

**SYLLABUS****Class - B.Com. (Hons.) II year****Subject -Advanced Statistics**

UNIT - I	Introduction to Statistics, Concept of Population and Sample, Types of data, Primary and Secondary data, Collection of data, Organization of data- Frequency tables and Frequency Distribution. Presentation of Data- Bar Digram, Pie Digram, Line Graph, Histograms & Frequency Polygons.
UNIT - II	Measurement of central tendency - Mode, Median and Geometric Mean. Measures of Dispersion- Range, Quartile Deviation, Mean Deviation, Standard Deviation and Basic Concept of Skewness and Kurtosis
UNIT - III	Theory of Probability - Experiments, Sample Spaces, and Events, Addition and Multiplication Theorum, Conditional Probability Concept Discrete and Continuous Random Variables. Probability Distributions — Binomial, Poisson and Normal Distributions.
UNIT IV	Sampling Distribution - Concept Parameter and Statistic. Sampling Distribution of Mean and Central Limit theorem, Point and Interval estimation of a Population Mean (Large and Small Sample Case) Basic Concepts of Hypothesis testing. Hypothesis Tests based on a Single Sample for Mean and Proportion — Z test, I test.
UNIT - V	Correlation — Meaning, Definition and Types of Correlation. Karl Pearson's Coefficient of Correlation, Coefficient of determination, Spearman's Rank Correlation Coefficient. Simple Linear Regression — Lines of Regression (Estimating Lines), Regression Coefficients and their Properties. Application of regression in forecasting



UNIT – I
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.

Definition: - “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.
- 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.

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:-

- (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. Scientific and systematic collection of data and their analysis with the help of various statistical method and their interpretation.

Stages of Statistical Investigation:-

- Planning of Investigation
- Collection of Data
- Editing of Data
- Presentation of Data
 - (a) Classification
 - (b) Tabulation
 - (c) Diagrams
 - (d) Graphs
- Analysis of Data
- 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.

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.

Types of Data:-

1. **Raw Data:-** When we collect data through schedules and questionnaires or some other method eg:- Classification, tabulation etc.
2. **Processed Data:-** When we use the above raw data for application of different methods of analysing of data. Like using correlation, Z-test, T-test on data. That will be known as processed data.

Sources of Data Collection:-

3. **Internal Data:** - When data is collected by problem the internal source for any specific



Its purpose.

- 4. **External Data:** - This type of data collected by the external source.
- 5. **Primary Data:** - It is original and collected first time. it is like raw material and it is required large sum of money, energy and time.
- 6. **Secondary Data:** - Secondary data are those already in existence and which have been collected for some other purpose than answering of the question at hand.
- 7. **Qualitative Data:** - Which can not be measurable but only their presence and absence in a group of individual can be noted are called qualitative data.
- 8. **Quantitative Data:** - The characteristics which can be measured directly are known as quantitative data.

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

Difference between Primary and secondary data:-

Points	Primary Data	Secondary Data
1. 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.
2. Organisation	Primary data are like raw material.	Secondary data are in the form of finished product. They have passed through statistical methods.
3. 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.
4. Expenditure	The collection of primary data require large sum, energy and time.	Secondary data are easily available from secondary sources (published or unpublished).
5. Precautions	Precautions are not necessary in the use of primary data.	Precautions are necessary in the use of secondary data.

Preparation of Questionnaires:-

This method of data collection is quite popular, particularly in case of big enquiries, it is adopted by individuals, research workers. Private and public organization and even by government also. A questionnaire consists of number of questions printed or type in a definite order on a form or set of forms. The respondents have to answer the questions on their own.

Importance:-

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

Demerits:-

- (i) Low rate of return
- (ii) Fill on educated respondents



(iii) Slowest method of Response

Preparation of Questionnaires: - It is considered as the heart of a survey operation. Hence it should be very carefully constructed. If it is not properly set up and carefully constructed.

Step I	:-	Prepare it in a general form.
Step II	:-	Prepare sequence of question.
Step III	:-	Emphasize on question formulation and wordings
Step IV	:-	Ask Logical and not misleading questions.
Step V	:-	Personal questions should be left to the end.
Step VI	:-	Technical terms and vague expressions should be avoided classification and Tabulation of Data

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, colour of hair, literacy, religion etc.

Population

- 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 a 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-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 algebraic 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 of Arithmetic Mean:

- 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 of Arithmetic Mean:

- 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.

Methods of Calculating Arithmetic Mean:

- Direct Method



- Short cut method
- Step deviation method

Use of Arithmetic Mean:

Arithmetic Mean is recommended in following situation:

- 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 of Median:

- 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 of Median:

- 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 of Median:

The use of median is recommended in the following situations:

- 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 of Mode:

- 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 of Mode:

- 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 of Mode:

The use of mode is recommended in the following situations:

- 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 n th 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 interest 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).

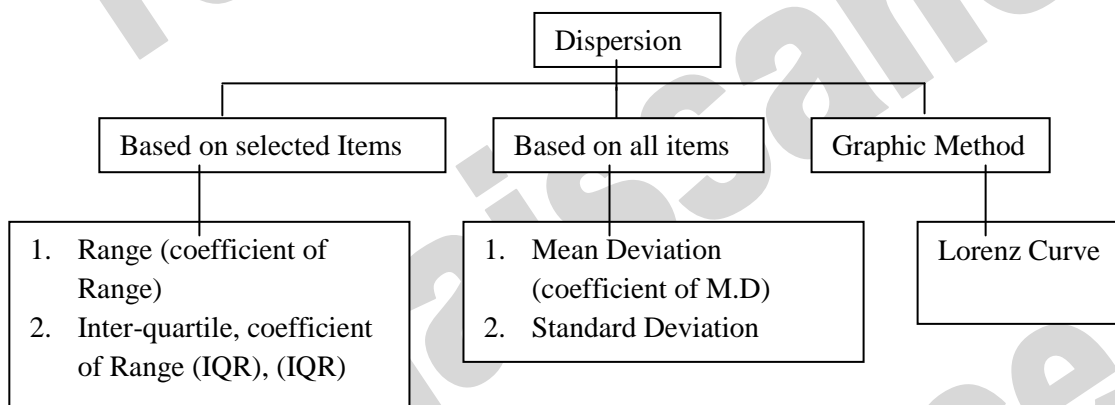
The two distributions of statistical data may be symmetrical and have common means, median or mode, yet they may differ widely in the scatter or their values about the measures of central tendency.

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 ($\delta\alpha$) 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.

Method of calculating standard Deviation-

1. Direct Method
2. Short-cut-Method
3. Step deviations Method

Properties



Fixed Relationship among measures of dispersion in a normal distribution there is a fixed relationship between quartile Deviation, Mean Deviation and Standard Deviation $Q.D = 2/3 \sigma$, Mean Deviation = $4/5\sigma$.

Distinction between mean deviation and standard deviation

Base	Mean Deviation	Standard Deviation
1. Algebraic 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.
2. Use of Measure	Mean deviation can be computed from mean, median, mode	Standard deviation is computed through mean only
3. Formula	M.D or $\delta = \frac{\sum fdx}{N}$	S.D or $\sigma = \frac{\sqrt{\sum fx^2}}{N}$
4. Further algebraic Treatment	It is not capable of further algebraic treatment.	It is capable of further algebraic treatment
5. Simplicity	M.D is simple to understand and easy to calculate	S.D is somewhat complex than mean deviation.
6. 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 σ^2 . Variance is also known as second moment 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 $\frac{\sigma}{\bar{X}} \times 100 =$ known as coefficient of variations.

Interpretation of Coefficient of Variance-

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

DISPERSION

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
$\delta M (\text{Median}) = \frac{\sum dM}{N}$	$\frac{\sum fdM}{N}$	$\frac{\sum fdM}{N}$
Coefficient of $\delta = \frac{\delta}{M}$	$\delta = \frac{\delta}{M}$	$\delta = \frac{\delta}{M}$
Mean $\delta \bar{X} = \frac{\sum dx}{N}$	$\frac{\sum fdx}{N}$	$\frac{\sum fdx}{N}$
Coefficient of $\bar{X} = \frac{\delta}{\bar{X}}$	$\frac{\delta}{\bar{X}}$	$\frac{\delta}{\bar{X}}$
(Mode) $\delta Z = \frac{\sum dz}{N}$	$\frac{\sum fdz}{N}$	$\frac{\sum fdz}{N}$
Coefficient of $\delta Z = \frac{\delta}{Z}$	$\frac{\delta}{Z}$	$\frac{\delta}{Z}$

Standard Deviation = σ can be calculated through mean only

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



UNIT-III

Introduction

Probability is a numerical measure of uncertainty. The general meaning of the word probability is likelihood. Every one of us use the phrases like

“There is a high chance of my getting the job next month.”

“Probably I will get elected.”

“Probably I will get good score in examination.”

“Possibly it will rain, to night.”

“India might win the cricket series against South Africa.”

“This year’s demand for the product is likely to exceed that of the last year’s.”

and so on.

All the above sentences, with words like ‘possibly’, ‘high chance’ ‘likely’ are expressions indicating a degree of uncertainty about the happening of the event. A numerical measure of uncertainty is provided by a very important branch of statistics called the ‘Theory of Probability’ Broadly, there are three possible states of expectation certainty by 1, impossibility by 0 and the various grades of uncertainties by coefficients ranging between 0 and 1.



Terms used in Theory of Probability

Random Experiment : Suppose we toss a coin or throw a die or draw a card from a pack of playing cards. In these examples there are a number of possible outcomes which can occur but there is an uncertainty as to which one of them will actually occur. Such experiments are called random experiment. A random experiment may be defined as an experiment which when repeated under essentially identical conditions does not give unique results but may result in any one of the several possible outcomes. These outcomes are known as events.

Properties of Random Experiment

- i) It can be repeated.
- ii) The number of possible outcomes is more than one.
- iii) In a single trial out of the several outcomes one and only one outcomes can happen i.e. one outcomes is certain to happen.
- iv) The outcomes of an individual trial cannot be predicted.

Trial : By Trial we mean performing of an experiment

Outcome : The result of a trial in a random experiment is called an outcome.

Sample Space : Each performance in a random experiment is called a trial. The result of a trial in a random experiment is called an outcome, an elementary event, or a sample point. The totality of all possible outcome (i.e., sample points) of a random experiment constitutes the sample space. Suppose $e_1, e_2, e_3, \dots, e_n$ are the possible outcomes of a random experiment E. We associate with each experiment E, a set $S = \{e_1, e_2, \dots, e_n\}$ of possible outcomes with the following properties :

- i) Each element of S denotes a possible outcomes of the experiment.
- ii) Any trial results in an outcomes that corresponds to one and only one element of the set S.

Thus a set of all possible outcomes in a random experiment is called sample space. It is denoted by S. For example,

- i) If a dice is thrown, the sample space will be : $S = \{1, 2, 3, 4, 5, 6\}$.
- ii) When a single coin is tossed : $S = \{H, T\}$.
- iii) When two coins are tossed simultaneously or one coin is tossed repeatedly twice, following will be the sample space : $S = \{HH, HT, TH, TT\}$.
- iv) Similarly, in case when either three coins are tossed or one coin is tossed repeatedly thrice, the sample space will be :

$$S = \{HHH, HHT, HTH, THH, TTT, TTH, THT, HTT\}.$$

Event : Any subset of a sample space is called an event. If an event contains only one sample point, then it is called a simple event or an elementary event.



If an event is the empty set (i.e., it does not contain any sample point) then it is called an impossible event. The sample space S is a subset of itself. Hence it is also an event. This event is called a certain event or a sure event, since it is sure to occur.

Compound Events : When two or more events are defined within the same experiment, they are called compound events. Following are the important events :

1) Equally Likely Events : Two (or more) events A and B are said to be equally likely if both are expected equally to happen. For example, when a coin is tossed, the appearance of head and tail is equally likely.

2) Mutually Exclusive Events : Two (or more) events A and B are said to be mutually exclusive events if they are so defined that both of them can not appear simultaneously. For example, in case of a coin, when it is tossed head and tail cannot appear simultaneously hence, they are mutually exclusive events. Mathematically,

$$A \cap B = \phi, \text{ a null set,}$$

i.e., nothing is common in both the events.

Note : When A and B are mutually exclusive events, they can't be independent or dependent events.

3) Independent Events : Two (or more) events A and B are called independent events if they are defined as, the appearance or non-appearance of one event does not affect the other event.

In case of a coin tossing experiment, $S = \{H, T\}$

if, $A = \{H\}$, $B = \{T\}$ are defined.

These A and B are not independent (but they are mutually exclusive) because if A appears, i.e., if head comes, tail can't come definitely, i.e., event B can't appear and vice-versa.

While if A and B are defined in the experiment of tossing of two coins as,

$A \rightarrow$ Head in first trial,

$B \rightarrow$ Head in second trial.

In this situation, both the events A and B are independent.

Note : When A and B are independent events, they cannot be mutually exclusive at the same time.

4) Dependent Events : Two (or more) events A and B are said to be dependent events if they are defined that the appearance of one event affect the other event.

When event A depends upon B , the notation used is B/A and when event B depends upon event A we denote it as A/B . It is not necessary that if A depends upon B then B will also depend upon A .

Note : Example of dependent events will be cited in the definition of conditional probability.



Conditional Probability

The simultaneous occurrence of two or more events is called a **compound event**. If A and B are two events, then AB denotes the simultaneous occurrence of A and B and $P(AB)$ denotes the probability of the simultaneous occurrence of two events A and B.

The probability of the happening of an event A when the event B has already happened is called the **conditional probability** and is denoted by $P(A/B)$. Similarly $P(B/A)$ means the probability of the happening of an event B when the event A has already happened.

Two events are said to be independent if the probability of the happening of one does not depend on the happening of the other.

Multiplication theorem of probability

(Theorem of compound probability)

According to this theorem, "If two events are mutually independent, and the probability of the one is p_1 while that of the other is p_2 , the probability of the two events occurring simultaneously is the product of p_1 and p_2 ."

Multiplication Rules

i) **When Events are Independent.** The probability that both independent events, A and B will occur is:

$$P(A \cap B) = P(A) \cdot P(B)$$

In other words, $P(A \text{ and } B) = P(A) \cdot P(B)$.

ii) **When Events are Dependent.** If events A and B are so related that the occurrence of B is affected by the occurrence of A, then A and B are called dependent events. The probability of event B depending on the occurrence of event A is called conditional probability and is written as $P(B/A)$.

The probability that both the dependent events A and B will occur, is given by:

$$P(A \cap B) = P(A) \cdot P\left(\frac{B}{A}\right)$$

In other words

$$P(A \text{ and } B) = P(A) \cdot P\left(\frac{B}{A}\right)$$



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, \dots, A_n and if B actually happens. Then the probability that it was preceded by the particular event $A_i (i = 1, 2, \dots, n)$ is given by:

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

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

The Formula for bayes is:

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



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 therefore 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 any 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, \dots, x_n with respective probabilities p_1, p_2, \dots, p_n such that (a) $0 \leq p_i \leq 1$ for $i = 1, 2, \dots, n$. (b) $p_1 + p_2 + \dots + p_n = 1$.

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 : $\text{var}(x) = E(x^2) - (E(x))^2$



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

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

But r successes in n trials can occur in ${}^n C_r$ 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 ${}^n C_r p^r q^{n-r}$ i.e.

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





Poisson Distribution

Concept of Poisson Distribution

Poisson distribution can be viewed as a limiting form of Binomial distribution when n approaches infinity ($n \rightarrow \infty$) and p approaches zero ($p \rightarrow 0$) 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 :

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



Normal Distribution

Concept of Normal Distribution

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

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

The normal distribution is the most versatile 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 be 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

INTRODUCTION

Statistical data are obtained either through a survey or experiment. Statistical surveys are most popular device of obtaining the desired data. A survey is a process of collecting data from existing population units with no particular control over factor that may affect the population characteristics of interest in the study. Experimental data are generally obtained in studies relating to natural or physical sciences. A statistical survey may be either a general purpose or a special purpose survey. So far as general surveys is concerned we may obtain data which are useful for several purpose. The best example of this type of survey is the population census taken every 10 years in India. Such a survey provides information not only about the total population but also about its division into males and females, literates and illiterates, employment and unemployment, age distribution, income distribution, etc. In many cases it is impossible or impracticable to make a complete survey, in such cases we have to depend on sample study only.

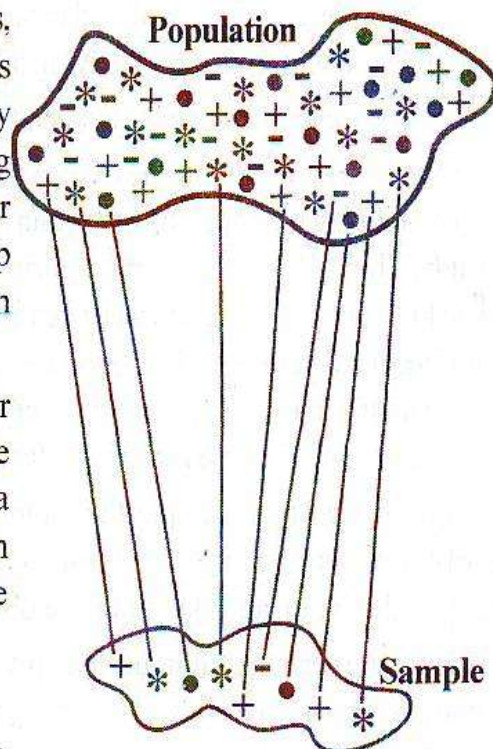
MEANING OF UNIVERSE OR POPULATION

In the field of statistics the term 'Universe' or 'Population' The meaning of the totality of cases or items constituting the field or enquiry. Thus, "In statistics, population is the aggregate of objects under study in any statistical investigation". In any statistical investigation the interest usually lies in studying the various characteristics relating to items or individuals belonging to a particular group. This group of individuals under study is known as the population or universe.

Finite and Infinite Population : The universe or population may be either finite or infinite. A finite population is hereby means a population having a determinable number of items. An infinite population is that in which the number of items cannot be determined.

MEANING OF SAMPLE

A sample is a part of universe or population selected for study to deduce some conclusions about population. According to L.R. Connor





PARAMETER AND STATISTIC

For any statistical analysis we have to find various statistical measures such as mean, mode, median, dispersion, moment, skewness and kurtosis etc. For both population as well as sample.

The characteristics of Population or Universe are mean, standard deviation and skewness. These characteristics are called Population **Parameters**. Thus any statistical measure computed population data is known as parameters.

The characteristics of sample data are mean, standard deviation and skewness and are called statistic. Thus any statistical measure computed sample data is known as **statistic**.

The usual notation used for parameter are capital letters and for statistic it is represented by small letter e.g.,

Statistical Measure	Mean	Standard deviation	Population	Size
Population Parameters	μ	σ	P	N
Sample statistic	\bar{x}	s	p	n



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 is 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 is 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, a 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 far 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. Neyman** 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 are 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 it is true.'

Symbol : It is denoted by H_0

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 different 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 hypothesize that it is true. For example: If we want to find the population mean has a specified value μ_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 occurs 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 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.

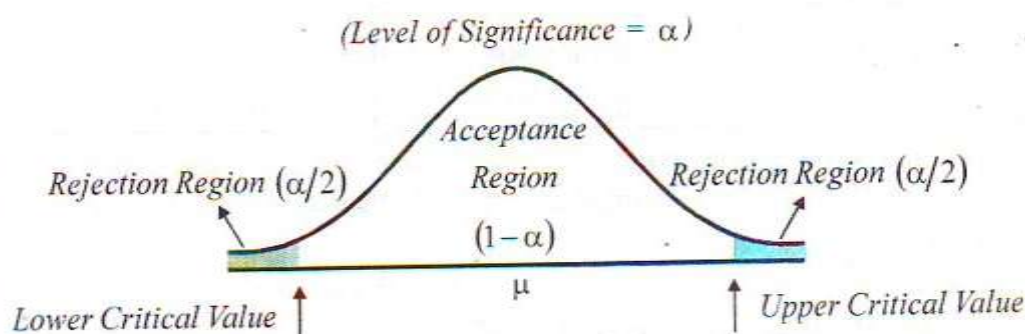
The region of the standard normal curve corresponding to a pre-determined level of significance that is fixed for knowing the probability of making the type I error or rejecting the hypothesis which is true, is known as the "rejection region" or "critical region". The region of standard normal curve 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): $\mu = 90$

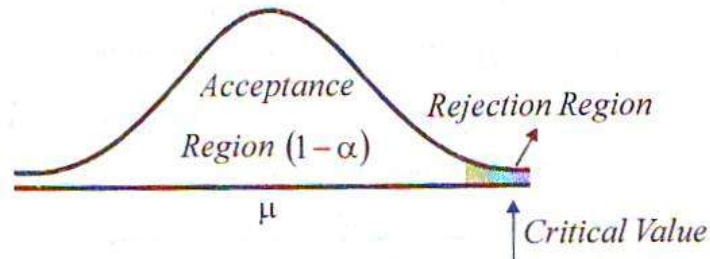


Alternative Hypothesis (H_1): $\mu \neq 90$ (i.e. either $\mu > 90$ or $\mu < 90$)

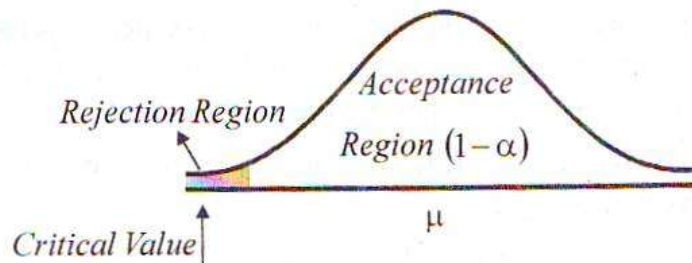
i) **One Tailed Test :** The one tail test is used in cases where it is considered that the population mean is at least as large as some specified value of mean or at least as small as some specified value of mean. There are two types of one tailed tests



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.



ii) **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 = α
H_0 is False	Wrong (Type II Error) Probability = β	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**

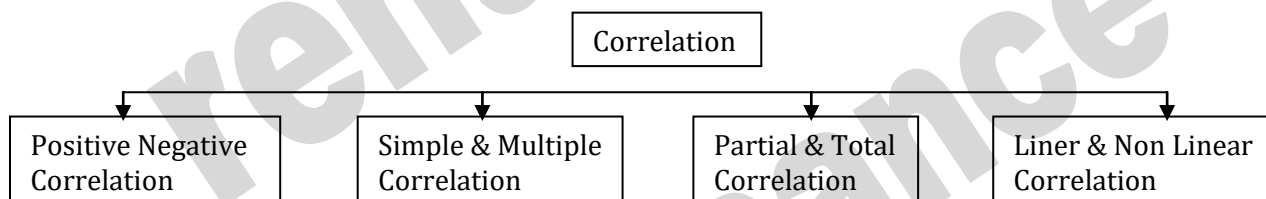
Introduction

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 (\uparrow or \downarrow) its demand also changed (\uparrow or \downarrow) so there is cause & 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 tree 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 $\uparrow \rightarrow$ Demand \downarrow
 Demand $\downarrow \rightarrow$ Price \uparrow
 So, both price & demand are affected by each other therefore we cannot tell in real sense which one is cause and which one is effect.

DEFINITIONS OF CORRELATION

1. “If 2 or more quantities vary in sympathy, so that movements in 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 connection”. **W.I 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. **Ya Lun chou**

TYPES OF CORRELATION


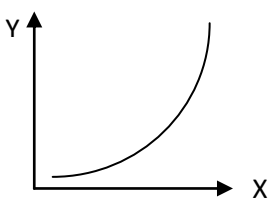


	POSITIVE CORRELATION	NEGATIVE CORRELATION
1	Value of 2 variables move in the same direction i.e. when increase/decrease in value of one variable will cause increase or decrease in value of other variable.	Value of 2 variables move in opposite direction i.e. when one variable increased, other variable decreases when one variable is decreased, other variable increase.
2	E.g. Supply & Price So, supply and price arecorrelated 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 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 Price = X_1 Price = X_2 Then Correlation between Y & X_1 Correlation between Y & X_2

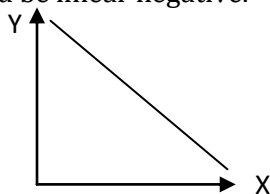
SIMPLE CORRELATION	MULTIPLE CORRELATION
In partial correlation though more than 2 factors are involved but correlation is studies only between to be constant. E.g. <div style="display: flex; align-items: center;"> <div style="margin-right: 10px;"> $Y \begin{cases} X_1 \\ X_2 \end{cases}$ </div> <div> $Y = \text{Demand}$ $X_1 = \text{Price}$ $X_2 = \text{Income}$ </div> </div>	In total correlation relationship between all the variables is studied i.e., none of item is assumed to be constant E.g. <div style="display: flex; align-items: center;"> <div style="margin-right: 10px;"> $Y \begin{cases} X_1 \\ X_2 \end{cases}$ </div> <div> $Y = \text{Demand}$ $X_1 = \text{Price}$ $X_2 = \text{Income}$ </div> </div>
If we study correlation between Y & X_1 & assume X_2 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, ceteris paribus assumption is relaxed in this case.

	LINEAR CORRELATION	NON-LINEAR CORRELATION																						
1	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.	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.																						
2	If the changed in 2 variables are in the same direction and in the constant ratio, it is linear positive correlation <table border="1" style="display: inline-table; margin-right: 20px;"> <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> 	X	Y	2	3	4	6	6	9	8	12	If the change in 2 variables is in the same direction but not in constant ratio, the correlation is non linear positive. <table border="1" style="display: inline-table; margin-right: 20px;"> <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> 	X	Y	50	10	55	12	60	15	90	30	100	45
X	Y																							
2	3																							
4	6																							
6	9																							
8	12																							
X	Y																							
50	10																							
55	12																							
60	15																							
90	30																							
100	45																							



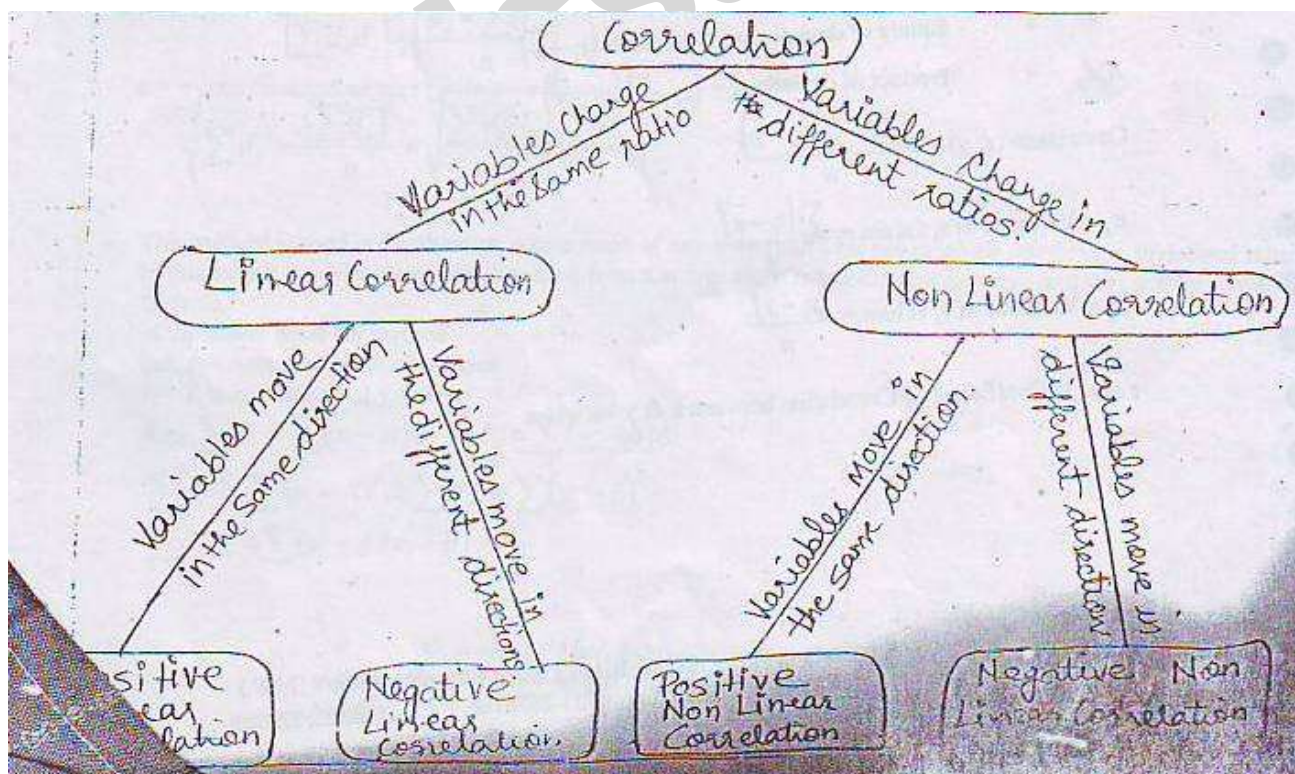
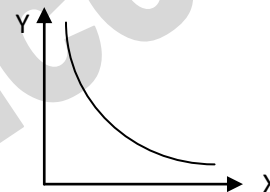
3 If changes in 2 variables are in the opposite direction but in constant ratio, the correlation is linear negative. For eg. every 5% ↑ is price of a good is associated with 10% decrease in demand the correlation between price and demand would be linear negative.

X	Y
2	21
4	18
6	15
8	12
10	9



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.

X	Y
80	50
55	60
50	75
90	130



TYPE - 1 [BASED ON KARL PEARSON'S COEFFICIENT OF CORRELATION]

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

- d_x = Deviations of x_i value from mean = $(x_i - \bar{x})$
- \bar{x} = Mean of x value [Average of X values] = $\frac{\sum x_i}{n}$
- n = No. of observations
- d_y = Deviation of y value from mean = $(y_i - \bar{y})$
- \bar{y} = Mean of y values = $\frac{\sum y_i}{n}$
- d_x^2 = Square of deviation of x values = $(x_i - \bar{x})^2$
- d_y^2 = Square of deviation of y values = $(y_i - \bar{y})^2$
- $d_x d_y$ = Product of deviations = $(x_i - \bar{x})(y_i - \bar{y})$

Covariance $(x,y) = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n}$

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



σ_y = Variance of y_i values = $\frac{(y_i - \bar{y})^2}{n}$
 r or r_{xy} = coefficient of correlation between x & y variables.

Direct Method for Karl Pearson's Coefficient of correlation

Direct Method for Karl Pearson's Coefficients of correlation (Product moment method)

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

Deviation from actual mean method

Deviation from actual mean method

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

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 moment method)

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

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 $\sum d_x = \sum (x_i - A)$ & $\sum d_y = \sum (y_i - B)$ &
 $\sum d_x^2 = \sum (x_i - A)^2$ & $\sum d_y^2 = \sum (y_i - B)^2$
 $\sum d_x d_y = \sum (x_i - A)(y_i - B)$

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 R^2 & coefficient of 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 -

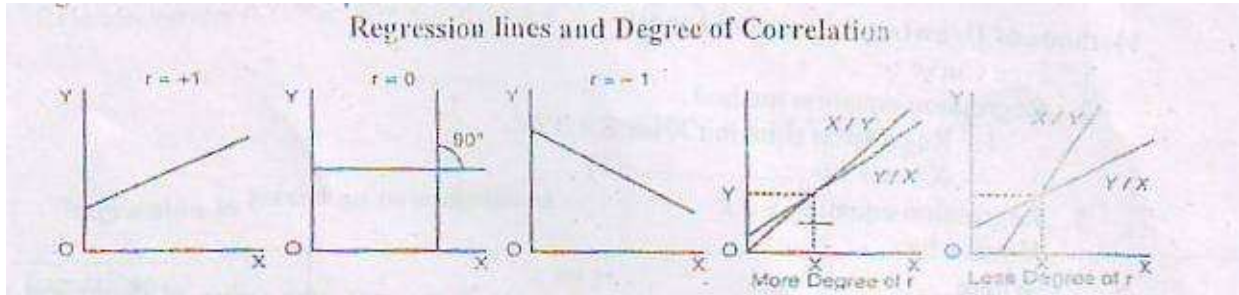
1. While constructing the lines of regression of x on y is treated as independent variables where as 'x' is treated as treated as dependent variable. This gives most probable values of 'X' for gives values of y. the same will be there for y on x.

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 ($r = 0$)> 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 ($r = \pm 10$ or 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 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 be 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 of 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 -

$$\Sigma(x = a + by) \text{ (by multiplying '}\Sigma\text{')}$$

$$\Sigma x = Na + b\Sigma y \tag{1}$$

$$\Sigma x (y = a + bx) \text{ (by multiplying } \Sigma x\text{)}$$

$$\Sigma xy = \Sigma xa + b\Sigma x^2 \tag{2}$$

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

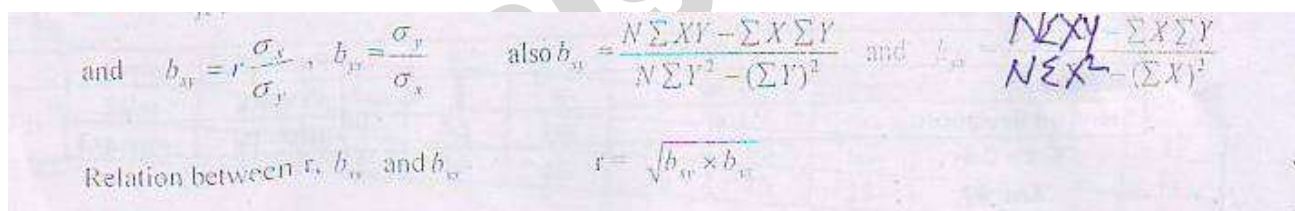
REGRESSION EQUATIONS -

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_{yx} (x - x) where b_{xy} = regression coefficient of Y on X





REGRESSION ANALYSIS

Regression is based on two equations -

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

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 or negative
- Both cannot be 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_{yx} - They all have same sign



renaissance

college of commerce & management

B.Com II Year. (Hons.)

Subject- Statistics

renaissance
renaissance
renaissance