regression coefficient of X on Y

b_2 or b_{yx} = regression coefficient of Y on X

Σxy = total of the products of deviations of the X and Y variables from their respective actual Means,

Σy^2 = total of the squares of the deviations of Y variable from its actual Mean

Σx^2 = total of the squares of the deviations of the X variable from its actual Mean.

The above two simplified formulae of the regression coefficients have been derived as under :

(*i*) $$b_{xy} = r\frac{\sigma_x}{\sigma_y} = \frac{\Sigma xy}{N\sigma_x\sigma_y} \times \frac{\sigma_x}{\sigma_y}$$

$$= \frac{\Sigma xy}{N\sigma_y^2} = \frac{\Sigma xy}{N \times \frac{\Sigma y^2}{N}} = \frac{\Sigma xy}{\Sigma y^2}$$

Also, $$= \frac{\Sigma xy / N}{\Sigma y^2 / N} = \frac{\text{Co-variance } xy}{\sigma_y^2}$$

(*ii*) $$b_{yx} = r \frac{\sigma_y}{\sigma_x} = \frac{\Sigma xy}{N\sigma_x\sigma_y} \times \frac{\sigma_y}{\sigma_x}$$

$$= \frac{\Sigma xy}{N\sigma_x^2} \times \frac{\Sigma xy}{N\frac{\Sigma x^2}{N}} = \frac{\Sigma xy}{\Sigma x^2}$$

Also, $$b_{xy} = \frac{\Sigma xy / N}{\Sigma x^2 / N} = \frac{\text{Co-variance } xy}{\sigma_x^2}$$

To simplifying further the process of finding the two regression coefficients, the following formula may be used advantageously, if the given values of the variables are of small and sound size

(*i*) $$b_{xy} = \frac{N\Sigma XY - \Sigma X \,.\, \Sigma Y}{N\Sigma Y^2 - (\Sigma Y)^2}$$

(*ii*) $$b_{yx} = \frac{N\Sigma XY - \Sigma X \,.\, \Sigma Y}{N\Sigma X^2 - (\Sigma X)^2}$$

In the above formula,

s of pairs of observation and all other factors carry the same meanings as before.

Properties of Regression Coefficients

The regression coefficients expalined as above have a number of valuable properties which may be cited as under :

1. The geometric Mean of the two regression coefficients gives the coefficients of correlation *i.e.*

$$r = \sqrt{b_{xy} \cdot b_{yx}}$$

Proof :

$\therefore$ $$b_{xy} \cdot b_{yx} = r \cdot \frac{\sigma_x}{\sigma_y} r \cdot \frac{\sigma_y}{\sigma_x} = r^2$$

$\therefore$ $$\sqrt{b_{xy} \cdot b_{yx}} = \sqrt{r^2} = r$$

2. **Both the regression coefficients must have the same algebraic sign *i.e.* both of them will have either + signs or – signs.** This is because, minus sign with one coefficient will make the product of the two coefficients minus and in that case we cannot find out its square root to obtain the correlation coefficient. **Thus, the first property mentioned above will be vitiated.**

3. The nature of the regression coefficients is reflected on the nature of the coefficient of correlation. This means that if the regression coefficients are positive the correlation coefficient will be

positive, and if the regression coefficients are negative then the correlation coefficient will be negative. Thus if $b_{xy} = -.5$ and $b_{yx} = -1.5$, then r would be $\sqrt{-5 \times -1.5} = -0.86$ and not + 0.86.

4. If one of the regression coefficients is greater than unity or 1, the other must be less than unity. This is because the value of coefficient of correlation 'r' must be in between ± 1. *i.e.* $r = \sqrt{b_{xy} \cdot b_{yx}}$ in between ±1. If both the regression coefficients happen to be more than 1, then their geometric mean will exceed 1 which will not give the correlation coefficient whose value never exceeds 1.

5. The arithmetic mean of the regression coefficients is either equal to or more than the correlation coefficient *i.e.* $\frac{b_{xy} + b_{yx}}{2} \geq r$

Proof : We know that $\overline{X} \geq$ G.M.

$$\Rightarrow \qquad \frac{b_{xy} + b_{yx}}{2} \geq \sqrt{b_{xy} \cdot b_{yx}} \text{ or } r.$$

6. From the regression coefficients we can find out the value of any factor forming part of it, if the value of the other 3 factors are given. Thus, if we are given, r=0.5, $\sigma_y = 3$, and $b_{yx} = 0.6$ we can find out the value of σ_x as under :

$$b_{yx} = r\frac{\sigma_y}{\sigma_x} \quad \therefore \quad \sigma_x = \frac{r\sigma_y}{b_{yx}} = \frac{0.5 \times 3}{0.6} = 2.5$$

Similarly, if we are given, $\sigma_x = 4, \sigma_y = 10$ and $b_{xy} = 0.2$, we can find out the value of r as follows :

$$b_{xy} = r\frac{\sigma_x}{\sigma_y} \qquad \therefore \qquad r = \frac{b_{xy} \cdot \sigma_y}{\sigma_x} = \frac{0.2 \times 10}{4} = 0.5$$

7. **Regression coefficients are independent of change of origin but not of scale.** This means that if the original values of the two variables are added or subtracted by some constant, the values of the regression coefficients will remain the same. But if the original values of the two variables are multiplied, or divided by some constant (common factors) the values of the regression equation will not remain the same.

The following illustrations will show how the two regression equations are formed under the method of deviation from the actual Means.

ILLUSTRATION 4. ***Using the method of deviations from the actual Means from the data given below find.***

(i) the two regression equations

(ii) the correlation coefficient and

(iii) the most probable value of Y when X = 30

X :	25	28	35	32	31	36	29	38	34	32
Y :	43	46	49	41	36	32	31	30	33	39

SOLUTION

Determination of the regression equations by the method of deviation from the Means

X	Y	(X–32) x	(Y–38) y	x^2	y^2	xy
25	43	–7	5	49	25	–35
28	46	–4	8	16	64	–32
35	49	3	11	9	121	33
32	41	0	3	0	9	0
31	36	–1	–2	1	4	2
36	32	4	–6	16	36	–24
29	31	–3	–7	9	49	21
38	30	6	–8	36	64	–48
34	33	2	–5	4	25	–10
32	39	0	1	0	1	0
ΣX=320	ΣY =380	$\Sigma x = 0$	$\Sigma y = 0$	Σx^2= 140	Σy^2= 398	$\Sigma xy = -93$

(*a*) (*i*) Regression equation of X on Y

This is given by

$$X = \overline{X} + r\frac{\sigma_x}{\sigma_y}(Y - \overline{Y})$$

where

$$\overline{X} = \frac{\Sigma X}{N} = \frac{320}{10} = 32$$

$$\overline{Y} = \frac{\Sigma Y}{N} = \frac{380}{10} = 38$$

$$\sigma_x = \sqrt{\frac{\Sigma x^2}{N}} = \sqrt{\frac{140}{10}} = 3.74 \text{ approx.}$$

$$\sigma_y = \sqrt{\frac{\Sigma y^2}{N}} = \sqrt{\frac{398}{10}} = 6.31 \text{ approx.}$$

And

$$r = \frac{\Sigma xy}{N\sigma_x\sigma_y} = \frac{-93}{10 \times 3.74 \times 6.31}$$

$$= \frac{-93}{10 \times 23.5994} = \frac{-93}{2235.99} = -0.394$$

Putting the respective values in the above equation we get,

$$X = 32 + -0.394 \times \frac{3.74}{6.31}(Y - 38)$$

$$= 32 - 0.2337\,(Y - 38)$$

$$= 32 + 8.8806 - 0.2337Y$$

$\therefore$ $$X = 40.8806 - 0.2337Y$$

Aliter

Substituting the regression coefficient of X on Y *i.e* . $r\frac{\sigma_x}{\sigma_y}$, by $\frac{\Sigma xy}{\Sigma y^2}$,

we get $$X = \overline{X} + \frac{\Sigma xy}{\Sigma y^2}(Y - \overline{Y})$$

$$= 32 + \frac{-93}{398}(Y - 38)$$
$$= 32 - 0.2337\,(Y - 38)$$
$$= 32 + 8.8806 - 0.2337Y$$
$$= 40.8806 - 0.2337Y$$

(*ii*) Regression equation of Y on X

This is given by $Y = \overline{Y} + r\frac{\sigma_y}{\sigma_x}(\overline{X} - \overline{X})$

Substituting the respective values in the above we get,

$$Y = 38 + -0.394 \times \frac{6.31}{3.74}(X - 32)$$
$$= 38 - 0.6643\,(X - 32)$$
$$= 38 + 21.2576 - 0.664\,3X$$
$$= 59.2576 - 0.6643X$$

$\therefore$ $$Y = 59.2576 - 0.6643\,X$$

Aliter

Replacing the formula of regression coefficient. $r\frac{\sigma_y}{\sigma_x}$ by $\frac{\Sigma xy}{\sigma x^2}$ we get,

$$Y = \overline{Y} + \frac{\Sigma xy}{\Sigma x^2}(X - \overline{X})$$

$$= 38 + \frac{-93}{140}(X - 32)$$
$$= 38 - 0.6643\,(X - 32)$$
$$= 38 + 21.2576 - 0.6643X$$

$\therefore$ $$Y = 59.2576 - 0.6643X$$

Thus, the two regression equations are :

X on Y : $X = 40.8806 - 0.2237Y$

and Y on X : $Y = 59.2576 - 0.6643$

(*b*) Coefficient of Correlation

The coefficient of correlation between the two variables, X and Y is given by

$$r_{xy} = \frac{\Sigma xy}{N\sigma_x\sigma_y} = \frac{-93}{10 \times 3,74 \times 6.31} = \frac{-93}{235.994} = -0.394$$

Alternatively

By the method of regression coefficients we have,

$$r_{xy} = \sqrt{b_{xy} \times b_{yx}}$$

$$= \sqrt{(-0.2337) \times (-0.6643)} = -\sqrt{0.1552} = -0.394 \ \ i.e. \ -0.394$$

Note. Since the regression coefficients are negative, the correlation coefficient has been negative.

(c) Probable value of Y when X = 30

This will be determined by the regression equation of Y on X as follows :

We have, $Y=59.2576 - 0.6643X$

Thus, when $X=30$, $Y = 59.2576 - 0.6643(30) = 59.2576 - 19.929 = 39.3286$.

ILLUSTRATION 5. ***From the data given below find,***

(*i*) ***the two regression coefficients,***

(*ii*) ***the correlation coefficient,***

(*iii*) ***the two regression equations,***

(*iv*) ***the standard deviations of X and Y.***

Expenditure on advertisement (in '000 ₹) X :	11	7	9	5	8	6	10
Volume of Sales (in lakhs ₹) Y :	10	8	6	5	9	7	11

Also, find the figure of sales when the expenditure on advertisement is ₹ 15000.

SOLUTION

Regression Analysis

X	Y	(X–8) x	(Y–8) y	x^2	y^2	xy	
11	10	3	2	9	4	6	
7	8	–1	0	1	0	0	
9	6	1	–2	1	4	–2	
5	5	–3	–3	9	9	9	
8	9	0	1	0	1	0	
6	7	–2	–1	4	1	2	
10	11	2	3	4	9	6	
$\Sigma X=56$	$\Sigma Y=56$	$\Sigma x = 0$	$\Sigma y = 0$	$\Sigma x^2 = 28$	$\Sigma y^2=28$	$\Sigma xy=21$	$N=7$

We have,

$$\overline{X} = \frac{\Sigma X}{N} = \frac{56}{7} = 8$$

$$\overline{Y} = \frac{\Sigma Y}{N} = \frac{56}{7} = 8$$

(*i*) (*a*) *Regression coefficient of X on Y*

This is given by $b_{xy} = \dfrac{\Sigma xy}{\Sigma y^2} = \dfrac{21}{28} = \dfrac{3}{4} = 0.75$

(*b*) *Regression coefficient of Y on X*

This is given by $b_{yx} = \frac{\Sigma xy}{\Sigma x^2} = \frac{21}{28} = \frac{3}{4} = 0.75$

(*ii*) ***Correlation Coefficient***

This is given by $r = \sqrt{b_{xy}.b_{yx}} = \sqrt{0.75 \times 0.75} = 0.75$

(*iii*) ***The two Regression Equations***

(*a*) X on Y: $X = \overline{X} + b_{xy}(Y - \overline{Y}$

$= 8 + 0.75\,(Y - 8)$

$= 8 - 6 + 0.75\,Y$

$= 2 + 0.75Y$

$\therefore \quad X = 2 + 0.75Y$

(*b*) Y on X: $Y = \overline{Y} + b_{yx}(X - \overline{X}$

$= 8 + 0.75\,(X - 8)$

$= 8 - 6 + 0.75\,X$

$= 2 + 0.75X$

$\therefore \quad Y = 2 + 0.75X$

(*iv*) Standard Deviation of X

This is given by $\sigma_x = \sqrt{\frac{\Sigma x^2}{N}} = \sqrt{\frac{28}{7}} = 2$

(*v*) Standard Deviation of Y

This is given by $\sigma_y = \sqrt{\frac{\Sigma y^2}{N}} = \sqrt{\frac{28}{7}} = 2$

Alternatively,

$$b_{xy} = r\frac{\sigma_x}{\sigma_y} \text{ or } \sigma_y = r\frac{\sigma_x}{b_{xy}} = 0.75 \times \frac{2}{0.75} = 2$$

(*vi*) Determination of Sales Y when Advertisement Expenditure X is ₹ 15000 :

This will be determined by the regression equation of Y on X as follows :

We have, $Y = 2 + 0.75X$

Thus, when $X = 15,\ Y = 2 + 0.75\,(15) = 2 + 11.25 = 13.25$

$\therefore$ When $X = 15000$, $Y = 13.25 \times 1000 =$ ₹ 13250

ILLUSTRATION 6. ***From the data given below, compute the two regression coefficients, and formulate the two regression equations :***

$$\Sigma X = 510, \Sigma Y = 7140, \Sigma X^2 = 4150, \Sigma XY = 54900,$$

$$\Sigma Y^2 = 740200 \text{ and } N = 102.$$

Also, determine the value of Y when X = 7.

SOLUTION

(*i*) Two Regression Coefficients

By the value based method,

(*a*) $$b_{xy} = \frac{N\Sigma XY - \Sigma X \,.\, \Sigma Y}{N\Sigma Y^2 - (\Sigma Y)^2}$$

Putting the respective values in the above we get,

$$b_{xy} = \frac{102(54900) - (510 \times 7140)}{102(740200) - (7140)^2} = \frac{1958400}{24520800} = \mathbf{0.08}$$

(*b*) $$b_{yx} = \frac{N\Sigma XY - \Sigma X.\Sigma Y}{N\Sigma X^2 - (\Sigma X)^2}$$

Putting the respective values in the above we get,

$$b_{yx} = \frac{102(54900) - (510 \times 7140)}{102\,(4150) - (510)^2} = \frac{1958400}{163200} = \mathbf{12}$$

(*ii*) Two Regression Equations

(*a*) X on Y : $X = \overline{X} + b_{xy}\,(Y - \overline{Y})$

where, $$\overline{X} = \frac{\Sigma X}{N} = \frac{510}{102} = 5$$

and $$\overline{Y} = \frac{\Sigma Y}{N} = \frac{7140}{102} = 70$$

Thus, $X = 5 + 0.08\,(Y - 70)$

$= 5 - 5.6 + .08\,Y$

$\therefore$ $X = -0.6 + 0.08\,Y$

(*b*) Y on X : $Y = (\overline{Y}) + b_{yx}\,(X - \overline{X})$

$= 70 + 12\,(X - 5)$

$= 70 - 60 + 12X$

$= 10 + 12X$

$\therefore$ $Y = 10 + 12X.$

(*iii*) Value of Y, when X = 7

When $X = 7$, $Y = 10 + 12\,(7)$

$= 10 + 84 = 94$

ILLUSTRATION 7. ***From the data given below, obtain the two regression coefficients and the two regression equations using the value based method :***

X :	2	4	6	8	10
Y :	5	7	9	8	11

SOLUTION

Computation of the Regression Coefficients and the Regression Equations by the Value Based Method

X	Y	X^2	Y^2	XY	
2	5	4	25	10	
4	7	16	49	28	
6	9	36	81	54	
8	8	64	64	64	
10	11	100	121	110	
$\Sigma X = 30$	$\Sigma Y = 40$	$\Sigma X^2 = 220$	$\Sigma Y^2 = 340$	$\Sigma XY = 266$	$N = 5$

(*a*) Regression Coefficients

(*i*) $$b_{xy} = \frac{N\Sigma XY - \Sigma X \,.\, \Sigma Y}{N\Sigma Y^2 - (\Sigma Y)^2} = \frac{(5 \times 266) - (30 \times 40)}{(5 \times 340) - (40)^2}$$

$$= \frac{1330 - 1200}{1700 - 1600} = \frac{130}{100} = \mathbf{1.3}$$

(*ii*) $$b_{yx} = \frac{N\Sigma XY - \Sigma X \,.\, \Sigma Y}{N\Sigma X^2 - (\Sigma X)^2} = \frac{(5 \times 266) - (30 \times 40)}{(5 \times 220) - (30)^2}$$

$$= \frac{130}{(1100 - 900)} = \frac{130}{200} = \mathbf{0.65}$$

(*b*) Regression Equations

(i) Regression Equation of X on Y

This is given by $X = \overline{X} + b_{xy}(Y - \overline{Y})$

where,

$$\overline{X} = \frac{\Sigma X}{N} = \frac{30}{5} = 6 \quad \text{and} \quad \overline{Y} = \frac{\Sigma Y}{N} = \frac{40}{5} = 8$$

Thus, $X = 6 + 1.3\,(Y - 8)$

$= 6 - 10.4 + 1.3Y$

$= -4.4 + 1.3Y$

$\therefore \quad X = -4.4 + 1.3\,Y$

(ii) Regression Equation of Y on X

This is given by $Y = \overline{Y} + b_{yx}(X - \overline{X})$

$= 8 + 0.65\,(X - 6)$

$= 8 - 3.9 + 0.65X$

$\therefore \quad Y = 4.10 + 0.65X$

3. METHOD OF DEVIATION FROM ASSUMED MEAN

This method is also otherwise known as short-cut method. **Under this method, the regression between any two related variables is studied on the basis of the deviations of the items from their**

respective assumed Means rather than their actual values or deviations from their respective actual Means.

The formula for the two regression equations remain the same as cited above under the method of deviation from the actual Means except that the two regression coefficients are determined by the following methods :

$$b_{xy} = \frac{N\Sigma d_x d_y - \Sigma d_x . \Sigma d_y}{N\Sigma d_y^2 - (\Sigma d_y)^2}$$

$$b_{yx} = \frac{N\Sigma d_x d_y - \Sigma d_x . \Sigma d_y}{N\Sigma d_x^2 - (\Sigma d_x)^2}$$

In the above formulae,

N = number of pairs of the observations,

d_x = deviation of items of X series from its assumed Mean,

d_y = deviation of items of Y series from its assumed Mean, and all other factors carry the same meanings as explained before.

Notes. 1. While taking deviations, if each of them has been divided by a common factor to reduce its magnitude, then each of the above formula will be modified slightly as under :

The formula of b_{xy} will be multiplied by $\frac{i_x}{i_y}$, and the formula of b_{yx} will be multiplied by $\frac{i_y}{i_x}$.

where i_x = common factor of X deviations

And i_y = common factor of Y deviations.

Regression of Frequency Distribution

1. In group data, a two way frequency table is formed and the formula stated above is slightly modified as under :

$b_{xy} = \frac{N\Sigma Fd_x d_y - \Sigma Fd_x . \Sigma Fd_y}{N\Sigma Fd_y^2 - (\Sigma Fd_y)^2}$	$b_{yx} = \frac{N\Sigma Fd_x d_y - \Sigma Fd_x . \Sigma Fd_y}{N\Sigma Fd_x^2 - (\Sigma Fd_x)^2}$

2. In case of grouped data given in two-way frequency tables with step deviations, the above formulae of the regression coefficients will be slightly modified as under :

$b_{xy} = \frac{N\Sigma Fd'_x d'_y - \Sigma Fd'_x . \Sigma Fd'_y}{N\Sigma Fd'^2_y - (\Sigma Fd'_y)^2} \times \frac{i_x}{i_y}$	$b_{yx} = \frac{N\Sigma Fd'_x d'_y - \Sigma Fd'_x . \Sigma d'_y}{N\Sigma Fd'^2_x - (\Sigma Fd'_x)^2} \times \frac{i_y}{i_x}$

In the above formula, the various factors involved carry the same meanings as explained in the previous formulae except that each factor is multiplied by its respective frequency.

The above formulae are multiplied respectively by $\frac{i_x}{i_y}$, *and* $\frac{i_y}{i_x}$ *as shown above only when step deviations are taken by dividing each deviation figure by the class interval of the respective series.*

9.5. STANDARD ERROR OF ESTIMATES

As stated before time and again, the regression lines or equations relating to the two variables are nothing but the lines or equations of estimates. With these equations or lines, we estimate the best probable value of one variable say X, on the basis of some given value of the other variable say, Y. But it must not be taken for sure that the values of a variable, which we obtain by such estimating lines or equations, are perfectly correct. Estimation is after all a matter of estimation and never exact. There must be some difference between the exact values, and the values we estimate with the regression equations. This difference is called error. Thus, while predicting the values of a variable with such equations it will be wise on our part to state always the probable amount of the error in these estimates. This probable amount of error expected to be in the estimates is called standard error of estimates.

Since, there are two estimating lines, or equations i.e. of X on Y, and of Y on X, we can calculate the standard errors for both these lines of estimates. This is calculated just in the manner of standard deviation. As the standard deviation measures the scatter, or dispersion of the items about their Mean, the standard error of estimates measures the deviations of the observed values of a variable from their estimated values.

Formulae

There are different types of formulae for computing the standard error of estimates which may be noted as under :

1. Fundamental formula. **The fundamental, or conceptual formulae for obtaining the standard** errors of estimates are as under :

$$(i)\ SE_{Y \text{ on } X} = \sqrt{\frac{\Sigma(Y - Y_e)^2}{N}}$$

$$(ii)\ SE_{X \text{ on } Y} = \sqrt{\frac{\Sigma(X - X_e)^2}{N}}$$

Where $SE_{Y \text{ on } X}$ = standard error of the estimates of Y on X

$SE_{X \text{ on } Y}$ = standard error of the estimates of X on Y

Y = observed value of the Y variable

X = observed value of the X variable

Y_e = estimated value of the Y variable.

X_e = estimated value of the X variable,

N = number of pairs of observations

Note. (*a*) If the size of the sample is small, the denominator in the above formula should be N – 2 in place of N as required by the principle of the degree of freedom.

(*b*) The above formulae may also be stated respectively as under :

$$(i)\ SE_{Y \text{ on } X} = \sqrt{\frac{\text{Unexplained variation in Y}}{N}}$$

$$(ii)\ SE_{X \text{ on } Y} = \sqrt{\frac{\text{Unexplained variation in X}}{N}}$$

2. Value based formulae. The above fundamental formulae involve some difficulties as they need computation of all the estimated values of the variables. To do away with such difficulties, the following value based formulae may be used advantageously.

$$(i)\ SE_{Y \text{ on } X} = \sqrt{\frac{\Sigma Y^2 - a\Sigma Y - b\Sigma XY}{N}}$$

$$(ii)\ SE_{X \text{ on } Y} = \sqrt{\frac{\Sigma X^2 - a\Sigma Y - b\Sigma XY}{N}}$$

In the above formulae,

X and Y represent respectively the given values of X and Y variables,

a and b represent the two constants, the values of which are got with reference to the first and second term respectively of the respective regression equations, or by solving the two relevant normal equations, and all other factors carry the same meanings as stated earlier.

3. Correlation coefficient based formulae

$$(i)\ SE_{y \text{ on } x} = \sigma_y \sqrt{1 - r^2}$$

$$(ii)\ SE_{x \text{ on } y} = \sigma_x \sqrt{1 - r^2}$$

In the above formulae, the factors carry the same meaning as explained earlier.

Interpretation of the Standard Error of Estimates

The standard error of estimates explained above is interpreted in the same manner as a standard deviation is interpreted about the Mean.

It indicates the significance of the estimated values. If the value of the S.E. is more, it will signify that there is a greater dispersion of the observed values from the estimated ones. On the other hand, if the value of the standard error is less, it will imply that the dispresion of the observed values from the estimated ones is less. Besides, it lays down the range within which a certain percentage of the total items of a sample will lie when there is a normal dispersion about the line of relationship.

The following are some of the ranges laid down by the standard error of estimates :

Standard error	Range of the items included
± 0.6745	50%
± 1	68.27%
± 2	95.45%
± 3	99.73%

It may be noted that a better measure of correlation can also be obtained by using the standard error of estimates as follows :

$$r = \sqrt{1 - \frac{SE^2_{y \text{ on } x}}{\sigma_y^2}}, \text{ or } = \sqrt{1 - \frac{SE^2_{x \text{ on } y}}{\sigma_x^2}}$$

The following illustrations will show how the standard error of estimates, and its related factors are calculated.

ILLUSTRATION 12. *From the following data, find the two regression equations, and calculate the standard error of estimates under different methods :*

X :	6	2	10	4	8
Y :	9	11	5	8	7

Also, determine the value of r on the basis of the standard error.

SOLUTION

Determination of the regression equations, and the standard error

X	Y	X^2	Y^2	XY	
6	9	36	81	54	
2	11	4	121	22	
10	5	100	25	50	
4	8	16	64	32	
8	7	64	49	56	
$\Sigma X = 30$	$\Sigma Y = 40$	$\Sigma X^2 = 220$	$\Sigma Y^2 = 340$	$\Sigma XY = 214$	$N = 5$

By the value based method

We have,
$$b_{xy} = \frac{N\Sigma XY - \Sigma X.\Sigma Y}{N\Sigma Y^2 - (\Sigma Y)^2} = \frac{(5\times 214) - (30\times 40)}{(5\times 340) - (40)^2}$$
$$= \frac{1070 - 1200}{1700 - 1600} = -\frac{-130}{100} = -1.3$$
$$b_{yx} = \frac{N\Sigma XY - \Sigma X.\Sigma Y}{N\Sigma X^2 - (\Sigma X)^2} = \frac{(5\times 214) - (30\times 40)}{(5\times 220) - (30)^2}$$
$$= -\frac{-130}{1100 - 900} = \frac{-130}{200} = -0.65$$
$$\bar{X} = \frac{\Sigma X}{N} = \frac{30}{5} = 6, \text{ and } \bar{Y} = \frac{\Sigma Y}{N} = \frac{40}{5} = 8$$

(*i*) Regression equation of X on Y

This is given by
$$X = \bar{X} + b_{xy}\,(Y - \bar{Y})$$
$$= 6 + -1.3\,(Y - 8) = 6 + 10.4 - 1.3Y$$
$$\therefore \quad X = 16.4 - 1.3Y$$

(*ii*) Regression equation of Y on X

This is given by
$$Y = \bar{Y} + b_{yx}\,(X - \bar{X})$$
$$= 8 + -0.65\,(X - 6)$$
$$= 8 + 3.90 - 0.65X$$
$$\therefore \quad Y = 11.9 - 0.65X.$$

(*iii*) Standard errors of estimate

(1) *Under the fundamental method*

(*a*) *Computation of the estimated values of X on Y*

When

$Y = 9, X = 16.4 - 1.3(9) = 16.4 - 11.7 = 4.7$

$Y = 11, X = 16.4 - 1.3(11) = 16.4 - 14.3 = 2.1$

$Y = 5, X = 16.4 - 1.3(5) = 16.4 - 6.5 = 9.9$

$Y = 8, X = 16.4 - 1.3(8) = 16.4 - 10.4 = 6$

$Y = 7, X = 16.4 - 1.3\,(7) = 16.4 - 9.1 = 7.3$

(*b*) *Computation of the estimated values of Y on X*

When

$X = 6, Y = 11.9 - 0.65(6) = 11.9 - 3.9 = 8$

$X = 2, Y = 11.9 - 0.65\,(2) = 11.9 - 1.3 = 10.6$

$X = 10, Y = 11.9. - 0.65\,(10) = 11.9 - 6.5 = 5.4$

$X = 4, Y = 11.9 - 0.65\,(4) = 11.9 - 2.6 = 9.3$

$X = 8, Y = 11.9 - 0.65\,(8) = 11.9 - 5.2 = 6.7$

Working Table

	X	Y	Xe	Ye	$(X-\bar{X})$	$(Y-\bar{Y})$	
	6	9	4.7	8.0	1.69	1.00	
	2	11	2.1	10.6	0.01	.16	
	10	5	9.9	5.4	0.01	.16	
	4	8	6.0	9.3	4.00	1.69	
	8	7	7.3	6.7	0.49	.09	
Total	–	–	–	–	6.20	3.10	N = 5

(*i*) Standard error of estimates of Y on X

This is given by

$$SE_{Y \text{ on } X} = \sqrt{\frac{\Sigma(Y-Y_e)^2}{N}} = \sqrt{\frac{3.10}{5}} = \sqrt{.62} = 0.787$$

(*ii*) Standard error of estimate of X on Y

This is given by

$$SE_{X \text{ on } Y} = \sqrt{\frac{\Sigma(X-X_e)^2}{N}} = \sqrt{\frac{6.20}{5}} = \sqrt{1.24} = 1.114$$

(2) *Under the correlation coefficient based method*

(*a*) $$SE_{y \text{ on } x} = \sigma_y \sqrt{1-r^2}$$

Where

$$\sigma_y = \sqrt{\frac{\Sigma Y^2}{N} - \left(\frac{\Sigma Y}{N}\right)^2} = \sqrt{\frac{340}{5} - \left(\frac{4.0}{5}\right)^2}$$

$$= \sqrt{68-64} = \sqrt{4} = 2$$

And $$r = \sqrt{b_{xy}.b_{yx}} = \pm\sqrt{(-1.3)(-0.65)} = \pm\sqrt{0.845} = +0.92$$

$\therefore$ $$SE_{y \text{ on } x} = 2\sqrt{1-(-0.92)^2}$$

$$= 2\sqrt{1-(0.845)} = 2\sqrt{0.155} = 2 \times 0.39 = 0.78 \text{ approx}$$

(*b*) $$SE_{X \text{ on } Y} = \sigma_x\sqrt{1-r^2}$$

where, $$\sigma_x = \sqrt{\frac{\Sigma X^2}{N} - \left(\frac{\Sigma X}{N}\right)^2} = \sqrt{\frac{220}{5} - \left(\frac{30}{5}\right)^2} = \sqrt{44-36} = \sqrt{8} = 2.82 \text{ approx.}$$

$\therefore$ $$SE_{X \text{ on } Y} = 2.82\sqrt{1-(-0.92)^2}$$

$$= 2.82 \times 0.39 = 1.10 \text{ approx.}$$

Note. The slight difference that appears between the results under the above two methods is due to aproximation.

(3) *Under the value based method*

(*i*) $$SE_{Y \text{ on } X} = \sqrt{\frac{\Sigma Y^2 - a\Sigma Y - b\Sigma XY}{N}}$$

Where, with reference to the regression equation of

Y on X, $a = 11.9$ and $b = -0.65$

Thus, putting the respective values in the above formula we have,

$$SE_{Y \text{ on } X} = \sqrt{\frac{340-11.9(40)-(-0.65)(214)}{5}}$$

$$= \sqrt{\frac{340-476-139.10}{5}} = \sqrt{\frac{340-336.90}{5}} = \sqrt{\frac{3.10}{5}}$$

$$= \sqrt{0.62} = 0.787$$

(*ii*) $$SE_{X \text{ on } Y} = \sqrt{\frac{\Sigma X^2 - a\Sigma X - b\Sigma XY}{N}}$$

Where, with reference to the regression equation of X on Y, $a = 16.4$ and $b = -1.3$.

Thus, putting the respective values in the above formula we have,

$$SE_{X \text{ on } Y} = \sqrt{\frac{220-16.4(30)-(-1.3)(214)}{5}}$$

$$= \sqrt{\frac{220-492+278.2}{5}} = \sqrt{\frac{220-213.8}{5}}$$

$$= \sqrt{\frac{6.2}{5}} = \sqrt{1.24} = 1.114$$

(4) *Determination of r on the basis of the Standard Errors*

By the formula we have,

$$= \sqrt{1 - \frac{0.62}{4}} = \sqrt{1 - 0.155}$$

$$= \sqrt{0.845} = \pm 0.92$$

Alternatively

$$r = \sqrt{1 - \frac{SE^2_{x \text{ on } y}}{\sigma_x^2}} = \sqrt{\frac{1 - (1.114)^2}{(2.82)^2}}$$

$$= \sqrt{1 - \frac{1.24}{8}} = \sqrt{1 - 0.155}$$

$$= \sqrt{0.845} = \pm 0.92$$

Note. Under this method, the nature of the correlation coefficient is not clear. However, with reference to the earlier calculation on the basis of the regression coefficients, its nature is negative.

Hence, $r = -0.92$

9.6. EXPLAINED AND UNEXPLAINED VARIATION

The total variation of a variable is the sum of the squares of deviations of its values from its arithmetic average. Symbolically, it is represented by

$$\boxed{\Sigma x^2 \text{ } i.e., \text{ } \Sigma(X - \overline{X})^2}$$

Where, X = Value of the variable,

And $\overline{X}, \overline{Y}$ = arithmetic average of the series, X and Y respectively

Thus, for X variable, total of variation $= \Sigma(X - \overline{X})^2$ and for the Y variable, total of variation $= \Sigma (Y - \overline{Y})^2$

This total of variation of a variable consists of two types of variations *viz* :

(*i*) Explained variation and (*ii*) Unexplained variation.

The explained variation of a variable say,

$$X = \Sigma (X_e - \overline{X})^2, \text{ and}$$

The unexplained variation of the variable say, $\overline{X} = \Sigma (X - X_e)^2$

Thus, the total variation = Unexplained variation + Explained variation

For the variable X, it can be symbolically represented thus,

$$\Sigma (X - \overline{X})^2 = \Sigma (X - X_e)^2 + \Sigma (X_e - \overline{X})^2$$

Similarly, for the variable Y, the above relationship can be represented as,

$$\Sigma (Y - \overline{Y})^2 = \Sigma (Y - Y_e)^2 + \Sigma (Y_e - \overline{Y})^2$$