STATISTICAL INFERENCE.

May be divided into two major areas:

(a) Estimation.

(b) Test of Hypotheses.

Types of estimates

(a) A point estimate is a single number that is used to estimate an unknown population parameter.

(b) An interval estimate is a range of values used to estimate a population parameter.

Statistical Inference

The process of drawing conclusions about population parameters based on a sample taken from the population.

It refers to the process of selecting and using a sample statistic to draw inference about a population parameter based on a subset of it - the sample drawn from the population.

Statistical inference treats 2 different classes of problems:

1. Hypothesis Testing: That is to test some hypothesis about parent population from which the sample is drawn.

2. Estimation: That is to use the "statistics" obtained from the sample as estimate of the unknown 'parameter' of the population from which the sample is drawn.

$H_0 \Rightarrow$ currently accepted Hypothesis.

Ha (Alternative Hypothesis / Research Hypothesis) : Involves the claim to be tested.

Q: It is believed that a candy machine makes chocolate bars that are on average 5g. A worker claims that the machine after maintenance no longer make 5g bars. Write H_0 and H_a.

H_0 & H_a are mathematically opposite.

Outcomes of this Test

- Reject Null Hypothesis
- fail to Reject Null Hypothesis.

Test Statistic - ~~calculate~~ calculate from sample Data

- Sample 50 Bars.
- Gert Avg.
- calculate Test statistic

Satistically significant.

Level of ~~significance~~ / confidence (c)

How confident are we in our decision.

C → 95%, 99%.

Level of significance

$\alpha = 1 - C$

$C = 0.95$

$\alpha = 0.05$

Estimation

When data are collected by sampling from a population, the most important objective of statistical analysis is to draw inferences or generalisations about that population from the information embodied in the sample. Estimation is concerned with the methods by which population characteristics are estimated from sample information.

The following are two types of estimates

1. Point Estimate.
2. Interval Estimate.

Point Estimate: A point Estimate is a single number which is used as an estimate of the unknown population parameter.

Interval Estimate: An interval estimate of a population parameter is a statement between which it is estimated that the parameter lies. An interval estimate would always be specified by two values that is the lower one and the upper one.

Eg: On the basis of sample study, if we estimate the average income off the people living in a village as Rs 875 it will be a point estimate. On the other hand, if we say that the average income would lie between Rs. 800 and Rs. 950, it will be an interval estimate.

HYPOTHESIS TESTING

Hypothesis testing begins with an assumption, called a Hypothesis, that we make about a population parameter. A Hypothesis is a supposition made as a basis of for reasoning.

Procedure of Testing Hypothesis

Procedure of testing hypothesis is briefly described below:

1. Set up a Hypothesis : The first thing in hypothesis testing is to set up a hypothesis about a population parameter. Then we collect sample data, produce

sample statistics, and use this information to decide how likely it is that our hypothesized population parameter is correct.

The two hypotheses in a statistical test are :-

(a) Null Hypothesis

(b) Alternative Hypothesis

Null Hypothesis is a very useful tool in testing the significance of difference.

~~It states~~ It asserts that there is no real difference in the sample and the population in the particular matter under consideration. Eg: if we want to find out whether extra coaching has benefited the students or not, we set up a null hypothesis that "extra coaching has not benefited the students".

As against the null hypothesis, the alternative hypothesis specifies those values tha the researcher believes to hold true & he hopes that the sample data lead to acceptance of this hypothesis is true.

EXAMPLE :

A psychologist who wishes to test whether or not a certain class of people have a mean I.Q higher than 100.

$H_o : \mu = 100$ (null Hypothesis)

$H_a : \mu \neq 100$ (alternative Hypothesis).

2. Set up a suitable significance level:

The next step is to test the validity of Ho against that of Ha at a certain level of significance. The significance level is the probability of rejecting the null Hypothesis if it is true. It is expressed as a percentage. ~~When the hypothesis in question is accepted at the 5-percent level, the statistician is running~~ the that, in the long run, he will be making the wrong decision about 5 percent of the time.

When the hypothesis is rejected at the 5 percent level, the statistician is running the risk of rejecting a true hypothesis in 5 out of every 100 occassions.

3. <u>Setting a test criterion</u>: This involves selecting an appropriate probability distribution for the particular test.

4. Doing computations :- These calculations include the testing statistic and the standard error of the testing statistic.

5. <u>Making Decisions</u> :- A statistical decision or conclusion is a decision either to reject or to accept the null Hypothesis.

Two types of Error.

Type I error : Type I error is committed by rejecting the null Hypothesis when it is True. It is denoted by α.

Type II error : It is committed by not rejecting (that is accepting) the null Hypothesis when it is false. It is denoted by β.

PARAMETER : A number that describes the data from a population.

STATISTIC : A ~~d~~ number that describes the data from a sample.

ST

the plotted points lie on a straight line parallel to the X-axis or in a haphazard manner, it shows absence of any relationship between the variables (*i.e.*, $r = 0$) as shown in diagram VII.

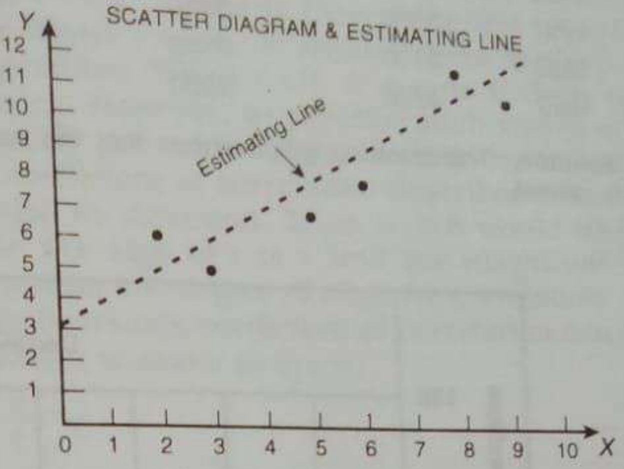
Illustration 1. Given the following pairs of values of the variables X and Y :

X:	2	3	5	6	8	9
Y:	6	5	7	8	12	11

(*a*) Make a scatter diagram.

(*b*) Is there any correlation between the variables X and Y?

(*c*) By graphic inspection, draw an estimating line*.



Solution. By looking at the scatter diagram we can say that the variables X and Y are correlated. Further, correlation is positive because the trend of the points is upward rising from the lower left-hand corner to the upper right-hand corner of the diagram. The diagram also indicates that the degree of relationship is higher because the plotted points are near to the line which shows perfect relationship between the variables.

Merits and Limitations of the Method

Merits Following are the merits of scatter diagram method :

- It is a simple and non-mathematical method of studying correlation between the variables. As such it can be easily understood and a rough idea can very quickly be formed as to whether or not the variables are related.
- It is not influenced by the size of extreme items whereas most of the mathematical methods of finding correlation are influenced by extreme items.
- Making a scatter diagram usually is the first step in investigating the relationship between two variables.

Limitations By applying this method we can get an idea about the direction of correlation and also whether it is high or low. But we cannot establish the exact degree of correlation between the variables as is possible by applying the mathematical methods.

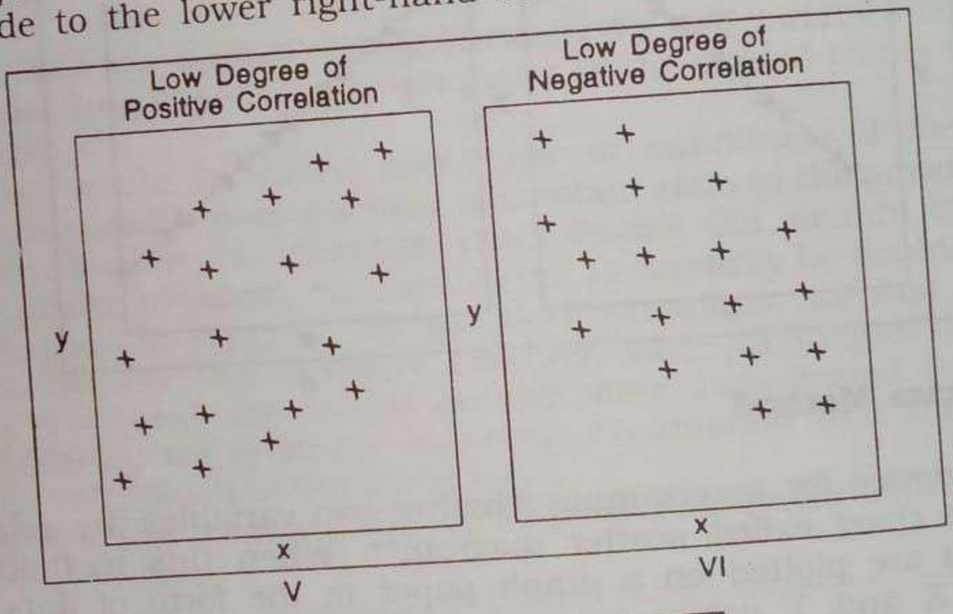
GRAPHIC METHOD

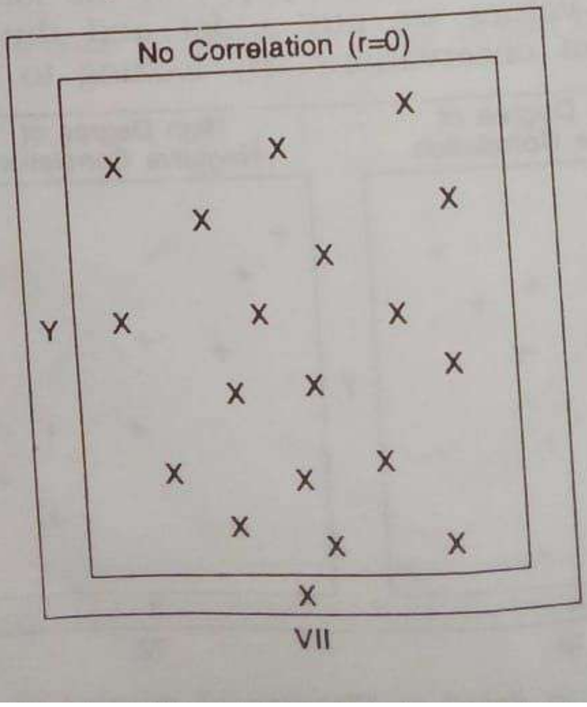
When this method is used the individual values of the two variables are plotted on the graph paper. We thus obtain two curves, one for X variable and another for Y variable. By examining the direction and closeness of the two curves so drawn we can infer whether or not the variables are related. If both the curves drawn on the graph are moving in the same direction (either upward or downward) correlation is said to be positive. On the other hand, if the curves are moving in the opposite directions correlation is said to be negative. The following example shall illustrate the method.

Illustration 2. From the following data ascertain whether the income and expenditure of the 100 workers of a factory are correlated :

* An estimating line or regression line is a line of average relationship. For details please see next chapter on 'Regression Analysis'.

various points we can form an idea as to whether the variables are related or not. The greater the scatter of the plotted points on the chart, the lesser is the relationship between the two variables. The more closely the points come to a straight line falling from the lower left-hand corner to the upper right-hand corner, correlation is said to be perfectly positive (*i.e.*, $r = +1$) (diagram I). On the other hand, if all the points are lying on a straight line rising from the upper left-hand corner to the lower right-hand corner of the diagram, correlation is said to be perfectly negative (*i.e.*, $r = -1$) (diagram II). If the plotted points fall in a narrow band there would be a high degree of correlation between the variables— correlation shall be positive, if the points show a rising tendency from the lower left-hand corner to the upper right-hand corner (diagram III) and negative if the points show a declining tendency from the upper left-hand corner to the lower right-hand corner of the diagram (diagram IV). On the other hand, if the points are widely scattered over the diagram it indicates very little relationship between the variables—correlation shall be positive if the points are rising from the lower left-hand corner to the upper right-hand corner (diagram V) and negative if the points are running from the upper left-hand side to the lower right-hand side of the diagram (diagram VI). If



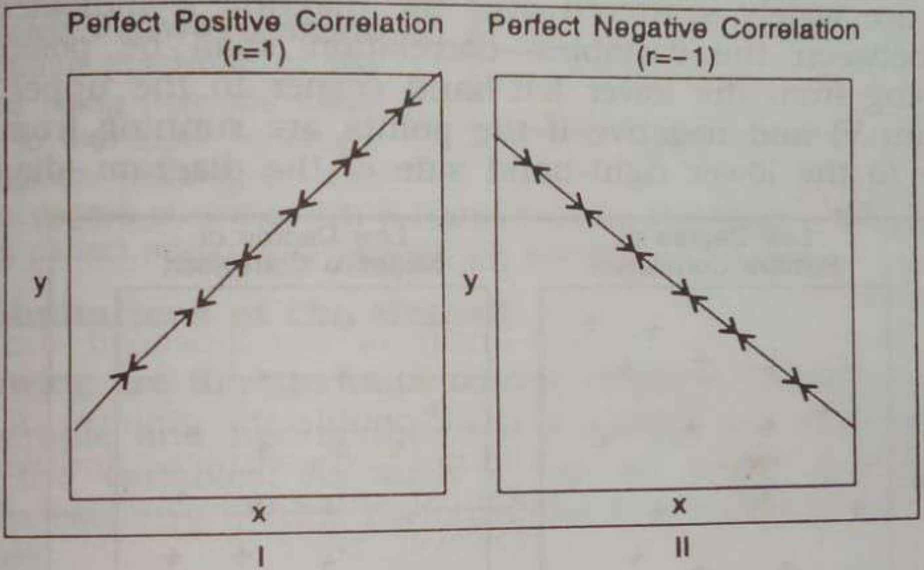


METHODS OF STUDYING CORRELATION

The various methods of ascertaining whether two variables are correlated or not are :

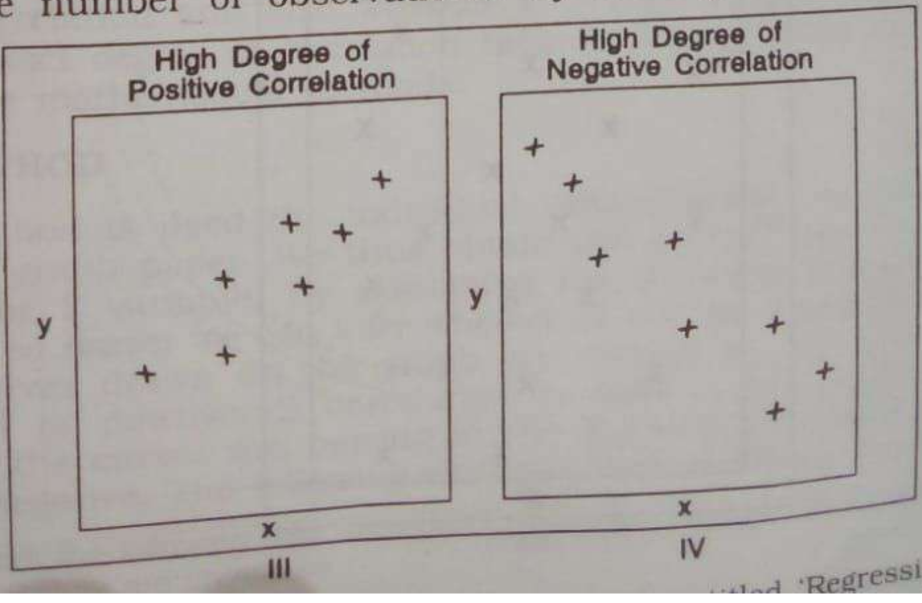
- Scatter Diagram Method
- Graphic Method
- Karl Pearson's Coefficient of Correlation
- Concurrent Deviation Method
- Method of Least Squares.*

Of these, the first two are based on the knowledge of diagrams and graphs, whereas the others are the mathematical methods. Each of these methods shall be discussed in detail in the following pages.



Scatter Diagram Method

The simplest device for ascertaining whether two variables are related is to prepare a dot chart called scatter diagram**. 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. By looking to the scatter of the



tled 'Regression Analysis'.
points.

either multiple or partial correlation. In multiple correlation three or more variables are studied simultaneously. For example, when we study the relationship between the yield of rice per acre and both the amount of rainfall and the amount of fertilizers used, it is a problem of multiple correlation. On the other hand, in partial correlation we recognize more than two variables, but consider only two variables to be influencing each other the effect of other influencing variables being kept constant. For example, in the rice problem taken above if we limit our correlation analysis of yield and rainfall to periods when a certain average daily temperature existed it becomes a problem relating to partial correlation only.

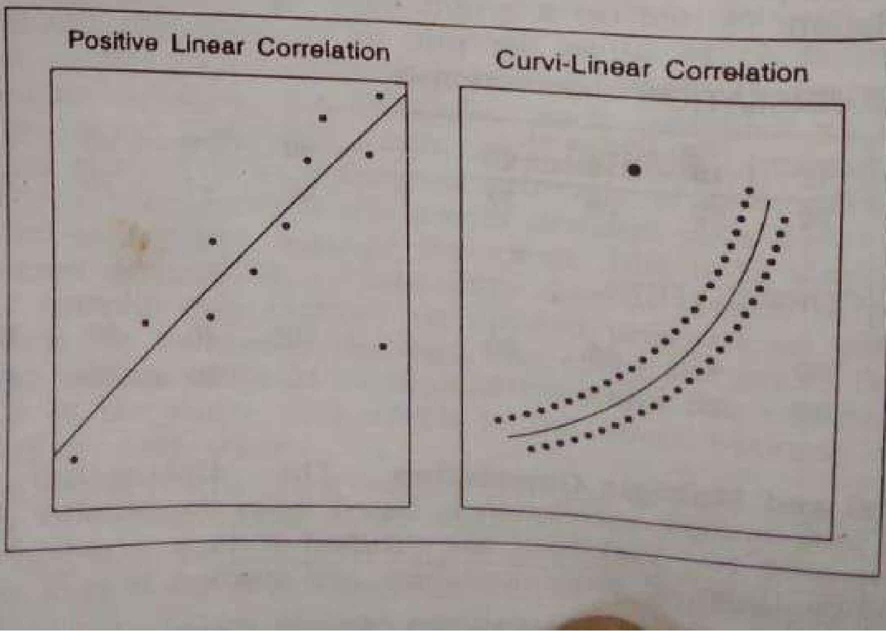
Linear and Non-Linear (Curvilinear) Correlation The distinction between linear and non-linear correlation is based upon the constancy of the ratio of change between the variables. If the amount of change in one variable tends to bear constant ratio to the amount of change in the other variable then the correlation is said to be linear. For example, observe the following two variables X and Y :

X:	10	20	30	40	50
Y:	70	140	210	280	350

It is clear that the ratio of change between the two variables is the same. If such variables are plotted on a graph paper all the plotted points would fall on a straight line.

Correlation would be called non-linear or curvilinear if the amount of change in one variable does not bear a constant ratio to the amount of change in the other variable. For example, if we double the amount of rainfall the production of rice or wheat, etc., would not necessarily be doubled. It may be pointed out that in most of the practical situations, we find a non-linear relationship between the variables. However, since techniques of analysis for measuring non-linear correlation are far more complicated than those for linear correlation, we generally make an assumption that the relationship between the variables is of the linear type.

The following two diagrams will illustrate the difference between linear and curvilinear correlation :



observed [illegible] called spurious [illegible] should remember that it is interpretation of the degree of correlation that is spurious, not the degree of correlation itself. The high degree of correlation indicates only the mathematical result. We should reach a conclusion based on logical reasoning and intelligent investigation of significantly related matters. It may also be pointed out that errors in correlation analysis include not only reading causation into spurious correlation but also interpreting spuriously a perfectly valid relationship.

TYPES OF CORRELATION

Correlation is described or classified in several different ways. Three of the most important ways of classifying correlation are :

1. Positive or negative.
2. Simple, partial and multiple.
3. Linear and non-linear.

Positive and Negative Correlation Whether correlation is positive (direct) or negative (inverse) would depend upon the direction of change of the variables. If both the variables are varying in the same direction, *i.e.*, if as one variable is increasing the other, *on an average*, is also increasing or, if as one variable is decreasing the other, *on an average*, is also decreasing, correlation is said to be positive. If, on the other hand, the variables are varying in opposite directions, *i.e.*, as one variable is increasing, the other is decreasing or *vice versa*, correlation is said to be negative. The following examples would illustrate the difference between positive and negative correlation.

I. POSITIVE CORRELATION

X :	10	12	15	18	20	*X* :	80	70	60	40	30
Y :	15	20	22	25	37	*Y* :	50	44	30	20	10

II. NEGATIVE CORRELATION

X :	20	30	40	60	80	*X* :	100	90	60	40	30
Y :	40	30	22	15	10	*Y* :	10	20	30	40	50

Simple, Partial and Multiple Correlation* The distinction between simple, partial and multiple correlation is based upon the number of variables studied. When only two variables are studied it is a problem of simple correlation. When three or more variables are studied it is a problem of

* For a detailed discussion of Partial and Multiple Correlation Analysis please refer to Chapter 9, Vol. II