SYLLABUS

Class – B.Com 2nd Year

Subject –Business Statistics

UNIT – I	Meaning and Definition of Statistics, Statistical Investigations, Laws of Statistics, Scope of Statistics, Limitations of Statistics.
UNIT II	Collection of Data, Presentation of Data, Frequency Distribution, Primary and Secondary Data.
UNIT – III	Measures of Central Tendencies: Mean, Median, Mode, Geometric Mean, Harmonic Mean.
UNIT IV	Measure of Variation: Standard Deviation, Mean Deviation and Skewness, Time Series Analysis.
UNIT – V	Correlation Analysis, Karl Pearson's Coefficient of Correlation, Spearman's Rank Correlation, Regression, Lines of Regression, Index Number.

UNIT — I
Introduction to STATISTICS

UNIT-1

Meaning of Statistics:

The word "Statistics" of English language has either been derived from the Latin word status or Italian word statistics and meaning of this term is "An organised political state.

Meaning: The science of collecting, analysing and interpreting such data or Numerical data relating to an aggregate of individuals.

E.g:- Statistics of National Income, Statistics of Automobile Accidents, Production Statistics, etc.

Statistics = Facts about Data

Definitions of Statistics:

The definition of Statistics is classified into two categories-

1) Statistics as Numerical Data:

"Statistics when used as plural, statistics means numerical set of data".

WHAT EXPERTS SAY ABOUT STATISTICS — SOME DEFINITIONS "STATISTICS AS NUMERICAL DATA"

1. "*Statistics are the classified facts representing the conditions of the people in a State…specially those facts which can be stated in number or in tables of numbers or in any tabular or classified arrangement*."—Webster.

2. "*Statistics are numerical statements of facts in any department of enquiry placed in relation to each other*."— Bowley.

3. "*By statistics we mean quantitative data affected to a marked extent by multiplicity of causes*".—Yule and Kendall.

4. "*Statistics may be defined as the aggregate of facts affected to a marked extent by multiplicity of causes, numerically expressed, enumerated or estimated according to a reasonable standard of accuracy, collected in a systematic manner, for a predetermined purpose and placed in relation to each other*."—Prof. Horace Secrist.

2. Statistics as Statistical Methods:

" Statistics when used in singular sense it means the science of statistical methods embodying the theory and techniques used for collecting, analysing and drawing inferences from the numerical data".

WHAT EXPERTS SAY ABOUT STATISTICS — SOME DEFINITIONS "STATISTICS AS STATISTICAL METHODS"

1. *Statistics may be called the science of counting*. —Bowley A.L .
2. *Statistics may rightly be called the science of averages*. —Bowley A.L.
3. *Statistics is the science of the measurement of social organism, regarded as a whole in all its manifestations*. —Bowley A.L .
4. "Statistics is the science of estimates and probabilities." —Boddington
5. *"The science of Statistics is the method of judging collective, natural or social phenomenon from the results obtained from the analysis or enumeration or collection of estimates."*—King
6. *Statistics is the science which deals with classification and tabulation of numerical facts as the basis for explanation, description and comparison of phenomenon."*—Lovin
7. "*Statistics is the science which deals with the methods of collecting, classifying, presenting, comparing and interpreting numerical data collected to throw some light on any sphere of enquiry."*—Selligman
8. *"Statistics may be defined as the science of collection, presentation, analysis and interpretation of numerical data*." —Croxton and Cowden
9. *"Statistics may be regarded as a body of methods for making wise decisions in the face of uncertainty."*—Wallis and Roberts
10. *"Statistics is a method of decision making in the face of uncertainty on the basis of numerical data and calculated risks."*—Prof. Ya-Lun-Chou
11. *"The science and art of handling aggregate of facts—observing, enumeration, recording, classifying and otherwise systematically treating them."*—Harlow

Nature /Features /Characteristics of statistics :

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

Division of Statistics

- ~~*Theoretical*~~:Mathematical theory which is the basis of the science of statistics is called theoretical statistics
- ~~*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:- (1) Collection of Data (2) Classification (3) Tabulation (4) Presentation (5) Analysis (6) Interpretation (7) Forecasting.
- *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.

Laws of Statistics

1. Law of Statistical Regularity

This law is derived from the mathematical theory of probability. It states that a **moderately large number of items chosen at random** from a large group are almost sure, on average, to possess the characteristics of the large group.

The Core Idea: You don't need to check every single customer to understand your market. If you pick a representative sample, the "average" behavior of that sample will mirror the "average" behavior of the whole population.

Conditions for Success:

Randomness: Every item must have an equal chance of being picked.

Size: The sample shouldn't be too small (e.g., asking two people isn't enough).

2. Law of Inertia of Large Numbers

This is a corollary (a natural consequence) of the first law. It states that **larger groups of data show higher stability** compared to smaller groups.

The Core Idea: While individual items might be unpredictable or "erratic," the aggregate (the total) stays relatively constant.

Business Example: A single person's death is unpredictable for an insurance company. However, the number of deaths among 1,000,000 policyholders stays very stable year-over-year. This "inertia" allows companies to calculate premiums accurately.

Why it happens: Movements in opposite directions tend to cancel each other out. If one store's sales drop, another is likely to be rise, keeping the total company revenue stable.

Summary Table for Quick Revision

Feature	Law of Statistical Regularity	Law of Inertia of Large Numbers
Focus	Relationships between Sample and Population.	Stability of large aggregates.
Main Point	A random sample represents the whole.	Large data is more stable than small data.
Business Use	Market Research, Quality Control.	Insurance, Macroeconomics, Risk Management.

Scope of Statistics

1. Statistics and state or govt.
2. Statistics and business or management.
    - Marketing
    - Production
    - Finance
    - Banking
    - Control
    - Research and Development
    - Purchases
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 of Statistics:

1. It is not deal with items but deals with aggregates.
2. Only on expert can use it.
3. It is not the only method to analyse the problem.
4. It can be misused etc.

Statistical Investigation

Meaning: In general it means as a statistical survey. In brief. Scientific and systematic collection of data and their analysis with the help of various statistical method and their interpretation.

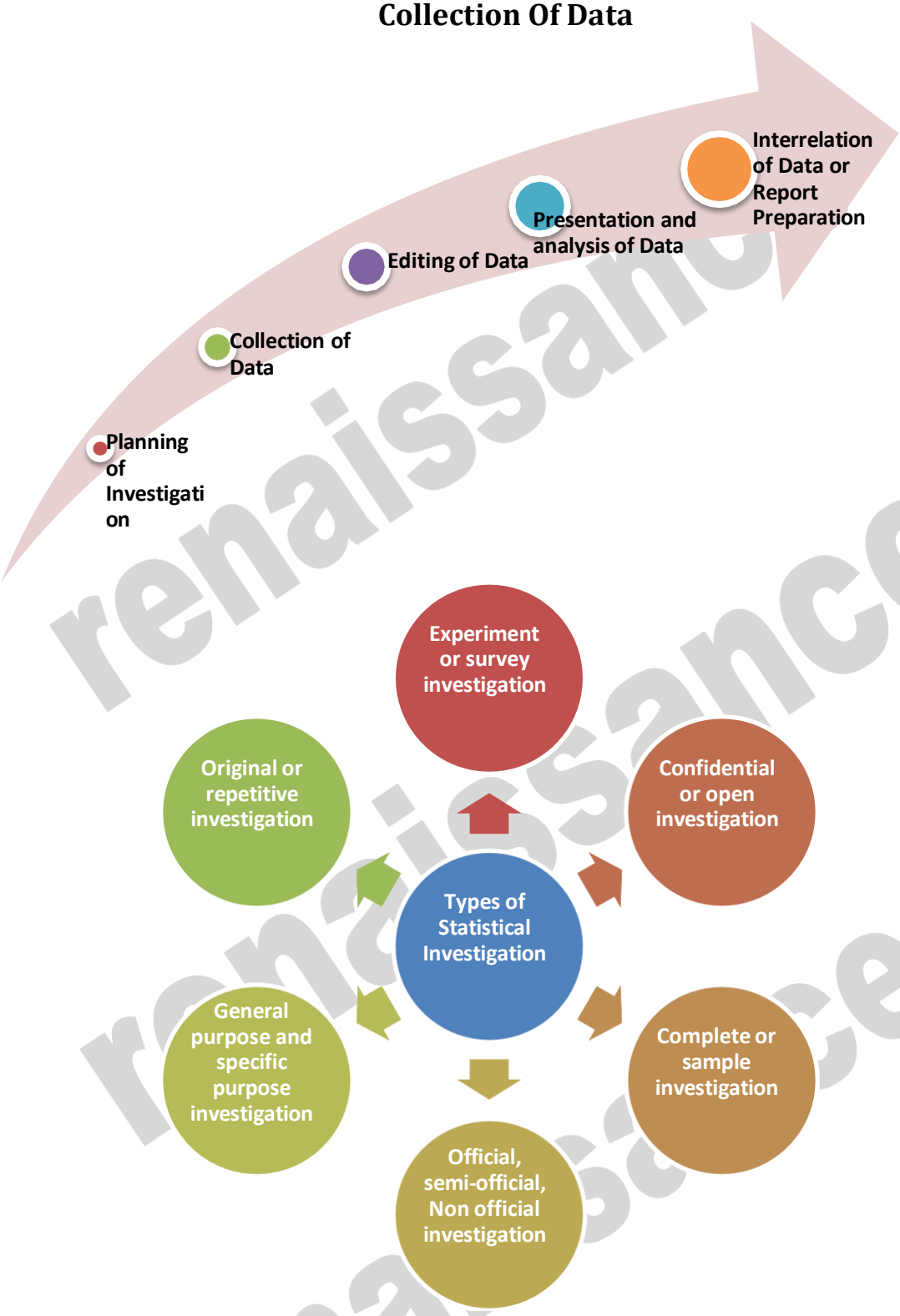
Stages of Statistical Investigation:-

1. Planning of Investigation
2. Collection of Data
3. Editing of Data
4. Presentation of Data
    - Classification
    - Tabulation
    - Diagrams
    - Graphs
5. Analysis of Data
6. Interrelation of Data or Report Preparation.

Types of Statistical Investigation:-

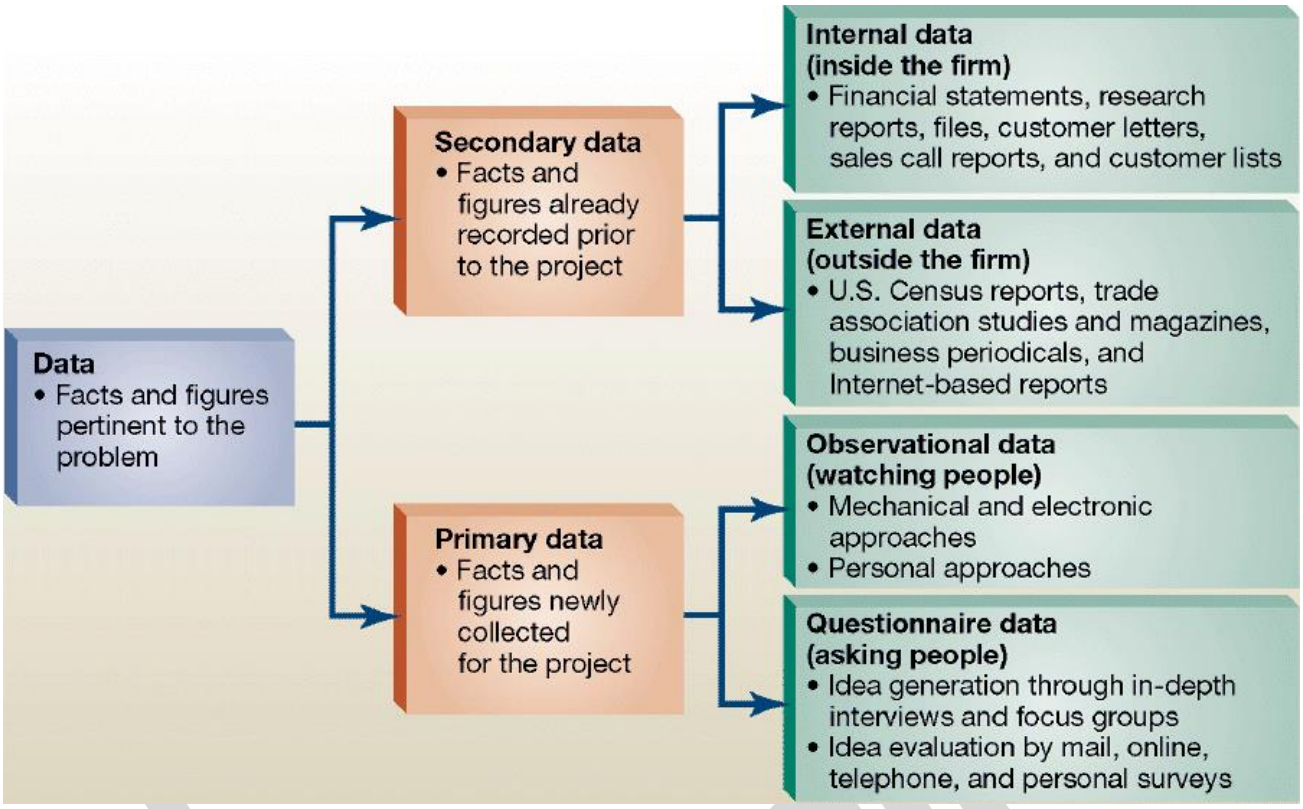
1. Experiment or survey investigation
2. Complete or sample investigation
3. Official, semi-official, Non official investigation
4. Confidential or open investigation
5. General purpose and specific purpose investigation
6. Original or repetitive investigation.

UNIT — II
Collection Of Data



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.

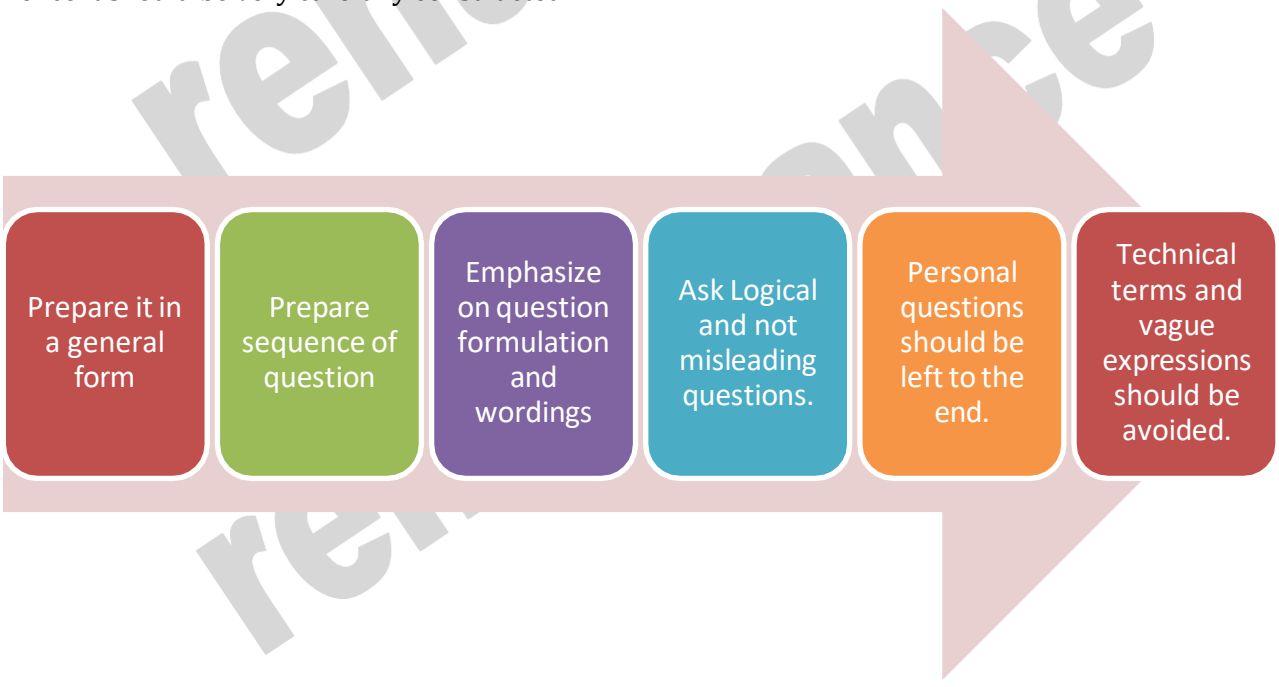
Importance:-

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



Example :

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

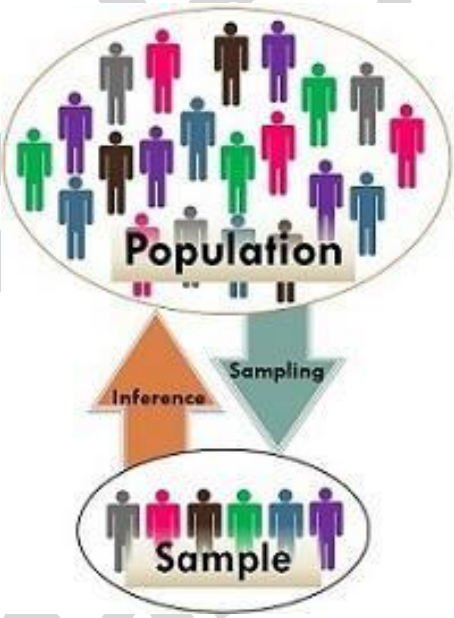
(b) Two way Table

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.

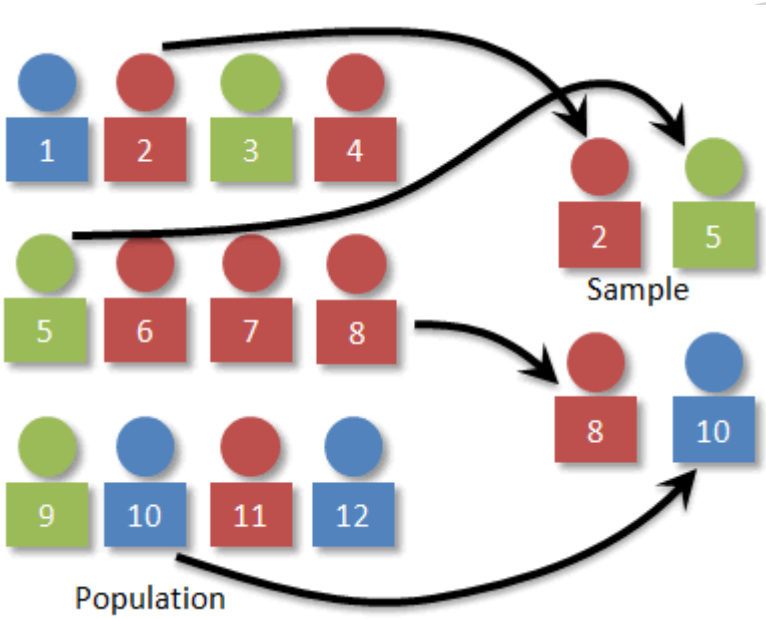


The various methods of sampling can broadly be divided into:

i. Random sampling method
ii. Non Random sampling method

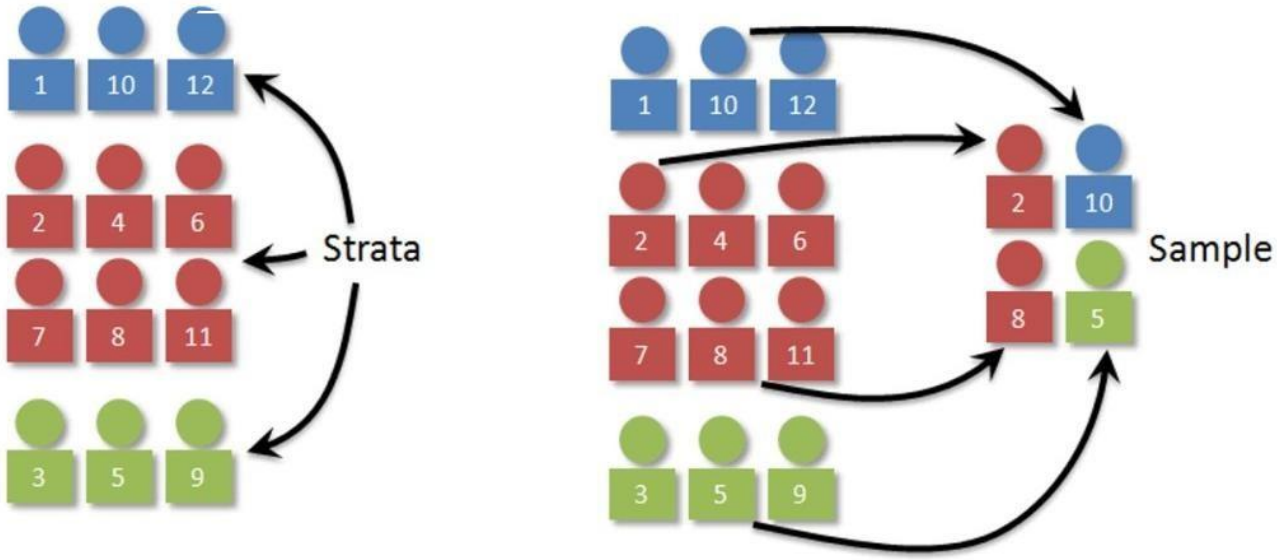
Random Sampling Method

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

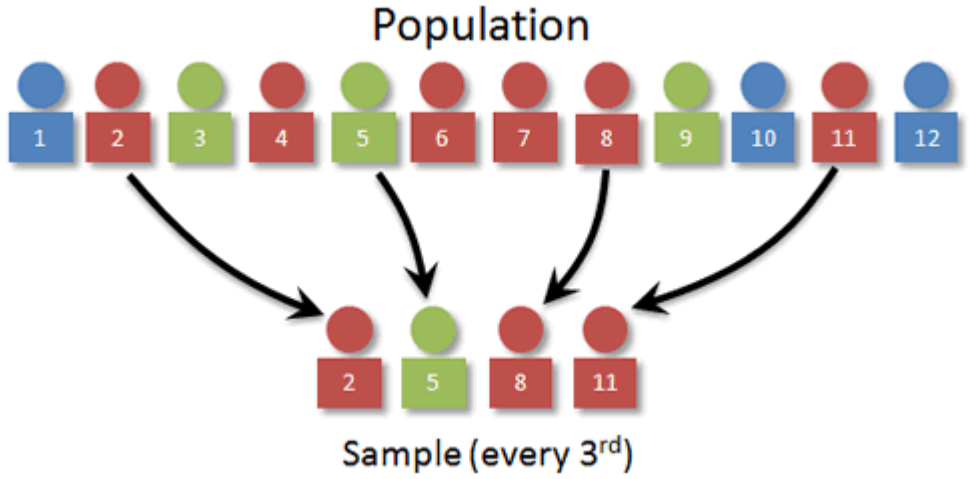
II Stratified Sampling:- In this it is important to divided the population into homogeneous group called strata. Then a sample may be taken from each group by simple random method.



Merit:- More representative sample is used.

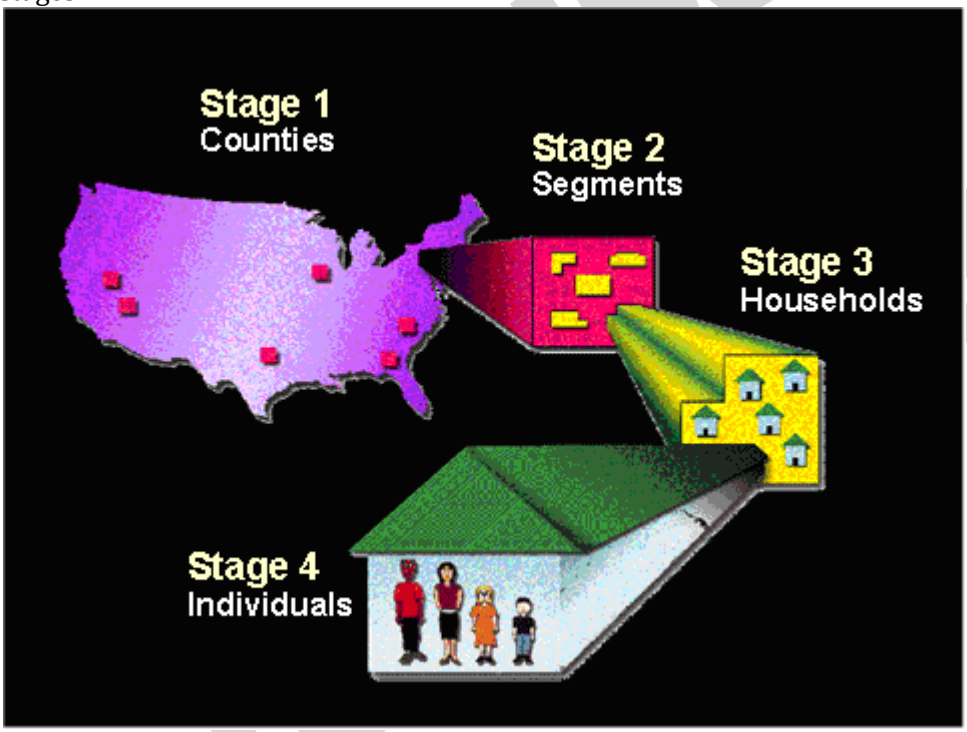
Grater accuracy
Geographically Concentrated
Limitations: Utmost care must be exercised due to homogeneous group deviation. In the absence of skilled supervisor sample selection will be difficult.

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.

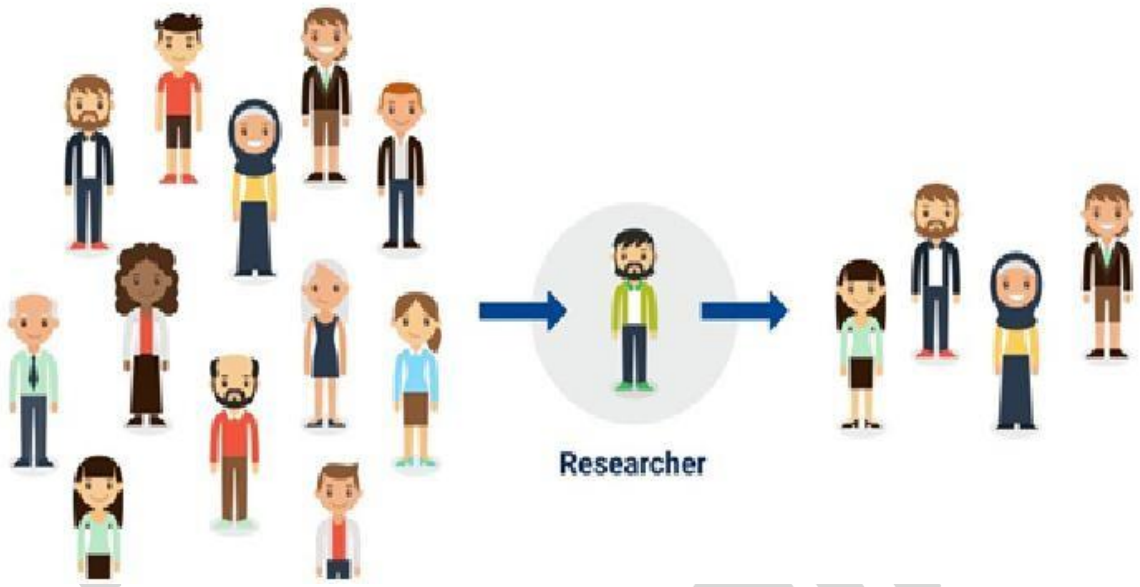
IV Multi- Stage Sampling: - This method refers to a sampling procedure which is carried out in several stages.



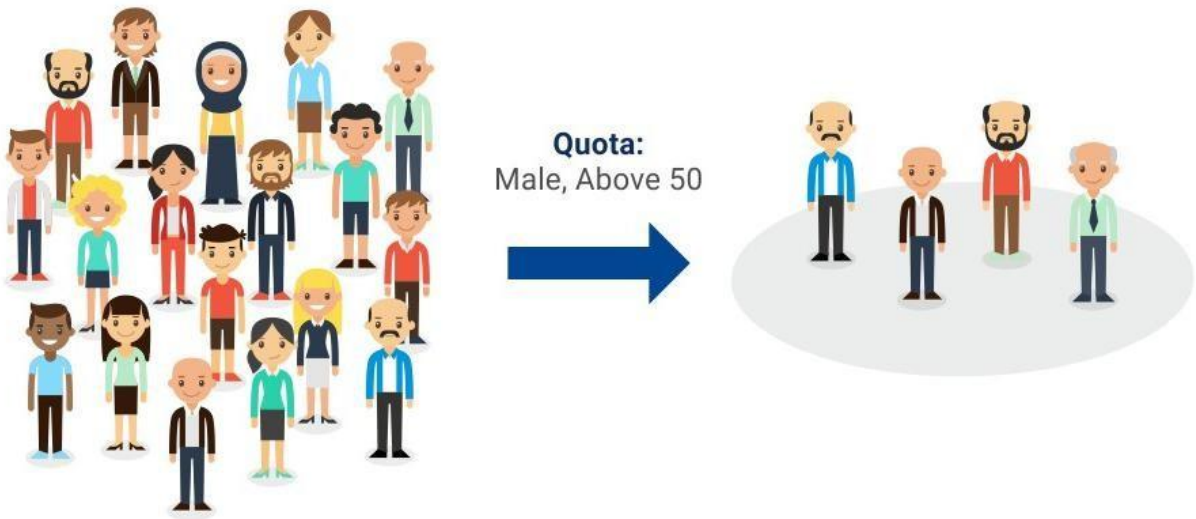
Merit: - It gives flexibility in Sampling
Limitation: - It is difficult and less accurate

Non Random Sampling Method:-

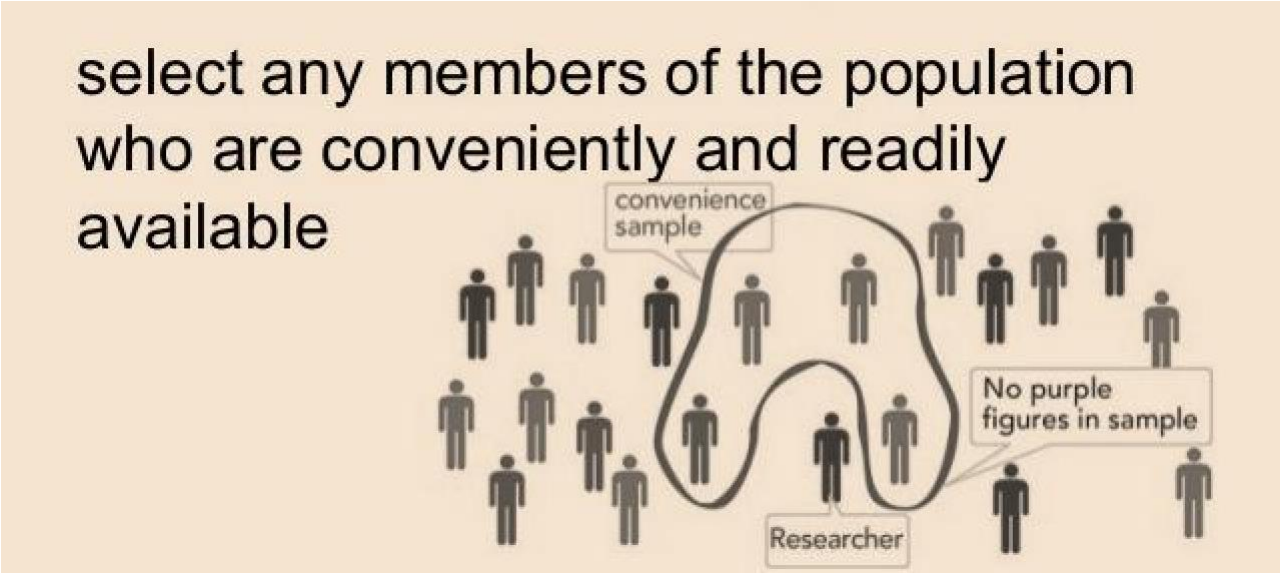
I. **Judgment Sampling: -** The choice of sample items depends exclusively on the judgment of the investigator or the investigator exercises his judgement in the choice of sample items. This is an simple method of sampling.



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.



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.

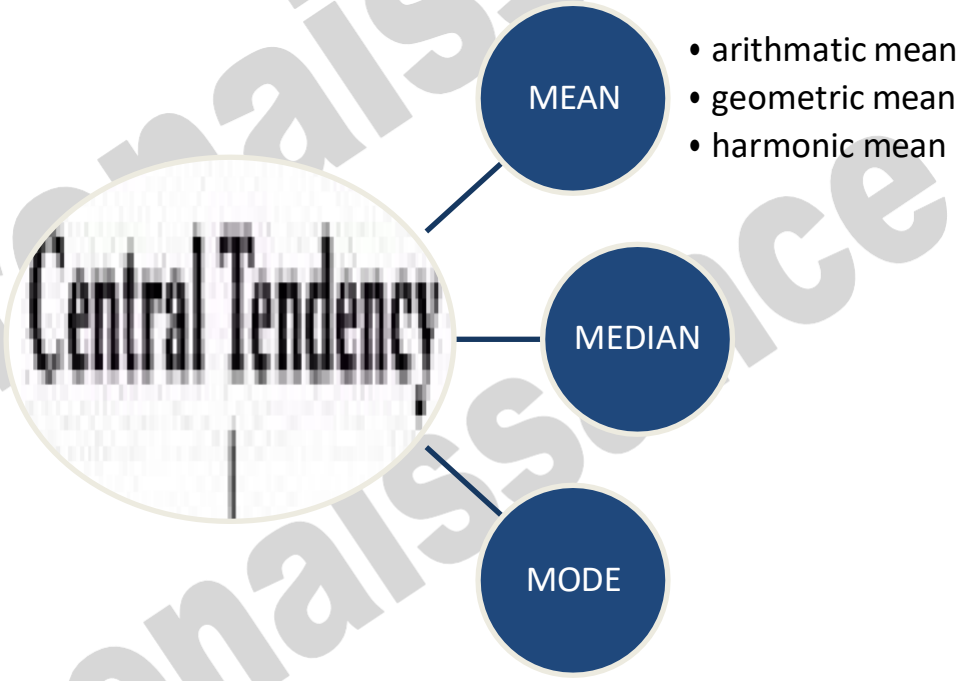
UNIT-III
MEASURES OF CENTRAL TENDENCY

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 $\bar{X}$.

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

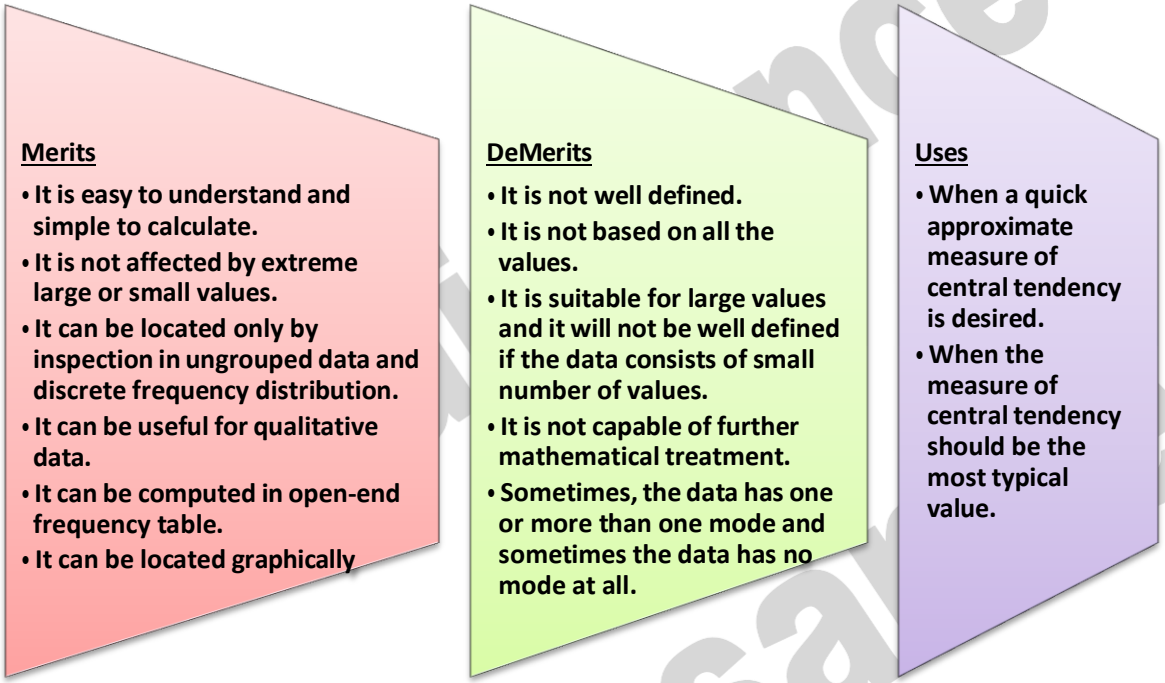
- For calculation, it is necessary to arrange data in ascending or descending order.
- Since it is a positional average, its value is not determined by each and every observation.
- It is not suitable for further algebric treatment.
- It is not accurate for large data.
- The value of median is more affected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



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.

UNIT IV- Measures of Variation

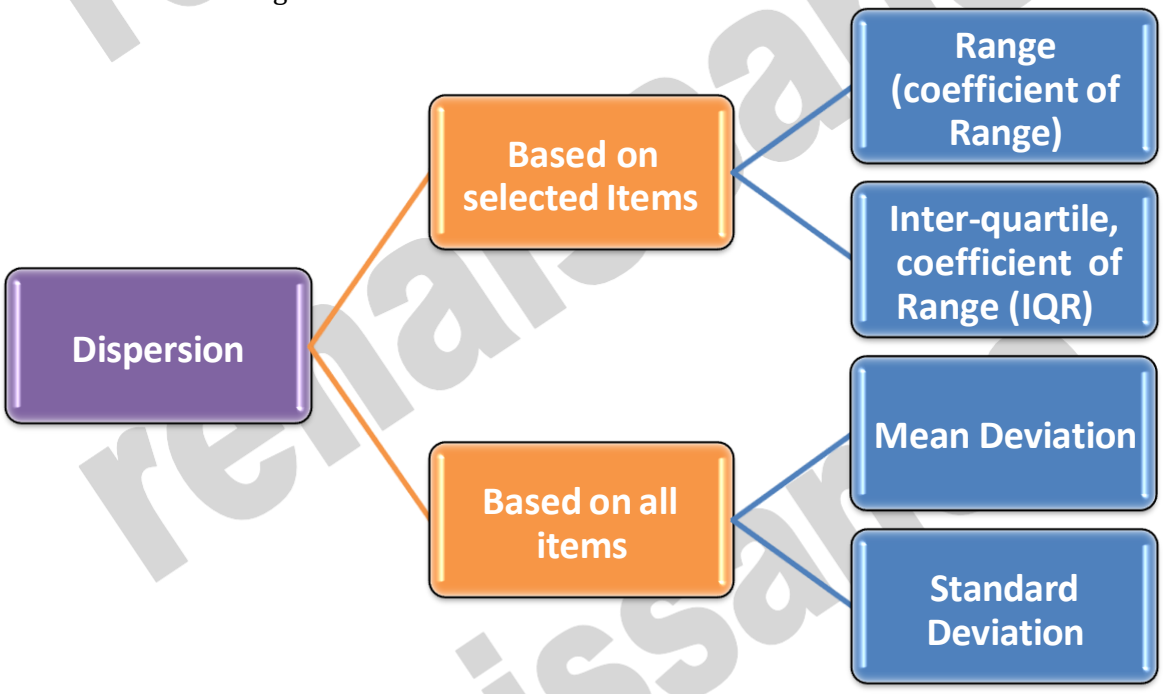
The Dispersion (Known as Scatter, spread or variations) measures the extent to which the items vary from some central value. The measures of dispersion is also called the average of second order (Central tendency is called average of first order).
The two distributions of statistical data may be symmetrical 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 $\sigma 2$. 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 $\frac{\sigma}{X}$ x 100 = known as coefficient of variations.

Interpretation of Coefficient of Variance-

Value of variance	Interpretation
Smaller the value of $\sigma 2$	Lesser the variability or greater the uniformity/ stable/ homogenous of population
Larger the value of $\sigma 2$	Greater the variability or lesser the uniformity/ consistency of the population

In statistics, **measures of variation** (also known as measures of dispersion) tell us how spread out a data set is. While the mean or median tells you where the "center" is, these measures tell you if the data points are huddled close to that center or scattered far away.

1. Mean Deviation (MD)

Mean deviation is the arithmetic average of the absolute differences between each data point and a central value (usually the mean or median).

Why absolute? We use absolute values (ignoring plus or minus signs) because otherwise, the positive and negative deviations would cancel each other out.

Formula (for ungrouped data):

$$MD = \frac{\sum |x - \bar{x}|}{n}$$

Where, x is each value, $\bar{x}$ is the mean, and n is the number of values.

Pros: Uses every item in the data set.

Cons: Ignoring signs makes it mathematically "inconvenient" for advanced statistical analysis.

2. Quartile Deviation (QD)

Also known as the **Semi-Interquartile Range**, this measure focuses on the middle 50% of the data. It is less affected by extreme outliers.

The Concept: It is half the distance between the third quartile (Q_3) and the first quartile (Q_1).

Formula:

$$QD = \frac{Q_3 - Q_1}{2}$$

Context:

Q1: The 25th percentile.

Q3: The 75th percentile.

Pros: Excellent for skewed distributions or data with extreme outliers.

Cons: It ignores the bottom 25% and top 25% of the data entirely.

3. Standard Deviation (SD)

The Standard Deviation is the most widely used and "gold standard" measure of variation. It measures the average distance of data points from the mean.

The Logic: Instead of ignoring signs (like Mean Deviation), we **square** the deviations to make them positive, then take the **square root** at the end to return to the original units.

Formula (Population SD - σ):

$$\sigma = \sqrt{\frac{\sum(x-\mu)^2}{N}}$$

The Variance: Note that the square of the standard deviation σ^2 is called the **Variance**.

Pros: Mathematically robust; used in almost all advanced statistical tests and the Normal Distribution.

Cons: Highly sensitive to outliers (since squaring a large difference makes it even larger).

Summary Comparison Table

Measure	Based on...	Sensitivity to Outliers	Best Used For...
Quartile Deviation	Positional values (Q1,Q3)	Low (Ignores them)	Skewed data
Mean Deviation	Absolute distances from mean	Moderate	Simple spread analysis
Standard Deviation	Squared distances from mean	High	Advanced stats & probability

DISPERSION

RANGE = R

Metric	Formula
Range (R)	R = L - S
Coefficient of Range	L - S/L + S

L: Largest value in the distribution.
S: Smallest value in the distribution.
Note: For Continuous Series, L is the upper limit of the highest class and S is the lower limit of the lowest class.

MEAN DEVIATION

Formula (For Direct Method)

(a) Mean Deviation from Mean:

(i) Individual Series:

$$\delta_{\bar{X}} = \frac{\sum |X - \bar{X}|}{N} = \frac{\sum d_X}{N}$$

(ii) Discrete Series:

$$\delta_{\bar{X}} = \frac{\sum f|X - \bar{X}|}{\sum f} = \frac{\sum f d_X}{N}$$

(iii) Grouped Series:

$$\delta_{\bar{X}} = \frac{\sum f|X - \bar{X}|}{\sum f} = \frac{\sum f d_X}{N}$$

(b) Mean Deviation from Median:

(i) Individual Series:

$$\delta_M = \frac{\sum |X - M|}{N} = \frac{\sum d_M}{N}$$

(ii) Discrete Series:

$$\delta_M = \frac{\sum f|X - M|}{\sum f} = \frac{\sum f d_M}{N}$$

(iii) Grouped Series:

$$\delta_M = \frac{\sum f|X - M|}{\sum f} = \frac{\sum f d_M}{N}$$

Here X = Mid-value.

(c) Mean Deviation from Mode:

(i) Individual Series:

$$\delta_Z = \frac{\sum |X - Z|}{N} = \frac{\sum d_Z}{N}$$

(ii) Discrete Series:

$$\delta_Z = \frac{\sum f|X - Z|}{\sum f} = \frac{\sum f d_Z}{N}$$

(iii) Grouped Series:

$$\delta_Z = \frac{\sum f|X - Z|}{\sum f} = \frac{\sum f d_Z}{N}$$

Here X = Mid-value.

COEFFICIENT OF MEAN DEVIATION

Let d_a denote mean deviation about any point a (may be mean, median, mode). Then coefficient of mean deviation is defined as $\frac{\delta_a}{a}$.

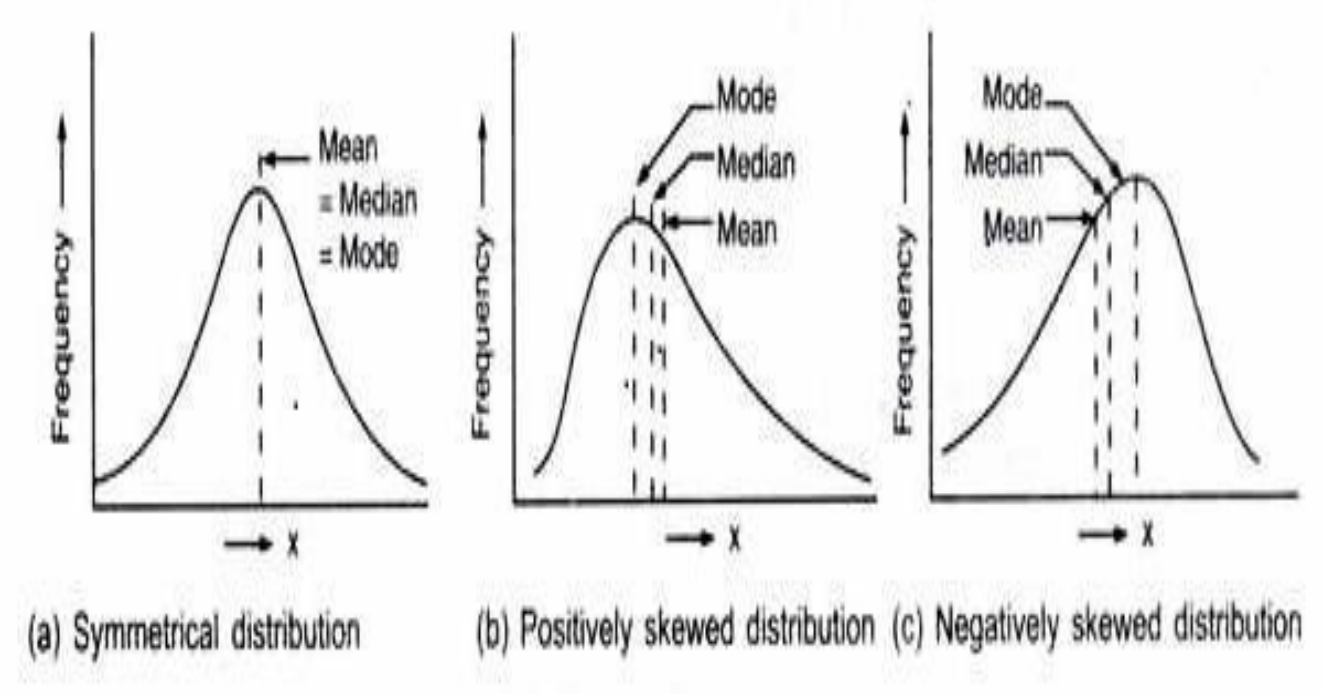
- **Thus, Coefficient of mean deviation from mean** $= \frac{\delta_{\bar{X}}}{\bar{X}}$
- **Coefficient of mean deviation from median** $= \frac{\delta_M}{M}$
- **Coefficient of mean deviation from mode** $= \frac{\delta_Z}{Z}$

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)	$\sqrt{\frac{\sum dx^2}{N} - \left(\frac{\sum dx}{N}\right)^2}$	$\sqrt{\frac{\sum fdx^2}{\sum f} - \left(\frac{\sum fdx}{\sum f}\right)^2}$	$\sqrt{\frac{\sum fdx^2}{\sum f} - \left(\frac{\sum fdx}{\sum f}\right)^2}$

SKEWNESS

Skewness is a measure of symmetry, or more precisely, the lack of symmetry. A distribution, or data set, is symmetric if it looks the same to the left and right of the center point.



Skewness is positive if the tail on the right side of the distribution is longer or fatter than the tail on the left side. The mean and median of positively skewed data will be greater than the mode. Skewness is negative if the tail of the left side of the distribution is longer or fatter than the tail on the right side. The mean and median of negatively skewed data will be less than the mode. If the data graph symmetrically, the distribution has zero skewness, regardless of how long or fat the tails are.

Karl Pearson developed two methods to find skewness in a sample:

1. Pearson's Coefficient of Skewness #1 uses the mode. The formula is:

$$Sk_1 = \frac{\bar{X} - Mo}{s}$$

Where $\bar{X}$ = the mean, Mo = the mode and s = the standard deviation for the sample.

2. Pearson's Coefficient of Skewness #2 uses the median. The formula is:

$$Sk_2 = \frac{3(\bar{X} - Md)}{s}$$

Where $\bar{X}$ = the mean, Mo = the mode and s = the standard deviation for the sample.
It is generally used when you don't know the mode.

TIME SERIES ANALYSIS

"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 timeseries 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

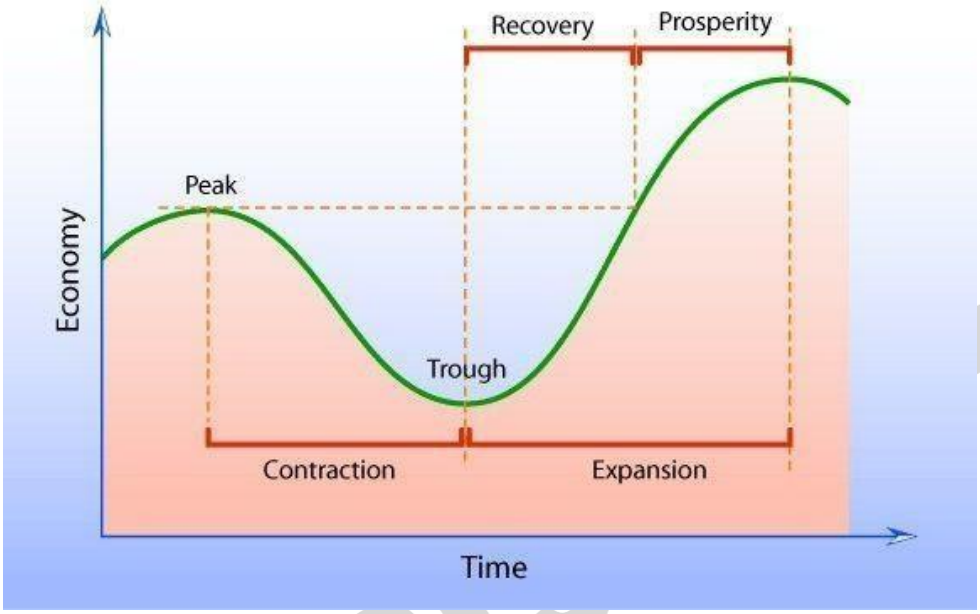
i. Sale of woolensgo up in winter.
ii. Sale of raincoat and umbrella go up in rainy season.
iii. Prices of food grains decrease with the arrival of new crop.
iv. Sale of cooler, refrigerator etc. rise during the summer season.

Another variation occurs due to man-made convention and customs, which people follow at different times like DurgaPooja, 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 DurgaPooja 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**. 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

$$U = T \times S \times C \times R$$

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 :

a+b+c3 b+c+d3 c+d+e3 d+e+f3

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: Addfirst two 4 yearly moving total place it against 3rd year. Leave the first moving total and then add nexttwo 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. (y-c)=0, i.e., the sum of the deviations of the actual values of y and computed values of y is zero.

ii. (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:
y sum of actual values of y variable.
y represents sum of deviations from the origin.
y x^2 represents sum of deviations from the origin.
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 x=0 the above normal equation would be reduced to the
y=na+x✇y=na+0=na✇Thus a = Yn
xy=ax+bx2+0+bx2= xy=b=x2 Thus b = xy=x2

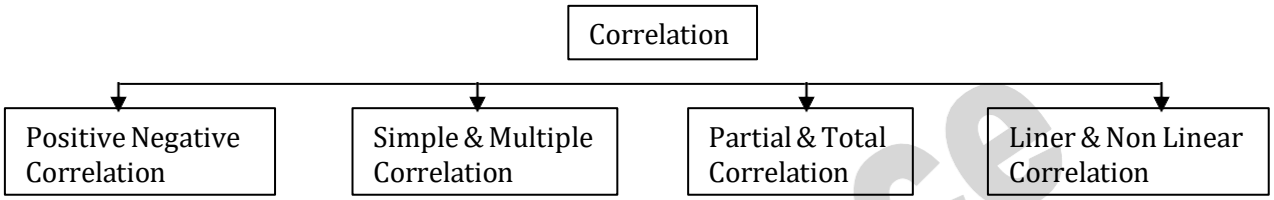
UNIT-V
CORRELATION
INTRODUCTION :-

1. 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.
2. 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 –
3. 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.
4. 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

1. "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.**
2. "Correlation means that between 2 series or groups of data there exists some casual correction". **WI King**
3. "Analysis of Correlation between 2 or more variables is usually called correlation."**A.M. Turtle**
4. "Correlation analysis attempts to determine the degree of relationship between variables. **YaLunchou**

TYPES OF CORRELATION

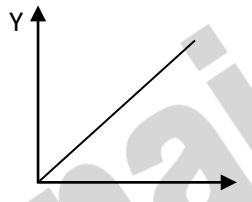
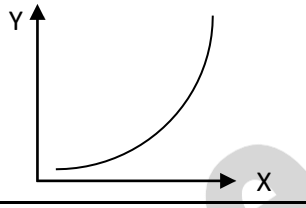
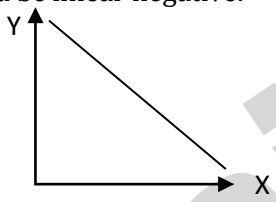
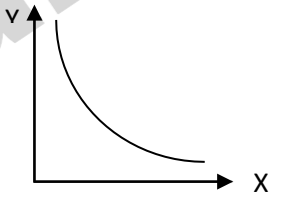


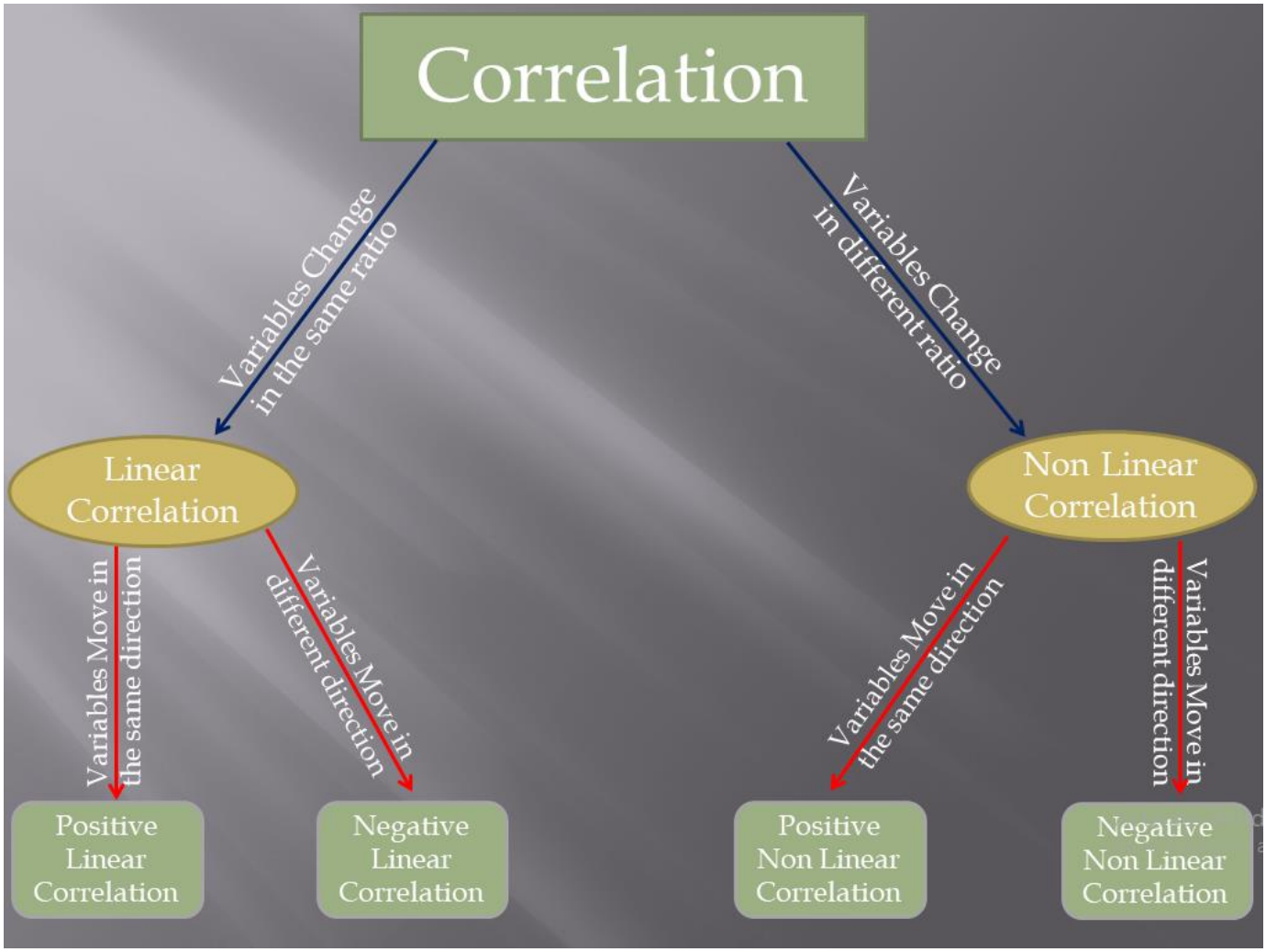
	POSITIVE CORRELATION	**NEGATIVE CORRELATION**
1	Value of 2 variables move in the same direction i.e. when increase/decrease in value of one variable will cause increase or decrease in value of other variable.	Value of 2 variables move in opposite direction i.e. when one variable increased, other variable decreaseswhen 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 Then, Correlation between Y & X	E.g. Demand & Price Demand depends on → Price Demand on → income This is case of multiplecorrelationsbecause 2factors (Price & Income) that affects demand are taken.We have to find correlation between demand & price. Demand & Price If, demand = Y Price = X_1 Price = X_2 Then Correlation between Y & X_1 Correlation between Y & X_2

SIMPLE CORRELATION	**MULTIPLE CORRELATION**
In partial correlation though more than 2 factors are involved but correlation is studies only between to be constant. E.g. Y, X_1 Y = Demand X_1 = Price X_2X_2 = Income	In total correlation relationship between all the variables is studied i.e., none of item is assumed to be constant E.g. Y, X_1 Y = Demand X_1 = Price X_2X_2 = Income

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)	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.

<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 linearor 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, itis 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 thechange 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>


Correlation
Variables Change in the same ratio
Variables Change in different ratio
Linear Correlation
Non Linear Correlation
Variables Move in the same direction
Variables Move in different direction
Variables Move in the same direction
Variables Move in different direction
Positive Linear Correlation
Negative Linear Correlation
Positive Non Linear Correlation
Negative Non Linear Correlation

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

σ_y = Variance of y 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 Coefficient of Correlation (Product Moment Method)

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

Deviation from Actual Mean Method

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

Put $\sum d_x = \sum d_y$

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

Deviation from Assumed Mean Method (Short Cut Method)

Direct Method for Karl Pearson's Coefficient of Correlation

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

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

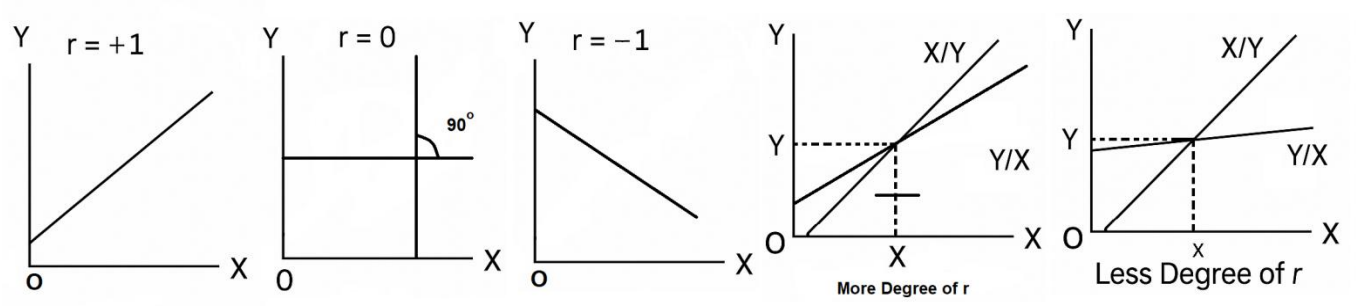
Why are two Regression lines –

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 = $\pm$ 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 = $\pm$70 or more the two regression line with be next to each other whereas when less degree of correction. Say (r=$\pm$ 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 the two variables. the degree of their co-variation is also reflected in correlation. but correlation does not study the nature of relationship.	Regression tells use about the 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 nonlinear 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. Tomakethebestestimate –
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
   $Y - Y = b_{xy}\ (x - x)$ whereb$_{xy}$ = regressioncoefficientofY onX

Regression Coefficients:

$$b_{xy} = r \cdot \frac{\sigma_x}{\sigma_y}, \quad b_{yx} = r \cdot \frac{\sigma_y}{\sigma_x}$$

$$\text{also } b_{xy} = \frac{N\sum XY - \sum X \sum Y}{N\sum Y^2 - (\sum Y)^2}, \quad b_{yx} = \frac{N\sum XY - \sum X \sum Y}{N\sum X^2 - (\sum X)^2}$$

Relation between r, b_{xy}, and b_{yx}:

$$r = \sqrt{b_{xy} \cdot b_{yx}}$$

Regression Analysis

Regression is based on two equations:

Equations	x on y	y on x
After elaborating them	$(x-\bar{x}) = r \cdot \frac{\sigma_x}{\sigma_y}(y-\bar{y})$	$(y-\bar{y}) = r \cdot \frac{\sigma_y}{\sigma_x}(x-\bar{x})$
Coefficient of Regression	$b_{xy} = r \cdot \frac{\sigma_x}{\sigma_y}$	$b_{yx} = r \cdot \frac{\sigma_y}{\sigma_x}$
To find out coefficient of regression through actual mean	$b_{xy} = \frac{\sum dx \cdot dy}{\sum dy^2}$	$b_{yx} = \frac{\sum dx \cdot dy}{\sum dx^2}$
Through assumed mean	$b_{xy} = \frac{\sum dxdy \cdot n - \sum dx \cdot \sum dy}{\sum dy^2 \cdot n - (\sum dy)^2}$	$b_{yx} = \frac{\sum dxdy \cdot n - \sum dx \cdot \sum dy}{\sum dx^2 \cdot n - (\sum dx)^2}$

Relation between r, b_{xy}, and b_{yx}:

$$r = \sqrt{b_{xy} \cdot b_{yx}}$$

REGRESSION COEFFICIENT –Thereare 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

INDEX NUMBERS

Index numbers are devices which measure the change in the level of a phenomenon with respect to time, geographical location or some other characteristic. The first index number was constructed in the year 1764 by an Italian named Carli to compare the changes in the price for the year 1750 with the price level of the year 1500. In present day situation changes in production, consumption, exports, imports, national income, cost of living, incidence of crimes, number of road accidents, business failures and a very wide variety of other phenomena are studied with the help of index numbers. Index numbers are supposed to be barometers which measure the change in the level of a phenomena.

"An index number is a statistical measure designed to show changes in variable or a group of related variables with respect to time, geographical location or other characteristics."

CHARACTERISTICS OF INDEX NUMBERS

1. **Index number are a specialised type of average.** Averages can be used to compare only those series which are expressed in the same units. However the device of index number helps us in comparing change in series which are in different units.
2. **Index numbers study the effects of such factors which cannot be measured directly.** Index numbers are meant to study the changes in the effects of such factors which cannot be measured directly.
3. **Index numbers being out the common characteristics of a group items.**
4. **Index number measure only relative changes in the values of a phenomenon.**

USES OF INDEX NUMBERS

1. **Help in Studying Trends.** Index numbers helps to find out the trend of exports, imports, balance of payments, industrial production, prices, national income and a variety of other phenomena.
2. **Help in policy formulation.** Index numbers help us in studying trends of various phenomena and these trends and tendencies are the bases on which may policy decisions are taken index number are used by the government in deciding the rates of D.A. and levy of excise duties.
3. **Help in measuring the Purchasing Power of Money.** Index numbers are helpful in finding out the intrinsic worth of money as contrasted with its nominal worth.
4. **Helps in deflating various value.** Index numbers are very helpful in deflating national income on the basis of constant prices.
5. **Act as economic barometers.** Index numbers measure the pulse of an economy and act as barometers to find the ups and down in the general economic condition of a country.

PROBLEMS IN THE CONSTRUCTION OF INDEX NUMBERS

1. **The selection of item-** The first problem which the marker of an index number of wholesale prices has to face is that of the selection of items from which the index number is to be constructed.
2. **The selection of the base year-** Second problem in the construction of index numbers is the selection of the base year and the conversion of current prices to price relatives based on the prices of the base year.
3. **The selection of the average-** The next step in the construction of wholesale price index number is to average the prices relatives of the various commodities.
4. **Selecting suitable weights.** All the items used in the construction of an index number are not of equal importance and as such if the index number is to be a representative one, weights

$$P_{01} = \frac{\sum p_1}{\sum p_0} \times 100$$

$$\sum p_1$$

$$\sum p_0$$

should be assigned to various items in relation to their importance.

METHODS OF CONSTRUCTING INDEX NUMBERS

Broadly speaking various methods of constructing index numbers can be classified in two groups viz.

A. Unweighted Index Numbers
B. Weighted Index Numbers
i. Simple Aggregative Method
ii. Simple Average of Relatives Method.

A. Unweighted Index Numbers

Where
P_{01} = Index number of the current year

= Total of the current year; price of all commodities.

= Total of the base year's price of all commodities.

Simple Average of Relatives Method. $$p_0 = \frac{\sum\left(\frac{p_1}{p_0}\times 100\right)}{N}$$

B. Weighted Index Numbers

1. **Laspeyres Method** - $$P01 = \frac{\sum p_1 q_o}{p_0 q_o}\times 100$$

2. Passche's Method $$P_{01} = \frac{\sum p_1 q_1}{\sum p_0 q_1}\times 100$$

3. Drobish and Bowleys Method $$P_{01} = \frac{\frac{\sum p_1 q_0}{\sum p_0 q_0} + \frac{\sum p_1 q_1}{\sum p_0 q_1}}{2}\times 100$$

4. **Fisher's Ideal Index**. $$P_{01} = \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0}\times\frac{\sum p_1 q_1}{\sum p_0 q_1}}\times 100$$

5. Marshall-Edgeworth formula

$$P_{01} = \frac{\sum (q_0 + q_1) p_1}{(q_0 + q_1) p_0}\times 100 \quad \text{or} \quad P_{01} = \frac{\sum p_1 q_o + p_1 q_1}{\sum p_0 q_0 + \sum p_0 q_1}\times 100$$

6. Walsch Formula. $$P_{01} = \frac{\sum p_1\sqrt{q_o q_1}}{\sum p_0\sqrt{q_o q_1}}\times 100$$

7. Kelly's Method. $$P_{01} = \frac{\sum p_1 q}{\sum p_0 q}\times 100$$

Quantity Index Number

Quantity index number measure the changes in the volume of production, construction or employment over a period of years.

Formula for simple or unweighted quantity index;

$$Q_{01} = \frac{q_1}{q_0}\times 100$$

Here q_1 = Quantity of current year

q_0 = Quantity of base year

Q_{01} = Current year's quantity index based on base year's quantity

Base Shifting:- Base shifting is generally required due to following reasons

i. The base year is too old to compare the current year.

ii. If different series of index numbers are based on different base years and they are to be compared from each other.

Deflation of index numbers

Computation of real wages from money income with taking the effect of price level changes is called as deflating of index numbers.

$$\text{Index No. of Real Income or Deflated Index Number} = \frac{\text{Real Income of Current year}}{\text{Real Income of Base year}}\times 100$$

Splicing: Sometimes series of index number based on a certain year is discontinued and a new series of index number is prepared by taking another year as base. Thus two series of index number would result. In this situation index number of these two series are not comparable because both are based

on different years. If these are to be compared then new series will be covered on the basis of old series or vice-versa; this conversion/shifting is called as spicing. Splicing may be taken as another form of base shifting.

Formula for splicing :-

a. Splicing of new series in old series (Forward splicing):

$$\text{Splicing Index Number} = \frac{\text{Index Number to be adjusted} \times \text{Old Index Number of new series}}{100}$$

Spliced Index Number

b. Splicing of old series in new series (backward splicing):

$$\text{Splicied Index Number} = \frac{100 \times \text{Index No. to be adjusted}}{\text{Old Index No. of New Series}}$$

TESTS OF ADEQUACY OF INDEX NUMBER FORMULAE

We have discussed a large number of formulae for the construction of both simple and weighted index numbers. We formula should be chosen for the construction of an index number is a question which can not be satisfactorily answered. However some tests have been suggested to determine the adequacy of an index number formula. These tests are:

1. **Unit Test –** This test requires that the formula for the construction of index numbers should be such which is not affected by the unit in which prices or quantities have been quoted. This test is satisfied by all index number formulae discussed above except the simple (unweighted) aggregative index formula. In this index as we have discussed earlier the units play an important part in determining the value of the index. If only the unit is changed (say from kg to quintal) the value of the index would change.

2. Time Reversal Test- In the worlds of Fisher: "The test is that the formulae for calculating an index number should be such that it will give the same ratio between one point of comparison and the other no matter which of the two is taken as base." This mean that the index number should work both backwards as well as forwards. Thus, if the index number of the current year is 4000 then the index number of the base year (based on the current year) should be 25. In other words, the two index numbers thus calculated (without the figure 100) should be reciprocals of each other. The reciprocal of 4 is .25 and the reciprocal of .25 is 4. The product of these two ratios would always be equal to one. Thus, if P_{10} represents the price change in the current year and P_{10} the price change of the base year (based on the current year) the following equation should be satisfied:-

3. Factor Reversal Test- In the words of Fisher: "Just as each formula should permit inter-changing the price and quantities without giving inconsistent result, i.e., the two results multiplied together should give the trust value ratio." It means that the changes in the prices multiplied by the changes in quantity should be equal to the total change in value. Change in value is the result of changes should represent the total change in value. Thus, if the price of a commodity has doubled during a certain period and if in this period the quantity has trebled the total change in the value should be six time the former level. In the other words, if p_1 and p_0 represent the prices and q_1 and q_0 the quantities in the current and the base years respectively, and if p_{01} represent the change in price in the current year and q_{01} the change in the quantity in the current year then

$$P_{01} \times q_{01} = \frac{\sum p_1 q_1}{\sum p_0 q_1}$$

The factor reversal test is satisfied only by the Fisher's Ideal Index Number.

The proof of it is given below:

Circular Test

Another test applied in index number studies is the circular test. It is a short of extension of the time reversal test. Suppose an index number is constructed for the year 1983 with the base of 1982 and another index number for 1982 on the base of 1981, then it should be possible for us to directly get an index number for 1983 on the base of 1981. If the index number calculated directly does not give an inconsistent value, the circular test is said to be satisfied. If p_{01} represent the price change of the current year on the base year and P_{12} the price change of the base year on some other base and p_{20} the price change of the current year on this second base then the following equation should be satisfied.