



Module 1: Correlation

Dr. A. N. Basugade

M.Sc. Ph.D

Head, Department of Statistics,
GopalKrishnaGokhaleCollege, Kolhapur

Email – arunb1961@gmail.com

Module 1:CORRELATION

Overview


- Definition of Correlation
- Types of Correlation
- Methods of obtaining Correlation
- Scatter diagram Method
- Karl Pearson's Correlation Coefficient (r)
- Spearman's Rank Correlation Coefficient (R)

➤ **Definition of Correlation:**


Correlation is defined as “ Study of existence, direction and magnitude of the relationship between two or more than two variables. e.g. There is correlation between heights & weights, income & expenditure, rainfall & yield.

➤ **Types of Correlation:** There are two types of correlation

- i) Positive & Negative correlation
- ii) Linear & non-linear Correlation.

- 
- **Positive & Negative correlation:** If two variables changes in same direction, i.e. increase (decrease) in one variable causes the increase (decrease) in other variable then there is positive correlation between the two variables. e.g. there is positive correlation between sales & profit, income & expenditure, rain fall & yield, education & salary etc.

If two variables changes in opposite direction, i.e. increase (decrease) in one variable causes the decrease (increase) in other variable then there is negative correlation between the two variables. . e.g. there is negative correlation between pressure & volume, development in science & death rate, stress & job satisfaction etc.



➤ **Linear & non-linear correlation:** Here the points are plotted in the graph by taking 1st variable on X-axis and 2nd on Y-axis with proper scale.

If all the plotted points in the graph lies on a straight then there is linear correlation between the two variable. e.g. There is linear correlation between mileage travelled & fuel used, sales & profit.

If all the plotted points in the graph lies on a curve then there is non-linear correlation between the two variable. e.g . Ages & weights, rain fall & yield etc.



➤ **Methods of obtaining Correlation:**


There are three methods of obtaining correlation.


- i) Scatter diagram method
- ii) Karl Pearson's correlation coefficient (r)
- iii) Charl Spearman's rank Correlation Coefficient (R).

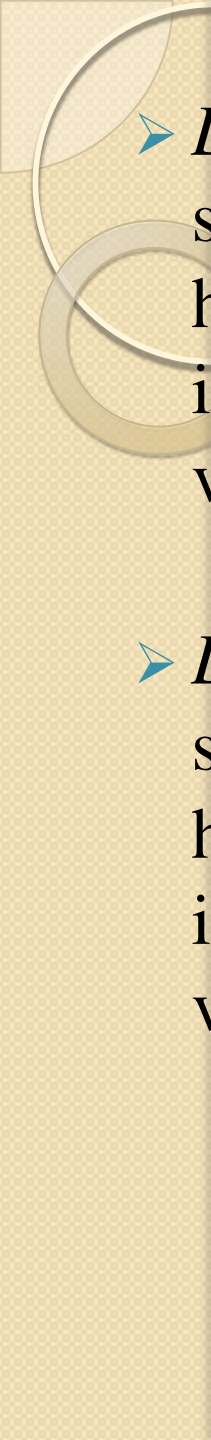
➤ Scatter Diagram Method:

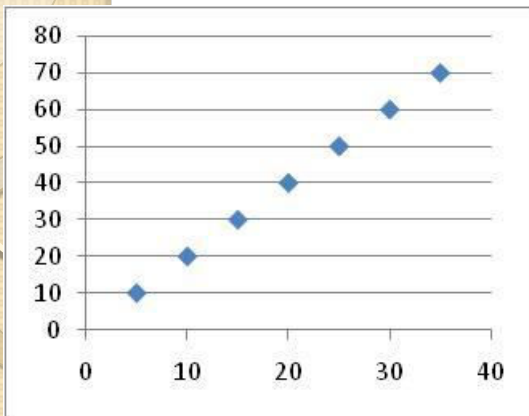
This is a graphical method of obtaining correlation. Here the points are plotted in the graph by taking 1st variable on X-axis and 2nd on Y-axis with proper scale. Generally these plotted points are scattered over the complete diagram. The manner in which these points are scattered in the scatter diagram gives the value of correlation.

- 1) *Perfectly positive correlation*: If all the points in the scatter diagram lies on a rising straight line from left hand bottom corner to right hand top corner then there is a perfectly positive correlation between the two variables and its value is +1 i.e. $r = 1$.

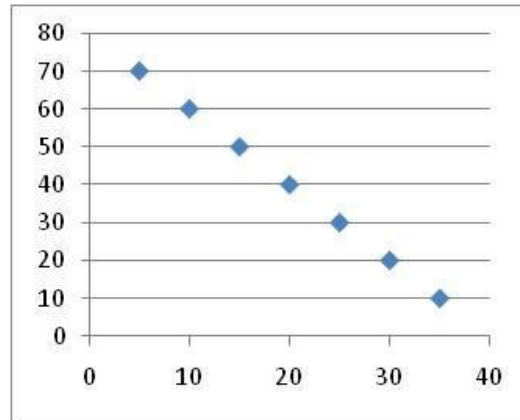
- 
- 1) *Perfectly negative correlation*: If all the points in the scatter diagram lie on a falling straight line from left hand top corner to right hand bottom corner then there is a perfectly negative correlation between the two variables and its value is -1 i.e. $r = -1$.
 - *Zero correlation*: If all the points in the scatter diagram are scattered over the complete diagram then there is zero correlation between the two variables and its value is 0 i.e. $r = 0$.

- 
- *High degree positive correlation:* If all the points in the scatter diagram lie on a rising narrow strip from left hand bottom corner to right hand top corner then there is a high degree positive correlation between the two variables and its value lies between 0 to +1.
 - *High degree negative correlation:* If all the points in the scatter diagram lie on a falling narrow strip from left hand top corner to right hand bottom corner then there is a high degree negative correlation between the two variables and its value lies between -1 to 0.

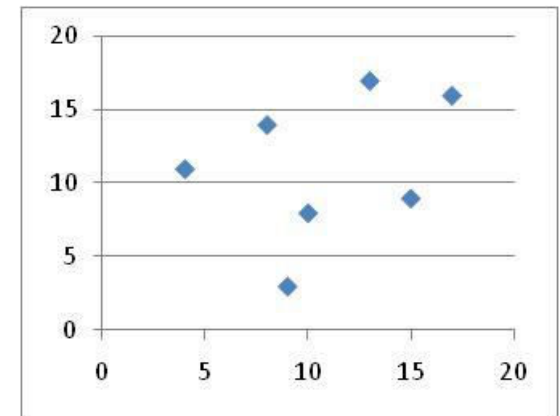
- 
- *Low degree positive correlation:* If all the points in the scatter diagram lies on a rising broad strip from left hand bottom corner to right hand top corner then there is a high degree positive correlation between the two variables and its value lies between 0 to +1.
 - *Low degree negative correlation:* If all the points in the scatter diagram lies on a falling broad strip from left hand top corner to right hand bottom corner then there is a high degree negative correlation between the two variables and its value lies between -1 to 0.



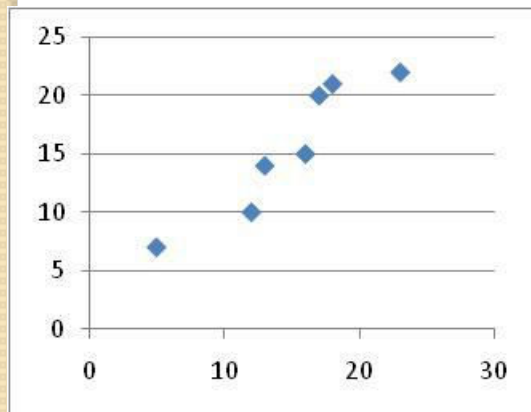
$$r = 1$$



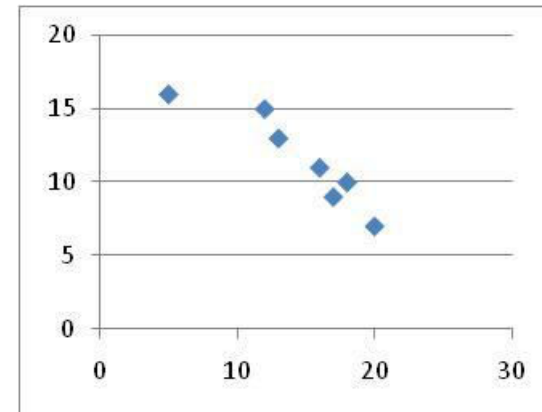
$$r = -1$$



$$r = 0$$



$$0 < r < 1$$



$$-1 < r < 0$$

➤ Karl Pearson's Correlation (r):

It is a mathematical method of obtaining correlation. It is defined as ratio of covariance between two variables X & Y to the product of standard deviation of X & standard deviation of Y.

$$r = \frac{\text{cov}(x, y)}{\sigma_x \sigma_y}$$

This correlation coefficient lies between -1 to +1.

➤ Interpretation:

1) $r = 1$ implies that there is perfectly positive correlation and one can estimate value of one variable knowing the value of other variable fairly accurate and shows closer relationship between two variables.

2) $r = -1$ implies that there is perfectly negative correlation and one can estimate value of one variable knowing the value of other variable fairly accurate and shows closer relationship between two variables.

3) $r = 0$ implies that there is no relationship between two variables.

➤ **Properties of Karl Pearson's Correlation Coefficient:**

- ❖ Correlation coefficient lies between -1 to +1.
- ❖ Correlation coefficient is independent change of origin & Scale.
- ❖ If $u = a x + b$ and $v = c y + d$ and if
 - i) signs of a & c are same then sign of $r(x, y)$ & $r(u, v)$ remains same.
 - ii) signs of a & c are different then sign of $r(x, y)$ & $r(u, v)$ is negative (minus).

➤ Spearman's Rank Correlation Coefficient (R):

When the data under study is of qualitative in nature then it is not possible to calculate Pearson's correlation coefficient r . In such cases we give the ranks to the variable values and calculate the correlation between the two variables. Charles Spearman has suggested the following formulae to obtain rank correlation coefficient & is denoted by R

1) When observations are Unrepeated: $R = 1 - \left[\frac{6 \sum d^2}{N^3 - N} \right]$

2) When the observations are Repeated:

$$R = 1 - \left[\frac{6 \left(\sum d^2 + \frac{m_1^3 - m_1}{12} + \frac{m_2^3 - m_2}{12} + \dots \right)}{N^3 - N} \right]$$

Where, d = Difference between ranks of x & ranks of y ,
 m_1 = no. of times first observation is repeated.
 m_2 = no. of times first observation is repeated.

➤ Derivation of Rank Correlation Coefficient when observations are unrepeatd:

Let x_i & y_i be the ranks of i^{th} member of two characteristics x & y , $i = 1, 2, \dots$. We assume that all n ranks of x as well as of y are distinct. Thus x & y will take values from 1 to n . Hence mean of $x = (1/n)(1+2+3+\dots+n) = (n+1)/2$, similarly mean of y is $(n+1)/2$ and

$$\sum (x_i - \bar{x})^2 = \sum x_i^2 - n\bar{x}^2 = (1^2 + 2^2 + \dots + n^2) - n\left(\frac{n+1}{2}\right)^2$$

$$\sum (x_i - \bar{x})^2 = (n^2 - n)/12 = \sum (y_i - \bar{y})^2$$

$$d_i = (x_i - y_i) = (x_i - \bar{x}) - (y_i - \bar{y}) = x_i' - y_i'$$

$$\sum d_i^2 = \sum x_i'^2 + \sum y_i'^2 - 2\sum x_i' y_i' = \frac{n^2 - n}{12} + \frac{n^2 - n}{12} - 2\sum x_i' y_i'$$

$$\sum x_i' y_i' = \frac{1}{2} \left[\frac{n^2 - n}{6} - \sum d_i^2 \right]$$

by definition of correlation coefficient $r = \frac{\sum x_i' y_i'}{\sqrt{\sum x_i'^2 \sum y_i'^2}} = \frac{(1/2) \left[\frac{n^2 - n}{6} - \sum d_i^2 \right]}{(1/12)(n^2 - n)}$

$$R = 1 - \frac{6\sum d_i^2}{(n^2 - n)}$$

Summary:

At the end of this module student must be able to

- Definition of Correlation
- Explain types of Correlation
- Explain Scatter diagram Method
- Define Karl Pearson's Correlation Coefficient (r)
- Derive the formula for Spearman's Rank Correlation Coefficient (R)