Unit - I

- **CURVE FITTING:**

Curve fitting is a method of estimating unknown constants in the equation $Y = f(X)$ on the basis of given n pairs of observations$(Xi, Yi), i = 1, 2 \ldots n$. Theoretically it is used in Correlation and Regression analysis and practically it is used to represent the functional relationship between two related variables. The various functional relations are: (i) $Y = a + bX$ (ii) $Y = a + bX + cX^2$ (iii) $Y = ab^X$ (iv) $Y = aX^b$

It is used to estimate the values of one variable for the given values of the other variable. It is used to predict the future value of one variable for the given value of other variable.

The general method used for the fitting is the Principle of least squares. In this method the unknown constants in the equation $Y = f(X)$ can be estimated in such a way that the error sum of squares is as minimum as possible i.e. E= $\sum_{i=1}^{n} ei^2$ = $\sum_{i=1}^{n} (Yi - \hat{Y}i)^2$ is as minimum as possible where Yi is observed values of Yi and $\hat{Y}i$ is expected value of Yi.

Fitting of a Straight Line:

We wish to fit a straight line $Y = a + bX$ on the basis of given n pairs of observations (Xi, Yi), i=1, 2…n. where a and b are two unknown constants. So the problem is to estimate the unknowns a and b which gives the best straight line fitted for the given data. Thus, according to Principle of least squares, we have to determine a and b in such a way that error sum of squares E= $\sum_{i=1}^{n} ei^2$ = $\sum_{i=1}^{n} (Yi - \hat{Y}i)^2$ = $\sum_{i=1}^{n} (Yi - (a + bXi))^2$ = $\sum_{i=1}^{n} (Yi - a - bXi)^2$ is minimum. From the principle of maxima and minima, differentiate E partially, w.r.to a and b and equate to zero i.e.

$$\frac{\partial E}{\partial a} = 0 \Rightarrow 2\sum_{i=1}^{n} (Yi - a - bXi)(-1) = 0$$

$$\Rightarrow (-2) \sum_{i=1}^{n} (Yi - a - bXi) = 0$$

$$\Rightarrow \sum_{i=1}^{n} Yi = na + b\sum_{i=1}^{n} Xi \text{ ---------- (I)}$$

And

$$\frac{\partial E}{\partial b} = 0 \Rightarrow 2\sum_{i=1}^{n} (Yi - a - bXi)(-Xi) = 0$$

$$\Rightarrow -2\sum_{i=1}^{n} Xi(Yi - a - bXi) = 0$$

$$\Rightarrow \sum_{i=1}^{n} XiYi = a\sum_{i=1}^{n} Xi + b\sum_{i=1}^{n} Xi^2 \text{ ---------- (II)}$$

And

$$\frac{\partial^2 E}{\partial a^2} = 2\sum_{i=1}^{n} 1 = 2n > 0$$

$$\frac{\partial^2 E}{\partial b^2} = 2\sum_{i=1}^{n} Xi^2 > 0$$

Equation (I) and (II) are called <u>normal equations</u> for estimating a and b. After solving the normal equations, we get the values of $\hat{a}$ and $\hat{b}$ which gives the minimum error sum of squares. After substituting these values of $\hat{a}$ and $\hat{b}$, we get the equation of straight line $\hat{Y} = \hat{a} + \hat{b}X$ which is the best for the given set of observations.

<u>Fitting of Second Degree Parabola:</u>

We wish to fit second degree parabola $Y = a + bX + cX^2$ on the basis of given n pairs of observations (Xi, Yi), i=1, 2…n. where a, b and c are unknown constants. So the problem is to estimate the unknowns a, b and c which gives the best equation fitted for the given data. Thus, according to Principle of least squares, we have to determine a, b and c in such a way that

E= $\sum_{i=1}^{n} ei^2 = \sum_{i=1}^{n} (Yi - \hat{Yi})^2 = \sum_{i=1}^{n} (Yi - (a + bXi + cXi^2))^2 = \sum_{i=1}^{n} (Yi - a - bXi - cXi^2)^2$ **is minimum. From the principle of maxima and minima, differentiate E partially, w.r.to a, b and c and equate to zero**

i.e. $\frac{\partial E}{\partial a} = 0 \Rightarrow 2\sum_{i=1}^{n} (Yi - a - bXi - cXi^2)(-1) = 0$

$$\Rightarrow (-2)\sum_{i=1}^{n} (Yi - a - bXi - cXi^2) = 0$$

$$\Rightarrow \sum_{i=1}^{n} Yi = na + b\sum_{i=1}^{n} Xi + c\sum_{i=1}^{n} Xi^2 \text{ ---------- (I)}$$

And

$$\frac{\partial E}{\partial b} = 0 \Rightarrow 2\sum_{i=1}^{n} (Yi - a - bXi - cXi^2)(-Xi) = 0$$

$$\Rightarrow (-2)\sum_{i=1}^{n} Xi(Yi - a - bXi - cXi^2) = 0$$

$$\Rightarrow \sum_{i=1}^{n} XiYi = a\sum_{i=1}^{n} Xi + b\sum_{i=1}^{n} Xi^2 + c\sum_{i=1}^{n} Xi^3 \text{ ---------- (II)}$$

And

$$\frac{\partial E}{\partial c} = 0 \Rightarrow 2\sum_{i=1}^{n} (Yi - a - bXi - cXi^2)(-Xi^2) = 0$$

$$\Rightarrow (-2)\sum_{i=1}^{n} Xi^2(Yi - a - bXi - cXi^2) = 0$$

$$\Rightarrow \sum_{i=1}^{n} Xi^2Yi = a\sum_{i=1}^{n} Xi^2 + b\sum_{i=1}^{n} Xi^3 + c\sum_{i=1}^{n} Xi^4 \text{ ---------- (III)}$$

$$\frac{\partial^2 E}{\partial a^2} = 2\sum_{i=1}^{n} 1 = 2n > 0$$

$$\frac{\partial^2 E}{\partial b^2} = 2\sum_{i=1}^{n} Xi^2 > 0$$

$$\frac{\partial^2 E}{\partial c^2} = 2\sum_{i=1}^{n} Xi^4 > 0$$

Equation (I) and (II) and (III) are called <u>normal equations</u> for estimating a, b and c. After solving the normal equations, we get the values of $\hat{a}$, $\hat{b}$ and $\hat{c}$ which gives the minimum error sum of squares. After substituting

these values of $\hat{a}$, $\hat{b}$ and $\hat{c}$ we get the equation of second degree parabola $\hat{Y} = \hat{a} + \hat{b}X + \hat{c}X^2$ is the best for the given set of observations.

<u>Fitting of Exponential curve</u>:

Sometimes the functional form is not linear but exponential. In such a situation first convert that relation to the linear by simple transformation, and then apply the technique of linear fitting.

e.g. let the relation be (i) $Y = ab^X$ (ii) $Y = aX^b$

(i) Fitting of $Y = ab^X$

Let us consider the problem of fitting an exponential curve $Y = ab^X$ to n pairs of observations (Xi, Yi), i=1, 2…n. So the problem is to estimate the unknowns a and b which gives the best equation fitted for the given data. Thus, according to Principle of least squares, we have to determine a and b in such a way that the error sum of squares is minimum.

Taking log to the base 'e' i.e natural log on both sides of $Y = ab^X$, we get

$\log_e Y = \log_e a + X \log_e b$

i.e $Y_1 = a_1 + b_1 X$ where $Y_1 = \log_e Y$, $a_1 = \log_e a$, $b_1 = \log_e b$

But the equation $Y_1 = a_1 + b_1 X$ is the equation of straight line. Thus applying the same technique as applied in fitting of straight line we have following normal equations.

$\sum Y_1 = na_1 + b_1 \sum X$ **---------- (I)**

$\sum XY_1 = a_1 \sum X + b_1 \sum X^2$ **---------- (II)**

After solving the normal equations (I) and (II) for a_1 and b_1, we get the estimates of a and b as $\hat{a} = \exp(a_1)$ and $\hat{b} = \exp(b_1)$. After substituting these values of $\hat{a}$ and $\hat{b}$ in equation $\hat{Y} = \hat{a}\hat{b}^X$, we get the equation of exponential curve which is the best for the given set of observations.

(ii) Fitting of $Y = aX^b$:

Let us consider the problem of fitting an exponential curve $Y = aX^b$ to n pairs of observations (Xi, Yi), i=1, 2…n. So the problem is to estimate the unknowns a and b b which gives the best equation fitted for the given data. Thus, according to Principle of least squares, we have to determine a and b in such a way that the error sum of squares is minimum.

Taking log to the base 'e' i.e natural log on both sides of $Y = aX^b$, we get

$\log_e Y = \log_e a + b \log_e X$

i.e. $Y_1 = a_1 + bX_1$ **where** $Y_1 = \log_e Y, a_1 = \log_e a, X_1 = \log_e X$

But the equation $Y_1 = a_1 + bX_1$ is the equation of straight line. Thus applying the same technique as applied in fitting of straight line we have following normal equations.

$\sum Y_1 = na_1 + b \sum X_1$ **---------- (I)**

$\sum X_1 Y_1 = a_1 \sum X_1 + b \sum X_1^2$ **---------- (II)**

After solving the normal equations (I) and (II) for a_1 and b, we get the estimates of as $\hat{a} = \exp(a_1)$. After substituting these values of $\hat{a}$ and $\hat{b}$ in equation, $Y = aX^b$, we get the equation of exponential curve which is the best for the given set of observations.

<u>Examples:</u>

1. **Fit a straight line to the following data:**

X	0	5	10	15
Y	1.80	1.45	1.18	1.00

Estimate Y when X = 20 and 22.

2. **The results of measurements of electric resistance of a copper bar at various temperatures are listed below:**

Temperature (X)	20	25	30	35	40	45	50
Resistance(^{0}c)(Y)	76	77	79	80	82	83	85

Fit an equation $Y = a + bX$ to the data given above using least squares method and estimate resistance when temperatures are 55 and 60^{0}c.

3. **For the data given below, fit straight line and estimate the literacy rate for the years 2011 and 2021.**

Year	1951	1961	1971	1981	1991	2001
Literacy rate	16.7	24.0	29.5	43.7	52.2	65.5

4. **From the following data, fit a straight line and predict the amount of export for the year 2012.**

Year	2003	2004	2005	2006	2007	2008	2009	2010
Export ('Crore Rs.)	196.2	217.4	241.9	302.4	385.2	468.8	535.5	521.2

5. **Fit a second degree parabola to the data given below:**

X	2	4	6	8	10
Y	3.07	12.85	31.47	57.38	91.29

Estimate Y when X = 12.

6. **The following table gives the results of measurements of train resistances; X is the velocity in miles per hour Y is the resistance in pounds per ton. Using method of least squares , fit a second degree parabola $Y = a + bX + cX^2$ to the data given below:**

X	20	40	60	80	100	120
Y	5.5	9.1	14.9	22.8	33.3	46.0

Estimate Y for x = 140, 150.

7. **The following table gives the results of measurements of train resistances; X is the velocity in miles per hour Y is the resistance in pounds per ton.**

X	20	40	60	80	100	120
Y	5.5	9.1	14.9	22.8	33.3	46.0

If Y is related to X by the relation $Y = a + bX + cX^2$, find a, b and c.

8. **The following table shows our urban population as the percentage of the total population**

Census year	1921	1931	1941	1951	1961
Urban population (%)	11.4	12.1	13.9	17.3	18.0

Fit an equation $Y = a + bX + cX^2$, to the data given above and estimate the urban population for the census year 1971 and 2011.

9. **Fit second degree parabola to the data given data below:**

X	1.0	1.5	2.0	2.5	3.0	3.5	4.0
Y	1.1	1.3	1.6	2.0	2.7	3.4	4.1

10. **Fit a curve $Y = aX^b$ to the data given below:**

X	1	2	3	4	5
Y	0.5	2	4.5	8	12.5

Estimate Y when X = 6 and 7.

11. Using method of least squares fit an equation of the form $Y = ab^X$ to the data given below:

X	2	3	4	5	6
Y	144	172.8	207.4	248.8	298.5

Predict Y when X = 7 and 9.

12. The sales of a company for the years (1995 – 1999) are given below:

Year (X)	1995	1996	1997	1998	1999
Sales ('Lacs Rs.) (Y)	65	92	132	190	275

Estimate sales for the years 2000 and 2001 using an equation of the form $Y = ab^X$.

- **Correlation**

Correlation is a statistical tool to study the relationship (cause and effect or causal) between two variables. Correlation analysis involves various methods and techniques used for studying and measuring the extent (or degree or amount) of relationship between two related variable.

In our day to day life, we observe that the following two variables are always occur together.

For example,

(a) Income and Saving (b) Price and Demand of a commodity (c) Running speed and Consumption of oxygen (d) Temperature and Butterfat (e) Age and B.P. (f) Cigarette smoking and Illness (g) Dosage of the drug and Reduction in B.P. h) Inflation and Unemployment etc.

One of the main objectives of a research study is to examine the relationships among different variables. The strength of relationship between any two variables is measured called correlation coefficient.

Definition: Two variables X and Y are said to be correlated if changes in one variable results the changes in other variable in such a way that an increase in one variable results an increase or decrease in the other.

Types of Correlation:

1. Positive correlation
2. Negative correlation
3. Zero correlation

1. **Positive correlation**:

When changes in two related variables takes place in the same direction i.e. increase (or decrease) in one variable results an increase (or decrease) in the other variable, then the two variables are positively correlated.

In positive correlation, if the rate of increase or decrease in two variables is constant then the correlation is perfect positive otherwise the correlation is partially positive.

e.g. (a) Consumption of electricity and Electricity bill

(b) Income and expenditure

(c) Running speed and Oxygen consumption

2. **Negative correlation**:

When changes in two related variables takes place in the opposite direction i.e. increase (or decrease) in one variable results a decrease (or increase) in the other variable, then the two variables are negatively correlated.

In negatively correlation, if the rate of increase or decrease in two variables is constant then the correlation is perfect negatively otherwise the correlation is partially negative.

e.g. (a) Temperature and Butterfat

(b) Price and Demand of a commodity

3. **Zero correlation**:

If a change in the values one variable doesn't affect the values of the other variable then we say that there is no correlations between two variables i.e. zero correlation.

Remark: If the two variables are not related by cause and effect relationship but the variable shows the correlation between them then such a correlation is called spurious or non-sense correlation.

Methods of studying correlation:

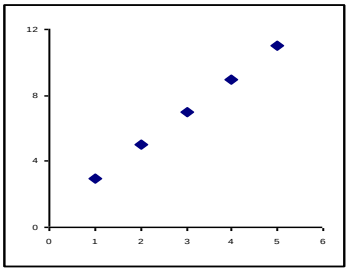
1. Scatter diagram method
2. Karl-Pearson's correlation coefficient method
3. Spearman's rank correlation coefficient method.

1. **Scatter diagram method**:

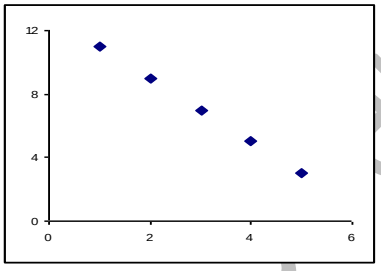
The simplest method of studying (observing) the correlation between two related variables is the scatter diagram method. In this method, the pairs of observations (Xi, Yi), i = 1, 2…, n be plotted on XY plane by taking values of X and Y along X-axis and Y-axis respectively by means of a point. Then the diagram of dots so obtained is called a Scatter diagram.

The scatter diagram method indicates the direction of correlation and tells how closely the two variables under study are related. It doesn't provide an extent (or degree or amount) of relationship between them.

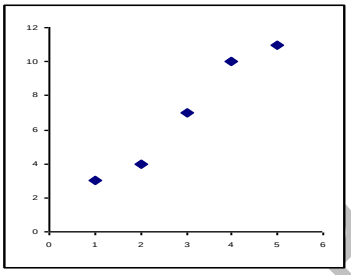
From this method we have following types of relationships:



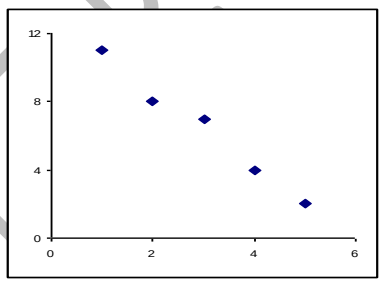
(Perfect positive correlation)



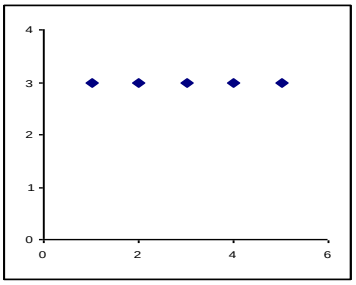
(Perfect negative correlation)



(Partial positive correlation)



(Partial Negative correlation)



(Zero correlation)

2. **Karl-Pearson's correlation coefficient method**:

Scatter diagram method gives only the direction of relationship but it is silent about the amount or extent or degree of correlation between the two related variables.

A numerical measure of linear relationship between two related variables X and Y is given by Karl-Pearson, which is known as correlation coefficient and is denoted by r or r_{XY} and is given by

$$r = \frac{\text{Cov}(X, Y)}{SxSy}$$

Where $\text{Cov}(X, Y) = \frac{1}{n}\sum_{i=1}^{n}\left(Xi - \overline{X}\right)\left(Yi - \overline{Y}\right)$

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} Xi$$

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Yi$$

$$Sx^2 = \frac{1}{n}\sum_{i=1}^{n}\left(Xi - \overline{X}\right)^2$$

$$Sy^2 = \frac{1}{n}\sum_{i=1}^{n}\left(Yi - \overline{Y}\right)^2$$

Properties of Correlation coefficient:

1. **Correlation coefficient is lies between -1 and 1 i.e. $-1 \le r \le 1$**

OR

The correlation coefficient r cannot exceed unity i.e. $|r| \le 1$

Proof: Let (Xi, Yi), i=1, 2…n be the given n pairs of observations then Karl-Pearson's correlation coefficient between two variables X and Y is given by

$$r = \frac{Cov(X, Y)}{SxSy}$$

Where $Cov(X, Y) = \frac{1}{n}\sum_{i=1}^{n}\left(Xi - \overline{X}\right)\left(Yi - \overline{Y}\right)$

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} Xi$$

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Yi$$

$$Sx^2 = \frac{1}{n}\sum_{i=1}^{n}\left(Xi - \overline{X}\right)^2$$

$$Sy^2 = \frac{1}{n}\sum_{i=1}^{n}\left(Yi - \overline{Y}\right)^2$$

Let us consider the sum of squares

$$\sum_{i=1}^{n}\left(Ui \pm Vi\right)^2 \ge 0$$

Where $Ui = \frac{Xi - \overline{X}}{Sx}$ and $Vi = \frac{Yi - \overline{Y}}{Sy}$

i.e. $\sum_{i=1}^{n}\left[Ui^2 \pm 2UiVi + Vi^2\right] \ge 0$

$$\Rightarrow \sum_{i=1}^{n}\left[\left(\frac{Xi - \overline{X}}{Sx}\right)^2 \pm 2\left(\frac{Xi - \overline{X}}{Sx}\right)\left(\frac{Yi - \overline{Y}}{Sy}\right) + \left(\frac{Yi - \overline{Y}}{Sy}\right)^2\right] \ge 0$$

$$\Rightarrow \frac{1}{Sx^2}\sum_{i=1}^{n}\left(Xi - \overline{X}\right)^2 \pm \frac{2}{SxSy}\sum_{i=1}^{n}\left(Xi - \overline{X}\right)\left(Yi - \overline{Y}\right) + \frac{1}{Sy^2}\sum_{i=`}^{n}\left(Yi - \overline{Y}\right)^2 \ge 0$$

Dividing both the sides by n we get,

$$\frac{1}{Sx^2}\frac{1}{n}\sum_{i=1}^{n}\left(Xi - \overline{X}\right)^2 \pm \frac{2}{SxSy}\frac{1}{n}\sum_{i=1}^{n}\left(Xi - \overline{X}\right)\left(Yi - \overline{Y}\right) + \frac{1}{Sy^2}\frac{1}{n}\sum_{i=`}^{n}\left(Yi - \overline{Y}\right)^2 \ge 0$$

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

$$\Rightarrow 2 \pm 2r \ge 0$$

$\Rightarrow 1 \pm r \geq 0$

$\Rightarrow 1 - r \geq 0$ and $1 + r \geq 0$

$\Rightarrow 1 \geq r$ and $r \geq -1$

$\Rightarrow r \leq 1$ and $r \geq -1$

$\Rightarrow \boxed{-1 \leq r \leq 1}$

2. **Correlation coefficient is an independent of change of origin and scale**.

Mathematically, if X and Y are the given variables and they are transformed to the new variables U and V by the change of origin and scale

i.e. $Ui = \frac{Xi - A}{c} and Vi = \frac{Yi - B}{h}$ where A, B, c and h are constants; c, h > 0 then the correlation coefficient between X and Y is same as the correlation coefficient between U and V.

i.e. $r_{XY} = r_{UV}$

Proof: Let (Xi, Yi), i=1, 2…n be the given n pairs of observations then Karl-Pearson's correlation coefficient between two variables X and Y is given by

$$r_{XY} = \frac{\text{Cov}(X, Y)}{SxSy} \text{ ---------- (A)}$$

Where $\text{Cov}(X, Y) = \frac{1}{n}\sum_{i=1}^{n}\left(Xi - \overline{X}\right)\left(Yi - \overline{Y}\right)$

$$Sx^2 = \frac{1}{n}\sum_{i=1}^{n}\left(Xi - \overline{X}\right)^2$$

$$Sy^2 = \frac{1}{n}\sum_{i=1}^{n}\left(Yi - \overline{Y}\right)^2$$

Let $Ui = \frac{Xi - A}{c} and Vi = \frac{Yi - B}{h}$, for i = 1, 2…n;

Thus, we have

$Xi = A + cUi$ and $Yi = B + hVi$, for i = 1, 2…n;

$\therefore \overline{X} = A + c\overline{U}$ and $\overline{Y} = B + h\overline{V}$

$\therefore \left(Xi - \overline{X}\right) = c\left(Ui - \overline{U}\right)$ and $\left(Yi - \overline{Y}\right) = h\left(\left(Vi - \overline{V}\right)\right.$, for i = 1, 2…n;

$$\therefore Cov(X, Y) = \frac{1}{n}\sum_{i=1}^{n}\left(c\left(Ui - \overline{U}\right)h\left(Vi - \overline{V}\right)\right)$$

$$= ch\frac{1}{n}\sum_{i=1}^{n}\left(Ui - \overline{U}\right)\left(Vi - \overline{V}\right)$$

$$= chCov(U, V) \text{ ---------- (I)}$$

i.e Covariance is an independent of change of origin but not of scale.

Also $Sx^2 = \frac{1}{n}\sum_{i=1}^{n}\left(Xi - \overline{X}\right)^2$

$$= \frac{1}{n}\sum_{i=1}^{n}c\left(Ui - \overline{U}\right)^2$$

$$= c^2\frac{1}{n}\sum_{i=1}^{n}\left(Ui - \overline{U}\right)^2$$

$$= c^2 S_U{}^2 \text{ ---------- (II)}$$

i.e. variance is an independent of change of origin but not of scale.

Similarly, $Sy^2 = h^2 Sv^2$ ---------- (III)

Thus from (A), (I), (II) and (III), we get,

$r_{XY} = r_{UV}$

3. **Two independent variables are uncorrelated but the converse is not true.**

Proof: We have

$$r_{XY} = \frac{\text{Cov}(X, Y)}{SxSy} \quad \text{---------- (I)}$$

Where Cov(X, Y) = $\frac{1}{n}\sum_{i=1}^{n}\left(Xi - \overline{X}\right)\left(Yi - \overline{Y}\right)$

$$= \frac{1}{n}\sum_{i=1}^{n} XiYi - \left(\overline{X}\right)\left(\overline{Y}\right)$$

$$= \frac{1}{n}\sum XY - \left(\overline{X}\right)\left(\overline{Y}\right)$$

$$= E(XY) - E(X)E(Y)$$

$\because$ Expected value of a variable is nothing but its mean.

Now, if X and Y are independent then

$$E(XY) = E(X)E(Y)$$

$$\therefore Cov(X, Y) = 0$$

$$\therefore r_{XY} = 0$$

i.e. X and Y are uncorrelated.

$\Rightarrow$ Independent variables are uncorrelated.

Conversely, the result is not true.

i.e. Uncorrelated variables need not necessarily be independent.

We explain it by giving counter example.

X	- 4	- 3	- 2	- 1	1	2	3	4	$\sum X = 0$
Y	16	9	4	1	1	4	9	16	$\sum Y = 60$
XY	- 64	- 27	- 8	- 1	1	8	27	64	$\sum XY = 0$

We have,

$$r_{XY} = \frac{\text{Cov}(X, Y)}{SxSy}$$

Where Cov(X, Y) = $\frac{1}{n}\sum_{i=1}^{n}\left(Xi - \overline{X}\right)\left(Yi - \overline{Y}\right)$

$$= \frac{1}{n}\sum_{i=1}^{n} XiYi - \left(\overline{X}\right)\left(\overline{Y}\right)$$

$$= \frac{1}{n}\sum XY - \left(\overline{X}\right)\left(\overline{Y}\right)$$

$$= 0 - 0 = 0 \qquad \because \overline{X} = \overline{Y} = 0$$

$$\therefore r_{XY} = 0$$

i.e. X and Y are uncorrelated.

But if we examine the data carefully we find that X and Y are not independent but are connected by the relation $Y = X^2$. The above example illustrates that uncorrelated variables need not be independent.

Remark: One should not be confused with the words of uncorrelation and independence. $r_{XY} = 0$ i.e. uncorrelation between the variables X and Y simply implies the **absence of any linear (straight line) relationship** between them. They may, however be related in some other form (other than linear relationship like quadratic, logarithmic or exponential).

Note:

(i) $-1 \leq r \leq 1$

(ii) Interpretation of r:

(a) If $r = 1$ implies that there is perfect positive correlation between the related variables.

(b) If $r = -1$ implies that there is perfect negative correlation between the related variables.

(c) If $r = 0$, the variables are uncorrelated i.e. there is no linear relationship between the two variables.

(d) If $0 < r < 1$ implies that there is partial positive correlation between the related variables.

(e) If $-1 < r < 0$ implies that there is partial negative correlation between the related variables.

The correlation coefficient merely tells us that a linear relationship between two variables; it does not specify whether the relationship is cause and effect.

3. Spearman's rank correlation coefficient method.

Some characteristics like honesty, intelligence, beauty etc. are not measurable numerically (quantitatively). Hence Karl-Pearson's correlation coefficient methods fail for **finding the relationship between individual possessing qualitative characteristics.**

For example, we want to find whether intelligence (I.Q) and beauty are related or not. Both the characteristics are incapable of quantitative measurements but we can arrange a group of n individuals in order of merits (ranks) with respect to proficiency in two characteristics. Let the variables X and Y denote the ranks of the individuals in the characteristics A and B respectively. If we assume that there is no tie i.e. no two individuals get the same rank in a characteristic then obviously, X and Y assume numerical values ranging from 1 to n.

Sometimes, it is convenient to consider the ranks of observations rather than its actual value. **The numerical measure of linear relationship between such a variables (ranks) is rank correlation coefficient**, which can be calculated as follows.

The observations in X-series are given ranks starting with the highest observation i.e. the largest observation is given rank 1, smaller than rank 2 and so on. The ranks are denoted by R_X. Similarly, assign the ranks to the observations in y-series, denoted by R_Y.

Spearman's rank correlation coefficient is denoted by ρ (when ranks are not repeated) and is given by

$$\rho = 1 - \frac{6\sum_{i=1}^{n} di^2}{n\left(n^2 - 1\right)}$$

Where $di =$ R_X - R_Y = difference in ranks assigned to the same individuals.

Derivation:

In usual notation, prove that

$$\rho = 1 - \frac{6\sum_{i=1}^{n} di^2}{n\left(n^2 - 1\right)}$$

Proof: Let (Xi, Yi), i = 1, 2… n be the ranks of n individuals in the two characteristics A and B respectively. Where Xi's and Yi's are different permutations (arrangements) of n numbers from 1 to n. Assuming that no ranks are repeated.

$$\sum_{i=1}^{n} Xi = \sum_{i=1}^{n} Yi = 1+2+\ldots+n = \sum_{i=1}^{n} i = \frac{n(n+1)}{2}$$

$$\therefore \overline{X} = \overline{Y} = \frac{n+1}{2} \text{ ---------- (I)}$$

Also $S_X{}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(Xi-\overline{X}\right)^2 = \frac{1}{n}\sum_{i=1}^{n} Xi^2 - \left(\overline{X}\right)^2 = \frac{1}{n}\sum_{i=1}^{n} i^2 - \left(\frac{n+1}{2}\right)^2 = \frac{1}{n}\times\frac{n(n+1)(2n+1)}{6} - \left(\frac{n+1}{2}\right)^2$

$$= \frac{n^2-1}{12} = S_Y{}^2 \text{ ---------- (II)}$$

Let $di =$ difference in ranks assigned to the same individuals

$$= Xi - Yi = \left(Xi-\overline{X}\right) - \left(Yi-\overline{Y}\right) \quad \because \overline{X} = \overline{Y}$$

$$\therefore di^2 = \left[\left(Xi-\overline{X}\right) - \left(Yi-\overline{Y}\right)\right]^2 = \left(Xi-\overline{X}\right)^2 - 2\left(Xi-\overline{X}\right)\left(Yi-\overline{Y}\right) + \left(Yi-\overline{Y}\right)^2$$

Taking summation over i from 1 to n and dividing by n, we get

$$\frac{1}{n}\sum_{i=1}^{n} di^2 = \frac{1}{n}\sum_{i=1}^{n}\left(Xi-\overline{X}\right)^2 - \frac{2}{n}\sum_{i=1}^{n}\left(Xi-\overline{X}\right)\left(Yi-\overline{Y}\right) + \frac{1}{n}\sum_{i=1}^{n}\left(Yi-\overline{Y}\right)^2$$

$$= S_X{}^2 - 2Cov(X,Y) + S_Y{}^2$$

Dividing both the sides by $S_X S_Y$ i.e. $S_X{}^2 or S_Y{}^2$ we get

$$\frac{\sum_{i=1}^{n} di^2}{nS_X{}^2} = 1 - 2\rho + 1 \quad \because \rho = \frac{\mathrm{Cov(X,Y)}}{\mathrm{SxSy}}$$

$$= 2 - 2\rho = 2(1-\rho)$$

$$\therefore 1-\rho = \frac{\sum_{i=1}^{n} di^2}{2nS_X{}^2}$$

$$\therefore \rho = 1 - \frac{\sum_{i=1}^{n} di^2}{2n\left(\frac{n^2-1}{12}\right)} = 1 - \frac{6\sum_{i=1}^{n} di^2}{n(n^2-1)}$$

Limits of rank correlation coefficient:

$$-1 \le \rho \le 1$$

Proof: We have

$$\rho = 1 - \frac{6\sum_{i=1}^{n} di^2}{n(n^2-1)} = 1 - \text{[Some non-negative quantity]}$$

Since the square of a real quantity is always non-negative, $\sum_{i=1}^{n} di^2$ being the sum of non-negative quantities is also non-negative.

$\Rightarrow \rho \le 1$ ---------- (I)

The sign of equality holds if and only if $\sum_{i=1}^{n} di^2 = 0$. Now $\sum_{i=1}^{n} di^2 = 0$ iff each $di = 0$. i.e. the ranks of an individual are same for both characteristics. One such possibility is as given in the following table:

X	1	2	----------	n
Y	1	2	----------	n

On the other hand, ρ Will be minimum if $\sum_{i=1}^{n} di^2$ is maximum. i.e. each of the deviation (difference) di is maximum and is possible if the ranks of n individuals for two characteristics are in the opposite directions as given in the following table:

X	1	2	----------	n
Y	n	n-1	----------	1

From the above table

Xi + Yi = n + 1 $\quad \forall i = 1,2,...n$

$\therefore di = Xi - Yi = Xi - \{(n+1) - Xi\}$ $\quad \because$ Xi + Yi = n + 1 $\quad \therefore Yi = ((n+1) - Xi)$

$= 2Xi - (n+1) \quad \forall i = 1,2,...n$

$\therefore di^2 = [2Xi - (n+1)]^2 \ \forall i = 1,2,...n$

Taking summation over i from 1 to n, we get

$$\therefore \sum_{i=1}^{n} di^2 = \sum_{i=1}^{n} [2Xi - (n+1)^2] = \sum_{i=1}^{n} [4Xi^2 - 4(n+1)Xi + (n+1)^2] = 4\sum_{i=1}^{n} Xi^2 - 4(n+1)\sum_{i=1}^{n} Xi + n(n+1)^2$$

$$= \frac{4n(n+1)(2n+1)}{6} - \frac{4(n+1)n(n+1)}{2} + n(n+1)^2$$

$$= n(n+1)\frac{[4n+2-3n-3]}{3} = \frac{n(n+1)(n-1)}{3} = \frac{n(n^2-1)}{3}$$

$$\therefore \rho = 1 - \frac{6\sum_{i=1}^{n} di^2}{n(n^2-1)} = 1 - \frac{6\frac{n(n^2-1)}{3}}{n(n^2-1)} = 1 - 2 = -1$$

i.e. the minimum value of ρ is – 1. i.e. $-1 \le \rho$ ---------- (II)

$\therefore$ From (I) and (II)

$\boxed{-1 \le \rho \le 1}$.

Interpretation:

1. When $\rho = 1$, the relationship between the sets of ranks is perfect. i.e. ranking are identical
2. When $\rho = -1$, the relationship between the sets of ranks is reverse. i.e. ranking are opposite.
3. When $\rho = 0$, there is no relationship between two sets of ranks.

Adjustment of ranks in case of repeated ranks:

Many times, we have data in which one or more values of a variable are equal. In such a situations, the usual formula cannot be used, but the common practice is assign each value the average of ranks which they jointly occupy. e.g. two similar values are given the ranks 4 and 5, then we would assign each value the rank (4+5)/2 = 4.5.

When equal ranks are assigning to same values, some adjustment is to be made in computing the coefficient of rank correlation.

The adjustment consisting of adding $\frac{m(m^2-1)}{12}$ to the values of $\sum_{i=1}^{n} di^2$. The quantity is added as many times as the no. of values having equal ranks. The adjusted formula is

$$\rho = 1 - \frac{6\left[\sum_{i=1}^{n} di^2 + \frac{m(m^2-1)}{12} + \frac{m(m^2-1)}{12} + \ldots\ldots\right]}{n(n^2-1)}$$

Where m = no. of values having same ranks.

EXAMPLES:

1. The table below gives the mean temperature for 20 successive days and the average butterfat content in the milk.

Temperature	64	65	65	64	61	55	39	41	46	59
Butterfat (%)	4.64	4.58	4.67	4.60	4.83	4.55	5.14	4.71	4.69	4.65
Temperature	56	56	62	37	37	45	57	58	60	55
Butterfat (%)	4.65	4.36	4.82	4.65	4.66	4.95	4.60	4.68	4.60	4.46

Calculate r, the correlation coefficient. Comment on it. Determine the role of an independent variable in the relationship, if exists, between two variables.

2. From a large no. of families living in a rural area each having youngest child aged between 5 and 7 years, a researcher selected a sample of 6 families and measured the I.Q of the youngest child. The researcher also recorded the total no. of children in each family and drew up the following table.

No. of children in family	2	3	3	3	4	5
I.Q. of youngest child	110	100	100	80	80	70

Calculate Karl-Pearson's correlation coefficient and comment on it.

3. In one stage of the development of a new drug for an allergy, an experiment is conducted to study how different dosage of the drug affect the duration of relief from the allergic symptoms. Ten patients are included in the experiment. Each patient receives a specified dosage of the drug and asked to report back as soon as the protection of the drugs seems to wear off. The observations are recorded in the table, which shows the dosage and duration of relief for 10 patients.

Dosage(in mg)	3	3	4	5	6	6	7	8	8	9
Duration of relief (no .of days)	9	5	12	9	14	16	22	18	24	22

(i) State the objective of the study.

(ii) Calculate an appropriate measure to study the said objective. Comment on your findings.

4. The values of serum cholesterol (mmol/litre) (Y) and Body Mass Index (BMI) (X) for the 10 subjects who participated in a study to examine the effect of oat-bran cereal on Y are:

X	7.29	8.04	8.43	7.96	5.43	5.77	6.96	6.23	6.65	8.2	6.21
Y	29.0	26.3	21.6	21.8	27.2	24.8	25.2	24.5	25.1	27.9	24.8

Calculate the correlation coefficient between serum cholesterol and BMI. Comment on your findings.

5. Find correlation coefficient between RBC counts and WBC count of 7 persons.

RBC	46	54	56	56	58	60	62
WBC	36	40	44	54	42	58	54

6. Find correlation coefficient between Systolic B.P. and Diastolic B.P. of 8 persons.

Systolic B.P.	135	148	145	139	142	150	152	144
Diastolic B.P.	89	91	86	88	85	83	93	85

7. In a study on 10 patients with hypertriglyceridemia were placed on a low-fat, high-carbohydrate diet. Before the start of the diet, cholesterol and triglyceridemia measurements were recorded for each subject.
(i) Construct a two-way scatter plot. (ii) Does there appear to be any evidence of a linear relationship between cholesterol and triglyceride levels prior to the diet? (iii) Compute r, the Pearson correlation coefficient.

Patient	Cholesterol Level (mmol /l)	Triglyceridemia Level (mmol/l)
1	5.12	2.30
2	6.18	2.54
3	6.77	2.95
4	6.65	3.77
5	6.36	4.18
6	5.90	5.31
7	5.48	5.53
8	6.02	8.83
9	10.34	9.48
10	8.51	14.20

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

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

It was, however, discovered at the time of checking that two pairs of observations were not correctly copied. They were taken as (6, 14) and (8, 6) while the correct values were (8, 12) and (6, 8). Find the correct value of the correlation coefficient.

9. Family income and its percentage spent on food in the case of hundred families gave the following bivariate frequency distribution. Calculate the coefficient of correlation and comment on it.

Food Expenditure (%)	Family Income(Rs.)				
	2000 - 3000	3000 - 4000	4000 - 5000	5000 - 6000	6000 - 7000
10 – 15	-	-	-	3	7
15 – 20	-	4	9	4	3
20 – 25	7	6	12	5	-
25 - 30	3	10	19	8	-

10. Calculate Spearman's rank correlation coefficient between advertisement cost and sales from the following data:

Advertisement cost ('000 Rs.)	39	65	62	90	82	75	25	98	36	78
Sales ('00000 Rs.)	47	53	58	86	62	68	60	91	51	84

11. The coefficient of rank correlation of the marks obtained by 10 students in two particular subjects was found to be 0.5. It was later discovered that the difference in ranks in two subjects obtained by one of the student was wrongly taken as 3 instead of 7. What should be the correct value of coefficient of correlation?

12. The coefficient of rank correlation between marks in Statistics and Mathematics obtained by a certain group of students is 0.8. If the sum of squares of the differences in ranks is given to be 33, find the number of students in the group.

13. Ten competitors in a debate contest are ranked by three different judges in the following order.

1st Judge	1	5	4	8	9	6	10	7	3	2
2nd Judge	4	8	7	6	5	9	10	3	2	1
3rd Judge	6	7	8	1	5	10	9	2	3	4

Which pair of judges has the nearest approach to debate competition? Justify your answer by calculating suitable statistical measure?

14. The following are the marks obtained by a group of students in two subjects. Calculate rank correlation coefficient.

Economics	78	36	98	25	75	82	92	62	65	36
Statistics	84	51	91	69	68	62	86	68	35	49