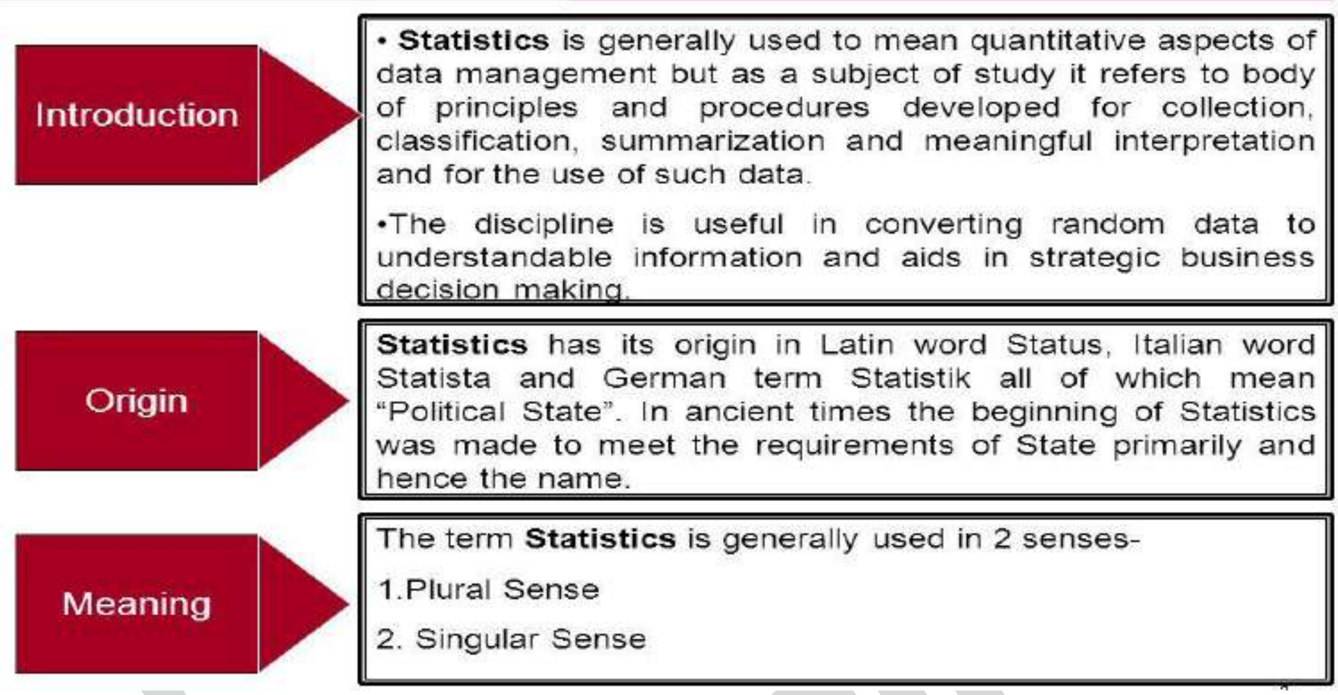
SYLLABUS

Class – B.B.A. III Sem.

Subject – Business Statistics

UNIT – I	Introduction to statistics – purpose of statistics – kinds of numbers, measurements, variables, attributes, UNITs, discrete and continuous data. Frequency distributions – continuous frequency distribution – graphic representation – Histogram – frequency polygon and their uses.
UNIT – II	Measures of central tendency – Mean, Median, Mode merits and demerits – Measure of Dispersion – Range, mean deviation, standard deviation, coefficient of variations.
UNIT – III	Concept of probability – Law of Addition – Law of Multiplication and conditional probabilities – Random variables – Its expectation and variances – Applications in Management. Probability distributions:- Binomial distribution, Poisson distribution and normal distribution.
UNIT – IV	Population parameters and sample statistics – standard error – sampling distributions of statistics – large sample tests – single mean, Difference of means, single proportion – Difference of proportions
UNIT – V	Correlation and regression – Scatter Diagram – Coefficient of correlation – Rank Correlation – Regression – Lines of Regression index nos. Simple, weighted – consumer price index – wholesale price index.
UNIT – VI	Time Series: Components – Trends – Seasonal – cyclical – irregular variables – Measurement of trend – Method of semi averages and method of moving averages principle of least squares – Measurement of seasonal fluctuations – simple averages – Ratio to trend method – Ratio to moving average method – simple problems.

UNIT — I
STATISTICS



Introduction

- **Statistics** is generally used to mean quantitative aspects of data management but as a subject of study it refers to body of principles and procedures developed for collection, classification, summarization and meaningful interpretation and for the use of such data.
- The discipline is useful in converting random data to understandable information and aids in strategic business decision making.

Origin

Statistics has its origin in Latin word Status, Italian word Statista and German term Statistik all of which mean "Political State". In ancient times the beginning of Statistics was made to meet the requirements of State primarily and hence the name.

Meaning

The term **Statistics** is generally used in 2 senses-

1.Plural Sense

2. Singular Sense

Definitions: - "The classified facts relating the condition of the people in a state specially those facts which can be stated in members or in tables of members or in any tabular or classified arrangements."

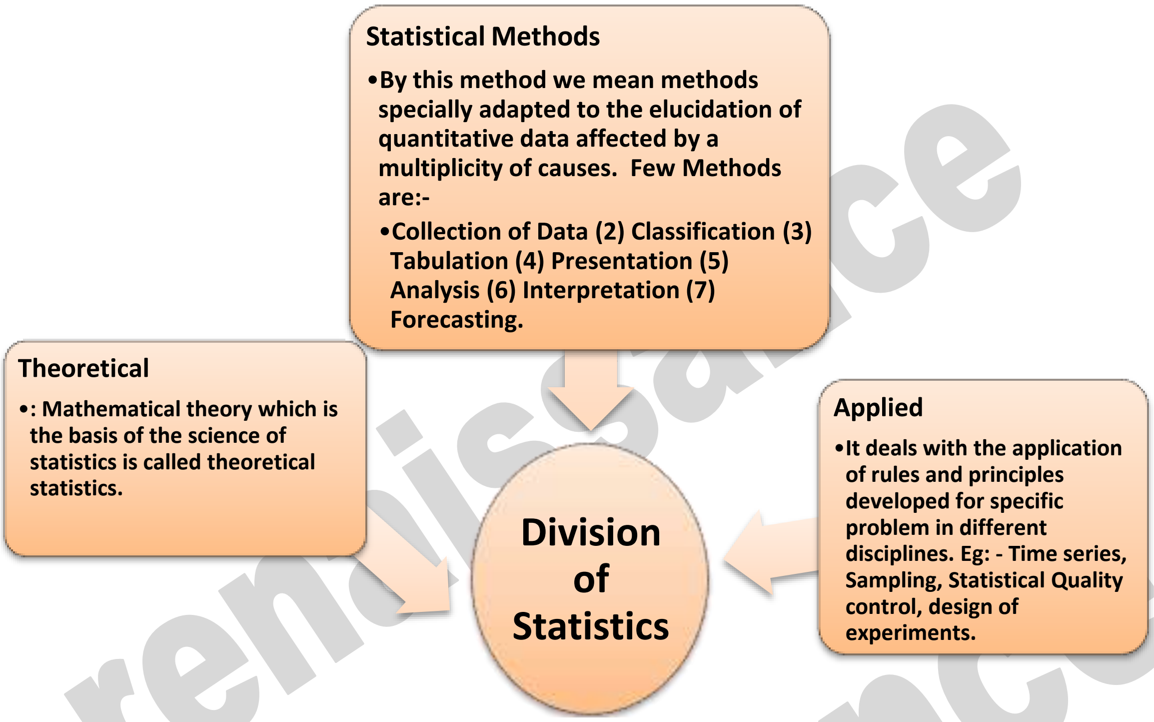
-Webster

"Statistics may be regarded as (i) the study of population (ii) The study of variation (iii) The study of method of reduction of data"

-R.A. Fisher

Nature /Features /Characteristics of statistics

- It is an aggregate of facts
- It is estimated according to reasonable standard of accuracy
- It is numerically expressed
- Analysis of multiplicity of causes.
- It is collected for pre-determined purpose
- It is collected in a systematic manner



Statistical Methods

- By this method we mean methods specially adapted to the elucidation of quantitative data affected by a multiplicity of causes. Few Methods are:-
  - Collection of Data (2) Classification (3) Tabulation (4) Presentation (5) Analysis (6) Interpretation (7) Forecasting.

Theoretical

- : Mathematical theory which is the basis of the science of statistics is called theoretical statistics.

Division of Statistics

Applied

- It deals with the application of rules and principles developed for specific problem in different disciplines. Eg: - Time series, Sampling, Statistical Quality control, design of experiments.

Functions of Statistics:-

- It presents facts in a definite form.
- It simplifies mass of figures
- It facilitates comparison
- It helps in prediction
- It helps in formulating suitable & policies.

Scope of Statistics:-

1. Statistics and state or govt.
2. Statistics and business or management.
   - Marketing
   - Production
   - Finance
   - Banking
   - Control
   - Research and Development
3. Statistics and Economics
   - Measures National Income
   - Money Market analysis
   - Analysis of competition, monopoly, oligopoly,
   - Analysis of Population etc.
4. Statistics and science
5. Statistics and Research

Limitations:-

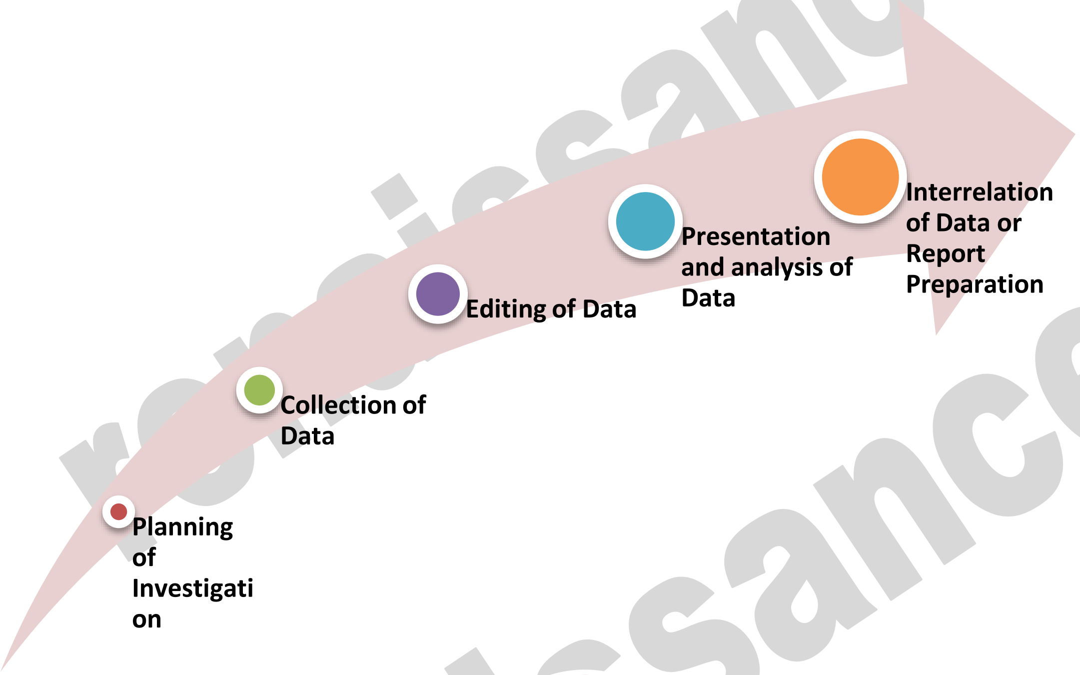
(i) It is not deal with items but deals with aggregates.
(ii) Only on expert can use it
(iii) It is not the only method to analyze the problem.
(iv) It can be misused etc.

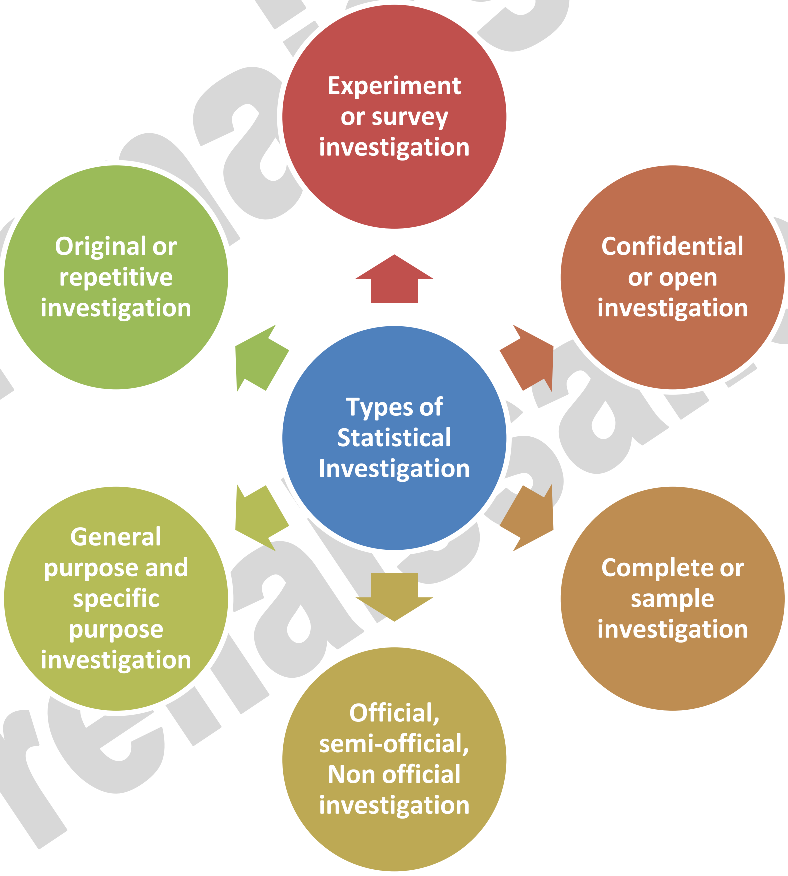
STATISTICAL INVESTIGATION

Meaning: In general it means as a statistical survey.

In brief, it is Scientific and systematic collection of data and their analysis with the help of various statistical method and their interpretation.

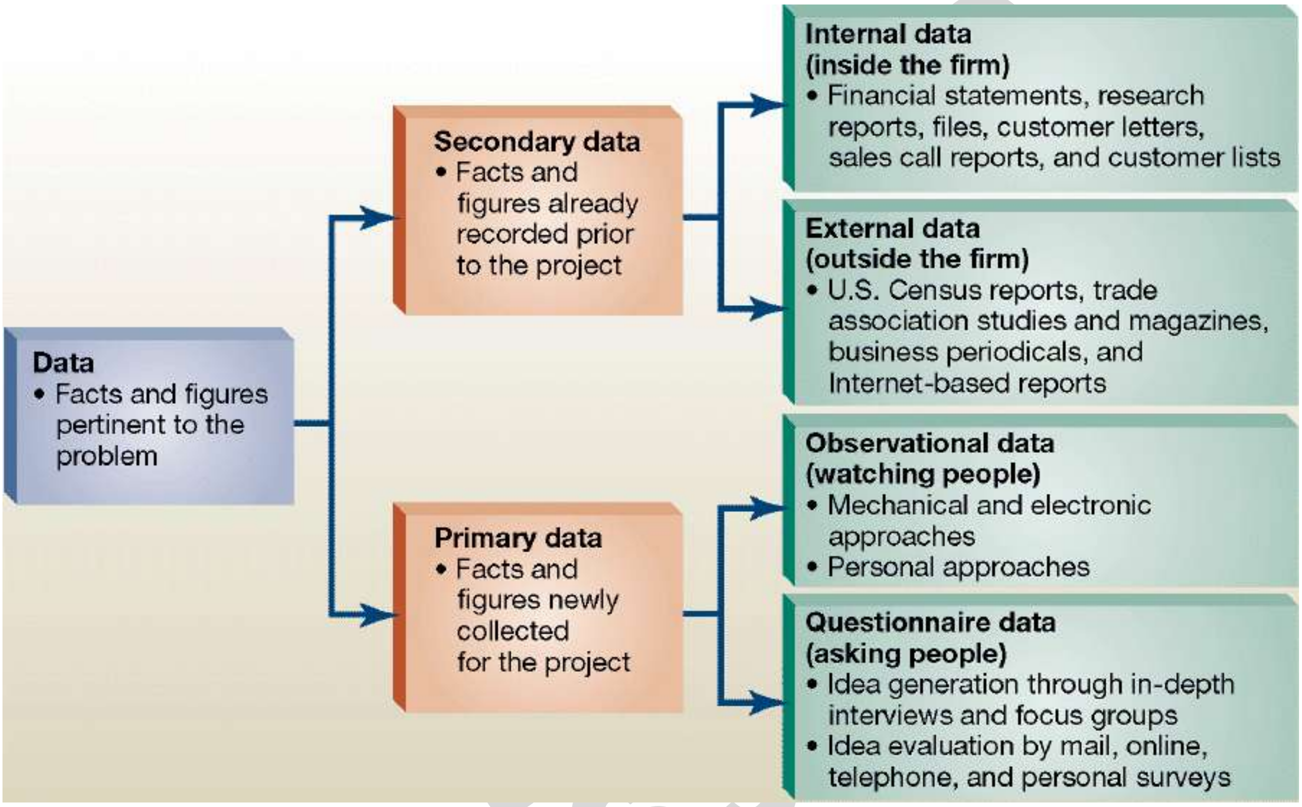
STAGES OF STATISTICAL INVESTIGATION:-





PROCESS OF DATA COLLECTION

Data: - A bundle of Information or bunch of information.
Data Collection: Collecting Information for some relevant purpose & placed in relation to each other.



Collection of Data: - It means the methods that are to be employed for obtaining the required information from the units under investigations.

Methods of Data Collection:- (Primary Data)

- Direct Personal Interviews
- By observation
- By Survey
- By questionnaires

Preparation of Questionnaires:-
This method of data collection is quit popular, particularly in case of big enquires, it is adopted by individuals, research workers. Private and public organization and even by government also.
A questionnaires consists of number of question printed or type in a definite order on a form or set of forms. The respondents have to answer the question on their own.

Importance:-

i. Low cost and universal
ii. Free from biases.
iii. Respondents have adequate time to respond
iv. Fairly approachable

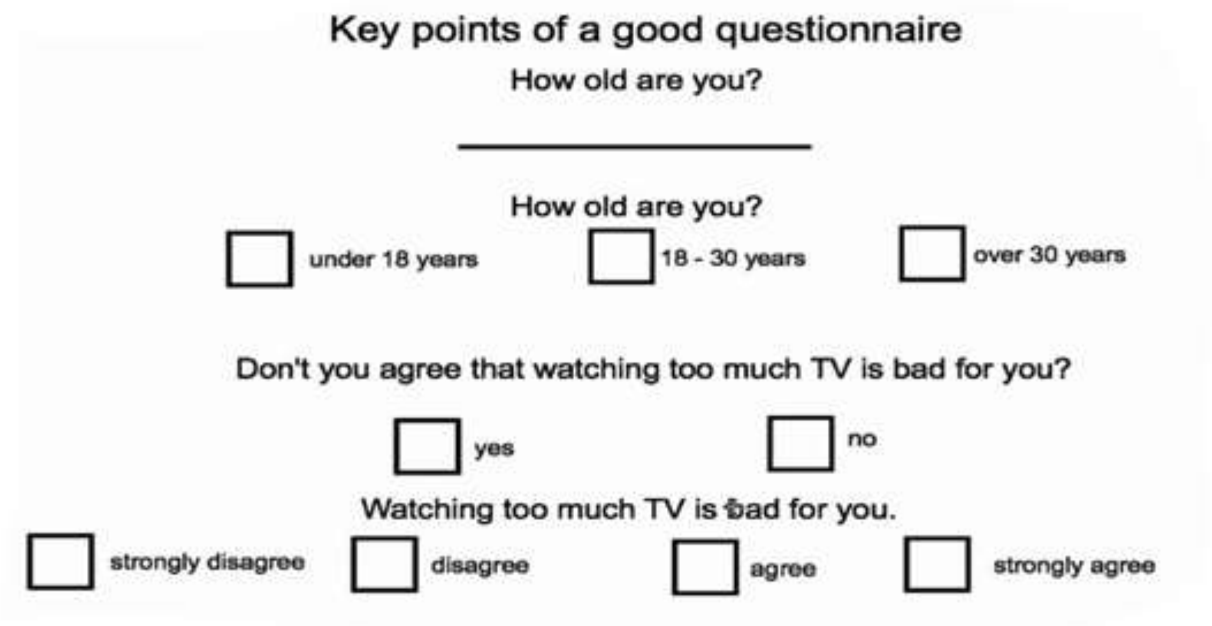
Basis of Difference	Primary Data	Secondary Data
Originality	Primary data are original i.e., collected first time.	Secondary data are not original, i.e.., they are already in existence and are used by the investigator.
Organisation	Primary data are like raw material.	Secondary data are in the from of finished product. They have passed through statistical methods.
Purpose	Primary data are according to the object of investigation and are used without correction.	Secondary data are collected for some other purpose and are corrected before use.
Expenditure	The collection of primary data require large sum, energy and time.	Secondary data are easily available from secondary sources (published or unpublished).
Precautions	Precautions are not necessary in the use of primary data.	Precautions are necessary in the use of secondary data.

Demerits:-
(i) Low rate of return
(ii) Fill on educated respondents
(iii) Slowest method of Response

Steps in construction of a questionnaire : It is considered as the heart of a survey operation. Hence it should be very carefully constructed

- Prepare it in a general form
- Prepare sequence of question
- Emphasize on question formulation and wordings
- Ask Logical and not misleading questions.
- Personal questions should be left to the end.
- Technical terms and vague expressions should be avoided.

Example :



Key points of a good questionnaire

How old are you?

____

How old are you?

☐ under 18 years ☐ 18 - 30 years ☐ over 30 years

Don't you agree that watching too much TV is bad for you?

☐ yes ☐ no

Watching too much TV is bad for you.

☐ strongly disagree ☐ disagree ☐ agree ☐ strongly agree

Classification & Tabulation of Data

After collecting and editing of data an important step towards processing that classification. It is grouping of related facts into different classes.

Types of classification:-

i. **Geographical:-** On the basis of location difference between the various items. E.g. Sugar Cave, wheat, rice, for various states.

ii. **Chronological:-** On the basis of time
e.g.-

Year	Sales
1997	1,84,408
1998	1,84,400
1999	1,05,000

iii. **Qualitative classification: -** Data classified on the basis of some attribute or quality such as, color of hair, literacy, religion etc.

iv. **Quantitative Classification: -** When data is quantify on some units like height, weight, income, sales etc.

Tabulation of Data

A table is a systematic arrangement of statistical data in columns and Rows.

Part of Table:-

1. Table number
2. Title of the Table
3. Caption
4. Stub
5. Body of the table
6. Head note
7. Foot Note

Types of Table:-

(i) Simple and Complex Table:-

(a) Simple or one-way table:-

Age	No. of Employees
25	10
30	7
35	12
40	9
45	6

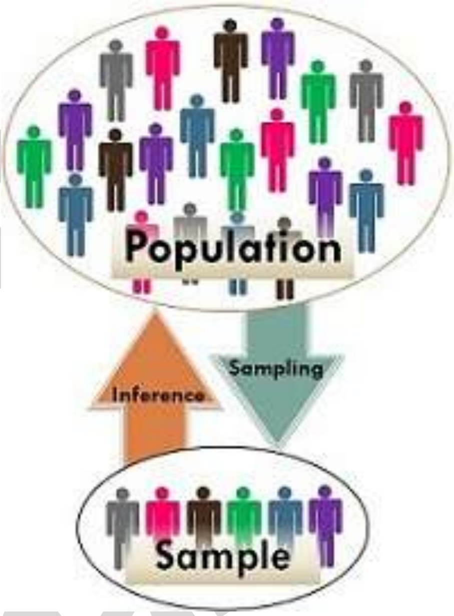
(b) Two way Table

Age	Males	Females	Total
25	25	15	40
30	20	25	45
35	24	20	44
40	18	10	28
45	10	8	18
Total	**97**	**78**	**175**

2) General Purpose and Specific Purpose Table:- General purpose table, also known as the reference table or repository tables, which provides information for general use or reference.
Special purpose are also known as summary or analytical tables which provides information for one particular discussion or specific purpose.

METHODS OF SAMPLING

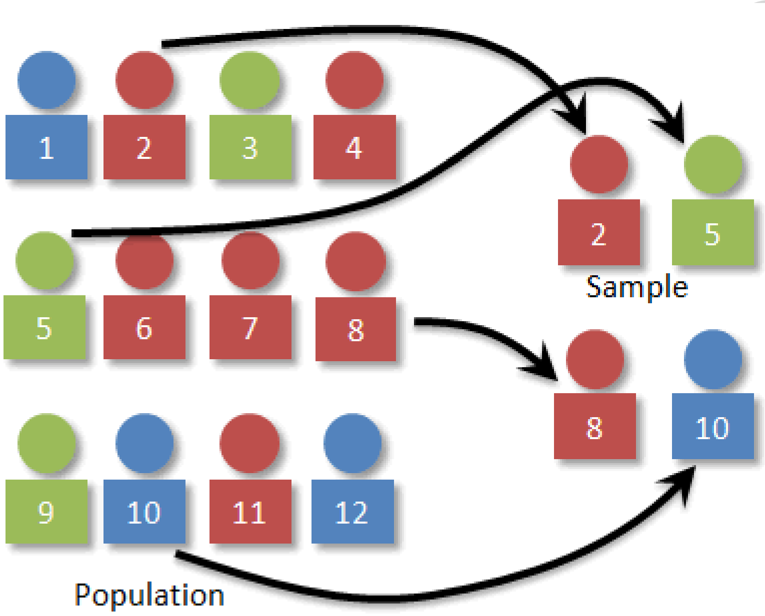
Meaning: - The process of obtaining a sample and its subsequent analysis and interpretation is known as sampling and the process of obtaining the sample if the first stage of sampling.



The various methods of sampling can broadly be divided into:
i. Random sampling method
ii. Non Random sampling method

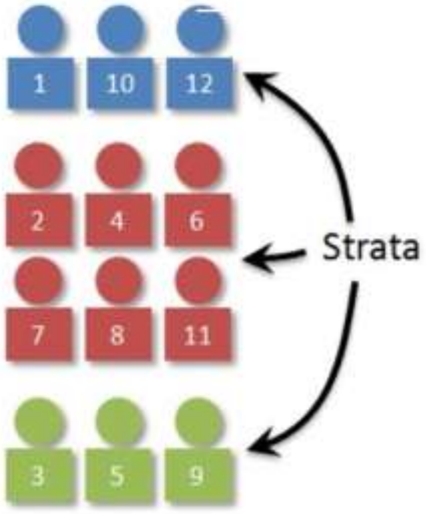
Random Sampling Method

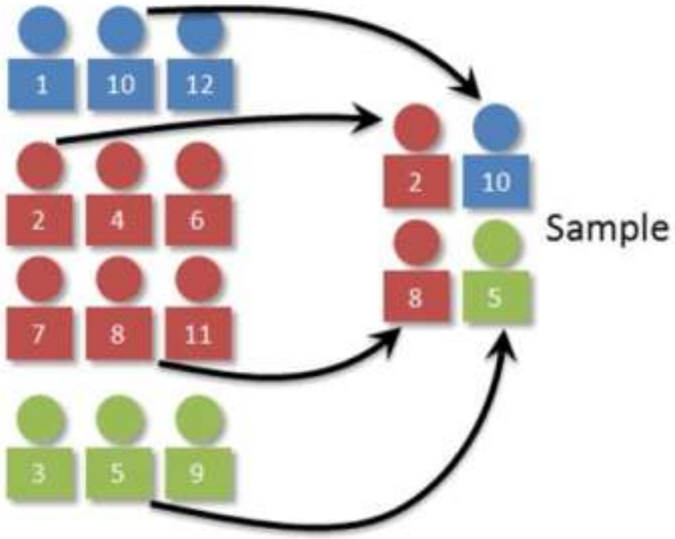
I Simple Random Sampling: - In this method each and every item of the population is given an equal chance of being included in the sample.
(a) Lottery Method (b) Table of Random Numbers



Merits:
Equal opportunity to each item.
Better way of judgment
Easy analysis and accuracy
Limitations:
Different in investigation
Expensive and time consuming
For filed survey it is not good

II Stratified Sampling:- In this it is important to divided the population into homogeneous group called strata. Then a sample may be taken from each group by simple random method.



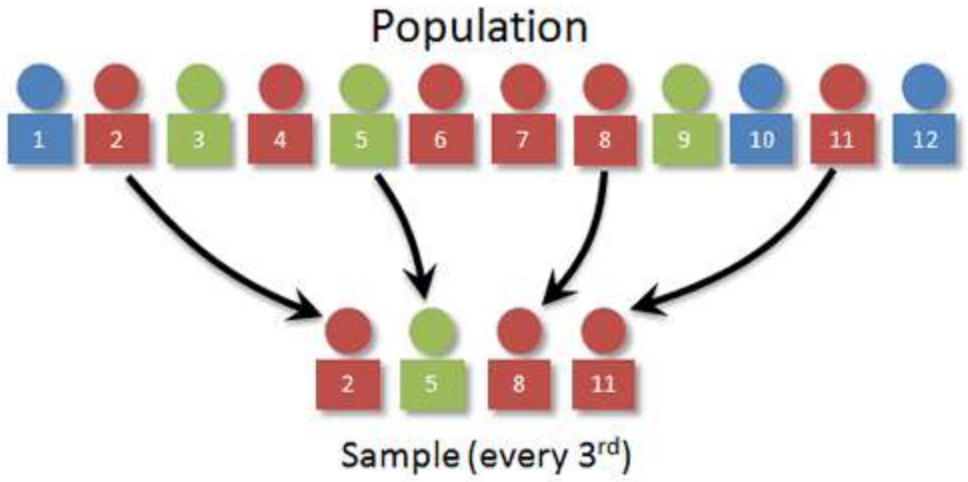


Merit:- More representative sample is used.

Grater accuracy
Geographically Concentrated

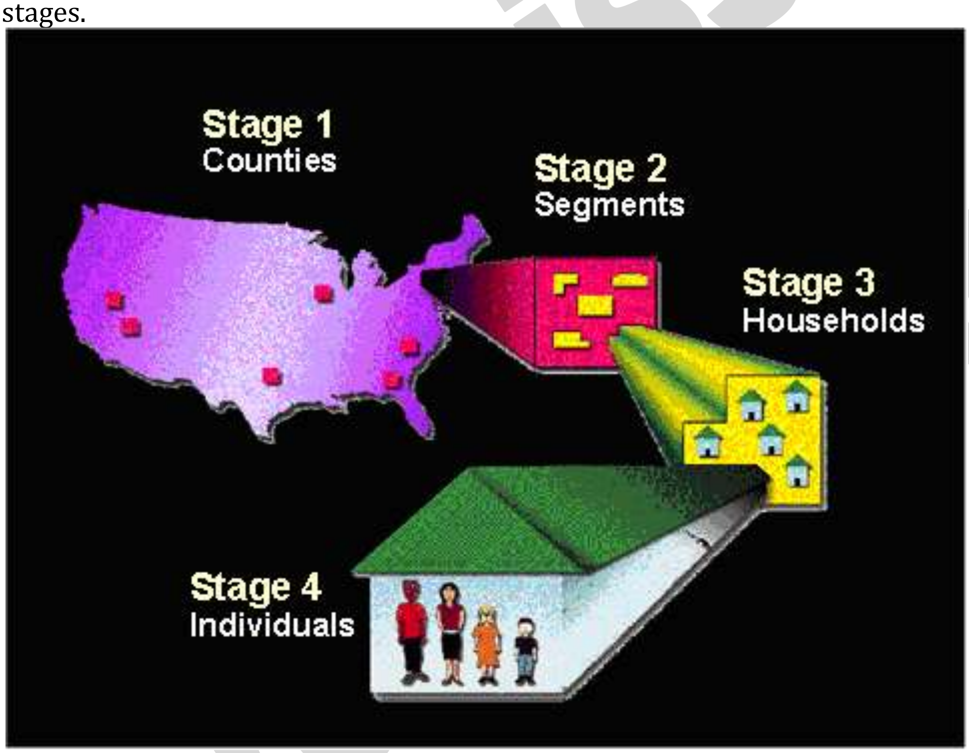
Limitations: Utmost care must be exercised due to homogeneous group deviation. In the absence of skilled supervisor sample selection will be difficult.

III Systematic Sampling:- This method is popularly used in those cases where a complete list of the population from which sampling is to be drawn is available. The method is to be select k th item from the list where k refers to the sampling interval.



Merits: - It can be more convenient.
Limitation: - Can be Baised.

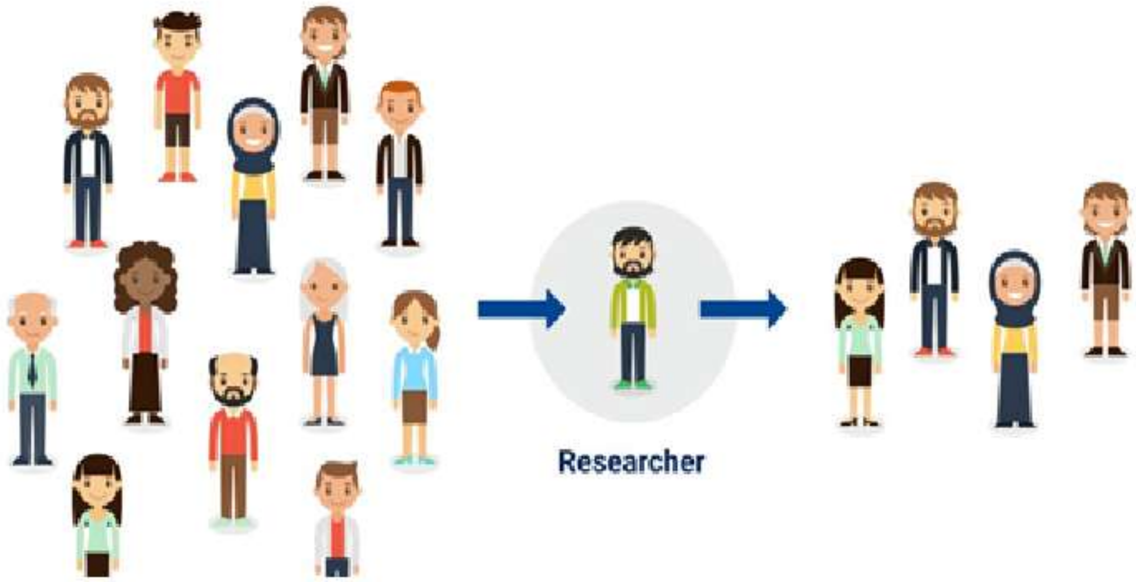
IV Multi- Stage Sampling: - This method refers to a sampling procedure which is carried out in several stages.



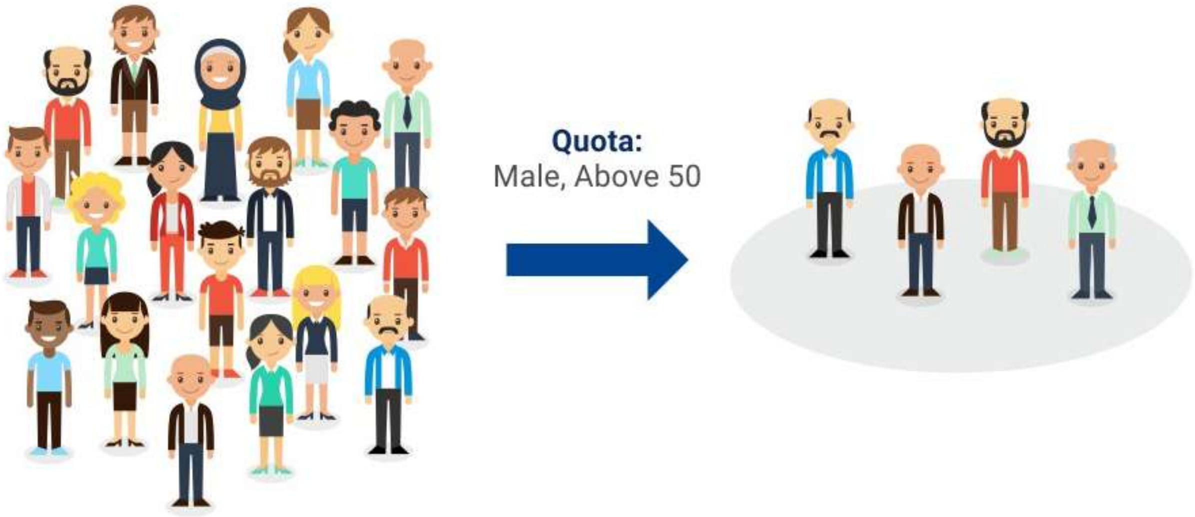
Merit: - It gives flexibility in Sampling
Limitation: - It is difficult and less accurate

Non Random Sampling Method:-

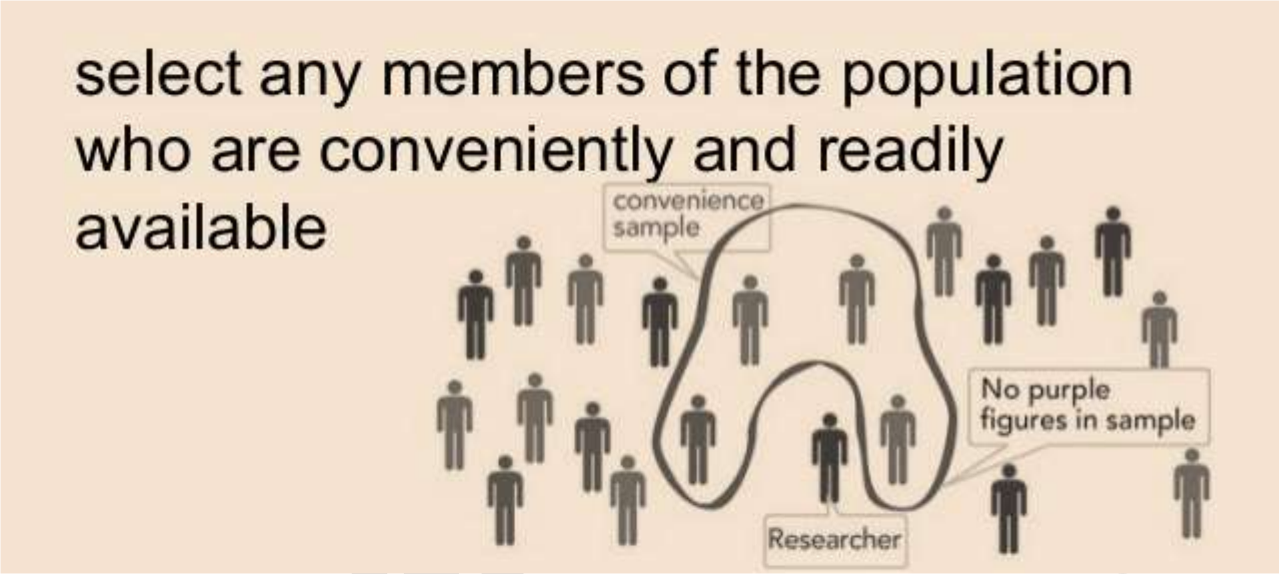
I. **Judgment Sampling: -** The choice of sample items depends exclusively on the judgment of the investigator or the investigator exercises his judgement in the choice of sample items. This is an simple method of sampling.



II. **Quota Sampling: -** Quotas are set up according to given criteria, but, within the quotas the selection of sample items depends on personal judgment.



III. **Convenience Sampling: -** It is also known as chunk. A chunk is a fraction of one population taken for investigation because of its convenient availability. That is why a chunk is selected neither by probability nor by judgment but by convenience.



select any members of the population who are conveniently and readily available

Size of Sample:- It depends upon the following things:-
Cost aspects.
The degree of accuracy desired.
Time, etc.
Normally it is 5% or 10% of the total population.

Limitation of overall sampling Method:-
Some time result may be inaccurate and misleading due to wrong sampling.
Its always needs superiors and experts to analyze the sample.
It may not give information about the overall defects. In production or any study.
It Becomes Biased due to following reason:-
(a) Faulty process of selection
(b) Faulty work during the collection of information
(c) Faulty methods of analysis etc.

UNIT-II
MEASURES OF CENTRAL TENDENCY

The point around which the observations concentrate in general in the central part of the data is called central value of the data and the tendency of the observations to concentrate around a central point is known as Central Tendency.

Objects of Statistical Average:

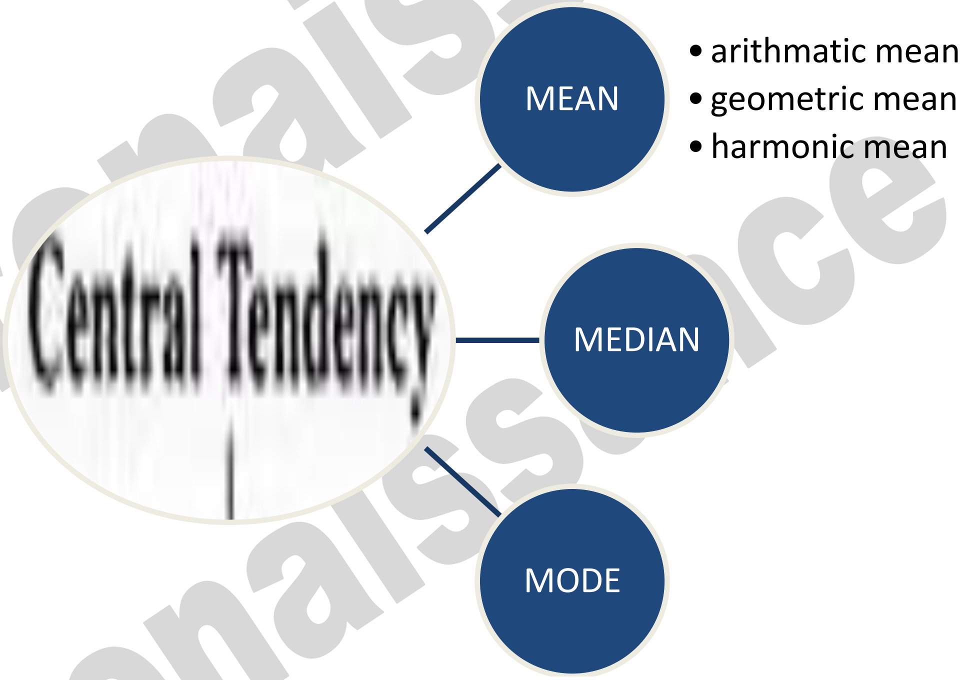
- To get a single value that describes the characteristics of the entire group
- To facilitate comparison

Functions of Statistical Average:

- Gives information about the whole group
- Becomes the basis of future planning and actions
- Provides a basis for analysis
- Traces mathematical relationships
- Helps in decision making

Requisites of an Ideal Average:

- Simple and rigid definition
- Easy to understand
- Simple and easy to compute
- Based on all observations
- Least affected by extreme values
- Least affected by fluctuations of sampling
- Capable of further algebric treatment



ARITHMETIC MEAN ($\bar{X}$)

Arithmetic Mean of a group of observations is the quotient obtained by dividing the sum of all observations by their number. It is the most commonly used average or measure of the central tendency applicable only in case of quantitative data. Arithmetic mean is also simply called “mean”. Arithmetic mean is denoted by $\overline{X}$.

Merits

- It is rigidly defined.
- It is easy to calculate and simple to follow.
- It is based on all the observations.
- It is readily put to algebraic treatment.
- It is least affected by fluctuations of sampling.
- It is not necessary to arrange the data in ascending or descending order.

DeMerits

- The arithmetic mean is highly affected by extreme values.
- It cannot average the ratios and percentages properly.
- It cannot be computed accurately if any item is missing.
- The mean sometimes does not coincide with any of the observed value.
- It cannot be determined by inspection.
- It cannot be calculated in case of open ended classes.

Uses

- When the frequency distribution is symmetrical.
- When we need a stable average.
- When other measures such as standard deviation, coefficient of correlation are to be computed later.

MEDIAN (M)

The median is that value of the variable which divides the group into two equal parts, one part comprising of all values greater and other of all values less than the median. For calculation of median the data has to be arranged in either ascending or descending order. Median is denoted by **M**.

Merits

- It is easily understood and easy to calculate.
- It is rigidly defined.
- It can sometimes be located by simple inspection and can also be computed graphically.
- It is positional average therefore not affected at all by extreme observations.
- It is only average to be used while dealing with qualitative data like intelligence, honesty etc.
- It is especially useful in case of open end classes since only the position and not the value of items must be known.
- It is not affected by extreme values.

DeMerits

- For calculation, it is necessary to arrange data in ascending or descending order.
- Since it is a positional average, its value is not determined by each and every observation.
- It is not suitable for further algebraic treatment.
- It is not accurate for large data.
- The value of median is more affected by sampling fluctuations than the value of the arithmetic mean.

Uses

- When there are open-ended classes provided it does not fall in those classes.
- When exceptionally large or small values occur at the ends of the frequency distribution.
- When the observation cannot be measured numerically but can be ranked in order.
- To determine the typical value in the problems concerning distribution of wealth etc.

MODE (Z)

Mode is the value which occurs the greatest number of times in the data. The word mode has been derived from the French word **'La Mode'** which implies fashion. The Mode of a distribution is the value at the point around which the items tend to be most heavily concentrated. It may be regarded as the most typical of a series of values. Mode is denoted by **Z.**

Merits

- **It is easy to understand and simple to calculate.**
- **It is not affected by extreme large or small values.**
- **It can be located only by inspection in ungrouped data and discrete frequency distribution.**
- **It can be useful for qualitative data.**
- **It can be computed in open-end frequency table.**
- **It can be located graphically**

DeMerits

- **It is not well defined.**
- **It is not based on all the values.**
- **It is suitable for large values and it will not be well defined if the data consists of small number of values.**
- **It is not capable of further mathematical treatment.**
- **Sometimes, the data has one or more than one mode and sometimes the data has no mode at all.**

Uses

- **When a quick approximate measure of central tendency is desired.**
- **When the measure of central tendency should be the most typical value.**

GEOMETRIC MEAN (G.M)

The geometric mean also called geometric average is the nth root of the product of n non-negative quantities. Geometric Mean is denoted by **G.M**.

Properties of Geometric Mean:

- The geometric mean is less than arithmetic mean, G.M<A.M
- The product of the items remains unchanged if each item is replaced by the geometric mean.
- The geometric mean of the ratio of corresponding observations in two series is equal to the ratios their geometric means.
- The geometric mean of the products of corresponding items in two series.

Merits of Geometric Mean:

- It is rigidly defined and its value is a precise figure.
- It is based on all observations.
- It is capable of further algebraic treatment.
- It is not much affected by fluctuation of sampling.
- It is not affected by extreme values.

Demerits of Geometric Mean:

- It cannot be calculated if any of the observation is zero or negative.
- Its calculation is rather difficult.
- It is not easy to understand.
- It may not coincide with any of the observations.

Uses of Geometric Mean:

- Geometric Mean is appropriate when:

  - Large observations are to be given less weight.
  - We find the relative changes such as the average rate of population growth, the average rate of intrest etc.
  - Where some of the observations are too small and/or too large.
- Also used for construction of Index Numbers.

HARMONIC MEAN (H.M)

Harmonic mean is another measure of central tendency. Harmonic mean is also useful for quantitative data. Harmonic mean is quotient of "number of the given values" and "sum of the reciprocals of the given values". It is denoted by **H.M.**

Merits of Harmonic Mean:

- It is based on all observations.
- It not much affected by the fluctuation of sampling.
- It is capable of algebraic treatment.
- It is an appropriate average for averaging ratios and rates.
- It does not give much weight to the large items and gives greater importance to small items.

Demerits of Harmonic Mean:

- Its calculation is difficult.
- It gives high weight-age to the small items.
- It cannot be calculated if any one of the items is zero.
- It is usually a value which does not exist in the given data.

Uses of Harmonic Mean:

- Harmonic mean is better in computation of average speed, average price etc. under certain conditions.

DISPERSION

The Dispersion (Known as Scatter, spread or variations) measures the extent to which the items vary from some central value. The measures of dispersion is also called the average of second order (Central tendency is called average of first order).

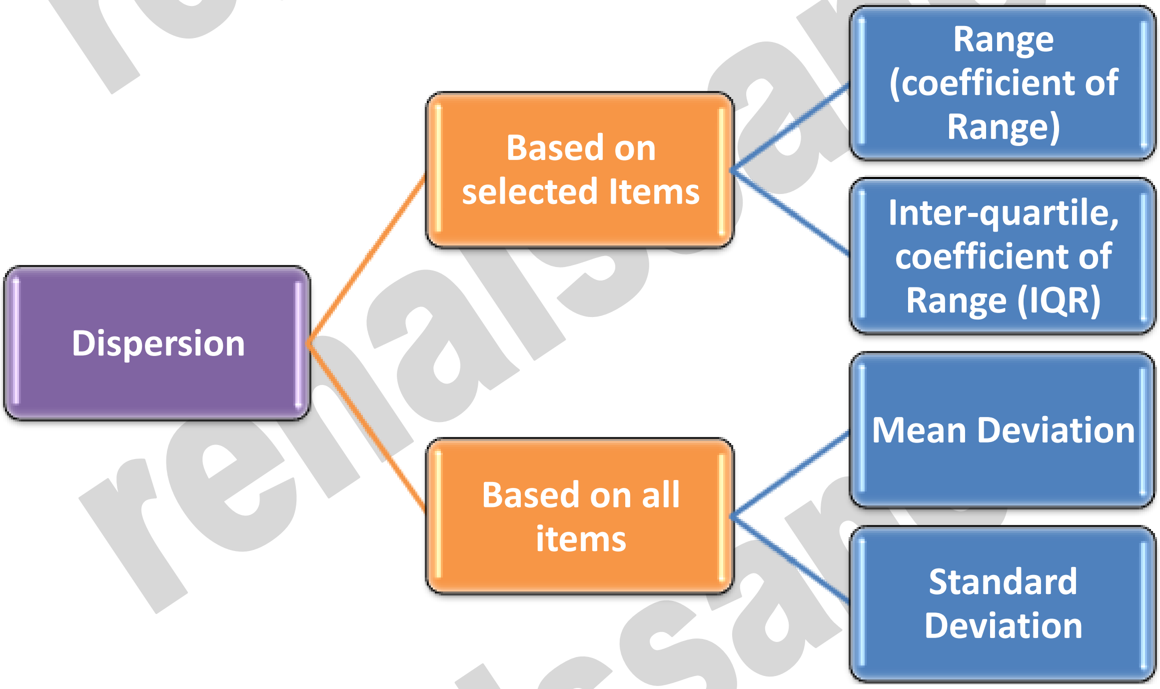
The two distributions of statistical data may be symmetrical and have common means, median or mode, yet they may differ widely in the scatter or their values about the measures of central tendency.

Significance/ objectives of Dispersion-

- To judge the reliability of average
- To compare the two an more series
- To facilitate control
- To facilitate the use of other statistical measures.

Properties of good Measure of Dispersion

- Simple to understand
- Easy to calculate
- Rigidly defined
- Based on all items
- Sampling stability
- Not unduly affected by extreme items.
- Good for further algebraic treatment



1. **Range: -** Range (R) is defined as the difference between the value of largest item and value of smallest item included in the distributions. Only two extreme of values are taken into considerations. It also does not consider the frequency at all series.
2. **Quartile Deviation: -** Quartile Deviation is half of the difference between upper quartile (Q3) and lower quartile (Q1). It is very much affected by sampling distribution.
3. **Mean Deviation: -** Mean Deviation or Average Deviation (δAlpha) is arithmetic average of deviation of all the values taken from a statistical average (Mean, Median, and Mode) of the series.

In taking deviation of values, algebraic sign + and – are also treated as positive deviations. This is also known as first absolute moment.

4. **Standard Deviation:-** The standard deviation is the positive root of the arithmetic mean of the squared deviation of various values from their arithmetic mean. The S.D. is denoted as σ Sigma.

Distinction between mean deviation and standard deviation

Base	Mean Deviation	Standard Deviation
Algebric Sign	Actual +, - Signs are ignored and all deviation are taken as positive	Actual signs +, - are not ignored whereas they are squared logically to be ignored.
Use of Measure	Mean deviation can be computed from mean, median, mode	Standard deviation is computed through mean only
Further algebraic Treatment	It is not capable of further algebraic treatment.	It is capable of further algebraic treatment
Simplicity	M.D is simple to understand and easy to calculate	S.D is somewhat complex than mean deviation.
Based	It is based on simple average of sum of absolute deviation	It is based on square root of the average of the squared deviation

Variance

The square of the standard deviation is called variance. In other words the arithmetic mean of the squares of the deviation from arithmetic mean of various values is called variance and is denoted as σ2. Variance is also known as second movement from mean. In other way, the positive root of the variance is called S.D.

Coefficient of Variations- To compare the dispersion between two and more series we define coefficient of S.D. The expression is $\frac{\sigma}{\bar{X}}$ x 100 = known as coefficient of variations.

Interpretation of Coefficient of Variance-

Value of variance	Interpretation
Smaller the value of σ2	Lesser the variability or greater the uniformity/ stable/ homogenous of population
Larger the value of σ2	Greater the variability or lesser the uniformity/ consistency of the population

RANGE = R

Individual Series	Discrete Series	Continuous Series
Range = L-S Where L=Largest, S=Smallest Observation	$R = L - S$	$R = L - S$
Coefficient of Range $\frac{L - S}{L + S}$	$\frac{L - S}{L + S}$	$\frac{L - S}{L + S}$

QUARTILE DEVIATION - Q.D.

Individual Series	Discrete Series	Continuous Series
$Q.D. = Q_3 - Q_1$	$Q.D. = Q_3 - Q_1$	$Q.D. = Q_3 - Q_1$
Coefficient of Q.D. $= \frac{Q_3 - Q_1}{Q_3 + Q_1}$	$= \frac{Q_3 - Q_1}{Q_3 + Q_1}$	$= \frac{Q_3 - Q_1}{Q_3 + Q_1}$

MEAN DEVIATION - M.D. δ ("Through actual Mean, Mode, Median)

Individual Series	Discrete Series	Continuous Series
$\delta M(Median) = \frac{\sum dM}{N}$	$\frac{\sum fdM}{N}$	$\frac{\sum fdM}{N}$
Coefficient of $\delta = \frac{\delta}{M}$	$\delta = \frac{\delta}{M}$	$\delta = \frac{\delta}{M}$
Mean $\delta \overline{X} = \frac{\sum dx}{N}$	$\frac{\sum fd\overline{x}}{N}$	$\frac{\sum fd\overline{x}}{N}$
Coefficient of $\overline{X} = \frac{\delta}{\overline{X}}$	$\frac{\delta}{\overline{X}}$	$\frac{\delta}{\overline{X}}$
(Mode) $\delta Z = \frac{\sum dz}{N}$	$\frac{\sum fdz}{N}$	$\frac{\sum fdz}{N}$
Coefficient of δZ $\frac{\delta}{Z}$	$\frac{\delta}{Z}$	$\frac{\delta}{Z}$

Standard Deviation = σ can be calculated through mean only

	Individual Series	Discrete Series	Continuous Series
Direct (Through actual mean)	$\sqrt{\frac{\sum d_x 2}{N}}$	$\sqrt{\frac{\sum fd^2}{\sum f}}$	$\sqrt{\frac{\sum fd^2}{\sum f}}$
Indirect (Through assumed mean)	$\sqrt{\frac{\sum dx^2}{N} - \left(\frac{\sum dx}{N}\right)^2}$	$\sqrt{\frac{\sum fdx^2}{\sum f} - \left(\frac{\sum fdx}{\sum f}\right)^2}$	$\sqrt{\frac{\sum fdx^2}{\sum f} - \left(\frac{\sum fdx}{\sum f}\right)^2}$

UNIT-III

Chance behavior is unpredictable in the short run, but has a regular and predictable pattern in the long run.

The **probability** of any outcome of a random phenomenon is the proportion of times the outcome would occur in a very long series of repetitions.

- Sample Space - the set of all possible outcomes of a random phenomenon
- Event - any set of outcomes of interest
- Pr(A) is the probability of event A
- A ∪ B is the event that either A or B or both occur.
- A ∩ B is the event that both A and B occur simultaneously.
- The complement of A is denoted by A'. A' is the event that A does not occur.
- Note that Pr(A') = 1 − Pr(A).

Example
Suppose we roll two die and take their sum S = {2, 3, 4, 5, .., 11, 12}
Pr(sum = 5) = 4/36
Because we get the sum of two die to be 5 if we roll a (1,4),(2,3),(3,2) or (4,1).

Let A and B denote two events.

A and B are mutually exclusive if both cannot occur at the same time.

A and B are independent events if and only if $Pr(A \cap B) = Pr(A)\, Pr(B)$

Laws of Probability

Multiplication Law: If A 1, · · · , A k are independent events, then Pr(A 1 ∩ A 2 ∩ · · · ∩ A k) = Pr(A 1) Pr(A 2)· · ·Pr(A k).

Addition Law: If A and B are any events, then Pr(A ∪ B) = Pr(A) + Pr(B) − Pr(A ∩ B)

Note: This law can be extended to more than 2 events.

Conditional Probability : The conditional probability of B given A Pr(B|A) = Pr(A ∩ B) Pr(A) A and B are independent events if and only if Pr(B|A) = Pr(B) = Pr(B|A)

Random Variable: A random variable is a variable whose value is a numerical outcome of a random phenomenon. It is usually denoted by X, Y or Z.

TYPES OF RANDOM VARIABLE

Discrete - a random variable that has finite or countable infinite possible values Example: the number of days that it rains yearly

Continuous - a random variable that has an (continuous) interval for its set of possible values

Example: amount of preparation time for the SAT

Probability Distributions the probability distribution for a random variable X gives the possible values for X, and the probabilities associated with each possible value (i.e., the likelihood that the values will occur)

Binomial Distribution

- Two possible outcomes: Success (S) and Failure (F).
- Repeat the situation n times (i.e., there are n trials).
- The "probability of success," p, is constant on each trial.

- The trials are independent

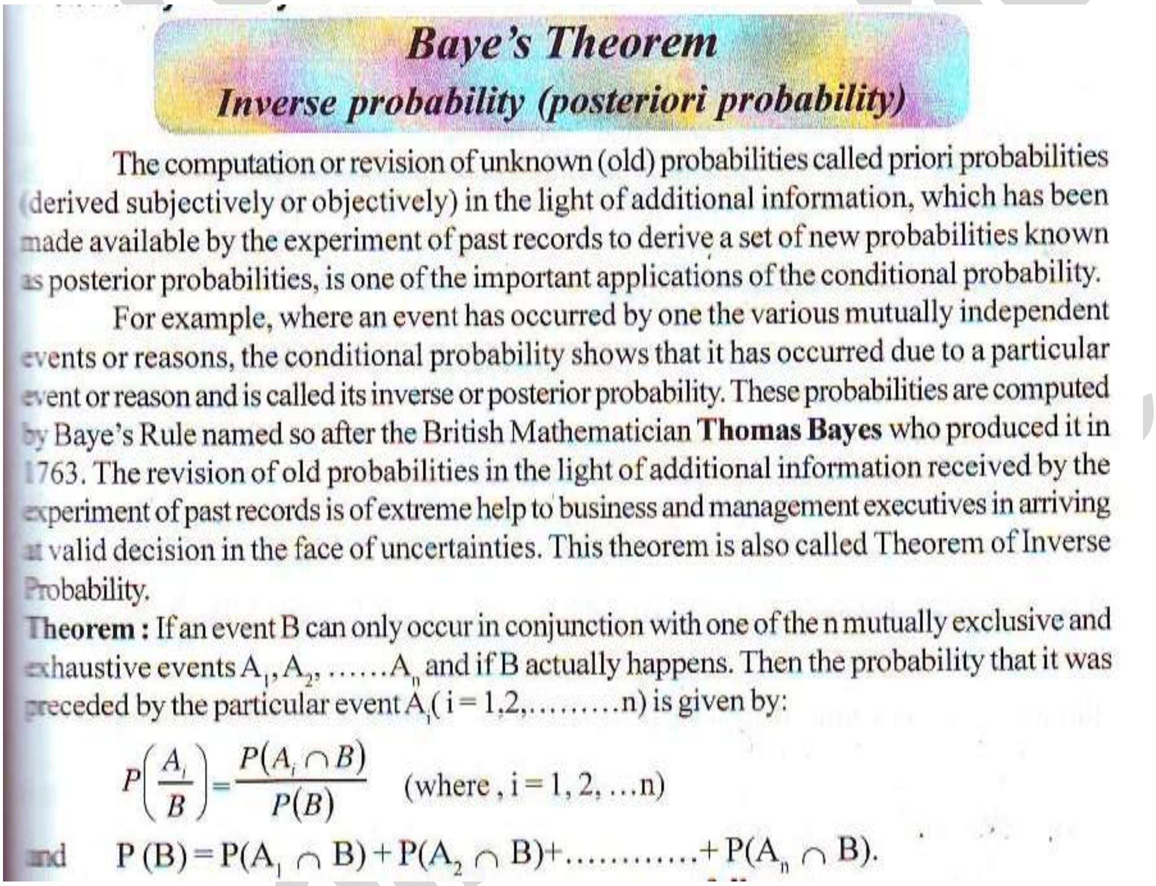
Let X = the number of S's in n independent trials. (X has values x = 0, 1, 2, · · · , n) Then X has a binomial distribution with parameters n and p.

The binomial probability mass function is

$$Pr(X = x) = {}^{n}C_{x}p^{x}(1 - p)^{n-x}, x = 0, 1, 2, \cdots, n$$

Expected Value: μ = E (X) = np

Variance: σ 2 = V ar (X) = np(1 − p)



Baye's Theorem

Inverse probability (posteriori probability)

The computation or revision of unknown (old) probabilities called priori probabilities (derived subjectively or objectively) in the light of additional information, which has been made available by the experiment of past records to derive a set of new probabilities known as posterior probabilities, is one of the important applications of the conditional probability.

For example, where an event has occurred by one the various mutually independent events or reasons, the conditional probability shows that it has occurred due to a particular event or reason and is called its inverse or posterior probability. These probabilities are computed by Baye's Rule named so after the British Mathematician **Thomas Bayes** who produced it in 1763. The revision of old probabilities in the light of additional information received by the experiment of past records is of extreme help to business and management executives in arriving at valid decision in the face of uncertainties. This theorem is also called Theorem of Inverse Probability.

Theorem : If an event B can only occur in conjunction with one of the n mutually exclusive and exhaustive events $A_1, A_2, \ldots\ldots A_n$ and if B actually happens. Then the probability that it was preceded by the particular event A_i (i = 1,2,.........n) is given by:

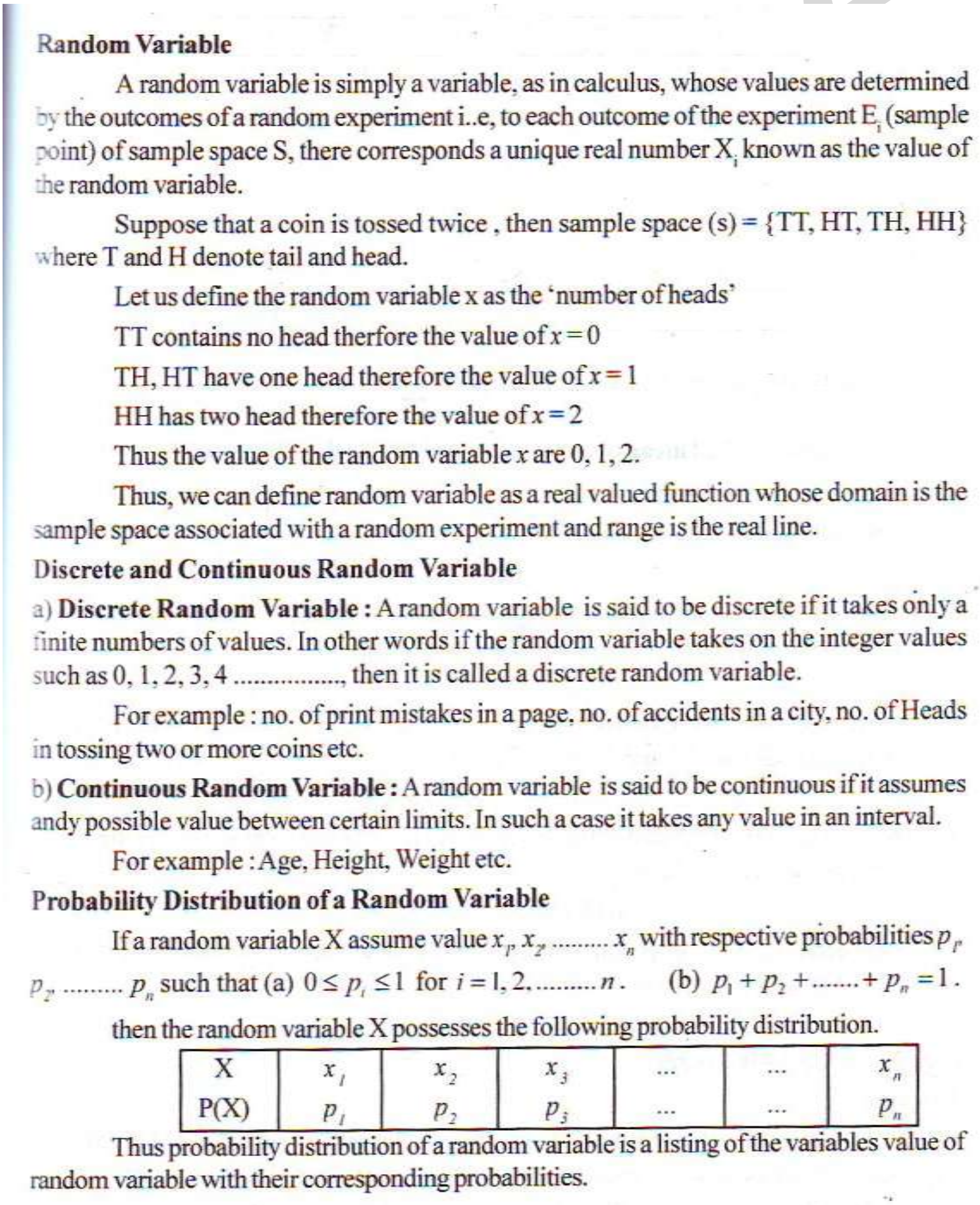
$$P\left(\frac{A_i}{B}\right)=\frac{P(A_i \cap B)}{P(B)} \quad \text{(where , i = 1, 2, ...n)}$$

and $$P(B) = P(A_1 \cap B) + P(A_2 \cap B) + \ldots\ldots\ldots\ldots + P(A_n \cap B).$$

The Formula for bayes is:

$$P\left(\frac{A_1}{B}\right)=\frac{P(A_1\cap B)}{P(B)}=\frac{P(A_1)P(B/A_1)}{P(A_1\cap B)+P(A_2\cap B)+....P(A_n\cap B)}$$

.



Random Variable

A random variable is simply a variable, as in calculus, whose values are determined by the outcomes of a random experiment i..e, to each outcome of the experiment E_i (sample point) of sample space S, there corresponds a unique real number X_i known as the value of the random variable.

Suppose that a coin is tossed twice , then sample space (s) = {TT, HT, TH, HH} where T and H denote tail and head.

Let us define the random variable x as the 'number of heads'

TT contains no head therfore the value of $x = 0$

TH, HT have one head therefore the value of $x = 1$

HH has two head therefore the value of $x = 2$

Thus the value of the random variable x are 0, 1, 2.

Thus, we can define random variable as a real valued function whose domain is the sample space associated with a random experiment and range is the real line.

Discrete and Continuous Random Variable

a) **Discrete Random Variable :** A random variable is said to be discrete if it takes only a finite numbers of values. In other words if the random variable takes on the integer values such as 0, 1, 2, 3, 4 ...., then it is called a discrete random variable.

For example : no. of print mistakes in a page, no. of accidents in a city, no. of Heads in tossing two or more coins etc.

b) **Continuous Random Variable :** A random variable is said to be continuous if it assumes andy possible value between certain limits. In such a case it takes any value in an interval.

For example : Age, Height, Weight etc.

Probability Distribution of a Random Variable

If a random variable X assume value x_1, x_2 .... x_n with respective probabilities p_1, p_2 .... p_n such that (a) $0 \le p_i \le 1$ for $i = 1, 2, ...n$. (b) $p_1 + p_2 + ... + p_n = 1$.

then the random variable X possesses the following probability distribution.

X	x_1	x_2	x_3	...	...	x_n
P(X)	p_1	p_2	p_3	...	...	p_n

Thus probability distribution of a random variable is a listing of the variables value of random variable with their corresponding probabilities.

Expectation : E(x) = $\sum_{i=1}^{n} p_i x_i$

Variance : var(x) = E(x²) – (E(x))²

Concept of Probability Distribution

In the population, the values of the random variable may be distributed according to some definite probability law which can be expressed mathematically on the basis of theoretical considerations and the corresponding probability distribution is known as Theoretical Probability Distribution. For these distributions, a random experiment is theoretically assumed to serve as model and the probability are given by a function of the random variable called 'Probability Function'. Only three distributions which are most popular and widely used, are discussed :

i) Binomial Distribution (Discrete Distribution);

ii) Poisson Distribution (Discrete Distribution);

iii) Normal Distribution (Continuous Distribution).

Binomial Distribution

Binomial distribution is associated with the name of James Bernoulli (1654-1705) but it was published in 1713, eight years after his death. It is also known as **Bernoulli Distribution** to honour its author. Binomial means two names hence the frequency distribution falls into two categories a dichotomous process. A binomial dostribution is a probability distribution expressing the probability of one set of dichotomous a alternatives for example success or failure.

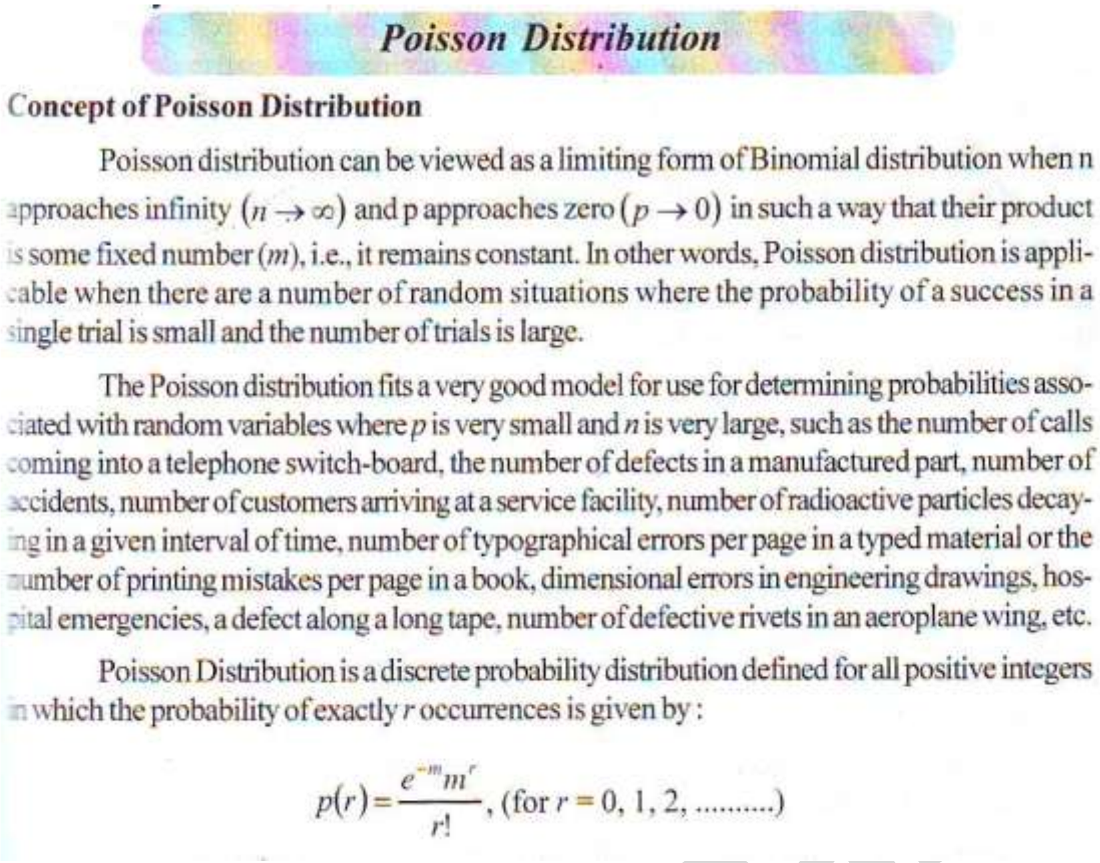
Concept of Binomial Distribution

Let us suppose that a trial is repeated n times (for example tossing a coin n times). We call the occurrence of an event a success and its non-occurrence a failure. Let p be the probability of a success and q be the probability of a failure in a single trial, so that $p + q = 1$ we shall assume that the trials are independent and p and q are same in every trial. By the theorem of Compound probability, the probability that the first r trials are successes and the remaining n-r trials are failures

$$\underbrace{p\times p........\times p}_{r\ times};\ \underbrace{q\times q........\times q}_{(n-r)\ times} = p^r q^{n-r}$$

But r successes in n trials can occur in ${}^nC_r$ mutually exclusive ways and the probability of each such way is $p^r q^{n-r}$ so by addition theorem of probability. The probability of r successes in n trials in any order is given by ${}^nC_r\ p^r q^{n-r}$ i.e.

$$p(r)={}^nC_r p^r q^{n-r}$$



Poisson Distribution

Concept of Poisson Distribution

Poisson distribution can be viewed as a limiting form of Binomial distribution when n approaches infinity $(n \to \infty)$ and p approaches zero $(p \to 0)$ in such a way that their product is some fixed number (m), i.e., it remains constant. In other words, Poisson distribution is applicable when there are a number of random situations where the probability of a success in a single trial is small and the number of trials is large.

The Poisson distribution fits a very good model for use for determining probabilities associated with random variables where p is very small and n is very large, such as the number of calls coming into a telephone switch-board, the number of defects in a manufactured part, number of accidents, number of customers arriving at a service facility, number of radioactive particles decaying in a given interval of time, number of typographical errors per page in a typed material or the number of printing mistakes per page in a book, dimensional errors in engineering drawings, hospital emergencies, a defect along a long tape, number of defective rivets in an aeroplane wing, etc.

Poisson Distribution is a discrete probability distribution defined for all positive integers in which the probability of exactly r occurrences is given by :

$$p(r) = \frac{e^{-m} m^r}{r!}, \text{(for } r = 0, 1, 2, \text{..........)}$$

Normal Distribution

Concept of Normal Distribution

Normal or Guassian distribution is the most important continuous probability distribution in Statistics and is defined by the probability density function (or simply density function) :

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} ;$$

The normal distribution is the most versutile of all Theoretical distributions. It is found to be useful in statistical inferences, in characterising uncertainties in many real life processes, and in approximating other probability distributions. Quite often we face the problem of making inferences about processes based on limited data. Limited data is basically a sample from the full body of data is distributed, it has been found that the Normal Distribution can be used to characterise the sampling distribution. This helps considerably in Statistical Inferences. Heights, weight and dimensions of a product are some of the continuous random variables which are found to the normally distributed. This Knowledge helps us in calculating the probabilities of different events in varied situations, which in turn is useful for decision, making.

The normal distribution was discovered first by **De Moivre** in 1733. **Laplace** also knew about it almost at the same time. It is also associated with the name of **Gauss** and is known as **Gaussian Distribution** .

The graphical shape of normal distribution called the normal curve, is the bell shaped smooth symmetrical curve as shown in T curve is asymptotic in both directions to the x-axis and depends on the two parameters mean and standard deviation of the normal curve distribution.

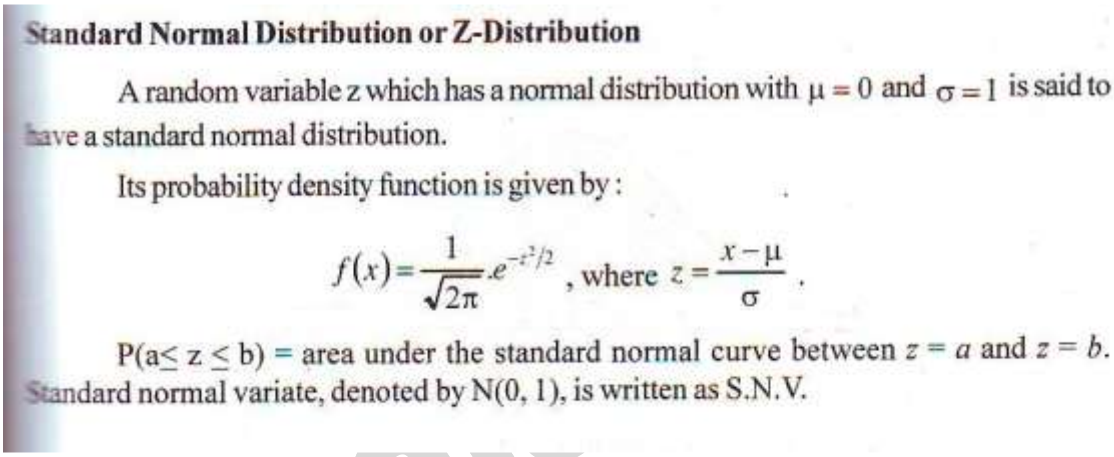
The equation of a normal curve is of exponential type. If X is a normal random variable with mean μ and standard deviation σ, then the equation of the normal curve is given by

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Where x can take any value in the range $(-\infty, +\infty)$ and π and e are two mathematical constants having approximate values, $\frac{22}{7}$ and 2.718 respectively.

μ and σ, the mean and standard deviation are known as parameter. The normal distribution with a mean μ and variance σ^2 may be denoted by the symbol $N(\mu, \sigma^2)$.

$p(x)$ is called the probability density function or simply density function.



Standard Normal Distribution or Z-Distribution

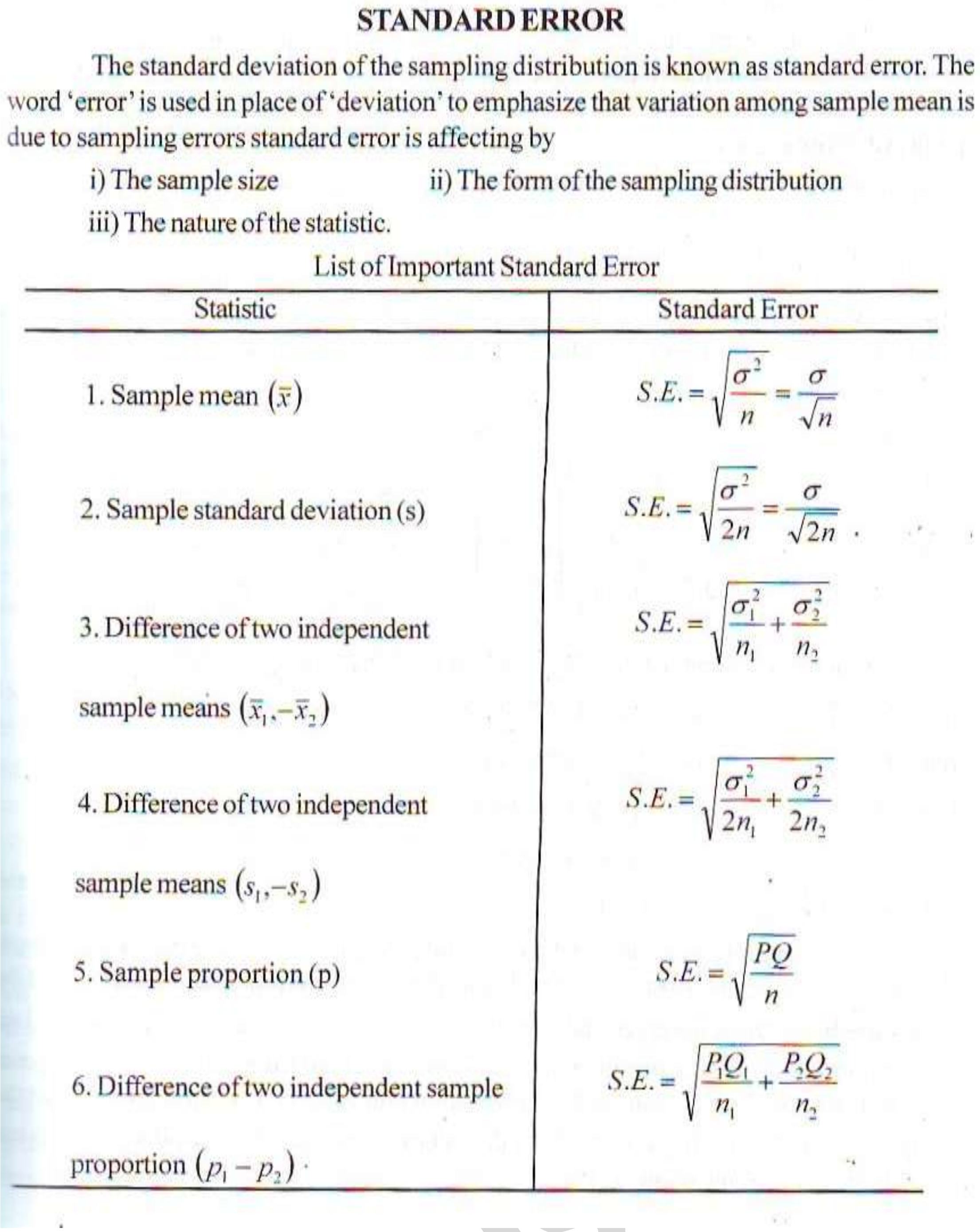
A random variable z which has a normal distribution with $\mu = 0$ and $\sigma = 1$ is said to have a standard normal distribution.

Its probability density function is given by :

$$f(x) = \frac{1}{\sqrt{2\pi}} . e^{-z^2/2} \text{ , where } z = \frac{x-\mu}{\sigma} .$$

$P(a \leq z \leq b)$ = area under the standard normal curve between $z = a$ and $z = b$. Standard normal variate, denoted by N(0, 1), is written as S.N.V.

Unit IV



STANDARD ERROR

The standard deviation of the sampling distribution is known as standard error. The word 'error' is used in place of 'deviation' to emphasize that variation among sample mean is due to sampling errors standard error is affecting by

i) The sample size ii) The form of the sampling distribution

iii) The nature of the statistic.

List of Important Standard Error

Statistic	Standard Error
1. Sample mean $(\bar{x})$	$S.E. = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$
2. Sample standard deviation (s)	$S.E. = \sqrt{\frac{\sigma^2}{2n}} = \frac{\sigma}{\sqrt{2n}}$
3. Difference of two independent sample means $(\bar{x}_1, -\bar{x}_2)$	$S.E. = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
4. Difference of two independent sample means $(s_1, -s_2)$	$S.E. = \sqrt{\frac{\sigma_1^2}{2n_1} + \frac{\sigma_2^2}{2n_2}}$
5. Sample proportion (p)	$S.E. = \sqrt{\frac{PQ}{n}}$
6. Difference of two independent sample proportion $(p_1 - p_2)$	$S.E. = \sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}$

STATISTICAL INFERENCE

Statistical Inference refers to the process of selecting and using a sample statistic to draw conclusion about the parameter of a population from which the sample is drawn statistical Inference broadly classified into following two heads.

i) Theory of Estimation

ii) Testing of Hypothesis

i) **Theory of Estimation :** It is used when no information is available about the parameters of the population from which the sample is drawn. Sample statistics (i.e. sample mean, variance etc.) are used to estimate the unknown population parameter (i.e. population mean, variance etc.) from which the sample drawn. It is divided into two groups.

i) Point Estimation

ii) Interval Estimation

In Point Estimation, a sample statistic is used to provide an estimate of the population parameter whereas in Interval Estimation, probable range is specified within which the true value of the parameter might be expected to lie.

ii) **Testing of Hypothesis :** It is used when some information is available about the population parameters from which the sample is drawn and it is required to test how for this information about the population parameters is tenable in the light of the information provided by the sample. The theory of testing hypothesis was given by **J.Nayman** and **E.S. Pearson**.

Meaning of Hypothesis : A hypothesis is an assumption about a population parameter to be tested. Assumptions or guesses are made about the population which may or may not be true, such Assumptions is known as hypothesis.

There are two types of hypothesis - Simple and Composite

i) **Simple Hypothesis :** A statistical hypothesis which specifies the population completely (i.e. the form of probability distribution and all parameters are known) is called a Simple Hypothesis.

ii) **Composite Hypothesis :** A statistical hypothesis which does not specify the population completely (i.e. either the form of probability distribution or some parameters remain unknown) is called a Composite Hypothesis.

Test of Hypothesis : The test of hypothesis discloses the fact whether the difference between the computed statistic and the hypothetical parameter is significant or otherwise. Hence the test of hypothesis is also known as the test of significance.

Null Hypothesis : A statistical hypothesis which is stated for the purpose of possible acceptance is known as null hypothesis. According to Prof. R.A. Fisher 'Null hypothesis is the hypothesis which is tested for possible rejection under the assumption that is true.'

Symbol : It is denoted by H_o

Setting up Null Hypothesis : The following steps must be taken into consideration while setting up a null hypothesis.

i) In order to test the significance of the difference between a sample statistic and the population parameter or between the two differnt sample statistics, we set up the null hypothesis H_0 that the difference is not significant. There may be some difference but that is solely due to sampling fluctuations.

ii) To test any statement about the population, we hypothesis that it is true. for example : If we want to find the population mean has a specified value u_0, then the null hypothesis H_0 is set as follows : $H_0 : \mu = \mu_0$

Acceptance : The acceptance of null hypothesis Implies that there is no significant difference between assumed and actual value of the parameter and that the difference occure is accidental arising out of fluctuations of sampling.

Rejection : It implies that there is significant difference between assumed and actual value of the parameter

Alternative Hypothesis : A hypothesis which is complementary to the null Hypothesis is called alternative hypothesis.

Symbol : It is denoted by H_1

Acceptance : Its acceptance depends on the rejection of the null hypothesis.

Rejection : Its rejection depends on the acceptance of the null hypothesis.

For Example : If null hypothesis that the population mean is 162 i.e., H_0 is = 162 an alternative alternative hypothesis may be any one of the following three :

i) $H_1 : \mu = 162$

ii) $H_1 : \mu > 162$

iii) $H_1 : \mu < 162$

Remark : H_0 and H_1 are mutually exclusive statements in the sense that both cannot hold good simultaneously. Rejection of one implies the acceptance of the other.

Level of Significance : Level of significance is the maximum probability of rejecting the null hypothesis when it is true or we can say that it is the maximum probability of making a **type I error** and it is denoted by α (alpha). It is usually expressed as % . Desired level of significance is always Fixed in advance before applying the test. Generally 5% or 1% level of significance is taken. Unless otherwise stated in the question, the students are advised to consider 5% level of significance.

For Example : 5% level of significance implies that there are about 5 chances in 100 of rejecting the H_0 when it is true or in other words, we are about 95% confident that we will make a correct decision.

Test Statistic : The next step is to compute suitable test statistic which is based on an appropriate probability distribution. It is used to test whether the null hypothesis set-up should be accepted or rejected.

Test - Statistic

i) Z-test	For test of Hypothesis involving large sample i.e. > 30
ii) t-test	For test of Hypothesis involving small sample i.e. ≤ 30
iii) χ^2-test	For testing the significant difference between observed frequencies and expected frequencies.
iv) F-test	For testing the sample variances.

The region of the standard normal curve corresponding to a pre-determined level of significance that is fixed for knowing the probability of making the type I error or rejecting the hypothesis which is true, is known as the "rejection region" or "critical region". The region of standard normal curve that is not covered by the rejection region, is called "accepted region". When the test statistic computed to test the hypothesis falls in the acceptance region, it is reasonable to accept the hypothesis as it is believed to be probably true.

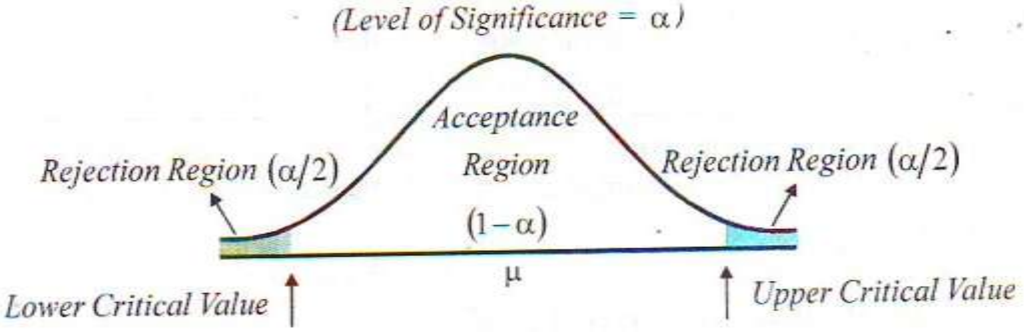
Two tailed test and one tailed test : The critical region may be shown by a portion of the area under the normal curve in two ways.

i) Two Tails

ii) One Tail (right tail or left tail)

i) **Two Tailed Test :** When the test of hypothesis is made on the basis of rejection region represented by both sides of the standard normal curve, it is called a two tailed test or two sides test for example :

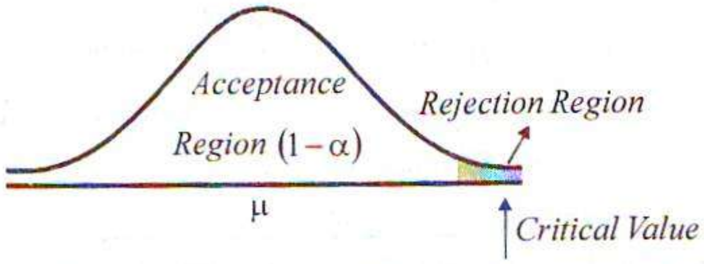
Null Hypothesis $(H_0) : \mu = 90$



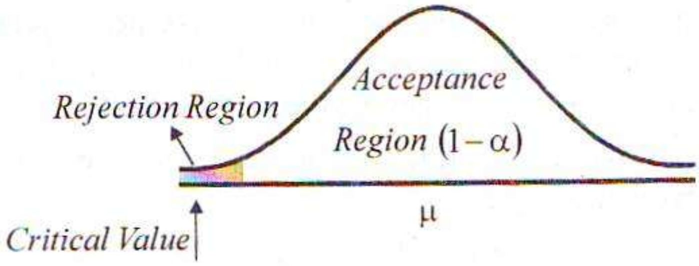
Alternative Hypothesis $(H_1) : \mu \neq 90$ (i.e. either $\mu > 90$ or $\mu < 90$)

i) **One Tailed Test :** The one tail test is used in cases where it is considered that the population mean is at least as large as some specified value of mean or at least as small as some specified value of mean. There are two types of one tailed tests

i) Right Tailed Test : In the right tailed test the rejection region or critical region lies entirely on the right tail of the normal curve.



i) Left Tailed Test : In the left tailed test the critical region or rejection region lies entirely on the left tail of the normal curve.



Actual	Decision	
	Accept H_0	Reject H_0
H_0 is True	Correct Decision (No Error) Probability $= 1 - \alpha$	Wrong **(Type I Error)** Probability $= \alpha$
H_0 is False	Wrong **(Type II Error)** Probability $= \beta$	Correct Decision (No Error) Probability $= 1 - \beta$

Relation Between Type I and Type II Error

i) The probability of making one type of error can be reduced only by allowing an increase in the probability of other type of error. The trade-off between these types of errors is made by assigning appropriate significance level after examining the costs or penalties attached to both type of errors.

ii) An increase in the sample size n will reduce the probability of committing both the types of errors simultaneously.

iii) The probability of committing a type I error, can always be reduce by adjusting the values of α.

iv) If the null hypothesis is false, β is a maximum when the true value of a parameter is close to the hypothesised value. The greater the distance between the true value and the hypothesised value, the smaller the β will be.

Procedure for testing of Hypothesis :

Step 1 : Set up the **null hypothesis.**

Step 2 : Set up the **alternative hypothesis.**

Step 3 : Identify the **sample statistic** to be used and its sampling distribution.

Step 4 : Test statistic : Define and compute the test statistic under H_0.

Step 5 : Specify the Level of significance such as 5% or 1%. If the level of significance is not specified in the question, generally 5% level is used.

Step 6 : Compute the value of test-statistic (e.g. Z, t, f, χ^2) used in testing.

Step 7 : Find the Critical Value of the Test Statistic used at the selected level of significance from the table of respective statistic distribution.

Step 8 : Specify the decision as follows :

a) **Acceptance :** Since the computed value is less than the critical value, we accept the null hypothesis (H_0) and conclude that difference is not significant and it could have arisen due to fluctuations of random sampling. **or**

b) **Rejection :** Since the computed value is greater than the critical value, we reject the null hypothesis (H_0) and conclude that the difference is significant and it could not have arisen due to fluctuations of random sampling.

UNIT-V

CORRELATION

Introduction

Correlation is a statistical tool & it enables us to measure and analyse the degree or extent to which two or more variable fluctuate/vary/change w.e.t. to each other.

For example – Demand is affected by price and price in turn is also affected by demand. Therefore we can say that demand and price are affected by each other & hence are correlated. the other example of correlated variable are –

While studying correlation between 2 variables use should make clear that there must be cause and effect relationship between these variables. for e.g. – when price of a certain commodity is changed (↑ or ↑) its demand also changed (↑ or ↑) so there is case & effect relationship between demand and price thus correlation exists between them. Take another eg. where height of students; as well as height of tree increases, then one cannot call it a case of correlation because neither height of students is affected by height of three nor height of tree is affected by height of students, so there is no cause & effect relationship between these 2 so no correlation exists between these 2 variables.

In correlation both the variables may be mutually influencing each other so neither can be designated as cause and the other effect for e.g. –

Price ↑→ Demand ↓
Demand ↓→ Price ↑

So, both price & demand are affected by each other therefore use cannot tell in real sense which one is cause and which one is cause and which one is effect.

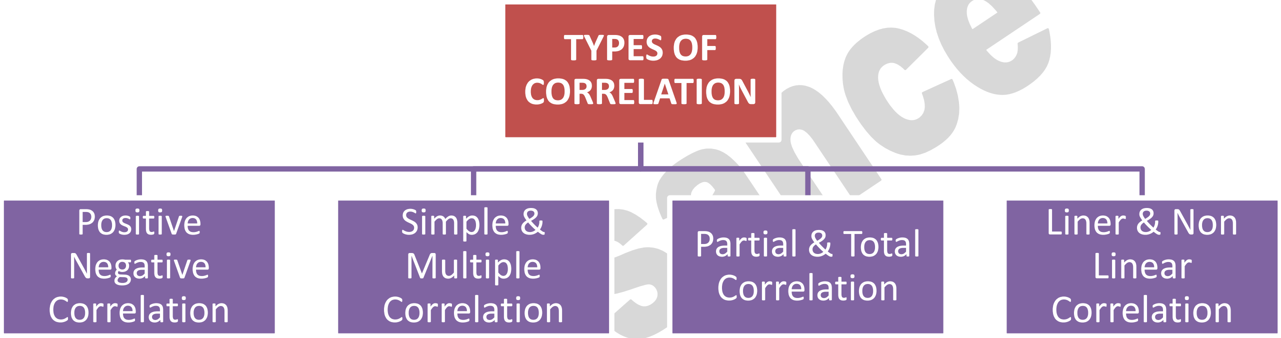
DEFINITIONS OF CORRELATION

"If 2 or more quantities vary is sympathy, so that movements is one tend to be accompanied by corresponding movements in the other(s), then they are said to be correlated". **Connor.**

"Correlation means that between 2 series or groups of data there exists some casual correction".
WI King

"Analysis of Correlation between 2 or more variables is usually called correlation." **A.M. Turtle**

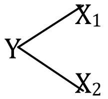
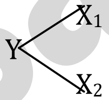
"Correlation analysis attempts to determine the degree of relationship between variables.
Ya Lun chou

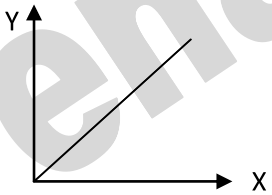
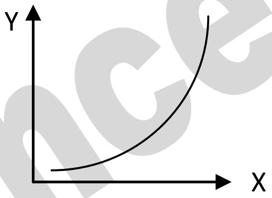
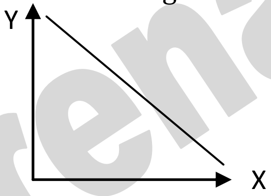
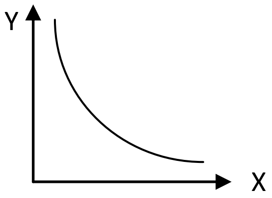


	POSITIVE CORRELATION	NEGATIVE CORRELATION
1	Value of 2 variables move in the same direction i.e. when increase/decrease in value of one variable will cause increase or decrease in value of other variable.	Value of 2 variables move in opposite direction i.e. when one variable increased, other variable decreases when one variable is decreased, other variable increase.
2	E.g. Supply & Price So, supply and price are …….correlated P = Price/Unit Q = quantity Supplied	E.g. Demand & Price So, Demand & Price vely correlated P = Price/Unit Q = quantity Supplied

	SIMPLE CORRELATION	MULTIPLE CORRELATION
1	In simple correlation, the relationship is confined to 2 variables only, i.e. the effect of only one variable is studied	The relationship between more than 2 variables is studied.
2	E.g. Demand & Price Demand depends on → Price This is case of simple correlation because relationship is confined to only one factor (that affects demand) i.e. price so we have to find correlation between demand & price. If, demand = Y If, demand – X	E.g. Demand & Price Demand depends on → Price Demand on → income This is case of multiple correlations because 2 factors (Price & Income) that affects demand are taken. We have to find correlation between demand & price. Demand & Price If, demand = Y

	Then, Correlation between Y & X	Price = X_1 Price = X_2 Then Correlation between Y & X_1 Correlation between Y & X_2

<table>
<tr><th>SIMPLE CORRELATION</th><th>MULTIPLE CORRELATION</th></tr>
<tr><td>In partial correlation though more than 2 factors are involved but correlation is studies only between to be constant.
E.g.

Y = Demand
X_1 = Price
X_2 = Income</td><td>In total correlation relationship between all the variables is studied i.e., none of item is assumed to be constant
E.g.

Y = Demand
X_1 = Price
X_2 = Income</td></tr>
<tr><td>If we study correlation between Y & X1 & assume X2 to be constant it is a case of partial correlation. this is what we do in law of demand – assume factors other than price as constant (Ceteris paribus – Keeping other things constant)</td><td>If we assume that income is not constant i.e. we study the effect of both price & income on demand, it is a case of total correlation.
In other words, cataris paribus assumption is relaxed in this case.</td></tr>
</table>

<table>
<tr><th></th><th>LINEAR CORRELATION</th><th>NON-LINEAR CORRELATION</th></tr>
<tr><td>1</td><td>In linear correlation, due to unit, change value of one variable there is constant change in the value of other variable. The graph for such a relationship is straight line. E.G. – If in a factory no of workers are doubled, the production output is also doubled, and correlation would be linear.</td><td>In non linear or curvilinear correlation, due to unit, change value of one variable, the change in the value of other variable is not constant. the graph for such a relationship is a curve. E.G. – The amount spent on advertisement will not bring the change in the amount of sales in the same ratio, it means the variation.</td></tr>
<tr><td>2</td><td>If the changed in 2 variables are in the same direction and in the constant ratio, it is linear positive correlation
<table><tr><td>X</td><td>Y</td></tr><tr><td>2</td><td>3</td></tr><tr><td>4</td><td>6</td></tr><tr><td>6</td><td>9</td></tr><tr><td>8</td><td>12</td></tr></table>
 </td><td>If the change in 2 variables is in the same direction but not in constant ratio, the correlation is non linear positive.
<table><tr><td>X</td><td>Y</td></tr><tr><td>50</td><td>10</td></tr><tr><td>55</td><td>12</td></tr><tr><td>60</td><td>15</td></tr><tr><td>90</td><td>30</td></tr><tr><td>100</td><td>45</td></tr></table>
 </td></tr>
<tr><td>3</td><td>If changes in 2 variables are in the opposite direction but in constant ratio, the correlation is linear negative. For eg. every 5% ↑ is price of a good is associated with 10% decrease in demand the correlation between price and demand would be linear negative.
<table><tr><td>X</td><td>Y</td></tr><tr><td>2</td><td>21</td></tr><tr><td>4</td><td>18</td></tr><tr><td>6</td><td>15</td></tr><tr><td>8</td><td>12</td></tr><tr><td>10</td><td>9</td></tr></table>
 </td><td>If changes in 2 variables are in opposite direction and not in constant ratio, the correlation is non linear negative. For eg: - every 5% ↑in price of good is associated with 20% to 10%↓in demand, the correlation between price & demand would be non linear negative.
<table><tr><td>X</td><td>Y</td></tr><tr><td>80</td><td>50</td></tr><tr><td>55</td><td>60</td></tr><tr><td>50</td><td>75</td></tr><tr><td>90</td><td>130</td></tr></table>
 </td></tr>
</table>

TYPE – 1 [BASED ON KARL PEARSON'S COFFICIENT OF CORRELATION]

Before use move to numerical, use understand the basic notions & concepts –

d_x = Deviations of xi value from mean = $(x_i - \bar{x})$

x = Mean of x value [Average of X values] = $\frac{\Sigma xi}{n}$

n = No. of observations

d_y = Deviation of y value from mean = $(y - \bar{y})$

$\bar{y}$ = Mean of y values = $\frac{\Sigma yi}{n}$

$d^2{}_x$ = Square of deviation of x values = $(x_i - \bar{x})^2$

$d^2{}_y$ = Square of deviation of x values = $(y_i - \bar{y})^2$

$d_x d_y$ = Product of deviations = $(x_i - \bar{x})(y_i - \bar{y})$

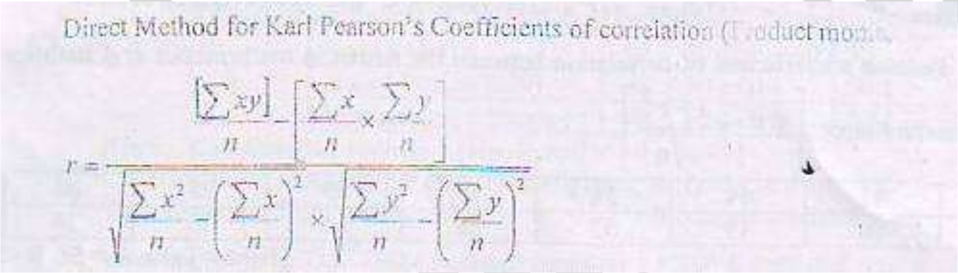
Covariance (x,y) = $\frac{\Sigma(xi - \bar{x})(yi - \bar{y})}{n}$

σ_x = Variance of x_i values = $\frac{\Sigma(xi - \bar{x})^2}{n}$

σ_y = Variance of y_i values = $\frac{(yi - \bar{y})^2}{n}$

r or r_{xy} = coefficient of correlation between x 7 y variables.

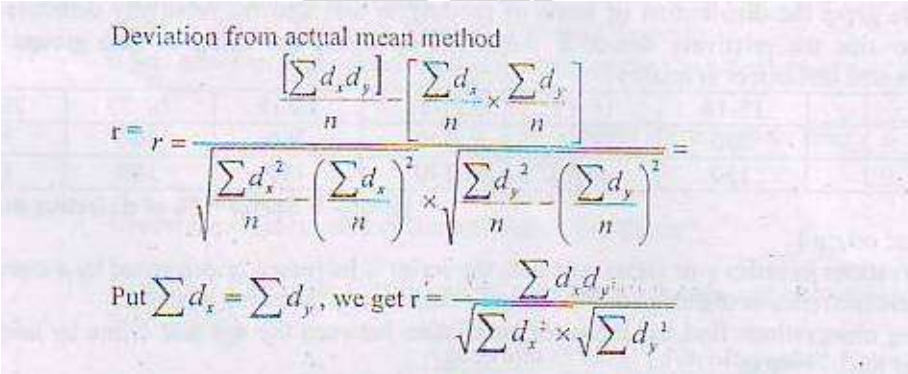
Direct Method for Karl Pearson's Coefficient of correlation



Direct Method for Karl Pearson's Coefficients of correlation (Product mo

$$r = \frac{\left[\frac{\sum xy}{n}\right] - \left[\frac{\sum x}{n} \times \frac{\sum y}{n}\right]}{\sqrt{\frac{\sum x^2}{n} - \left(\frac{\sum x}{n}\right)^2} \times \sqrt{\frac{\sum y^2}{n} - \left(\frac{\sum y}{n}\right)^2}}$$

Deviation from actual mean method

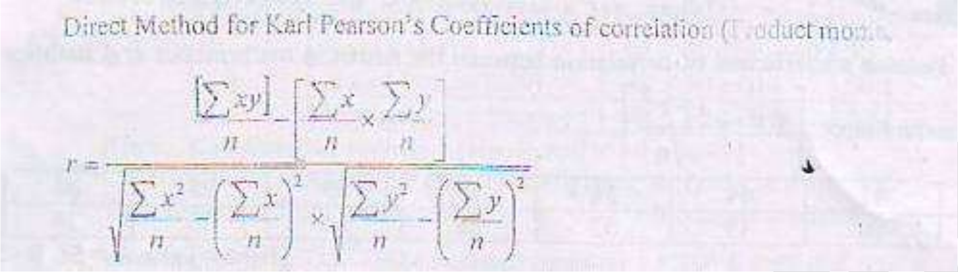


Deviation from actual mean method

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

Put $\sum d_x = \sum d_y$, we get $r = \frac{\sum d_x d_y}{\sqrt{\sum d_x^2} \times \sqrt{\sum d_y^2}}$

Deviation from assumed mean method (Short Cut Method)



Direct Method for Karl Pearson's Coefficients of correlation (Product mo

$$r = \frac{\frac{\left[\sum xy\right]}{n} - \left[\frac{\sum x}{n} \times \frac{\sum y}{n}\right]}{\sqrt{\frac{\sum x^2}{n} - \left(\frac{\sum x}{n}\right)^2} \times \sqrt{\frac{\sum y^2}{n} - \left(\frac{\sum y}{n}\right)^2}}$$

This method is used in the situation where mean of any series (x or y) is not in whole number, i.e. in decimal value. in this case it is advisable to take deviation from assumed mean rather than actual mean and then use the above formula.

In the above short cut method

Let, A = Assumed mean of X series

B = Assumed mean of y series

then $\Sigma d_x = \Sigma(x_i - A)$ & $\Sigma d_y = \Sigma(y_i - B)$ &

$\Sigma d_x{}^2 = \Sigma(x_i - A)^2$ & $\Sigma d_y{}^2 = \Sigma(y_i - B)^2$

$\Sigma d_x d_y = \Sigma(x_i - A)(x_i - B)$

REGRESSION ANALYSIS

The dictionary meaning of regression is "Stepping Back". The term was first used by a British Biometrician" Sir Francis Galton 1822 – 1911) is 1877. He found in his study the relationship between the heights of father & sons. In this study he described "That son deviated less on the average from the mean height of the race than their fathers, whether the father's were above or below the average, son tended to go back or regress between two or more variables in terms of the original unit of the data.

Meaning

Regression Analysis is a statistical tool to study the nature extent of functional relationship between two or more variable and to estimate the unknown values of dependent variable from the known values of independent variable.

Dependent Variables – The variable which is predicted on the basis of another variable is called dependent or explained variable (usually devoted as y)

Independent variable – The variable which is used to predict another variable called independent variable (denoted usually as X)

Definition

Statistical techniques which attempts to establish the nature of the relationship between variable and thereby provide a mechanism for prediction and forecasting is known as regression Analysis.

– Ya-lun-Chon"

Importance/uses of Regression Analysis

- Forecasting
- Utility in Economic and business area
- Indispensible for goods planning
- Useful for statistical estimates.
- Study between more than two variable possible
- Determination of the rate of change in variable
- Measurement of degree and direction of correlation
- Applicable in the problems having cause and effect relationship

- Regression Analysis is to estimate errors
- Regression Coefficient (b_{xy} & b_{yx}) facilitates to calculate of determination ® & coefficient or correlation (r)

Regression Lines

The lines of best fit expressing mutual average relationship between two variables are known as regression lines – there are two lines of regression

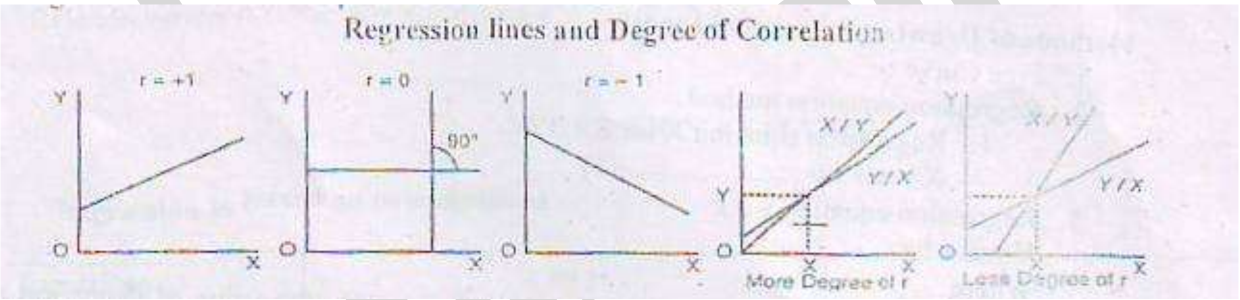
Why are two Regression lines –

1. While constructing the lines of regression of x on y is treated as independent variables where as 'x' is treated as treated as dependent variable. This gives most probable values of 'X' for gives values of y. the same will be there for y on x.

RELATIONSHIP BETWEEN CORRELATION & REGRESSION

1. When there is perfect correlation between two series (r = ± 1) the regression with coincide and there will be only one regression line.
2. When there is no correction (r = o)> Both the lines will cut each other at point.
3. Where there is more degree of correction, say (r = ± 70 or more the two regression line with be next to each other whereas when less degree of correction. Say (r=± 10 on less) the two regression line will be a parted from each other.

REGRESSION LINES AND DEGREE OF CORRELATION



DIFFERENCE BETWEEN CORRELATION AND REGRESSION ANALYSIS

The correlation and regression analysis, both, help us in studying the relationship between two variables yet they differ in their approach and objectives. The choice between the two depends on the purpose of analysis.

S.NO	BASE	CORRELATION	REGRESSION
1	MEANING	Correlation means relationship between two or more variables in which movement in one have corresponding movements in other	Regression means step ping back or returning to the average value, i.e., it express average relationship between two or more variables.
2	RELATIONSHIP	Correlation need not imply cause and effect relationship between the variables under study	Regression analysis clearly indicates the cause and effect relationship. the variable(s) constituting causes(s) is taken as independent variables(s) and the variable constituting the variable consenting the effect is taken as dependent variable.
3	OBJECT	Correlation is meant for co-variation of	Regression tells use about the

		the two variables. the degree of their co-variation is also reflected in correlation. but correlation does not study the nature of relationship.	relative movement in the variable. We can predict the value of one variable by taking into account the value of the other variable.
4	NATURE	There may be nonsense correlation of the variable has no practical relevance	There is nothing like nonsense regression.
5	MEASURE	Correlation coefficient is a relative measure of the linear relationship between X and Y. It is a pure number lying between 1 and +1	The regression coefficient is absolute measure representing the change in the value of variable. We can obtain the value of the dependent variable.
6	APPLICATION	Correlation analysis has limited application as it is confined only to the study of linear relationship between the variables.	Regression analysis studies linear as well as non linear relationship between variables and therefore, has much wider application.

Why least square is the Best?

When data are plotted on the diagram there is no limit to the number of straight lines that could be drawn on any scatter diagram. Obviously many lines would not fit the data and disregarded. If all the points on the diagram fall on a line, that line certainly would the best fitting line but such a situation is rare and ideal. Since points are usually scatters, we need a criterion by which the best fitting line can be determined.

Methods of Drawing Regression Lines –

1. Free curve –
2. Regression equation x on y,
   X = a + by ....(1)
3. Regression equation y on x
   Y = a + bx
   Where
   ‘a’ is that point where regression lines touches y axis (the value of dependent variable value when value or independent variable is zero)
   ‘b’ is the slop of the said line (The amount of change in the value of the dependent variable per unit change)

Change in independent variable)
A and b constants can be calculated through –

$\Sigma(x = a + by)$ (by multiplying ‘Σ’)
$\Sigma x = Na + b\Sigma y$ (1)

$\Sigma x\ (y = a + bx)$ (by multiplying Σx)
$\Sigma xy = \Sigma xa + b\Sigma x^2$ (2)

KINSDS OF REGRESSION ANALYSIS

1. Linear and Non- Linear Regression
2. Simple and Multiple Regression

FUNCTIONS OF REGRESSION LINES –

1. To make the best estimate –
2. To indicate the nature and extent of correlation

REGRESSION EQUATIONS –

The regression equation's express the regression lines, as there are two regression lines there are two regression equations –
Explanation is given in formulae –

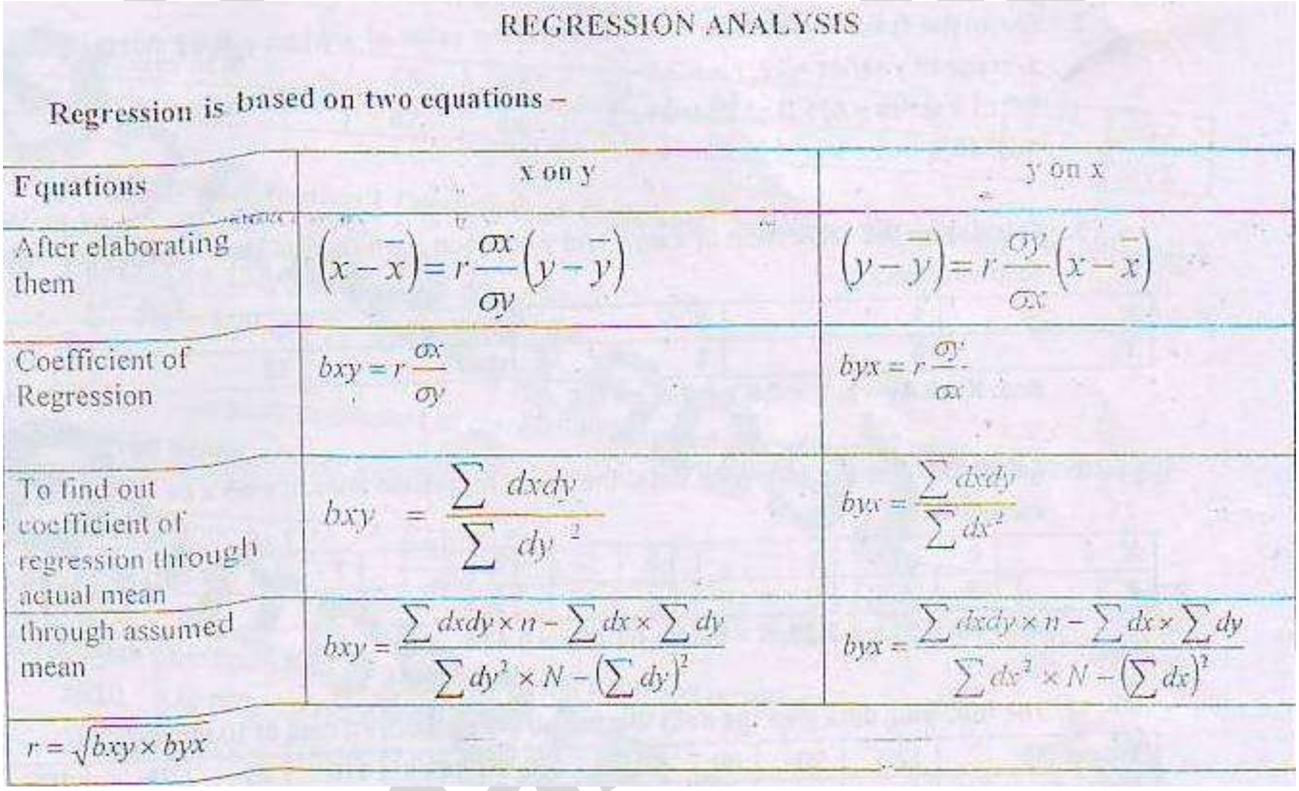
REGRESSION LINES

1. Regression equation of x on y
   $X - X = b_{xy} (y - y)$
   Where b_{xy} = regression coefficient of X on Y
2. Regression equation of y on x
   $Y - Y = b_{xy} (x - x)$ where b_{xy} = regression coefficient of Y on X

REGRESSION COEFFICIENT – There are two regression coefficient like regression equation, they are (b_{xy} and b_{yx})

Properties of regression coefficients –

- Same sign – Both coefficient have the same either positive on negative
- Both cannot by greater than one – If one Regression is greater than "One" or unity. Other must be less than one.
- Independent of origin – Regression coefficient are independent of origin but not of scale.
- A.M.> 'r' – mean of regression coefficient is greater than 'r'
- R is G.M. – Correlation coefficient is geometric mean between the regression coefficient
- R, b_{xy} and b_{xy} – They all have same sign



REGRESSION ANALYSIS

Regression is based on two equations –

Equations	x on y	y on x
After elaborating them	$\left(x-\bar{x}\right)=r\frac{\sigma x}{\sigma y}\left(y-\bar{y}\right)$	$\left(y-\bar{y}\right)=r\frac{\sigma y}{\sigma x}\left(x-\bar{x}\right)$
Coefficient of Regression	$bxy=r\frac{\sigma x}{\sigma y}$	$byx=r\frac{\sigma y}{\sigma x}$
To find out coefficient of regression through actual mean	$bxy = \frac{\sum dxdy}{\sum dy^2}$	$byx=\frac{\sum dxdy}{\sum dx^2}$
through assumed mean	$bxy=\frac{\sum dxdy\times n-\sum dx\times\sum dy}{\sum dy^2\times N-\left(\sum dy\right)^2}$	$byx=\frac{\sum dxdy\times n-\sum dx\times\sum dy}{\sum dx^2\times N-\left(\sum dx\right)^2}$
$r=\sqrt{bxy\times byx}$		

Unit VI

TIME SERIES

"A Time Series" is a series of statistical data recorded in accordance with their time of occurrence. Here it is noted that it is a set of observation taken at specified times usually (but not always) at equal intervals. Thus a set of data depending on the time (which may be year, quarter, month, day etc.) is called a "Time Series".
Today the use of time series analysis is not merely confined to economists and businessmen, but it extensively used by scientists, sociologist, biologists, geologists, research workers etc.

Some example of time series are

(i) The population of a country in different years.
(ii) The annual production of coal in India over the last ten years.
(iii) Deposits received by bank in a year.
(iv) The daily closing price of a share in the Bombay Stock Exchange.
(v) The monthly sales of departmental store for the last six months.
(vi) Hourly temperature recorded by the store for the last six months.

According to Patterson "A time series consists of statistical data which are collected. Recorded or observed over successive increments.

Utility or importance of Time Series
The very important use of time series analysis is its use in forecasting future information and behavior.

(i) It enables us to predict or forecast the behavior of the phenomenon in future, which is very essential for business planning. On the basis of past information, the trend can be estimated and projections can also be made for the uncertain future. It assists in reducing, the risk and uncertainties of business and industry.

(ii) It helps in the evaluation of current achievement by review and evaluation of progress made through a plan can be done on the basis of time series.

(iii) It helps in the analysis of past behavior of the phenomenon under consideration. What changes had taken place in the past, what factor were responsible for these changes, under that conditions these changes took place, etc. are certain issues which could be studied and analyzed by time series.

(iv) It helps in making comparative studies in the values of different phenomenon at different times or place. It provides a scientific basis for making comparison by studying and isolating the effects of various components of a time series.

(v) The segregation and study of the various components of time series is of paramount importance to a businessman in the planning of future operations and the formulation of executive and policy decisions.

(vi) On the basis of the past performance of the various sectors of economy, we can determine future requirements and a suitable policy can be formulated to get desired and predetermined objectives.

Causes of variation in time series

If the values of a phenomenon are observed at different periods of time, the values so obtained will show appreciable variations.
The following factors are generally affecting any time series are :

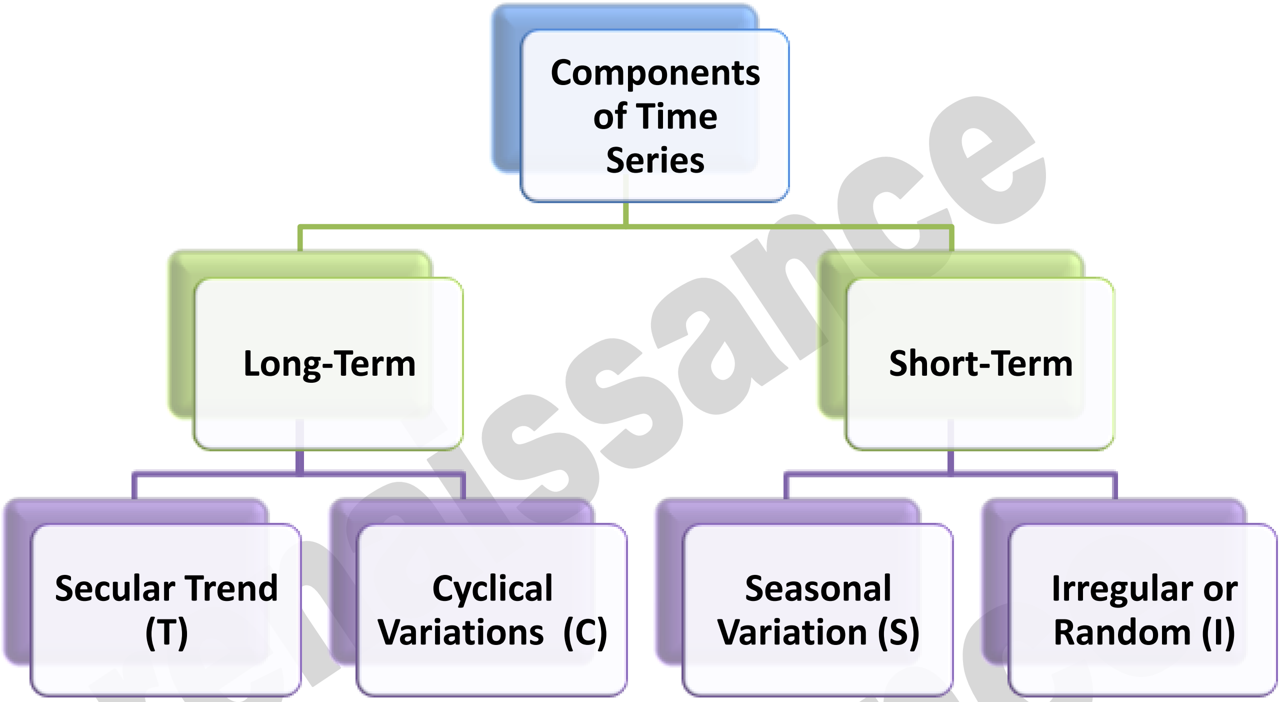
(i) Changing of tastes, habits and fashions of the people.
(ii) Changing of customs, conventions of the people.
(iii) Rituals and festivals.
(iv) Political movements, government policies.
(v) War, Famines, Drought, Flood, Earthquakes and Epidemic etc.
(vi) Unusual weather or seasons.

Components of Time Series

A time series may be defined as a collection of readings belonging to different time periods of some economic variable or composite of variable.
Eg. The retail price of a particular commodity are influenced by a number of factors namely the crop yield which further depends on weather conditions, irrigation facilities, fertilizers used, transportation facilities, consumer demand etc.

The various forces affecting the values of a phenomenon in a time series may be broadly classified into the following four categories, commonly known as the components of a time series.

i. Secular Trend (i.e. long-term smooth, regular movement)
ii. Seasonal Variation (periodic movement, the period being not greater than one year)
iii. Cyclical Variation (periodic movement with period greater than one year)
iv. Irregular or Random Variation.



1. **Secular Trend**: - It is the matter of common sense that there might be violent variations in a time series during a short span of time, however in a long run, it has a tendency either to rise or fall. This tendency or trend of variation may be either upward or down set on over a long time period. This is known as 'Secular trend' or 'Simple trend. It is but natural that population growth, Technological progress medical facilities production, prices etc. are not judge over a day, month or year they shores. The movement are upward, downward or constant over a fairly long period.

Broadly the trends are divided under two heads:

1. Linear Trends: - If we plot the values of time series on graph it shows the straight line i.e. growth rate is constant. Although in practice linear trend is commonly used but it is rearely found in economics and business data.

2. Non-Linear Trends: In business or economics generally growth is slow in the begging and them it is rapid for some time period after which it becomes stable for some time period and finally retards gradually. It is not linear it forms a curve known as non linear trends.

Seasonal Variation: As we read season the first things comes in our mind is spring, summer, autumn and winter. Generally seasonal variations occur due to changes in weather condition, customer, tradition fashion etc.

Seasonal variations represent a periodic movement where the period is not longer than one year. The factors, which mainly cause this type of variation in time series, are the climatic changes of the different seasons. For example

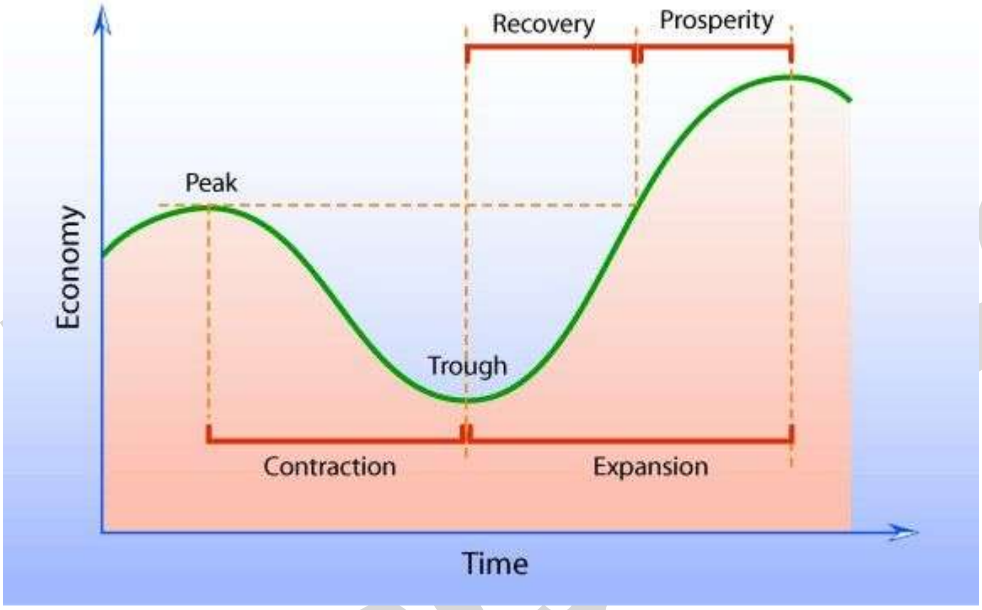
(i) Sale of woolens go up in winter.
(ii) Sale of raincoat and umbrella go up in rainy season.
(iii) Prices of food grains decrease with the arrival of new crop.
(iv) Sale of cooler, refrigerator etc. rise during the summer season.

Another variation occurs due to man-made convention and customs, which people follow at different times like Durga Pooja, Dashehra, Deepawali, Ide. X-Max etc. The seasonal variations may take place per day per week or per month. For example:

(i) Sale of departmental stores go up in festivals.

(ii) Sale of cloths and Jewelry pick up in marriages.
(iii) Sale of Paint, furniture and electronics goes up during festivals like, Deepawali, Ide, X-max etc.
(iv) Sale of vehicles increase considerably during Durga Pooja and Dasherhra.

Cyclical Variations: Most of the business activities are often characterized by recurrence of periods of prosperity and slump constituting a business cycle. Cyclical variations are another type of periodic movement, with a period more than one year. Such movements are fairly regular and oscillatory in nature. One complete period is called a 'cycle' cyclical variations are not as regular as seasonal variation, but the sequence of changes, marked by prosperity, decline, depression and recovery, remains more of less regular.



PHASES OF A BUSINESS CYCLE

Irregular or Random Variation: Irregular or random variation are such variation which are completely unpredictable in character. These are caused by factors which are either wholly unaccountable or caused by such unforeseen events like Earthquakes, flood, drought famines, epidemic etc, and some man-made situations like strikes lock-outs wart etc.

Mathematical Models for Analysis of Time Series

Though there are many models by which a time series can be analyzed, two models commonly used for decomposition of a time series into various components are

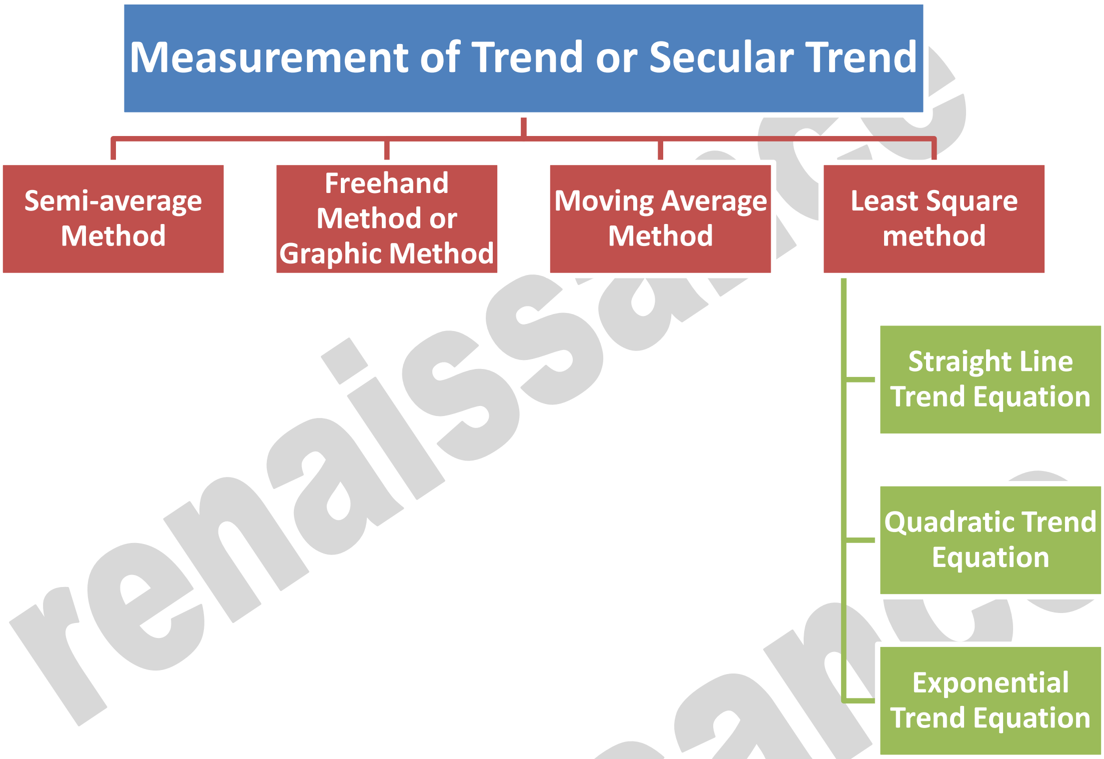
1. **Additive Model: -** According to the additive model, the decomposition of time series is done on the assumption that the effect of various components are additives in nature, i.e. $\mathbf{U = T+S+C+R}$
Where, U, is the time series value and T, S, C, and R stand for trend seasonal, cyclical and random variation.
In this model 'S, C and R are absolute quantities and can have positive or negative values. The model assumes that the four components of the time series are independent of each other and non-has any affect whatsoever on the remaining three components.

2. **Multiplication Model:** According to the multiplication model, the decomposition of a time series on the assumption that the effects of the four components of a time series (T, S, C and R) are not necessarily independent of each other. In fact, the model presumes that their effects are interdependent

$$\mathbf{U = T \times S \times C \times R}$$

Measurement of Trend or Secular Trend

The different methods of determining the trend component of a time series are:



1. Moving Average Method: Moving average method is very commonly used for the isolation of trend and in smoothing out fluctuations in time series. In this method, a series of arithmetic means of successive observation, known as moving averages, as calculated from the given data, and these moving average are used as trend values. Yearly moving average is given by :

$$\frac{a+b+c}{3} \qquad \frac{b+c+d}{3} \qquad \frac{c+d+e}{3} \qquad \frac{d+e+f}{3}$$

Illustration1 Calculate 3 yearly moving averages:

Years	:	1979	1980	1981	1982	1983	1984	1985	1986
Earning(Lakhs)	:	80	90	70	60	110	50	40	30

Working Rule

(i) Add the values of the first3 years (namely 1979, 1981 i.e., 80+90+70=240) and place the total against the middle year1980.

(ii) Leave the first year's value and add up the values of the next 3 years (i.e., 1980, 1981, 1982, viz., 90+70+70+60 = 220) and place the total against the middle year i.e., year 1981.

Illustration2 Calculate 5 yearly moving averages and seven year moving average for the following data:

Year	:	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990
Sales ('000 Rs.)	:	123	140	110	98	104	133	95	105	150	135

Calculation of Moving Averages when the Period is Even:

If the period of the moving average is even, centre point of the group will lie between two years. It is, therefore, necessary to adjust or shift (technically known as centre) these average so that they coincide with the years. For example

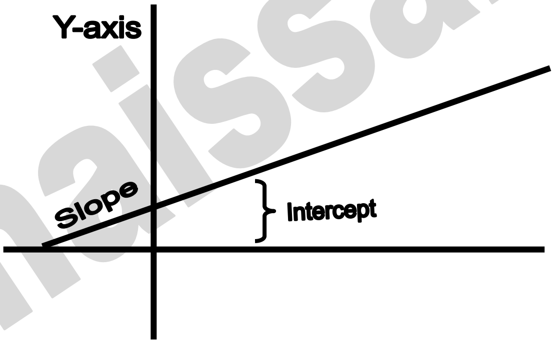
4-yearly moving average is calculated as:
Step 1 : Add the values of first four year, and place the total between the 2nd and 3rd year.
Step 2 : Leave the first year value and then add the for values of the next four years and place the total in between the 3rd and 4th year **Continue this process until the last year is taken into account.**
Step 3 : Divide 4 yearly moving totals 4. It will give **4 yearly moving average.**
Step 4 : Add first two moving averages and divide it by 2 to get the **moving average centered.** Place it against 3rd year. Leave the first moving average and then add next two moving average and divide by 2 to get the next moving average centered. Place it against the 4th Year. Continue this process till the last moving average is included.
Alternative Procedure: In this procedure step 1 and 2 are same as above.
Step 3: Add first two 4 yearly moving total place it against 3rd year. Leave the first moving total and then add next two moving total to get the next moving total centred. Place it against the 4th year. Continue this process till the last moving total is included.
Step 4 : Diving these centered moving totals by 8. It will give 8 yearly moving average. This procedure will more clear by following illustration.

Illustration Construction a four-yearly centered moving average from the following data :

Year	:	1970	1975	1980	1985	1990	1995	2000
Imported Cotton (in '000)	:	129	131	106	91	95	84	93

Method of Least Squares

It is an appropriate mathematical technique to determined an equation which best fits on a given observation relating to two variables. In this procedure for fitting a live to a set of observation the sum of the squared deviations between the calculated and observed values in minimised. Therefore the technique is named as "Least-Squares method." And the line so obtained is known as 'Best fit line".
We know that sum of the deviations from the arithmetic mean is zero. Therefore the sum of the deviations from the line of the best fit is zero.
(i) $\sum(y - c) = 0$, i.e., the sum of the deviations of the actual values of y and computed values of y is zero.
(ii) $\sum(y - y)2$ =is least, i.e., the sum of the squares of deviations from the actual and the computed value of y is least.
That is why it is called the method of least squares and the line obtained by this method is called the 'line of best fit'
This method may be used either to fit a straight line trend or parabolic trend straight line trend is represented by the equation y= a + bx where y represents the estimated values of the trend x represents the deviations in the time period. A and b are constants.
'a' represents intercept of the line of the y no is and 'b' represent the slope of the line i.e. it gives the changes in the value of y for per unit change in the value of x if b>0 it show and growth rate and if b<0 it shows decline rate.



Merits:

1. This is the only method of measuring trend which provides the future values authentically very convincing and reliable.

2. This method is used for forecasting the series for example.
3. If other factors are not so effective no share market, this method can provide very reliable information about the movement of the share of a company.
4. This method has no scope for personal bias of the Investigator.
5. It is only method which gives the rate of growth per annum.

Demerits:-

1. The method required mathematical ability. Some items it involves tedious and complicated calculations.
2. The method has no flexibility i.e. if even a single term is added to series it makes necessary to do all the calculations again.
3. Estimations and predictions by this method are based only on long term variations and the impact of cyclical, seasonal and irregular variations are completely ignored.

Computation of Trend Values by the Least Squares Method

We know straight lines trend is given by y= a+bx in order to determine the values of the constants and b the following two normal equations are to be solved.

$$\sum Y=na+b\sum X$$

$$\sum XY = a\sum X+b\sum X^2$$

Where n represents number of years (months or any other period) for which data are given:
$\sum y$ sum of actual values of y variable.
$\sum y$ represents sum of deviations from the origin.
$\sum y\ x^2$ represents sum of deviations from the origin.
$\sum xy$ represents sum of the deviations from the origin and actual values.

Remarks :- The variable x can be measured from any point of time as origin. But if middle time period is taken as origin and deviations are taken from the middle time period it provides $\sum x = 0$ the above normal equation would be reduced to the

$\sum y = na + \sum x \rightarrow \sum y = na+0=na \rightarrow$ Thus a = $\frac{\sum Y}{n}$

$\sum xy = a\sum x + b\sum x^2+0+\sum bx^2 = \sum xy = b\sum = x^2$ Thus b = $\frac{\sum xy}{\sum = x^2}$