ISBN: 978-93-47587-95-5

INTRODUCTION TO

STATISTICAL METHODS

DR. GIRISH MAHAJAN

Bhumi Publishing, India
First Edition: April 2026

Introduction to Statistical Methods

(ISBN: 978-93-47587-95-5)

DOI: https://doi.org/10.5281/zenodo.20110553

Dr. Girish Mahajan

Department of Agricultural Economics,

Krishi Vigyan Kendra- Bara- Hamirpur (H.P.)

Corresponding author E-mail: lovely_nickname@rediffmail.com

Bhumi Publishing

April 2026

Title: Introduction to Statistical Methods

Author: Dr. Girish Mahajan

First Edition: April 2026

ISBN: 978-93-47587-95-5

ISBN 978-93-475-8795-5

9 789347 587955

DOI: https://doi.org/10.5281/zenodo.20110553

Published by Bhumi Publishing,

a publishing unit of Bhumi Gramin Vikas Sanstha

Bhumi Publishing

Nigave Khalasa, Tal – Karveer, Dist – Kolhapur, Maharashtra, INDIA 416 207

E-mail: bhumipublishing@gmail.com

PREFACE

The book *Introduction to Statistical Methods* has been carefully designed to provide a clear, concise, and practical foundation in statistics for students and beginners across diverse disciplines. In an era driven by data and evidence-based decision making, statistical literacy has become an essential skill for understanding patterns, drawing meaningful conclusions, and solving real-world problems. This book aims to simplify statistical concepts while maintaining academic rigor, making it accessible to undergraduate learners and useful for self-study.

The structure of the book follows a logical progression from fundamental to more analytical aspects of statistics. It begins with Descriptive Statistics, which introduces methods for organizing, summarizing, and presenting data in an informative manner. This is followed by Probability Theory, where the foundational principles of probability are discussed to help readers understand randomness and risk.

The book then advances to Inferential Statistics, focusing on estimation, hypothesis testing, and drawing conclusions about populations based on sample data. The chapter on Sampling and Sampling Fundamentals highlights various sampling techniques and their importance in ensuring the accuracy and reliability of statistical analysis. Finally, Correlation and Regression explore relationships between variables and introduces predictive modeling techniques that are widely used in research and applied sciences.

Special emphasis has been placed on clarity of explanation, practical examples, and step-by-step approaches to problem-solving. The content is tailored to bridge the gap between theory and application, enabling readers to develop both conceptual understanding and analytical skills.

This book will be valuable for students of science, commerce, social sciences, and allied fields, as well as for anyone seeking to build a strong foundation in statistics. It is hoped that this work will inspire learners to appreciate the power and relevance of statistical methods in academic research and everyday life.

- Dr. Girish Mahajan

TABLE OF CONTENT

INTRODUCTION TO STATISTICS

Statistics: The word 'Statistics' comes from the Italian word 'statista' meaning 'statement' or the German word 'statistic' each of which means a political state. According to Croxton and Cowden have defined statistics as the collection, presentation, analysis and interpretation of statistical data. It is a science of inductive inference.

Inductive means to induce; deductive means to deduce or to derive; and inference means to conclude.

Aim or the objective of Statistics or Why should we study Statistics?

a. To study the population, we go for statistics.
b. To study the variation, we go for statistics.
c. To study the reduction in data, we go for statistics.

Functions of Statistics:

a. It presents facts in definite form.
b. It simplifies mass of figures.
c. It facilitates comparison.
d. It helps in formulating and testing hypothesis.
e. It helps in prediction.
f. It helps in formulation of suitable policies.

Five stages of Statistical Investigation:

a. Collection
b. Organisation
c. Presentation: diagrams and graph
d. Analysis
e. Interpretation

Limitations of Statistics:

a. It deals with population and not with individual.
b. Like physical law, statistical laws are not exact. But they are true on an average and in the long run.
c. Statistical methods or statistical tools are not applicable to qualitative type of data whereas, they are applicable to quantitative type of data. In other words, qualitative data are converted into quantitative data and then we use statistical tools.
d. Statistical tools must be handled by expert only. These are very dangerous tools in the hands of inexpert.

Scope of Statistics: It is applicable in all fields. There is a variation in all fields and whenever there is variation, we study statistics.

Uses of Statistics in Economics:

a. In solving various economic problems such as poverty, unemployment, disparities in the distribution of income and wealth, statistical data and tools play a vital role.
b. Analyses of population, land economics, economic geography are basically statistical in their approach.
c. Operational studies of public utilities require both statistical and legal tool of analysis.
d. Studies of competition, oligopoly, and monopoly require statistical comparison of market prices, cost and profits of individual firm.
e. Financial statistics are basic in the field of money and banking, short term credit, consumer finance and public finance.
f. Measures of GNP and input-output analysis.

Main Sources of the Origin of Statistics:

a. Government records.
b. Mathematics

Collection of Data: There are two methods for the collection of data. These are:

a. **Direct Method:** By personal contact, schedules, questionnaires, sampling technique.
b. **Indirect Methods:** It is a secondary data which can be gathered from government records.

Representation of Statistical Data:

a. **Line Diagram:** Vertical lines on x-axis representing line diagrams.
b. **Bar Diagram**: These bars are having the uniform width.
c. **Pie Diagram**
d. **Picto Graph**: when the data are given in picter.
e. **Statistical Diagram:** Whenever quantitative data are representing through certain picture, it is called statistical diagram.

Chapter 1

DESCRIPTIVE STATISTICS

Definition of Key Terms

- **Unit of Analysis**: Also referred to as cases. The most elementary part of what is being studied or observed. Example: Individuals, households, Countries, states, firms, industries, etc.
- **Variables:** A real valued function which takes any value. In other words, concepts, characteristics, or properties which can vary or change, from one unit of analysis to another. Please note that all variables must vary, if there is no variation among the different cases then it is not a variable. Some examples of variable include gender, social class, education, age, etc.
- **Dependent variable-DV:** If one variable depends upon or is a consequence of other variable, it is termed as dependent variable.
- **Independent Variable-IV:** A variable that is antecedent to the dependent variable is termed as independent variable. For instance, if we say height depends upon age, then height is dependent variable and age is independent variable. Another example is yield of wheat is dependent upon fertilizer consumption then,

$$Y= f (X)$$

In the above equation, Y is wheat yield in Kg/ha which is a dependent variable and X is the amount of fertilizer in Kg/ha which is a independent variable. Thus, from the above equation we say that yield of wheat is dependent upon fertilizer intake.

- **Random variable/ Variate:** A random variable is a real valued function which can be defined on the outcome of an experiment. In other words, when a variable takes certain value at certain probability. It is also known as stochastic variable, chance variable. For example, if we throw a coin or dye.
- **Continuous variable:** A variable which takes any value between two limits. For example-Height, weight, etc.
- **Discrete variable**: A variable which takes specified value. For example- no. of children in a family, no of grains in a cob whether in fraction or in whole number.
- **Extraneous variable:** The independent variables that are not related to the purpose of study, but may affect the dependent variable are termed as extraneous variable. E.g.

$$Y= f (HYV, AREA, FERTILIZER) \text{ and also rainfall.}$$

Here, rainfall is not included, so it may be extraneous variable.

A study must always be so designed that the effect on the dependent variable is attributed entirely to the independent variable(s) and not to some extraneous variable or variable(s).

Level of measuring variables:

(a) Nominal: Nominal scales are adopted for non-quantitative (containing no numerical implication) labeling of variables which are unique and different from one another.

Types of Nominal Scales:

Dichotomous/dummy: A nominal scale that has only two labels is called dichotomous. E.g. yes/no

Nominal with order: the labels on a nominal scale are arranged in an ascending or descending order. E.g. Excellent, Good, Average, Poor, Worst.

Nominal without order: Such nominal scale has no sequence. E.g. Black, White

(b) Ordinal: An ordinal variable has qualitative categories that are ordered in terms of degree or magnitude. E.g. A nominal variable includes class or degree obtained. The variable degree obtained may include the following categories: None, High school diploma, College/University degree, Masters, Advanced degree Ph.D. All of these categories are qualitative and are ordered in terms of the amount of education each individual has completed.

Another example-At Amazon .in every product has a customer review section where the buyers rate the listed products according to their buying experience, product features quality, usage, etc. The rating so provided are as follows: 5Star-Excellent; 4Star-Good; 3Star-Average; 2Star-Poor; 1-Star-Worst.

(c) Interval Scale: An interval scale is also called as cardinal scale which is the numerical labeling with the same difference among the consecutive measurements units. With the help of this scale, researcher can obtain a better comparison between the objects. E.g. A survey conducted by an automobile company to know the number of vehicles owned by the people living in a particular area who can be its prospective customers in future. In adopting the interval scaling technique for the purpose and provided the units as 1, 2, and 3,4,5,6 to select from. In this scale every unit has the same difference, i.e.1, whether it is between 2 and 3 or 4 and 5.

(d) Ratio Scale: One of the most superior measurement techniques is the ratio scale. Similar to an interval scale, a ratio scale is an abstract number system. It allows measurement at proper intervals, order, categorization and distance with an added property of originating from a fixed zero. Here, the comparisons can be made in terms of the acquired ratio.

Descriptive Statistics:

Descriptive statistics are often used to describe variables. Descriptive statistics are performed by analyzing one variable at a time (univariate analysis). All researchers perform these descriptive statistics before beginning any type of data analysis.

Frequency: A number of times a thing happens.

Frequency distribution: If a large number of data are summarized into different classes and the class frequency are also given. Then, this representation of classes along with class frequency is called frequency distribution or frequency tables.

Weight/ classes	40	50	60
16-18 years	10		
18-20 years		20	
20-22			5

Table shows that for 35 students (in total) 10 is the frequency of the students having weight 40 Kg and age 16-18 and so on. This is called frequency distribution or distribution table. 16 are the lower limit of the class and 18 is the upper limit of the class.

Open Class: When ever either upper or lower or some time both limits are not given, we call them as open class. E.g.

Classes	Frequency
0-18 years	10
18-20 years	20
20 and above	22

Class mid point: It is the middle point of the class and is calculated by:

(Upper limit+ lower limit)/2

Class interval or width of the class: Upper limit minus lower limit is called class interval.

Absolute Frequency: This tells you how many times a particular category in your variable occurs.

Relative Frequency: This tells you the percentage of each category/ value relative to the total number of cases.

Cumulative Frequency: This is simply a commutation of the relative frequency for each category/ value

For Example: Table provides an example of a frequency table for an ordinal variable (note it is ordinal because the categories are qualitative and ordered) named socioeconomic class. If there were numbers assigned to each category that were also ordered, we could treat this as an interval level variable.

Frequency Table: Socioeconomic Class

Socioeconomic Class	Frequency	Per cent	Cumulative Percent
Upper	50	7.14%	7.14%
Upper Middle	150	21.43%	28.57%
Middle	300	42.86%	71.43%
Lower Middle	150	21.43%	92.86%
Lower	50	7.14%	100%
Total	700	100%	

Cross tabulation: This is also referred to as grouped frequency table for two variables. A cross tabulation simply presents the absolute frequency broken down by categories of two or more variables. It is also possible to find percentages in these types of tables. For instance, using the example below, we can find the percentage of young people that listen to music.

Cross tabulation of music preference and age

	Age		
Preference	Young	Middle Age	Old
Music	14	10	3
News-Talk	4	15	11
Sports	7	9	5

There are four characteristics by which we can compare two or more than two distribution. These are:

1. Central Tendency or central value or central class
2. Dispersion
3. Skewness
4. Kurtosis

We will discuss each one by one below.

1. Measures of Central Tendencies:

Central Tendency: Central tendency is the value of the variable which represents the distribution thoroughly. It may or may not be one of the values of the distribution. Central tendency must be measured in certain unit. E.g. Class Representative.

Measures of Central Tendency: These are three namely mean, median, and mode.

Mean: Mean is of different types:

i) Arithmetic Mean(AM)
ii) Geometric Mean (GM)
iii) Harmonic Mean(HM)
iv) Weighted Mean(WM)
v) Pooled Mean(PM)

(i) Arithmetic Mean (AM): It is calculated by

$$\bar{x} = \frac{X_1 + x_2 + x_3 \ldots\ldots\ldots + x_n}{N}$$

$$\text{Or} \quad \bar{x} = \frac{f_1x_1 + f_2x_2 + \ldots\ldots\ldots f_nx_n}{f_1 + f_2 + \ldots\ldots\ldots\ldots + f_n}$$

$$\text{Or} \quad \bar{x} = \frac{\sum_{I=1}^{n} f_i x_i}{N}$$

Advantage of Arithmetic Mean: Even one or two values are missing; even then we can calculate arithmetic mean.

Disadvantage of Arithmetic Mean: It gives more weight age to the extreme values.

(ii) Geometric Mean (GM): GM of 'n' number is the n^{th} root of the product. Suppose the numbers are x_1, x_2, x_3……..,x_n, then

GM= ($x_1.x_2$……..x_n) $^{1/n}$

Take log on both sides, we get:

Log GM= 1/n {log x_1+ log x_2+……+ log x_n}

If x_1 is having frequency f_1

X_2 is having frequency f_2

.x_n is having frequency f_n then,

$$GM= (x_1^{f1}.x_2^{f2}\ldots\ldots\ldots x_n^{fn})^{1/\sum fi=N}$$

n

$$\text{Or Log GM} = 1/N \sum (f_i \log x_i)$$

I=1

For example, GM of 2, 4, 8 is $(2.4.8)^{1/3} = (64)^{1/3} = 4$

Merits of Geometric Mean:

a) It is rigidly defined.

b) It is based on all the observation of a series.

c) It is capable of further algebraic treatment.

d) GM is not much affected by the fluctuation of sampling.

Demerits of Geometric Mean:

a) It is neither easy to calculate nor easy to understand.

b) If any value in a series is zero, then we cannot calculate the Geometric Mean. E. g. GM= 2x4x0x8=0

c) If one of the value or the item in a series is negative then also we cannot find the GM.

d) Like arithmetic average, it may be a value which does not exist in a series.

e) It gives more weight age to the smaller value.

(iii) Harmonic Mean (HM): It is the reciprocals of AM's of the reciprocals. Suppose there are 'n' observation, then

$$\text{AM's of reciprocals} = \frac{1/x_1+1/x_2+\ldots\ldots+1/x_n}{N}$$

$$\text{Or } 1/H = \frac{1/x_1+1/x_2+\ldots\ldots+1/x_n}{N}$$

$$\text{Or } H = \frac{N}{1/x_1+1/x_2+\ldots\ldots+1/x_n}$$

Example: Find out the AM, GM, &HM of 2, 4, 8.

Solution: AM = (2+4+8)/3 =14/3 =4.66

GM = $(2.4.8)^{1/3} = (64)^{1/3} = 4$

HM = 1/H = (1/2+1/4+1/8)/3 = (4+2+1)/24 = 7/24

or H = 24/7 =3.42

It is clear from the above example that AM ≥ GM ≥ HM

When we should make use of Harmonic Mean?

Whenever we deals with quantum of time and rate, then we use HM. E.g. A cyclist cover 1st mile @ 3miles/hour and 2nd miles @ 4miles/hour, then

HM = 1/H = (1/3 +1/4)/2 = 7/24 = 3.42

HM gives more weight age to the smaller values.

Merits of Harmonic Mean (HM):

i) It is rigidly defined.
ii) It is based on all the observation of the series.
iii) Capable of further algebraic treatment.
iv) It is not much affected by the fluctuation of sampling.
v) It gives more weight age to the smaller values.
vi) It measures the relative change.

Demerits of Harmonic Mean:

i) It is not readily understood nor can be calculated with ease.
ii) It gives very high weight age to smaller value.
iii) It is usually a value which does not exist in a series.
iv) Generally it is not a good representative of statistical technique.

(iv) Pooled Mean (PM): Pool up all the means and then take average. Suppose $n_1 = \bar{x}_1$, $n2 = \bar{x}$,…….. $n_n = \bar{x}_n$

$$\textbf{Therefore, PM} = \bar{x} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2 + \ldots + n_n \bar{x}}{n_1 + n_2 + n_3 + \ldots + n_n}$$

$$\text{or PM} = \bar{x} = \frac{\sum_{i=1}^{n} n_i \bar{x}_i}{\sum_{i=1}^{n} n_i}$$

(v) Weighted Mean (WM): Suppose weight w_1 is assign to x_1, weight w_2 is assign to x_2 and weight w_n is assign to x_n then, weighted mean is calculated by:

$$WM = \frac{w_1x_1 + w_2x_2 + \ldots\ldots\ldots\ldots + w_nx_n}{w_1 + w_2 + \ldots\ldots\ldots\ldots + w_n}$$

$$\text{Or} \quad WM = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

Difference between mean and average: In means, we have different types of means like AM, GM, HM, PM, WM. While average gives only average mean.

Average: It is an attempt to find one single figure to describe the whole of figures. In other word, an average value is a single value within the range of the data that is used to represent all the values in the series. Since average is somewhat within the range of the data, it is also called a measure of central value/ tendency.

Characteristics of a good average:

a) It should be rigidly defined.
b) It should be based on all the observation of the series.
c) It should be capable of further algebraic treatment.
d) It should be easy to calculate and simple to flow.
e) It should not be affected by the fluctuation of sampling.

Drawbacks of Arithmetic Average:

a) It gives greater importance to higher items of a series and lesser importance to smaller items.
b) It gives fallacious conclusion.
c) Arithmetic average some time gives such results which appear almost absurd.

Mathematical Properties of Arithmetic Average:

i) The sum of the deviation of the items from the mean is always zero.
ii) The sum of the squared deviation of the items from the mean is less than the sum of the squared deviation of the items from any other value.

Cumulative Series: Cumulative series can be either “more than type” or “less than type” In more than type the frequencies are cumulative upwards so that the first class interval has the highest frequency and it goes on decline in the subsequent class. In less than type, the frequencies are cumulative downwards so that the first class has the lowest frequency and the subsequent class has the higher cumulative frequency. Hence we can say frequency above or below a particular point is the cumulative frequency.

Change of origin and change of scale: Whenever we are adding or subtracting the unit in any item or variable, then it is called change of origin and when we divide the unit in a particular item or variable then it is called change of scale.

Median: **Median is the value of the variables which divide the total frequencies into two equal half such that half of this lies above this value of variable and other half below this value of variable**. (Either in ascending or descending order). E. g.

X =1, 2, 4, 6, 7.

F = 1, 1, 1, 1, 1.

Hence 4 is the median value.

If x = 1, 2, 4, 6, 7, 8, and

f =1, 1, 1, 1, 1, 1 then take mean = (4+6)/2 = 5 Therefore, 5 is the median value.

If the date is given in distribution form then the median is calculated as:

$$M = L + \frac{(N/2 - C)}{f} h$$

Where, L = Lower limit of the median class,

N = Total frequency,

C = Cumulative frequency just above the median class,

f = Frequency of the median class,

h = Width of the median class.

Merits of Median:

i) Even if the value of extreme is not known, median can be calculated if the number of item is known.

ii) It can be easily calculated and understood without any difficulty.

iii) It is rigidly defined.

iv) It is not affected by the values of extreme items.

v) It gives best results in a study of those phenomenon's which are incapable of direct quantitative measurement.

Drawbacks of Median:

i) It is not suitable of further algebraic treatment.

ii) It is more likely to be affected by the fluctuation of sampling then the average.

iii) Arrangement of items in ascending order is somewhat difficult.

Mode: Mode is the value of the variable which occurs more frequently.

E.g. 1, 1, 2, 2, 2, 3, 4, 5. Here, 2 is the value of the mode.

In group distribution, mode can be found out by:

$$\text{Mode} = L + \frac{\blacktriangle_1}{\blacktriangle_1 + \blacktriangle_2} h$$

Where L = Lower limit of the model class;

$\blacktriangle_1$ = Difference in the frequency of the model class and the preceding class;

$\blacktriangle_2$ = Difference in the frequency of the model class and the following class;

h = width of the model class.

Model Class: It is that class in which mode lies.

Suppose by chance, if there are two equal frequency's i.e. 4&5 then, we use approximation method

Mean – Mode = 3 (Mean- Median).

Merits of Mode:

i) It possesses the merits of simplicity.

ii) It is commonly understood.

iii) It cannot be a value which cannot found in a series.

iv) It is affected by the values of extreme items.

v) For the determination of mode, it is not necessary to know the values of all the items of series.

Drawbacks of Mode:

i) Mode is ill-defined, indeterminate and indefinite.

ii) Not based on all the observation of a series.

iii) Mode is not capable of further mathematical treatment.

iv) Mode may be unrepresentative in many cases.

v) In many cases, it may be impossible to get a definite value of mode.

Note: Mode is the best measurement of normal data i.e. data normally distributed and median is the best measure of skewed distribution and arithmetic mean is affected mostly with extreme values. Standard deviation is most important measure of variation for any type of data.

2. Dispersion:

Suppose

A	10	15	75	1000	Average-275
B	200	300	150	450	Average-275

Here, average figure of A&B is equal i.e. 275. Here, the measure of central tendency fails and in this case we go for dispersion. Dispersion means spread or scattered of values around the central tendency is called dispersion. In the above example workers in factory 'A' are more spread than in factory 'B'. Dispersion is also measured in certain unit.

Measures of Dispersion:

i) Range;

ii) Inter quartile range;

iii) Inter decile range;

iv) Quartile deviation/ semi inter quartile range;

v) Standard deviation;

vi) Mean deviation from mean;

vii) Mean deviation from mode;

viii) Mean deviation from median.

Why should we go for dispersion? Whenever Central tendency fails to resort two or more than two distribution, then we go for dispersion. For instance, let there are two factories 'A' &'B'

A	Rs.50	Rs.100	Rs.75	Rs.175	Average- Rs. 100
B	Rs. 10	Rs. 50	Rs.40	Rs. 300	Average- Rs. 100

From here we can draw the conclusion that although the average wages are the same. But different workers in these two factories possess different wages for their work.

i) Range: Difference between higher limit and the lower limit in a class e.g. 300-10 = 290

Note: For calculating range, the values of the variables are taken into account and the frequencies are completely ignored.

Coefficient of range: The relative measure corresponding to range is called coefficient of range.

Coefficient of range = (L-S) / (L +S) where, L = Largest item and S = Smallest item

Coefficient: means a pure number that is independent of the unit of measurement.

Quartile: The value which divide the data into four equal parts. In other words, it is the value of the variable below which 25% of the frequency lies and is known as first quartile or lower quartile. (Q_1).

Second Quartile: it is the value of the variable below which 50% of the frequencies lies and is denoted by (Q_2).

Third Quartile: It is the value of the variable below which 75% of the frequencies lies and is denoted by (Q_3).

F 0----------25(Q_1) ----------50(Q_2)----------75(Q_3)--------100

Caution: We don't go for the value of the variable below which 100% of the frequency lies.

ii) Inter quartile range: Difference between the third quartile and the first quartile i.e. Q_3-Q_1

Quartile deviation/ semi inter quartile range:

$Q = (Q_3-Q_1)/2$

How to calculate quartile if the data is given in distribution forms?

Quartile (Qi) = L + { iN/4—C}.h /f

i= 1,2,3

When i =1, it is first quartile;

When i =2, it is second quartile;

When i =3, it is third quartile.

L = lower limit of the class in which quartile lies;

N = Total frequency;

C = Cumulative frequency just above the quartile class;

h = Width of the quartile class;

F = Frequency of the quartile class.

The above formula is valid only when data are to be arranged in ascending order and the cumulative frequency is of "less than type".

In a series of individual observation and in a discrete series, the value of lower and upper quartile would be the value of $\{(N + 1)/4\}^{th}$ and $3\{(N +1)/4\}^{th}$ item respectively.

Quartile Class: it is that class in which the desired quartile lies.

Decile: These are the values which divide a series into ten equal parts. In a series of individual observation and in a discrete series, the value of lower and upper decile would be the value of D_1 = value of $\{(N + 1)/10\}$th and D_2 =value of $2\{(N +1)/10\}^{th}$ and D_9 = value of $9\{ (N +1)/10\}^{th}$ item respectively.

Note: In a continuous series the value of quartile and decile is calculated by N/4 and N/10 respectively.

iii) Inter decile range: it is the value of the variable below which ten % of the frequency lies and is represented by D.

Inter decile range: (D) = D_9 -- D_1

If the data is given in distribution forms, then

Di = L + { iN/10—C}.h /f

Where, i= 1,2,3------9;

L = lower limit of the decile class;

N = Total frequency;

F = Frequency of the decile class;

C = Cumulative frequency just above the decile class;

H = Width of the decile class.

iv) **Mean deviation from mean:** Deviation from mean and then take the mean

$$\textbf{MD from mean} = \frac{1}{N}\sum f_i \left| x_i - \bar{x} \right|$$

$$\text{Where, } N = \sum_{i=1}^{n} f_i$$

Here, deviation from mean = $(x_1 — \bar{x})$, $(x_2— \bar{x})$,----------$(x_n— \bar{x})$,----- $(x_i--- \bar{x})$

v) Mean deviation from mode:

$$\textbf{MD from mode} = \frac{1}{N}\sum f_i \left| (x_i - \text{mode}) \right|$$

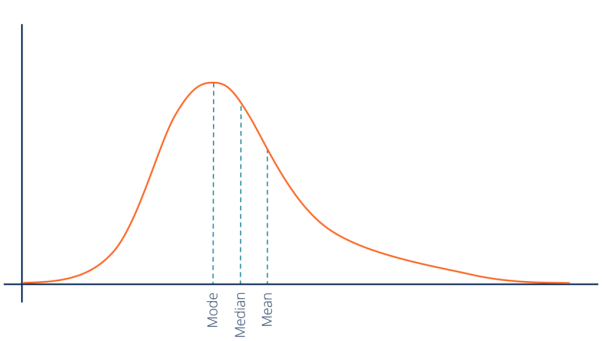
vi) Mean deviation from median:

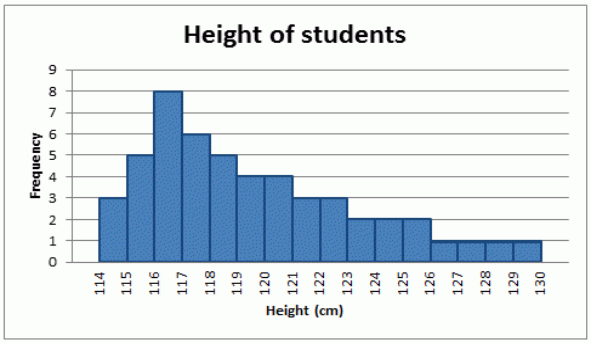
$$\textbf{MD from median} = \frac{1}{N}\sum f_i \left| (x_i - \text{median}) \right|$$

3. Skewness:

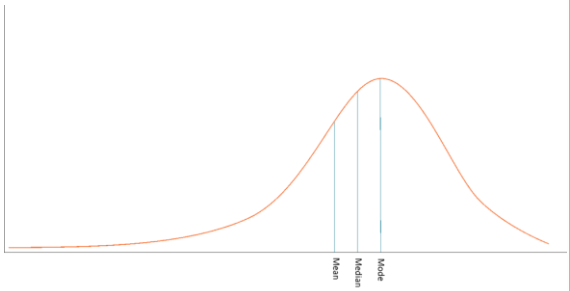
Departure from symmetry is known as skewness and when a distribution is not symmetrical (or is asymmetrical) it is called skewed distribution. It is of two types:

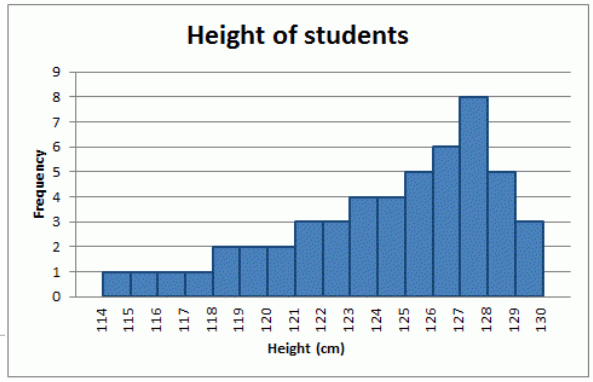
i) Positive skewness/positively skewed distribution: In a positively skewed distribution, the value of mean is maximum and that of mode is least- the median lies in between the two. In this case, excess tail is on the right hand side.



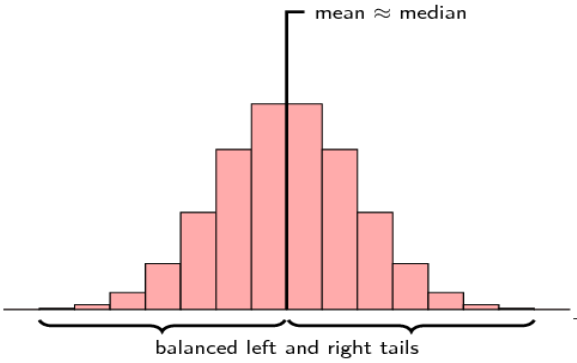


ii) Negative skewness/ negatively skewed distribution: In a negatively skewed distribution, the value of mode is maximum and that of mean least- the median lies in between the two. Here, the excess tail is on the left hand side.

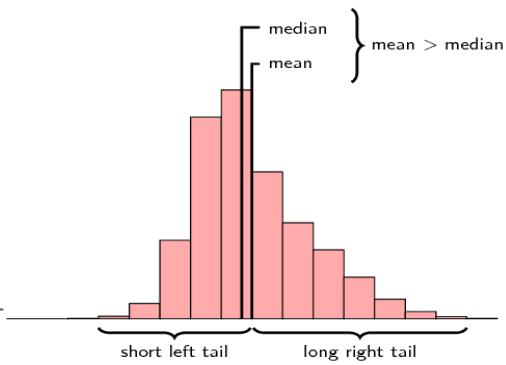




iii) Symmetrical and Asymmetrical Distribution: Two distribution may have the same mean and standard deviation but, may differ widely in their overall appearance as can be seen from below



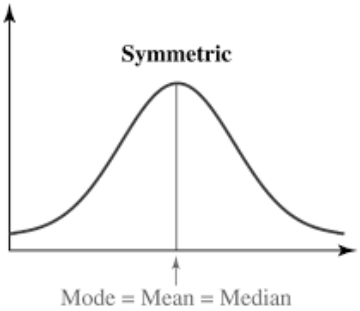
Symmetrical Distribution



Asymmetrical distribution

Measures of skewness help us to distinguish different types of distribution. Skewness refers to lack of symmetry.

Symmetrical Distribution:



It is clear from the above diagram that in a symmetrical distribution, the value of mean, median and mode coincide. The spread of the frequencies is the same on both sides of the centre point of the curve.

Measures of Skewness: It is measured by coefficient of skewness or moment coefficient of skewness.

Skewness - Measures and Interpretation

Skewness is a key statistical measure that shows how data is spread out in a dataset. It tells us if the data points are skewed to the left (negative skew) or to the right (positive skew) in relation to the mean. It is important because it helps us to understand the shape of the data distribution which is important for accurate data analysis and helps in identifying outliers and finding the best statistical methods to use for analysis. In this article, we will see skewness, different types of skewness and its core concepts.

Types of Skewness

Skewness describes the direction and degree of asymmetry in a dataset's distribution. Various types are as follows:

1. Positive Skewness (Right Skew)

In a positively skewed distribution, the right tail is longer than the left which means most data points are on the left with a few large values pulling the distribution to the right.

Relationship:

Mean > Median > Mode

Examples: Income distribution, exam scores and stock market returns.

2. Negative Skewness (Left Skew)

In a negatively skewed distribution, the left tail is longer which means most data points are on the right with a few smaller values pulling the distribution to the left.

Relationship:

Mean < Median < Mode

Examples: Test scores on easy exams, age at retirement and gestational age at birth.

3. Zero Skewness (Symmetrical Distribution)

Zero skewness shows a perfectly symmetrical distribution where the mean, median and mode are equal. In a symmetrical distribution, the data points are evenly distributed around the central point.

Relationship:

Mean = Median = Mode

Example: A perfectly balanced dataset with equal frequencies of all values.

Tests of Skewness

There are several ways to find the skewness of a dataset which can help to find whether the data is positively skewed, negatively skewed or roughly symmetric. Below are some common methods used to measure skewness:

1. Visual Inspection

This is the simplest and quickest method for assessing skewness by creating a histogram or a density plot of the given data.

- If the plot has a long tail on the right, the data is positively skewed (right-skewed).
- If the plot has a long tail on the left, the data is negatively skewed (left-skewed).
- If the plot is roughly symmetric, the data has no skewness (zero skew).

2. Skewness Coefficient (Pearson's First Coefficient of Skewness)

This is a numerical measure of skewness based on the relationship between the mean and mode. It helps us to find if the data is skewed when the mean and mode are not equal.

Formula: Skewness=Mean−Mode *Skewness=Mean−Mode*

- Positive Skew: If the mean is greater than the mode, the skewness is positive.
- Negative Skew: If the mean is smaller than the mode, the skewness is negative.
- Zero Skew: If the mean is equal to the mode, the skewness is zero which indicates a symmetric distribution.

3. Skewness Based on Quartiles

This method checks the distances between the quartiles to find skewness. If the quartiles are not equidistant, it suggests skewness:

- The third quartile (Q3) minus the median (Me) should ideally be equal to the median (Me) minus the first quartile (Q1) in a symmetric distribution.
- If this condition is not met, it shows either a positive or negative skew which depends on which side is longer.

Interpretation of skewness:

- **Skewness = 0:** The distribution is symmetric means the mean, median and mode are equal.
- **Skewness > 0:** The distribution is positively skewed (right-skewed) with the tail on the right side longer than the left.
- **Skewness < 0:** The distribution is negatively skewed (left-skewed) with the tail on the left side longer than the right.

Difference between Dispersion and Skewness

While dispersion and skewness may seem similar but they measure different aspects of data distribution. Dispersion refers to the extent to which data points are spread out from the central

value (mean or median). Dispersion helps understand the variability of data while skewness helps to identify the shape and asymmetry of data.

4. Kurtosis

It refers to the relative flatness of the top of the curve as compared to symmetrical curve.

Another characteristic of a frequency distribution is Kurtosis.

Kurtosis refers to the flatness or the peakedness of a distribution.

A distribution can be Leptokurtic, Mesokurtic or Platykurtic.

Leptokurtic Distribution:

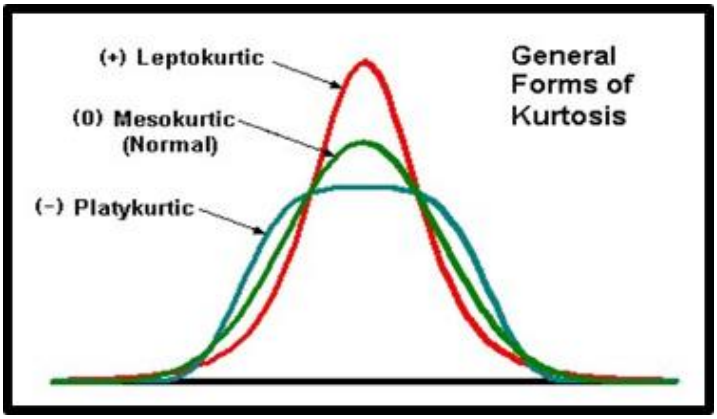
A distribution with a high peak. It is more peaked than the normal distribution.

Mesokurtic Distribution:

A distribution with a medium peak and resembles a bell shape. It is also called a normalcurve.

Platykurtic Distribution:

A distribution with low peak. It is flatter than the normal curve.



The standard value of kurtosis is taken as 3 and the curve with value of kurtosis less than 3 are called platikurtic curve and the curve with value of kurtosis more than 3 are called leptokurtic. In normal or mesokurtic curve, the value of kurtosis is 3. The bigger the value of kurtosis in a frequency distribution, greater its departure from normality.

Measures of Kurtosis: is known as the moment coefficient of kurtosis or coefficient of kurtosis. Kurtosis is equal to the fourth moment about the mean divided by square of the second moment about the mean.

Condition for symmetrical distribution: Coefficient of skewness is zero and coefficient of kurtosis is 3. In symmetrical distribution, the value of mean, median and mode coincide. The spread of frequency is same on both side of the center point of the curve.

Measures of Variability

Measures of variability are defined as the dispersion (or deviation) away from the mean for each variable. Measures of variability only exist for internal level variables. There are four measures of variability- range, standard deviation, variance, and coefficient of variation. We will discuss each one by one below:

a) **Range**: The range is found by taking the highest value of a variable minus the lowest value of that variable.

b) **Standard deviation**: The standard deviation exists for all interval variables. It is the average distance of each value away from the sample mean. The larger the standard deviations, the farther away the values are from the mean; the smaller the standard deviation, the closer the values are to the mean. Suppose you passed out a questionnaire asking randomly selecting individuals to rate the Prime Minister Modi's job performance on a scale from 1 to 10. And suppose you find that average these individual give the PM a rating of 5.8 and suppose this variable i.e. employment has a standard deviation of 1.2. This mean that on an average, each rating of the PM is approximately 1.2 points away from 5.8 (the sample mean).

Standard deviation is the square root of the mean of the squared deviation from arithmetic mean. It is also known as root mean squared deviation for the reason that it is the square root of the mean of the squared deviation from arithmetic mean introduced by Karl Pearson in 1823. A smaller value of SD means a high degree of uniformity of the observation as well as homogeneity of a series; a larger SD means just the opposite. It is extremely useful in judging the representativeness' of the mean.

c) **Variance:** it is nothing but the square of standard deviation i.e. variance = σ^2

$$\text{Where } \sigma = \text{Standard deviation and Variance} = \sigma^2 = \frac{\sum(x_i - \bar{x})^2}{N}$$

The variance cannot be interpreted as meaning anything other than the standard deviation squared.

d) **Coefficient of variation:** developed by Karl Pearson is the most commonly used measure of relative variation. It is used in such problem where we want to compare the variability of two or more than two series. It is unit less as it is the ratio.

CV = (Standard Deviation / mean) x100

It measures the spread i.e. percentage in variation. More the variation more will be the heterogeneous field; less the variation, more will be the homogeneous field. E.g. For Boys, CV=4.41 and for Girls, CV = 3.08

In this example CV is more for boys; hence height of boys shows greater variation. In field CV varies from 15-20%

Standard Error: SE is the standard deviation in the sampling distribution of desired statistics. Standard error is used only for sampling distribution. Standard error plays a larger role in testing for significance, and can drastically affect the output. For instance, large standard error will cause variables to be insignificant, which may indicate an incorrect use of a statistical method or analysis. The standard error (SE) of the mean measures the accuracy of a sample mean compared to the true population mean. It is calculated by dividing the sample standard deviation by the square root of the sample size.

Difference between standard error and standard deviation: Standard deviation is concerned with original values while Standard error is concerned with statistics computed from the sample of original values.

Difference between variance and standard deviation: Both the variance and SD are the measures of the variability in a population. These two measures are closely related as variance $=^2$. Variance is average squared deviation from the arithmetic mean and SD is the square root of the variance. The smaller the value of variance (σ^2) less the variability or greater the variability in a population.

Difference between mean deviation and standard deviation:

i) Algebraic signs are ignored while calculating the mean deviation whereas in the calculation of SD, signs are taken into account.

ii) MD can be computed from either mean or mode while, the SD is always computed from arithmetic mean.

Conclusion

Based on the introduction to descriptive statistics, the primary conclusion is that descriptive statistics are used to summarize, organize, and simplify large datasets into meaningful, interpretable, and manageable summaries without drawing conclusions beyond the specific data analyzed. Key conclusions and takeaways from the introduction include: they transform raw, complex data into easy-to-understand formats (tables, graphs, and numerical metrics); they describe data through three main categories: distribution (frequency), central tendency (mean, median, mode), and variability/dispersion (range, standard deviation, variance); a critical conclusion is that descriptive statistics only describe the collected sample or population data and cannot be used to make inferences, predictions, or generalizations about a larger population; and they serve as the first step in data analysis, allowing for the detection of outliers and errors, which helps to prepare data for more complex inferential statistical techniques.

Chapter 2

PROBABILITY THEORY

Probability:

In common parlance, the term probability refers to the chance of happening or not happening of an event. The moment we use the word chance, we indicate that there is an element of uncertainty about the statement that has been made or is being made. Thus, we say that (a) The probability of India winning a hockey match against Pakistan is poor, (b) or the chance that India will put a man on the moon are low.

Here, we are making statement about which we are not certain. There is an element of uncertainty associated with each of the above statements. We have not assigned any numerical value to these statements. The theory of probability provides a numerical measure of chance is called probability. It enables us to take decision under conditions of uncertainty with a calculated risk. E.g. if we throw a coin, the chance of getting head or tail is 1/2 and if we throw a dice, the chance of getting 1 or 2 or 3 or 4 or 5 or 6 is 1/6.

Random Experiment: It is an experiment which if conducted repeatedly under homogeneous conditions does not give the same result. The result may be any one of the various possible "outcomes". Here, the result is not unique (or the same every time). For instance, if a disc is thrown it would not always fall with no. 6 up. It would fall in any one of the six ways which are possible i.e. with any one of the six numbers on the disc.

Trial and Event: The performance of a random experiment is called trial and the outcome an event. Thus throwing of a disc would be called a trial and the result (falling with numbers 1, 2,3,4,5, or 6) an event.

Exhaustive Cases or Event: All possible outcomes of an event are known as exhaustive cases. In the throw of a disc the exhaustive cases are six as the disc has only six faces each marked with different numbers. However, if two disc are thrown, the exhaustive cases are 36 (6x6) as there are 36 ways in which the two disc can fall. Similarly, the number of exhaustive cases in the throw of two coin would be four (2x2) - HH, TT, HT, TH (if H stands for head and T for tail).

Mutual Exclusive Cases or Event: Two or more cases are said to be mutually exclusive if the happening of any one of them excludes the happening of all others in a single experiment. Thus in the throw of a disc, the event 5 & 6 are mutually exclusive, if the event 5 happens no other event is possible in the same experiment. Here, one and only one of the events can take place excluding all others.

Equally Likely Cases or Event: Two or more event is said to be equally likely if the chance of their happening is equal i.e. there is no preference of any one event over the other. Thus, in a

throw of a disc, the coming up of 1, 2, 3, 4, 5, or 6 is equally likely. Likewise, in the throw of a coin the coming up of head or tail is equally likely.

Independent Events: An event is said to be independent if it's happening is not affected by the happening of other events. Thus, in the throw of a disc repeatedly, the coming up of 5 on the first throw is independent of the coming up of 5 again in the second throw.

Dependent Events: If we are successively drawing cards from a pack (without replacement) the event would be dependent. The chance of getting a king in the first draw is 4/52 (as there are four kings in a pack). If this card is not replaced before the second draw, the chance of getting a king is again 3/51 as there are now only 51 cards left and they contain only 3 kings.

If however the card is replaced after the first draw, the event would remain independent. In each of the successive draws, the chance of getting a king would be 4/52.

Favorable Cases: The numbers of outcomes which result in the happening of a desired event are called favorable cases. Thus in a single throw of a disc, the number of favorable cases of getting an odd number are three 1, 3, & 5.Similarly, in drawing a card from a pack, the cases favorable to getting a spade are 13 (as there are 13 spade card in the pack).

Mathematical or Classical or a "Priori" definition of Probability: If a random experiment results in N-exhaustive mutually exclusive and equally likely outcome (cases) out of which 'm' are favorable to the happening of an event 'A'. Then the probability of occurrence of 'A' usually denoted by P (A) is given by

$$P(A) = \frac{\text{Favorable number of cases to A}}{\text{Exhaustive number of cases}}$$

$$\text{Or } P(A) = m/N$$

Statistical / Empirical Probability:

$$P(A) = \lim_{N \to \infty} m / N$$

It means probability of happening of the limit.

Remember:

1. In statistics, probability of happening of 'A' is written as P (A) or simply (A).
2. Probability of not happening of the limit 'A' is written as P ($\bar{A}$).
3. Suppose, there are two events A& B. A+B means either A happens or B happens or both happen. Both happens means both simultaneously happen.
4. Probability always lies between 0-1. Never say probability is negative or positive.

Question: Suppose A & B are mutually exclusive events. Then A & B means what?

Solution: It means either A happens or B happens or it means if A happens then B is not happens. Here both happen are zero if they are independent.

Permutations and Combination: The words permutation refers to the "arrangement" and combination refers to "group". These terms are used in the calculation of probability. Some simple rules of Permutation and Combination are given below:

1. The permutation of 'n' dissimilar things taken all at a time is n! Thus, if there are three letters A, B, & C, the total number of ways in which they can be arranged is 3! = 3x2x1 = 6 i.e. ABC, ACB, BCA, CAB, BAC & CBA.
2. The permutation of 'n' dissimilar things taken 'r' at a time is ${}^{n}P_{r}$ or n! / n-r! Thus, if we have to arrange any two letters out of three (A, B, & C), we can arrange them as ${}^{3}P_{2}$ or 3! / (3-2)! Or 3x2x1 / 1 = 6ways i.e. AB, BA, BC, CB, CA, AC.
3. The number of permutations of n things when n_1 of them are of one kind and n_2 of another kind is given by n! /n_1! n_2!. Thus, if we have to find out the permutation of the letter of the word 'FARIDABAD' (where 'A' occurs three times and 'D' occurs two times), the answer would be 9! / 3! 2! = 9x8x7x6x5x4x3x2x1 / 3x2x1x2x1.
4. The number of combination of 'n' different things taken 'r' at a time is given by ${}^{n}C_{r}$ or n! /r! (n-r)! Thus, if we have to pick up two alphabets out of the three A, B, & C; we can pick up as ${}^{3}C_{2}$ = 3! / 2! (3-2)! = 3x2x1 /2x1x1 = 3 i.e. AB, AC, or BC. We had seen that the number of permutations in this case was 6 because each combination can be arranged in two ways as AB, BA, AC, CA, BC, and CB. Thus, the number of permutation is equal to the number of combinations multiplied by 'r'. In other words, ${}^{n}P_{r} = {}^{n}C_{r}$ x r or ${}^{n}C_{r} = {}^{n}P_{r}$ / r

Remember: 5! = 5x4x3x2x1 = 120; 0! = 1; 1! =1

Question: There are three students. I want a combination of either A student or B student or C student. Then what is the probability of A.

Solution: Since, Probability = Favorable Cases / No. of Cases = 4 /7.

A +B +C

A, B, C	AB, BC, AC	ABC
${}^{3}C_{1}$	${}^{3}C_{2}$	${}^{3}C_{3}$

Question: What is the probability of a random event?

Solution: ½

Question: What is the probability of getting 53 Sunday in a leap year.

Solution: Leap year is always after 4 year in which 29 days are there in February month.

Leap Year = 366 days.

Probability = 366 /7 = 52. (2 /7) = 2/7

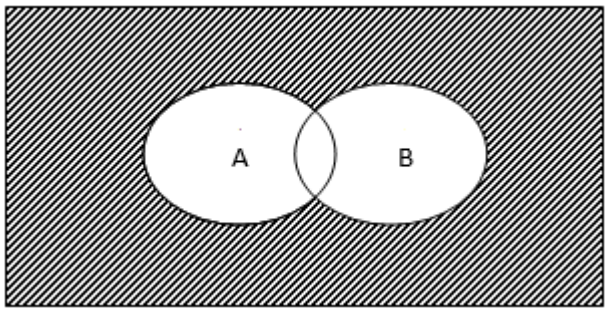
Theorem of Probability: There are three theorems of probability. These are:

- Theorem of addition
- Theorem of multiplication
- Baye's theorem

Theorem of addition: The probability of happening of at least one of the several mutually exclusive events is the sum of the probability of each event. If E_1, E_2, E_3, E_4 are four mutually exclusive events then,

$P(E_1+E_2+E_3+E_4) = P(E_1) + P(E_2) + P(E_3) + P(E_4)$

Caution: The theorem of addition will not be applicable when the events are not mutually exclusive (when two or more event can take place together).The following diagram explains the nature of two events which are not mutually exclusive or are overlapping.



The addition formula in such cases then takes the following shapes:

P (A or B) = P (A) + P (B) – P (AB)

If there are three such events then,

P (A or B or C) = P (A) + P (B) + P (C) – P (AB) – P (AC) – P (BC) + P (ABC).

Theorem of multiplication: If A, B are two events occur simultaneously then probability of simultaneously occurrence of two events A & B means

P (AB) = P (A). P (B /A)

(B /A) means event B happens when A has already occur.

Conditional occurrence of B = B / A when A has already occurred. This is called theorem of multiplication.

This can also be written as P (AB) = P (B). P (A / B). It means conditional occurrence of A if B has already occurred.

Note: AB means simultaneously occurrence of both the event. ABC means simultaneously occurrence of all the three events.

Theoretical distribution: it is also known as distribution of count because of specific countable number.

Difference between theoretical distributions and frequency distribution: A distribution which is based on observation and experimentation is called frequency distribution. The variable in frequency distribution does not follow the exact law. E.g. if we throw a coin, the distribution of

head or a distribution of tail is a frequency distribution. While, theoretical distribution is based on probability distribution. In theoretical distribution, variables follow an exact mathematical law.

- **Distribution:** It shows the trends on which the variables taking different values.
- **Variable:** A thing which takes any value.
- **Variate:** When a variable takes a certain value at certain probability, it is known as variate. Other name given to variate is random variable, stochastic variable, chance variable.
- **Probability distribution:** How the probability values are falling in a particular pattern.
- **Probability model:** Specific value of probability distribution is called probability model. E.g. model may be a set which is taken from a particular trial. In model, we chose a particular distribution.

Types of theoretical distribution: Two

i. Theoretical discrete distribution,
ii. Theoretical continuous distribution.

Theoretical discrete distribution: These include

i) Bernoulli's distribution, ii) Binomial distribution,

iii) Poisson distribution, iv) Negative binomial distribution

Theoretical continuous distribution: Normal distribution

Remark: For every distribution, we can calculate mean, variance, skewness and kurtosis because these are helpful in making comparison.

Why we study distribution? We study distributions for comparison and interpretation of data.

Continuous random variable: A random variable X is said to be continuous if it can take all the possible values between certain limit.

Discrete random variable: A random variable X is said to be discrete if it can take a specific value.

Probability density function: Whenever the variate is continuous, the probability function is known as probability density function (p.d.f). It will generate probability between two limits.

Probability mass function: Whenever the variate is discrete, the probability function is known as probability mass function (p.m.f). It will generate probability at a particular point.

Bernoulli's Distribution: A random variable X is said to be a Bernoulli's variate if it takes the value of head as X=1 (successful event) with probability 'p' and takes the value of tail as X=0 (failure event) with probability 'q' then, this distribution is called Bernoulli's distribution. Bernoulli's distribution takes the value of 0 and 1. For example, germination of seed follows Bernoulli's distribution.

Conditions of Bernoulli's distribution: There are four conditions of Bernoulli's distribution. These are: 1. each trial has independent event. 2. Probability of successful event is 'p'. 3.

Probability of failure event is 'q'.4. An experiment is performed under the same conditions for a fixed number of trial.

Why should we study Theoretical distribution? We study theoretical distribution because these parameters are known to us and are helpful for us in comparison. We study Bernoulli's distribution for comparison and interpretation.

Remark: For every distribution, we can calculate mean, variance, skewness and kurtosis because these are helpful in comparison.

Binomial Distribution: It is again a discrete type of theoretical distribution and is more advanced than the Bernoulli's distribution. Here, 'Bi' means two and 'Nomial' means expression. The expansion of binomial expression is known as Binomial distribution.

If a coin is tossed once, there are two outcomes, namely head or tail. The probability of obtaining head i.e. p =1/2 and the probability of obtaining tail i.e. q = 1/2. Thus, (q+p) = 1. These are the terms of the binomial (q+p). The general form of binomial distribution thus is given by

$$P(r) = {}^{n}C_{r}\, q^{n-r}\, p^{r} \quad \text{--------(1)}$$

Where, p = probability of success in a single trial.

q= 1-p; n= number of trial; and r = number of success in 'n' trial.

Equation (1) is called the probability mass function of the binomial distribution. N and p are known as the parameters of the binomial distribution. 'n' is also known as degree of binomial distribution.

Properties of Binomial Distribution:

1. Shape and location of binomial distribution changes as 'p' changes for a given 'n'.
2. The mode of binomial distribution is equal to the value of x which has higher probability.
3. As 'n' increases for a fixed 'p', the binomial distribution moves to the right, flattens and spread out.
4. If 'n' is large and if neither 'p' nor 'q' is too close to zero, the binomial distribution can be closely approximated by a normal distribution with standard variate given by

$$Z = X\text{-}np / \sqrt{npq}$$

The approximation becomes better with increasing 'n'.

1. Mean of the binomial distribution is equal to 'np'.
2. Variance of the binomial distribution is equal to 'npq'.
3. Mean is always greater than variance in binomial distribution.
4. Coefficient of skewness of binomial distribution is given by 'β_1' and is equal to $(q\text{-}p)^2 / npq$.
5. Coefficient of kurtosis of binomial distribution is given by 'β_2' and is equal to $3 + (1\text{-}6pq) /npq$.

Importance of Binomial Distribution:

1. The outcome of each trial in the process is characterized as one of the two types of possible outcomes. In other words, they are attributes.
2. The possibility of outcome of any trial does not change and is independent of the result of previous trial.
3. Sum of the independent binomial variate is not a binomial variate. In other words, binomial distribution does not possess the additive or reproductive properties.

Poisson distribution: The Poisson distribution is defined as

$$P(r) = e^{-m} m^r / r! \text{ ---------- (1)}$$

Where, r =0, 1, 2, 3…….

$$E = 2.7183 \text{ (base of natural log)}$$

m = mean of the Poisson distribution i.e. 'np' or the average number of occurrences of an event. 'm' is also known as the parameter of the distribution and is always greater than zero.

This is the probability mass function or the probability model of Poisson distribution because it will generate a population of probability for different values of 'r'. It is a discrete theoretical distribution with a single parameter 'm'. As 'm' increases, the distribution shifts to the right. All the Poisson distribution is skewed to the right. This is the reason why Poisson distribution has been called the "probability distribution of rare events". The probability tends to be high for small number of occurrence.

Postulates of Poisson distribution: Successful event occur in time obeying the following the following points.

1. The number of time the success occurs in any interval of time is independent of the number of the occurrence of the successes in any other disjoint time interval.
2. The chances of two or more occurrences happening simultaneously can be assumed to be zero.
3. The average number of occurrence per unit of time is constant and it does not change with time.
4. When average number of successes are known per unit of space than we use Poisson distribution.

Poisson distribution as a limiting case of Binomial distribution:

Poisson distribution is a limiting case of Binomial distribution under the following conditions:

1. Number of trial i.e. 'n' is independently large i.e. $n \to \infty$
2. 'p' the constant probability of success for each trial is independently small i.e. $p \to 0$
3. 'np' =m (say) is finite. Thus p = m /n and q = 1 –m/n. where 'm' is a positive real number.

4. The Poisson distribution is a good approximation of the Binomial distribution when $n \geq 20$ and $p \leq 0.5$

If the above conditions hold well, we can substitute the mean of the Binomial distribution (np) in place of the mean of the Poisson distribution (m) so that the formula becomes

$P(r) = e^{-np} (np)^r / r!$

Other Characteristics of Poisson distribution:

1. Mean and variance of the Poisson distribution are same i.e. 'm'.
2. Coefficient of skewness of Poisson distribution $\beta_1 = 1 / \sqrt{m}$.
3. Coefficient of Kutosis of Poisson distribution $\beta_2 = 1 / m$.
4. Poisson distribution is always a skewed distribution and its standard deviation is $\sqrt{m}$ and standard Poisson variate = $x - m / \sqrt{m}$.
5. Sum of the independent Poisson variate is also a Poisson variate.

Negative Binomial Distribution: The equality of the mean and variance is an important characteristic of the Poisson distribution, whereas for Binomial distribution mean is always greater than variance. But occasionally, observable phenomenon gives rise to empirical discrete distributions which shows a variance larger than mean. Some of the commonest examples of such behavior are the frequency distribution of plant density obtained by quadrant sampling when the clustering of plants makes the simple Poisson model inapplicable. In such a case, negative Binomial distribution provides an excellent model because this distribution has a variance larger than mean. For example, death of insect, number of insects bite lead to a negative binomial distribution and the distribution also arises in inverse sampling from a binomial population or as a weighted average of Poisson distribution. Poisson distribution as a limiting case of Negative Binomial distribution.

Normal Distribution: A random variable X is said to have a normal distribution with parameter u (called mean) and σ^2(called variance) if its density function is given by the probability law:

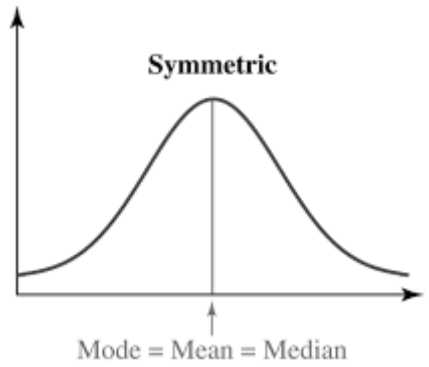
$$f(x; u, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp.\{-1/2 (x - u / \sigma)^2\}$$

$$\text{or} \quad f(x; u, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-1/2 (x-u)^2 / 2\sigma}, \quad -\infty < x < \infty \ \& \ -\infty < u < \infty, \ \sigma > 0$$

Chief characteristics of the normal distribution and the normal probability curve:

1. The curve is bell shaped and symmetrical about the line x =u
2. Mean, median and mode of the distribution coincide

3. As x increases numerically, f(x) decreases rapidly, the maximum probability occurring at a point x =u and is given by
4. $[p(x)]_{max} = 1/\sigma\sqrt{2\pi}$
5. Coefficient of skewness(β_1) is equal to zero and coefficient of kurtosis (β_2) is equal to 3
6. Since f(x) being the probability, can never be negative. No portion of the curve lies below x-axis.
7. X-axis is asymptotic to the curve
8. Linear combination of independent normal variate is also normal.



Normal Probability Curve

8. The point of inflection of the curve are given by

   [x = u + r, f(x) = $\{1 / \sigma\sqrt{2\pi}\}\ e^{-1/2}$]

9. Mean deviation about mean is given by

$\sqrt{2/\pi}.\ \sigma = 4/5.\ \sigma$ (approx.)

10. Q.D. = $(Q_3 - Q_1)/2 \approx (1/2)\ \sigma$

11. In normal distribution, the sample size is large.

12. Normal distribution is a particular case or a limiting case of Binomial distribution when $n \to \infty$ and neither p nor q is very small.

Area property:

P ($u - \sigma < X < u + \sigma$) = 0.6826

P ($u - 2\sigma < X < u + 2\sigma$) = 0.9544

P ($u - 3\sigma < X < u + 3\sigma$) = 0.9973

The table gives the area under the normal probability curve for some important values of standard normal variate Z.

Distance from the mean ordinates in terms of ± σ	Area under the curve
Z = ± 0.745	50% = 0.50
Z = ± 0.196	95% =0.95
Z =±0.2.58	99% = 0.99

Note: In normal distribution, we have probability density function (pdf). Probability density function will generate probability value between two limits. Probability mass function (pmf) will generate probability at a particular point. Any distribution when its size is very large then, it becomes normal distribution. Normal distribution is also called the normal probability distribution or symmetrical distribution or distribution of error happens to be the most useful theoretical distribution for continuous variable. In normal distribution, sample size is large. The whole data are uniformly distributed for normal distribution. Sometimes, the data are not uniformly distributed for normal distribution then; we transform the data into normal form. There are different transformations by which we transform the data into normal form. These transformations are:

Under-root transformation,

Arc-sine transformation,

$\sqrt{(x +1 /4)}$ transformation.

We use transformation just to convert the data into normal form. With the help of scattered diagram, the shape of the curve go on changes as the value of σ either increase or decrease (σ= error). We choose smaller value of σ for normal distribution.

Chapter 3
INFERENTIAL STATISTICS

Statistical Inference:

It is a branch of statistics which is concerned with using probability concepts to deal with uncertainty in the decision making.

Forms of statistical inference:

1. Hypothesis testing i.e. to test some hypothesis about parent population from which the sample is drawn;
2. Estimation i.e. to use the 'statistics' obtained from the sample as estimate of the unknown 'parameter' of the population from which the sample is drawn.

In both these problems, inferences can be drawn from the sample data.

Procedure of testing hypothesis:

1. Set up a hypothesis
2. Set up a suitable level of significance
3. Setting a test criterion
4. Doing computation
5. Making decision.

Hypothesis: Any tentative statement

Statistical hypothesis: Any statement regarding population parameter is known as statistical hypothesis.

Simple hypothesis: If the hypothesis specifies all the population parameter completely, it is known as simple hypothesis. E.g. In normal population, $\mu = \mu_0$ and $\sigma^2 = \sigma_0^2$.

Composite hypothesis: If in a hypothesis some of the parameters are unspecified then, it is known as composite hypothesis. E.g. Mean of the normal distribution is $\mu = \mu_0$ is a composite hypothesis because it does not tell anything about. σ^2.

Null hypothesis: Any statement which is to be nullified where no personnel biasness is there is called null hypothesis. E .g. To write H_0: u=50 means that we are to test the null hypothesis that the mean of the distribution is 50.

Alternate hypothesis: Any statement opposite to null hypothesis is called alternate hypothesis. (H_1) .E.g. H_1: u≠50 means that we are to test the alternate hypothesis that mean of the distribution is not 50.

Caution: Always remember when H_0 is rejected then, alternate hypothesis H_1 is accepted. Always remember when we accept H_1, then, we say that there is no evidence against H_0. Never use the word accepts or reject.

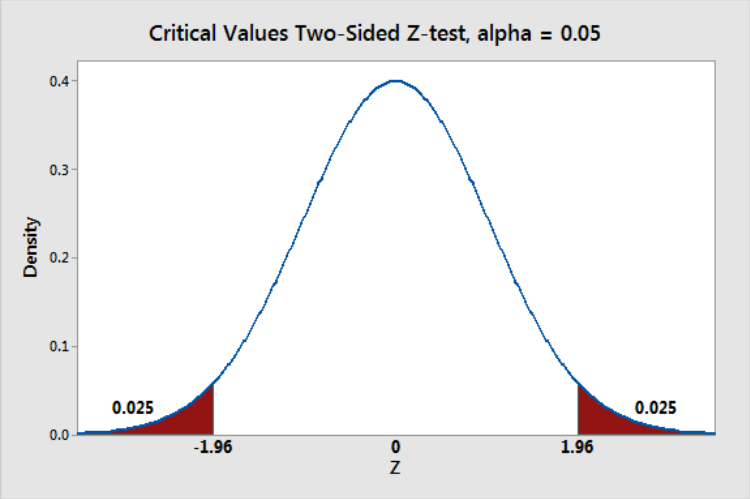
Test: A test of a statistical hypothesis is a procedure of deciding whether to reject or accept H_0.

Coefficient of confidence: It mean with what degree of confidence, the estimate is correct or wrong. If the estimate is 90 % correct then, coefficient of confidence is 0.9 or 90 % and is denoted by (1- α).

Level of significance (LOS): When the critical region is expressed in terms of probability, it is known as level of significance and is denoted by 'α'.

Critical region: Region of rejection is known as critical region. It is that region if our test statistics falls in it we reject our null hypothesis. It is expressed in terms of area or probability.

Test Statistics: It is any statistics which is used to test a thing. Test statistics are generally based on some probability distribution. Some of the common probability distributions which are used in hypothesis testing are Z, t, F and $\chi 2$ etc. Choice of test statistics would depends upon the nature of the distribution and the size of the sample.

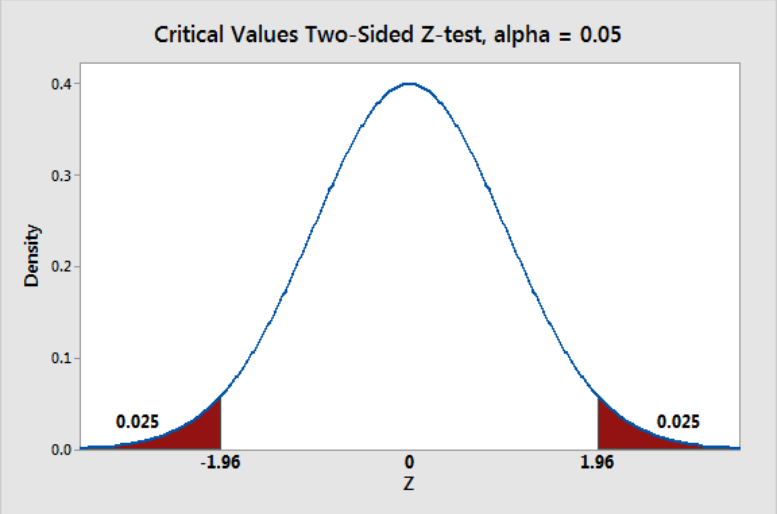


Note: Always remember probability lies between 0 and 1. Area under the normal curve is represented by probability and is equal to 1.

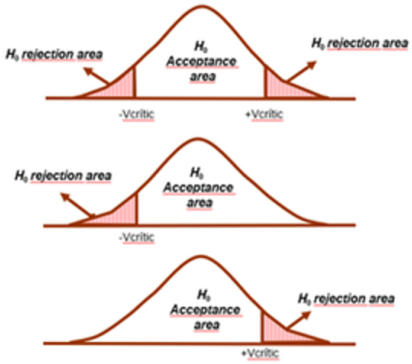
Confidence level and significance level: The confidence level or reliability is the expected percentage of times that the actual value fall within the stated precision limits. Thus, if we take a confidence level of 95 %, then, we mean that there are 95 chances in 100 (0.95 in 1) that the sample results represent the true condition of the population within a specified precision range against 5 chances in 100 (0.05 in 1) that it does not. Precision is the range within which the answer may vary and still be acceptable; confidence level indicates the likelihood that the answer will fall within the range, and the significance level indicates the likelihood that the answer will fall outside the range. One should always remember that if the confidence level is 95% then, the significance level will be (100-95) i.e. 5 %; and if the confidence level is 99%, then the significance level is (100-99) i.e. 1% and so on. We should also remember that the area of the normal curve within precision limits for the specified confidence level constitutes the acceptance region and the area of the curve outside these limits in either direction constitutes the rejection region.

One-tailed and two-tailed test: If the critical region (region of rejection) lies on both side of the tail then it is known as two-tailed test and if the critical region (CR) lies only one side of the tail then, it is known as one-tail test. E.g. Whenever H_1: $\bar{x} \neq u$, It is always advised to use two tail test because it is either $\bar{x} > u$ or $\bar{x} < u$. .

If H_1: $\bar{x} > u$, it means that we are using only one side of this tail i.e. RHS. And if H_1: $\bar{x} < u$, it means we are using the LHS of the tail or one tail test.



"Two –tail test Diagram"



H0: sample mean = population mean
Ha: sample mean <> population mean

H0: sample mean >= population mean
Ha: sample mean < population mean

H0: sample mean <= population mean
Ha: sample mean > population mean

Type-1 error: When we are rejecting something whereas it is correct, it is known as Type -1 error. Type-1 error of rejecting null hypothesis when it is true is dented by α.

α =Probability of Type-1 error = Probability of rejecting H_0 when H_0 is true.

Type-2 error: Type-2 error of accepting null hypothesis when it is not true is denoted by 'β'.

β = Probability of type-2 error = Probability of accepting H_0 when H_0 is not true.

To conclude from the above, the following points should also be remembered:

1. A thing which is rejected at 1% LOS will always be rejected at 5% LOS and a thing which is rejected at 5% LOS may not be rejected at 1% LOS.
2. The probability of committing Type-1 error is known as level of significance (power of a test).
3. 5% LOS means that result may come to be true up to 95% and there are chances that 5% may go wrong.

4. Type-1 error is more dangerous in terms of money. Whereas Type-2 error is more dangerous in because of medicine. Both these error cannot be minimize simultaneously because they are interdependent. E.g. when Type-1 error is 90 % then, Type-2 error is 10%. It is not possible to reduce both types of error simultaneously. If we reduce the probability of Type-1 error from 0.05 to .01, we simultaneously increase the probability of Type-2 error. The value of α can be reduced only by increasing the value of β. Type-1 error is more serious for economic point of view. Fix Type-1 error, and then minimize Type-2 error.
5. Type-1 error is committed when we accept a wrong or incorrect hypothesis. i.e. Type-1 error is committed when we reject a correct or true hypothesis.
6. Region of rejection in terms of probability is known as level of significance (LOS). It gives us the probability of rejection of null hypothesis.
7. Type-1 and Type-2 error: To summarize, decision depending upon sample can be of two type and is given in the following table:

	Decision	
	Rejection H_0	**Accept H_0**
H_0 : True	Type-1 error	Correct decision
		(No error)
H_0 : False	Correct decision	Type-2 error
	(No error)	

Level of Significance: After setting up the hypothesis, the next step is to determine the level of significance at which the hypothesis would be tested. It means that we have to determine the level of confidence with which a particular hypothesis is accepted or rejected. The level of confidence will determine the probability of our being right (or wrong) in accepting or rejecting a hypothesis. Normally, the level of confidence set for most hypothesis testing is 5%. This mean that our decision of accepting or rejecting a hypothesis would be correct 95 times out of 100 and the chances of our going wrong are only 5 %. This further means that there is a 5 per cent chance of our rejecting a null hypothesis which is true or 5 per cent chances of our accepting an alternate hypothesis which is wrong. If we want a greater precision, the hypothesis can be tested at 1 % LOS in which case the chances of our going wrong are only 1%.

Procedure of hypothesis testing:

1. The first step is to set up a hypothesis. The hypothesis is a set of inferences that is drawn concerning the parameter of population.
2. After setting up the hypothesis, the next step is to determine the level of significance at which the hypothesis would be tested.
3. The third step in hypothesis testing is to decide the test statistics.
4. The last stage in hypothesis testing is to draw conclusions about accepting or rejecting a hypothesis.

Degree of freedom: If 'n' is the independent observations and 'k' the independent constraints then (n-k) is called the degree of freedom. E.g. 2,3,4,5 are the four independent observations. The total is 14. But, the constraint is there. Therefore, df =4-1=3.

In other words, the number of unspecified parameters is known as degree of freedom. That is those parameters which are not stated in detail is called the degree of freedom.

Parameter: These are the constant of population which specify completely.

Statistics: It is a function of all the sample observation. Suppose, we have got only three observations, x_1, x_2, x_3 then

$Y= x_1- x_2+ x_3$

$Y_1= x_1+ x_2- x_3$

These functions are nothing but sample statistics. Among these functions, the function which gives very close value of population parameters, that statistics is selected.

Functions of sample observation or estimating function: The best estimate would be one which fall nearest to the true value of the parameter to be estimated i.e. statistics whose distribution concentrate as closely as possible near the true value of the parameter is regarded as the best estimate.

Characteristics of a good estimator: i) unbiased; ii) consistent; iii) efficiency; iv) sufficiency; v) minimum variance; vi) invariance; vii) completeness. If all these qualities are there in the estimator, then we say it is an ideal estimator.

Amount of information (AI): AI = 1/ Variance

Efficiency of A over B: Amount of information of A / Amount of information B.

= Var.B/ Var. A. Efficiency are also measured in terms of variance.

Test of a statistical hypothesis: A test of a statistical hypothesis is two action decision problems after the experimental sample value have been obtained. The two actions being the acceptance or rejection of the hypothesis under consideration.

Test of a Null hypothesis: For example,

Reject H_0, when $X \geq 15$

Accept H_0, when $X<15$

This technique is called test of a null hypothesis. X is called test statistics. Let X=60 i.e. it is >15, it means there is no evidence against H_1 (reject H_0).

Test of significance: We can find out the significance level of a variable by means of following statistics given below:

Z-statistics

t-statistics

F-statistics

X^2 –statistics

When performing any type of inferential statistics and any type of statistical testing, a value is generated based on the data (either t, F, Z, or X^2), and this value is being compared to some corresponding critical value (these critical values can be found by looking at the table in the back of any statistics textbook.) and this value is being compared to some corresponding critical value (t, F, Z, or X^2) in order to determine the statistical significance.

Z-statistics: Z is nothing but standard normal variate. Whenever large sample is there then we use Z-statistics also called large sample test.

$$Z = \frac{|\bar{x} - \mu|}{\sigma/\sqrt{n}}$$

Where, $\bar{x}$ = sample mean; = population mean

σ = population standard deviation; n = sample size

Whether we use one-tail test or two-tail test, we must fix up H_0 & H_1 by looking into the problem before solution. If the calculated value of Z ≤ table value of Z at a desired LOS accept H_0. On the other hand if the calculated value of Z is > the table value of Z, reject H_0.

Assumption of Z-statistics:

1. Parent population is normal
2. Sample is a simple random sample.
3. Sample size is large (>30).
4. Standard deviation of population is known.

Z –test for two samples:

For a single sample mean, the standardized normal variate (Z) is given by:

$$Z = \frac{|\bar{x} - \mu|}{\sigma/\sqrt{n}}$$

where

$\bar{x}$= sample mean, μ= population mean, σ= population standard deviation, n= sample size.

Thus,

$$Z = \frac{|\bar{x} - \mu|}{\text{S.E. of } \bar{x}}$$

Z-test for difference between two sample means

For two independent samples, the Z-value is:

$$Z = \frac{|\bar{x}_1 - \bar{x}_2|}{\text{S.E. of } (\bar{x}_1 - \bar{x}_2)}$$

The standard error of the difference between two means is:

$$Z = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Case 1: Equal population variances ($\sigma_1^2 = \sigma_2^2 = \sigma^2$)

$$Z = \frac{|\ \bar{x}_1 - \bar{x}_2\ |}{\sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}}}$$

This simplifies to:

$$Z = \frac{|\ \bar{x}_1 - \bar{x}_2\ |}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Case 2: Equal sample sizes ($n_1 = n_2 = n$)

$$Z = \frac{|\ \bar{x}_1 - \bar{x}_2\ |}{\sigma\sqrt{\frac{1}{n} + \frac{1}{n}}}$$

Further simplifying:

$$Z = \frac{|\ \bar{x}_1 - \bar{x}_2\ |}{\sigma\sqrt{\frac{2}{n}}}$$

Final simplified form

$$Z = \frac{|\ \bar{x}_1 - \bar{x}_2\ |}{\sigma\sqrt{2/n}}$$

Inference: If $Z_{calculated} \leq Z_{table}$ at desired level of significance, accept H_0. And If $Z_{calculated} > Z_{table}$ at desired level of significance, reject H_0.√

Additional information about Z-test:

Sometimes sample size is large; population is normal; samples are simple random sample; but standard deviation of the sample is unknown. In that case

$$Z = \frac{|\ \bar{x} - \mu\ |}{s/\sqrt{n}}$$

Where, $s$is estimate of variance =

$$\frac{1}{n-1}\sum((x_1 - \bar{x}_2)^2)$$

For two sample,

$$Z = \frac{|\ \bar{x}_1 - \bar{x}_2\ |}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{|\ \bar{x}_1 - \bar{x}_2\ |}{\text{SE of } (\bar{x}_1 - \bar{x}_2)}$$

Where,

$$s_1^2 = \frac{1}{n_1 - 1}\sum((x_{1i} - \bar{x}_1)^2)$$

Inference: If $Z_{calculated} \leq Z_{table}$ at desired level of significance, accept H_0. And If $Z_{calculated} > Z_{table}$ at desired level of significance, reject H_0.

Student's t –statistics: Student was the pen name of the scientist. t-distribution is used when sample size is 30 or less and the population standard deviation is unknown. It is also called as small sample test.

Assumptions of t-distribution:

1. Population is a normal population.
2. Sample is a simple random sample.
3. Sample size is small (n≤30).
4. Population standard deviation is unknown.

$$t_{(n-1)\,df} = \frac{|\,\bar{x} - \mu\,|}{\text{SE of } (\bar{x} - \mu)} = \frac{|\,\bar{x} - \mu\,|}{\text{SE of } \bar{x}} = \frac{|\,\bar{x} - \mu\,|}{s/\sqrt{n}}$$

$\bar{x}$ = sample mean, u = population mean

s = standard deviation of population, n = sample size

Where,

$$s^2 = \frac{1}{n-1}\sum((x_i - \bar{x})^2)$$

t has got (n-1) degree of freedom.

Inference: If $t_{calculated} \leq t_{table}$ at desired level of significance and for a given number of degree of freedom, accept H_0. And If $t_{calculated} > t_{table}$ at desired level of significance and for a given number of degree of freedom, reject H_0.

Fisher's t-test or t-test for two samples:

Assumptions:

1. Population is normal.
2. Samples are simple random samples.
3. There is homogeneity of variances in two populations.
4. Sample sizes are small.
5. Standard deviations of the population are statistically same but unknown.

(Statistically same means 3=4 i.e. here we put probability. Numerically same means 3=3 i.e. they are exactly same.)

$$t_{(n1+n2-2)df} = = \frac{|\,\bar{x_1} - \bar{x_2}\,|}{\text{SE of } (\bar{x_1} - \bar{x_2})} \qquad \frac{|\,\bar{x_1} - \bar{x_2}\,|}{s\sqrt{1/n_1 + 1/n_2}}$$

$$\text{Where,} \quad s^2 = \frac{(n_1\text{-}1)s_1^2 + (n_2\text{-}1)s_2^2}{n_1 + n_2 - 2}$$

$$s_1^2 = \frac{1}{n_1-1}\sum((\bar{x_i}-\bar{x_1})^2 = \{\sum x_i^2 - (\sum x_1)^2/n\}$$

$$s_2^2 = \frac{1}{n_2-1}\sum((\bar{x_i}-\bar{x_2})^2 = \{\sum x_i^2 - (\sum x_2)^2/n\}$$

's^2' is known as pooled variance. Pooling is feasible only if the variances are homogeneous. Otherwise not. This test is used only when the variances are homogeneous. This't' is called Fisher's t. Fisher is known as the father of statistics.

Inference: If $t_{calculated} > t_{table}$ at desired level of significance, the difference between the sample mean is said to be significant at 5% LOS. Otherwise, data are said to be consistent with the hypothesis.

Matched Paired t-test: When two samples are given indication of correlation and are not independent, then we use paired t-test. In other words, there is 1-1 correspondence between two values so as to minimize any differential effect due to different factors.

$$t_{(n-1)df} = \frac{|\bar{d}|}{s/\sqrt{n}} \quad \text{or} \quad s = \frac{1}{n-1}\{\sum di^2 - (\sum di)^2/n\}$$

Where, d = difference between the paired value

$|\bar{d}|$ = mean of the difference

n = number of pairs

s = standard deviation of difference = $\frac{1}{n-1}\sum(di-\bar{d})^2$

This paired "t" is based on (n-1) degree of freedom.

Inference: If $t_{calculated} \leq t_{table}$ for a desired level of significance at (n-1) df then, accept H_0 . And If $t_{calculated} > t_{table}$ for a desired level of significance at (n-1) df then reject H_0.

Snedecor's F-test:

$F = e^{2Z}$ where, Z is Fisher's Z.

This statistics is used to test the homogeneity of two variances.

$F_{(n1-1)(n2-1)df} = s_1^2/s_2^2$ provided $s_1^2 > / s_2^2$

$F_{(n1-1)(n2-1)df} = s_2^2/s_1^2$ provided $s_2^2 > / s12$

F' will be having two degree of freedom (n_1-1) and (n_2-1)

For (n_2-1)df, we see horizontal degree of freedom and for (n_2-1) df, we see vertical degree of freedom.

Inference: If $F_{calculated} \leq F_{table}$, accept H_0 and if $F_{calculated} > F_{table}$ value, then reject H_0.

Assumptions: H_0: $\sigma_1^2 = \sigma_2^2$ i.e. variances are same.

H_1: $\sigma_1^2 \neq \sigma_2^2$ i.e. variances are not same. If F calculated < F table, accept H_0 i.e. variances are same.

Two samples have been drawn from two different population having same variance and same mean. Also, two samples have been drawn from two different population having same variance but different means.

If there are more than two samples, then we use ANOVA table for testing the homogeneity of variances. This is called Bartlett test for testing the homogeneity of variances for more than two samples.

In t-test, we test the means. In F-test we test the variances whether they are significant or not. If the variances are homogeneous, they can be pooled, otherwise not.

Fisher's and Behreem's'd' test: For testing the difference in means when the variances are heterogeneous or when the variances are not same, then we use this test. In other words, if in the above F-test, H_1 is accepted, then we use this test.

$$\tan\theta = \frac{s_1/\sqrt{n_1}}{s_2/\sqrt{n_2}} = \frac{\text{SE of } \bar{x_1}}{\text{SE of } \bar{x_2}}$$

$$\theta = \tan^{-1}\left\{\frac{s_1/\sqrt{n}}{s_2/\sqrt{n_2}}\right\}$$

Here, we have three parameters viz. (n_1-1), (n_2-1) and θ. We use Fisher's and Yates's table.

1. Calculate θ (see natural tan table).

2. $$d = \frac{|\bar{x_1}-\bar{x_2}|}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

(Table value of d is seen from table no. 6 keeping in view (n_1-1)df & (n_2-1)df.)

Inference: If $d_{calculated} \leq d_{table}$, accept H_0 and if $d_{calculated}$ is> d_{table}, reject H_0.

Approximate method:

Cochran and Cox't' test: This is an approximate method to Fisher's and Behreem's d-test.

$$d = \frac{|\bar{x_1}-\bar{x_2}|}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

Where, $s_1^2 = \frac{1}{n_1-1}\sum((\bar{x_i}-\bar{x_2})^2$ And $s_2^2 = \frac{1}{n_2-1}\sum((\bar{x_i}-\bar{x_2})^2$

$$t' = \frac{t_1 s_1^2/n_1 + t_2 s_2^2/n_2}{s_1^2/n_1 + s_2^2/n_2}$$

Where, t_1 = t value at 5 % LOS for (n_1-1)df

t_2 = t value at 5% LOS for (n_2-1) df

The value of t′ at 5 % LOS is given approximately by the weighted mean of the two value of t i.e. t_1 &t_2 and the weights are the variances of means. Weighted mean is denoted by t′.

Inference: if t calculated ≤ accept H_0 at 5% LOS and if t calculated > t′, reject H_0 at 5% LOS.

Assumptions of Cochran and Cox t test and Fisher's and Behreem's d test:

1. Population is normal.
2. Samples are simple random sample but they are independent.
3. Samples sizes are small (≤30).
4. Variances of the population are heterogeneous (different) and unknown.

Note: 1. Z-test is used to test the significance of the correlation coefficient.

2. F-test also called variance ratio test as it is based on the ratio of the two variances.

$$\frac{\text{Large estimate of variance}}{\text{Small estimate of variance}}$$

$v_1 = n_1-1$ and $v_2 = n_2-1$

v_1 = degree of freedom for samples having large variance.

v_2 = degree of freedom for samples having smaller variance.

Assumptions of F –test:

1. Normality, i.e. the values in each group is normally distributed.
2. Homogeneity, i.e. variance within each group should be equal for all groups ($\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_n^2$).
3. Independence of error. It states that the error (variation of each value around its own group mean) should be independent for each other.

Remember: if the variances are homogeneous, use Fisher's' test for testing the means. And if the variances are not homogeneous, use Cochran and Cox't' test for testing the means.

Analysis of Variance: The analysis of variance frequently referred to by the contraction 'ANOVA' is a statistical technique specially designed to test whether the means of more than two quantitative populations are equal. This technique was developed by R.A. Fisher in 1920s.

Assumptions in the Analysis of Variance: The assumptions in the analysis of variance are the same as discussed in F-test, i.e.,

1. Normality;
2. Homogeneity; and

3. Independence of error.

Technique of the Analysis of Variance: For the sake of clarity the technique of the analysis of variance has been discussed separately for (a) one-way classification and (b) two-way classification.

One-way Classification: In one way classification, the data are classified according to only one criterion. The null hypothesis is:

H_0: $u_1 = u_2 = u_3 = \ldots\ldots u_k$

H_1: $u_1 \neq u_2 \neq u_3 \neq \ldots\ldots \neq u_k$

All the means are not equal i.e. the arithmetic means of populations from which the k samples are randomly drawn are equal to one another. The steps in carrying out the analysis are:

1. Calculate the variance between the samples.
2. Calculate the variance within the samples.
3. Calculate the ratio F as follows

$$F = \frac{\text{Between column variance}}{\text{Within column variance}}$$

Compare the calculated value of F with the table value of F for the degree of freedom at a certain critical level. If $F_{calculated} > F_{table}$, it is concluded that the difference in sample means is significant i.e. it could not have arisen due to the fluctuation of simple sampling. On the other hand, if $F_{calculated}$ is $\leq F_{table}$ that means the difference is not significant and has arisen due to fluctuations of simple sampling.

It is customary to summarize calculations for sum of squares, together with their number of degree of freedom and mean squares in a table, called the analysis of variance table, generally abbreviated ANOVA. The specimen of ANOVA table is given below:

Analysis of Variance (ANOVA) Table: One Way Classification Model:

Source of variation	**SS(Sum of Square)**	**V (degree of freedom)**	**MS (Mean Square)**	**Variance ratio of F**
Between samples	SSC	$v_1 = C-1$	MSC=SSC /(c-1)	MSC /MSE
Within samples	SSE	$v_2 = n-c$	MSE =SSC /(n-c)	
Total	SST	n-1		

Where, SST = Total sum of squares of variations

SSC = Sum of squares between samples (column)

SSE = Sum of squares within samples (rows)

MSC = Mean sum of squares between samples

MSE = Mean sum of squares within samples.

Coding: Coding refers to addition, subtraction, division and multiplication of data by a constant.

Analysis of Variance in Two Way Classification Model: In a two way classification, the data are classified according to two different criterion or factors. The procedure for analysis of variance is somewhat different than the one followed while dealing with the problems of one-way classification. In two-way classification, the analysis of variance table takes the following forms:

Source of variation	**Sum of square**	**Degrees of freedom**	**Mean sum of squares**	**Ratio of F**
Between samples	SSC	(c-1)	MSC = SSC/(c-1)	MSC/MSE
Between rows	SSR	(r-1)	MSR = SSR/ (r-1)	MSR/MSE
Residual or error	SSE	(c-1)(r-1)	MSE = SSE/(c-1)(r-1)	
Total	SST	n-1		

Where, SSC = Sum of squares between column

SSR = Sum of squares between rows

SSE = Sum of squares due to error

SST = Total sum of squares

c = Number of column

r = Number of rows

Remember: Z, t and F statistics are known as **'parametric tests'** while X^2 is a '**non-parametric test'**.

Chi-square test: The X^2 test is one of the simplest and most widely used non-parametric tests in statistical work. It was first used by Karl Pearson in the year 1990s. the quantity X^2 describes the magnitude of the discrepancy between theory and observation. It is defined as:

$\chi^2 = \sum (O-E)^2/E$

Where, O = Observed frequencies and E = Expected frequencies

Steps: To determine the value of chi square, the steps required are:

1. Calculate the expected frequencies. In general the expected frequencies of any cell can be calculated from the following equation:

$$E = \frac{RT \times CT}{N}$$

Where, E = Expected frequency

RT = the row total for the row containing the cell.

CT= the column total for the column containing the cell.

N = the total number of observations.

2. Take the difference between the observed and expected frequencies and obtain the squares of these differences, i.e., obtain the value of $(O-E)^2$.

3. Divide the values of $(O-E)^2$ obtained in step 2 by the respective expected frequencies and obtain the total $\sum (O-E)^2/E$. This gives the value of χ^2 which can range from 0 to infinity. If χ^2 is zero it means that the observed and the expected frequencies completely coincide. The greater the discrepancy between the observed and the expected frequencies, greater shall be the value of χ^2.

4. The calculated value of χ^2 is compared with the table value of $\chi^2$for given degree of freedom at a certain specified level of significance. If at the stated level (generally 5% level is selected), the calculated value of χ^2 > table value of χ^2, the difference between theory and observation is considered to be significant, i. e. , it could not have arisen due to fluctuations of simple sampling. If on the other hand, the calculated value of χ^2 is < table value of χ^2, then the difference between theory and observation is not considered as significant, i.e., it is regarded as due to fluctuations of simple sampling and hence ignored.

Constants of χ^2 distribution:

1. The mean of the χ^2 distribution is equal to the number of degrees of freedom.
2. The variance of χ^2 distribution is twice the degrees of freedom
3. Total area under the curve in chi-square distribution is 1. Square of the standard normal variate is a χ^2 variate with 1 degree of freedom.

Applications and uses of χ^2:

1. It is used as a test of significance whether a given sample has been taken from a population of specified variate.
2. It is used to test the homogeneity of sample variances. This test is called Bartlett test.
3. It is used as a test of goodness of fit, i.e., to test whether the sample has been taken from a specified population or not.
4. It is used to test the association of attributes.
5. It is extensively used in the analysis of genetics especially to test the genetic hypothesis and to test the linkage.

Contingency table: When the data are classified into ‘m’ rows representing ‘n’ classes according to one attribute and ‘n’ columns and ‘m’ classes according to the other attribute. There is an mxn classes. The resulting table is known as mxn contingency table. For example,

	Blind	Non-blind
Dumbness		
Non-dumbness		

This is a 2x2 contingency table because, one attribute is blind and the other is dumbness. In 2x2 contingency table, Yates's correction is always applied.

Yates's Correction: χ^2 (corrected) = $\sum (\left| \text{O-E} \right| - 0.5)^2 / \text{E}$

Bivariate distribution: Those distribution having two variances, we call it as bivariate distribution**.**

Sampling distribution: A sampling distribution is an array of sample studies relating to a universe. If, for example, we wish to know the average income of industrial workers, we can have a large number of samples of individual workers and study their income. These samples which have been taken from the universe of individual workers would be a "Sampling distribution".

Conclusion: The introduction of inferential statistics marks the transition from merely describing a dataset (descriptive statistics) to making educated, probabilistic guesses about a larger population based on a smaller sample. It allows researchers to draw conclusions that go beyond the immediate data, enabling generalization and prediction. Key inferences drawn from the introduction of inferential statistics include:

1. Inferences can be drawn about a large population based on a representative sample, which is more cost-effective and faster than studying an entire population.
2. It allows us to determine if patterns observed in a sample represent real population trends or are merely due to chance.
3. Inferential statistics (through confidence intervals and p-values) accounts for sampling errors and acknowledges that sample data will not perfectly represent the population.
4. Researchers can accept or reject hypotheses about population parameters (like mean or variance) using tests such as t-tests, z-tests, and ANOVA.
5. It enables the identification of relationships between variables (e.g., via regression analysis) to predict future outcome.

In essence, inferential statistics transforms raw sample data into actionable knowledge and evidence-based decision-making.

Chapter 4

SAMPLING AND SAMPLING FUNDAMENTALS

Introduction:

Studying sampling is necessary because it is cost-effective and time-efficient, allowing researchers to draw generalizable conclusions about a large population by studying a smaller, representative subset. It makes studies that would otherwise be impossible to conduct because of their sheer scale, and by selecting an appropriate sample, researchers can achieve more accurate and precise measurements than they could with a less careful approach to data collection.

Benefits of Studying Sampling:

a. **Time and cost savings:** It is often impractical and expensive to collect data from an entire population. Sampling provides a more manageable and affordable way to get the information needed for a study;
b. **Feasibility:** For very large populations, it can be physically impossible to collect data from every individual. Sampling makes research feasible by allowing studies that would otherwise be impossible;
c. **Accuracy and precision:** A well-selected sample allows for more accurate measurements and more precise findings. The larger the sample size (within reason), the more confident you can be that the results reflect the entire population;
d. **Generalizability:** If a sample is truly representative of the population, researchers can generalize their findings to the entire group. This allows for a broader understanding of the subject of the study and
e. **Efficiency:** A study on a sample can be completed more quickly than one on a whole population, allowing for faster results and decision-making.

Sampling:

The word sampling is because of sample. It is the process of obtaining information about an entire population by examining only a part of it. In simple terms, it is the process of selection of sample.

Universe/Population:

From statistical point of view, the term ‘universe’ refers to the total of the items or units in any field of inquiry whereas the term ‘population’ refers to the total of the items about which information is desired. In simple terms, population is nothing but the number of individuals.

Types of Universe or Population:

a. **Finite population**: if the population is countable, it is called finite population. E. g. human population
b. **Infinite population**: If the population cannot be counted or it is limitless population. Then, it is known as infinite population. E.g. no of stars in the galaxy.
c. **Existence population**: Existence universe refers to a population of concrete objects like the number of persons having a certain income or the number of books with a certain number of pages etc.
d. **Hypothetical universe**: Hypothetical universe is one which does not consist of concrete objects. E.g. if a disc is tossed, each draw is an individual unit and we can construct a universe by throwing the disc a large number of times and recording its result.

Sample: Small portion of population is called sample. Small portion is taken either from finite or infinite or existence or hypothetical population.

Sampling theory: Theory which deals with the relationship between sample and the population are called sampling theory.

Sampling error: Sampling errors are those errors which arise on account of sampling and they generally happen to be random variations (in case of random sampling) in the sample estimates around the true population values. The magnitude of the sampling error depends upon the nature of universe. The more homogeneous the universe, the smaller the sampling error. Sampling error is inversely related to size of the sample i.e. sampling error decreases as the sampling size increases and vice-versa. Sampling error is worked out by:

Sampling error= Critical value at certain level of significance X Standard error

In other words, the difference between the sample value and the population value is called sampling error.

Why do we go for sampling? Sometimes, it is very difficult to study the whole population because of shortage of time and funds then, we resort to sampling. This implies that we study the population through sampling or we want to draw the inference about the population through sampling. E.g. soil sampling.

How should to draw sample? Our sample should be a through representative of the population. Otherwise, purpose of drawing sample about the population fails.

Object of sampling: The most important aim of sampling study is to obtain maximum information about the population phenomenon under study with the least sacrifice of money, time and energy.

Sampling design: It refers to the techniques or the procedure; the researcher would adopt in selecting some sampling units from which inferences about the population is drawn. These are census survey and sample survey.

Statistics and parameter: Statistics is a characteristic of the sample whereas; parameter is a characteristic of the population. E.g. Population mean 'u' is a parameter whereas; sample means '$\bar{x}$ 'is a statistic. To obtain the estimate of a parameter from a statistic constitutes the prime objective of sampling analysis.

Sampling and non-sampling error: Sampling error is that error which arises on account of sampling and they generally happen to be random variation in the sample estimate around the true population value. The sampling error usually decreases with increase in sampling size (number of units selected in the sample) and in fact in many situations decrease is inversely proportional to the square root of the sample size. Sampling error is not existed in case of complete enumeration survey, since the whole population is surveyed. However, the error mainly arising at the stages of ascertainment and processing of data, which are termed as non-sampling error, are common in both complete enumeration and sample surveys. Non-sampling error is likely to me more in case of complete enumeration survey than in case of sample survey, since it is possible to reduce the non-sampling error to a great extent by using better organization and suitably trained personnel at the field and tabulation stage. The non-sampling error is likely to be increased with increase in sample size.

Sample space: Sample space is the totality of all the sample points. E.g. suppose the set {1, 2, 3, and 4} and take any two numbers of these together which is called the sample point and the total number is called the sample space.

Sampling fraction: The ratio of the sampling size to the population size is known as sampling fraction. If the sample of size 'n' is taken from a population of size 'N' then, the sampling fraction is denoted by n/N.

Sampling procedure: It is the method of selection of sample from the population. Sampling procedure is said to be random if it is governed by the law of probability, i.e. its unit in the population is assigned some pre-determined probability of selection. The sample which is not selected by the random process is known as non-random

Accuracy and Precision: Accuracy refers to the amount of deviation of the estimate from the true value, whereas precision refers to the size of this deviation by repeated application of sampling procedure. Precision is the range within which the population average (or other parameter) will lie in accordance with the reliability specified in the confidence level as a percentage of the estimate ± or as a numerical quantity. For instance, if the estimate is Rs. 4000

and the precision desired is ± 4 per cent, then the true value will be no less than Rs. 3840 and no more than Rs. 4160. This is the range (Rs. 3849 to Rs. 4160) within which the true answer should lie.

Relative accuracy: The relative accuracy of two samples which differ in respect of method of sampling or the size of sample or both may be defined as the reciprocal of the ratio of sampling variance of the estimates given by the two methods when the same number of units is taken.

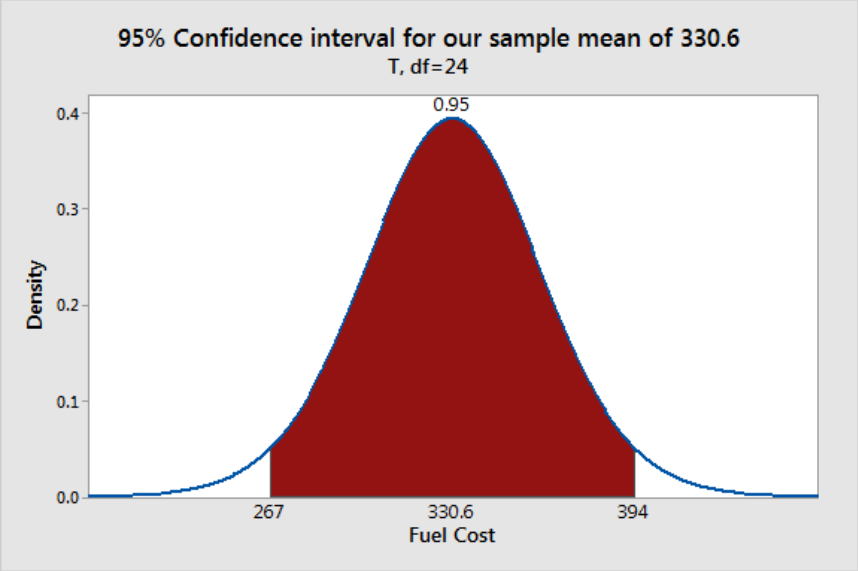
Efficiency and relative efficiency: The efficiency is measured by the inverse of the sampling variance of the estimator. The relative efficiency of the two different methods of sampling on the same type of sampling unit may be defined as the reciprocal of the ratio of the number of units required to attain a given accuracy with the two methods.

Estimator and estimate: The random variable such as sample mean ($\bar{\mathbf{x}}$) and sample variance (σ_s^2) used to estimate the population parameter such as population mean ($\boldsymbol{\mu}$) and population variance ($\boldsymbol{\sigma_p^2}$) are called estimator while, specific value of these say sample mean ($\bar{\mathbf{x}}$)=105, and sample variance (σ_s^2)=21.44 are referred to as estimate of the population parameter. Estimate always has a numerical value.

Estimation: It is the process of finding the value of population parameter.

Confidence level and Significance level: The confidence level or reliability is the expected percentage of times that the actual value fall within the stated precision limits. Thus, if we take a confidence level of 95%, then we mean that there are 95 chances in 100 (0.95 in 1) that the sample results represent the true condition of the population within specified precision range against 5 chances in 100 (or 0.05 in 1) that it does not. Precision is the range within which the answer may vary and still be acceptable; confidence level indicates the likelihood that the answer will fall within that range. One should always remember that if the confidence level is 95%, then the significance level will be 5% (100-95) and if the confidence level is 99%, then the significance level will be 1% (100-99) and so on. We should also remember that area under the normal curve within precision limits for the specified confidence level constitutes the acceptance region and the area under the normal curve outside these limits in either direction constitutes the rejection regions. The region of rejection is known as critical region and when the critical region is expressed in terms of probability, it is known as level of significance (LOS).

A normal curve diagram for a 95% confidence level shows the central area as the acceptance region and the two outer tails as the rejection regions. The acceptance region is shaded and represents 95% of the data, while the two unshaded tails, each representing 2.5% (LOS), are the rejection regions.



Sampling Distribution: We are often concerned with sampling distribution in sampling analysis. It is a probability distribution of a statistics that is obtained through repeated sampling of a specified population. It describes a range of possible outcomes from a statistic such as mean or mode of some variable, of a population. Distribution shows the trend on which the variable taking the different values. Some of the important sampling distributions, which are commonly used, are:

(a) Sampling distribution of mean; (b) Sampling distribution of proportion;

(c) Student's t-distribution; (d) F-distribution; and (e) Chi-square distribution

How we can increase the precision of the sample estimate? Precision can be increased (a) by increasing the sample size but, we can't increase the sample size beyond certain limit and (b) to reduce the heterogeneity in the population. More the precision of the sample estimate more will be the reliable estimation for the population.

Note: we estimate the value of population mean (**μ**) with the help of sample mean (**$\bar{x}$**). Closer the value of sample means (**$\bar{x}$**)) to population mean (μ), better is the estimate. The accuracy or precision of the sample depends upon-(a) sample size (n) & (b) heterogeneity in the population. More the value of sample size (n), closer the value of population parameters. But we can't increase the value of 'n' beyond certain limit. If the population is homogeneous, sample estimates give better parameter i.e. heterogeneity is removed. The conclusion in sampling studies is based not no certainty but on probability

Effective Sample Size: It is the actual distinct unit in that particular sample. E.g. In 111, ESS =1 and in 121, ESS =2.

Estimation of population parameter: There are two types of estimates pertaining to estimation of population parameter.

Point estimate and Interval estimate: The estimate of a population parameter may be one single value or it could be a range of values. In the former case, it is referred as point estimate, whereas in the later case it is termed as interval estimate.

a. Point estimate: A particular value of sample statistics which is used to estimate the parameter value is called the point estimate. E.g. average production of wheat in India is 40 q/ha.

b. Interval estimate: The limit within which the parameter value is estimated is called interval estimate. E.g. average production of wheat in India lies between 40-50 q/ha.

Small sample size and large sample size: When we have $n \leq 30$, it means it is a small sample and when $n>30$, it means we have large sample size

Sampling frame: It consists of a list of items from which the sample is to be chosen (source list).

Sampling Design

Census Survey: All the items under consideration in any field of inquiry constitute a "universe or population". A complete enumeration of all the items in the "population" is known as census survey. As all the items are included in this type of survey, therefore it must be having the highest accuracy. But, in practice this may not be true. Even the slightest element of bias in such an inquiry will get larger and larger as the number of observation increases. This method requires lot of time, money and energy. Therefore, when the field of inquiry is large, this method becomes difficult to adopt because of the resources involved. When the universe or the population is large, then this method is beyond the reach of an ordinary researcher. Government is the only institution which can get the complete enumeration carried out. Even the government adopts this method in very cases such as population census conducted once in a decade.

Sample Survey: When field studies are undertaken in practical life, consideration of time and cost almost invariably lead to a selection of respondents i.e. selection of only a few items. The respondents selected should be a representative of total population as possible in order to produce a miniature cross-section. The selected respondents constitute what is technically called s "Sample" and the selection process is called "Sampling Technique". The survey so conducted is known as "Sample Survey". For example, let the size of the population be "N". if a part of population say "n" units i.e $n<N$ is selected according to some rule then, n is called the sample.

Implications of a Sampling Design: It refers to the technique or the procedure the researcher would adopt in selecting items for the sample. Sample design indicates the number of items to be included in the sample i.e the size of the sample. Sample design is determined before the data are collected.

Steps in Sampling Design:

(a)Type of universe: First of all, the universe to be studied is defined. The universe can be finite or infinite. In finite universe, the number of items is certain, but in case of infinite universe, the number of items is infinite i.e. we cannot have the idea about the total number of items. The populations of a city, the number of workers in a factory are the example of finite universe. Whereas, the number of stars in the sky, listeners of a specific radio programme, throwing of a disc etc are the example of infinite universe.

(b) Sampling Unit: Sampling unit is to be decided before selecting sample. Sampling unit may be a geographical one such as state, district, village, etc. or it may be a social unit such as family, club, school, etc, or it may be an individual.

(c) Source list: It is also known as "Sampling Frame" from which sample is to be drawn. It contains the names of all the items of the universe (in case of finite universe only). If the source list is not available, the researcher has to prepare it. Such a list should be comprehensive, correct, reliable and appropriate.

(d) Size of Sample: This refers to the number of items to be selected from the universe to constitute a sample. The size of the sample should neither be excessively large, not too small. It should be optimum. An optimum sample fulfills the requirements of efficiency, representativeness, reliability and flexibility. Costs, time and budgetary constraints dictate the size of the sample.

(e) Parameters of interest: In determining the sample design, one must consider the question of the specific population parameters which are of interest.

(f) Budgetary Constraints: Cost considerations, from practical point of view, have major impact upon the decisions relating to not only the size of the sample but also to the type of sample.

(g) Sampling Procedure: Finally, the researcher must decide the type of sample he will use i.e. he must decide about the technique to be used in selecting the items for the sample. There are several sample designs out of which the researcher must choose the one for his study.

Criteria for Selecting a Sampling Procedure: In this context, one must remember that two costs are involved in the sampling analysis viz., the cost of collecting the data and the cost of incorrect inference resulting from the data. The researcher must keep in view the two causes of incorrect inferences viz., systematic bias and sampling error. A systematic bias result from errors in the sampling procedure and it cannot be reduced or eliminated by increasing the sample size. A systematic bias is due to the following one or more factors: - inappropriate sampling frame; -

defective measuring device; -non-respondents; -indeterminacy principle; and –natural bias in the reporting of data.

Sampling errors are the random variations in the sample estimates around the true population parameters. Since they occur randomly and are equally likely to be in either direction, their nature happens to be of compensatory type and the expected value of such errors happens to be equal to zero. Sampling error decreases with increase in sample size and it happens to be of smaller magnitude in case of homogeneous population. The measurement of sampling error is usually called the "precision of the sampling plan". If we increase the sample size, the precision can be improved. But, increasing the size of sample leads to increases the cost of collecting data and so enhances the systematic bias. Therefore, to increase precision, one must select a better sampling design which has a smaller sampling error for a given sample size at a given cost. In practice, however, people prefer a less precise design because it is easier to adopt the same and also because of the fact that systematic bias can be controlled in a better way in such a design.

Characteristics of a Good Sample Design:

(**a**) it should be representative one;

(**b**) it must be free from sampling error;

(**c**) it must be viable in the context of funds available for research study;

(**d**) it must be of the type that systematic bias can be controlled in a better way;

(**e**) it should be such that the results of the sample study can be applied, in general, for the universe with a reasonable level of confidence.

Types of sampling/Different methods of selection of sample: There are different techniques for the selection of sample so that our sample should be a thorough representative of the population. These techniques are:

1. Random sampling or Probability sampling;

2. Purposive or deliberate sampling;

3. Stratified random sampling**;**

4. Quota sampling;

5. Multistage sampling;

6. Convenience sampling;

7. Self selected sampling;

8. Cluster sampling;

9. Systematic sampling and

10. Sequential sampling

Now we shall discuss each of these sampling methods one by one.

1. Random sampling or Probability sampling: According to Pertan-"Random sampling is a form applied when the method of selection assured each individual or element in the universe an equal chance of being chosen." A random sample is more suitable in more homogeneous and comparatively large groups. When the universe is composed of differing groups of extremely varied sizes, then this method cannot be successfully used.

Methods of Drawing a Random Sample: Following four methods are generally used for drawing out a sample on random basis.

a. **Lottery Method:** Under this system, numbers or names of various units of the universe are written on chits, capsules or balls. They are put in a container. Then required numbers of chits are drawn from the container. The practical utility of such a method is very much limited.

b. Table or Random Number: Table or random number sampling have been constructed by Tippett's (1927), Fisher& Yates (1963), and Smith and Kendell (1939). Generally, Tippett's random number table is used for this purpose. Tippett gave 10,400 four figure number. He selected 41,600 digits from the census reports and combined them into fours to give his random numbers which may be used to obtain a random sample. **For Example:** let us we have first 30 sets of Tippett numbers

2953	**6641**✓	**3992**✓	**9792**	**7979**✓	**5911**✓
3170✓	**5624**✓	**4167**✓	**9525**	**7545**✓	**1396**
7203✓	**5356**✓	**1300**	**2693**	**2370**	**7483**✓
3408	**2729**	**3563**	**6107**	**6913**	**7691**
0506	**5246**	**1112**	**9025**	**0008**	**8126**

Let us suppose that we are to take a sample of ten units from a population of 5000 units which bears a numbers from 3001 -8000. So, we are to select 10 such figures from the above table which are not less than 3001 and greater than 8000.

c. Selecting from Sequential List: Under this method, all the units of the universe are arranged according to some particular order which may be alphabetical, geographical or simply serial. Then, every 5^{th}, 10^{th}, 100^{th} or nth is selected as random number from the list. For example, if every tenth unit is to be selected, the selection may begin from 7^{th} and then 17^{th}, 27^{th}, 37^{th} etc. may be selected.

d. Grid System: It is used for selecting a sample of area. According to this method, a map of entire area is prepared. Then a screen with squares is placed upon the map. Some of the squares

are selected random. Then the screen is placed upon the map and areas falling within the selected squares are taken as sample.

Advantage of Random Sampling: --(a) It is free from bias; (b) Generally more representative; (c) It is very simple and (d) Assessment of sampling error can be made.

Disadvantage of Random Sampling: (a) It is difficult to have a completely catalogued universe and thus selection according to strictly random basis is frequently not possible; (b) Cases selected may be too widely dispersed or even impossible to contact and thus adherence to whole sample may not be possible; (c) If the universe is of heterogeneous in nature then, this method is unsuitable.

2. Purposive or Deliberate Sampling: When the researcher deliberately or purposively selects certain units for study from the universe, it is known as purposive selection. In this type of sample selection, the choice of the selection is supreme and nothing is left to chance.

Merits: -(a)t can be widely used in business decision making; (b) Useful in stratified random sampling; (c) Very simple method of sampling.

Demerits: (a) There is possibility of inquiry being influenced by personal prejudice (judgment) and bias; (b) It is impossible to have an idea of degree of accuracy.

3. Stratified Sampling: It is a combination of both random sampling and purposive selection. Under this method, the universe is first divided into a number of groups or strata. Then from each group or stratum certain numbers of items are taken on random basis. Thus in the selection of strata we use purposive selection method, but in selecting actual units from each stratum, random method is used.

Process of Stratification: Following points may be kept in mind while constructing strata:

(i) First of all we should note the different variables involved in the study of the problem. The common variables used for stratification are generally region; income, sex etc. then divide the population into these groups and get the sample of required units from each group. In selecting the variables, care should be taken that they are related to study. (ii) The size of each stratum in the universe should be large enough to provide selection of items on random basis. (iii) There should be perfect homogeneity in the different units of strata. (iv) It is desirable that the number of items to be selected from each stratum should be in the same ratio as the total number of units in the stratum bear to the unit in the whole universe. (v) The strata should be clear cut and free from overlapping of units.

Kinds of Stratified Sampling: Stratified sampling itself is of the following three types:

a.) Proportional Stratified Sample: In this method, the number of units to be drawn from each stratum is in the same proportion as they stand in the universe. i.e. if P_i represents the proportion

of population included in stratum ' i' and 'n' represent the total sample size. Then the number of elements selected from each stratum 'i' will be equal to n (P_i).

Example: let us suppose that we want a sample of size n=30 to be drawn from a population of size N=8000 which is divided into three strata of sizes N_1=4000, N_2=2400, N_3=1600. Adopting proportional allocation, we shall get the sample sizes as under for three strata:

For strata with N_1, we have P_1= 4000/8000

Therefore, n_1= n.P_1= 30.(4000/8000)=15

Similarly, for strata with N_2=2400, we have

n_2= n.P_2 = 30. (2400/8000)= 9 and

For strata with N3=1600, we have

n_3= n. P_3= 30(1600/8000) =6

Thus, using proportional allocation, the sample sizes for different strata are 15, 9, and 6 respectively which is in proportion to the sizes of the strata viz., 4000: 2400: 1600.

Proportional allocation is considered most efficient and an optimal design when the cost of selecting an item is equal for each stratum. There is no difference in within stratum variances, and the purpose of sampling happens to be to estimate the population value of some characteristics.

b.) Disproportionate Stratified Sample: But in case the purpose happens to be to compare the differences among the strata, then equal sample selection from each stratum would be more efficient even if the strata differ in sizes. In case where the strata differ not only in size but also in variability and it is considered reasonable to take larger samples from more variable strata and smaller sample from a less variable strata, then we can account for both (differences in stratum size and differences in stratum variability) by using disproportionate sampling design.

Disproportionate sampling design requiring

$$n_1/N_1\sigma_1 = n_2/N_2\sigma_2 =----------= nk/N_k\sigma_k$$

Where, k is the number of strata

σ_1, σ_2…….. are the standard deviations of the strata

$n_1, n_2 \ldots n_k$ denote the sample size of k strata

N_1, N_2----N_K denote the sizes of k-strata.

This is called optimum allocation in the context of disproportionate sampling. The allocation in such a situation results in the following formula for determining the sample sizes for different strata.

$$ni = \frac{n_i N_i \boldsymbol{\sigma_i}}{N_1 \sigma_1 + N_2 \sigma_2 + \text{-----} + N_k \sigma_k}$$

Example: A population is divided into three strata so that $N_1 = 5000$, $N_2 = 2000$, $N_3 = 3000$. The respective standard deviation is: $\sigma_2 = 15$, $\sigma_2 = 18$ and $\sigma_3 = 5$. How should a sample of size n=84 be allocated to the three strata, if we want optimum allocation using disproportionate sampling design.

Solution: Using the disproportionate sampling design for optimum allocation, the sample sizes of different strata will be determined as under:

Sample size for strata with $N_1 = 5000$

$$n_1 = \frac{84(5000)(15)}{(5000)(15) + (2000)(18) + (3000)(5)} = 63,00,000/1,26,000 = 50$$

Sample size for strata with $N_2 = 2000$

$$n_2 = \frac{84(2000)(18)}{(5000)(15) + (2000)(18) + (3000)(5)} = 30,24,000/1,26,00 = 24$$

Sample size for strata with $N_3 = 3000$

$$n_3 = \frac{84(3000)(5)}{(5000)(15) + (2000)(18) + (3000)(5)} = 12,60,000/1,26,00 = 10$$

In addition to differences in stratum sizes and differences in stratum variability, we may have differences in stratum sampling cost, and then we can have cost optimum disproportionate stratified sampling design by requiring

$$n_1/N_1.\sigma_1.\sqrt{C_1} = n_2/N_2.\sigma_2.\sqrt{C_2} = \text{----------} = nk/N_k.\sigma_k.\sqrt{C_k}$$

where,

C_1 = Cost of sampling in strata 1

C_2 = Cost of sampling in strata2

C_k = Cost of sampling in strata 3

The allocation in such a situation results in the following formula for determining the sample sizes of different strata:

$$ni = \frac{n_i N_i \sigma_i / \sqrt{C_i}}{N_1 \sigma_1/\sqrt{C_1} + N_2 \sigma_2/\sqrt{C_2} + \text{------- -----} + N_k \sigma_k/\sqrt{C_k}}$$

For i = 1, 2, 3…..k

Merits of Stratified Sampling:- (a) In case of highly skewed population, this method is most suitable; (b)It gives better representation to characteristics of population in the sample; (c)It gives high degree of accuracy; (d) Less time is required because stratified sample are most concentrated geographically.

Demerits of Stratified Sampling: - (a) Subjected judgments often affect the accuracy of results; (b) The results may be misleading in case the basis of stratification is not properly determined; (c) Random selection of items for each stratum is essential.

4. Quota Sampling: This is a special form of stratified sampling. Under this method, the universe is first divided into different strata. Then the number to be selected from each stratum is decided. This number is known as quota. The field investigators are generally asked to select the quota from the stratum according to their will. Example, Let us suppose a sample of 500 families is to be selected, then the houses to be approached will be decided first and the field investigators will be asked to select one family from each house at their will.

Merits: -(a) Helpful in making research institutions and public opinion studies; (b) Satisfactory results are obtained.

Demerits: (a) There are chances of personal bias and prejudice entering the enquiry; (b) Accurate results are not obtained.

5. Multi-stage Sampling: This method is generally used in selecting a sample from a very large area, like an entire country, state etc. Under this method, the selection of the sample is made in different stages. Under multi-stage sampling, the first stage may be to select large primary sampling units such as states, then district, then towns and finally certain families within town. If the technique of random sampling is applied at all stages, the sampling procedure is described as multi-stage random sampling.

6. Convenience Sampling: It is generally known as unsystematic, careless, accidental or opportunistic sampling. According to this method, a sample is selected according to the convenience of the sampler. This convenience may be in respect of source list, accessibility of the units etc. although this method is most unscientific, yet quite a large number of samples are selected according to this method. This method is used in any one or more of the following cases:

-When the universe is not clearly defined; -Sampling unit is not clear; -A complete source list is not available.

7. Self- Selected Sampling: Sometimes a sample is not actually selected but people themselves opt to be included or not to be included in the sample. Suppose, for example, an enquiry has to be made about the people's liking for particular radio program me, and an announcement to this

effect is made on the radio. In such a case, the sample is not fixed. Those who care to reply from the part of the sample. Such a sample is known as self –selected sample.

8. Cluster Sampling and Area Sampling: Clusters refers to the particular area and thus cluster sample implies area sample. Cluster sample is mainly concerned with the particular geographical area or a particular aspect of population. The region of a country, blocks, and countries may constitute the cluster and within each group all units present may be included. In this method, the universe is first divided into some recognizable sub-groups which are called clusters. After this a simple random sample of these clusters is drawn and then all the units belonging to the selected clusters constitute the sample. Cluster or area sampling is practiced in sample survey.

9. Systematic Sampling: In some instances, the most practical way of sampling is to select every i^{th} item on a list. Sampling of this type is known as systematic sampling. In this method, first item is selected randomly from the list and then every i^{th} unit is selected automatically at fixed intervals. Example, if 4 per cent sample is to be drawn, then first item would be selected randomly from the first 25 and thereafter every 25^{th} item would automatically be included in the sample.

10. Sequential Sampling: Under this method, a small number of items are tested and the whole lot from which this small number is taken is either selected or rejected on the basis of the results obtained from the list.

Merits: (a) This is widely used in quality control; (b) In the selection of manufactured products or raw materials, this method is used.

Demerits:-This method gives confusing results.

Sampling with Probability Proportional to Size: If the cluster sampling units do not have the same number or approximately the same number of elements, it is considered appropriate to use a random selection process where the probability of each cluster being included in the sample is proportional to the size of the cluster. The following procedure is to be used.

Procedure: - (a) list the number of elements in each cluster irrespective of the method of ordering the cluster; (b) Take a sample systematically of the appropriate number of elements from the cumulative totals; (c) The actual number selected in this way do not refer to individual elements but indicate which clusters and how many from the cluster are to be selected by simple random sampling or by systematic sampling; (d) The results of this type of sampling are equivalent to those of a simple random sampling and the method is less cumbersome and is also relatively less expensive. We can illustrate this with the help of an example.

Following are the number of departmental stores in 15 cities:
5,17,10,32,70,28,26,19,26,66,37,44,33,29,and 28. If we want to select a sample of ten stores, using cities as clusters and selecting within clusters proportional to size, how many stores from each city should be chosen? (Use a starting point of 10).

City Number	No. of Departmental Stores	Cumulative Total	Sample
1	35	35	10
2	17	52	
3	10	62	60
4	32	94	
5	70	164	110 160
6	28	192	
7	26	218	210
8	19	237	
9	26	263	260
10	66	329	310
11	37	366	360
12	44	410	410
13	33	443	
14	29	472	460
15	28	500	

Since in the given problem, we have 500 departmental stores from which we have to select a sample of 10 stores.

The appropriate sampling interval= 500/10 = 50

As we have to use the starting point of 10, so we add successively increments of 50 till 10 numbers have been selected. The numbers, thus obtained are: 10, 60, 110, 160, 210, 260, 310, 360, 410, and 460 which have been shown in the last column of the above table against the concerning cumulative totals.

From this we can say that two stores should be selected randomly from the city number five and one each from city number 1, 3, 7, 9, 10, 11, 12, and 14.

This sample of 10 stores is the sample with probability proportional to size.

Note: if the starting point is not mentioned, then same can randomly be selected.

Limitation of sampling: Sample studies can give better results only if the samples are drawn systematically, their size is adequate and an appropriate sample design is used. Besides, these

sample survey suffer from certain limitations and if they are not properly conducted, they may give erroneous inferences.

If, for example, some selected units of sample did not respond and are left out or if the person conducted the survey are not qualified then, the sample results may be highly misleading. Sometimes, sample survey may need more time and labor also if the sample size is large and the sampling technique is complicated.

Sample size and its determination: In sampling analysis, the most ticklish question is: what should be the size of the sample or how large or small should be "n"? If the sample size (n) is too small, it may not serve to achieve the objectives and if it is too large, we may incur huge cost and waste resources. Therefore, the size of the sample should be determined keeping in view the following points:

(a)Nature of universe: Universe may be either homogeneous or heterogeneous. If the items of the universe are homogeneous, a small sample can serve the purpose. But, if the items are heterogeneous, a large sample would be required; **(b)** Number of classes proposed; **(c)** Nature of study; **(d)** Type of sampling; **(e)** Standard of accuracy and acceptable confidence level; **(f)** Availability of finance and **(g)** other considerations.

There are two alternative approaches for determining the size of the sample.

(i) Determination of sample size through approach based on precision rate and confidence level and **(ii)** Bayesian Statistical Approach.

The first approach is capable of giving mathematical solution and as such is a frequently used technique of determining the size of the sample (n). The second approach is theoretical optimal, but it is seldom used because of the difficulty involved in measuring the value of information.

Determination of sample size through the approach based on precision rate and confidence level: In the sampling study process there are chances of sampling error which can be controlled by selecting a sample of adequate size. Therefore, the researcher must specify the precision that that he wants in respect of his estimates concerning the population parameters. For instance, a researcher must like to estimate the mean of the universe within ± 3 of the true mean with 95% confidence. So, it means the desired precision is ±3 i.e. if the sample mean is 100, the true value of the mean will be no less than Rs. 97 and no more than Rs. 103. This means that the acceptable error (e) is equal to 3. Keeping this in view, we can now explain the determination of sample size so that specified precision is ensured.

Sample size (When estimating a mean): The confidence interval for the universe mean, **μ**, is given by

$$\bar{X} \pm Z.\ \sigma_p / \sqrt{n}$$

Where, $\bar{X}$ is sample mean; Z is the value of standard variate at a given confidence level (to be read from the table giving area under the normal curve and it is 1.96 for 95% confidence level; n is the size of the sample and σ_p is the standard deviation of population.

Suppose we have σ_p= 4.8 and if the difference between population mean **(**μ) and sample mean **(**$\bar{X}$) or the acceptable error is to be kept within ±3 of the sample mean with 95% confidence, then we can express the acceptable error (e) as equal to

$$e = Z.\ \sigma_p/\ \sqrt{n} \text{ or } 3 = 1.96.\ (4.8)/\ \sqrt{n}$$

$$\text{Hence, } n = (1.96)^2\ (4.8)^2/\ (3)^2 = 9.834 = 10$$

Generally, if we want to estimate μ in a population with standard deviation σ_p with an error no greater than "e" by calculating a confidence interval with confidence corresponding to Z, the necessary sample size becomes:

$$n= [(Z.\sigma_p)/e]^2$$

The above formula is applicable when the population happens to be infinite. But, in case of finite population, the above stated formula for determining the sample size becomes:

$$n= (N.Z^2.\sigma_p^2)/[(N-1)e^2 + Z^2.\sigma_p^2]$$

Where, N is the size of the population; n is the size of the sample; e is the acceptable error or precision; σ_p is the standard deviation of population; and Z is the value of standard variate at the given confidence level.

Illustration: Determine the size of the sample for estimating the true weight of cereal containers for the universe with N = 5000 on the basis of the following information:

(1) The variance of weight = 4 ounces on the basis of past records.

(2) Estimate should be within 0.8 ounces of the true average weights with 99% probability.

Will there be a change in the size of the sample if we assume infinite population in the given case? If so, explain by how much.

Solution: Given that N = 5000;

σ_p = 2 ounces (since the variance of weight = 4 ounces);

e = 0.8 (since the estimate should be within 0.8 ounces of the true average weight);

Z = 2.57 (as per the table of area under the normal curve for the given confidence level of 99 %)

Hence, the confidence interval for μ is given by:

$$\bar{X}\pm Z.\ \boldsymbol{\sigma_p}\ /\ \sqrt{\mathbf{n}}.\ \sqrt{\ \mathbf{(N-n)/\ (N-1)}}$$

And accordingly the sample size can be worked out as under:

$$n= (N.Z^2.\sigma_p^2)\ /\ [(N-1)e^2 + Z^2.\sigma_p^2]$$

By substituting the values we get:

$$n = (2.57)^2.(5000).(2)^2\ /\ [\ (5000-1)(0.8)^2 + (2.57)^2(2)^2\ ]$$

$$n = 132098/ [3199.36 + 26.4196] =132098 / 3225.7796 = 40.95 = 41$$

Hence, the sample size, n, = 41 for the given precision and confidence level in the above question with finite population. But, if we take population to be infinite, the sample size will be worked out as under:

$$\mathbf{n= [(Z.\sigma_p)/e]^2}$$

By substituting the values we get: $n = (2.57)^2 (2)^2 / (0.8)^2 = 26.4196 / 0.64 = 41.28 = 41$.Thus, in the given case, the sample size remains the same even if we assume infinite population.

Sample size (when estimating a percentage or proportion): Here, we shall have to specify the precision and the confidence level and then we will work out the sample size as under:

Since the confidence interval for the universe proportion, is given by

$$p \pm Z.\sqrt{pq/n}$$

Where p is the sample proportion; q=1-p; Z is the value of standard variate at a given confidence level and to work out from the table showing area under the normal curve; and n is the size of the sample.

With the given precision rate, the acceptable error 'e' can be expressed as under:

$$e = Z.\sqrt{pq/n} \text{ or } e^2 = Z^2pq/n$$

Hence, $n= (Z^2.p.q)/e^2$

The above formula gives the size of the sample in case of infinite population when we are to estimate the proportion in the universe. But, in case of finite population, the formula becomes:

$$n= (Z^2.p.q.N)/[e^2(N-1)+ Z^2.p.q]$$

Illustration: What should be the size of the sample if a simple random sample from a population of 4000 items is to be drawn to estimate the per cent defective within 2 per cent of the true value with 95.5 per cent probability? What would be the size of the sample if the population is assumed to be infinite in the given case?

Solution: We have, N = 4000;

e = 0.02 (since the estimate should be within 2% of true value);

Z = 2.005(as per the table of area under the normal curve for the given confidence level of 95.5 %)

As we have not been given the p value being the proportion of defectives in the universe, let us assume it to be p = 0.2 (This may be on the basis of our experience or on the basis of past data or may be the result of a pilot study).

Since $n= (Z^2.p.q.N)/[e^2(N-1)+ Z^2.p.q]$

By substituting the values, we get:

$$n = (2.005)^2 (0.2)(1-.02) (4000) / [(0.02)^2 (4000-1) + (2.005)^2 (0.2) (1-.02)]$$

$$n = 315.1699 / [1.5996 +0.788] = 315.1699 / 1.6784 = 187.78 = 188$$

This is the sample size in case of finite population. But, if the population happens to be infinite, then our sample size will be as under:

$$n = (Z^2.p.q)/e^2$$

By substituting the values, we have:

$$n = (2.005)^2(0.02)\ (1\text{-}0.02) / (0.02)^2 = 0.0788 / 0.0004 = 196.98 = 197$$

Thus, the sample size in case of infinite population is 197.

Case Study: The method of scientific social research may broadly be divided into two parts: - The statistical methods and the case study methods.

Statistical methods are based on large scale collection of facts, while case study is based on intensive study of comparatively fewer persons, sometimes confined to a very small number of cases only. The case study is thus more intensive in nature. The field of study is comparatively limited but has more of depth in it. It aims at studying everything about something rather than something about everything as in the case of statistical methods.

According to P.V.Young describes case study is a method of exploring and analyzing the life of a social unit, be that a person, a family, an institution, cultural group or even entire community.

Census versus sample enumeration

Sample data can be collected either on the basis of census type of enquiry or on the basis of sample method.

Census: It means complete enumeration. Human /population census is always conducted after ten years while, livestock census is always conducted after five years.

Census method/ complete enumeration: In the census type of enquiry, there is a complete enumeration of all the items of the universe and the question of taking a sample does not arise. It is used when the complete information is needed about the entire universe. It is also used when the size of the universe is not big and the need for accurate result is great. It gives exact and accurate results. It is suitable where area of inquiry is not vast; it is suitable where there is enough time available for data collection; it is also suitable where there is higher degree of accuracy is required; it is also suitable where there is enough finance available to meet the expenditure on the collection of statistics.

Limitations of this method: very expensive if the size of universe is large; -it requires more time for completion; -it needs substantial man power and administrative control.

Sample survey/ sample census: In this, we study some selected items from the universe for drawing general inferences. It is suitable where the census method cannot be used especially where area of inquiry is wide; it is suitable where financial constraint is there on the collection of statistics; it is suitable where there is difficult to get the complete data; and it is also suitable where the goods or the items of inquiry is very changeable.

Merits/ advantage of sampling method: This method has the advantages of speed, economy, adaptability, time saving and it has a scientific approach. It takes less time; it is less expensive; and it is more dependable.

Demerits of Sampling Method: (a) Change of bias- A bias in the sample may be caused either by faulty method of sampling or the nature of the phenomenon itself**; (b)** Difficulties of a representative sample; **(c)** Need for specialized knowledge**; (d)** Difficulties in sticking to sample; and **(e) Impossibilities of Sampling-** Sometimes the universe is too small or too heterogeneous so that it is impossible to draw a representative sample. In such cases, census study is the only alternative.

Different steps in large scale sample survey:

i) Planning stage;

ii) Execution stage;

iii) Analysis and reporting stage.

(i) Planning stage consists of the following steps:

a. Defining the objectives; **b.** Defining the population; **c.** Determination of data should be collected; **d.** Deciding on the method of data collection i.e. whether interview method (house to house enquiry for the collection of data) or mail questionnaire method (mailing of the questionnaire to individual of population for filling in and returning them); **e.** Choice of sampling unit; **f.** designing the survey; **g.** drawing the sample; **h.** Training of personnel

(ii) The execution stage involves the identification of the sampled individual in the field and the filling up the questionnaire.

(iii) The analysis and reporting stage again consists of the following step:

a. Scrutiny of data; **b.** Tabulation of data; **c.** Statistical analysis; **d.** Reporting; **e.** Storing of information for future survey.

Data and the different lines of data collection

Data summarizes the fact about the phenomenon under investigation. These facts may be of various types and derive from various sources. They may be fundamentally qualitative, quantitative or both**.**

Types of Data:

Primary and secondary data: Primary data are those which are collected a fresh and for the first time and thus happens to be original in character. While, the facts and figures that have already been collected by someone else and which have already been passed through the statistical process are called the **secondary data.**

Qualitative and Quantitative data: Qualitative data refers to situations, attitudes, positions, qualitative characteristics, marital status, and condition of country (industrialized, poor,

developing, etc), sex, and education. No numerical measure exist for such qualitative facts whereas, quantitative data are numerically measured.

Time series and cross-sectional data: Quantitative data are of two types i.e. Time-series and cross-sectional. Time series data refers to a set of observation on a particular variable at different points of time, while cross-sectional data refers to a set of observation on different variables at a given point of time. **For example**- production of wheat in Kangra during 1990-91 to 2000-01 is time series while, the production of wheat in different district of Himachal Pradesh during one agricultural year is cross-sectional data.

Experimental and non-experimental data: Experimental data are obtained from controlled experiments conducted by the researcher whereas; non-experimental data are obtained when the problem under investigation is not subject to any control. Generally the data obtained in the natural or biological sciences are the experimental data or controlled data, whereas in social sciences like agricultural economics where the underlying conditions are not subject to control are the non-experimental data.

Sources of Data:

Published sources:

International publications (IMF,ILO,UNO.World Bank); Governmental publications (Central, State. Local); Semi-governmental publications (Boards and Corporations); Reports of Commissions and Committees; Research Publications; Research institutions; News papers, Periodicals, Magazines; and Individuals.

Unpublished sources:

Scholars; Research workers; Trade associations; Chamber of commerce; and Labor bureau

Chapter 5

CORRELATION AND REGRESSION

Abstract

Correlation and regression are essential statistical tools used to analyze the relationship between variables. Correlation measures the strength and direction of a linear relationship between two variables, indicating how one variable changes in response to another. Regression, on the other hand, goes a step further by not only measuring this relationship but also predicting the value of a dependent variable based on one or more independent variables

Keywords: Independent Variable, Dependent Variable, Correlation, Regression, Cause and Effect.

Introduction:

It is a common experience in our country that the level of production depends on the monsoon. In a year of good rainfall, the agricultural production is high and when monsoon fail, the agricultural output is low. We also know that farm with assured irrigation are generally a higher productivity per hectare than the un- irrigated one. Many other examples can be given off the relationship between variables. Economic theory teaches us that there is a relationship between income and consumption, cost & output. These all are the examples of two variables. Very often the relationship might extend to three or more variables. For example, productivity per hectare may not be only related to irrigation but also on the quality of seed, pesticides etc. the quantity demanded of a commodity may be related to disposable income, taste, prices of related commodities in addition to its own price. Similarly, in addition to output, production cost is related to wage rate and type of technology used. The consumption expenditure of a household will be related besides household income to household size, consumption habit and social status etc. There are various methods for measuring the relationship existing between the economic variables. The simplest are the correlation and regression analysis.

The measurement of degree of relationship between variables is called simple correlation and the degree of relationship connecting three or more variables is called multiple correlations. The correlation only indicated the degree and direction of relationship between two variables. However, it does not apply to cause-effect relationship, even when there are grounds to believe that casual relationship exist, correlation does not tell us which variable is the cause and which one is the effect. For example, the demand for a commodity and its price will generally be found to be highly co-related. But, the question whether demand depends on price or price depend s on demand is not be answered by correlation. In the example on demand and price, economic theory tells us that other thing being equal, quantity demanded of a commodity depends on its price. It is possible to test this hypothesis and further to find out by how much on an average is demand expected to change if price change by certain percentage. This is done through regression

analysis which describes the functional relationship and enables us to make estimates of one variable from another.

Correlation and Causation: Correlation analysis helps us in determining the degree of relationship between two or more variables. But it does not tell us anything about cause-and-effect relationship. Even a high degree of correlation does not necessarily mean that a relationship of cause and effect exists between the variables. The explanation of a significant degree of correlation may be any one of the following factors:

a. One variable being the cause of other: When variable causes the change of other variable, the variable which is the cause is called the subject or the independent variable and the other variable is called as relative or dependent variable.

b. Both variables being the result of a common cause or both the co-related variables may be influenced by one or more other variables: For example-(i) a high degree of correlation between the yield per hectare of rice and tea may be due to the fact that both are related to the amount of rainfall. But, none of the two variables is the cause of other. (ii) Suppose, the correlation of teacher's salaries and the consumption of liquor over a period of year comes out to be 0.9. This does not prove that neither teachers drink, nor does it prove that liquor sale increases teacher salaries. Instead both variables move together because both are influenced by third variable i.e long run growth in national income and population.

c. Both the variables may be mutually influencing each other so that neither can be designated as the cause and other the effect: There may be high degree of correlation between variables. But, it is very difficult to pin-point as to which the cause is and which the effect is. This is especially likely to be so in case of economic variables. Examples of such variables are such as demand and supply, price and production etc. These are mutually interacting. But, it is also possible that increase demand of a commodity due to growth of population or an upward pressure on price. Now, the cause is the increase demand and effect is price. That means price is a function of demand.

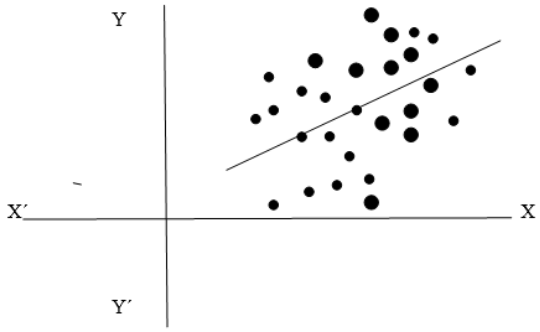
d. Chance: Sometime, it has been seen that between two variables a fair degree of correlation may be observed when one exist in the universe. It is just possible that the existence of correlation may be by-chance or accident. So, such a correlation observed between variables that cannot be casually related is called spurious or non-sense correlation. If there is a cause and effect, there is a correlation linear or otherwise. But, the reverse is not necessarily true. For example- there is a extremely high correlation between series represented by the production of pig and the production of iron. Yet, no one has ever believed that this correlation has any meaning or that it indicated the existence of cause-effect relationship.

Correlation: The term correlation indicates the relationship between two such variables in which with changes in the value of one variable, the value of the other variable also changes. On the other hand, it gives the relationship between two variables whether positive or negative. Relationship depends upon how thick is the correlation i.e. it is perfect one or limited perfect one.

Measures of Correlation:

1. Scattered diagram;
2. Graphic method;
3. Karl Pearson's coefficient of correlation;
4. Coefficient of rank correlation;
5. Partial correlation coefficient;
6. Zero order correlation coefficient
7. Regression line (method of determining average relationship between variable

1. Scattered Diagram: The scatter diagram helps us to visualize the relationship between two phenomenons**.** It indicates the strength of relationship between two variables. If the point lies close to the line, the correlation is strong, on the other hand, greater dispersion of points about the lines implies weaker correlation. It gives only rough idea of the relationship between variables X&Y.

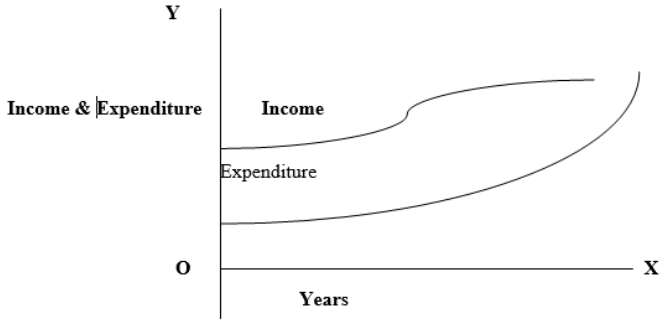


If the values are overlapping then, this measure is not a reliable because it simply gives us perfect correlation or direction of correlation. Correlation should be quantified.

Merits: It is very easy to draw a scattered diagram; it can be easily understood and interpreted; and values of extreme items do not affect this method, such points are always isolate in the diagram.

Demerits: It only gives a visual picture of the relationship of two variables; it only tells us whether there is correlation between the variables and if so, then in which direction positive or negative; it does not gives an idea about the precise degree of relationship as it is not amenable to mathematical treatments.

2. Graphic Method:

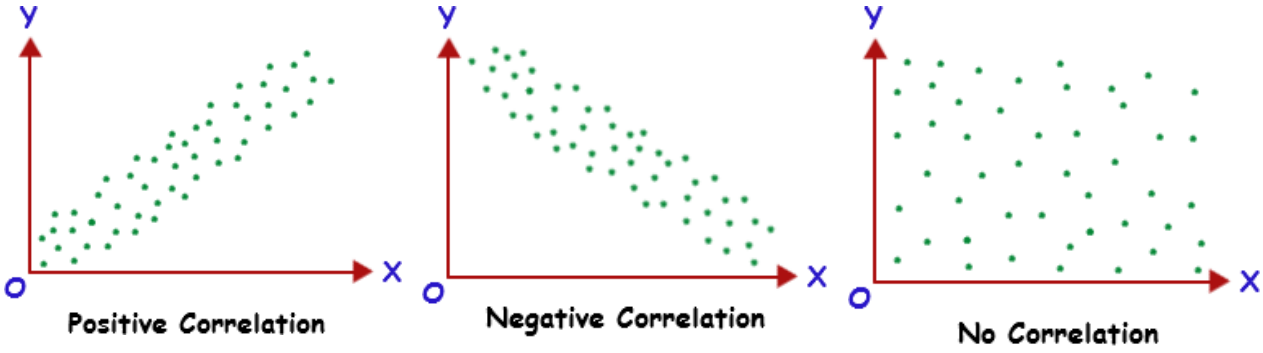


3. Pearson's Coefficient of Correlation: For precise quantitative measurement of the degree of relationship between X&Y, we use Karl Pearson's coefficient of correlation denoted by "r".

$$r_{xy} = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2} \sqrt{\sum y_i^2}}$$

The value of the correlation coefficient may assume vary from -1 to +1. When r = +ve, It implies that there is a perfect correlation between X&Y. When r = -ve, it implies that there is perfect negative correlation between X&Y. When r =0, then the variables are uncorrelated.

The graphical representation of positive, negative and no correlation are shown below:



The Pearson correlation coefficient is denoted by the letter "r"

4. Spearmen's Rank Correlation Coefficient: If the variables are of qualitative or binary in nature then, it is possible to use this statistics. Rank correlation coefficient is calculated by the formula given below:

$$r' = \frac{6\sum D^2}{N(N^2-1)}$$

Where, D = difference between rank of corresponding pairs of X&Y.

N =number of observations.

5. Partial correlation coefficient: It measures the relationship between any two variables, when all other variables connected with two are kept constant. The simple correlation coefficient between two variables can be written as:

r_{12} = correlation coefficient between X_1 and X_2.

r_{13} = correlation coefficient between X_1 and X_3

r_{23} = correlation coefficient between X_2 and X_3

Therefore, Partial correlation coefficient between X_1 and X_2 when X_3 is kept constant is given by:

$$r_{12.3} = \frac{r_{12} - (r_{13})(r_{23})}{\sqrt{1 - r_{13}^2}\sqrt{1 - r_{23}^2}}$$

6. Zero-Order Correlation Coefficient (r_{x1x2}): The relationship between two variables at a time and none of them held constant. For more than two variables, we use partial correlation.

Types of Correlation:

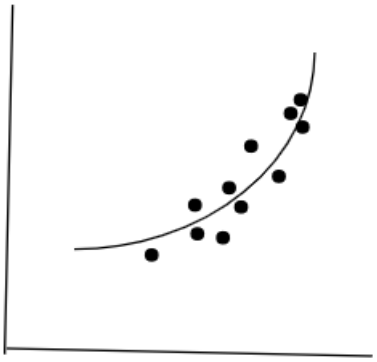
a. Simple correlation: Measurement of degree of relationship between variables.

b. Multiple correlations: Degree of relationship connecting three or more variables is called multiple correlations.

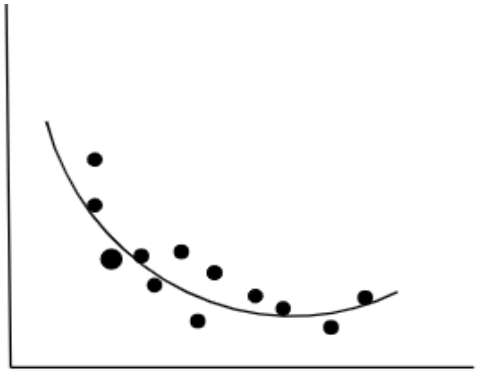
c. Positive correlation: The correlation is said to be positive when the higher values of one variable are associated with the higher values of the other variable. Or when the lower values of one variable are accompanied by the lower values of the other variable. In such cases the movements of the two variables are in the same direction. For example, quantity of commodity supplied and its price. When the price rises, the quantity offered increase and when price fall, the quantity supplied decreases. If all the point lies on a line or curve, the correlation is said to be perfect positive.

d. Negative or inverse correlation: It happens, when the higher values of one variable are associated with the lower values of the other variable, the correlation is said to be negative. In such a case, the movements of the two variables are in the opposite direction. For example, quantity demanded and their prices are negatively correlated. If all the points lie on the line or curve, the correlation is said to be perfect negative.

e. No correlation or zero correlation: The correlation is said to be zero when two variables tends to change with no connection to each other. In the scattered diagram, the points are dispersed all over the surface of XY plane. For example, one should expect zero correlation between the height of inhabitants of a country and the production of steel or the weight of the students and the color of their hair.



Positive non-linear correlation



Negative non-linear correlation

f. Linear and non-linear (curvilinear) correlation: The distinction between the linear and the non-linear correlation is based upon the ratio of change between the variables.

If the amount of change in one variable tends to bear constant ratio to the change in the other variable, then the correlation is said to be linear. For example,

X	10	20	30	40
Y	70	140	210	280

The correlation would be called non-linear or curvilinear if the amount of change in one variable does not bear a constant ratio to the amount of change in other variable. For example, if the amount of rainfall doubles, the production of wheat or rice etc. would not necessarily be doubled**.** The techniques of analysis for measuring non-linear correlation are more complicated than those

for the linear correlation. Therefore, we generally make an assumption that the relationship between the variables is of linear type.

g. Spurious or non-sense correlation: The variables that cannot concise ably be casually related is called spurious correlation or non-sense correlation. For example, there is a extremely high correlation between the series represented by the production of pig and the production of iron .Yet, no one have ever believed that this correlation has any meaning or that it indicates the existence of cause-effect relationship.

Limitations of Linear Correlation:

The formula

$$\text{i) } r_{xy} = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2} \cdot \sqrt{\sum y_i^2}}$$

is applicable only when we have linear correlation between the variables. Though, the two variables may be strongly connected with non-linear relationship.

ii) The linear or simple correlation does not tell us to estimate the cause and effect relationship between the variables.

iii) Linear correlation coefficient does not help us in estimating the numerical values of the parameters i.e. regression coefficients.

Test of significance 'r' when the true population (ϱ) =0: If true population (ϱ) =0, the sampling distribution of 'r' (estimate of ϱ) is symmetrical.

$$t^* = \text{variable} / \text{SE (variable)} = r / \sqrt{(1-r^2)/n-1}$$

Therefore, $t^* = r(\sqrt{n-2}) / \sqrt{1-r^2}$

Then we compare the calculated value of 't' with the table value of 't' with (n-2) d.f. at 5% or 1% LOS.

Test of significance of "r" when the true population, is not equal to 0: it means the distribution is not normal but it is skewed. The higher value of the true population, more skewed the sampling distribution of 'r', then Fisher Test is applied.

Test of significance of rank correlation coefficient (r´) when true population, ϱ, =0: In this case the sampling distribution of r´ has a normal distribution. To test this statistics, Z-test is applied.

To conclude from the concept of correlation, we say that correlation only indicates the degree and direction of relationship between two variables. However, it does not apply to cause-effect relationship i.e. correlation does not tell us which variable is the cause and which one is the effect. For example, demand for a commodity and its price generally be found to be highly correlated. But, the question whether demand depends on price or vice-versa is not be answered by correlation theory.

Regression Analysis: Regression analyses describe the functional relationship between the variables and enable us to make estimate of one variable from another. For example, economic theory tells us that, other thing being equal, quantity demanded of a commodity depends on its price. It is possible to test this hypothesis and further to find out by how much on an average is demand expected to change if the price change by certain percentage. This is done through

regression analysis. It tells us the cause and effect relationship. The variable which is the cause is called the subject or the independent variable and the other variable is called as relative or dependent variable.

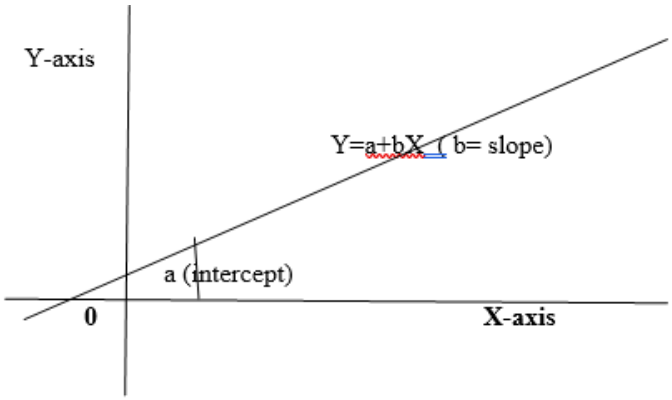
Thus, regression analysis attempts to establish the nature of relationship between the variables i.e. to study the functional relationship between the variable and thereby provide a mechanism for prediction or forecasting.

Regression: Regression is the measure of the average relationship between two or more variables in terms of the original units of the data.

Lines of Regression: Literary meaning of regression is 'stepping back towards average'. This concept of regression was first used in biometry. But, nowadays, regression word is used in statistics. For convenience if two variables are measured on the same individual then the corresponding scattered diagram will show that they are clustered around a curve what we called as regression line. It the curve is a straight line then that is known as regression line. There are two types of regression lines in case of bi-variate data:

i) To predict the value of say 'y' variable for a given value of x-variable.

ii) To predict the value of 'x' variable for a given value of y-variable.



Properties of Regression line:

i) Regression coefficient is affected by change of scale but not by the change of origin.

ii) Means of observed and the estimated value from the regression line are the same. .

iii) The magnitude of correlation coefficient between two variables denotes the proportion of variability in the observed value which is accounted for by the regression equation.

$Y = a + bX$

iv) Correlation coefficient is the geometric means of the regression coefficient.

$$GM = (x_1.x_2)^{1/2}$$

$$r_{xy} = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2} . \sqrt{\sum y_i^2}}$$

v) If one of the regression coefficient is <1, the other will be >1.

i.e. $b_{yx} < 1$ then $b_{xy} > 1$

vi) The arithmetic means (AM) of the regression coefficient is always greater than the value of the correlation coefficient (r), i.e. AM of two regression line is given by:

$$(b_{yx}+ b_{xy})/2 > r$$

Also note that

- If two regression lines are perpendicular to each other, it means there is no correlation between the variables.
- If the regression lines moves from right to left, then, there is negative correlation.
- Linear correlation coefficient between two independent variables is zero.
- Correlation analysis does not provide numerical values for the coefficients of the functional relationship. i.e. slope and intercept cannot be worked out.

Equation: Statement of equality between the two quantities. For example, 5X =15 then, for X =3, there is equality**.**

Identity: If the equality is true for all values assigned to X. Then it is called identity. For example,

$(X^2 -4) = (X-2) (X+2)$

Regression Equation: Algebraic expression of the regression line is the regression equation. For example:

Y = a + bX > this is regression line of Yon X.

X = a + bY > this is regression line of X on Y.

Model: It is the miniature of real world economic complexity.

Function: Functional relationship between dependent and independent variable.

Difference between Correlation and Regression:

i) Correlation coefficient is a measure of degree of co variability between and Y while, regression analysis is to study the nature of relationship between the variables so that we may able to predict the value of one variable on the basis of other.

ii) In correlation analysis, we cannot study the cause and effect relationship, while in regression analysis it is possible to study the cause and effect relationship.

iii) There may be a non-sense correlation between the two variables which is purely due to chance and has no practical relevance such as increase in income and increase in weight of a group of people. However, there is nothing like in regression analysis.

iv) Correlation coefficient is independent of change of scale and origin while, regression coefficients are independent of change of origin but not of scale.

Conclusion

To conclude from the concept of correlation and regression, we say that correlation only indicates the degree and direction of relationship between two variables. However, it does not apply to cause-effect relationship i.e. correlation does not tell us which variable is the cause and which one is the effect. For example, demand for a commodity and its price generally be found to be highly correlated. But, the question whether demand depends on price or vice-versa is not be answered by correlation theory. While regression analysis attempt to establish the nature of relationship between the variables i.e. to study the functional relationship between the variable and thereby provide a mechanism for prediction or forecasting.

REFERENCES

Mahajan, G. (2026, February). An introduction to descriptive statistics. In P. K. Ray, P. Gore, R. Deshmukh, & M. A. Gud (Eds.), *Advances in sustainable agriculture and allied sciences* (First Edition, pp. 45–61). © Bhumi Publishing, India. ISBN: 978-93-47587-75-7. https://doi.org/10.5281/zenodo.18814888

Mahajan, G. (2026, March). Probability and random experiments: Basics of uncertainty. In M. R. Verma, D. V. N. Sanjana Veni, P. V. Shelke, & V. S. Nirmalkar (Eds.), *Advances in sustainable agriculture and allied sciences* (Vol. II, pp. 84–93). © Bhumi Publishing, India. ISBN: 978-93-47587-43-6. https://doi.org/10.5281/zenodo.19370408

Mahajan, G. (2026, March). Inferential statistics: From sample to population insights. In M. R. Verma, D. V. N. Sanjana Veni, P. V. Shelke, & V. S. Nirmalkar (Eds.), *Advances in sustainable agriculture and allied sciences* (Vol. II, pp. 69–83). © Bhumi Publishing, India. ISBN: 978-93-47587-43-6. https://doi.org/10.5281/zenodo.19370408

Mahajan, G. (2025, November). Introduction to sampling and sampling fundamentals. In P. Chand, G. Handique, S. Bhagawati, & D. Meena (Eds.), *Precision and digital agriculture* (First Edition, pp. 31–51). © Bhumi Publishing, India. ISBN: 987-93-48620-67-5. https://doi.org/10.5281/zenodo/17864737

Mahajan, G. (2026, February). An introduction to correlation and regression. In P. K. Ray, P. Gore, R. Deshmukh, & M. A. Gud (Eds.), *Advances in sustainable agriculture and allied sciences* (First Edition, pp. 36–44). © Bhumi Publishing, India. ISBN: 978-93-47587-75-7. https://doi.org/10.5281/zenodo.18814888

Introduction to Statistical Methods

(ISBN: 978-93-47587-95-5)

About Author

Dr. Girish Mahajan

Extension Specialist (Agricultural Economics),
Department of Agricultural Economics,
Krishi Vigyan Kendra- Bara- Hamirpur (H.P.)

Dr. Girish Mahajan is an Extension Specialist (Agricultural Economics) at CSKHPKV–Krishi Vigyan Kendra, Hamirpur (Bara), Himachal Pradesh. He holds a Ph.D. in Agricultural Economics from CSKHPKV, Palampur, and specializes in Production Economics and Farm Management. He completed his B.Sc. (Agriculture) in 1992, M.Sc. in 1994, and Ph.D. in 1998 from the same institution, securing top academic distinctions including a Gold Medal and merit scholarships. He qualified ICAR-NET and has gained international exposure as a Post Doctoral Researcher at the Centre for International Environment and Development Studies, Norway. With over 25 years of professional experience, he has served in various capacities including Senior Research Associate, Research Officer with the Second Punjab Finance Commission, and Research Associate in multiple agricultural economics projects. Since 2007, he has been working as an Extension Specialist, first at KVK Kangra and currently at KVK Hamirpur, contributing extensively to agricultural extension and farmer outreach programs.

Dr. Mahajan has actively handled several national and institutional projects such as ARYA, NICRA, PKVY, CFLD (Oilseeds and Maize), and Mera Gaon Mera Gaurav as Co-PI and Nodal Officer. He has published 24 research papers in reputed journals and authored books, book chapters, teaching manuals, and five major technical reports. His extension contributions include over 40 articles in newsletters, pamphlets, booklets, and success stories aimed at disseminating agricultural knowledge in simple language. He has attended numerous conferences, workshops, and trainings, and has taught both undergraduate and postgraduate courses while guiding students. His achievements include the Norwegian Government Scholarship, Santosh Shiksha Puruskar, and several academic distinctions. Through his work, he continues to strengthen sustainable agriculture, farmer capacity building, and policy-oriented research in hill agriculture systems.

Contact Details: Mobile No. 8988293678
E-mail: lovely_nickname@rediffmail.com

Bhumi Publishing

ISBN 978-93-475-8795-5