Introduction to Correlation Analysis

There are many situations where the objective of the study is to see the relationship among variables. Consider the following situations-

- Is there any relation between shoe size and height?
- Is there any relation between number of hours you exercise every day and your heart rate?
- Is there any relation between math scores and overall scores in exam?
- Is there any relation between temperature and earthquake?
- Is there any relation between temperature in summer and water consumption?

Correlation was first developed by **Sir Francis Galton** (1822-1911) and then reformulated by **Karl Pearson** (1857-1936).

Correlation analysis enables us to measure or quantify the relationship between the variables.

In correlation analysis, the statistical problem is to find out if there is a correspondence of variation (or movement) between two or more variables. It is concerned with two (or more) sets of data, not with only one set. When we measure interrelationship between two variables, then it is called simple correlation and relationship among more than two variables is termed as multiple correlation.

Definition of Correlation Analysis:

The correlation is the measure of the extent and the direction of the relationship between two variables in a bivariate distribution.

Correlation analysis involves various methods and techniques used for studying and measuring the extent of the relationship between the two variables. It also helps us to decide the strength of the linear relationship or association between two variables.

Advantages of Correlation Analysis:

1. Correlation analysis measures the strength of relationship between the variables.
2. It also provides the direction of the relationship between the variables.
3. It reduces the range of uncertainty in matter of prediction. Prediction is not perfectly accurate here, but helpful.

Limitations of Correlation Analysis:

- Correlation only describes a relationship. It does not tell us why.
- Correlation is affected greatly by the range of the data set.
- If there is outlier in the dataset, it significantly alters the correlation. It can result in apparent statistical evidence that a relationship exists when in fact there is no relation.
- Correlation does not capture strong non-linear relationship between variables.

Types of Correlation:

- **Positive correlation** – Variables change in the same direction.
  - As X is increasing, Y is also increasing
  - As X is decreasing, Y is also decreasing
- **Negative correlation** – Variables change in the opposite directions.
  - As X is increasing, Y is decreasing
  - As X is decreasing, Y is increasing

Examples:

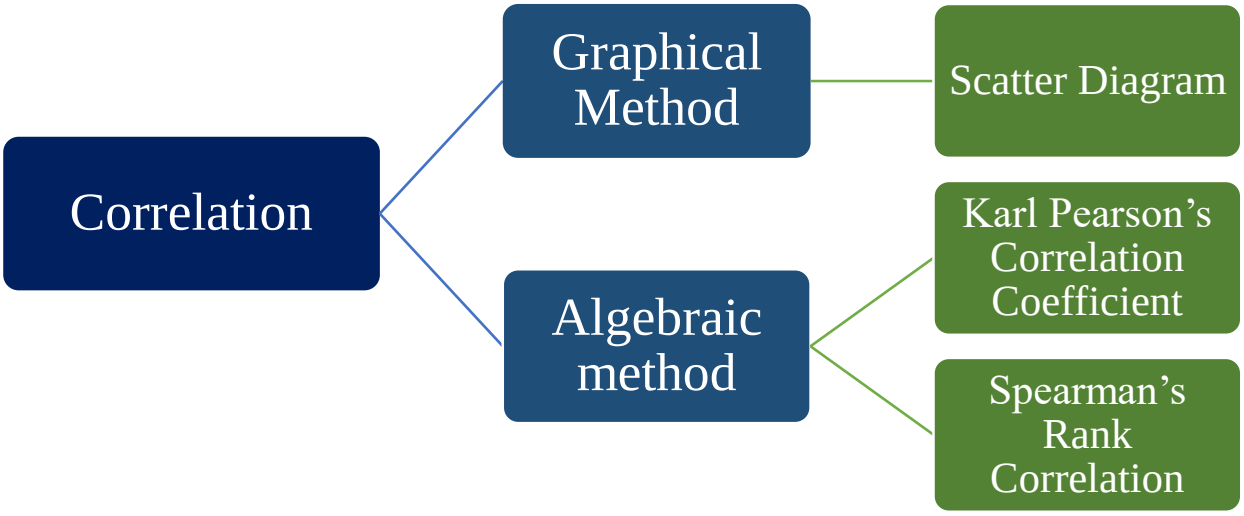
Positive Correlation

- Income and expenditure on luxury goods
- A student's study hour and his test score
- Advertising and sales
- Water consumption and temperature

Negative Correlation

- Price and demand of a commodity
- Students absence in class and their grades
- TV registration and cinema attendance
- Alcohol consumption and driving ability

Methods in Studying/Determining Correlation:



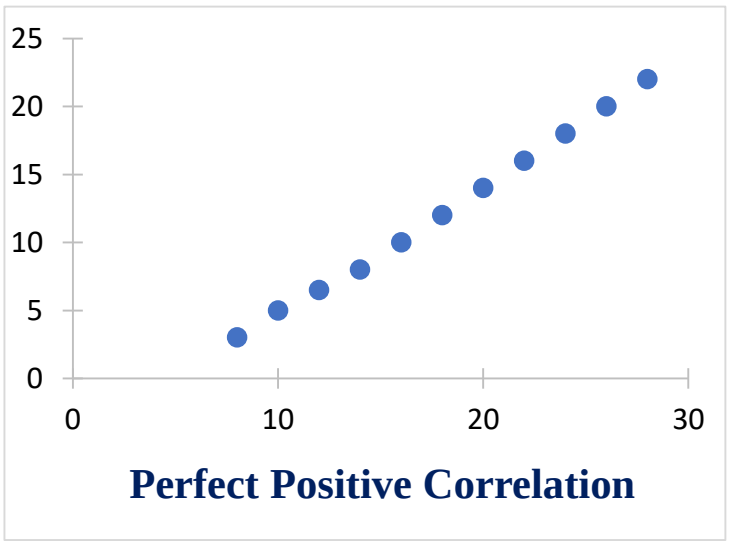
Scatter Diagram Method:

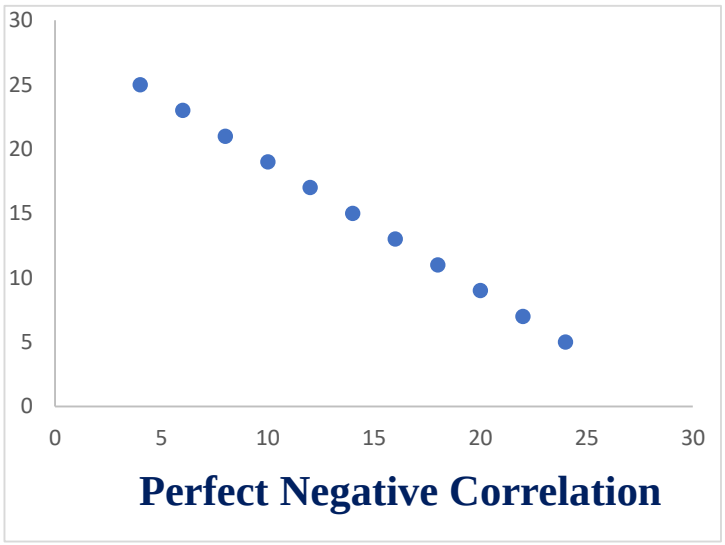
It is the simplest device for ascertaining if the two variables are related or not. This is a graphical method of finding out relationship between the variables.

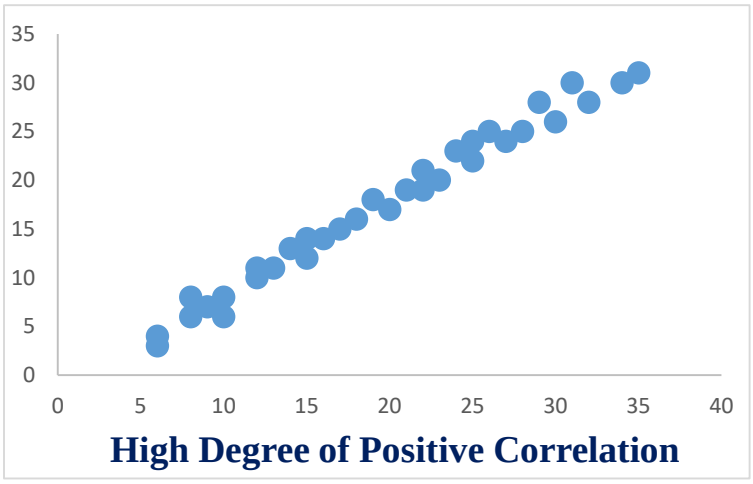
When this method is used the given data are plotted on a graph paper in the form of dots i.e. for each pair of x and y values we put a dot and thus obtain as many points as the number of observations.

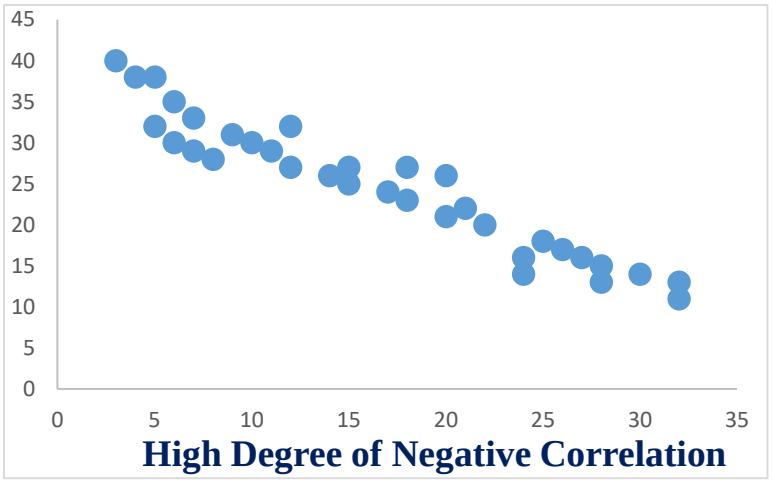
The greater the scatter of points over the graph, the lesser is the relationship between the variables.

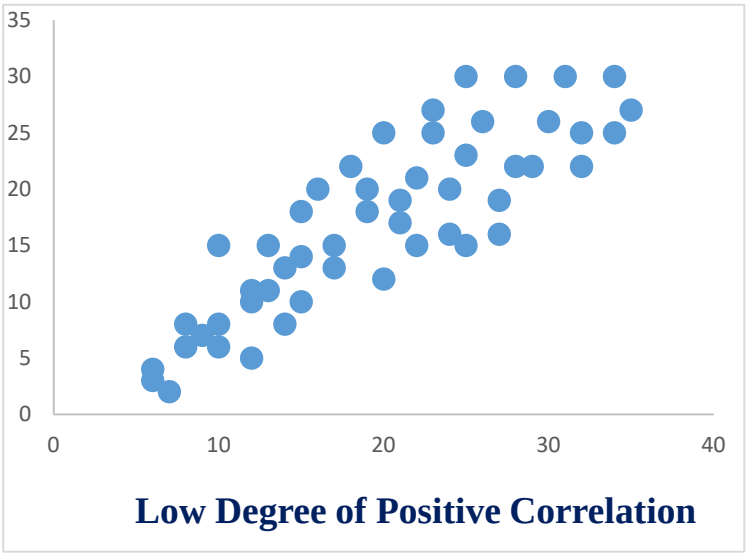
Several Scatter Diagrams:



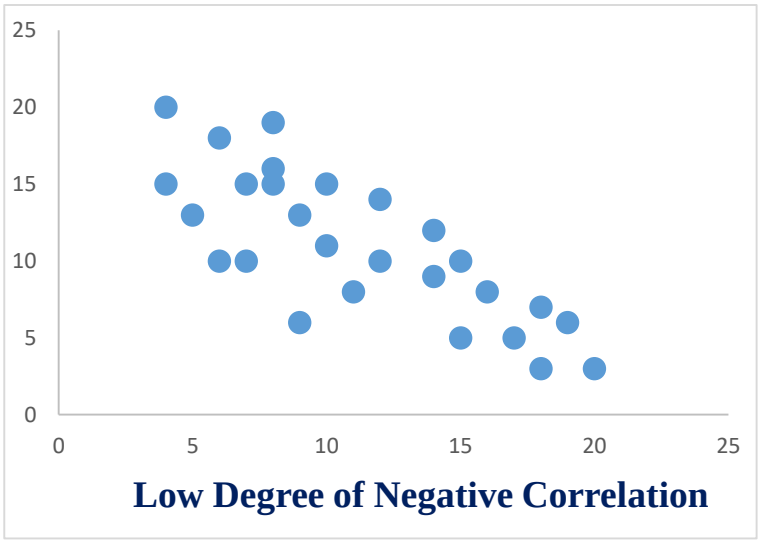




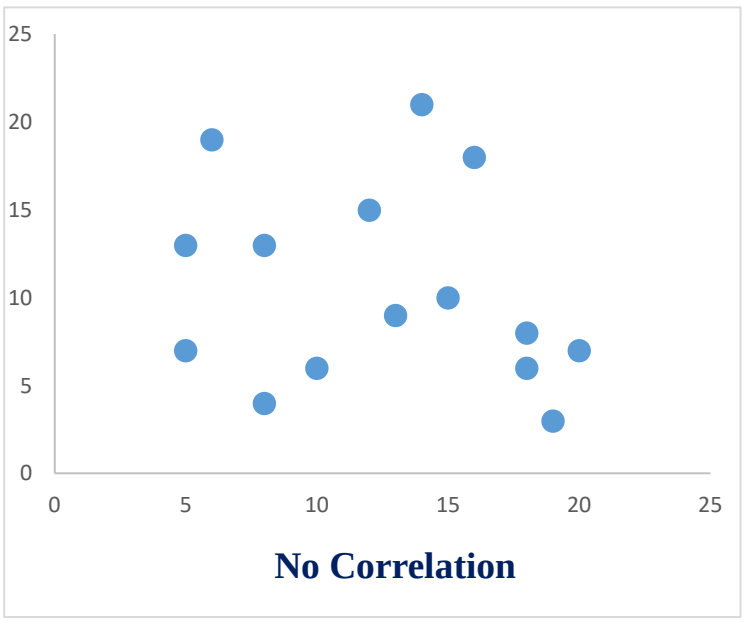




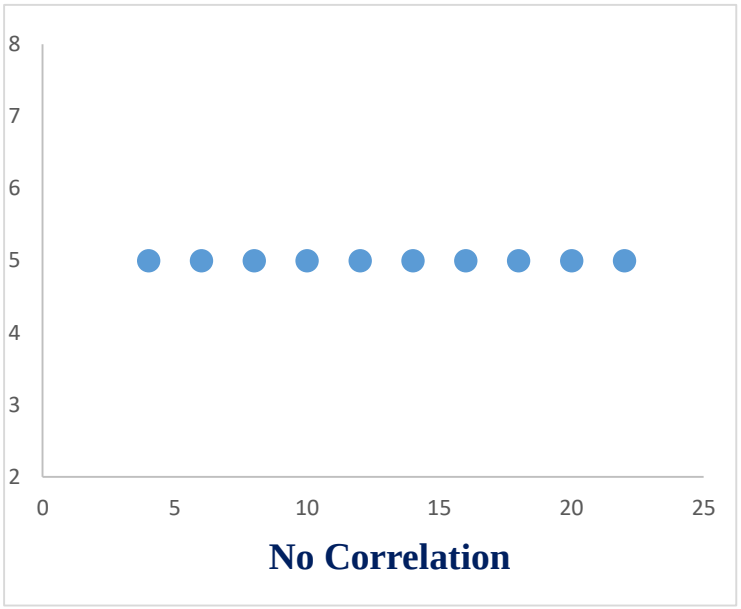
Low Degree of Positive Correlation



Low Degree of Negative Correlation



No Correlation



No Correlation

Merits and Demerits of Scatter Diagram:

Merits

- Simple and non-mathematical method.
- Easily understandable.
- Not influenced by extreme values.

Demerits

- Can't establish the exact degree of correlation between the variables as is possible by applying the mathematical methods.

Karl Pearson's Correlation Coefficient:

It is also known as Pearson's product-moment correlation coefficient. It was introduced by Galton in 1877 and developed later by Karl Pearson in 1895.

The coefficient of correlation is a number which indicates to what extent two variables are related, to what extent variation in one go with the variation in the other.

This is the most popular measure of correlation for measuring the degree or strength of linear relationship between two variables.

Pearson's correlation coefficient between two variables is defined as the covariance of the two variables divided by the product of their standard deviations.

Assumptions of Pearson's Correlation Coefficient:

- The underlying relationship between two variables is linear, which is the assumption of linearity.
- The variables under consideration should have a bivariate normal distribution.

Mathematical Formula:

The Pearson's correlation coefficient is denoted by r or r_{xy}

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}} = \frac{Cov(x,y)}{S_x S_y}$$

Or the algebraic equivalent

$$r = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{x}\,\overline{y}}{\sqrt{(\sum_{i=1}^{n} {x_i}^2 - n\overline{x}^2)(\sum_{i=1}^{n} {y_i}^2 - n\overline{y}^2)}}$$

Properties of Pearson's Correlation Coefficient (r):

- It lies between -1 and 1.
- It is symmetrical with respect to x and y, i.e. $r_{xy} = r_{yx}$
- It is just a pure number and independent from the unit of measurement of x and y.

Interpretation of Pearson's Correlation Coefficient (r):

- If $r = +1\ or\ -1$, the sample points lie on a straight line.
- If $r = +1$, it means there is perfect positive correlation between the variables.
- If $r = -1$, it means there is perfect negative correlation between the variables.
- If $r = 0$, it means there is no correlation between the variables.
- If r is near to $+1$ or -1, there is a strong linear association between the variables.
- If r is small (close to 0), there is low degree of correlation between variables.

Degrees of Correlation:

High degree, moderate degree or low degree are the three categories of correlation. The following table reveals the effect of coefficient or correlation:

Degrees	Positive	Negative
Absence of correlation	0	0
Perfect Correlation	+1	-1
High Degree	+0.75 to +1	-0.75 to -1
Moderate Degree	+0.25 to +0.75	-0.25 to -0.75
Low degree	0 to +0.25	0 to -0.25

Coefficient of Determination:

The coefficient of determination is the square of the coefficient of correlation (r^2).

It is the measure of the strength of the relationship between two variables. It shows the percentage of variation of one variable which can be described by another variable and it is a measure for the goodness of fit for lines passing through plotted points.

The value of determination coefficient lies between 0 and 1.

Suppose $r = 0.9$, then $r^2 = 0.81$. This would mean that 81% of total variation in y can be explained by x.

Merits and Demerits of r:

Merits:

- It summarizes in one value the degree of correlation and direction of correlation also.
- It gives us exact measure of degree of association.
- It is independent of units of the variables.

Demerits:

- Value of "*r*" is affected by extreme values.
- Calculation process of this method is long, tedious and time consuming.
- Pearson correlation coefficient is computed only for those attributes which have quantitative measurements.
- When the population is not normally distributed or when the shape of the population is not known, then *r* may underestimate the relationship between the variables.

Example:

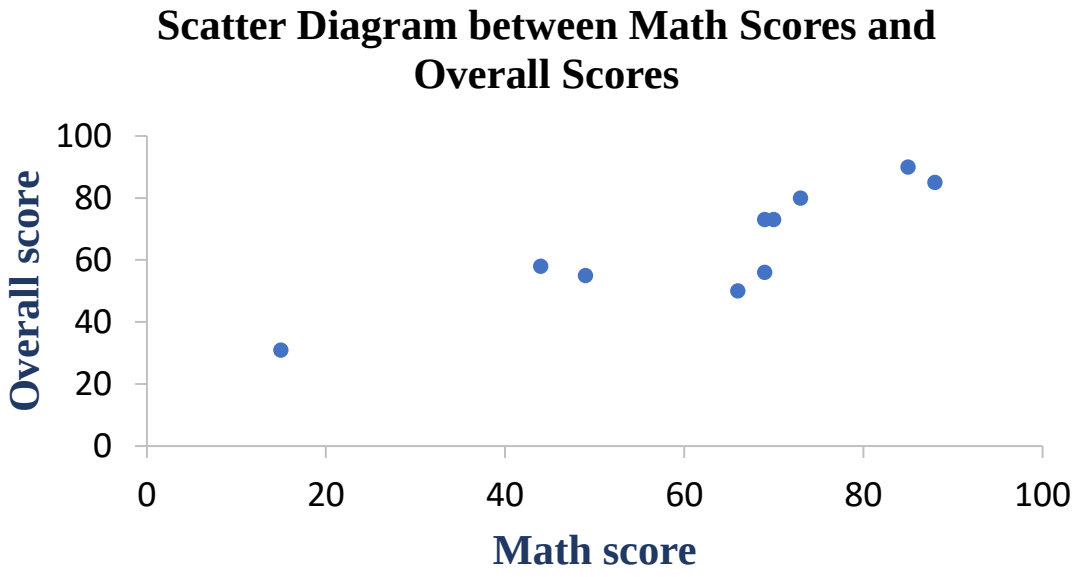
A small experiment is conducted involving 10 students to investigate the association between students' math scores and overall scores in exam.

Math Scores (%)	70	85	66	15	69	49	73	44	88	69
Overall scores (%)	73	90	50	31	56	55	80	58	85	73

Our aim is to find out whether there is any linear association between math score and overall score.

Let us denote x = Math Scores and y = Overall Scores

The correlation between x and y can be shown in a scatter Diagram.



Comment: This graph shows positive correlation between math scores and overall scores.

Calculation of *"r"*:

x	y	xy	x^2	y^2
70	73	5110	4900	5329
85	90	7650	7225	8100
66	50	3300	4356	2500
15	31	465	225	961
69	56	3864	4761	3136
49	55	2695	2401	3025
73	80	5840	5329	6400
44	58	2552	1936	3364
88	85	7480	7744	7225
69	73	5037	4761	5329
$\sum x = 628$	$\sum y = 651$	$\sum xy = 43993$	$\sum x^2 = 43638$	$\sum y^2 = 45369$

$$r = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{x}\,\overline{y}}{\sqrt{(\sum_{i=1}^{n} x_i{}^2 - n\overline{x}^2)(\sum_{i=1}^{n} y_i{}^2 - n\overline{y}^2)}}$$

Here,

$$\overline{x} = \frac{\sum x_i}{n} = \frac{628}{10} = 62.8$$

$$\overline{y} = \frac{\sum y_i}{n} = \frac{651}{10} = 65.1$$

So, $r = 0.878$

Interpretation

- $\boldsymbol{r = 0.878}$

Interpretation: There exists high degree of positive correlation between math scores (x) and overall scores (y) of students.

- $\boldsymbol{r^2 = 0.771}$

Interpretation: 77.1% of total variation in overall score (y) can be explained by the linear relationship between math scores (x) and overall scores (y). The other 22.9% of the total variation in overall score (y) remains unexplained.

Practice Problems:

Textbook: Probability and Statistics for Engineering and the Sciences (Devore)

CHAPTER 12 Simple Linear Regression and Correlation

Page 516: 59(a, d).

Textbook: Statistical Techniques in Business & Economics (LIND MARCHAL WATHEN)

Chapter 13: CORRELATION AND LINEAR REGRESSION

Page 446-447: 3, 4, 5, 6