Unit

3 Correlation and Regression Analysis

3.1 Introduction

Descriptive statistics such as central tendency. Dispersion, skewness and kurtosis studied so far were concerned with the distributions of data associated with a single variable, however in some practical problem there are two or more than two variables found which tend to change together. Even in our practical life, we come, across certain conditions, where changes in one variable are accompanied by changes in other variables. For example, the expenditure is very much related to the income of the concerned family. An increase in income is expected to cause an increase in expenditure. To ascertain the such association between two variables, the statistical analysis is required. Correlation analysis is a statistical tool used to measure the degree of association (relationship) between two or more variables.

Correlation analysis is a statistical tool which studies the association or relationship between two or more variables. Two variables said to have correlation, if the change (increase or decrease) in one variable is accompanied by the change (increase or decrease/decrease or increase) in the other. For example, demand and supply of commodity, Height and Weight of children, quantity of fertilizer used and production of crops etc. The measure of correlation is 'correlation co-efficient' generally denoted by 'r' refers to the degree and direction of association but it does not give cause and effect of association or relationship and number of changes. It only enables us to have an idea about the degree and direction of the relationship between the variables under the study.

Some important definitions are given below:

"If two or more quantities vary so that movements in the one tend to be accompanied by corresponding movements in the other (s) then they are said to be correlated." ***-L.R. Cannor***

“Correlation is an analysis of co-variation between two or more variables” ***- A.M. Tuttle***

"When the relationship is of a quantitative nature, the appropriate statistical tool for discovering and measuring the relationship and expressing it in brief formula is known as correlation." ***- Croxton and Cowden***

“The measure of relationship between two or more variables is termed as correlation”***-Sir Francis Galton.***

The importance of studying correlation are as follows:

a. Related to quantity of fertilizer used, types of soil, quality of seeds, amount of rainfall and so on. Correlation helps in quantifying precisely the degree and direction of such relationships.

b. Correlation analysis contributes to the understanding of economic behaviour, aids in locating the critically important variables on which others depend, may reveal to the economist the connections by which disturbances spread and suggest to him the paths through which stabilizing forces may become effective. (W.A. Neiswanger)

c. In theory of economics and business studies, we come across several types of variables which show some kind of relationship. For example, there exists a relationship between price, supply and quantity demanded; advertising expenditure and sales promotion etc.

d. The concepts of regression and ratio of variation are based on measure of correlation.

e. Measures of correlation give us the more reliable prediction of variables. According to Tippett, "the effect of correlation is to reduce the range of uncertainty of our prediction".

f. In most researches in social sciences, correlation analysis helps in arriving at very important conclusions.

Note:- **Correlation** is the degree measure of relationship between any two variables.

Types of Correlation

Following are the types of correlation.

1. **Positive Correlation:** Positive Correlation indicates that value of variables deviated in same direction. Two variables said to have positive correlation, if increase (or decrease) in the value of one variable results increase (or decrease) in the value of other variable. For example, family income and expenditure in luxury items, Height and Weight etc.

i)	Increasing ↑	*X*:	17	20	25	30	34
	Increasing ↑	*Y*:	8	12	15	18	22
ii)	Decreasing ↓	*X*:	60	51	40	35	30
	Decreasing ↓	*Y*:	18	17	10	7	5

2. **Negative Correlation:** Negative Correlation indicates that value of variables deviated in opposite direction. Two variables said to have negative correlation, if increase (or decrease) in the value of one variable results decrease (or increase) in the value of other variable. it is also known as inverse correlation. For example, price and demand of commodity, temperature and sale of woolen garments etc.

i)	Increasing ↑	*X*:	17	20	25	30	34
	Decreasing ↓	*Y*:	22	18	15	12	8
ii)	Decreasing ↓	*X*:	60	51	40	35	30
	Increasing ↑	*Y*:	5	7	10	17	18

3. **Linear Correlation:** Two variables said to have linear correlation if unit change in the value of one variable results constant change in the value of other variable over the entire range of values. For example, the correlation between 'the number of students admitted' and the 'monthly fee collected' is linear in nature.

X:	1	2	3	4	5
Y:	5	7	9	11	13

4. **Non-Linear Correlation:** Two variables said to have non-linear correlation, if unit change in one variable does not result the change at a constant rate but at fluctuating rate of other variable. It is also called curvilinear correlation. Such correlation is common in the data relating to economics and social sciences. This is however, beyond the scope of this textbook. For example, Profit and expenditure in advertisement.

X:	1	2	3	4	5
Y:	7	8	15	20	23

5. **Simple, Partial and Multiple Correlation**

Simple correlation: The degree of relationship between only two variables is called simple correlation. e.g. (i) A study on the yield of crop with respect to only amount of fertilizer, (ii) sales revenue with respect to amount of money spent on advertisement.

Partial Correlation: It is also called net correlation. For some situation, we have to study more than two correlated variable. When we study correlation between only two variables among the variables taking remaining constant, it is called partial correlation. For example, correlation between deposit and income level keeping interest rate constant.

Multiple Correlations: The simultaneous study of correlation among more variables is called multiple correlation. For example, correlation among land, labour, capital and production of crop.

Methods of Studying Simple Correlation

The commonly used methods for studying correlation (linear relationship) between two variables are:

1. Scatter Diagram Method or Graphic method.
2. **Karl Pearson's Coefficient of Correlation (Covariance method)**
3. Bivariate Correlation Method (Two way frequency table)
4. Spearman's Rank Correlation Method

1. Scatter Diagram (Graphic method)

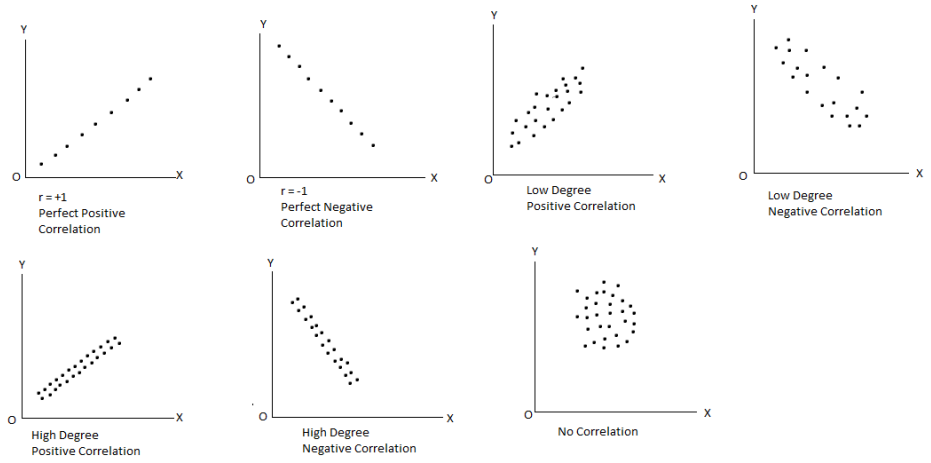
Scatter diagram is one of the simplest diagrammatic representations of bivariate distribution consisting of dots or points. It is a graphical method of studying the correlation between two variables. In this method, the points are represented by dots by keeping values of one variable on X-axis and the values of other variable on Y-axis in the XY-plane. Then, the diagram of dots so obtained is called scatter diagram. Scatter diagram is used to observe the existence of correlation between two variables. The following are the different types of pattern in scatter diagram of a bivariate data.

a. The pattern of points on a scatter diagram reveals on upward trend rising from bottom left to top right. The variables in this case shows positive correlation between them.

b. The pattern of points on a scatter diagram reveals a downward trend falling from top left to bottom right. The variables in this case show the negative correlation between them.

c. When we can not trace any trend, there is no correlation between them.

The following are some scatter diagrams depicting different types of correlation between two variables.



Observations on Scatter Diagrams

a. If the points on the scatter diagram show a very little spread, then the fair good amount of correlation can be expected between the two variables and if the points on scatter diagram show a widely spread, then the poor correlation may be expected.

b. The scatter diagram provides a rough measure of correlation only. It gives the direction and degree of correlation between the two variables.
c. It enables us to locate the line of best fit.

Merits:

a. It is the simplest method of measuring correlation.
b. It is least affected by the extreme observations.
c. It can be easily understood by non –statistician.
d. If helps us to detect abnormal variates in the data.

Demerits:

a. It gives only rough idea.
b. It can not be numerically expressed.
c. It is not amenable to algebraic treatment.

2. Karl Pearson's Coefficient of Correlation (Covariance Method)

Karl Pearson's developed a widely used mathematical formula to measure the degree (intensity) of linear relationship between two variables is called Karl Pearson's correlation coefficient. It is also called **product moment correlation** coefficient or simple correlation coefficient. This method is based on the linear relationship between two variables (Series). Karl Pearson's correlation coefficient is especially useful when data are quantitatively measured.

According to Pearson, correlation coefficient between two variables X and Y is denoted by $r(X, Y)$ or r_{XY} or simply r and is defined as the ratio of the covariance between them to the product of their corresponding standard deviations.

$$\therefore \text{ Coefficient of correlation, } r = \frac{\text{Cov(X,Y)}}{\sqrt{Var(X)}\sqrt{Var(Y)}} = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y} \quad \text{... (i)}$$

where, Cov (X, Y) = Covariance between two variables X & Y

$$= \frac{1}{n}\Sigma(X - \bar{X})(Y - \bar{Y})$$

Covariance measures the simultaneous changes between two variables.

$$V(X) = \text{Variance of } X = \sigma_X^2 = \frac{1}{n}\Sigma(X - \bar{X})^2$$

Standard deviation of $X = \sigma_X = \sqrt{\frac{1}{n}\Sigma(X - \bar{X})^2}$

$$V(Y) = \text{Variance of } Y = \frac{1}{n}\Sigma(Y - \bar{Y})^2$$

Standard deviation of $Y = \sigma_Y = \sqrt{\frac{1}{n}\Sigma(Y - \bar{Y})^2}$

Substituting Cov(X, Y), σ_X and σ_Y in (i) we have,

$$r = \frac{\frac{1}{n}\Sigma(X - \bar{X})(Y - \bar{Y})}{\sqrt{\frac{1}{n}\Sigma(X - \bar{X})^2}\sqrt{\frac{1}{n}\Sigma(Y - \bar{Y})^2}}$$

or,
$$r = \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{\sqrt{\Sigma(X - \bar{X})^2}\sqrt{\Sigma(Y - \bar{Y})^2}} \quad \text{... (ii)}$$

It is also called Product Moment Formula.

On simplification, we get

$$r = \frac{n\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{n\Sigma X^2 - (\Sigma X)^2}\sqrt{n\Sigma Y^2 - (\Sigma Y)^2}} \quad \text{... (iii)}$$

This method is also known as direct method.

Different formulae for calculating Karl Pearson's Coefficient of Correlation

1. We have,

$$r = \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{\sqrt{\Sigma(X - \bar{X})^2}\sqrt{\Sigma(Y - \bar{Y})^2}} = \frac{\Sigma xy}{\sqrt{\Sigma x^2}\sqrt{\Sigma y^2}} \quad \text{... (iv)}$$

where, x and y denote the deviations of X and Y from their arithmetic means $\bar{X}$ and $\bar{Y}$ respectively, i.e., $x = X - \bar{X}$, $y = Y - \bar{Y}$. This method is known as Actual mean method.

2. Product moment formula (Direct method)

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

3. When actual means of X and Y are in fraction, the calculation of Pearson's correlation coefficient can be simplified by taking deviations of X and Y values from their assumed means A and B respectively. That is $U = X - A$ and $V = Y - B$, when A and B are assumed means of X and Y series. The formula (iii) becomes as given below and known as short cut method.

   Assumed mean method (i.e. Change of origin)

$$r = \frac{n\Sigma UV - (\Sigma U)(\Sigma V)}{\sqrt{n\Sigma U^2 - (\Sigma U)^2}\sqrt{n\Sigma V^2 - (\Sigma (V)^2)}} \quad \text{... (v)}$$

4. Step deviation method (i.e. Change of origin &scale)

$$r = \frac{n\Sigma U'V' - (\Sigma U')(\Sigma V')}{\sqrt{n\Sigma U'^2 - (\Sigma U')^2}\sqrt{n\Sigma V'^2 - (\Sigma V')^2}} \quad \text{... (vi)}$$

where, $U' = \frac{X - A}{h}$, $V' = \frac{Y - B}{k}$

h = Common factor for variable X, A = Assumed mean of X-series,

n = Number of paired observations,

k = Common factor for variable Y, B = Assumed mean of Y-series.

Note: Karl Pearson's correlation coefficient is also called product moment formula because

$$\text{Cov}(X, Y) = E[\{X - E(X)\}\{Y - E(Y)\}]$$

Interpretation of Correlation Coefficient

The degrees of correlation coefficient according to Karl Pearson's formula's are as follows:

Degree of Correlation	Direction	
	Positive	**Negative**
Perfect correlation	+1	–1
Very high degree of correlation	+ 0.9 or more	– 0.9 or more
High degree of correlation	+ 0.75 to + 0.9	– 0.75 to – 0.9
Moderate degree of correlation	+ 0.50 to + 0.75	– 0.50 to – 0.75
Low degree of correlation	+ 0.25 to + 0.50	– 0.25 to – 0.50
Very low degree of correlation	less than + 0.25	less than – 0.25
No correlation	0 (zero)	

Properties of Karl Pearson's Correlation Coefficient

1. Pearson's correlation coefficient (r) cannot exceed 1 numerically. In other words, it lies between -1 and +1. Symbolically, $\mathbf{-1 \le r \le +1}$
2. Correlation coefficient is independent of change of origin and scale.
   Mathematically, if x and y are given variables and they are transferred into new variables u and v by change of origin and scale viz. $u = \frac{x - A}{h}$ & $v = \frac{y - B}{k}$.

   Where A, B, h & k are constants, $h > 0$, $k > 0$; then the correlation coefficient between x & y is same as correlation coefficient between u & v, r(x, y) = r(u, v) i.e. $r_{xy} = r_{uv}$
3. Correlation coefficient is the geometric mean between two regression coefficients i.e. $r = \pm \sqrt{b_{yx} \cdot b_{xy}}$

   where, b_{yx} = Regression coefficient of regression line of y on x

   b_{xy} = Regression coefficient of regression line of x on y

   Note: Sign of correlation coefficient 'r' is determined according to the sign of regression coefficients. If both regression coefficients are positive then 'r' is also positive and if both regression coefficients are negative then 'r' is also negative.
4. Correlation coefficient is a relative statistical measure, so it is a pure number independent of unit of measurement.
5. Two independent variables are uncorrelated but converse may not be true i.e. two uncorrelated variables need not be independent.

Some Worked Out Examples

Example 3.1 Find the Karl Pearson's Co-efficient of correlation from the following data

X	5	7	1	3	4
Y	2	2	4	5	6

Solution: Calculation of Correlation Co-efficient (Product moment method)

X	Y	$x = X - \bar{X}$	$y = Y - \bar{Y}$	xy	x^2	y^2
5	2	1	–1.8	–1.8	1	3.24
7	2	3	–1.8	–5.4	9	3.24
1	4	–3	0.2	–0.6	9	0.04
3	5	–1	1.2	–1.2	1	1.44
4	6	0	2.2	0	0	4.84
$\Sigma x = 20$	$\Sigma Y = 19$	$\Sigma x = 0$	$\Sigma y = 0$	$\Sigma xy = -9$	$\Sigma xy^2 = 20$	$\Sigma y^2 = 12.8$

We have $\bar{X} = \frac{\Sigma x}{n} = \frac{20}{5} = 4$ and $\bar{Y} = \frac{\Sigma y}{n} = \frac{19}{5} = 3.8$

Now, Karl Pearson's Co-efficient of Correlation is given by

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2}\sqrt{\Sigma y^2}} = \frac{-9}{\sqrt{20} \,.\, \sqrt{12.8}} = -0.5624 \text{ [Using actual mean]}$$

Alternatively, (Direct method/using product moment formula)

X	Y	XY	X^2	Y^2
5	2	10	25	5
7	2	14	49	4
1	4	4	1	16
3	5	15	9	25
4	6	24	16	36
$\Sigma X = 20$	$\Sigma Y = 19$	$\Sigma XY = 67$	$\Sigma X^2 = 100$	$\Sigma Y^2 = 86$

Now,

Karl Pearson's co-efficient of correlation is given by

$$r = \frac{n\,\Sigma XY - \Sigma X \cdot \Sigma Y}{\sqrt{n\,\Sigma X^2 - (\Sigma X)^2} \cdot \sqrt{n\,\Sigma Y^2 - (\Sigma Y)^2}} \quad \text{[Product moment formula]}$$

$$= \frac{5 \times 67 - 20 \times 19}{\sqrt{5 \times 100 - 20^2}\sqrt{5 \times 86 - 19^2}} = -0.5417$$

$\therefore$ There is moderate negative correlation between X and Y.

Merits and Demerits of Karl Pearson's Coefficient

Merits of correlation Co-efficient are

a. It is based on all the observations.
b. It indicates magnitude as well as direction of linear correlation between the variables.
c. There is no chance of personal bias in its computation.
d. It is the best method of computing simple linear correlation.

Demerits of correlation co-efficient are

a. It is applicable only when the correlation between the variables is linear.
b. It is the affected by the extreme values.
c. Its interpretation is not an easy attempt.

Probable Error (P.E)

Probable error of the correlation co-efficient is the measure of testing the reliability of the calculated value of r. It is generally denoted by P.E. (r). If r be the calculated value from a sample on 'n' pair of observations. Then P.E. (r) is given by

$$\text{P.E. } (r) = 0.6745 \times \frac{1 - r^2}{\sqrt{n}}$$

The probable error of r may be used to determine the limits within which the population correlation co-efficient lies. **Limits (range)** for population correlation co-efficient are $r \pm \text{P.E. } (r)$.

Another use of P.E. (r) is to test whether value of sample correlation co-efficient is **significant** for any correlation in the population, for these following results arises:

i) If $|r| < P.E.$ (r), then r is not significant (insignificant) at all.

ii) If $|r| > 6\ P.E.$ (r), then r is definitely significant.

iii) In other situations, nothing can be concluded with certainty.

Example 3.2 If correlation co-efficient of 10 pair of observations is 0.4. Test whether the value of r is significant or not. Also compute the limits within which the population correlation co-efficient may be expected to lie.

Solution: Here,

$n = 10$ and $r = 0.4$

We have, $\text{P.E. } (r) = 0.6745 \times \frac{1 - r^2}{\sqrt{n}} = 0.6745 \times \frac{1 - (0.4)^2}{\sqrt{10}} = 0.6745 \times \frac{1 - 0.16}{\sqrt{10}} = 0.18$

and $6 \times \text{P.E. } (r) = 6 \times 0.18 = 1.08$

We see that neither $|r| < \text{P.E. } (r)$ nor $|r| > 6 \times \text{P.E. } (r)$. Thus, no conclusion can be drawn with certainty.

Now, Limits for population correlation co-efficient are

$r \pm \text{P.E. } (r) = 0.4 \pm 0.18$ i.e., 0.58 and 0.22

Example 3.3 If sample size n is 50, variance of X is 9, S.D. of Y is 4, then covariance is 9.8, find Karl Pearson's correlation coefficient.

Solution: We have,

$n = 50$, variance of $X = \sigma^2_x = 9$, S.D. of $X = \sigma_x = 3$, S.D. $(Y) = \sigma_y = 4$

Cov $(X, Y) = 9.8$

Correlation coefficient (r) = ?

Now, $r = \frac{\text{Cov}(X, Y)}{\sigma_X \cdot \sigma_Y} = \frac{9.8}{3 \times 4} = 0.82.$

There is high degree of positive correlation between two variables X & Y.

Example 3.4 Compute the coefficient of correlation from the following results obtained between two variables.

	Variable *X*	**Variable *Y***
Number of sets	7	7
Arithmetic mean	4	8
Sum of squares of deviations from arithmetic mean	28	76

Summation of products of deviation of variables X and Y from this respective means is 46.

Solution: In the usual notations, we have given

$$n = 7,\ \bar{X} = 4,\ \bar{Y} = 8,$$

$$\Sigma(X - \bar{X})^2 = \Sigma x^2 = 28, \qquad \Sigma(Y - \bar{Y})^2 = \Sigma y^2 = 76$$

$$\Sigma(X - \bar{X})\ \Sigma(Y - \bar{Y}) = \Sigma xy = 46$$

Then, Karl Pearson's correlation coefficient between X-series and Y-series is given by:

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2} \cdot \sqrt{\Sigma y^2}} = \frac{46}{\sqrt{28} \cdot \sqrt{76}} = 0.997$$

There is very high degree of positive correlation between two variables X & Y.

Example 3.5 Find the Karl Pearson's correlation coefficient for the following data:

X	2	4	6	8	10
Y	6	8	9	10	12

Use: a) Direct method
b) Deviation from assumed mean method
c) Deviation from actual mean method

Solution:

a) Direct method

X	Y	XY	X^2	Y^2
2	6	12	4	36
4	8	32	16	64
6	9	54	36	81
8	10	80	64	100
10	12	120	100	144
$\Sigma X = 30$	$\Sigma Y = 45$	$\Sigma XY = 298$	$\Sigma X^2 = 220$	$\Sigma Y^2 = 425$

$$r = \frac{n\Sigma XY - \Sigma X \Sigma Y}{\sqrt{n\Sigma X^2 - (\Sigma X)^2}\sqrt{n\Sigma Y^2 - (\Sigma Y)^2}} = \frac{5(298) - (30)\ (45)}{\sqrt{5(220) - (30)^2}\sqrt{5(425) - (45)^2}} = 0.99.$$

There is very high degree of positive correlation between two variables X & Y.

b) Deviation from assumed mean method

Let, Assumed mean of X-series (A) = 4, Assumed mean of Y-series (B) = 10

Now,

X	Y	$U = X - 4$	$V = Y - 10$	UV	U^2	V^2
2	6	–2	–4	8	4	16
4	8	0	–2	0	0	4
6	9	2	–1	–2	4	1

8	10	4	0	0	16	0
10	12	6	2	12	36	4
		$\Sigma U = 10$	$\Sigma V = -5$	$\Sigma UV = 18$	$\Sigma U^2 = 60$	$\Sigma V^2 = 25$

Karl Pearson's co-efficient of correlation is given by,

$$r = \frac{n\Sigma UV - \Sigma U\, \Sigma V}{\sqrt{n\Sigma U^2 - (\Sigma U)^2}\sqrt{n\Sigma V^2 - (\Sigma V)^2}} = \frac{5(18) - (10)(-5)}{\sqrt{5(60) - (10)^2}\sqrt{5(25) - (-5)^2}} = 0.99$$

c) Deviation from actual mean method

Computation of Correlation Coefficient

X	Y	$x = X - 6$	$y = Y - 9$	x^2	y^2	xy
2	6	–4	–3	16	9	12
4	8	–2	–1	4	1	2
6	9	0	0	0	0	0
8	10	2	1	4	1	2
10	12	4	3	16	9	12
$\Sigma X = 30$	$\Sigma Y = 45$	$\Sigma x = 0$	$\Sigma y = 0$	$\Sigma x^2 = 40$	$\Sigma y^2 = 20$	$\Sigma xy = 28$

Here, $\bar{X} = \frac{\Sigma X}{n} = \frac{30}{5} = 6, \quad \bar{Y} = \frac{\Sigma Y}{n} = \frac{45}{5} = 9$

Now, $r = \frac{\Sigma xy}{\sqrt{\Sigma x^2} \cdot \sqrt{\Sigma y^2}} = \frac{28}{\sqrt{40} \cdot \sqrt{20}} = 0.99,$

There is very high degree of positive correlation between X & Y.

Example 3.6 If $r = 0.5$, $\Sigma xy = 120$, $\sigma_y = 8$ and $\Sigma x^2 = 90$, find the number of items, where x and y are deviation from their respective means**.**

Solution: We have given,

$$r = 0.5,\ \Sigma xy = 120,\ \sigma_y = 8,\ \Sigma x^2 = 90$$

$$\left[\because x = X - \bar{X} \text{ and } y = Y - \bar{Y},\ \sqrt{\frac{1}{n}\Sigma(Y - \bar{Y})^2} = \sigma_y\right]$$

Given, $\sqrt{\frac{1}{n}\Sigma(Y - \bar{Y})^2} = 8$

$\Rightarrow \sqrt{\frac{1}{n}\Sigma y^2} = 8$

$\Rightarrow \frac{1}{n} \cdot \Sigma y^2 = 64$

$\Rightarrow \Sigma y^2 = 64n$

Number of items (n) = ?

Then Karl Pearson's correlation coefficient between variables X and Y is given by

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

$\Rightarrow 0.5 = \frac{120}{\sqrt{90} \cdot \sqrt{64n}}$

$$\Rightarrow \qquad (0.5)^2 = \frac{(120)^2}{90 \times 64 \times n} \qquad [\square \text{ squaring on both sides}]$$

$$\Rightarrow \qquad n = \frac{(120)^2}{90 \times 64 \times (0.5)^2} = \frac{14400}{5760 \times 0.25} = 10.$$

Example 3.7 Compute the product moment correlation coefficient for the following:

Height of father	67	65	68	64	66
Height of Son	69	65	69	65	67

Solution: Let X be a variable of father's height and Y be the variable of son's height

Computation of Correlation Coefficient

Height of Father (in inches) (X)	Height of Son (in inches) (Y)	$x = X - \bar{X} = X - 66$	$y = Y - \bar{Y} = Y - 67$	x^2	y^2	xy
67	69	1	+ 2	1	4	2
65	65	–1	–2	1	4	2
68	69	2	2	4	4	4
64	65	–2	–2	4	4	4
66	67	0	0	0	0	0
$\Sigma X = 330$	$\Sigma Y = 335$	$\Sigma x = 0$	$\Sigma y = 0$	$\Sigma x^2 = 10$	$\Sigma y^2 = 16$	$\Sigma xy = 12$

Here, $\bar{X} = \frac{\Sigma X}{n} = \frac{330}{5} = 66, \qquad \bar{Y} = \frac{\Sigma Y}{n} = \frac{335}{5} = 67$

Then, Pearson's co-efficient of correlation is given by

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2} \cdot \sqrt{\Sigma y^2}} = \frac{12}{\sqrt{10} \cdot \sqrt{16}} = 0.95$$

There is very high degree of positive correlation between X & Y.

Example 3.8 Find the Karl Pearson's correlation coefficient between two variables X and Y from 12 pairs of observations,

$$\Sigma X = 30,\ \Sigma Y = 5,\ \Sigma X^2 = 670,\ \Sigma Y^2 = 285,\ \Sigma XY = 385$$

Solution: We have given,

Number of observation (n) = 12

$$\Sigma X = 30,\ \Sigma Y = 5,\ \Sigma X^2 = 670,\ \Sigma Y^2 = 285,\ \Sigma XY = 385$$

Karl Pearson's correlation coefficient between two variables X and Y is

$$r = \frac{n\Sigma XY - \Sigma X \cdot \Sigma Y}{\sqrt{n\Sigma X^2 - (\Sigma X)^2} \cdot \sqrt{n\Sigma Y^2 - (\Sigma Y)^2}} = \frac{12 \times 385 - 30 \times 5}{\sqrt{12 \times 670 - 30^2} \cdot \sqrt{12 \times 285 - 5^2}} = 0.91$$

There is very high degree of positive correlation between X & Y.

Example 3.9 Calculate the Karl Pearson's correlation coefficient for the following data of sales and expenses in thousands of rupees of 5 firms.

Sales	43	41	36	34	50
Expenses	12	24	15	21	19

Solution: Let X be the variable of sales and Y be the variable of expenses in '000' Rs.

Also let Assumed Mean of X-series (A) = 41;

Assumed Mean of Y-series (B) = 19

Computation of Correlation Coefficient

Sales (X)	$U = X - 41$	U^2	Expenses (Y)	$V = Y - 19$	V^2	$U \cdot V$
43	2	4	12	– 7	49	– 14
41	0	0	24	5	25	0
36	–5	25	15	–4	16	20
34	–7	49	21	2	4	–14
50	9	81	19	0	0	0
	$\Sigma U = -5$	$\Sigma U^2 = 159$		$\Sigma V = -4$	$\Sigma V^2 = 94$	$\Sigma UV = -8$

Therefore, The Karl Pearson's correlation coefficient between sales (X) and expenses (Y) is given by

$$r = \frac{n\Sigma U\,V - \Sigma U \cdot \Sigma V}{\sqrt{n\Sigma U^2 - (\Sigma U)^2} \cdot \sqrt{n\Sigma V^2 - (\Sigma V)^2}}$$

$$r = \frac{5 \times (-8) - (-5) \times (-4)}{\sqrt{5 \times 159 - (-5)^2} \,.\, \sqrt{5 \times 94 - (-4)^2}} = -0.10.$$

There is very low degree of negative correlation between sales and expenses.

Example 3.10 Compute the coefficient of correlation from the following data:

No. of items (X):	200	270	340	360	400	300
No. of defective items (Y):	150	160	165	180	195	155

Solution: Let Assumed mean of X-series (A) = 300 and h = 10;

Assumed mean of Y-series (B) = 165 and K = 5

U' and V' are obtained as $U' = \frac{X - A}{h}$ and $V' = \frac{Y - B}{k}$.

Computation of Correlation Coefficient

X	Y	$U' = \frac{X - 300}{10}$	$V' = \frac{Y - 165}{5}$	$(U')^2$	$(V')^2$	$U' \cdot V'$
200	150	–10	–3	100	9	30
270	160	–3	–1	9	1	3
340	165	4	0	16	0	0
360	180	6	3	36	9	18
400	195	10	6	100	36	60
300	155	0	–2	0	4	0
		$\Sigma U' = 7$	$\Sigma V' = 3$	$\Sigma U'^2 = 261$	$\Sigma V'^2 = 59$	$\Sigma U'V' = 111$

Karl Pearson's correlation coefficient between no. of items (X) and no. of defective items (Y) is given by

$$r = \frac{n\Sigma U'V' - (\Sigma U')\,(\Sigma V')}{\sqrt{n\Sigma U'^2 - (\Sigma U')^2}\,\sqrt{n\Sigma V'^2 - (\Sigma V')^2}}$$

$$= \frac{6 \times 11 - 7 \times 3}{\sqrt{6 \times 261 - 7^2} \,.\, \sqrt{6 \times 59 - 3^2}}$$

$$= \frac{666 - 21}{\sqrt{1566 - 49} \cdot \sqrt{354 - 9}}$$

$$= \frac{645}{\sqrt{1517} \cdot \sqrt{345}} = \frac{645}{723.44} = 0.89.$$

There is high degree of positive correlation between number of items and no. of defective items.

Example 3.11 Compute the coefficient of correlation between age and playing habits from the following data:

Age in years	Population	No. of players
15 and less than 20	150	120
20 and less than 25	200	156
25 and less than 30	400	228
30 and less than 35	300	150
35 and less than 40	250	100
40 and less than 45	100	30
45 and less than 50	80	20

Note: We should calculate the percentage of players out of the population in the respective age group which is defined as the playing habit.

Solution: Let us consider the percentage of no. of players as their playing habit.

Calculation of percentage no. of player.

Population	**No. of players**	**Percentage of Players**
150	120	$\frac{120}{150} \times 100 = 80$
200	156	$\frac{156}{200} \times 100 = 78$
400	228	$\frac{288}{400} \times 100 = 57$
300	150	$\frac{150}{300} \times 100 = 50$
250	100	$\frac{100}{250} \times 100 = 40$
100	30	$\frac{30}{100} \times 100 = 30$
80	20	$\frac{20}{80} \times 100 = 25$

Let X and Y be the mid values of age and percentage no. of players respectively. Also let assumed mean of X-series (A) = 32.5 and assumed mean of Y-series (B) = 50. Also h = 5 and k = 1, then U' and V' are obtained as $U' = \frac{X-A}{h}$ and $V' = \frac{Y-B}{k}$.

Computation of Correlation Coefficient

Age in years	Mid value (X)	Percentage of Players (Y)	$U' = \frac{X - 32.5}{5}$	$V' = \frac{Y - 50}{1}$	U'^2	V'^2	$U'V'$
15-20	17.8	80	–3	30	9	900	–90
20-25	22.5	78	–2	28	4	78	–56
25-30	27.5	57	–1	22	1	484	–22
30-35	32.5	50	0	0	0	0	0
35-40	37.5	40	1	-10	1	100	–10

40-45	42.5	30	2	-20	4	400	–40
45-50	47.5	25	3	30	9	900	90
			$\Sigma U' = 0$	$\Sigma V' = 20$	$\Sigma U'^2 = 28$	$\Sigma V'^2 = 3658$	$\Sigma U'.V' = -308$

Then, correlation coefficient between age and percentage of players is given by

$$r = \frac{n\Sigma U'V' - (\Sigma U')(\Sigma V')}{\sqrt{n\Sigma U'^2 - (\Sigma U')^2}\sqrt{n\Sigma V'^2 - (\Sigma V')^2}}$$

$$= \frac{7\times(-308) - 0\times 10}{\sqrt{7\times 28 - 0^2}\sqrt{7\times 3658 - 20^2}}$$

= –0.969, which is very near to –1.

There is very high degree of negative correlation between age and the playing habits.

Example 3.12 The following table gives the distribution of the total population and those who are wholly or partially blind among them. Find out if there is any relationship between age and blindness.

Age	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80
No. of students	100	60	40	36	24	11	6	3
No. of Blinds:	55	40	40	40	36	22	18	15

Solution:

Here we have to compute Karl Pearson's correlation coefficient between age (X) and blindness of the given total population in each age group (Y). Where blindness is obtained as percentage of blind students among the total number of students of the respective age group.

Calculation of percentage number of blind

Age	**No. of students**	**No. of Blinds**	**Percentage of blinds**
0-10	100	55	$\frac{55}{100} \times 100 = 55$
10-20	60	40	$\frac{40}{60} \times 100 = 67$
20-30	40	40	$\frac{40}{40} \times 100 = 100$
30-40	36	40	$\frac{40}{36} \times 100 = 111.11 = 111$
40-50	24	36	$\frac{36}{24} \times 100 = 150$
50-60	11	22	$\frac{22}{11} \times 100 = 200$
60-70	6	18	$\frac{18}{6} \times 100 = 300$
70-80	3	15	$\frac{15}{3} \times 100 = 500$

Computation of Correlation Coefficient between age and blindness

Age	Mid value (X)	Blindness (Y)	$U' = \frac{X-35}{10}$	$V = Y - 200$	U'^2	V'^2	$U'V'$
0-10	5	55	–3	–145	9	21025	435
10-20	15	67	–2	–133	4	17689	266
20-30	25	100	–1	–100	1	10000	100
30-40	35	111	0	–89	0	7921	0
40-50	45	150	1	–50	1	2500	-50
50-60	55	200	2	0	4	0	0
60-70	65	300	3	100	9	10000	300
70-80	75	500	4	300	16	90000	1200
			$\Sigma U' = 4$	$\Sigma V' = -117$	$\Sigma U'^2 = 44$	$\Sigma V'^2 = 159135$	$\Sigma U'V' = 2251$

Karl Pearson's co-efficient of correlation is given by

$$r = \frac{n\Sigma U'V' - \Sigma U' \cdot \Sigma V'}{\sqrt{n\Sigma U'^2 - (\Sigma U')^2} \cdot \sqrt{n\Sigma V'^2 - (\Sigma V')^2}}$$

$$= \frac{8 \times 2251 - 4 \times (-117)}{\sqrt{8 \times 44 - (4)^2}\sqrt{8 \times 15935 - (-117)^2}}$$

$$= \frac{18008 + 468}{\sqrt{336} \cdot \sqrt{1259391}} = \frac{18476}{20570.74} = 0.898.$$

Hence, there is very high positive correlation between age and blind.

Example 3.13 Compute the Karl Pearson's coefficient of correlation from the following data by the Karl Pearson's method

Price A	25	28	35	20	22	30	31	22
Price B	35	39	48	29	30	38	40	32

Also (a) Calculate its probable error

(b) Interpret if the value of r is significant or not

(c) Determine the limits within which the population correlation coefficient may be expected to lie.

Solution: Let X be the price of tea and Y be the price of coffee in Rupees.

Computation of Correlation Coefficient

X	Y	$U = X - 28$	$V = Y - 38$	U^2	V^2	UV
25	35	–3	–3	9	9	9
28	39	0	1	0	1	0
35	48	7	10	49	100	70
20	29	–8	–9	64	81	72
22	30	–6	–8	36	64	48
30	38	2	0	4	0	0
31	40	3	2	9	4	6
22	32	–6	– 6	36	36	36
		$\Sigma U = -11$	$\Sigma V = -13$	$\Sigma U^2 = 207$	$\Sigma V^2 = 295$	$\Sigma UV = 241$

Karl Pearson's correlation of coefficient is

$$r = \frac{n\sum UV - \sum U \cdot \sum V}{\sqrt{n\sum U^2 - (\sum U)^2} \cdot \sqrt{n\sum V^2 - (\sum V)^2}}$$
$$= \frac{8 \times 241 - (-11) \times (-13)}{\sqrt{8 \times 207 - (-11)^2} \cdot \sqrt{8 \times 295 - (-13)^2}}$$
$$= \frac{1928 - 143}{\sqrt{1535} \cdot \sqrt{2191}} = 0.9733$$

a) Probable error of correlation coefficient is given by

$$\text{P.E. } (r) = 0.6745 \times \frac{1 - r^2}{\sqrt{n}} = 0.6745 \times \frac{1 - (0.9733)^2}{\sqrt{8}} = 0.0125$$

b) To test significance of r

$$6 \times \text{P.E. } (r) = 6 \times 0.0125 = 0.0753$$

Since, r is much greater than 6 × P.E. (r), the value of r is highly significant.

c) Limit of population correlation coefficient

$$r \pm 6 \times \text{P.E. (r)}$$
$$= 0.9733 \pm 0.0753 = (0.9733 - 0.0753, 0.9733 + 0.0753)$$
$$= (0.8980, 1.048) = (0.8980, 1.048)$$

Example 3.14 A computer while calculating the correlation coefficient between two variants X and Y from 25 pairs of observation obtained the following information:

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

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

X	Y
6	14
8	6

While the correct values were

X	Y
8	12
6	8

Obtain correct value of the correlation coefficient between X and Y.

Solution: We have given,

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

Now,

$$\text{Corrected } \Sigma X = 125 - 6 - 8 + 8 + 6 = 125$$
$$\text{Corrected } \Sigma Y = 100 - 14 - 6 + 12 + 8 = 100$$
$$\text{Corrected } \Sigma X^2 = 650 - 6^2 - 8^2 + 8^2 + 6^2 = 650$$
$$\text{Corrected } \Sigma Y^2 = 460 - 14^2 - 6^2 + 12^2 + 8^2 = 436$$
$$\text{Corrected } \Sigma XY = 508 - 6 \times 14 - 8 \times 6 + 8 \times 12 + 6 \times 8 = 520$$
$$\text{Corrected } r = \frac{n\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{n\Sigma X^2 - (\Sigma X)^2}\sqrt{n\Sigma Y^2 - (\Sigma Y)^2}}$$
$$= \frac{25 \times 520 - 125 \times 100}{\sqrt{25 \times 650 - (125)^2} \cdot \sqrt{25 \times 436 - (100)^2}}$$
$$= \frac{500}{\sqrt{325} \cdot \sqrt{900}} = \frac{500}{750} = 0.67$$

There is moderate positive correlation between X & Y.

Example 3.15 From the following data, find out if there is any relationship between density of population and death rate:

District	Area in sq. km.	Population	No. of deaths
A	120	24000	288
B	150	75000	1125
C	80	48000	768
D	50	40000	720
E	200	50000	650

Solution: Here, we have to compute the coefficient of correlation between density of population and death rate. Therefore, we should first calculate density and death rates using the formula,

$$\text{Density} = \frac{\text{Population}}{\text{Area}} \text{ and Death rate} = \frac{\text{No. of deaths}}{\text{Population}} \times 1000$$

Calculation of Correlation between Density and Death Rate

District	Area in sq. km.	Population	No. of deaths	Density	Death rate
A	120	24000	288	$\frac{24000}{120} = 200$	$\frac{288}{24000} \times 1000 = 12$
B	150	75000	1125	$\frac{75000}{150} = 500$	$\frac{1125}{75000} \times 1000 = 15$
C	80	48000	768	$\frac{48000}{80} = 600$	$\frac{768}{48000} \times 1000 = 16$
D	50	40000	720	$\frac{40000}{50} = 800$	$\frac{720}{40000} \times 1000 = 18$
E	200	50000	650	$\frac{50000}{200} = 250$	$\frac{650}{50000} \times 1000 = 13$

Let X and Y be the density of population and the death rate respectively.

X	Y	$x = X - \bar{X}$	$y = Y - \bar{Y}$	x^2	y^2	xy
200	12	–270	–2.8	72900	7.84	756
500	15	30	0.2	900	0.04	6
600	16	130	1.2	16900	1.44	156
800	18	330	3.2	108900	10.24	1056
250	14	–220	–1.8	48400	3.24	396
$\Sigma x = 2350$	$\Sigma Y = 74$			$\Sigma x^2 = 248000$	$\Sigma y^2 = 22.8$	$\Sigma xy = 2370$

$$\bar{X} = \frac{\Sigma X}{n} = \frac{2350}{5} = 470 \text{ and } \bar{Y} = \frac{\Sigma Y}{n} = \frac{74}{5} = 14.8$$

Then,

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2} \cdot \sqrt{\Sigma y^2}} = \frac{2370}{\sqrt{248000}\sqrt{22.8}} = 0.996$$

There is high degree positive correlation between population density and death rate.

Example 3.16 The following are the monthly figures of advertising expenditure and sales of a firm. It is generally found that advertising expenditure has its impact on sales generally after 2 months. Allowing for this time-lag calculate coefficient of correlation.

Month	Advertising Expenditure (Rs. Lakhs)	Sales (Rs. Lakhs)
Jan	50	1200
Feb	60	1500
Mar	70	1600
Apr	90	2000
May	120	2200
Jun	150	2500
Jul	140	2400
Aug	160	2600
Sep	170	2800
Oct	190	2900
Nov	200	3100
Dec	150	3900

Solution: Allow for a time-lag of 2 months i.e. link advertising expenditure of January with sales for March and so on.

Let x and Y be two variables of advertising expenditure (in lakhs) and sales (in lakhs) respectively.

Computation of Correlation Coefficient

Months	X	Y	$U' = \frac{X - 120}{10}$	$V' = \frac{Y - 2600}{100}$	U'^2	V'^2	$U' \cdot V'$
Jan	50	1600	–7	–10	49	100	70
Feb	60	2000	–6	–6	36	36	36
Mar	70	2200	–5	–4	15	16	20
Apr	90	2500	–3	–1	9	1	3
May	120	2400	0	–2	0	4	0
Jun	150	2600	3	0	9	0	0
Jul	140	2800	2	2	4	4	4
Aug	160	2900	4	3	16	9	12
Sep	170	3100	5	5	25	25	25
Oct	190	3900	7	13	49	169	91
					$\Sigma U'^2 = 222$	$\Sigma V'^2 = 364$	$\Sigma U'.V' = 261$

The Karl Pearson's co-efficient of correlation is given by

$$r = \frac{n\Sigma U' V' - \Sigma U' \cdot \Sigma V'}{\sqrt{n\Sigma U'^2 - (\Sigma U')^2} \cdot \sqrt{n\Sigma V'^2 - (\Sigma V')^2}}$$

$$= \frac{10 \times 261 - 0 \times 0}{\sqrt{10 \times 222 - 0} \cdot \sqrt{10 \times 364 - 0}}$$

$$= \frac{2610}{\sqrt{2220} \cdot \sqrt{3640}}$$

$$= \frac{2610}{2842.67} = 0.92$$

There is high degree of positive correlation between advertising expenditure and sales.

Karl Pearson's Correlation for Bivariate Frequency Distribution (Two way Frequency Table)

When the number of observations in a bivariate distribution is fairly large, then in order to facilitate the calculation of correlation coefficient, the data are often classified according to two measurements in a two-way frequency table called a bivariate frequency table or bivariate frequency distribution. In this distribution, the values of one variable are kept in rows and values of another variable are kept in columns. This values can be discrete or continuous. The frequencies for each cell of the table are determined by tally bar or tally marks.

The correlation coefficient of bivariate distribution is computed by using following formulae:

Direct method

$$r = \frac{N\,\Sigma fXY - (\Sigma fX)\,(\Sigma fY)}{\sqrt{N\Sigma fX^2 - (\Sigma fX)^2}\,\sqrt{N\Sigma\, fY^2}}$$

Shortcut method (Change of origin/Deviation method)

$$r = \frac{N\,\Sigma fUV - (\Sigma fU)\,(\Sigma fV)}{\sqrt{N\Sigma fU^2 - (\Sigma fU)^2}\,\sqrt{N\Sigma\, fV^2}}$$

Where, $U = X - A$ & $V = Y - B$

Step deviation method (or Change of origin and scale)

$$r = \frac{N\,\Sigma fU'V' - (\Sigma fU')\,(\Sigma fV')}{\sqrt{N\Sigma fU'^2 - (\Sigma fU')^2}\,\sqrt{N\Sigma\, fV'^2 - (\Sigma fV')^2}}$$

Where, N = total frequency U' = $\frac{X - A}{h}$ & $V' = \frac{Y - B}{k}$

A = Assumed mean of variable X

B = Assumed mean of variable Y

h = Class size of variable X

k = Class size of variable Y

Steps

a. List the class intervals of two variables X and Y, one is in column heading and another is in row heading.

b. Calculate mid-points of class intervals of variables X and Y and then take deviations (or step-deviations) from their assumed means which are denoted by U and V (or U' and V') respectively.

c. For each class of X, add the frequencies of total cells. Similarly for each class of Y.

d. Multiply the frequency of X variable, with the corresponding value of U and the products are summed up to obtain ΣfU. Similarly, we obtain ΣfV.

e. Again multiply fU with U and fV with V to obtain fU^2 & fV^2 and then obtain ΣfU^2 & ΣfV^2.

f. Multiply f, U and V of each cell and write the figure so obtained in the right-hand corner of each cell.

g. All the values in the top corner are added to get the last column (or row) fUV to obtain ΣfUV. Substitute all sum of values in formula to calculate Correlation Coefficient 'r'.

Example 3.17 Compute the coefficient of correlation between income and expenditure from the following bi-vatiate table.

Expenditure in Rs.	Income Rs.				
	0 - 500	500 - 1000	1000 - 1500	1500 - 2000	2000 - 2500
0 – 400	12	6	8	-	-
400 – 800	2	18	4	5	1
800 – 1200	-	8	10	2	4
1200 – 1600	-	1	10	2	1
1600 – 2000	-	-	1	2	3
Total	14	33	33	11	9

Solution: Let X-series and Y-series be the expenditure (in Rs.) and income (in Rs.) respectively. Also

Assumed mean of X-series (A) =1250

Size of class interval of X-series (h) = 500

Assumed mean of Y-series (B) = 1000

Size of class interval of Y-series (k) = 400

Then, U' and V' are obtained as

$$U' = \frac{X - 1250}{500} \text{ and } V' = \frac{Y - 1000}{400}$$

Income			0-500	500-1000	1000-1500	1500-2000	2000-2500	f	fV'	fV'^2	$fU'V'$
Mid Value (X)			250	750	1250	1750	2250				
Exp.	Mid value (Y)	U' \ V'	– 2	– 1	0	1	2				
0-400	200	– 2	48 | 12	12 | 6	0 | 8	–	–	26	– 52	104	60
400-800	600	– 1	4 | 2	18 | 18	0 | 4	– 5 | 5	–2 | 1	30	– 30	30	15
800-1200	1000	0	–	0 | 8	0 | 10	0 | 2	0 | 4	24	0	0	0
1200-1600	1400	1	–	– 1 | 1	0 | 10	2 | 2	2 | 1	14	14	14	3
1600-2000	1800	2	–	–	0 | 1	4 | 2	12 | 3	6	12	24	16
		f	14	33	33	11	9	$\Sigma f = 100$	$\Sigma fV' = -56$	$\Sigma fV'^2 = 172$	$\Sigma fU'V' = 94$
		fU'	– 28	– 33	0	11	18	$\Sigma fU' = -32$			
		fU'^2	56	33	0	11	36	$\Sigma fU^2 = 136$			
		$fU'V'$	52	29	0	1	12	$\Sigma fU'V' = 94$			

Now, Karl Pearson's co-efficient is given by

$$r = \frac{N\Sigma fU'V' - (\Sigma fU')(\Sigma fV')}{\sqrt{N\Sigma fU'^2 - (\Sigma fU')^2}\sqrt{N\Sigma fV'^2 - (\Sigma fV')^2}}$$

$$= \frac{100 \times 94 - (-32) \times (-56)}{\sqrt{100 \times 136 - (-32)^2} \cdot \sqrt{100 \times 172 - (-56)^2}} = \frac{7608}{13299.205} = 0.57.$$

There is moderate positive correlation between income and expenditure.

Example 3.18 Calculate the coefficient of correlation from the following bivariate frequency distribution. Also, test the significance of *r*.

Sales Revenue (Rs. in Lakh)	Advertising Expenditure in Rs.			
	5000-10000	10000-15000	15000-20000	20000-25000
75-125	4	1	-	-
125-175	7	6	2	1
175-225	1	3	4	2
225-275	1	1	3	4

Solution: Let *X*-series and *Y*-series be the sales revenue (in Rs. lakhs) and advertising expenditure (in Rs. thousands) respectively. Also

Assumed mean of *X*-series (A) =150

Size of class interval of *X*-series (h) = 50

Assumed mean of *Y*-series (B) = 12.5

Size of class interval of *Y*-series (k) = 5

Then, U' and V' are obtained as

Let, $U' = \frac{X - 150}{50}$ and $V' = \frac{Y - 12.5}{5}$

Adv. Exp (Rs. in '000)			5-10	10-15	15-20	20-25	f	fU'	fU'^2	$fU'V'$
Mid Value *Y*			7.5	12.5	17.5	22.5				
Sales Revenue in Lakh Rs.	Mid value *X*	V' \ U'	– 1	0	1	2				
75-125	100	– 2	8 | 4	0 | 1	–	–	5	– 10	20	8
125-175	150	– 1	7 | 7	0 | 6	– 2 | 2	– 2 | 1	16	– 16	16	3
175-225	200	0	0 | 1	0 | 3	0 | 4	0 | 2	10	0	0	0
225–275	250	1	– 1 | 1	0 | 1	3 | 3	8 | 4	9	9	9	10
		f	13	11	9	7	Σf = 40	$\Sigma fU'$ = – 17	$\Sigma fU'^2$ = 45	$\Sigma fU'V'$ = 21
		fV'	– 13	0	9	14	$\Sigma fV'$ = 10			
		fV'^2	13	0	9	28	$\Sigma fV'^2$ = 50			
		$fU'V'$	14	0	1	6	$\Sigma fU'$·			

					V' = 21

Karl Pearson's coefficient of correlation is given by

$$r = \frac{N\Sigma fU'V' - (\Sigma fU')(\Sigma fV')}{\sqrt{N\Sigma fU'^2 - (\Sigma fU')^2}\sqrt{N\Sigma fV'^2 - (\Sigma fV')2}}$$

$$= \frac{40(21) - (-17)(10)}{\sqrt{40(45) - (-17)^2}\sqrt{40(50) - (10)^2}} = 0.596$$

$\therefore \quad r = 0.596.$

$\therefore$ Probable Error (P.E.) $= 0.6745\frac{(1-r^2)}{\sqrt{N}} = 0.6745\frac{(1-0.596^2)}{\sqrt{40}} = 0.69$

Since, $|r| < P.E.$, coefficient of correlation (r) between sales revenue and advertising expenditure is not significant.

Coefficient of Determination

The square of correlation coefficient is called coefficient of determination. It is denoted by r^2. Coefficient of correlation measures the degree of linear relationship (association) between two variables series whereas the coefficient of determination measures percentage of total variation in one variable has been explained by the variation in the other variable. In other words, the coefficient of determination is defined as the ratio of the explained variance to the total variance. Thus,

Coefficient of determination, $r^2 = \frac{\text{Explained variance}}{\text{Total variance}}$

Coefficient of determination is highly applicable in regression analysis in order to measure the Percentage of total variation in dependent variable has been explained by the independent variable.

Suppose, $r = 0.82$, $r^2 = 0.6724 = 0.6724 \times 100\% = 67.24\%$

It implies that 67.24 % of total variation in dependent variable has been explained by independent variable and remaining (100- 67.24)% = 32.76% of the variation due to other factors.

y = f (x)

Dhan = f(Mal)

Dependent= Dhan and independent Mal r^2 = 80% 20% -other factors

Since co-efficient of determination is always positive, so it does not tell us about the direction of the relationship whether it is positive or negative between the two variables. And co-efficient of non-determination is usually denoted by k^2 and is given by $K^2 = 1 - r^2 = \frac{\text{Un-explained variance}}{\text{Total variance}}$.

Rank Correlation (Not in our Syllabus)

In practical problem, we may be faced by the problems of computing correlation between the variables, which are not quantitative in nature. For example, correlation between the variables 'honesty' and 'smartness' among a group of students. Here the variables honesty' and smartness are not of quantitative measurement. These are qualitative in nature. But ranking is possible in case of qualitative variables.

Karl Pearson's co-efficient of correlation cannot measure the relation between the variables which are qualitative in nature. For this situation, British psychologist **Charles Edward Spearman** developed a formula to obtain the correlation co-efficient between the ranks of the variables under the study. This formula (method) works for both quantitative and qualitative variables.

Methods of Studying Rank Correlation Coefficient

There are three cases while computing Spearman's rank correlation coefficient:

Case (1): When the Actual Ranks are given

When the actual ranks are given then the following steps have to be followed:

(i) Find the difference of ranks $d = R_1 - R_2$.

(ii) Compute d^2 to get Σd^2

(iii) Then, find the rank correlation coefficient by using the formula:

$$R = 1 - \frac{6\Sigma d^2}{n(n^2 - 1)}$$

Where,

$d = R_1 - R_2$ = Difference between the pair of ranks.

R_1 = Ranks of items of one variable

R_2 = Rank of items of second variable

n = Number of pair of observations.

Example 3.20 Ten industries of some state have been ranked as follows according to profit earned in and working capital for that year:

Industry	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*J*
Rank of Profit	1	2	3	4	5	6	7	8	9	10
Rank of Working capital	3	2	5	1	4	6	9	10	8	7

Solution: Computation of rank correlation coefficient

Industry	Profit rank (R_1)	Working capital rank (R_2)	$d = R_1 - R_2$	d^2
A	1	3	–2	4
B	2	2	0	0
C	3	5	–2	4
D	4	1	3	9
E	5	4	1	1
F	6	6	0	0
G	7	9	–2	4
H	8	10	–2	4
I	9	8	1	1
J	10	7	3	9
	$n = 10$	$n = 10$	$\Sigma d = 0$	$\Sigma d^2 = 36$

Rank correlation coefficient is given by,

$$R = 1 - \frac{6\sum d^2}{n(n^2-1)} = 1 - \frac{6 \times 36}{10(100-1)}$$

$$= 1 - \frac{216}{990} = 1 - 0.2182 = 0.782$$

There is high degree of positive correlation between the ranks of profit and working capital.

Distinction between Karl Pearson's coefficient of correlation and Spearman's rank correlation coefficient

Karl Pearson's coefficient of correlation	Spearman's rank correlation coefficient
1. Karl Pearson's correlation evaluates the linear relationship between two continuous variables. A relationship is linear when a change in one variable is associated with a proportional change in the other variable.	**1.** Spearman's correlation evaluates the monotonic relationship between two continuous or ordinal variables. In a monotonic relationship, the variables tend to change together but not necessarily at a constant rate. The Spearman's rank correlation coefficient is based on the rank values for each variable rather than the raw data.
2. Karl Pearson's correlation is often used to evaluate relationship involving continuous variables.	**2.** Spearman correlation is used to evaluate relationship involving ordinal variables.
3. Karl Pearson's correlation coefficient is used in regression analysis.	**3.** But Spearman's rank correlation can not be used.
4. It measures the strength of the linear relationship between normally distributed variables.	**4.** When the variables are not normally distributed or the relationship between the variables is not linear. It is more appropriate to use Spearman's rank correlation.
5. Pearson's correlation coefficient between variables is defined as the covariance of the two variables divided by the product of their standard deviations.	**5.** The Spearman correlation coefficient is defined as the Pearson correlation coefficient between the ranked variables.

Regression Analysis

The theory of regression analysis was first developed by British Biometrician **Sir Francis Galton** in 1877. The literal or dictionary meaning of the word "Regression" is "Stepping back" or "Returning back" towards the average. He used the term "regression" in report of his research on heredity (on estimating the nature of relationship between height of fathers and sons).. He made a study and found that the off-springs having either short or tall parents tend to "regress" or "step back" towards the average height of general population. But the term "regression" as now used in Statistics is only a convenient term without having any reference to biometry. Nowadays, it is widely used in business and economics. In Statistics the concept of regression analysis is applicable to all those fields where two or more variables have tendency to move back to the average behaviour.

Galton studied the average relationship between these two variables graphically and called the line of regression.

In statistics, regression analysis is concerned with the **measure of average relationship** between the variables where one variable is **cause** (independent) and other are **effect**s (dependent). Regression explains the nature of relationship between variables. Thus, it can be said that regression is the estimation or prediction of one variables value from the given value of others variables.

In regression analysis, there are two variables dependent and independent. The variable whose value is influenced or is to be predicted is called dependent variable. It is also known as regressed or predicted

or explained variable. On the other hand, the variable which influences the value of dependent variable or the variable which value is used for prediction or estimation of dependent variable is called independent variable. It is also known as regressor, or predictor or explainator. Thus, prediction or estimation is possible in regression analysis.

Prediction or estimation is an activity. For example, estimation of future production, consumption, prices, sales, profits etc. Regression analysis is one of the very scientific techniques for making such predictions which are paramount importance to a manager, decision maker, businessman or economist.

Uses of Regression Analysis

The tools of regression analysis are definitely more useful and important in statistics. Some of the important uses of regression analysis are as follows:

i) Regression analysis helps in establishing relationship between dependent and independent variable.

ii) Regression analysis is very useful for prediction. For example, prediction of sale, profit, income, population etc.

iii) A very important branch of economics, called 'Econometrics' is solely based on the techniques of regression analysis.

iv) The average and correlation co-efficient between two variables can obtained easily by using the regression lines.

v) In social and economic field, it is used for projection of population, birth rates, death rates, marital status, planning etc.

vi) In the business field, it is widely used. For example, businessmen are interested to predict the future production, consumption, investment, prices, profit, sales etc.

vii) It can be used to estimate unknown value of a variable (dependent variable) from the given value of other variable (independent variable) which are interrelated.

viii) Regression analysis helps to explore cause and effect relationship between variables.

Types of Regression

If there are only two variables under consideration, then the regression is called simple regression. For example, the study of regression between income and expenditure. If there are more than two variables under consideration, then the regression is called multiple regression. If there are more than two variables under consideration and relation between only two variables is established excluding the effect of other variables is called partial regression. The simple regression is called linear regression, if the points (dots) on the scatter diagram lies almost along a line, otherwise it is termed as non-linear regression or curvilinear regression. If the graph of dependent and independent variables shows a linear trend (i.e. straight line) then it is called linear regression. Therefore, the linear regression is a first degree equation in the variables X and Y. In case of linear regression, the value of the dependent variable will increase by a constant, absolute amount for a unit of change in the value of the independent variable. But, if it is not in a straight line then it is called non-linear or curvi- linear regression. The non-linear regression equation will be a functional relationship between X and Y involving terms in X and Y of degree higher than one, i.e. involving terms of the type X^2, Y^2, XY, etc. However, in this chapter we will discuss linear regression between two variables only.

Comparison between Correlation and Regression

There is no doubt the fact that there are some differences between the statistical techniques correlation and regression which are as follows:

Correlation	Regression
1. Correlation analysis is the statistical tool (statistical measure) which is used to study or describe the degree of relationship to which the variables are linearly related.	1. Regression analysis is the mathematical measure of the average relationship between two or more variables in terms of original units of data whether the variables are linearly related or non linearly related.
2. It does not necessarily imply cause and effect relationship between the two variables under study.	2. It necessarily shows (indicates) the cause and effect relationship between the variables such that the cause is taken as independent variable and effect is taken as dependent variable.
3. Correlation coefficient i.e. r_{XY} is a relative measure of linear relationship between two variables and is independent of the units of measurement.	3. Regression coefficients b_{xy} & b_{yx} are absolute measures. so, regression coefficients have units of measurement of the variable.
4. Correlation coefficients are symmetric i.e. $r_{XY} = r_{YX}$	4. Regression coefficients are not symmetric. i.e. $b_{xy} \neq b_{yx}$
5. Correlation analysis is confined only to the study of linear relationship between the variables	5. Regression analysis studies linear as well as non linear (curvilinear) relationship between the variables.
6. Correlation coefficients is a pure number lying between -1 & +1.	6. But regression coefficients are absolute measures and if one of the regression coefficients is greater than unity (one), the other must be less than unity. i.e. if $\boldsymbol{b_{yx} > 1}$ then $\boldsymbol{b_{xy} < 1}$.
7. Correlation coefficient is independent of change of origin and scale.	7. Regression coefficients are independent of change of origin but not the scale.
8. Correlation coefficient is independent of units of the variables.	8. Regression coefficient is expressed in the units of dependent variable.
9. The correlation co-efficient cannot be used for prediction.	9. The regression line can be used for prediction.

Regression Lines

A line of regression gives the best estimate of one variable for any given value of the other variable. It two variables x and y are under consideration, there are two lines of regression, one is line of regression of y on x and other is the line of regression of x on y. The line used to estimate the value of x for a given value of y is called the regression line of x on y. Similarly, the line used to estimate the value of y for a given value of x is called the regression line of y on x. The regression lines are also known as estimating lines.

In case of perfect correlation between the variables, the regression lines will be coincident. The angle between the regression lines will be coincident. The angle between the regression lines will increase from 0^o to 90^o as the correlation co-efficient numerically decreases from 1 to 0. If two regression lines are

perpendicular, the pair of variables have correlation co-efficient $r = 0$. The regression lines are determined by using the principle of least square.

Note: The regression lines of Y on X and X on Y intersect at the point $(\bar{X}, \bar{Y})$.

Regression Line and Regression Co-efficient

The regression lines in terms of algebraic expression (linear equation of two variables) are known as the regression equations. Line of regression is the line which gives the best estimate of one variable for any given value of the other variable. In case of simple regression only two variables X & Y are studied. If X and Y are two variables, the algebraic expressions of regression equations in terms of X and Y are called regression equations (lines).

Regression Equation/Line of Y on X

The regression equation of Y on X which describes the variation in the values of dependent variable Y for given changes in independent variable X. So, the line of regression of Y on X is the line which gives the best estimate for the value of Y for any specified value of X.

Since, the line of regression is the line of best fit. so, the term 'best fit' is interpreted in accordance with the principle of least square which consists in minimizing the sum of squares of the residuals or the errors of estimates i.e. the deviations between the given observed values of the variable and their Corresponding estimated values as given by the line of best fit

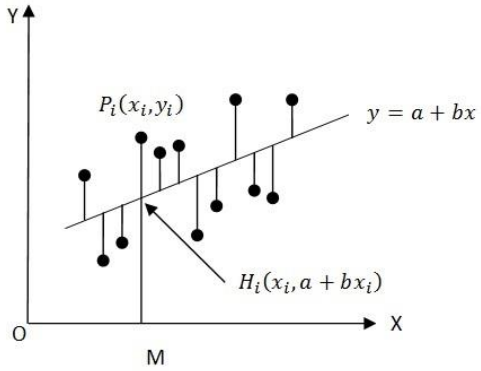
The sum of squares of the errors either parallel to Y-axis or parallel to X-axis may be minimized. The equation of the line of regression of Y on X is obtained by minimizing the sum of squares of errors parallel to Y- axis.

The regression equation/line of y on x is given by $Y = a + b\,X$

Where 'a' and 'b' are constants or parameters to be determined to find the position of the regression line. The parameter 'a' determines the distance of the line above or below the origin and 'b' the **slope** (change in dependent variable per unit change in independent variable) of line as shown below.

Let regression equation of Y on X be

$$y = a + bx \qquad \text{... (i)}$$



Where, Y = dependent variable (Effect)

X = independent variable (Cause)

a = value of Y when $X = 0$

= Intercept of the line (Y - intercept)

$b = b_{YX}$ = **regression coefficient of *Y* on *X*.**

= slope of line

= rate of change in *Y* due to unit change in *X*

By using the techniques of least square, the parameters '*a*' and '*b*' can be obtained by solving two equations. We have, **the regression equation/ line of Y on X** is

$$Y = a + b\,X \qquad \text{…(i)}$$

Taking Σ on both sides of (i)

$$\Sigma Y = n\,a + b\Sigma X \qquad \text{... (ii)}$$

Again, multiplying both sides of (i) by X,

$$XY = a\,X + b\,X^2$$

Taking Σ on both sides,

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

Solving (ii) and (iii) we get **a** and **b** and put in equation (i) to get required equation.

This method is called **Least Square Method**.

For Alternative Method

from equation (i), $\Sigma Y = na + b\,\Sigma X$

$$\Rightarrow \quad \frac{\Sigma Y}{n} = \frac{na}{n} + b\,\frac{\Sigma X}{n}$$

$$\Rightarrow \quad \frac{\Sigma Y}{n} = a + b\,\frac{\Sigma X}{n}$$

$$\therefore \quad a = \bar{Y} - b\bar{X}$$

Putting $a = \bar{Y} - b\bar{X}$ in equation (ii), we get

$$\Sigma XY = (\bar{Y} - b\bar{X})\cdot\Sigma X + b\,\Sigma X^2$$

$$\Rightarrow \quad \Sigma xy = \frac{\Sigma Y}{n} - b\,\frac{\Sigma X}{n} + b\Sigma X^2$$

$$\Rightarrow \quad \Sigma XY - \frac{\Sigma X\cdot\Sigma Y}{n} = b\,\Sigma X^2 - b\left(\frac{\Sigma X}{n}\right)^2$$

$$\Rightarrow \quad \frac{n\Sigma XY - \Sigma X\cdot\Sigma Y}{n} = b\,\frac{[n\,\Sigma X^2 - (\Sigma X)^2]}{n^2}$$

$$\therefore \quad b = \frac{n\Sigma XY - \Sigma X\cdot\Sigma Y}{n\Sigma X^2 - (\Sigma X)^2}$$

i.e. $$b_{YX} = \frac{n\Sigma XY - \Sigma X\cdot\Sigma Y}{n\Sigma X^2 - (\Sigma X)^2}$$

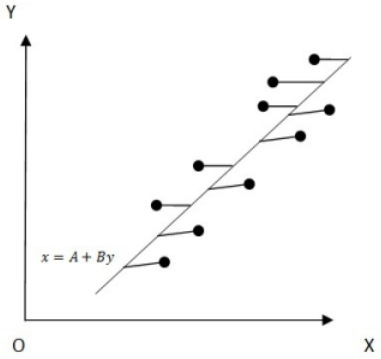
Alternatively, the regression equation of *Y* on *X* is $Y - \overline{Y} = byx\,(X - \overline{X})$

Regression Equation of *X* on *Y*

The regression equation of *X* on *Y* which describes the variation in the values of dependent variable *X* for given changes in independent variable *Y*. Thus, the line of regression of *X* on *Y* is the line which gives the best estimate for the value of *X* for any specified value of *Y*.

Since, the equation of the line of regression of *X* on *Y* is obtained by minimizing the sum of squares of errors parallel to *X*- axis

Then, the required equation of the line of regression of *X* on *Y* becomes. Let regression equation of *Y* on *X* be



$$X = a + bY$$

Where '*a*' and '*b*' are constants or parameters to be determined to locate the position of the regression line. The parameter '*a*' determines the distance of line above or below the origin and '*b*' the slope of line as shown below.

Where

X = dependent variable

Y = independent variable

a = value of X when $Y = 0$

= Intercept of the line (X - intercept)

$b = b_{XY}$ = regression coefficient of X on Y.

= slope of line

= rate of change in X due to unit change in Y

By using the techniques of least square, the parameters '*a*' and '*b*' can be obtained by solving two equations We have, **the regression equation of X on Y** is

$$X = a + bY \qquad \text{…(i)}$$

Taking Σ on both sides of (i)

$$\Sigma X = n\,a + b\Sigma Y \qquad \text{... (ii)}$$

Again, multiplying both sides of (i) by Y,

$$XY = a\,Y + b\,Y^2$$

Taking Σ on both sides,

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

Solving (ii) and (iii) we get a and b and put in equation (i) to get required equation.

This method is called **Least Square Method**.

For **Alternative Method**

from equation (i), $\Sigma X = na + b\Sigma Y = \frac{\Sigma x}{n} = \frac{na}{n} + b\,\frac{\Sigma Y}{n}$

$$\therefore\ a = \left(\frac{\Sigma X}{n} - b\,\frac{\Sigma Y}{n}\right)$$

substituting $a = \frac{\Sigma X}{n} - b\,\frac{\Sigma Y}{n}$ in equation (ii) we get

$$\Sigma XY = \left[\frac{\Sigma X}{n} - b\,\frac{\Sigma Y}{n}\right]\Sigma Y + b\,\Sigma i^2$$

$$\Rightarrow \quad \Sigma XY = \frac{\Sigma X \cdot \Sigma Y}{n} - b\left(\frac{\Sigma Y}{n}\right)^2 + b\,\Sigma Y^2$$

$$\Rightarrow \quad \frac{n\,\Sigma XY - \Sigma X \cdot \Sigma Y}{n} = b\left[\frac{nSY^2 - (SY)^2}{n^2}\right]$$

$$\therefore \quad b = \frac{n\,\Sigma XY - \Sigma X \cdot \Sigma Y}{n\Sigma Y^2 - (\Sigma Y)^2}$$

i.e. $$b_{XY} = \frac{n\Sigma XY - \Sigma X \cdot \Sigma Y}{n\,\Sigma Y^2 - (\Sigma Y)^2} = r\,\frac{\sigma_x}{\sigma_y}$$

Alternatively, the regression equation of X on Y is $X - \bar{X} = b_{xy}\,(Y - \bar{Y})$

In above regression equations b_{yx} and b_{xy} are called regression co-efficient of Y on X and X on Y respectively. Both the regression co-efficient always have same algebraic sign. The regression co-efficient determines the correlation co-efficient by the relation $r = \sqrt{b_{xy} \times b_{yx}}$. Since $|r| \le 1$, so the product of the regression co-efficient must be less than or equal to 1. i.e. $b_{XY} \cdot b_{YX} \le 1$.

Properties of Regression Coefficients

Let X and Y be two variables and two regression coefficients Y on X i.e. b_{YX} and X on Y i.e. b_{XY}. Then

1. The correlation coefficient is the geometric mean between the regression coefficients.

   OR

   Geometric mean between two regression coefficients is equal to the correlation coefficient.

   Mathematically, $r = \pm\sqrt{b_{yx}.b_{xy}}$

2. If one of the regression coefficients is greater than unity (one), the other must be less than unity because $b_{XY} \cdot b_{YX} \leq 1$.

   i.e. if $b_{yx} > 1$ then $b_{xy} < 1$.

3. Arithmetic mean between two regression coefficients is greater than or equal to correlation coefficient

   i.e. $\frac{1}{2}\left(b_{yx} + b_{xy}\right) \geq r$

4. Regression coefficients are independent of change of origin but not of scale.

   Symbolically, $U = \frac{X-a}{h}$, & $V = \frac{Y-b}{k}$

   Then $b_{YX} = \frac{k}{h} \cdot b_{VU}$, & b_{XY}, & $b_{XY} = \frac{h}{k} \cdot b_{UV}$ where a, b, $h(>0)$ & $k(>0)$ are constants.

5. Both regression coefficients must have same sign. The sign of correlation coefficient is same as that of regression coefficients.

6. Regression lines (equations) always pass through their mean values $(\bar{X}, \bar{Y})$ which is also the intersection point of two regression lines.

7. If $r = \pm 1$, the regression lines become identical.

8. If $r = 0$, the regression lines are perpendicular to each other.

Example 3.26 Find the equations of two line of regression if the following results obtained for 5 pair of observations.

$$n = 5, \Sigma X = 15, \Sigma Y = 18, \Sigma X^2 = 55, \Sigma Y^2 = 74, \Sigma XY = 58$$

Solution: The regression equation of X on Y is given by

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

The values of X and Y is obtained using principles of least square as below

$$\Sigma X = n\,a + b\,\Sigma Y$$

or, $$15 = 5a + 18b \qquad \text{… (ii)}$$

and $$\Sigma XY = a\,\Sigma Y + b\,\Sigma Y^2$$

or, $$58 = 18a + 74\,a \qquad \text{... (iii)}$$

Solving equation (ii) and (iii), we get

$$a = 1.45 \text{ and } b = 0.43$$

Putting the value of a and b in equation (i), we get

$$X = 1.45 + 0.43\,Y.$$

Also the regression equation of Y on X is given by

$$Y = a + b\,X \qquad \text{... (ii)}$$

The values of a and b are obtained using principle of least square as below.

$$\Sigma Y = n\,a + b\,\Sigma X$$

or, $18 = 5a + 156$ ... (i)

And $\Sigma XY = a\,\Sigma X + b\Sigma X^2$

or, $58 = 15a + 55b$ ... (ii)

Solving (i) and (ii),we get

$$a = 2.4 \text{ and } b = 0.4$$

Putting the value of a and b in equation (ii), we get

$$Y = 2.4 + 0.4X$$

$\therefore$ The lines of regression are $X = 0.434\ Y + 1.45$ and

$$Y = 0.4\,X + 2.4$$

Alternatively,

$$\bar{X} = \frac{\Sigma X}{n} = \frac{15}{5} = 3 \text{ and } \bar{Y} = \frac{\Sigma Y}{n} = \frac{18}{5} = 3.6$$

Regression Co-efficient of Y on X is given by

$$b_{YX} = \frac{n\,\Sigma XY - \Sigma X \cdot \Sigma Y}{n\,\Sigma X^2 - (\Sigma X)^2} = \frac{5 \times 58 - 15 \times 18}{5 \times 55 - (15)^2} = 0.4$$

and regression co-efficient of X on Y is given by

$$b_{XY} = \frac{n\,\Sigma XY - \Sigma X \cdot \Sigma Y}{n\,\Sigma Y^2 - (\Sigma Y)^2} = \frac{5 \times 58 - 15 \times 18}{5 \times 74 - (18)^2} = 0.43$$

Now, regression equation of Y on X is

$$Y - \bar{Y} = b_{YX}\,(X - \bar{X})$$

$\Rightarrow$ $Y - 3.6 = 0.4\,(X - 3)$

$\Rightarrow$ $Y - 3.6 = 0.4X - 1.2$

$\therefore$ $Y = 0.4X + 2.45$

And regression equation of X on Y is

$$X - \bar{X} = b_{XY}\,(Y - \bar{Y})$$

$X - 3 = 0.43\,(Y - 3.6)$

$\Rightarrow$ $X - 3 = 0.43Y - 1.548$

$\therefore$ $X = 0.43Y + 1.45$

Example 3.27 Given that x and y are correlated variables. Ten observations of value of (x, y) have the following results. n = 10, $\Sigma x = 55$, $\Sigma y = 55$, $\Sigma xy = 350$, $\Sigma x^2 = 385$. Estimate the value of y when the value of x is 6.

Solution: We have to find the value of y when $x = 6$, it requires regression equation of y on x for this regression co-efficient of y on x is given by

$$b_{yx} = \frac{n\Sigma xy - \Sigma x.\Sigma y}{n\,\Sigma x^2 - (\Sigma x)^2} = \frac{10 \times 350 - 55 \times 55}{10 \times 385 - (55)^2} = \frac{475}{825} = 0.58$$

$\bar{x} = 55/10 = 5.5$ and $\bar{y} = 55/10 = 5.5$

$\therefore$ The regression equation of y on x is given by

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

$$\text{or, } y - 5.5 = 0.58\,(x - 5.5)$$

$$\therefore \quad y = 0.58\,x + 2.31$$

Now, when $x = 6$, the estimated value of y is

$$y = 0.58 \times 6 + 2.31 = 5.79$$

Example 3.28 From the following data

	Price (Rs)	Demand of commodity (suitable units)
Arithmetic Mean	36	85
Standard deviation	11	8

Co-efficient of correlation = 0.66

i) Find the equation of regression lines.

ii) Estimate the likely price of commodity when quantity of demanded commodity is 75.

Solution: Let x denotes prices of commodity and y denotes quantity of commodity demanded.

Here, $\bar{x} = 36$, $\bar{y} = 85$, $\sigma_x = 11$, $\sigma_y = 8$ and $r = 0.66$

i) Regression co-efficient of y on x is given by

$$b_{yx} = r\frac{\sigma_y}{\sigma_x} = 0.66 \times \frac{8}{11} = 0.48$$

and Regression co-efficient of x on y is given by

$$b_{xy} = r\frac{\sigma_x}{\sigma_y} = 0.66 \times \frac{11}{8} = 0.91$$

Now Regression equation of y on x is

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

$$y - 85 = 0.48\,(x - 36)$$

$$\therefore \quad y = 0.48x + 67.72$$

and regression equation of x on y is

$$x - \bar{x} = b_{xy}(y - \bar{y})$$

$$x - 36 = 0.91\,(y - 85)$$

$$\therefore \quad x = 0.91y - 41.35$$

ii) To estimate the value of x when the value of $y = 75$ putting $y = 75$ in regression equation of x on y.

i.e. $x = 0.91\ y - 41.35$

or, $x = 0.91 \times 75 - 41.35$

$\therefore \quad x = 26.9$

$\therefore$ The most likely price of commodity is Rs 26.9 when the quantity of commodity demanded is 75.

Example 3.29 Find correlation co-efficient and regression equations for the following data.

Ages (in yrs) :	75	89	97	69	59	79	68	61
Blood Pressure:	125	137	156	112	107	136	123	108

Solution:

Let x denotes ages and y denotes *B.P.*

Again let Assumed mean of x–series (a) = 80

and assumed mean of y–series (b) = 130

Calculation of correlation co-efficient & regression lines

x	y	$u = x - a$	$v = y - b$	uv	u^2	v^2
75	125	–5	–5	25	25	25
89	137	9	7	63	81	49
97	156	17	26	442	289	676
69	112	–11	–18	198	121	324
59	107	–21	–23	483	441	529
79	136	–1	6	–6	1	36
68	123	–12	–7	84	144	49
61	108	–19	–22	418	361	484
		$\Sigma u = -43$	$\Sigma v = -36$	$\Sigma uv = 1707$	$\Sigma u^2 = 1463$	$\Sigma v^2 = 2172$

Now, Karl Pearson's correlation co-efficient is given by

$$r = \frac{n\,\Sigma uv - \Sigma u \cdot \Sigma v}{\sqrt{n\Sigma u^2 - (\Sigma u)^2} \times \sqrt{n\,\Sigma v^2 - (\Sigma v)^2}}$$

$$= \frac{8 \times 1707 - (-43) \times (-36)}{\sqrt{8 \times 1463 - (-43)^2} \times \sqrt{8 \times 2172 - (-36)^2}}$$

$$= \frac{12109}{\sqrt{9855} \times \sqrt{16080}} = 0.9618$$

For Regression lines,

$$\bar{x} = a + \frac{\Sigma u}{n} = 80 + \frac{(-43)}{8} = 74.63$$

and

$$\bar{y} = b + \frac{\Sigma v}{n} = 130 + \frac{(-36)}{8} = 125.5$$

Regression co-efficient of y on x is given by

$$b_{yx} = \frac{n\Sigma uv - \Sigma u \cdot \Sigma v}{n\,\Sigma u^2 - (\Sigma u)^2} = \frac{8 \times 1707 - (-43) \times (-36)}{8 \times 1463 - (-43)^2}$$

$\therefore$ $b_{yx} = 1.22$

Thus, Regression equation of y on x is

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

$$\Rightarrow \quad y - 125.5 = 1.22\,(x - 74.63)$$

$$\therefore \quad y = 1.22x + 34.45$$

Again, Regression co-efficient of x on y is given by $b_{xy} = \frac{n\sum v^2 - \sum U \cdot \sum u}{n\sum v^2 - (\sum u)^2}$.

$$= \frac{8 \times 1707 - (-43).(-36)}{8 \times 2172 - (-36)^2} = 0.75$$

and Regression equation of x on y is

$$\Rightarrow \quad x - \bar{x} = b_{xy}(y - \bar{y})$$

$$\Rightarrow \quad x - 74.63 = 0.75(y - 125.5)$$

$$\Rightarrow \quad x = 0.75y - 19.49$$

Example 3.30 Find the two regression equations from the following marks in Statistics and Mathematics obtained by 8 students.

Marks in Statistics (X)	57	58	59	59	60	61	62	64
Marks in Mathematics (Y)	77	78	75	78	82	82	79	81

i) Estimate the marks in Mathematics when the marks in Statistics is 65.

ii) Find correlation coefficient between marks in Statistics and marks in Mathematics.

Solution: Regression equations of Y on X and X on Y are given by

$$Y - \bar{Y} = b_{YX}(X - \bar{X}) \quad \text{... (i)}$$

$$X - \bar{X} = b_{XY}(Y - \bar{Y}) \quad \text{... (ii)}$$

Calculation of regression equations

X	Y	$x = X - \bar{X}$	$y = Y - \bar{Y}$	xy	x^2	y^2
57	77	–3	–2	6	9	4
58	78	–2	–1	2	4	1
59	75	–1	–4	4	1	4
59	78	–1	–1	1	1	1
60	82	0	3	0	0	9
61	82	1	3	3	1	9
62	79	2	0	0	4	0
64	81	4	2	8	16	4
$\Sigma X = 480$	$\Sigma Y = 632$	$\Sigma x = 0$	$\Sigma y = 0$	$\Sigma xy = 24$	$\Sigma x^2 = 36$	$\Sigma y^2 = 32$

Now,

$$\bar{X} = \frac{\Sigma X}{n} = \frac{480}{8} = 60 \quad \bar{Y} = \frac{\Sigma Y}{n} = \frac{632}{8} = 79$$

Regression coefficient of Y on X

$$b_{YX} = \frac{\Sigma xy}{\Sigma x^2} = \frac{24}{36} = 0.667$$

Regression coefficient of X on Y

$$b_{XY} = \frac{\Sigma xy}{\Sigma y^2} = \frac{24}{32} = 0.75$$

Regression equation of Y on X is given line

$$Y - \bar{Y} = b_{YX}(X - \bar{X})$$

or, $\quad Y - 79 = 0.667 + (X - 60)$

or, $\quad Y = 79 + 0.667X - 40.02$

or, $\quad Y = 39.98 + 0.667X$

Regression equation of X on Y is given line

$$X - \bar{X} = b_{XY}(Y - \bar{Y})$$

$$\Rightarrow \quad X - 60 = 0.75\,(Y - 79)$$

$$\Rightarrow \quad X = 60 + 0.75Y - 0.75 \times 79$$

$$\Rightarrow \quad X = 60 + 0.75Y - 59.25$$

$$\Rightarrow \quad \hat{X} = 0.75 + 0.75Y$$

i) When $X = 65$, then

$$\hat{Y} = 38.98 + 0.667 \times 65$$

$$= 38.98 + 43.355 = 82.34$$

ii) $r = \sqrt{b_{YX} \times b_{XY}} = \sqrt{0.667 \times 0.75} = 0.707$. Since both the regression co-efficient b_{XY} and b_{YX} are positive, so r is also positive. i.e. $r = 0.707$.

There is high degree positive correlation between marks in Statistics and marks in Mathematics.

Example 3.31 Given is the following information:

	X (in Rs.)	Y (in Rs.)
Arithmetic mean	6	8
Standard deviation	5	40/3

Coefficient of correlation between X and Y is $\frac{8}{15}$

Find

a) The regression coefficient of Y on X and X on Y
b) The two regression equations.
c) The most likely value of Y when $X = 100$ rupees.

Solution: We have

$$\text{Mean of } X = \bar{X} = 6$$

$$\text{Mean of } Y = \bar{Y} = 8$$

$$\text{S.D. of } X = \sigma_x = 5$$

$$\text{S.D. of } Y = \sigma_y = \frac{40}{3}$$

Coefficient of correlation between X and $Y(r) = \frac{8}{15}$

Now,

a) Regression coefficient of Y on X is

$$b_{yx} = r\frac{\sigma_y}{\sigma_x} = \frac{8}{15} \times \frac{40/3}{5} = \frac{8}{15} \times \frac{40}{3} \times \frac{1}{5} = 1.422.$$

Similarly, regression coefficient of X on Y is

$$b_{xy} = r\frac{\sigma_x}{\sigma_y} = \frac{8}{15} \times \frac{5}{40/3} = \frac{8}{15} \times \frac{3}{40} \times \frac{5}{1} = 0.20$$

b) The regression equation of Y on X is

$$Y - \bar{Y} = b_{yx}(X - \bar{X})$$

or, $Y - 8 = 1.422\,(X - 6)$

or, $Y = 8 + 1.422X - 1.422 \times 6$

or, $\hat{Y} = 1.422X - 0.532$

Similarly, the regression equation of X on Y is

$$X - \bar{X} = b_{xy}(Y - \bar{Y})$$

or, $X - 6 = 0.2\,(Y - 8)$

or, $X = 6 + 0.2Y - 0.2 \times 8$

or, $\hat{X} = 0.2Y + 4.4$

c) When $X = 100$, then

$\hat{Y} = 1.422 \times 100 - 0.532 = 142.2 - 0.532 = \text{Rs. } 141.67.$

$\therefore$ The mostly like value of Y is Rs. 141.67.

Example 3.32 In a partially destroyed record, the following data are available:

Variance of $X = 25$, the regression lines are

$5x - Y = 22$ and $64X - 45Y = 24$

Find

a) Mean value of X and Y.

b) Coefficient of correction between X and Y.

c) Standard deviation of Y.

Solution:

a) The regression equations of X on Y and Y on X are

$$5x - Y = 22 \quad \text{... (i)}$$

$$64X - 45Y = 24 \quad \text{... (ii)}$$

Since, both the lines of regression pass through the mean values, the point $(\bar{X}, \bar{Y})$ must satisfy equations (i) and (ii).

Hence, we get

$$5\bar{X} - \bar{Y} = 22 \quad \text{... (iii)}$$

$$64\bar{X} - 45\bar{Y} = 24 \quad \text{... (iv)}$$

Multiplying equation (i) by 45 and subtracting equation (iv) from equation (i), we get

$$\begin{array}{llll} 225\,\bar{X} & -\ 45\bar{Y} & = & 990 \\ 64\,\bar{X} & -\ 45\bar{Y} & = & 24 \\ - & + & = & \\ \hline 161\bar{X} & +\quad 0 & = & 966 \end{array}$$

or, $\bar{X} = \dfrac{966}{161} = 6$

Substituting the value of $\overline{X}$ in equation (iii), we get

$$5 \times 6 - \overline{Y} = 22$$

or, $$\overline{Y} = 30 - 22$$

or, $$\overline{Y} = 8$$

Hence, the mean of values are $\overline{X} = 6,\ \overline{Y} = 8$

b) The regression equation of X on Y is

$$5X - Y = 22$$

or, $$5X = 22 + Y$$

or, $$X = \frac{22}{5} + \frac{1}{5} \cdot Y$$

Comparing this equation with $X = a + bY$

$$b_{XY} = \frac{1}{5}$$

Similarly, the regression equation of Y on X is

$$64X - 45Y = 24$$

or, $$45Y = -24 + 64X$$

or, $$Y = \frac{-24}{45} + \frac{64}{45} X$$

Comparing this equation with $Y = a + bX$

$$b_{YX} = \frac{64}{45}$$

Hence, correlation coefficient between two variables X and Y is given by

$$r = \pm \sqrt{b_{YX} \cdot b_{XY}} = \pm \sqrt{\frac{64}{45} \times \frac{1}{5}} = \pm \sqrt{\frac{64}{225}} = \pm \sqrt{0.2844} = \pm\ 0.533$$

Since, both regression coefficients are positive, r must be positive.

Hence $r = 0.533$.

c) We have, $\sigma_X^2 = 25$ then $\sigma_x = 5$

we know, $$b_{YX} = r \frac{\sigma_y}{\sigma_x}$$

or, $$\frac{64}{45} = 0.533 \times \frac{\sigma_Y}{5}$$

or, $$\sigma_Y = \frac{64 \times 5}{45 \times 0.533} = 13.333$$

Example 3.33 The correlation coefficient between two variables X and Y is $r = 0.60$. If

Average of variable $X = 10$,

Average of variable $Y = 20$,

Coefficient of variation of $X = 15$,

Coefficient of variation of $Y = 10$

Find

a) The regression equation of Y on X

b) The Most likely value of Y when $X = 18$

Solution:

a) The regression equation of Y on X is

$$Y - \bar{Y} = b_{YX}(X - \bar{X}) \qquad \text{... (i)}$$

Here, $r = 0.60$, $\bar{X} = 10$, $\bar{Y} = 20$, C.V. $(X) = 15$, C.V. $(Y) = 10$

For variable X,

$$\text{C.V. } (X) = \frac{\sigma_X}{\bar{X}} \times 100$$

$$\Rightarrow \quad 15 = \frac{\sigma_X}{10} \times 100 \quad \Rightarrow \quad \sigma_X = \frac{15 \times 10}{100} = 1.5$$

For variable Y,

$$\text{C.V. } (Y) = \frac{\sigma_Y}{\bar{Y}} \times 100$$

$$\Rightarrow \quad 10 = \frac{\sigma_y}{10} \times 100 \quad \Rightarrow \sigma_y = \frac{10 \times 20}{100} = 2$$

$$\therefore \quad b_{YX} = r\frac{\sigma_Y}{\sigma_X} = 0.60 \times \frac{2}{1.5} = 0.8$$

The regression equation of Y on x is

$$Y - \bar{Y} = b_{YX}(X - \bar{X})$$

$$\Rightarrow \quad Y - 20 = 0.8(X - 10)$$

$$\Rightarrow \quad Y = 20 + 0.8X - 0.8 \times 10$$

$$\Rightarrow \quad \hat{Y} = 12 + 0.8X$$

b) When $X = 18$, the mostly likely value of Y is

$$\hat{Y} = 0.8 \times 18 + 12 = 14.4 + 12 = 26.4$$

Regression Equation for a Bivariate Frequency Distribution

The Calculation for obtaining regression equation are more or less same as the calculation of correlation co-efficient 'r' from a bivariate frequency which has been already discussed in this lesion. However, since the regression co-efficient b_{YX} and b_{XY} are independent of change of origin but not scale, so they are computed as

$$u = \frac{x - a}{h},\ v = \frac{y - b}{k},\ \text{then } b_{yx} = \frac{k}{h} b_{VU} \text{ and } b_{xy} = \frac{h}{k} b_{UV}$$

where a = assumed mean of x–series

b = assumed mean of y–series

h = common constant or width of classes of x–series

k = common constant or width of classes of y –series

That is

$$b_{YX} = \frac{n\,\Sigma fuv - \Sigma fu \cdot \Sigma fv}{n\,\Sigma fu^2 - (\Sigma fu)^2} \times \frac{k}{h} \;\&\; b_{XY} = \frac{n\,\Sigma fuv - \Sigma u.\,\Sigma fv}{n\,\Sigma fv^2 - (\Sigma fv)^2} \times \frac{h}{k}$$

Example 3.34 Family income and its percentage spent on food in the case of hundred families gives the following bivariate frequency distribution.

Food expenditure in %	Family income (Rs)				
	200 – 300	300 – 400	400 – 500	500 – 600	600 – 700
10 – 15	–	–	–	3	7
15 – 20	–	4	9	4	3
20 – 25	7	6	12	5	–
25 – 30	3	10	19	8	–

i) Find regression co-efficient

ii) Find regression equations

iii) Estimate the income of a family whose food expenditure is 21%

iv) Calculate correlation co-efficient also test the significance of r.

Solution: Let x denotes food expenditure and y denotes income of family. Also let assumed mean of X–series (a) = 22.5 and assumed mean of y–series (b) = 450. We have,

$$h = 5 \text{ and } k = 100, \text{ then } u = \frac{x - 22.5}{5} \text{ and } v = \frac{y - 450}{100}$$

income / expenditure	x	y / u	200–300 250	300–400 350	400–500 450	500–600 550	600–700 650	f	fu	fu^2	fuv
C.I			–2	–1	0	1	2				
10 - 15	12.5	–2	–	–	–	3 –6	7 –28	10	-20	40	34
15 – 20	17.5	–1	–	4 4	9 0	4 –4	3 –6	20	-20	20	–6
20 – 25	22.5	0	7 0	6	12 0	5 0	–	30	0	0	0
25 – 30	27.5	1	3 –3	10 –10	19 0	8 8	–	40	40	40	–8
		f	10	20	40	20	10	$N = 100$	$\Sigma fu = 0$	$\Sigma fu^2 = 100$	$\Sigma fuv = -48$
		fv	–20	–20	0	20	20	$\Sigma fv = 0$			
		fv^2	40	20	0	20	40	$\Sigma fv^2 = 120$			
		fuv	–3	–6	0	–2	–34	$\Sigma fuv = -48$			

i) Regression co-efficient of y on x is

$$b_{xy} = \frac{N\Sigma fuv - \Sigma fu\Sigma fv}{N\Sigma fu^2 - (\Sigma fu)^2} \times \frac{k}{h} = \frac{100 \times (-48) - 0 \times 0}{100 \times 100 - 0^2} \times \frac{100}{5}$$

$= -9.6$

and Regression co-efficient of x on y is

$$b_{xy} = \frac{N\Sigma fuv - \Sigma fu\Sigma fv}{N\Sigma fv^2 - (\Sigma fv)^2} \times \frac{h}{k} = \frac{100\times(-48) - 0\times 0}{100\times 120 - 0^2} \times \frac{5}{100}$$

$$= -0.02$$

ii) Regression equations are $y - y - \bar{y} = b_{yx}(x - \bar{x})$ and $x - \bar{x} = b_{XY}(y - \bar{y})$

We have, $\bar{x} = a + \frac{\Sigma fu}{N} \times h = 22.5 + \frac{0}{100} \times 5 = 22.5$

& $\bar{y} = b + \frac{\Sigma fv}{N} \times k = 450 + \frac{0}{100} \times 100 = 450$

Regression equation of y on x is $y - \bar{y} = b_{xy}(x - \bar{x})$

$$y - 450 = -9.6\,(x - 22.5)$$

$$\therefore \quad y = -9.6x + 666$$

And Regression equation of x on y is

$$x - \bar{x} = b_{xy}\,(y - \bar{y})$$

$$x - 22.5 = -0.02\,(y - 450)$$

$$\therefore \quad x = -0.02y + 31.5$$

iii) Here food expenditure of family $(x) = 21\%$

To estimate likely income of family, putting $x = 21$ in regression equation of y on x i.e

$$y = -9.6x + 666$$

$$y = -9.6 \times 21 + 666$$

$$y = 464.4$$

$\therefore$ Estimated income of the family is Rs. 463.5

iv) We have, Two regression co-efficient are

$$b_{yx} = -9.6 \text{ and } b_{xy} = -0.02$$

Correlation co-efficient $(r) = \sqrt{b_{xy} \times b_{yx}} = -\sqrt{(-9.6) \times (-0.02)} = -0.43$

Since both the regression co-efficient negative, so correlation co-efficient is also negative.

i.e. r = -0.43

Also, to test the significance of 'r'

$$\text{P.E.}\,(r) = 0.6745 \times \frac{1 - r^2}{\sqrt{N}}$$

$$= 0.6745 \times \frac{1-(-0.43)^2}{\sqrt{100}}$$

$$= 0.6745 \times 0.0875 = 0.055$$

Since $|r| > P.E.(r)$, $6P.E.(r) = 6 \times 0.056 = 0.329$

Thus r is significant because $|r| > 6P.E.(r)$.

Example 3.35 It the correlation co-efficient between two variables x and y is 0.9, then what is the co-efficient of determination? Also, interpret it.

Solution: Here, correlation co-efficient $(r) = 0.9$

We have, co-efficient of determination $= r^2 = (0.9)^2 = 0.81$. This implies that 81% of the variation in dependent variable has been explained by the independent variable and the remaining 19% of the variation is due to the other variable.

Example 3.36 A correlation between two variables X and Y is 0.6 and correlation between other two variables U and V is 0.3. Does it mean that first correlation is twice as strong as the second? Give reason.

Solution: Here, Correlation between X and Y is $r_{XY} = 0.6$ and correlation between U and V is $r_{UV} = 0.3$

No, it does not mean that first correlation i.e., correlation between X and Y is twice as strong as the second correlation i.e. correlation between U and V because correlation co-efficient measure only linear relationship between the variables. To compare, the correlation, it is required to obtain percentage of variation of dependent variable explained by the variation of independent variable.

Example 3.37 The data gives the marks in statistics and Mathematics obtained by 7 students. Find i) Correlation co-efficient ii) By what percent the variation of marks in Statistics is due to variation of marks in Mathematics?

Students	x	y	xy	x^2	y^2
A	22	41	902	484	1681
B	24	44	1056	576	1936
C	25	45	1125	625	2025
D	27	48	1296	729	2304
E	21	40	840	441	1600
F	22	42	924	484	1764
G	23	44	1012	529	1936
Total	$\Sigma x = 164$	$\Sigma y = 304$	$\Sigma xy = 7155$	$\Sigma x^2 = 3868$	$\Sigma y^2 = 13246$

i) Karl Pearson's co-efficient of correlation is given by

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

$$= \frac{7 \times 7155 - 164 \times 304}{\sqrt{7 \times 3868 - (164)^2} \times \sqrt{7 \times 13246 - (304)^2}}$$

$$= \frac{229}{\sqrt{180} \times \sqrt{306}}$$

$$= 0.9757$$

ii) Co-efficient of determination = $r^2 = (0.9757)^2 = 0.952$

Hence 95.2% variation in marks obtained in Statistics is explained by the marks obtained in Mathematics.

Exercise 3.1

Theoretical Questions

1. What is meant by correlation? Write the measures of correlation between two variables.

2. Define correlation. Explain various types of correlation.

3. Does the degree of correlation between two variables signify the existence of cause and effect relationship between the variables? Explain it.

4. Define Karl Pearson's co-efficient of correlation. What are the special characteristics of Pearsonian Correlation co-efficient?

5. What do you mean by probable error of correlation co-efficient? Write its uses.

6. Define Spearman's rank correlation co-efficient. How is it different from Karl Pearson's correlation co-efficient? Discuss.

7. What is co-efficient of determination? writes its uses.

8. What is regression analysis? How does it differ from correlation?

9. What do you mean by regression? Write its uses dealing with business.

10. Define regression co-efficient. Write its properties.

11. Differentiate between

i) Positive correlation and negative correlation.

ii) Partial correlation and multiple correlation.

iii) Linear regression and curvilinear regression.

iv) Correlation co-efficient and regression co-efficient.

12. Write down the equation of the two regression lines, explaining the constants used in it. Also write the properties of regression co-efficient.

13. Define scatter diagram. State its merits and demerits.

Exercise 3.2

Numerical and Practical Problems

1. From the following data, ascertain with the help of satterdiagram, whether the income and expenditure of the workers of a industry are correlated or not.

Years :	1979	1980	1981	1982	1983	1984	1985
Average income (in Rs.)	210	215	215	222	230	236	245
Average expenditure (in Rs):	205	208	212	218	225	230	237

2. Find Karl Pearson's Co-efficient of Correlation, when

i) Cov $(X, Y) = 10$, Var $(X) = 6.25$ and var $(Y) = 13.36$

ii) $\bar{X} = 25$, $\bar{Y} = 18$, $\Sigma(X - \bar{X})^2 = 136$, $\Sigma(Y - \bar{Y})^2 = 138$,

$\Sigma(X - \bar{X})(Y - \bar{Y}) = 122$ and $N = 15$

iii) $n = 10$, $\Sigma x = 18$, $\Sigma y = 25$, $\Sigma x^2 = 90$, $\Sigma y^2 = 120$ and $\Sigma xy = 65$

iv) $n = 10$, $\Sigma xy = 120$, $\sigma_x = 3$, and $\sigma_y = 8$ where $x = (X - \bar{X})$ and $y = (Y - \bar{Y})$.

3. Find the Karl Person's Correlation co-efficient between x and y if the observation (x, y) are follows: (9, 8), (8, 10), (6, 9), (5, 7), (10, 5), (6, 6), (4, 2), (3, 0), (2, 2), (1, 1)

4. Calculate 'r' for the following data:

x:	–3	–2	–1	0	1	2	3
y:	9	4	1	0	1	4	9

5. Find the co-efficient of correlation from the following data:

X:	300	350	400	450	500	550	600	650	700
Y:	800	900	1000	1100	1200	1300	1400	1500	1600

Also draw the scatter diagram and interpret it.

6. Find the Karl Pearson's co-efficient of correlation between the sale and expenditure of a firm for six month.

Months:	Jan	Feb	Mar	Apr	May	June
Sale (in '00'):	50	55	60	65	60	70
Expenses (in Rs'00') :	12	12	15	14	13	14

7. Compute the co-efficient of correlation from the following data.

Age	56	42	36	47	49	42	60	72	63	55
Blood pressure	147	125	118	128	145	140	155	160	149	150

8. From the following table giving the distribution of boys and also regular football players among them according to age group, find out correlation between 'age' and playing habit of boy.

Age (in years):	15	16	17	18	19	20
No. of boys:	200	270	340	360	400	300
No. of regular players:	150	162	170	180	180	120

9. (a) The following table gives the distribution of the total population and those who are totally or partially blind among them. Find out if there is any relation between age and blindness.

Age (in years) :	0 – 10	10 – 20	20 – 30	30 – 40	40 – 50	50 – 60	60 – 70	70 – 80
No. of people ('000'):	100	60	40	36	24	11	6	3
No. of blind:	55	40	40	40	36	22	18	15

(b) The following table provides the result of some examinations of different levels of a college.

Age of the students	15–17	17–19	19–21	21–23	23–25
No. of attendance	12	15	20	25	30
No. of successful students	9	12	14	20	27

Find the correlation coefficient between the age of students and the success rate in the different level.

10. With the following data in 6 Localities, calculate the co-efficient of correlation by Karl Pearson's method between the density of population and the death rate.

Localities	**Area in Sq. km.**	**Population ('000')**	**No. of deaths**
A	150	30	300
B	180	90	1440
C	100	40	560
D	60	42	840

E	120	72	1224
F	80	24	312

11. Co-efficient of correlation between X and Y for 20 items is 0.3, mean of X-series is 15 and that of Y is 20, standard deviations are 4 and 5 respectively. At the time of calculation, one item 17 was wrongly copied instead of 27 in case of X–series and 35 instead of 30 in case of Y– series. Find the correct co-efficient of correlation.

12. A computer while calculating the correlation co-efficient between the variable x and y obtained the following results.

$$N = 35, \quad \Sigma x = 120, \quad \Sigma x^2 = 550$$
$$\Sigma y = 105, \quad \Sigma y^2 = 500, \quad \Sigma xy = 350$$

If was, however, later discovered at the time of checking that it had copied down two pair of items as (8, 7) and (12, 13) instead of (7, 8) and (13, 12) respectively obtain the correct value of the correlation co-efficient between x and y.

13. Following information's given below are related with the ages of husband (X) and wife (Y) for married couples living together in a sample survey. Calculate the co-efficient of correlation between the age of husband and that of his wife. Test the significance of calculate value of r.

$$N = 72, \Sigma fx = 3560, \Sigma fx^2 = 196800, \Sigma fy = 3260, \Sigma fy^2 = 168400 \; \Sigma fxy = 172000.$$

14. From the data given below, find the co-efficient of correlation between the drivers age and the number of accidents made by them.

Number of accidents	**Drivers age**				
	25 – 30	30 – 35	35 – 40	40 – 45	45 – 50
0	–	3	5	7	8
1	–	–	9	4	1
2	3	5	10	3	–
3	4	9	6	–	–
4	12	7	3	1	–

15. The following table gives the distribution of sales (in Rs. 00) and profit (in Rs '00') of 100 shops. Find the co-efficient of correlation and its probable error. Also state whether correlation co-efficient is significant or not.

Sales (in Rs '000')	**Profit (in Rs. '00')**				
	10 – 20	**20 – 30**	**30 – 40**	**40 – 50**	**50 – 60**
350 – 450	–		–	–	5
450 – 550	–	–	1	10	9
550 – 650	–	4	12	25	3
650 – 750	4	16	2	2	–
750 – 850	2	5	–	–	–

16. Correlation co-efficient between the ages of fathers and sons is 0.9031. Discuss if the value of r is significant or not. Also compute the limits for population correlation.

17. Ten students were examined in Accountancy and Statistics. The ranks obtained by the students are given below. Find the Spearman's rank Correlation co-efficient.

Ranks in	1	2	3	4	5	6	7	8	9	10

Accountancy:										
Ranks in Statistics:	2	4	1	5	3	9	7	10	6	8

18. In Board Examination, 10 students got the following percentages in Economics and Accountancy papers.

Economics	81	36	98	25	75	82	92	62	65	39
Ranks in Statistics:	84	51	91	60	68	62	86	58	35	49

Find the rank correlation co-efficient

19. Find the rank Correlation co-efficient from the following data.

X:	54	57	61	64	64	72	72	79	86	89
Y:	31	36	50	60	83	93	90	85	99	122

20. Ten competitors in a beauty contest are ranked by three judges in the following order.

I judge:	9	3	7	5	1	6	2	4	10	8
II judge:	9	1	10	4	3	8	5	2	7	6
III judge:	6	3	8	7	2	4	1	5	9	10

Calculation appropriate rank correlation and answer the following question.

i) Which pair of judges disagree the most?

ii) Which pair of judges has the nearest approach to commontaste of beauty?

21. The co-efficient of rank correlation between beautiness and intelligence of 10 girls was found to be 0.5. It was later discovered that the difference in ranks of a girl was wrongly taken as 3 instead of 7. Find the correct correlation co-efficient.

22. If $r = 0.5$ then find the co-efficient of determination and interpret the result.

23. Calculate the co-efficient of correlation between *X* and *Y*.

X:	7	6	5	4	3	2	1
Y:	18	16	14	12	10	6	8

Also find the percentage of variation explained.

24. Find the regression lines for the following pairs (x, y) for the values of x and y.
(1, 6), (5, 1), (3, 0), (2, 0), (1, 1), (1, 2), (7, 1), (3, 5)

25. From the following data, obtain the two regression equations

Sales (in Rs.):	91	97	108	121	67	124	51	73	111	57
Purchase (in Rs.):	71	75	69	97	70	91	39	61	80	47

26. The following table gives the ages and blood pressure of 10 women.

Age:	56	42	36	47	49	42	60	72	63	55
B.P.	147	125	118	128	145	140	155	160	149	150

i) Find the correlation co-efficient between age and B.P.

ii) Determine the least square regression equations.

iii) Estimate the blood pressure of a woman of age 45 years.

27. The advertisement expenses and the sales of a product are recorded as below.

Adv. exp. (Rs. '000')	1	5	6	8	10
Sales (Rs '000')	50	60	80	100	110

Estimate the sales when advertisement expenses is Rs. 15,000.

28. The following data gives the experience of machine operators in years and their performance as given by the number of good parts turned out per 100 pieces.

Operator:	1	2	3	4	5	6	7	8
Experience:	16	22	18	4	3	10	5	12
Performance:	87	88	89	68	78	80	75	83

Determine regression equation and estimate the experience of the operator having performance of 158.87.

29. From the following data

X:	40	34	?	30	44	38	31
Y:	32	39	26	30	38	34	28

The arithmetic mean of X – series is calculated as $\bar{x} = 35$. Find the regression equation of Y on X. and estimate Y when $X = 36$.

30. In a survey of ages 10 pairs of husband and wife, the following data is recorded.

Age of husband:	25	22	28	26	35	20	22	40	20	18
Age of wife:	18	15	20	17	22	14	16	21	15	14

The age of eleventh husband is 30 years. He deny to give the importation of his wife's age, predict the age of his wife assuming that the same degree of relationship exists between their ages.

31. A Computer while calculating the correlation co-efficient between two variables X and Y from 8 pair of observations obtained the following results $N = 8$, $\Sigma x = 562$, $\Sigma x^2 = 39602$, $\Sigma Y = 561$, $\Sigma Y^2 = 39851$, $\Sigma XY = 39815$. It was however later discovered at the time of checking that the computer had copied down two pairs as

X	Y		X	Y
76	56	While correct values were	67	65
76	86		67	68

Calculate the correct values of co-efficient of correlation &

i) Test whether the value of calculated co-efficient of correlation is significant or not.

ii) Find regression co-efficient of X on Y

iii) Find regression co-efficient of Y on X.

iv) Find regression line of Y on X.

v) Estimate Y when $X = 65$.

32. Given the information:

Sum of $x = 5$, Sum of $Y = 4$

Sum of square of deviations from mean of $X = 40$

Sum of square of deviations from mean of $Y = 50$

Sum of the products of deviations from the means of X on $Y = 32$, no. of pair of observations = 10.

i) Find regression co-efficient of Y on X and X on Y.

ii) Find Pearsonian co-efficient of correlation.

33. While calculating the co-efficient of correlation between two variables X and Y, the following results obtained.

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

Later it was found that two pair of observations (X, Y) were copied (6, 14) and (8, 6) at the time of checking while the correct values were (8, 12) and (6, 8) respectively. Determine the

i) Correct values

ii) Correct Correlation co-efficient

iii) Correct equations of the lines of regression

iv) Probable error and interpret the significance of the co-efficient of correlation.

34. For a bivariate data, the mean value of X is 20 and mean value of Y is 45. The regression co-efficient of Y on X is 4 and that of X on Y is $\frac{1}{9}$.

Find

i) Co-efficient of correlation

ii) The standard deviation of X if standard deviation of Y is 12.

iii) The two regression equations

iv) Estimate the value of X when $Y = 25$.

35. You are given the following information about profit and Sales:

	Profit (Rs. in lakhs)	Sale (Rs in Lakhs)
A.M	10	90
S.D	3	12

$r = 0.8$

i) Find the regression co-efficient

ii) Find the equations of lines of regression.

iii) Find the estimated sale when profit is Rs. 15 Lakhs

iv) What should be the profit if a company wants to attain sales target of Rs. 120 Lakhs.

36. Out of the following two regression lines, identify the line of regression of x on y and line of regression of y on x. Why?

$2x + 3y - 7 = 0$ and $5x = 4y - 9 = 0$

37. The equations of two regression lines obtained in a regression analysis are as follows:

$3x + 12y - 19 = 0$ and $9x + 3y - 46 = 0$, obtain

i) The means of x and y

ii) The regression co-efficient of y on x and x on y.

iii) Correlation co-efficient between x and y.

38. For 50 students in a class of *BBS*, the regression equation of marks in Statistics (y) on the marks in Economics (x) is $5x - 4y + 8 = 0$. Average marks in Economics is 44 and the ratio of standard deviations $\sigma_y : \sigma_x$ is 5: 2 find the average marks in statistics and co-efficient of correlation between marks in two subjects.

39. The two regression lines are given by

$3x + 2y = 6$ and $7x + 5y = 12$

i) Indentify the lines of regression

ii) Estimate y when $x = 10$

iii) Calculate the co-efficient of correlation between x & y.

iv) What percentage of total variation remains unexplained by the regression equation of y on x?

40. Find the correlation between the two variables from the following bi-variate frequency table.

Marks in Science	**Marks in English**			
	0 – 25	25 – 50	50 – 75	75 – 100
0 – 25	2	–	–	–
25 – 50	1	3	1	1
50 – 75	–	1	4	5
75 – 100	–	–	5	17

Also estimate the marks in Science of a student who secured 95 in English.

41. Following is the distribution of students according to their height and weight.

i) Calculate the two regression co-efficients.

ii) Obtain two regression equations.

iii) Estimate the weight of student whose height is 58 inch.

iv) obtain the height of student whose weight is 115 lbs.

Height (in inches)	**Weight in (lbs)**			
	90 – 100	100 – 110	110 – 120	120 – 130
50 – 55	4	7	5	2
55 – 60	6	10	7	4
60 – 65	6	12	10	7
65 – 70	3	8	6	3

Answers

Numerical and Practical Problems:

1. High degree positive linear correlation.

2. i) 0.72 ii) 0.89 iii) 0.35 iv) 0.5

3. 0.754 **4.** r = 0 **5.** 1, perfect positive linear correlation

6. 0.70 **7.** 0.892, high degree positive correlation

8. –0.92 **9.** (a)0.89 (b) 0.64 **10.** 0.99 **11.** 0.92

12. 0.093 **13.** $r = 0.52$ r is significant **14.** –0.699 **15.** $r = –0.823$

16. r is significant and limits are 0.8626 and 0.9436. **17.** 0.7575 **18.** 0.7575

19. 0.952 **20.** i. II and III ii. I and III **21.** 0.258,

22. $r^2 = 0.25$ and 25% of the data are explained. **23.** $r = 0.96$ and 92.16% of variation explained

24. $y = – 0.3042x + 2.8745$ and $x = – 0.27784y + 3.4306$

25. $y = 0.6132$ x + 14.812 and $x = 1.361y – 5.27$

26. i) $r = 0.89$ ii) $y = 83.758 + 1.11x$ iii) B.P. = 134 when age = 45

27. Rs, 144,557 **28.** $y = 1.133x + 69.67$, 78.73

29. $y = 0.59x + 32$, 53.24 **30.** 19 years nearly **31.** $r = 0.60$ i) no. conclusion

ii) bxy = 0.545 iii) byx = 0.667 iv) $y = 0.667x + 23.644$

v) 67

32. i) $b_{yx} = 0.80$. $b_{xy} = 0.64$ ii) $r = 0.7156$

33. i) $\Sigma x = 125$, $\Sigma y = 100$, $\Sigma x^2 = 650$, $\Sigma Y^2 = 43$, $\Sigma XY = 520$

ii) $r = 0.667$ iii) $Y = 0.8x$, $x = 0.556Y + 2.776$

iv) P.E. (r) = 0.075, significant.

34. i) $r = 0.67$ ii) $\sigma y = 2$ iii) $y = 4x - 35$ and $x = \frac{1}{9}y + 15$

iv) $x = 17.78$

35. i) $b_{yx} = 3.2$ and $b_{xy} = 0.2$ ii) $y = 3.2 + 58$ and $x = 0.2y - 8$

iii) Estimated sale = Rs. 106 Lakhs.

36. x on y is $5x + 4y - 9 = 0$ and y on x is $2x + 3y - 7 = 0$ because $b_{yx} \times b_{xy} < 1$

37. i) $\bar{x} = 5$, $\bar{y} = \frac{1}{3}$ ii) $b_{yx} = -\frac{1}{4}$ & $b_{xy} = -\frac{1}{3}$

iii) $r = -0.29$

38. $\bar{y} = 52$ and $r = 0.5$

39. i) x on y is $3x + 2y = 6$ and y on x is $7x + 5y = 12$

ii) $y = -11.6$ iii) r = – 0.96 iv) 7.84%

40. $r = 0.736$, marks in science = 87.5

41. i) $b_{yx} = 0.0405$ and $b_{xy} = 0.1518$

ii) $y = 0.0405x = 55.9314$ and $x = 0.15187 + 99.94$

iii) 108.47 lbs

iv) 60.6 inches.

Exercise 3.3

Analytical Answer Questions

1. The data on sales and promotion expenditures on a newly launched product for 6 years are given below:

Year	2003	2004	2005	2006	2007	2008
Sales (in Rs.00,000)	16	20	18	24	20	22
Promotion expenses (Rs.'000'	4	4	6	10	10	12

(a) Calculate the two regression coefficients from the above data of sales and expenses.

(b) Compute the correlation coefficient between sales and promotional expenses and interpret it.

(c) Test the significance of the correlation coefficient.

(d) Develop the estimating equation that describes the effect of promotion expenses on sales. Estimate sales if the promotional expenses is Rs. 20,000.

(e) Explain the meaning of each parameter of the equation; in terms of above information.

2. A researcher is interested in seeking how accurately a new job performance index measures what is important for corporation. One way to check is to look at the relationship between job evaluation index and an employee's salary is significant or not. A sample of eight employees was taken and information about salary in thousand rupees and job performance index (1-10, 10 is best) was collected.

Job Performance index :	9	7	8	4	7	5	5	6

Salary '000' Rs.:	36	25	33	15	28	19	20	22

Develop an estimating equation that best describes these data and estimate the salary of an employee whose job performance is 10 and 2 . Also find whether there exists significant relationship between the variables.

3. A computer while calculating the correlation coefficient between two variants X and Y from 8 pairs of observation obtained the following results:

$n = 8$, $\Sigma X = 562$, $\Sigma X^2 = 39602$, $\Sigma Y = 561$, $\Sigma Y^2 = 39815$, $\Sigma XY = 39441$

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

X	Y
76	56
76	86

While the correct values were

X	Y
67	65
67	68

Calculate the correct of value of the correlation coefficient between X and Y and

(a) Test whether calculated coefficient of correlation is significant or not.

(b) Regression coefficient of X on Y.

(c) Regression coefficient of Y on X.

(d) Regression line of Y on X

(e) Estimate Y when $X = 65$

4. While calculating the correlation coefficient between two variables X and Y, the following results were obtained.

$X = 8$ and 12, $Y = 10$ and 7 were copied wrongly, the corresponding correct values being $X = 8$ and 10, $Y = 12$ and 8. Obtain the correct values and then develop the regression equation of X on Y and Y on X and finally find out whether there exists any relationship between these two variables or not.

5. The income and expenditure of 100 families is given below:

Income (Rs.)	**Expenditure (Rs.)**			
	0-500	500-1000	1000-1500	1500-2000
0 -1000	-	-	-	3
1000- 2000	-	4	9	4
2000-3000	3	10	19	8
3000-4000	7	6	12	5
4000-5000	3	7	-	-

Find (a) Two regression coefficients

(b) Coefficient of correlation between income and expenditure.

(c) Mode value of expenditure and income.

(d) Coefficient of variation of expenditure and income.

6. From the following bi-variate table.

Expenditure (Rs.)	**Income (Rs.)**				
	0 – 500	500 – 1000	1000 – 1500	1500 – 2000	2000 – 2500
0 – 400	2	6	8	–	–
400 – 800	2	18	4	5	1
800 – 1200	–	8	10	2	4
1200 – 1600	–	1	10	2	1

1600 – 2000	–	–	1	2	3

i) Compute two regression co-efficient

ii) Co-efficient of correlation between income and expenditure.

iii) Estimate the expenditure of person, when his income is Rs 4,000.

iv) Which is more uniform, income distribution or expenditure distribution?

v) Find the modal expenditure.

vi) Determine the median value of frequency distribution of income.

Answers

Analytical and Comprehensive Problems:

1. (a) $bXY = 0.545$, (b) $r = 0.6$, (c) no conclusion

(d) $\hat{Y} = 23.644 + 0.667\ X$ (e) $\hat{Y} = 67{,}00{,}000$

2. $Y = -2.11321 + 4.2138\ X$, $r = 0.9853$, r is highly significant

3. $b_{xy} = 0.607$, $b_{yx} = 0.9$, $r = 0.739$, nothing can be concluded, $x = 15.34 + 0.607$, Y; Rs. 2748000

4. $Y = -2.11321 + 4.2138X$, $r = 0.9853$, r is highly significant.

5. (a) $b_{YX} = 0.208$, $b_{XY} = -0.865$ (b) $r = -0.424$, , Negative correlation

(c) Mode (income) = Rs.2696.97, Mode (expenditure) = Rs.1196.97

(d) C.V. (Income) = 34.58%, C.V. (Expenditure) = 43.265%

6. (i) $b_{yx} = 0.484$ and $b_{xy} = 0.676$ (ii) $r = 0.572$

(iii) Estimated expenditure = Rs. 2184.44 (iv) expenditure is more uniform than income

(v) M_o = Rs 480 (vi) M_d = Rs 1045.45

Exercise 3.4

Multiple Choice Questions circle (O) the correct answer.

1. What is the range of correlation coefficient?

(a) 0 to ∞ (b) –∞ to ∞ **(c)** –1 to 1 (d) 0 to 1

2. If r = 0.3 then coefficient of determination implies that

(a) 30% of total variation in dependent variable has been explained by in dependent variable.

(b) 60% of total variation in dependent variable has been explained by in dependent variable.

(c) 3% of total variation in dependent variable has been explained by in dependent variable.

(d) 4% of total variation in dependent variable has been explained by in dependent variable.

3. The regression line of X on Y and Y on X are intersect at the point.

(a) (μ, 0) (b) (a, b) (c) (X, Y) **(d)** $(\bar{X}, \bar{Y})$

1. The term regression was introduced by:

(a) R.A. Fisher **(b)** Sir Francis Galton (c) Karl Pearson (d) None of above

2. If X and Y are two variates, there can be at most:

(a) one regression line **(b)** two regression lines

(c) three regression lines (d) an infinite number of regression lines

3. In a regression line of Y on X, the variable X is known as:

(a) independent variable (b) regressor

(c) explanatory variable (d) All the above

4. Regression equation is also named as:

(a) prediction equation (b) estimating equation

(c) line of average relationship (d) all the above

5. Scatter diagram of the variate values (X, Y) gives the idea about:

(a) functional relationship **(b)** regression model
(c) distribution of errors (d) none of the above

6. The estimate of β in the regression equation $Y = \alpha + \beta X + e$ by the method of least squares is:
(a) biased **(b)** unbiased (c) consistent (d) efficient

7. The formula for the estimation of β in the regression equation $Y = \alpha + \beta X + \varepsilon$ is:
(a) cov $(X, Y) / V(X)$ (b) $r \cdot \frac{\sigma_Y}{\sigma_X}$
(c) $\Sigma(X_i - \bar{X})(Y_i - \bar{Y}) / \Sigma(X_i - \bar{X})^2$ **(d)** All the above

8. In the regression line $Y = \alpha + \beta X$, β is called the:
(a) slope of the line (b) intercept of the line
(c) neither (a) nor (b) (d) both (a) and (b)

9. In the regression line $Y = \beta_o + \beta_1 X$, β_o is the:
(a) slope of the line **(b)** intercept of the line
(c) both (a) and (b) (d) neither (a) nor (b)

10. If β_{YX} and β_{XY} are two regression coefficients, they have:
(a) same sign (b) opposite sign
(c) either same or opposite signs (d) nothing can be said

11. The property that β_{YX} and β_{XY} and ρ have same signs, it called:
(a) fundamental property **(b)** signature property
(c) magnitude property (d) none of the above

12. The average of two regression coefficients is always greater than or equal to the correlation coefficient is called:
(a) fundamental property (b) signature property
(c) magnitude property **(d)** mean property

13. If $\beta_{YX} > 1$, then β_{XY} is:
(a) less than 1 (b) greater than 1 (c) equal 1 (d) equal to 0

14. If $\beta_{YX} < 1$, then β_{XY} is:
(a) less than 1 **(b)** greater than 1 (c) equal to 1 (d) equal to 0

20. If $\rho = \pm 1$, the two lines of regression are
(a) coincident (b) parallel (c) perpendicular to each other
(d) none of the above

21. If ρ = 1, he angle between the two line of regression is:
(a) zero degree (b) ninety degree (c) sixty degree (d) thirty degree

22. If ρ = 0, the lines of regression are:
(a) coincident (b) parallel **(c)** perpendicular to each other
(d) none of the above

***