CHAPTER 7

CORRELATION

- **Correlation** is a statistical tool that measures the quantitative relationship between different variables. For example, there exists a positive correlation between the price and quantity supplied.

- **Correlation does not Imply Causation**
  - The correlation between the two variables does not imply that one variable causes the other.
  - Correlation only measures the degree and intensity of the relationship between the two variables, but surely not the cause and effect relationship between them.

- **Types of Correlation- On the Basis of Nature of Relationship**

  - ***Positive and Negative Correlation***

    Positive Correlation- When two variables move in the same direction, such a relationship is called positive linear correlation. In this case, the value of r_{xy} is positive. For example, price of a commodity and its supply.

    Negative Correlation- When the two variables move in different directions, such a relationship is called negative linear correlation. In this case, the value of r_{xy} is negative. For example, price of a commodity and its demand.

  - ***Linear and Curvilinear Correlation***

    Linear Correlation- If the ratio of change between the two variables is constant (or linear), then the two variables are said to be linearly correlated. Such type of correlation is depicted by a straight line.

    Curvilinear (or Non-linear Correlation) - If the ratio of change between the two variables is not constant, then the two variables are said to be non-linearly correlated. This can be shown by quadratic graph, parabola, hyperbola, etc.

  - **Simple and Multiple Correlation**

    Simple Correlation- The study of relationship only between two variables is known as simple correlation. For example, relationship between price and demand.

    Multiple Correlation- The study of relationship among three or more than three variables is called Multiple Correlation. For example, study of relationship between price, demand, tastes and income of the consumers.

- **Degrees of Correlation- Perfect Correlation and Zero Correlation**

  Perfect Correlation- When two variables change in the same proportion it is called perfect correlation. It is of two types:

1. When the proportional change in two variables is in the same direction, it is called **perfect positive correlation**. The coefficient of correlation is positive (+).

2. When the proportional change in the two variables is in opposite direction, then the correlation found is **perfect negative correlation**. The coefficient of correlation is negative (–).

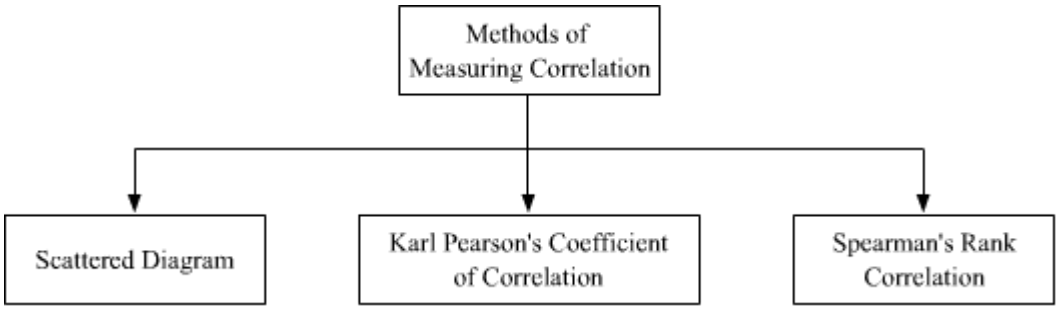
   The degree of correlation between 0 and 1 is a situation of **limited degree of correlation**.

Degrees of Correlation	Positive	Negative
Perfect Correlation	+ 1	– 1
Very High Degree	Between + 0.75 and + 1	Between – 0.75 and – 1
Moderate Degree	Between + 0.25 and + 0.75	Between – 0.25 and – 0.75
Low Degree	Between 0 and + 0.25	Between 0 and – 0.25
Zero	0	0

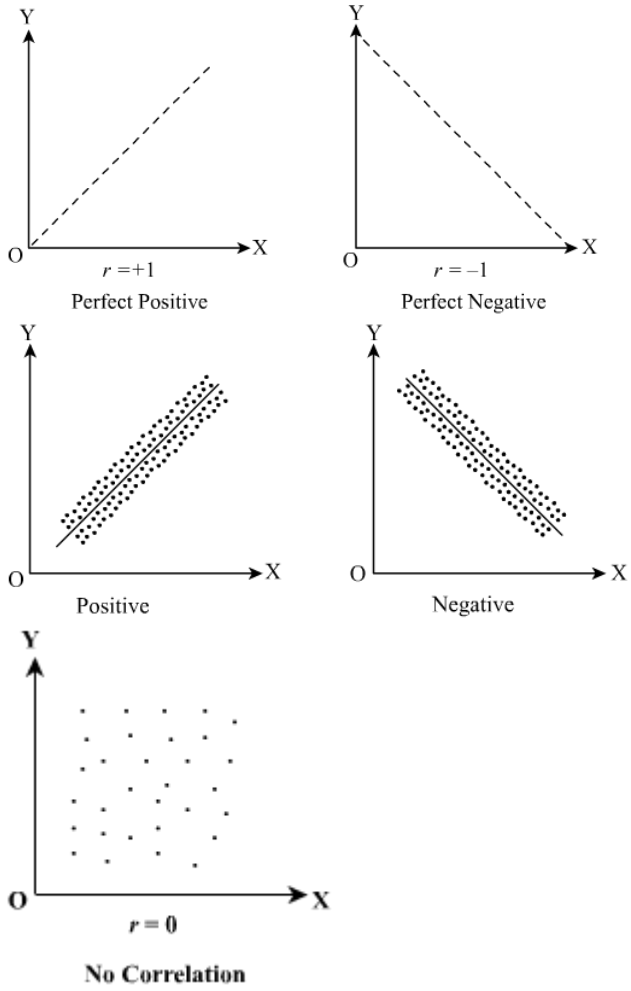
Zero Correlation- If there is no relation between two variables, i.e. change in one variable has no effect on the change in the other, then the variable lacks correlation.

NOTE: *A zero correlation between any two variables should not be mistakenly assumed as there is no relationship at all between them. In fact, it should be interpreted that the two variables are not linearly related, however, it may be possible that the two variables may be non-linearly related with each other.*

- **Methods of Measuring Correlation**
  - Scattered diagram
  - Karl Pearson's Coefficient of correlation
  - Spearman's Rank correlation coefficient



- **Scattered diagram** is a graphic means of measuring the direction, magnitude and degree of correlation.



- **Merits of Scatter Diagram**
  - It is easy to draw.
  - It does not include any tedious and difficult calculation process like Karl Pearson's method.
  - It is not affected by the presence of the extreme value in the series.
  - It reflects an unambiguous picture of proportionate change in *Y* values due to change in *X* values.

- **Demerits of Scatter Diagram**
  - It presents only a rough estimation of correlation between the variables. Hence, we cannot ascertain exact degree of correlation.
  - It helps us in knowing only the type of the correlation, i.e. whether, positive or negative. But fails to reveal anything about the magnitude and degree of correlation.
  - This method of correlation fails in case of ascertaining correlation between more than two variables.
  - It fails to reveal the direction of causation. That is, whether, *X* causes *Y* or *Y* causes *X* remains unknown.

Karl Pearson's Coefficient of Correlation (or Product Moment Correlation/ Simple Correlation Coefficient)

It measures the degree of linear relationship between two variables. It is denoted by '*r*'

- **Actual Mean Method**

$$r = \frac{\sum xy}{N\sigma_x\sigma_y}$$

OR

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

OR

$$r = \frac{\sum\left(X-\bar{X}\right)\left(Y-\bar{Y}\right)}{\sqrt{\sum\left(X-\bar{X}\right)^2}\sqrt{\sum\left(Y-\bar{Y}\right)^2}}$$

where,

$r =$ Coefficient of Correlation

$x = X - \bar{X}$

$y = Y - \bar{Y}$

$\sigma_x =$ Standard deviation of X series

$\sigma_y =$ Standard deviation of Y series

$N =$ Number of Observations

- **Direct Method**

$$r = \frac{\sum XY - N\left(\frac{\sum X}{N}\right)\times\left(\frac{\sum Y}{N}\right)}{\sqrt{\frac{\sum X^2}{N} - \left(\frac{\sum X}{N}\right)^2} \times \sqrt{\frac{\sum Y^2}{N} - \left(\frac{\sum Y}{N}\right)^2}}$$

OR

$$r = \frac{N\sum XY - \sum X\sum Y}{\sqrt{N\sum X^2 - \left(\sum X\right)^2} \times \sqrt{N\sum Y^2 - \left(\sum Y\right)^2}}$$

where,

$\sum XY =$ Sum of multiples of X and Y values

$\sum X^2 =$ Sum of squares of X values

$\sum Y^2 =$ Sum of squares of Y values

$N =$ Total numbers of observations

If the arithmetic means of X and Y are not in whole numbers, then the following formula should be used.

$$r = \frac{\sum XY - N\left(\overline{X}\right)\left(\overline{Y}\right)}{\sqrt{\sum X^2 - N\left(\sum \overline{X}\right)^2} \times \sqrt{\sum Y^2 - N\left(\sum \overline{Y}\right)^2}}$$

where,

$\sum XY = \text{Sum of multiples of } X \text{ and } Y \text{ values}$

$\sum X^2 = \text{Sum of squares of } X \text{ values}$

$\sum Y^2 = \text{Sum of squares of } Y \text{ values}$

$\overline{X} = \frac{\sum X}{N}$ and $\overline{Y} = \frac{\sum Y}{N}$

$N = \text{Total numbers of observations}$

- **Short-cut Method/Assumed Mean Method**

This method is applied when:

1. the actual mean value is not in whole number (but in fraction).
2. the series is large.

$$r = \frac{\sum d_x d_y - \frac{\left(\sum d_x\right) \times \left(\sum d_y\right)}{N}}{\sqrt{\sum {d_x}^2 - \frac{\left(\sum d_x\right)^2}{N}} \times \sqrt{\sum {d_y}^2 - \frac{\left(\sum d_y\right)^2}{N}}}$$

where,

d_x = Deviation of X series from assumed mean

d_y = Deviation of Y series from assumed mean

$\sum d_x d_y = \text{Sum of multiples of } d_x \text{ and } d_y$

$\sum {d_x}^2 = \text{Sum of squares of } d_x$

$\sum {d_y}^2 = \text{Sum of squares of } d_y$

$\sum d_x = \text{Sum of deviations of } X \text{ series}$

$\sum d_y = \text{Sum of deviations of } Y \text{ series}$

$N =$ Total numbers of observations

- **Step-Deviation Method**

This method can be used instead of short-cut method to avoid time-consuming calculations.

$$r = \frac{\sum d'_x d'_y - \frac{\left(\sum d'_x\right) \times \left(\sum d'_y\right)}{N}}{\sqrt{\sum {d'_x}^2 - \frac{\left(\sum d'_x\right)^2}{N}} \times \sqrt{\sum {d'_y}^2 - \frac{\left(\sum d'_y\right)^2}{N}}}$$

where,

$d'_x = \frac{d_x}{h}$ and $d'_y = \frac{d_y}{i}$

$h =$ common factor for X series

$i =$ common factor for Y series

d_x = Deviation of X series from assumed mean

d_y = Deviation of Y series from assumed mean

$\sum d'_x d'_y$ = Sum of multiples of d'_x and d'_y

$\sum d'^2_x$ = Sum of squares of d'_x

$\sum d'^2_y$ = Sum of squares of d'_y

$\sum d_x$ = Sum of deviations of X series

$\sum d_y$ = Sum of deviations of Y series

$N =$ Total numbers of observations

Properties of Correlation Coefficient (*r*)

- It is free from any units, i.e. it is a pure number.
- A negative *r* indicates inverse relation, and positive *r* a positive relation
- If *r* is zero, then it implies that there is no linear relation between the two variables.
- The value of *r* lies between –1 and +1 i.e. $-1 \leq r \leq +1$.
- If $r = +1$ (or –1), then it indicates perfect positive (or negative) correlation.
- The value of *r* is independent of origin and any change of the scale of the graph.

Merits of Karl Pearson's Correlation Coefficient

- It is the most common and an ideal method of calculating correlation.
- The value of the correlation coefficient helps in assessing the type and magnitude of the linear relationship between the two variables.
- It helps in measuring the exact correlation between the two variables.

Demerits of Karl Pearson's Correlation Coefficient

- It is affected by the presence of extreme items.
- It involves a tedious and time-consuming calculation process.
- It only studies the linear relationship between the two variables and fails to study non-linear relationship such as, quadratic relations, etc.

Spearman's Rank Correlation Coefficient

- Named after British psychologist C.E. Spearman.
- This method is most suitable when the variable cannot be measured quantitatively.
- It is used to calculate the correlation between two qualitative variables, such as, beauty, honesty, etc.
- It indicates the correlation between ranks.

Merits of Spearman's Rank Correlation Coefficient

- The calculation of Rank Correlation is easier than the Pearson's Correlation.
- It can be used for qualitative variables such as, honesty, beauty, intelligence, etc.
- The value of the correlation coefficient helps us in assessing the type and magnitude of the linear relationship between the two variables.
- This method can also be used even when only ranks are given and not the actual values of the observations.

Demerits of Spearman's Rank Correlation Coefficient

- This method cannot be used in case of open-ended series, i.e. continuous series. That is, this method only works out for discrete and individual series.
- It cannot be used for large number of observations. It works if the number of observation is less than 30.
- Compared to Karl Pearson's method, the Rank Correlation method lacks precision. This is because, it does not use all the information (i.e. the actual observations) rather uses only ranks.

Formula to Calculate Spearman's Rank Correlation Coefficient

- ***Case1: When ranks <u>are given</u> and when ranks <u>are not given</u>***

$$r_k = \frac{1-6\sum D^2}{N^3 - N}$$

where,

r_k = Coefficient of Rank Correlation

D = Rank Differences

N = Number of Observations

- ***Case2: When the ranks are given and are <u>equal or repeated</u>***

$$r_k = \frac{1-6\left[\sum D^2 + \frac{1}{12}\left(M_1^3 - M_1\right) + \frac{1}{12}\left(M_2^3 - M_2\right) + \ldots\ldots\ldots\ldots\right]}{N^3 - N}$$

where,

m = Number of items of equal ranks

$D = R_1 - R_2$

D^2 = Sum of squares of difference of Rank1 and Rank2

N = Number of observations

Similarities and Dissimilarities Between Karl Pearson's Correlation Coefficient (r_{xy}) and Spearman's Rank Correlation Coefficient (r_k)

- Generally, all the properties of Karl Pearson's coefficient of correlation are similar to that of the rank correlation coefficient.
- Rank correlation coefficient is generally lower or equal to Karl Pearson's coefficient.
- Rank correlation coefficient is usually preferred to measure the correlation between the two qualitative variables.
- The difference between the two coefficients is because the rank correlation coefficient uses ranks whereas the Karl Pearson's coefficient uses full set of observations.
- If the precisely measured data are available, then both the coefficients will be identical.
- If extreme values are present in the data, then the rank correlation coefficient is more precise and reliable and consequently its value differs from that of the Karl Pearson's coefficient.
- When the values are not repeated, in this case, the value of Karl Pearson's correlation coefficient will be same as that of the Spearman's Rank correlation coefficient.