

KARNATAKA STATE OPEN UNIVERSITY

**MANASAGANGATRI
MYSURU-570006
KARNATAKA, INDIA**

Third Semester M. Sc. MATHEMATICS

MMDSC 3.4(C) - MATHEMATICAL STATISTICS

DR. BHAT SATISH SHANKAR
Associate Professor, Department of Statistics
Yuvaraja's College
University of Mysore, Mysuru-57005

MATHEMATICAL STATISTICS

CONTENTS

	Title	Page No.
BLOCK - I	Descriptive Statistics	
Unit 1	Introduction to Statistics	4-116
Unit 2	Measures of Central tendency	17-43
Unit 3	Measures of Dispersion	44-58
Unit 4	Moments, Skewness and Kurtosis	59-67
BLOCK - II	Probability and random Variables	
Unit 5	Introduction to Probability Theory	69-87
Unit 6	Random Variable and Probability distributions	88-102
Unit 7	Mathematical Expectation of a random variable	103-112
Unit 8	Central limit Theorem	113-117
BLOCK - III	Probability & Sampling distributions and Estimation	
Unit 9.	Standards Discrete Probability Distributions	118-128
Unit 10.	Standard Continuous Probability Distributions	129-137
Unit 11.	Sampling Distributions	138-143
Unit 12.	Point and interval Estimation	144-160
BLOCK - IV	Test of significance	
Unit 13	Introduction to Testing of Hypothesis	162-166
Unit 14	Large Sample Test	167-174
Unit 15	Small Sample Test - I	175-186
Unit 16	Small Sample Test - II	187-203
	Statistical Tables	204-208
	References	209

PREFACE

This study material is intended to serve as text for reference for the course Mathematical Statistics for the first semester course of M. Sc. Degree in Mathematics offered by Karnataka State Open University(KSOU), Mysuru. Here, they study the topics on descriptive statistics and inferential statistics at the post graduate level. It is designed primarily for students who have no prior knowledge of probability and or statistics is assumed. It provides a well balanced introduction to mathematical statistics and probability theory.

It consists of four blocks; every block consists of four units each. Block I, covers basic knowledge of descriptive statistics, the block - II consist of introduction to probability theory, random variable and probability functions, mathematical expectation and the concepts of central limit theorem. Block - III consists of four units such as some standard discrete and continuous Probability Distributions, Sampling Distributions and theory of Estimation. Lastly, the block IV, consists of concepts regarding testing of hypothesis for both large and small samples.

Finally, in this study material there are bound to be misprints, errors and or ambiguities in presentation. I am grateful to any reader who notices these to my attention.

Dr. Bhat Satish Shankar

BLOCK – I
(DESCRIPTIVE STATISTICS)

UNIT 1: INTRODUCTION TO STATISTICS

UNIT 2: MEASURES OF CENTRAL TENDENCY

UNIT 3: MEASURES OF DISPERSION

UNIT 4: MOMENTS, SKEWNESS, AND KURTOSIS

UNIT 1

INTRODUCTION TO STATISTICS

1.1 Objectives

The main objective of statistics is to collect, interpret, present and analyse a data obtained from statistical surveys. In this unit our aim is to give an idea about data types, methods of collection, interpretation etc.

1.2 Introduction

In ancient days statistics was just pertaining to a state, where state head, or the emperor, or the king, would like to know about total population of his state, their economy levels, i.e., how many people are of rich class, middle class, poor class, etc. of the population, so that how much tax could be collected from them in a year to run state affairs for various reasons. Also, he would like to know how much fertile land present in his state, so that how much agricultural production could be expected and so on. But nowadays statistics is used in all walks of our life. For eg., it is used in various fields such as agriculture, industries, five year planning, defence, sports, budget preparation, social science, clinical trials, census-to know total population, birth rate, death rate, literacy rate, economy level, category, etc. Thus science without statistics does not bear fruit, i.e., in present days statistics is an integral part of our life.

1.3 Definitions of statistics:

'Statistics is literally defined as a science of averages or a science of counting'. Several authors have been defined the term statistics in different ways in the literature. Statistics are defined mainly in two senses, namely, singular sense and plural sense. Here important but collective thoughts of all the authors have been given in the following two definitions. They are

- a. Croxton –Cowden's definition of statistics (**In Singular sense**).
- b. Prof. Horace Secrist's definition of Statistics (**In Plural sense**).

1.3a Croxton –Cowden's definition of statistics: *Statistics may be defined as the science of collection, presentation, analysis, and interpretation of numerical data.*

1.3b Prof. Horace Secret's definition of Statistics - *Statistics may be defined as an aggregate of facts, affected to a marked extent by multiplicity of causes, numerically expressed, enumerated or estimated according to reasonable standards of accuracy, collected in a systematic manner, for a predetermined purpose and placed in relation to each other.*

1.4 Functions of Statistics

Following are the some important functions of statistics

- i. It simplifies the complexity of the data
- ii. It indicates trends and tendencies
- iii. It compares one set of data with other
- iv. It establishes the relationship between two sets of data

- v. It guides the management in planning
- vi. It measures the effects of government policies

1.5 Limitations of Statistics:

- i. It does not deal with qualitative phenomenon
- ii. It does not deal with single item
- iii. Statistical laws are not exact
- iv. It liable to be misused
- v. It does not reveal entire story
- vi. Statistical results are true only an average

1.6 Some Terminologies

i. Population: It is the collection of all facts taken in to consideration under study. Population can be finite or infinite. If a population consists of countable number of units, then it is called as finite population otherwise, n it is called as infinite population. For example, large scales industries in Karnataka state constitute a finite sample where as number of stars in the sky constitute a infinite population.

ii. Sample: It is the representative part of the population. It possess the characteristics of the population.

iii. Enumerator: The field agents who put the questions in questionnaire

iv. Data: A collection of numerical observations of known facts is called a data.

1.7 Types of Data

i. Primary data: It is a firsthand data. It is fresh and originally collected by the investigator.

ii. Secondary data: The data which are published or unpublished or processed by some agency already. It is second hand data.

iii. Qualitative data: The data which cannot be measured or expressed numerically but presence or absence can be felt are called qualitative data. For eg., blindness, literacy, beauty etc.

iv. Quantitative data: The data which can be measured or expressed numerically are called quantitative data. For eg., height and weight of persons. Wages, prices etc

1.8 Collection or Sources of primary data

- i. Direct personal investigation
- ii. Indirect oral interviews
- iii. Information received through agencies
- iv. Mailed questionnaire method
- v. Schedules sent through enumerators

i. Direct personal investigation: The investigator has to go to the field personally for making enquiries and soliciting information from the respondents. It is used only if the investigation is generally local area.

ii. Indirect oral interviews: This is to be applied when direct personal investigation is not practicable either because of unwillingness or reluctance of the individuals. For example, we want to solicit information on certain social evils like, if a person addicted to drinking, gambling or smoking etc., the person may not respond correctly, in such a case, habits of an individual can best be obtained by interviewing his friends, relatives who know him better.

iii. Mailed questionnaire method: This method consists in preparing questionnaire which is mailed to the respondents with a request for quick response within the specified time. The success of this method based upon is the skill, efficiency, care and the wisdom.

1.8a Sources of Secondary data:

- i. Published sources
- ii. Unpublished sources

i. Published sources: There are a number of national such as CSO, NSSO etc., and international organizations such as UNO, IMF, WHO etc., which collect information regarding business, trade, labour, prices, production, income, health and so on, and publish their findings in statistical reports on a regular basis like weekly, monthly, quarterly and yearly. These publications of the various offices serve as a very powerful source of secondary data. News papers, magazines, internet and periodicals are also come under published sources.

ii. Unpublished sources: These are not openly circulated in the public. They are mentioned as records by various government and private organizations, research institutes, research scholars and so on.

1.9 Other types of data

i. Nominal data: Data representing the presence or absence of attributes in a group of items are termed as nominal data. Here, no importance is given to the units assumed. A nominal scale represents the absence of an attribute by '0' and its presence by '1' where 0 and 1 have no specific meaning. For eg., roll numbers allotted to students of a class.

ii. Ordinal data: Data representing ordering or ranking of units are called ordinal data. Here, importance is given to the units present under study. An ordinal scale arranges the units in either ascending order or in descending order and assigns ranks. For eg., allotment of medical or engineering seats.

3. Interval data: On interval measurement scales, one unit on the scale represents the same magnitude on the trait or characteristic being measured across the whole range of the scale. For example, if anxiety were measured on an interval scale, then a difference between a score of 10

and a score of 11 would represent the same difference in anxiety as would a difference between a score of 50 and a score of 51. Interval scales do not have a "true" zero point, however, and therefore it is not possible to make statements about how many times higher one score is than another.

4. Ratio data: When a scale consists not only of equidistant points but also has a meaningful zero point, then it refers as a *ratio scale*. If we ask respondents their ages, the difference between any two years would always be the same, and 'zero' signifies the absence of age or birth. Hence, a 100-year old person is indeed twice as old as a 50-year old one. Sales figures, quantities purchased and market share are all expressed on a ratio scale. Ratio scales are the most sophisticated of scales, since it incorporates all the characteristics of nominal, ordinal, and interval scales. As a result, a large number of descriptive calculations are applicable.

1.10 Questionnaire: It is a list of questions relating to the field of enquiry and providing space for the answers to be filled by respondents.

Drafting or Framing a good Questionnaire:

- i. The size of the questionnaire should be as small as possible.
- ii. The questions should be simple, clear, brief and unambiguous.
- iii. The questions should be arranged in sequential order .
- iv. The questions should not be lengthy.
- v. Too personal and sensitive questions should be avoided.
- vi. The questions should not be complex

Schedule: It is the device of obtaining answers to the questions in a form which is filled by the enumerators or interviewers in a face to face situation with the respondents.

1.11 Statistical Surveys

1.11a. Censuses Survey or Enumeration: The complete information of each and every unit of the population is called census survey or census enumeration. Here, data are collected from each and every unit of the population. The results obtained are generally not very accurate and reliable.

1.11b. Sample Survey: In a sample survey, only a part of the population is considered. The results obtained from sample can be used to estimate the population value. It is less expensive, less time consuming, requires small number of skilled labours, and the results obtained are generally accurate and reliable.

Variables: A quantitative characteristics which varies from unit to unit is called a variable. For eg., height, weight, price, sales, purchase, volume etc.

Discrete variables: Variables which take only distinct or fixed values are called discrete variables. For eg., number of accidents occurring in a city, no. of patients admitted to a hospital etc.

Continuous variables: Variables which take any numerical value within the specified range are called continuous variables. For eg., height, weight, prices, time etc.

Attribute: A qualitative characteristic which varies from unit to unit is called an attribute. For eg., richness, colour, beauties etc.

1.12. Classification. It is defined as a process of systematic arrangement of data according to common characteristics. Classification of data makes data readable and understandable easily.

1.12a. Types of Classification

There are mainly four types of classification. They are-

i. Qualitative Classification: Classification with respect to qualitative characteristics (i.e., an attribute) is called qualitative Classification. For eg., Classification of population according to economy conditions i.e., Rich, middle and poor class people.

ii. Quantitative Classification: Classification with respect to quantitative characteristics(i.e., a variable) is called qualitative Classification. For eg., Classification regarding sales, purchase, production and prices of commodities.

iii. Temporal Classification: Classification with respect to time factor is called temporal Classification. Time may be in hours, minutes, seconds, years etc.

iv. Spatial Classification: Classification with respect to geographical area or location is called spatial classification.

Further we divide the classification in to following ways. They are-

Dichotomy or dichotomous classification: The process of dividing the data into two classes (or categories) with respect to an attribute is said to be dichotomous Classification. For eg., Population is divided into sex wise as male and female.

Manifold Classification: The process of dividing the data into more than two categories with respect to an attribute is said to be manifold classification. For eg., for the attribute intelligence, the various classes may be, say genius, intelligent, average intelligent, below average, dull etc.

1.13. Tabulation: It is the process of systematic arrangement of classified data in to rows and columns or in a tabular form.

Parts of a Good Statistical table: In general, a good statistical table should contain the following parts. They are-

- i. Table Number
- ii. Title
- iii. Captions
- iv. Stubs
- v. Body of the table
- vi. Head note
- vi. Footnote

Table Number: Each and every table should be given a number. There is no specific place allotted for this number. It can be given at the centre, on top or bottom of the table or even towards the top left hand side.

Title: Every table must have a suitable title. The title must describe briefly, the contents of the table.

Captions: Captions refer to column readings. This explains what the column represents, in the table. There can be one or more columns, depending on the data. The caption headings are written in smaller letters when compared to the title. This mainly helps to save space.

Stubs: Stubs refer to row readings. These are written to the left extreme of the table. They explain what the row represents. Usually in a table, we find more rows than columns.

Body of the Table: The body of the table gives numerical information. This is the most important part of the table. It contains information represented by the captions and stubs.

Head note: Head note gives a brief explanation of the information in the table. This is placed below the title and enclosed within brackets. The units of measurement is written in head note, like (in million tones), (in '000s of rupees'), (in '00 of kgs), etc.

Foot note: Foot note is written below the body of the table. Sometimes complete explanation may be lacking in some parts of the table. The same can be provided in the footnote. Also, if there is any clarification needed in any part of the table it can be done in the footnote.

1.14 Frequency and Frequency distribution

Frequency: The number times an item or a value of a variable is repeated is called a frequency. For eg. A student scored 85 marks in four subjects out of six subjects in an examination, here, 85 is the value of a variable(Marks) and 4 is the frequency.

Frequency distribution: The systematic allocation of frequencies along with their variable values is called frequency distribution.

1.14a.Types of Frequency distribution

1.14a.1 Discrete frequency distribution: In a frequency distribution if a variable takes distinct values along with their frequencies, then it is said to be discrete frequency distribution.

Variable :	12	20	35	47	and so on
Frequency :	5	13	9	12

Example: Prepare a discrete frequency distribution for the following data representing height(inches) of 40 persons: 60, 62, 63, 68, 65, 62, 61, 63, 68, 66, 63, 65, 64, 67, 68, 66, 65, 64, 63, 62, 61, 60, 66, 67, 68, 63, 66, 64, 63, 67, 68, 60, 63, 63, 64, 66, 67, 62, 61, 65.

Height in inches	Tally mark	Frequency (No. of persons)
60		3
61		3
62		4
63		8
64		4
65		4
66		5
67		4
68		5
		Total = 40

1.14a.2 Grouped frequency distribution: The distribution of frequencies along with their classes, where classes are derived by dividing the entire range of values of the variable in to a suitable number of groups is called grouped frequency distribution. For eg.,-

Class : 10 - 19 20 - 29 30 - 39 and so on
 Frequency : 5 13 9

Example: Prepare a grouped frequency distribution for the following data representing marks in Mathematics of 30 students: 50, 92, 63, 88, 65, 62, 61, 63, 68, 66, 63, 95, 64, 67, 68, 96, 86, 64, 63, 77, 68, 60, 73, 63, 64, 66, 67, 72, 61, 90.

Solution: Maximum value = 96; Min.=50; Range = Max – Min = 96-50=46, let class width =10, then No. of class \approx Range/Class width = 46/10 = 4.6 ~5 class.

Height in inches	Tally mark	Frequency(No. of persons)
50 - 60		1
60 - 70		20
70 - 80		3
80 - 90		2
90 - 100		4
	Total=	30

1.14a.3 Continuous Grouped frequency distribution: The presentation of data into continuous classes along with the corresponding frequencies is known as continuous frequency distribution.

Class : 10 - 20 20 - 30 30 - 40 and so on
 Frequency : 5 13 9

Class interval: Dividing the entire range of values of the variable in to a suitable number of groups is called classes. A class consists of all number between the lower and the upper class limits. Hence it it's an interval.

Class limits: The end points of classes are known as class limits-the lower end point of the class being the lower class limit and the upper end point being the upper class limit. For eg.,in a class(10 – 20), **lower limit =10 and upper limit= 20.**

Width of the class interval: The difference between lower and upper class limits is called width or magnitude of the class interval.

For eg., in the class (10 – 20) class width is equal to 10 i.e,(20 - 10=10).

In the class (10-14)class width is equal to 5, here both 10 and 14 are included.

Class frequency: The number of observations corresponding to a particular class is called the frequency of the class.

Class mark (or midpoint of the class): Class mark is the midpoint of a class and is given by

$$\text{Class mark} = (\text{upper limit} + \text{lower limit}) / 2$$

Inclusive Class: If both the upper and lower limits of a class are included, then such a class interval is known as inclusive class.

For eg., Class: 10 – 19, 20 – 29, 30 –39, ...; and 0 - 4, 5 – 9, 10-14 etc., are the inclusive classes. Inclusive class are of discontinuous in nature.

Exclusive class : If only the lower limit of a class is included whereas the upper limit is excluded in that class, then such a class is called exclusive class. **For eg.,** Class : 10 – 20, 20 – 30, 30 – 40, and so on are exclusive classes. Here, 20 is considered in 2nd class and 30 is considered in 3rd and so on. Exclusive classes are of continuous type.

Note: If the values are in decimals or fraction, better consider exclusive class instead inclusive class, because we may not be able include some of those values in one or the other inclusive class.

Open end class: If one of the class limits is missing then such a class is known as open end class. For eg., class: 20 and above, below 10 etc., are open end classes.

1.14b Basic Principles for framing a grouped frequency distribution:

1. Define range(**R**) = **H - L** , where H->Highest or maximum value, and L ->lowest or smallest value in the given set of values of a data.
2. Generally, define a class width (**C.W**) as a multiple of five(5)each.
3. Define number of classes. i.e. **size of the class intervals(i or h) = R/ (1+ 3.322log₁₀N** or approximately, **No. of class = R/C.W.**
4. As far as possible prefer exclusive classes instead inclusive class. Inclusive class can be preferred when given values are of integers.
5. As far as possible minimum 5 class intervals and maximum of 20 classes to be framed.
6. Class intervals should be so fixed that each class has a convenient mid point.

7. As far as possible classes are of uniform in size.

1.15 Graphical Representation of Data

Objective of studying graphical representation of data is to interpret the data in terms of graphs, so that even common man able to understand the changes taken place. When the variables are of continuous type or for a continuous frequency distribution suppose to be depicted, then we use various graphs to represent them. Graphs are easy to understand and they are attractive too.

Types of Graphs

- i. Histogram
- ii. Frequency polygon
- iii. Frequency curves
- iv. Ogive curves (cumulative frequency curves)

i. Histogram: Histogram is drawn in such a way that the set of rectangular bars are placed in adjacent to each other, so that the total area is directly proportional to the height of the rectangular bars. Height of rectangular bars represents the frequency and breadth represents the width or magnitude of the class. The skeleton of histogram can be drawn as under.

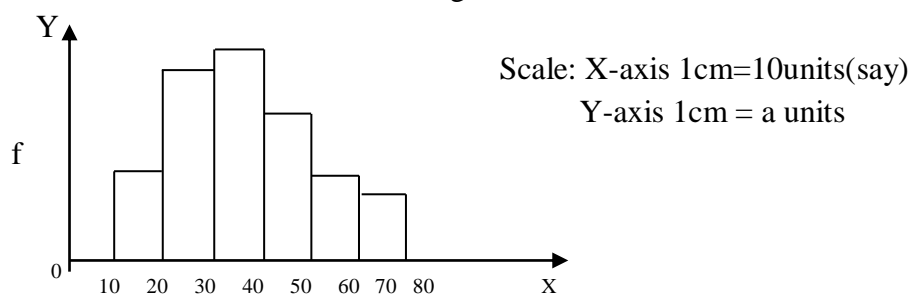


Fig. 6. Histogram

Construction rules of Histogram

- i. If all the given classes having equal or uniform widths then draw rectangular bars vertically and placed them adjacent to each other
- ii. If the given classes having unequal class widths, then compute frequency density, defined by

$$\text{Frequency density} = \frac{\text{frequency of the class}}{\text{Width of the corresponding class}}$$

Then, draw rectangular bars vertically and placed them adjacent to each other to get the histogram.

Important remark. Mode can be computed through Histogram. As in the following mark A, B, C and D on Histogram, by considering tallest rectangular bar and the just neighbouring bars. Then join A & C and join B & D by straight lines as in the diagram. Straight lines intersect at the point O, and then draw OZ, a perpendicular to X axis, and the point at Z on X-axis gives the Mode value of the given distribution.

Example. Draw histogram and find the mode from the following

Age in years: 10-20 20-30 30-40 40-50 50-60 60-70

No. of persons: 3 10 16 12 5 2

Solution. Histogram is drawn below for the above data

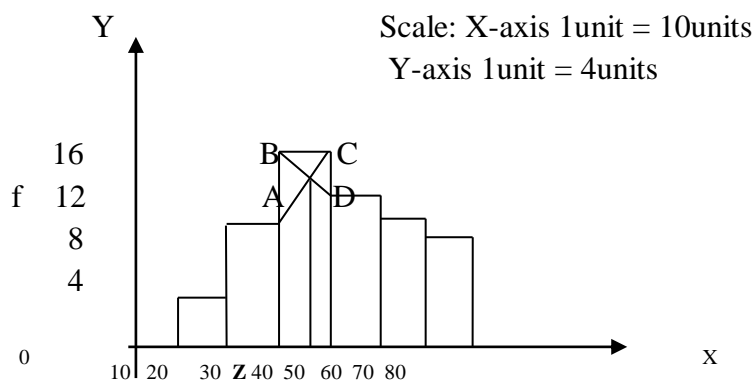


Fig. 7. Histogram in which mode is located

From the above histogram, value of Mode (Z) = 36 years.

ii. Frequency polygon

Here, as a first step, the class frequencies are to be marked against the class mid points on X-Y plane. Then frequency polygon is to be obtained by joining the class frequencies using straight lines. This can also be obtained by joining the mid points on the upper side of the rectangular bars of histogram. To complete the curve beginning and the end mid points which are joined by dotted lines.

iii. Frequency curve

Here, as a first step, the class frequencies are to be marked against the class mid points on X-Y plane. Then frequency curve is to be obtained by joining the class frequencies using smooth line or by free hand curve. This can also be obtained by joining the mid points on the upper side of the rectangular bars of histogram by smooth line. To complete the curve beginning and the end mid points can be joined by dotted lines.

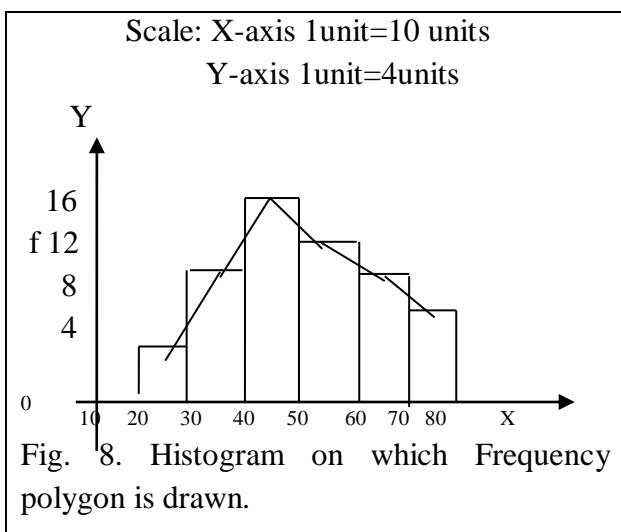


Fig. 8. Histogram on which Frequency polygon is drawn.

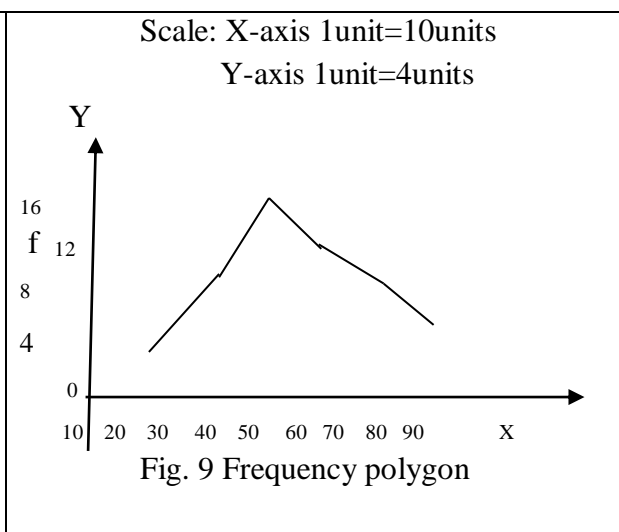


Fig. 9 Frequency polygon

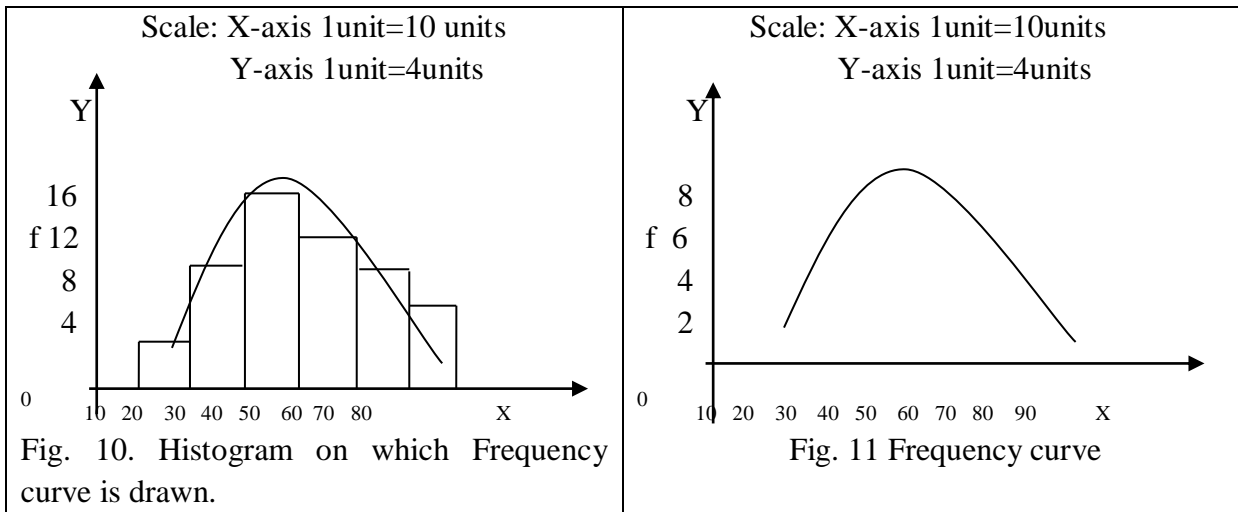


Fig. 10. Histogram on which Frequency curve is drawn.

Fig. 11 Frequency curve

iv. Ogive(Cumulative frequency Curve)

Ogive curve is also known as cumulative frequency curve. There, are of two types, one less than ogive, which is to be drawn by joining the less than cumulative frequencies against the upper limits of the class interval using smooth line; and secondly, more than ogive curves, which is to be drawn by joining more than cumulative frequencies against the lower limits of the class interval using a smooth line. The intersection of these two curves will yield ‘median’, value of the given distribution. That is, draw a perpendicular from the intersection point say O, to the X-axis, which cuts or meets the X-axis at the point M, and it is the required median value.

Example. Draw less than and more than ogive and hence find median from it.

Age in years:	10-20	20-30	30-40	40-50	50-60	60-70
No. of persons:	3	10	16	12	5	2

Solution. Consider,

Age in years:	10-20	20-30	30-40	40-50	50-60	60-70
No. of persons:	3	10	16	12	5	2
Less than cf:	3	13	29	41	46	48
upper limit:	20	30	40	50	60	70
More than cf:	48	45	35	19	7	2
Lower limit:	10	20	30	40	50	60

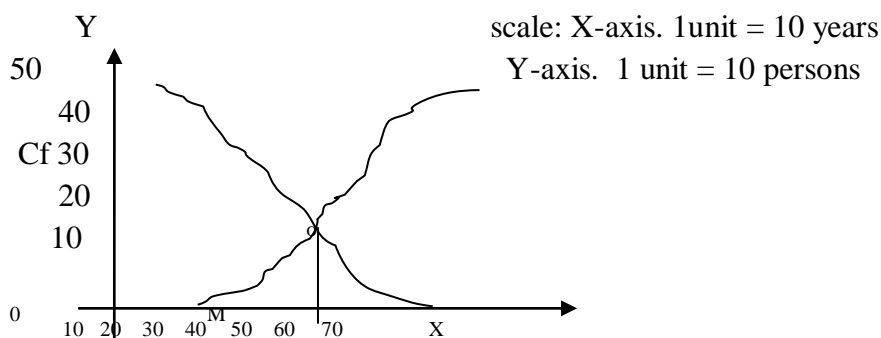


Fig. 12. Less than and more than ogive curves

Objective questions

1. Mode can be obtained by the graph called
a. ogive curves b. Histogram c. Simple bar d. Frequency curves
2. Median can be obtained from
a. ogive curves b. Histogram c. Simple bar d. Frequency curves
3. Class in which only lower limit is considered is called
a. inclusive class b. exclusive class d. open end class d. All the above
4. Class in which both upper and lower limit are considered is called
b. inclusive class b. exclusive class d. open end class d. All the above
5. Qualitative characteristic is called
a. variable b. constant c. attribute d. Average

Exercise

1. Form a discrete frequency distribution from the following
Weight(kgs.) of new born babies:2.0, 2.0, 2.0, 2.5, 3.0, 3.3, 3.5, 3.1, 2.2, 2.0, 4.0, 3.8, 2.75, 2.5, 3.0, 3.3, 3.5, 3.1, 2.2, 2.0, 4.0, 3.8, 2.75, 2.0, 2.3, 2.4.
2. Form a grouped frequency distribution from the following
Height in cms: 120, 122, 145, 156, 160, 170, 180, 187, 165, 156, 135, 146, 155, 160, 180, 182, 168, 164, 164, 163, 152, 160, 158, 159, 150, 160, 165, 162, 168, 163, 158, 159, 168, 164, 165, 160, 166, 170, 172, 169.[hint take class 120-130, 130-140, so on].
Hence draw histogram and locate mode.
3. Form a grouped frequency distribution from the following
Weight in kgs: 60, 52, 45, 56, 60, 70, 80, 87, 65, 56, 35, 46, 55, 60, 80, 82, 68, 64, 64, 163, 52, 60, 58, 59, 50, 60, 65, 62, 68, 63, 58, 59, 68, 64, 65, 60, 66, 70, 72, 69, use class 35-39,40-44, 45-49 and so on.

4. Draw Histogram and locate the mode from the following

Marks obtained	10-25	25-40	40-55	55-70	70-85	85-100
No. of students	2	3	10	6	2	3

5. Draw ogive curves and locate the median from the following

Age in years	10-20	20-30	30-40	40-50	50-60	60-70
No. of persons	12	30	54	36	21	8

UNIT 2

MEASURES OF CENTRAL TENDENCY

2.1 Objective

Objective of studying central values is to know the concentration or overall information with regard to mass of a data set. In this unit our aim is to give knowledge about various measures of central tendency.

2.2 Introduction

From the previous chapters it is understood that how to collect, classify, analyse and interpret a given data either through graphical or by diagrammatic representations. All of them give some crude idea about accuracy of the data and thereby it may not be able to draw meaningful and reliable conclusion about the distribution of data. Therefore it is the time to think, and interpret data more rigorously i.e., either mathematically or algebraically so that more sophisticated inference could be drawn. *'Central tendency' is one such mathematical technique which deals with the study of concentration or density of observations lie at the centre part of the given distribution. In other words, central value is a single entity, which gives overall information with regard to mass of the data.* These are generally termed as 'averages'. Thus, averages are the mathematical formulations which are used to characterize given set of data. Sometimes, these averages are also known as location measures as they locate at some specific positions almost.

2.3 Types of Central Measures

In real life there are a number of data sets available which may or may not have sampling fluctuations or they may or may not contain extreme values. In such cases a suitable measure is to be used in order to have meaningful conclusions. With this point of view, there are mainly five different types of measures of central tendency are defined in the literature. Namely,

- | | | |
|--------------------|------------------|-----------|
| i. Arithmetic Mean | ii. Median | iii. Mode |
| iv. Geometric mean | v. Harmonic Mean | |

Here we concentrate only on Arithmetic mean, median and mode.

2.3.1 Characteristics of a good/ideal measure of central tendency

1. It should be rigidly defined.
2. It should be based on all the observations.
3. It should be easy to understand and easy to calculate.
4. It should be used for further mathematical or statistical analysis.
5. It should be least affected by sampling fluctuations.
6. It should be least affected by extreme or abnormal values.

2.4 Arithmetic Mean: *Arithmetic mean or simply, mean is defined as the ratio of sum of the given set of observations say, x_1, x_2, \dots, x_n to the number of observations (n). It is denoted by 'AM or A or \bar{x} '. Symbolically,*

$$\text{Arithmetic Mean}(\bar{x}) = \frac{\text{Sum of the given observations}}{\text{Number of observations}} = \frac{x_1 + x_2 + \dots + x_n}{n}$$

$$\bar{x} = \sum_{i=1}^n x_i / n$$

Note: Above formula for *arithmetic mean or simply mean* is used for '**raw data**'.

For a frequency data, i.e., if x_1, x_2, \dots, x_n are the set of n observations with respective frequencies f_1, f_2, \dots, f_n , then the arithmetic mean of the frequency data is given by

$$\text{Arithmetic Mean}(\bar{x}) = \frac{f_1x_1 + f_2x_2 + \dots + f_nx_n}{f_1 + f_2 + \dots + f_n}$$

$$\bar{x} = \sum_{i=1}^n f_i x_i / N$$

where $N = \sum_{i=1}^n f_i$.

Example 1: Find the mean of the following

Marks in English: 67, 78, 65, 74, 72, 70, 75, 80.

Solution: Let X: Marks in English. Then mean(\bar{x}) marks is given by

$$\bar{x} = \sum_{i=1}^n x_i / n = 581 / 8 = 72.625 \text{ marks}$$

Example 2: Find the mean of the following

Height in inches	60	61	62	63	64	65	66	67	68
No. of Students	5	4	10	12	16	10	7	4	3

Solution: Here, let X: height in inches and f: Number of students. Then, we have

X_i	60	61	62	63	64	65	66	67	68	Total
f_i	5	4	10	12	16	10	7	4	3	71
$f_i x_i$	300	244	620	756	1024	650	462	268	204	4528

The mean height is given by

$$\bar{x} = \sum_{i=1}^n f_i x_i / N, \text{ where } N = \sum_{i=1}^n f_i$$

$$= 4528 / 71 = 63.7746 \approx 64 \text{ inches}$$

Example 3: Find the mean of the following

Height in cms.(X)	130-135	135-140	140-145	145-150	150-155	155-160	160-165	165-170	170-175
No. of Persons (f)	4	3	6	10	17	25	13	10	5

Solution: Here, let X: height in inches and f: Number of students. Then, we have

f_i	4	3	6	10	17	25	13	10	5	93=N
Mid point(x_i)	132.5	137.5	142.5	147.5	152.5	157.5	162.5	167.5	172.5	-

of class										
$f_i x_i$	530	412.5	855	1475	2592.	3937.	2112.	1675	862.5	14452
					5	5	5			.5

The mean height is given by

$$\begin{aligned}\bar{x} &= \frac{\sum_{i=1}^n f_i x_i}{N}, \text{ where } N = \sum_{i=1}^n f_i \\ &= 14452.5/93 \approx 155.4032 \text{ cms}\end{aligned}$$

2.4.1 Properties of Arithmetic Mean

Property 1. *The algebraic sum of the deviations of set of values taken from their mean is zero.*

Symbolically, $\sum_{i=1}^n (x_i - \bar{x}) = 0$, for raw data & $\sum_{i=1}^n f_i (x_i - \bar{x}) = 0$, for frequency data.

Proof: Here we prove it in more general case. i.e., consider a frequency data $X_i / f_i, i = 1, 2, \dots, n$ of a set of n values. Then by definition, we have

$$\text{Arithmetic Mean}(\bar{x}) = \frac{\sum_{i=1}^n f_i x_i}{N} . \quad (1)$$

Consider, $\sum_{i=1}^n f_i (x_i - \bar{x}) = \sum_{i=1}^n f_i x_i - \bar{x} \sum_{i=1}^n f_i$, since \bar{x} is a constant.

$$= N\bar{x} - N\bar{x} = 0, (\because N = \sum_{i=1}^n f_i, \text{ and by using equation (1)})$$

Hence proved.

Note: In the similar way one can prove it for raw data.

Property 2. *The algebraic sum of squared deviations of set of values taken from their mean is least. Symbolically, $\sum_{i=1}^n f_i (x_i - \bar{x})^2 \leq \sum_{i=1}^n f_i (x_i - A)^2$, where A , is a constant.*

Proof: Consider a frequency data $X_i / f_i, i = 1, 2, \dots, n$ of a set of n values. Then, we have

$$\begin{aligned}\sum_{i=1}^n f_i (x_i - \bar{x})^2 &= \sum_{i=1}^n f_i (x_i - A + A - \bar{x})^2, \text{ where, } A \text{ is a constant.} \\ &= \sum_{i=1}^n f_i [(x_i - A) - (\bar{x} - A)]^2 \\ &= \sum_{i=1}^n f_i (x_i - A)^2 + \sum_{i=1}^n f_i (\bar{x} - A)^2 - 2(\bar{x} - A) \sum_{i=1}^n f_i (x_i - A),\end{aligned}$$

Since $(\bar{x} - A)$ is constant, and $N = \sum_{i=1}^n f_i$, we have

$$= \sum_{i=1}^n f_i (x_i - A)^2 + N(\bar{x} - A)^2 - 2(\bar{x} - A) \left(\sum_{i=1}^n f_i x_i - NA \right)$$

$$\begin{aligned}
&= \sum_{i=1}^n f_i(x_i - A)^2 + N(\bar{x} - A)^2 - 2(\bar{x} - A)(N\bar{x} - NA) \\
&= \sum_{i=1}^n f_i(x_i - A)^2 + N(\bar{x} - A)^2 - 2N(\bar{x} - A)^2 \\
&= \sum_{i=1}^n f_i(x_i - A)^2 - N(\bar{x} - A)^2,
\end{aligned}$$

which implies,

$$\sum_{i=1}^n f_i(x_i - \bar{x})^2 \leq \sum_{i=1}^n f_i(x_i - A)^2.$$

Hence proved.

Property 3. Effect of change of origin and change of scale on Arithmetic Mean

Statement: The arithmetic mean is not independent of change of origin and not independent of change of scale. i.e, when $u_i = (x_i - A)/h$, where A , the origin and h , the scale are two positive constants then $\bar{x} = A + h\bar{u}$.

Proof: Consider a frequency data X_i / f_i , $i = 1, 2, \dots, n$ of a set of n values. Then, we have

$$\text{Arithmetic Mean}(\bar{x}) = \frac{\sum_{i=1}^n f_i x_i}{N}. \quad (1)$$

Let u_i be a new variable such that $u_i = (x_i - A)/h$, where A , the origin and h , the scale, both A and h are two positive constants. Then,

$$x_i = A + hu_i \quad (2)$$

On multiplying both sides of equation (2) by f_i , and then taking sum over $i = 1, 2, \dots, n$, we get

$$\sum_{i=1}^n f_i x_i = A \sum_{i=1}^n f_i + h \sum_{i=1}^n f_i u_i$$

Now, dividing throughout by N , we get

$$\sum_{i=1}^n f_i x_i / N = A + h \sum_{i=1}^n f_i u_i / N, \text{ where } N = \sum_{i=1}^n f_i$$

Therefore, by equation(1), Arithmetic mean(\bar{x}) is given by

$$\boxed{\bar{x} = A + h\bar{u}} \quad (3)$$

Where, $\bar{u} = \sum_{i=1}^n f_i u_i / N$. Since, both 'A' and 'h' are present in the eqn. (2), it is concluded that the mean \bar{x} , is not independent of change of origin 'A', and not independent of change of scale 'h'.

Example 4: For the data given in example 3, find the mean using property change of origin and change of scale.

Solution: Let $u_i = (x_i - A)/h$, where A , the origin and h , the scale and these are two positive constants. Here, we assume $A = 152.5$ (middle value in x_i , called assumed mean) and $h = 5$ (class width), then we have

f_i	4	3	6	10	17	25	13	10	5	93=N
Mid point(x_i) of class	132.5	137.5	142.5	147.5	152.5	157.5	162.5	167.5	172.5	-
$u_i = (x_i - 152.5)/5$	-4	-3	-2	-1	0	1	2	3	4	
$f_i u_i$	-16	-9	-12	-10	0	25	26	30	20	54

The mean height is given by

$$\bar{x} = A + h\bar{u},$$

where $\bar{u} = \frac{\sum_{i=1}^n f_i u_i}{N}$

Therefore,

$$\begin{aligned} \bar{x} &= 152.5 + 5 \times (54/93) \\ \Rightarrow \bar{x} &= 155.4032 \text{ cms} \end{aligned}$$

Property 4. Combined Mean (Mean of k -sets of data)

Let there be k sets of random samples of sizes n_i , ($i = 1, 2, \dots, k$), each with respective means \bar{x}_i , $i = 1, 2, \dots, k$. Then the combined mean of k -sets of data is given by

$$\text{Combined Mean}(\bar{x}_c) = \frac{n_1\bar{x}_1 + n_2\bar{x}_2 + \dots + n_k\bar{x}_k}{n_1 + n_2 + \dots + n_k} = \frac{\sum_{i=1}^k n_i\bar{x}_i}{\sum_{i=1}^k n_i}.$$

Proof: Let $(x_{11}, x_{12}, \dots, x_{1n_1}), (x_{21}, x_{22}, \dots, x_{2n_2}), (x_{31}, x_{32}, \dots, x_{3n_3}), \dots, (x_{k1}, x_{k2}, \dots, x_{kn_k})$ be the k -sets of random samples with respective sample sizes n_i , and sample means \bar{x}_i , $i = 1, 2, \dots, k$. Then, we have

$$\bar{x}_1 = \frac{\sum_{i=1}^{n_1} x_{1i}}{n_1} \Rightarrow \sum_{i=1}^{n_1} x_{1i} = n_1\bar{x}_1, \tag{1}$$

Similarly, for 2nd set through k -set, we have

$$\sum_{j=1}^{n_2} x_{2j} = n_2\bar{x}_2, \sum_{r=1}^{n_3} x_{3r} = n_3\bar{x}_3, \dots, \sum_{l=1}^{n_k} x_{kl} = n_k\bar{x}_k \tag{2}$$

Therefore the combined mean of these k -sets of data is given by

$$\text{Combined Mean}(\bar{x}_c) = \frac{\sum_{i=1}^{n_1} x_{1i} + \sum_{j=1}^{n_2} x_{2j} + \sum_{r=1}^{n_3} x_{3r} + \dots + \sum_{l=1}^{n_k} x_{kl}}{n_1 + n_2 + \dots + n_k}$$

Then by using equations (1) and (2), we have

$$\text{Combined Mean}(\bar{x}_c) = \frac{n_1\bar{x}_1 + n_2\bar{x}_2 + \dots + n_k\bar{x}_k}{n_1 + n_2 + \dots + n_k} = \frac{\sum_{i=1}^k n_i\bar{x}_i}{\sum_{i=1}^k n_i}$$

Note. In particular, if $k = 2$, i.e., for two sets of data, combined mean of (n_1+n_2) observations is given by

$$\text{Combined Mean}(\bar{x}_c) = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$$

Example 5: In a class, of 55 students, the mean marks of 25 girls is 68.3 and mean marks of 30 boys is 65.8. Find the mean marks of 55 students.

Solution: Given $n_1 = 25$ girls and $n_2 = 30$ boys; $\bar{x}_1 = 68.3$, the mean marks of girls and $\bar{x}_2 = 65.8$, the mean marks of boys. Therefore the combined mean of $(n_1+n_2) = 55$ students is given by

$$\begin{aligned} \bar{x}_c &= \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2} \\ &= \frac{25 \times 68.3 + 30 \times 65.8}{55} = 66.94 \text{ marks} \end{aligned}$$

2.5 Weighted Arithmetic Mean

In real life, not all observations have the same importance. i.e., each observation has its own relative importance. In such a case simple arithmetic mean over estimates the average. Therefore, in order to have more stable result for the average one could use weighted arithmetic mean.

Thus, if x_1, x_2, \dots, x_n are a set of n observations with respective weights w_1, w_2, \dots, w_n , then weighted arithmetic mean of this data is given by

$$\text{Weighted Arithmetic Mean}(\bar{x}_w) = \frac{w_1x_1 + w_2x_2 + \dots + w_nx_n}{w_1 + w_2 + \dots + w_n}$$

$$\bar{x}_w = \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i}$$

Example 6: Find simple and weighted arithmetic means for the following

x_i	60	65	70	64	66	67	68
w_i	5	4	10	16	7	4	3

Solution: To find simple mean and weighted arithmetic mean, we have

x_i	60	65	70	64	66	67	68	$460 = \sum x_i$
w_i	7	5	1	6	4	3	2	$28 = \sum w_i$
$x_i w_i$	420	325	70	384	264	201	136	$1800 = \sum w_i x_i$

Simple arithmetic mean $\bar{x} = \sum_{i=1}^n x_i / n = 460/7 = 65.71$ units (1)

Weighted Arithmetic Mean is given by

$$\bar{x}_w = \sum_{i=1}^n w_i x_i / \sum_{i=1}^n w_i = 1800/28 = 64.2857 \text{ units} \quad (2)$$

From (1) and (2), it is observed that $(\bar{x} > \bar{x}_w)$, which means, simple mean slightly over estimates the average value.

Example 7: Find simple arithmetic mean and weighted arithmetic mean of first n natural numbers, where weights being the corresponding numbers.

Solution: We know that first n natural numbers are 1, 2, 3, . . . , n ; and since weights are the corresponding numbers, we have

x_i	1	2	3	.	.	.	n
w_i	1	2	3	.	.	.	n

Simple arithmetic mean $\bar{x} = \sum_{i=1}^n x_i / n$

$$= [1 + 2 + 3 + \dots + n] / n$$

$$= [n(n+1)/2] / n$$

$$= (n+1) / 2, \text{ units.}$$

Now, to find weighted Arithmetic Mean (\bar{x}_w), we have

$w_i x_i$	1^2	2^2	3^2	.	.	.	n^2
-----------	-------	-------	-------	---	---	---	-------

Therefore, we have

$$\bar{x}_w = \sum_{i=1}^n w_i x_i / \sum_{i=1}^n w_i$$

$$= [1^2 + 2^2 + \dots + n^2] / (1 + 2 + \dots + n)$$

$$= \frac{n(n+1)(2n+1)/6}{n(n+1)/2}$$

$$= (2n+1) / 3, \text{ units.}$$

Example 7: Find weighted arithmetic mean of first n natural numbers, where weights being the corresponding but opposite numbers.

Solution: We know that first n natural numbers are 1, 2, 3, . . . , n ; and since weights are the corresponding opposite numbers, we have

x_i	1	2	3	.	.	$n-1$	N
w_i	n	$n-1$	$n-2$.	.	2	1

Consider,

$w_i x_i$	n.1	2.(n-1)	3(n-2)	.	.	2.(n-1)	n.1
-----------	-----	---------	--------	---	---	---------	-----

Therefore, $\sum_{i=1}^n w_i x_i = n.1 + 2(n-1) + 3(n-2) + \dots + 2(n-1) + n.1$

$$\begin{aligned}
&= \sum_{i=1}^n i(n - [i - 1]) = \sum_{i=1}^n [i(n + 1) - i^2] \\
&= (n + 1) \sum_{i=1}^n i - \sum_{i=1}^n i^2 \\
&= (n + 1)n(n + 1)/2 - n(n + 1)(2n + 1)/6 \\
&= n(n + 1)(n + 2)/6
\end{aligned}$$

Now, to find weighted Arithmetic Mean (\bar{x}_w), we have

$$\begin{aligned}
\bar{x}_w &= \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i} \\
&= \frac{[n(n + 1)(n + 2)/6]}{(1 + 2 + \dots + n)} \\
&= \frac{n(n + 1)(n + 2)/6}{n(n + 1)/2} \\
&= (n + 2)/3, \text{ units.}
\end{aligned}$$

2.6 Merits and demerits of Arithmetic Mean

Arithmetic mean has few merits (advantages) and demerits (disadvantages). They are given in the form of table.

Merits	Demerits
1. It is based on all the observations.	1. It cannot be calculated even if one observation is missing.
2. It is easy to understand and easy to calculate.	2. It cannot be calculated for frequency distributions with 'open end class' at the tails, for eg., less than 20, more than 80, etc
3. It is rigidly defined.	3. It cannot be used to analyse qualitative characteristics such as honesty, beauty, etc.
4. It can be used for further algebraic treatment.	4. It is highly sensitive to extreme values.
5. It is least affected by sampling fluctuations.	5. It cannot be calculated graphically.

2.7 Median (M or Md)

Definition: Median is the value of a variable which divides the entire distribution into two equal parts. In other words, median is the value which exceeds 50% and exceeded by 50% of the given set of values, such that it lies exactly at the centre part of the given distribution. In simple, it is the middle most value in the given distribution. It is denoted by M or Md.

2.7.1 Merits and demerits of Median

Median has few merits (advantages) and demerits (disadvantages). They are given in the form of table.

Merits	Demerits
i. It is rigidly defined.	a. Since it is not based on all the values, result may not be reliable and thus sometimes it is called insensitive.
ii. It can be calculated even if one observation is missing.	b. It cannot be used for further algebraic treatment.
iii. It can be calculated for frequency distributions with open end class.	c. It is difficult to calculate as it requires ordered data.
iv. It is not at all sensitive to extreme values	d. It is highly affected by fluctuations of values.
v. It can be calculated for both quantitative and qualitative data.	
vi. It can be obtained graphically.	
vii. In some cases it can be located merely by inspection	

2.7.2 Computation of Median

Case (1): Raw Data

As a first step, arrange the data (i.e., array) either in ascending and descending order of magnitude. Then,

- a. Median is the *Middle value*, if there are '*odd number*' of values in the data
- b. Median is the mean of two middle values, if '*even number*' of values are present in the data.

Or, we can use the formula,

$$M = \left(\frac{n+1}{2} \right)^{nd} \text{ term in the array}$$

where n , the number of observations in the data.

Example 22: Find the median of 10, 22, 15, 16, 18

Solution: Array: 10, 15, 16, 18, 22

Since $n=5$ is odd, we have,

$$\text{Median}(M) = \text{middle value in array} = 3^{\text{rd}} \text{ term} = 16 \text{ units}$$

Or,

$$M = \left(\frac{n+1}{2} \right)^{nd} \text{ term in the array}$$

$$= \left(\frac{5+1}{2} \right)^{nd} = \frac{6}{2} = 3^{\text{rd}} \text{ term in the array}$$

Implies, Median = 16 units.

Example 23: Find the median of 10, 22, 15, 16, 18, 25, 46, 22

Solution: Array: 10, 15, 16, 18, 22, 22, 25, 46

Since $n = 8$ is even, we have,

Median(M) = Mean(two middle values) in array = (18+22)/2 =20 units

Or,
$$M = \left(\frac{n+1}{2}\right)^{nd} \text{ term in the array}$$

$$= \left(\frac{8+1}{2}\right)^{nd} = \frac{9}{2} = 4.5^{th} \text{ term in the array}$$

Implies,

$$\text{Median}(M) = \left(\frac{4^{th} \text{ term} + 5^{th} \text{ term}}{2}\right) \text{ in array} = \frac{18+22}{2} = \frac{40}{2} = 20 \text{ units}$$

Case 2: Median for discrete frequency data

Median for discrete frequency data $X_i / f_i, i=1,2,\dots,n$; can be obtained by the following steps.

Step 1. Find the cumulative frequencies(**CFs**) for the discrete frequency data.

Step 2. Find $\left(\frac{N+1}{2}\right)$, where $N = \sum_{i=1}^n f_i$.

Step 3. Find a $CF \geq \left(\frac{N+1}{2}\right)$, i.e., a CF which is just more than or equal to $\left(\frac{N+1}{2}\right)$.

Step 4. Thus median(M) is a value of the variable X (say), which corresponds to **CF**, obtained in 3rd step.

Example 23: Find the median of the following

x_i :	7	10	15	18	20	22
f_i :	5	7	10	12	8	4

Solution: Given discrete frequency data

x_i :	7	10	15	18	20	22
f_i :	5	7	10	12	8	4
CFs:	5	12	22	34	42	46

Since, $N = \sum_{i=1}^n f_i = 46$, we have

$$\left(\frac{N+1}{2}\right) = \frac{46+1}{2} = \frac{47}{2} = 23.5$$

(23.5)th term lies in the CF 34.(i.e., as $CF = 34 > 23.5$).

=> Median(M) = 18 units.

Case 3. Median for grouped frequency data

Consider the grouped frequency data $X_i - X_j / f_i, i=1,2,\dots,n$; then, median can be obtained by the following steps.

Step 1. Find the cumulative frequencies(**CFs**) to the given data.

Step 2. Find $\left(\frac{N}{2}\right)$, where $N = \sum_{i=1}^n f_i$.

Step 3. Find a $CF \geq \left(\frac{N}{2}\right)$, i.e., a CF which is just more than or equal to $\left(\frac{N}{2}\right)$.

Step 4. Find a class interval, which corresponds to CF , obtained in 3rd step, called median(M) class.

Step 5. Once getting the median class, use the following formula to compute median value. i.e.,

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

Where, L , denote the lower limit of median class

C , the cumulative frequency of class, just preceding(previous) to median class

h , the width of median class

f , frequency of the median class

Note: The above median formula holds good only for the grouped frequency distribution with continuous (i.e., exclusive type) class intervals. If the classes are of inclusive type, then they must be get converted to inclusive type before applying the median above formula. This will influence on the value of ' L ', in the above formula.

Example 24: Find the median of the following

Weight in kgs.	10-20	20-30	30-40	40-50	50-60	60-70	70-80
No. of Persons	3	5	10	18	9	7	2

Solution: To compute median, we have

Weight in kgs.	10-20	20-30	30-40	40-50	50-60	60-70	70-80	Total
No. of Persons f_i	3	5	10	18	9	7	2	54= N
Cumulative Frequency(CFs)	3	8	18	36	45	52	54	

Now, $N/2=54/2=27$, \Rightarrow 27th term lies in the CF 36.(i.e. $CF=36 > 27=N/2$).

\Rightarrow Median class= 40-50.

Here, $L=40$, $h=10$, $C=18$, $f=18$. Therefore, the Median(M) weight is given by

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

$$= 40 + (27 - 18) \times \frac{10}{18} = 45kgs$$

Example 24: Find the missing frequency if the median of the distribution is 3.76 units

Family size	1-3	3-5	5-7	7-9	9-11
No. of Persons	7	8	?	2	1

Solution: Let the missing frequency be ' y '

Family size	1-3	3-5	5-7	7-9	9-11
-------------	-----	-----	-----	-----	------

No. of Persons	7	8	y	2	1
Less than cumulative frequency	7	15	15+ y	17+ y	18+ y

Given, median(M) = 3.76 units

=> Median class= 3-5. =>L=3, h=2, f =8, C = 7.

The median is given by

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

$$3.76 = 3 + [(18 + y)/2 - 7] \times \frac{2}{8}$$

$$3.76 - 3 = [(4 + y)] \times \frac{1}{8}$$

$$\Rightarrow 4 + y = 6.08 \Rightarrow y = 2.08 \approx 2$$

=> missing frequency(y) = 2, since frequency is a positive integer.

Example 24: The following table gives the information on price of groceries in a supermarket.

Price of groceries	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89	90-99
No. of variety of groceries	7	4	10	28	29	17	10	5	6

Solution: First convert the inclusive classes into exclusive classes before computing median, then we have

Price of groceries	9.5-19.5	19.5-29.5	29.5-39.5	39.5-49.5	49.5-59.5	59.5-69.5	69.5-79.5	79.5-89.5	89.5-99.5
No. of variety of groceries(f)	7	4	10	28	29	17	10	5	6
Cum. Frequency	7	11	21	49	78	95	105	110	116

Now, $N/2 = 116/2 = 58$.

=> 58th term lies in the CF 78, => Median class= 49.5-59.5.

Here, $L=49.5$, $h=10$, $C=49$, $f=29$.

Therefore, the median price is given by

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

$$= 49.5 + (58 - 49) \times \frac{10}{29} = 52.6034$$

=> Median(M) price of groceries is ~ Rs.53/-

Example 24: Find the median for the following data related to the observed survival times (in years) of Indians taken from various states

survival times	Below 20	Below 40	Below 60	Below 80	Below 100	Below 120
----------------	----------	----------	----------	----------	-----------	-----------

(in years)						
No. of persons	25	57	92	168	196	200

Solution: Since the given frequencies (No. of persons) are of less than cumulative type, we rewrite the table with exclusive classes before computing median. i.e.,

survival times (in years)	0-20	20-40	40-60	60-80	80-100	100-120
No. of persons(f)	25	32	35	76	28	4
Cumulative frequency	25	57	92	168	196	200

Now, $N/2 = 200/2 = 100$.

=> 100th term lies in the CF 168, => Median class= 60-80.

Here, $L= 60, h = 20, C = 92, f = 76$.

Therefore, the median survival time in years is given by

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

$$= 60 + (100 - 92) \times \frac{20}{76} = 62.1052$$

=> Median(M) survival time of Indians is = 62 years.

Example 24: A survey on was conducted to know the size of television (in inches) set using in households of a locality

Size of television (in inches)	above 10	above 20	above 30	above 40	above 50	above 60	above 70
No. of televisions	193	185	168	117	83	47	13

Solution: Since the given frequencies (No. of televisions) are of more than cumulative type, we rewrite the table with exclusive classes before computing median. i.e.,

Size of television (in inches)	10-20	20-30	30-40	40-50	50-60	60-70	70-80
No. of televisions(f)	8	17	51	34	36	34	13
Cum. frequency	8	25	76	110	146	180	193

Now, $N/2 = 193/2 = 96.5$.

=> 96.5th term lies in the CF 110, => Median class= 40-50.

Here, $L=40, h=10, C=76, f=34$.

Therefore, the median size of television is given by

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

$$= 40 + (96.5 - 76) \times \frac{10}{34} = 45.6944$$

=> Median(M) size of TV using in that locality is ≈ 46 inches.

2.8 Mode (Z or M_o): Mode is the most frequently occurring or most repeated value in the given set of observations. It is usually denoted by either Z or M_o .

2.8.1 Merits and demerits of Mode

Mode has few merits (advantages) and demerits (disadvantages). They are given in the form of table.

Merits	Demerits
<ul style="list-style-type: none"> a. It can be calculated even if one observation is missing. b. It is least affected by extreme values. c. It can be calculated for frequency distributions with open end class. d. It can be calculated for both quantitative and qualitative data. e. It can be obtained graphically. f. Like median it can be located merely by inspection. g. It is not at all sensitive to extreme values. h. It can be calculated even if a frequency distribution with unequal class width provided modal class, and its preceding and succeeding classes have the same class width. 	<ul style="list-style-type: none"> a. Since it is not based on all the values, result may not be stable. b. It cannot be used for further algebraic treatment. c. It is difficult to calculate as compared to mean d. It is very much affected by fluctuations of sampling. e. Mode can be ill defined. i.e., it is not always possible to find a clearly defined mode. If a distribution has two modes then it is said to be bimodal and if a distribution has more than two modes then it is said to be multimodal.

2.8.2 Computation of Mode

Case (1): Raw Data

Mode is the most repeated value in the given set of observations.

Example 25: Find the mode of 6 8 10 12 8 9 8.

Solution: Here, most repeated value is 8, => Mode(Z) = 8 units.

Example 26: Find the mode of the following

Marks in Statistics: 60 80 70 65 70 82 83 72 55 72 70 73 72.

Solution: Here, both 70 and 72 are repeated 3 (most number of) times each.

Therefore,

$$\text{Mode}(Z) = 70 \text{ or } 72$$

=> given distribution is a ‘bimodal’ distribution as it has two modes.

Case 2: For discrete frequency data

Consider a discrete frequency data $X_i / f_i, i = 1, 2, \dots, n$; then mode can be obtained using the following steps.

- i. Find the Maximum(largest/highest) frequency
- ii. The value of the variable X(say), which corresponds to largest frequency is the required Mode(Z) of that distribution, provided the largest frequency is not at the extremes.

Note 1. In a frequency distribution, if the *neighbouring frequencies are very close(usually a difference of 1 or 2)* to the highest frequency, then *above method fails* to give Mode. In this case, we use the method of ‘*grouping to evaluate Mode*’.

Note 2. If the highest frequency is at the extremes of a frequency distribution, then mode is ill defined. In this case, graphical method also will fail to give Mode, and we need to use different approach, and it will be discussed later.

Example 27: Find the mode from the following data

x_i :	10	15	20	25	30	35	40
f_i :	3	7	8	12	7	6	4

Solution: In the given data, Highest frequency(f) =12, => Mode(Z) =25 units.

Case 3. Mode for Grouped frequency data

Consider the continuous grouped frequency data $LL-UL/f_i, i = 1,2,\dots,n$; then mode can be obtained using the following steps.

- i. Find the Maximum(largest/highest) frequency
- ii. The class interval, which corresponds to largest frequency is the required ‘modal class’ of that distribution.
- iii. The mode is then given by

$$Z = L + \left[\frac{(f_1 - f_0) \times h}{2f_1 - f_0 - f_2} \right]$$

Where, L , denote the lower limit of Modal class

f_1 , the highest frequency in a frequency distribution

f_0 , frequency of the class, just preceding to modal class.

f_2 , frequency of the class just succeeding the modal class

h , the width of modal class

This is the required formula of Mode for continuous grouped frequency distribution

Example 24: Find the mode of the following

Weight in kgs.	10-20	20-30	30-40	40-50	50-60	60-70	70-80
No. of Persons	3	5	10	18	9	7	2

Solution: To compute mode, we have

Weight in kgs.	40-45	45-50	50-55	55-60	60-65	65-70	70-75	Total
No. of Persons	3	5	10	18	9	7	2	54=N
f_i								

Here, highest frequency $f_1=18$, \Rightarrow Modal class= 55-60.

Here, $L=55$, $h=5$, $f_1=18$, $f_0=10$, $f_2=9$.

The mode is given by

$$Z = L + \left[\frac{(f_1 - f_0) \times h}{2f_1 - f_0 - f_2} \right]$$

$$= 55 + \left[\frac{(18 - 10) \times 5}{2(18) - 9 - 10} \right] = 55 + 2.3529 = 57.3529$$

Implies, Modal weight $Z \cong 57$ kgs

Example 24: Find the missing frequency if the mode of the distribution is 3.286 units

Family size	1-3	3-5	5-7	7-9	9-11
No. of Persons	7	8	?	2	1

Solution: Let the missing frequency be 'y' and given, Mode(Z) = 3.286 units

\Rightarrow Modal class= 3-5.

Therefore, we have $L=3$, $h=2$, $f_1=8$, $f_0=7$, $f_2=y$ (say).

The mode is given by

$$Z = L + \left[\frac{(f_1 - f_0) \times h}{2f_1 - f_0 - f_2} \right]$$

$$3.286 = 3 + \left[\frac{(8 - 7) \times 2}{2(8) - 7 - y} \right]$$

$$3.286 - 3 = \left[\frac{2}{9 - y} \right]$$

$$\Rightarrow y = 9 - 6.99 = 2.01$$

\Rightarrow missing frequency(y) = 2, as frequency can not be a decimal number.

Example 24: The following table gives the information on the marks obtained in Statistics of 225 students

Marks in statistics	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89	90-99
No. of Students	3	5	10	28	59	67	32	15	6

Solution: First convert the inclusive classes into exclusive classes before computing mode. i.e.,

Marks in statistics	9.5-19.5	19.5-29.5	29.5-39.5	39.5-49.5	49.5-59.5	59.5-69.5	69.5-79.5	79.5-89.5	89.5-99.5
No. of Students	3	5	10	28	59	67	32	15	6

Here, highest frequency $f_1=67$, \Rightarrow Modal class= 59.5-69.5.

Here, $L=59.5$, $h=10$ $f_1=67$, $f_0=59$, $f_2=32$.

The mode is given by

$$\begin{aligned}
 Z &= L + \left[\frac{(f_1 - f_0) \times h}{2f_1 - f_0 - f_2} \right] \\
 &= 59.5 + \left[\frac{(67 - 59) \times 10}{2(67) - 59 - 32} \right] \\
 &= 59.5 + \left[\frac{80}{43} \right] = 61.36 \text{ marks}
 \end{aligned}$$

Example 24: Find the mode for the following data related to the observed lifetimes (in hours) of electrical components

lifetimes (in hours)	Below 20	Below 40	Below 60	Below 80	Below 100	Below 120
No. of electrical components	10	45	97	158	196	225

Solution: Since the given frequencies (No. of electrical components) are of less than cumulative type, we rewrite the table with exclusive classes before computing mode. i.e.,

Marks in statistics	0-20	20-40	40-60	60-80	80-100	100-120
No. of Students	10	35	52	61	38	29

Here, highest frequency $f_1=61$, \Rightarrow Modal class= 60-80

Here, $L=60$, $h=20$ $f_1=61$, $f_0=52$, $f_2=38$.

The mode(Z) is given by

$$\begin{aligned}
 Z &= L + \left[\frac{(f_1 - f_0) \times h}{2f_1 - f_0 - f_2} \right] \\
 &= 60 + \left[\frac{(61 - 52) \times 20}{2(61) - 52 - 38} \right]
 \end{aligned}$$

On simplification, we have mode(Z) = 65.625hours.

Example 24: The following table gives the information about monthly salary of 200 engineers of a software company

Wages in('000')	above 10	above 20	above 30	above 40	above 50	above 60	above 70	above 80	above 90
No. of engineers	200	185	168	147	133	87	54	26	5

Solution: Since the given frequencies (No. of engineers) are of more than cumulative type, we rewrite the table with exclusive classes before computing mode. i.e.,

Wages	in	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90	90-100
-------	----	-------	-------	-------	-------	-------	-------	-------	-------	--------

('000'Rs)									
No. of engineers	15	17	21	14	46	33	28	21	5

Here, highest frequency $f_1=46$, \Rightarrow Modal class= 50-60

Here, $L=50$, $h=10$ $f_1=46$, $f_0=14$, $f_2=33$.

The mode(Z) is given by

$$Z = L + \left[\frac{(f_1 - f_0) \times h}{2f_1 - f_0 - f_2} \right]$$

$$= 50 + \left[\frac{(46 - 14) \times 10}{2(46) - 14 - 33} \right]$$

On simplification, we have mode(Z) = 57.1111('000' Rs.) = Rs. 57111.1/-

Example 24: Find the mode for the following

Wages in ('00'Rs)	10-20	20-23	23-40	40-55	55-60	60-65	65-80	80-90	90 & above
No. of workers	5	17	23	24	43	31	28	21	5

Solution: Here, highest frequency $f_1=43$, \Rightarrow Modal class= 55-60

Here, $L=55$, $h=5$ $f_1=43$, $f_0=24$, $f_2=31$.

The mode(Z) is given by

$$Z = L + \left[\frac{(f_1 - f_0) \times h}{2f_1 - f_0 - f_2} \right]$$

$$= 55 + \left[\frac{(43 - 24) \times 5}{2(43) - 24 - 31} \right]$$

On simplification, we have mode(Z) = 58.0645('00'Rs.) \approx Rs.5806/-.

2.9 Partition values

Partition values are of special type of averages, they divide the entire distribution into some fixed number of equal parts. Viz., four or ten, or hundred equal parts. These are also known as 'location values' as they lie at some specific position in the given distribution.

There are mainly three types of partition values. They are

- i. Quartiles(Qr, r=1,2,3.)
- ii. Deciles(Dr, r=1,2,3,...,9)
- iii. Percentiles(Pr, r=1,2,3,...,99)

2.9.1 Quartiles($Q_r, r = 1,2,3$): There are three quartiles, divide the entire distribution into four equal parts. Quartiles are usually denoted by Q_r , $r=1, 2$, and 3 . When $r = 1$, i.e., Q_1 , is called the 'lower' or 'first' quartile. Q_1 is a value which exceeds 25% and exceeded by 75% of the given set observations. In other words, Q_1 is a value, which is more than 25% of the given observations and less than the remaining 75% of observations given. When $r = 2$, i.e., Q_2 , is called the 'second' quartile. Q_2 is the value which exceeds 50% and exceeded by 50% of the given set

observations, i.e., Q_2 is the value, which is more than 50% of the given observations and less than the remaining 50% of observations given. Q_2 is also known as '*median quartile*' as it lies exactly at centre part of given distribution. When $r = 3$, i.e., Q_3 , is called the 'upper' or 'third' quartile. Q_3 is a value which exceeds 75% and exceeded by 25% of the given set observations. That is, Q_3 is a value, which is more than 75% of the given observations and less than the remaining 25% of observations given.

2.9.2. Deciles ($D_r, r = 1, 2, 3, \dots, 9$): There are nine deciles, divide the entire distribution into ten equal parts. Deciles are usually denoted by $(D_r, r = 1, 2, 3, \dots, 9)$. When $r = 1$, i.e., D_1 , is called the 'lower' or 'first' Decile. D_1 is a value which exceeds 10% and exceeded by 90% of the given set observations. In other words, D_1 is a value, which is more than 10% of the given observations and less than the remaining 90% of observations given. When $r = 2$, i.e., D_2 , is called the 'second' decile. D_2 is the value which exceeds 20% and exceeded by 80% of the given set observations, i.e., D_2 is the value, which is more than 20% of the given observations and less than the remaining 80% of observations given. D_5 is the 5th decile and also known as '*median decile*' as it lies exactly at centre part of given distribution. D_5 is the value which exceeds 50% and exceeded by 50% of the given set observations. Similarly, when $r = 9$, i.e., D_9 , is called the 9th or 'upper' decile. D_9 is a value which exceeds 90% and exceeded by 10% of the given set observations. That is, D_9 is a value, which is more than 90% of the given observations and less than the remaining 10% of observations given.

2.9.3. Percentiles ($P_r, r = 1, 2, 3, \dots, 99$): There are ninety-nine deciles, divide the entire distribution into hundred equal parts. Percentiles are usually denoted by $P_r, r = 1, 2, 3, \dots, 99$. When $r = 1$, i.e., P_1 , is called the 'lower' or 'first' percentile. P_1 is a value which exceeds 1% and exceeded by 99% of the given set observations. In other words, P_1 is a value, which is more than 1% of the given observations and less than the remaining 99% of observations given. When $r = 2$, i.e., P_2 , is called the 'second' percentile. P_2 is the value which exceeds 2% and exceeded by 98% of the given set observations, i.e., P_2 is the value, which is more than 2% of the given observations and less than the remaining 98% of observations given and so on. When $r = 50$, i.e., P_{50} , is called the 50th percentile and also known as '*median percentile*' as it lies exactly at centre part of given distribution. P_{50} is the value which exceeds 50% and exceeded by 50% of the given set observations. Similarly, when $r = 99$, i.e., P_{99} , is called the 99th percentile. P_{99} is a value which exceeds 99% and exceeded by 1% of the given set observations. That is, P_{99} is a value, which is more than 99% of the given observations and less than the remaining 1% of observations given.

2.9.4 Computation of Partition values

Computation of partition values is similar to that of median. Thus we have,

Case (1): Raw Data

As a first step, arrange the data (i.e., array) either in ascending and descending order of magnitude. Then, use the formula,

$$Q_r = r \times \left(\frac{n+1}{4} \right)^{\text{th}} \text{ term in the array for Quartiles, } r=1, 2, 3.$$

$$D_r = r \times \left(\frac{n+1}{10} \right)^{\text{th}} \text{ term in the array for Deciles, } r= 1, 2, \dots, 9.$$

$$P_r = r \times \left(\frac{n+1}{100} \right)^{\text{th}} \text{ term in the array for Deciles, } r = 1, 2, 3, \dots, 99.$$

where n , the number of observations in the data.

Example 22: Find quartiles, D_2 , and P_{98} , from the following.

Weight of infant in kgs. 4.0, 2.2, 3.5, 2.6, 5.8, 4.5, 5.2, 6.5, 3.6, 4.8

Solution: Array: Weight of infant in kgs. 2.2, 2.6, 3.5, 3.6, 4.0, 4.5, 4.8, 5.2, 5.8, 6.5

To find quartiles, we have

$$Q_r = r \times \left(\frac{n+1}{4} \right)^{\text{th}} \text{ term in the array for Quartiles, } r=1, 2, 3.$$

Now, when $r=1$, i.e.,

$$\begin{aligned} Q_1 &= 1 \times \left(\frac{10+1}{4} \right)^{\text{th}} \text{ term} = (11/4) = (2.75)^{\text{th}} \text{ term} \\ &= 2\text{nd term} + 0.75(3\text{rd term} - 2\text{nd term}) \text{ in array} \\ \Rightarrow Q_1 &= 2.6 + 0.75(3.5 - 2.6) = 2.6 + 0.75(0.9) = 2.6 + 0.675 = 3.275 \text{ kgs} \end{aligned}$$

Now, when $r=2$, i.e.,

$$\begin{aligned} Q_2 &= 2 \times \left(\frac{10+1}{4} \right)^{\text{th}} \text{ term} = 2(11/4) = (5.5)^{\text{th}} \text{ term} \\ &= 5\text{th term} + 0.5(6\text{th term} - 5\text{th term}) \text{ in array} \\ \Rightarrow Q_2 &= 4.0 + 0.5(4.5 - 4.0) = 4.0 + 0.5(0.5) = 4.0 + 0.25 = 4.25 \text{ kgs} \end{aligned}$$

Or, from second step

$$\begin{aligned} Q_2 &= [5\text{th term} + 6\text{th term}]/2, \text{ in array} \\ \Rightarrow Q_2 &= (4.0+4.5)/2 = 8.5/2 = 4.25 \text{ kgs} \end{aligned}$$

Now, when $r=3$, i.e.,

$$\begin{aligned} Q_3 &= 3 \times \left(\frac{10+1}{4} \right)^{\text{th}} \text{ term} = 3(11/4) = 33/4 = (8.25)^{\text{th}} \text{ term} \\ &= 8\text{th term} + 0.25(9\text{th term} - 8\text{th term}) \text{ in array} \\ \Rightarrow Q_3 &= 5.2 + 0.25(5.8 - 5.2) = 5.2 + 0.25(0.6) = 5.2 + 0.15 = 5.35 \text{ kgs} \end{aligned}$$

To find deciles, we have

$$D_r = r \times \left(\frac{n+1}{10} \right)^{\text{th}} \text{ term in the array for Deciles, } r = 1, 2, \dots, 9.$$

Now, when $r=2$, i.e.,

$$\begin{aligned} D_2 &= 2 \times \left(\frac{10+1}{10} \right)^{\text{th}} \text{ term} = 2(11/10) = 22/10 = (2.2)^{\text{th}} \text{ term} \\ &= 2^{\text{nd}} \text{ term} + 0.2(3^{\text{rd}} \text{ term} - 2^{\text{nd}} \text{ term}) \text{ in array} \\ \Rightarrow D_2 &= 2.6 + 0.2(3.5 - 2.6) = 2.6 + 0.2(0.9) = 2.6 + 0.18 = 2.78 \text{ kgs} \end{aligned}$$

To find percentiles, we have

$$P_r = r \times \left(\frac{n+1}{100} \right)^{\text{th}} \text{ term in the array for Deciles, } r = 1, 2, 3, \dots, 99.$$

Now, when $r=98$, i.e.,

$$\begin{aligned} P_{98} &= 98 \times \left(\frac{10+1}{100} \right)^{\text{th}} \text{ term} = 98(11/100) = 1078/100 = (10.78)^{\text{th}} \text{ term} \\ &\approx 10^{\text{th}} \text{ term in array} \\ \Rightarrow P_{98} &= 6.5 \text{ kgs} \end{aligned}$$

Case 2: Partition values for discrete frequency data

Quartiles, Deciles and Percentiles for discrete frequency data $X_i / f_i, i = 1, 2, \dots, n$; can be obtained by the following steps.

Step 1. Find the cumulative frequencies (**CFs**) for the discrete frequency data.

Step 2. Find $Q_r = r \times \left(\frac{N+1}{4} \right)^{\text{th}}$ term lies in the CF, for Quartiles, $r=1, 2, 3$.

$D_r = r \times \left(\frac{N+1}{10} \right)^{\text{th}}$ term lies in the CF, for Deciles, $r=1, 2, \dots, 9$.

$P_r = r \times \left(\frac{N+1}{100} \right)^{\text{th}}$ term lies in the CF, for Deciles, $r=1, 2, 3, \dots, 99$.

where $N = \sum_{i=1}^n f_i$.

Step 3. Find a **CF** $\geq r \times \left(\frac{N+1}{4} \right)$ for quartiles, **CF** $\geq r \times \left(\frac{N+1}{10} \right)$ for deciles, and **CF** $\geq r \times \left(\frac{N+1}{100} \right)$ for percentiles.

Step 4. The Q_r or D_r or P_r is a value of the variable X (say), which corresponds to **CF**, obtained in 3rd step.

Example 23: Find, Q_3, D_1 , and P_{68} of the following

$xi:$	7	10	15	18	20	22
$fi:$	5	7	10	12	8	4

Solution: Given discrete frequency data

$xi:$	7	10	15	18	20	22
$fi:$	5	7	10	12	8	4
CFs:	5	12	22	34	42	46

Here, $N = \sum_{i=1}^n f_i = 46$.

To find quartiles

$$Q_r = r \times \left(\frac{N+1}{4} \right)^{\text{th}} \text{ term lies in the CF}$$

Now when $r=3$, we have

$$Q_3 = 3 \times \left(\frac{46+1}{4} \right)^{\text{th}} = 3(47/4) = (35.25)^{\text{th}} \text{ term lies in the CF} = 42, \text{ (i.e., as CF} = 42 > 35.25).$$

=> corresponding to CF=42, we have, $xi = 20$,

=> $Q_3 = 20$ units.

Now to find deciles, we have

$$D_r = r \times \left(\frac{N+1}{10} \right)^{\text{th}} \text{ term lies in the CF, for Deciles, } r = 1, 2, \dots, 9.$$

Now, when $r=1$, we have

$$D_1 = 1 \times \left(\frac{46+1}{10} \right)^{\text{th}} = (47/10) = (4.7)^{\text{th}} \text{ term lies in the CF} = 5, \text{ (i.e., as CF} = 5 > 4.7).$$

=> corresponding to CF=5, we have, $xi = 7$,

=> $D_1 = 7$ units.

Now when $r=68$, we have

$$P_{68} = 68 \times \left(\frac{46+1}{100} \right)^{\text{th}} = 68(47/100) = 544/100 = (5.44)^{\text{th}} \text{ term lies between 5th and 6th terms.}$$

=> $P_{68} = 5^{\text{th}} \text{ term} + 0.44(6^{\text{th}} \text{ term} - 5^{\text{th}} \text{ term})$

$P_{68} = 7 + 0.44(10 - 7)$, (because, 5^{th} term lies in the CF=5, and 6^{th} term lies in the CF=12)

=> $P_{68} = 7 + 0.44(3) = 7 + 1.32 = 8.32$ units;

Case 3. Quartiles, Deciles and Percentiles for grouped frequency data

Consider the grouped frequency data $X_i - X_j / f_i$, $i = 1, 2, \dots, n$; then partition values can be obtained by the following steps.

Step 1. Find the cumulative frequencies(CFs) for the discrete frequency data.

Step 2. Find $Q_r = \left(\frac{rN}{4} \right)^{\text{th}}$ term in the CF, for Quartiles, $r=1, 2, 3$.

$$D_r = \left(\frac{rN}{10}\right)^{\text{th}} \text{ term in the CF, for Deciles, } r= 1, 2, \dots, 9.$$

$$P_r = \left(\frac{rN}{100}\right)^{\text{th}} \text{ term in the CF, for Deciles, } r= 1, 2, 3, \dots, 99.$$

where $N = \sum_{i=1}^n f_i$.

Step 3. Find a $CF \geq \left(\frac{rN}{4}\right)$ for quartiles, $CF \geq \left(\frac{rN}{10}\right)$ for deciles, and $CF \geq \left(\frac{rN}{100}\right)$ for percentiles.

Step 4. Find a class interval, which corresponds to CF , obtained in 3rd step, gives the required Quartile, or Decile or Percentile class.

Step 5. Once getting the class for a partition value, use the following formula to compute the required partition value. i.e.,

$$Q_r = L + \left(\frac{rN}{4} - C\right) \times \frac{h}{f}, \text{ for quartiles, } r=1, 2, 3.$$

$$D_r = L + \left(\frac{rN}{10} - C\right) \times \frac{h}{f}, \text{ for deciles, } r=1, 2, \dots, 9.$$

$$P_r = L + \left(\frac{rN}{100} - C\right) \times \frac{h}{f}, \text{ for percentiles, } r=1, 2, \dots, 99.$$

Where, L , denote the lower limit of quartile/decile/percentile class

C , the cumulative frequency of class, just preceding(previous) to partition value class

h , the width of partition value class, and f , frequency of the partition value class.

Example 24: Find Q_1 , Q_3 , D_3 , and P_{99} of the following

Weight in kgs.	10-20	20-30	30-40	40-50	50-60	60-70	70-80
No. of Persons	3	5	10	18	9	7	2

Solution: To compute median, we have

Weight in kgs.	10-20	20-30	30-40	40-50	50-60	60-70	70-80	Total
No. of Persons f_i	3	5	10	18	9	7	2	$54=N$
Cumulative Frequency(CFs)	3	8	18	36	45	52	54	

To find respective quartile class, we have

$$Q_r = \left(\frac{rN}{4}\right)^{\text{th}} \text{ term lies in the CF, for Quartiles, } r=1, 2, 3.$$

Now when $r=1$,

$$Q_1 = \left(\frac{1 \times 54}{4}\right)^{\text{th}} = (13.5)^{\text{th}} \text{ term lies in the CF} = 18$$

$\Rightarrow Q_1 \text{ class} = 30-40.$

Here, $L=30, h=10, C=8, r=1, f=10$.

Therefore, Q1 weight is given by

$$\begin{aligned} Q_1 &= L + \left(\frac{N}{4} - C \right) \times \frac{h}{f} \\ &= 30 + (13.5 - 8) \times \frac{10}{10} \end{aligned}$$

$$\Rightarrow Q1 = 30 + 5.5 = 35.5 \text{ kgs}$$

Now when $r=3$,

$$Q_3 = \left(\frac{3 \times 54}{4} \right)^{\text{th}} = (40.5)^{\text{th}} \text{ term lies in the CF} = 45$$

\Rightarrow Q3 class = 50-60.

Here, $L=50, h=10, C=36, r=3, f=9$.

Therefore, Q3 weight is given by

$$\begin{aligned} Q_3 &= L + \left(\frac{3N}{4} - C \right) \times \frac{h}{f} \\ &= 50 + (40.5 - 36) \times \frac{10}{9} \end{aligned}$$

$$\Rightarrow Q3 = 50 + 5 = 55 \text{ kgs}$$

Now when $r = 3$,

$$D_3 = \left(\frac{3 \times 54}{10} \right)^{\text{th}} = (16.2)^{\text{th}} \text{ term lies in the CF} = 18$$

\Rightarrow D3 class = 30-40.

Here, $L=30, h=10, C=8, r=3, f=10$.

Therefore, D3 weight is given by

$$\begin{aligned} D_3 &= L + \left(\frac{3N}{10} - C \right) \times \frac{h}{f} \\ &= 30 + (16.2 - 8) \times \frac{10}{10} \\ &= 30 + 6.2 = 36.2 \end{aligned}$$

$$\Rightarrow D3 = 36.2 \text{ kgs}$$

Now when $r = 99$,

$$P_{99} = \left(\frac{99 \times 54}{100} \right)^{\text{th}} = (53.46)^{\text{th}} \text{ term lies in the CF} = 54$$

\Rightarrow P99 class = 70-80.

Here, $L=10, h=10, C=52, r = 99, f=2$.

Therefore, P99 weight is given by

$$P_{99} = L + \left(\frac{99N}{100} - C \right) \times \frac{h}{f}$$

$$= 70 + (53.46 - 52) \times \frac{10}{2}$$

$$\Rightarrow P_{99} = 77.3 \text{ kgs}$$

Objective questions

- The average which is affected by extreme values is
 - Median
 - Mode
 - Mean
 - None
- The average which lies exactly at the centre part of the given distribution is
 - Median
 - Mode
 - Mean
 - all
- The average which is highly useful for businessmen is
 - Median
 - Mode
 - Mean
 - None of the above
- The average which is calculated for qualitative data also is
 - Median
 - Mode
 - Mean
 - both a and b.
- The average which is calculated for quantitative data only is
 - Median
 - Mode
 - Mean
 - None

Exercise

- Define central tendency. Write the chief characteristics of a good measure of central tendency.
- Write the various measures of central tendency. Explain any one of them.
- Write the properties of mean. Prove any one of them.
- Deduce the effect of change of origin and change of scale on arithmetic mean.
- Derive the expression for combined arithmetic mean of two sets of data.

Or show that, $\bar{x}_c = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$

- Find simple arithmetic mean and weighted arithmetic mean of first n natural numbers, where weights being the corresponding numbers.
- Write a note on partition values. Or, what are partition values?
- Find the mean, median and mode from the following data
Heart beats/min: 72 78 80 75 79 70 71 77 75 74
- Find the mean, median and mode for the following data
Student of 10th standard: A B C D E F G
Height in cms: 162 168 160 175 169 170 171
- Find the mean, median and mode for the following data
Height in cms: 160 162 168 169 170 171 175
No. of Students: 5 8 15 20 9 6 2
- The marks obtained by 30 students of a class in mathematics are given below. Find the mean marks of that class.

Marks obtained	10-25	25-40	40-55	55-70	70-85	85-100
No. of students	2	3	10	6	2	3

Also, compute median and mode.

12. Find missing frequency of the data given below which shows the mean daily pocket allowance of college students of a town is Rs.180/-

Daily pocket allowance (Rs)	110-130	130-150	150-170	170-190	190-210	210-230	230-250
No. of students	7	6	9	13	-	5	4

13. Find missing frequencies of the distribution given below which shows the mean daily wages of 50 labours is Rs.545.2/-.

Daily wages (Rs)	500-520	520-540	540-560	560-580	580-600
No. of labours	12	?	?	6	10

14. A survey was conducted by a group of students as a part of their environment awareness programme. During the survey, they have collected the following data regarding height of rose plants in a rose garden. Find suitable average height of rose plants.

Height of rose plants(feet)	0-0.5	0.5-1.0	1.0-1.5	1.5-2.0	2.0-2.5	2.5-3.0	3.0-3.5
Number of plants	1	2	4	5	6	2	3

15. A physician examined and recorded the heartbeats(per minute) of 30 pregnant women in his hospital. Find the average heartbeats/minute per women.

Heartbeats/minute	65-69	70-74	75-79	80-84	85-89	90-94
No. of pregnant women	2	5	11	8	3	1

Also, compute median and mode.

16. Find the mean for the following data

Daily wage(Rs.)	below 150	below 200	below 250	below 300	below 350	below 400	below 450
N0. of workers	5	16	27	65	80	93	100

Also, compute median and mode.

17. Find the mean weight of 100 children in pounds(lbs) from the following data.

Weight(lbs)	>10	>15	>20	>25	>30	> 35	>40
No. of children	100	86	67	40	28	15	4

Also, compute median and mode.

18. A school conducts a mid-term examination for X-standard students, in which 45 boys scored an average marks of 64.5 and that of 32 girls is 68.2. Find the mean marks of all the students taken together.

19. The mean height (in cms.) of 80 students of a class is 168.5cms. The mean height of 30 girls is 162.2cms. Find the mean height of boys.

20. The mean of height of 20 boys is 66.5inches. While calculating the average height, a person recorded one value as 63 inches, instead of 68inches. Find the actual mean height.

21. The mean of marks of 25 boys is 66.5%. While calculating, a person recorded wrongly the values as 63 % and 68%, instead 65% and 73% respectively. Find the correct mean.

22. A person travels from Mysore to Bangalore by Bus. The speed of first 50kms it runs at a speed of 45kms/hr., and the next 50 kms at a speed of 55kms/hr. The remaining 40kms the bus runs at a speed of 30kms/hr. Find the average speed of the Bus.
23. A person travels from Mysore to Chennai. First, he travels 50 kms by car at a speed of 65 kms/hr., next 20 kms by bike at a speed of 50kms/hr., and the remaining 250 kms by train at a speed of 45 kms/hr. Find the average speed of the total journey.
24. Out of 80 students who took an examination, 35 passed in the second class (50 to 59%) and 18 passed in the 1st class. (60% and above). Find the median of the marks.
25. A train runs 25 miles at a speed of 30 m. p.h. another 50 miles at a speed of 40m.p.h. then due to repairs of the track, travels for 6 minutes at a speed of 10 mph and finally covers the remaining distance of 24 miles at a speed of 24mph. What is the average speed in mph? Also, verify your answer with actual formula for speed.
26. The numbers 3.2, 5.8, 7.9, and 9.5 has frequencies x , $x + 2$, $x - 3$ and $x + 6$ respectively. If the A.M. is 4.876, find the value of x .

UNIT 3 MEASURES OF DISPERSION

3.1 Objective

The main objective of dispersion or variation is to determine the reliability or consistency of the given data set and also, to find out the variability within the sample. In this unit we study various measures of dispersion, and their merits, demerits and their applications.

3.2 Introduction

In the previous chapter ‘measures of central tendency’, we have discussed about various types of averages. These averages indicate i) the concentration of observations at the centre part of the given distribution, and ii) variations between two or more samples, i.e., they do not indicate variations within the sample clearly. Therefore averages are not just enough to justify the stability or uniformity or consistency of the given distribution. Thus it is necessary to study different methods to have an idea about variations within the sample. This could be achieved through ‘measures of dispersion’. For eg., consider three leading cricketers say A, B and C and their scores in first five one-day international matches are as follows.

Cricketer	I	II	III	IV	V	Average(\bar{x})
A	30	45	50	55	90	54
B	20	55	98	25	72	54
C	06	47	14	63	140	54

From the above table, it is observed that the average (arithmetic mean) scores of all the three cricketers remain same. But by looking at the data carefully, the scores of ‘B’ and ‘C’, varied(i.e., ups and downs) lot as compared to the scores of ‘A’, who has shown considerable improvement in his performance from match to match. Thus, averages alone will not give complete idea about ‘stability’ or ‘consistency’ of the data, and they do not take into consideration of ‘dispersion’ or ‘spreadness’ or ‘variations’ within the given set of observations. Therefore, one should not take decision blindly or simply using averages, instead have a look at the given distribution of values, and this will lead to measures of variability at great extent before drawing some conclusion about the distribution of values.

Thus, *‘measures of dispersion’ is defined as a statistical measure, deals with the study of ‘spreadness’, i.e., how far the given set of observations away from a central value, such as mean, median, mode etc.*

3.3 Types of Dispersion Measures

There are mainly two types of measure of dispersion. They are

- a. Absolute measures
- b. Relative measures

Absolute measures	Relative measures
a. Range(R) b. Quartile deviation(QD) c. Mean deviation(MD) d. Standard deviation(SD)	a. Coefficient of Range b. Coefficient of Quartile deviation c. Coefficient of Mean deviation d. Coefficient of Variation(CV)

3.3.1 Difference between Absolute and Relative measures of Dispersion

Absolute measures	Relative measures
1. Based on units of measurements 2. Based on actual values 3. These are not expressed in terms of rates, ratios and percentages 4. Not good for comparative study	1. Not Based on units of measurements 2. Not Based on actual values 3. These are expressed in terms of rates, ratios and percentages 4. Good for comparative study

3.3.2 Characteristics of a good or an ideal measure of central tendency

1. It should be rigidly defined.
2. It should be based on all the observations.
3. It should be easy to understand.
4. It should be used for further mathematical or statistical analysis.
5. It should be least affected by sampling fluctuations.
6. It should be least affected by extreme or abnormal values.

3.3a. Range(R): Range is the difference between largest value and the smallest value of the given set of observations. It is denoted as 'R'. Symbolically, range(R) is then given by

$$R = H - L$$

Where H , is the largest or highest value and L , is the smallest or lowest value.

The relative measure of range is coefficient of range and is given by

$$\text{Coefficient of Range} = \frac{H - L}{H + L}$$

Application of range

- a. It is useful in measuring quality of products. i.e., in statistical quality control of items.
- b. In finance, say difference between the low and high prices of a commodity over a period of time. For eg., shares, gold prices
- c. In equity reports
- d. Risk analysis in investments

3.3b. Quartile Deviation (QD): It is half times the difference between upper quartile(Q3) and the lower quartile(Q1). Symbolically, it is given by

$$\text{Quartile Deviation} = \frac{Q_3 - Q_1}{2}$$

The relative measure of quartile deviation is coefficient of quartile deviation and is given by

$$\text{Coefficient of Quartile Deviation} = \frac{Q_3 - Q_1}{2} \bigg/ \frac{Q_3 + Q_1}{2} = \frac{Q_3 - Q_1}{Q_3 + Q_1}.$$

3.3c. Mean Deviation(MD): It is the mean of absolute deviations of set of values taken from a central value, like mean, median etc. Symbolically, if x_1, x_2, \dots, x_n are the set of n observations and A be any constant, then the mean deviation about A , is given by

$$\text{Mean Deviation}(A) = \frac{\sum_{i=1}^n |x_i - A|}{n}$$

where $A = \text{mean}(\bar{x})$ or $\text{median}(M)$, or $\text{mode}(Z)$, etc.

If x_1, x_2, \dots, x_n are the set of n observations with respective frequencies f_1, f_2, \dots, f_n , and A be any constant, then the mean deviation about A , for a frequency data is given by

$$\text{Mean Deviation}(A) = \frac{\sum_{i=1}^n f_i |x_i - A|}{N}$$

where, $N = \sum_{i=1}^n f_i$.

If deviation is taken from mean, then it is known as mean deviation from mean, usually denoted by $MD(\bar{x})$ and if the deviation is taken from median, then it is known as mean deviation about median, usually denoted by $MD(M)$, etc.

The relative measure of mean deviation is called the coefficient of mean deviation and thus coefficient of mean deviation about A is given by

$$\text{Coefficient Mean Deviation}(A) = \frac{MD(A)}{A}.$$

Thus, coefficient of mean deviation about mean is given by

$$\text{Coefficient Mean Deviation}(\bar{x}) = \frac{MD(\bar{x})}{\bar{x}}.$$

Coefficient of mean deviation about median is given by

$$\text{Coefficient Mean Deviation}(M) = \frac{MD(M)}{M},$$

and etc.

3.3d. Standard Deviation(SD): It is the positive square root of mean of algebraic sum of squared deviations of set of values taken from their mean. It 'is' denoted by Greek letter

' σ ' (sigma). Thus if x_1, x_2, \dots, x_n are the set of n observations with mean (\bar{x}), then the standard deviation is given by

$$\text{Standard Deviation}(\sigma) = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}}, \text{ for raw data}$$

If x_1, x_2, \dots, x_n are the set of n observations with respective frequencies f_1, f_2, \dots, f_n , with mean ($\bar{x} = \sum_{i=1}^n f_i x_i / N$), the standard deviation for a frequency data is then given by

$$\text{Standard Deviation}(\sigma) = \sqrt{\frac{\sum_{i=1}^n f_i (x_i - \bar{x})^2}{N}}.$$

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

Note: Variance is the square of standard deviation (SD). That is, $(\text{SD})^2 = \sigma^2 = \text{Variance}$.

Note: For practical point of view above formulae can be written as

$$\sigma = \sqrt{\frac{\sum_{i=1}^n x_i^2}{n} - \left(\frac{\sum_{i=1}^n x_i}{n}\right)^2} = \sqrt{\frac{\sum_{i=1}^n x_i^2}{n} - (\bar{x})^2}, \text{ for raw data,}$$

where $\bar{x} = \sum_{i=1}^n x_i / n$

And, for frequency data $f_i/X_i, i=1,2,\dots,n$ we write

$$\sigma = \sqrt{\frac{\sum_{i=1}^n f_i x_i^2}{N} - \left(\frac{\sum_{i=1}^n f_i x_i}{N}\right)^2} = \sqrt{\frac{\sum_{i=1}^n f_i x_i^2}{N} - (\bar{x})^2},$$

where, $\bar{x} = \sum_{i=1}^n f_i x_i / N$ and $N = \sum_{i=1}^n f_i$.

Merits and demerits of Standard Deviation (SD)

Standard Deviation has few merits (advantages) and demerits (disadvantages). They are given in the form of table.

Merits	Demerits
1. It is rigidly defined.	1. It cannot be calculated even if one observation is missing.
2. It is based on all the observations.	2. It cannot be calculated for frequency distributions with 'open end class' at the
3. It can be used for further	

algebraic treatment. 4. It is least affected by sampling fluctuations. 5. It is least affected by extreme values.	tails, for eg., less than 20, more than 80, etc. 3. It cannot be used to analyse qualitative characteristics such as honesty, beauty, etc. 4. It cannot be calculated graphically. 5. It is difficult to calculate as compared to range
---	--

Note: Since standard deviation satisfies most of the requisites of a good measure of dispersion, and hence it is to be called as an **'Ideal' measure of dispersion**.

Applications of standard deviation

It is one of the widely used measures of variation. Namely,

- a. It is used in statistical quality control of products.
- b. It is used find the consistency of two or more sets of data.
- c. In finance **standard deviation** is used as a measure of volatility. For eg. Price data.
- d. In pooling it is the key factor of calculating margins of error.

Variance: The square of a standard deviation(σ) is called the variance. That is, Variance = σ^2 . Thus,

$$\text{Variance} = \sigma^2 = \frac{\sum_{i=1}^n f_i (x_i - \bar{x})^2}{N}$$

Coefficient of Standard deviation: Coefficient of standard deviation is defined as the ratio of standard deviation to the arithmetic mean of the given data. Symbolically,

$$\text{Coefficient of standard deviation} = \frac{\text{standard deviation}}{\text{mean}} = \frac{\sigma}{\bar{x}}$$

This is practically more seldom used and thus we define coefficient of variation based on standard deviation.

Coefficient of Variation(C.V.): It is hundred times the coefficient of standard deviation. Symbolically,

$$C.V. = \frac{\sigma}{\bar{x}} \times 100$$

Remark. Coefficient of variation is highly useful for comparative study of two or more data sets. If a data set having lesser C.V. is called more consistent or more reliable or more uniform. That is, if $C.V.(A) < C.V.(B)$, then data set A is more consistent than the data set B, i.e., observations in data set A is more closer than in set B.

3.4 Properties of standard deviation

Property 1. Standard deviation is independent of change of origin but not independent of change of scale. That is, if $u_i = (x_i - A)/h$, where A, the origin and h, the scale are two positive constants, then

$$\sigma_x = h\sigma_u = h \times \sqrt{\frac{\sum_{i=1}^n (u_i - \bar{u})^2}{n}}.$$

Proof: Let x_1, x_2, \dots, x_n are the set of n observations with mean(\bar{x}), then the standard deviation is given by

$$\sigma = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}}. \quad (1)$$

Let u_i be a new variable such that $u_i = (x_i - A)/h$, where A , the origin and h , the scale are two positive constants. Then,

$$x_i = A + hu_i \quad (2)$$

Summing over $i = 1, 2, \dots, n$ on both sides and dividing by n , we get

$$\begin{aligned} \sum_{i=1}^n x_i / n &= A \sum_{i=1}^n (1) / n + h \sum_{i=1}^n u_i / n \\ \bar{x} &= nA / n + h\bar{u} \\ \Rightarrow \bar{x} &= A + h\bar{u} \end{aligned} \quad (3)$$

Therefore from equations (2) and (3), is given by

$$x_i - \bar{x} = h(u_i - \bar{u}) \quad (4)$$

Squaring both sides of(4) and on taking sum over $i = 1, 2, \dots, n$, we get

$$\sum_{i=1}^n (x_i - \bar{x})^2 = h^2 \sum_{i=1}^n (u_i - \bar{u})^2 \quad (5)$$

Dividing equation (5) throughout by 'n', we get

$$\begin{aligned} \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n} &= h^2 \times \frac{\sum_{i=1}^n (u_i - \bar{u})^2}{n} \\ \Rightarrow \sigma_x^2 &= h^2 \sigma_u^2 \end{aligned}$$

Taking square root both sides, we get

$$\Rightarrow \sigma_x = h\sigma_u = h \times \sqrt{\frac{\sum_{i=1}^n (u_i - \bar{u})^2}{n}}. \quad (6)$$

Which is independent of 'A', the origin but not independent of change of scale(h). Hence Standard deviation is independent of change of origin but not independent of change of scale.

Remark: Above result can be extended to frequency data in similar lines. This could be achieved through multiplying equation(2) by f_i , throughout and later dividing by $N = \sum_{i=1}^n f_i$, wherever necessary.

Property 2. Standard deviation is not less than mean deviation from mean

Proof: Here we have to show, σ not less than $MD(\bar{x})$, implies,

$$\sigma \geq MD(\bar{x})$$

That is,

$$\sigma^2 \geq (MD(\bar{x}))^2 \quad (1)$$

Let $Z_i = |x_i - \bar{x}|$, then

$$MD(\bar{x}) = \frac{\sum_{i=1}^n f_i Z_i}{N} \quad \text{and} \quad SD(\sigma) = \sqrt{\frac{\sum_{i=1}^n f_i Z_i^2}{N}} \quad (2)$$

Therefore equation (1) gives,

$$\begin{aligned} \frac{\sum_{i=1}^n f_i Z_i^2}{N} &\geq \left(\frac{\sum_{i=1}^n f_i Z_i}{N} \right)^2 \Rightarrow \frac{\sum_{i=1}^n f_i Z_i^2}{N} - \left(\frac{\sum_{i=1}^n f_i Z_i}{N} \right)^2 \geq 0 \\ &\Rightarrow \frac{\sum_{i=1}^n f_i Z_i^2}{N} - \bar{Z}^2 \geq 0 \Rightarrow \frac{\sum_{i=1}^n f_i (Z_i - \bar{Z})^2}{N} \geq 0, \end{aligned} \quad (3)$$

which is true always. i.e., $\text{Var}(Z) \geq 0$, always $\Rightarrow SD(\sigma) \geq 0$, since standard deviation is always nonnegative, and thus,

$$\sigma \geq MD(\bar{x})$$

i.e., standard deviation is not less than mean deviation from mean.

Property 3. Standard Deviation of Combined Series(Combined SD)

Let there be k sets of random samples of sizes n_i ($i=1,2,\dots,k$), each with respective means \bar{x}_i , and standard deviations σ_i , Then the combined standard deviation of k -sets of data is given by

$$\text{Combined SD}(\sigma_c) = \sqrt{\frac{n_1(\sigma_1^2 + d_1^2) + n_2(\sigma_2^2 + d_2^2) + \dots + n_k(\sigma_k^2 + d_k^2)}{n_1 + n_2 + \dots + n_k}},$$

where $d_i = \bar{x}_i - \bar{x}_c$, for all $i=1,2,\dots,k$, and $\bar{x}_c = \frac{n_1\bar{x}_1 + n_2\bar{x}_2 + \dots + n_k\bar{x}_k}{n_1 + n_2 + \dots + n_k}$.

Proof: Let $(x_{11}, x_{12}, \dots, x_{1n_1}), (x_{21}, x_{22}, \dots, x_{2n_2}), (x_{31}, x_{32}, \dots, x_{3n_3}), \dots, (x_{k1}, x_{k2}, \dots, x_{kn_k})$ be the k -sets of random samples with respective sample sizes n_i , and sample means \bar{x}_i , $i=1,2,\dots,k$.

Then the combined mean of k -sets of data is given by

$$\bar{x}_c = \frac{n_1\bar{x}_1 + n_2\bar{x}_2 + \dots + n_k\bar{x}_k}{n_1 + n_2 + \dots + n_k} \quad (1)$$

We know that,

$$\sigma_1^2 = \frac{\sum_{i=1}^{n_1} (x_{1i} - \bar{x}_1)^2}{n_1} \Rightarrow n_1 \sigma_1^2 = \sum_{i=1}^{n_1} (x_{1i} - \bar{x}_1)^2$$

$$\text{Similarly, we have } n_2 \sigma_2^2 = \sum_{j=1}^{n_2} (x_{2j} - \bar{x}_2)^2, \dots, n_k \sigma_k^2 = \sum_{k=1}^{n_k} (x_{kk} - \bar{x}_k)^2 \quad (2)$$

The combined standard deviation for k sets of data is given by

$$\sigma_c = \sqrt{\frac{\sum_{i=1}^{n_1} (x_{1i} - \bar{x}_c)^2 + \sum_{j=1}^{n_2} (x_{2j} - \bar{x}_c)^2 + \dots + \sum_{k=1}^{n_k} (x_{kk} - \bar{x}_c)^2}{n_1 + n_2 + \dots + n_k}} \quad (3)$$

$$\text{Let } d_i = (\bar{x}_i - \bar{x}_c), \text{ for } i=1,2,\dots,k \quad (4)$$

Now, we write,

$$\sum_{i=1}^{n_1} (x_{1i} - \bar{x}_c)^2 = \sum_{i=1}^{n_1} (x_{1i} - \bar{x}_1 + \bar{x}_1 - \bar{x}_c)^2 = \sum_{i=1}^{n_1} (x_{1i} - \bar{x}_1)^2 + n_1 (\bar{x}_1 - \bar{x}_c)^2, \quad (\text{by eqn. 4}) \quad (5)$$

Where the cross product term vanish because, $\sum_{i=1}^{n_1} (x_{1i} - \bar{x}_1) = 0$. Thus we have

$$\sum_{i=1}^{n_1} (x_{1i} - \bar{x}_c)^2 = n_1 \sigma_1^2 + n_1 d_1^2 \Rightarrow n_1 (\sigma_1^2 + d_1^2) \quad (6)$$

$$\sum_{j=1}^{n_2} (x_{2j} - \bar{x}_c)^2 = \sum_{j=1}^{n_2} (x_{2j} - \bar{x}_2)^2 + n_2 (\bar{x}_2 - \bar{x}_c)^2 = n_2 (\sigma_2^2 + d_2^2), \dots, \sum_{k=1}^{n_k} (x_{kk} - \bar{x}_c)^2 = n_k (\sigma_k^2 + d_k^2), \quad (7)$$

Using equations (6) and (7), eqn.(3) reduces to

$$\sigma_c = \sqrt{\frac{n_1 (\sigma_1^2 + d_1^2) + n_2 (\sigma_2^2 + d_2^2) + \dots + n_k (\sigma_k^2 + d_k^2)}{n_1 + n_2 + \dots + n_k}}, \quad (8)$$

which is the required formula for combined standard deviation.

Note. In particular, if $k = 2$, i.e., for two sets of data, combined standard deviation of (n_1+n_2) observations is given by

$$\sigma_c = \sqrt{\frac{n_1 (\sigma_1^2 + d_1^2) + n_2 (\sigma_2^2 + d_2^2)}{n_1 + n_2}}$$

Proof is just like above result.

Example: Find the mean, standard deviation and variance of first n natural numbers

Solution: Given the first n natural numbers, $x: 1, 2, \dots, n$. Then the mean is given by

$$\bar{x} = \sum_{i=1}^n x_i / n$$

$$\begin{aligned}
&= [1 + 2 + 3 + \dots + n] / n \\
&= [n(n+1)/2] / n \\
&= (n+1) / 2, \text{ units.}
\end{aligned}$$

Consider,

x_i^2	1^2	2^2	3^2	\cdot	\cdot	\cdot	n^2
---------	-------	-------	-------	---------	---------	---------	-------

then,
$$\sum_{i=1}^n x_i^2 = [1^2 + 2^2 + \dots + n^2] = \frac{(n+1)(2n+1)}{6}$$

$$\begin{aligned}
\text{Variance} = \text{Var}(x) &= \sum_{i=1}^n x_i^2 / n - (\bar{x})^2 \\
&= \frac{(n+1)(2n+1)}{6} - \left(\frac{n+1}{2}\right)^2 \\
&= \left(\frac{n+1}{2}\right) \left(\frac{2n+1}{3} - \frac{n+1}{2}\right) \\
&= \left(\frac{n+1}{2}\right) \left(\frac{n-1}{6}\right) = \frac{n^2 - 1}{12}.
\end{aligned}$$

Therefore, Standard deviation(SD) = $\sqrt{\text{var}(x)} = \sqrt{\frac{n^2 - 1}{12}}$.

Example 2: Find the mean and standard deviation of $\frac{ax+b}{c}$, where a , b , and c are constants when the random variable X has the mean ' m ' and sd ' σ '.

Solution: Given mean and sd of X is ' m ' and ' σ ', respectively. Now to find mean and sd of

$y = \frac{ax+b}{c}$, we have

$$\begin{aligned}
\text{Mean}(y) &= \sum_{i=1}^n y / n = \sum_{i=1}^n \left(\frac{ax+b}{nc}\right) \\
&= \frac{a \sum_i x + nb}{nc} \Rightarrow \frac{a \left(\sum_i x / n\right) + b}{c} \Rightarrow \frac{am + b}{c} \\
\Rightarrow \text{Mean}(y) &= \frac{am + b}{c}
\end{aligned}$$

Example: Find the mean deviation from mean and standard deviation of AP, $a, a+d, a+2d, \dots, a+2nd$, and verify that SD is greater than MD(mean).

Solution: We know that mean of a series in AP is the mean of its first and last terms. Hence the mean of the given series is $\bar{x} = \frac{a + (a + 2nd)}{2} = a + nd$

X	$ x - \bar{x} $	$(x - \bar{x})^2$
a	nd	n^2d^2
$a+d$	$(n-1)d$	$(n-1)^2d^2$
$a+2d$	$(n-2)d$	$(n-2)^2d^2$
\cdot	\cdot	\cdot
\cdot	\cdot	\cdot
\cdot	\cdot	\cdot
$a+(n-2)d$	$2d$	2^2d^2
$a+(n-1)d$	d	d^2
$a+nd$	0	0
$a+(n+1)d$	d	d^2
$a+(n+2)d$	$2d$	2^2d^2
\cdot	\cdot	\cdot
\cdot	\cdot	\cdot
\cdot	\cdot	\cdot
$a+(2n-2)d$	$(n-2)d$	$(n-2)^2d^2$
$a+(2n-1)d$	$(n-1)d$	$(n-1)^2d^2$
$a+2nd$	Nd	n^2d^2

$$\begin{aligned} \text{Mean deviation(Mean)} = MD(\bar{x}) &= \frac{\sum_{i=1}^n |x - \bar{x}|}{2n+1} \\ &= \frac{2d(1+2+3+ \dots + n)}{2n+1} \\ &= \frac{2d(n(n+1)/2)}{2n+1} = \frac{n(n+1)d}{2n+1}. \end{aligned}$$

$$\begin{aligned} \text{Variance}(\sigma^2) &= \frac{\sum_{i=1}^n (x - \bar{x})^2}{2n+1} \\ &= \frac{2d^2(1^2 + 2^2 + 3^2 + \dots + n^2)}{2n+1} \\ &= \frac{n(n+1)d^2}{3} \end{aligned}$$

$$\Rightarrow \text{Sd}(\sigma) = \sqrt{\text{variance}} = d \times \sqrt{\frac{n(n+1)}{3}}.$$

To verify, $SD > MD(\text{Mean})$, we suppose,

$$(SD)^2 > [MD(\bar{x})]^2$$

Then we have,

$$\Rightarrow \frac{n(n+1)d^2}{3} > \left[\frac{n(n+1)d}{2n+1} \right]^2$$

$$\Rightarrow (2n+1)^2 > 3n(n+1)$$

$\Rightarrow n^2 + n + 1 > 0$, which is true always, since n is a positive integer.

Thus, $SD > MD(\text{mean})$.

3.5 Standard deviation and Root mean square deviation

Root mean square deviation is usually denoted by 's' and is defined as

$$s = \sqrt{\frac{\sum_{i=1}^n f_i (x_i - A)^2}{N}}$$

Where A , is any arbitrary constant, and s^2 , is called the mean square deviation.

By definition, we have,

$$\begin{aligned} s^2 &= \frac{\sum_{i=1}^n f_i (x_i - A)^2}{N} = \frac{\sum_{i=1}^n f_i (x_i - \bar{x} + \bar{x} - A)^2}{N} \\ &= \frac{1}{N} \left[\sum_{i=1}^n f_i \{ (x_i - \bar{x})^2 + (\bar{x} - A)^2 + 2(\bar{x} - A)(x_i - \bar{x}) \} \right] \\ &= \frac{1}{N} \left[\sum_{i=1}^n f_i (x_i - \bar{x})^2 + N(\bar{x} - A)^2 + 2(\bar{x} - A) \sum_{i=1}^n f_i (x_i - \bar{x}) \right] \\ &= \frac{1}{N} \sum_{i=1}^n f_i (x_i - \bar{x})^2 + (\bar{x} - A)^2, \left(\because \sum_{i=1}^n f_i (x_i - \bar{x}) = 0, \text{ by property of arithmetic mean} \right) \\ &= \sigma^2 + (\bar{x} - A)^2 \\ &\Rightarrow s^2 \geq \sigma^2, \text{ always.} \end{aligned}$$

Remark: Note that, $s^2 = \sigma^2$, only if $A = \bar{x}$. That is, mean square deviation is least when the deviations are taken from $A = \bar{x}$, the mean. Hence variance is the least value of mean square deviation and consequently, standard deviation is the minimum/least value of root mean square deviation.

Example: Find the range, quartile deviation, mean deviation from mean, standard deviation and their relative measures from the following data.

Weight of students in kgs. 40, 22, 35, 26, 58, 45, 52, 65, 36, 48

Solution: Array: Weight (x_i) of students in kgs. 22, 26, 35, 36, 40, 45, 48, 52, 58, 65

$$\text{Range}(R) = H - L,$$

where H , the highest value & L , the lowest value. Thus,

$$R = 65 - 22 = 43,$$

$$\text{coefficient of range} = (H - L) / (H + L) = 43 / 87 = 0.4943$$

To find quartile deviation, we have

$$Q_1 = 1 \times \left(\frac{10+1}{4} \right)^{\text{th}} \text{ term} = (11/4) = (2.75)^{\text{th}} \text{ term}$$

$$= 2^{\text{nd}} \text{ term} + 0.75(3^{\text{rd}} \text{ term} - 2^{\text{nd}} \text{ term}) \text{ in array}$$

$$\Rightarrow Q_1 = 26 + 0.75(35 - 26) = 32.75 \text{ kgs}$$

$$\text{And, } Q_3 = 3 \times \left(\frac{10+1}{4} \right)^{\text{th}} \text{ term} = 3(11/4) = 33/4 = (8.25)^{\text{th}} \text{ term}$$

$$= 8^{\text{th}} \text{ term} + 0.25(9^{\text{th}} \text{ term} - 8^{\text{th}} \text{ term}) \text{ in array}$$

$$\Rightarrow Q_3 = 52 + 0.25(58 - 52) = 52 + 0.25(6) = 53.5 \text{ kgs}$$

$$\text{Quartile Deviation} = \frac{Q_3 - Q_1}{2} = \frac{53.5 - 32.75}{2} = 10.375 \text{ kgs}$$

$$\text{Coefficient of QD} = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{53.5 - 32.75}{53.5 + 32.75} = 0.2405$$

To find mean deviation from mean, and its coefficient, we have

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} = (22+26+\dots+65)/10 = 427/10 = 42.7$$

$$x_i : \quad 22, \quad 26, \quad 35, \quad 36, \quad 40, \quad 45, \quad 48, \quad 52, \quad 58, \quad 65$$

$$|x_i - \bar{x}| : 20.7, 16.7, 7.7, 6.7, 2.7, 2.3, 5.3, 9.3, 15.3, 22.3$$

$$\text{Thus, } MD(\bar{x}) = \frac{\sum_{i=1}^n |x_i - \bar{x}|}{n} = \frac{109}{10} = 10.9 \text{ kgs,}$$

$$\text{and Coefficient of MD about mean} = \frac{MD(\bar{x})}{\bar{x}} = \frac{10.9}{42.7} = 0.2553$$

To find standard deviation & coefficient of variation(C.V.) we have

$$x_i^2 : 484 \quad 676, \quad 1225, \quad 1296, \quad 1600, \quad 2025, \quad 2304, \quad 2704, \quad 3364, \quad 4225 \Rightarrow \sum_{i=1}^n x_i^2 = 19903$$

$$\text{standard deviation } \sigma = \sqrt{\frac{\sum_{i=1}^n x_i^2}{n} - (\bar{x})^2} = \sqrt{\frac{19903}{10} - (42.7)^2} = 12.92 \text{ kgs}$$

$$\text{Coefficient of variation(C.V.)} = 100 \sigma / \bar{x} = 100(12.92)/42.7 = 30.2576.$$

Example 2. Find standard deviation and coefficient of variation

$$x_i : \quad 10 \quad 22 \quad 26 \quad 35 \quad 36 \quad 40$$

$$f_i : \quad 3 \quad 12 \quad 20 \quad 7 \quad 8 \quad 5$$

$$\text{Solution: We know that, mean}(\bar{x}) = \frac{\sum_{i=1}^n f_i x_i}{N}, \text{ where } N = \sum_{i=1}^n f_i$$

$$\sum_{i=1}^n f_i x_i = 1547, \text{ and } N = \sum_{i=1}^n f_i = 55, \text{ so that}$$

$$\text{mean}(\bar{x}) = \sum_{i=1}^n f_i x_i / N = 1547/55 = 28.1273$$

$$\text{and, standard deviation } \sigma = \sqrt{\frac{\sum_{i=1}^n f_i x_i^2}{N} - (\bar{x})^2} = \sqrt{\frac{46571}{55} - (28.1273)^2} = 7.4566$$

$$\text{Coefficient of variation(C.V.)} = 100 \sigma / \bar{x} = 100(7.4566)/28.1273 = 26.51.$$

Example 3. The scores of two golfers for 5 rounds were as follows:

Golfer A: 33 38 35 40 34

Golfer B: 22 30 40 28 35

Find which golfer may be considered to be more i) better, ii) consistent player?

Solution: we have

Golfer A scores (x)	33	38	35	40	34	$\sum x = 180$
Golfer A scores (y)	22	30	40	28	35	$\sum y = 155$
x^2	1089	1444	1225	1600	1156	$\sum x^2 = 6514$
y^2	484	900	1600	784	1225	$\sum y^2 = 4993$

i. Mean score of golfer A = $\bar{x} = \sum_{i=1}^n X / n = 180/5 = 36$

Mean score of golfer A = $\bar{y} = \sum_{i=1}^n Y / n = 155/5 = 31$

Since $\bar{x} > \bar{y} \Rightarrow$ Golfer A is the better player.

ii. For golfer A: $\sigma_A = \sqrt{\frac{\sum_{i=1}^n x_i^2}{n} - (\bar{x})^2} = \sqrt{\frac{6514}{5} - (36)^2} = 2.61$

Coefficient of variation(C.V.(A)) = $100 \sigma_A / \bar{x} = 100(2.61)/36 = 7.25\%$

For golfer B: $\sigma_B = \sqrt{\frac{\sum_{i=1}^n y_i^2}{n} - (\bar{y})^2} = \sqrt{\frac{4993}{5} - (31)^2} = 6.13$

C.V.(B) = $100 \sigma_B / \bar{y} = 100(6.13)/31 = 19.77\%$

$\Rightarrow CV(A) < CV(B)$, i.e., $7.25 < 19.77 \Rightarrow$ golfer A is more consistent player.

Objective questions

1. Absolute measure of dispersion is

- a.mean b.range c. coefficient of variation d. None

2. Relative measures of dispersion has

- a.units b. no units c. either a or b d. None

3. Standard deviation is----- than root mean square deviation

- a.less b.more c.equal d.all the above

4. Mean deviation is least when it is measured from

a. mode b. mean c. range d. median

5. The ideal measure of dispersion is

a. range b. mean deviation c. Standard deviation d. All the above

Questions

1. Define dispersion. Write the chief characteristics of a good measure of dispersion.
2. Write the various measures of dispersion. Explain any one of them.
3. Differentiate between absolute and relative measures of dispersion.
4. Which is the ideal measure of dispersion? Write its characteristics.
5. Write the properties of standard deviation.
6. Deduce the effect of change of origin and change of scale on standard deviation.
7. Derive the expression for standard deviation of two sets of data.
8. Find simple standard deviation and weighted standard deviation of first n natural numbers, where weights being the corresponding numbers.
9. Find standard deviation and standard deviation of first n natural numbers, where weights being the corresponding opposite numbers.
10. Find the range, quartile deviation, mean deviation from mean, standard deviation from the following data

Heart beats/min: 72 78 80 75 79 70 71 77 75 74

11. Find the range, quartile deviation, mean deviation from mean, standard deviation from the following data

Student of 10th standard: A B C D E F G
 Height in cms: 162 168 160 175 169 170 171

12. Find the range, quartile deviation, mean deviation from mean, standard deviation from the following data

Height in cms: 160 162 168 169 170 171 175
 No. of Students: 5 8 15 20 9 6 2

13. Find the quartile deviation, mean deviation from median, from the following data

Weight in kgs: 50 56 60 64 66 70 75 80
 No. of persons: 5 7 13 16 9 6 2 3

14. The marks obtained by 30 students of a class in mathematics are given below. Find the range, quartile deviation, mean deviation from mean, standard deviation marks from the following data

Marks obtained	10-25	25-40	40-55	55-70	70-85	85-100
No. of students	2	3	7	6	6	6

15. The distribution below shows the total number of runs scored by leading batsmen in first fifty one-day international cricket matches. Find mean deviation from mean and standard deviation number of runs.

Runs scored(00's)	10-15	15-20	20-25	25-30	30-35	35-40
No. of batsmen	7	5	16	12	2	3

16. Find missing frequency of the data given below which shows the mean daily pocket allowance of college students of a town is Rs.180/-. Hence obtain standard deviation from the following data.

Daily pocket allowance (Rs)	110-130	130-150	150-170	170-190	190-210	210-230	230-250
No. of students	7	6	9	13	-	5	4

17. Compare the variability of the series A and B

Series A: 348, 457, 424, 682, 524, 388, 380, 438

Series B: 487, 508, 620, 382, 408, 266, 186, 218

18. An analysis of monthly wages of the workers of two organisations X and Y gave the following results.

	X	Y
No. of worker	50	60
Av. monthly wage	60	48
Variance	100	144

i) Which organization pays better wage?

ii) Which organization has more homogeneity in wages?

UNIT 4

MOMENTS, SKEWNESS AND KURTOSIS

4.1 Objective: Here the objective is to study the symmetry, asymmetry and the peakedness of a given data.

4.2 Introduction: Moments are of special type of averages, these indicate the concentration of values at the centre part of the given distribution, spreadness, asymmetry, peakedness etc. Moments are usually denoted by Greek letter μ_r . There are two types of moments, they are

- a. Central Moments or Moments about mean
- b. Raw Moments

4.2a The r^{th} order central moments: *The r^{th} order central moments are the arithmetic mean of the sum of the r^{th} power of deviations of set of n observations taken from their mean.* That is, if x_1, x_2, \dots, x_n are the set of n observations with mean (\bar{x}), then the r^{th} order central moments or r^{th} order moments about mean is given by

$$\mu_r = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^r, \quad r = 1, 2, \dots \text{ (for raw data)}$$

For a frequency data $X_i / f_i, i = 1, 2, \dots, n$ of a set of n values, r^{th} order moments about mean is given by

$$\mu_r = \frac{1}{N} \sum_{i=1}^n f_i (x_i - \bar{x})^r, \quad r = 1, 2, \dots, \text{ and } N = \sum_{i=1}^n f_i.$$

Note that, $\mu_0 = 1$.

4.2b Properties of Central moments

- a. First order central moment is always zero. i.e., when $r = 1$, first order moment about mean is zero. i.e.,

$$\mu_1 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}) = 0, \text{ being the algebraic sum of deviations of set of values taken}$$

from their mean is zero always.

- b. Second order central moment i.e., μ_2 is called the variance of the distribution. i.e., when $r = 2$, we have,

$$\mu_2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = \sigma^2 = \text{variance.}$$

- c. Third order central moment i.e., μ_3 is the *measure of skewness* of the distribution.
- d. Fourth order central moment i.e., μ_4 is the *measure of kurtosis* of the distribution.

Remark. *Moments about origin zero* is given by

$$\mu_r = \frac{1}{n} \sum_{i=1}^n (x_i - 0)^r = \frac{1}{n} \sum_{i=1}^n x_i^r, \quad r=1, 2, \dots$$

4.3 The r^{th} order raw moments: The r^{th} order raw moments is the arithmetic mean of sum of the r^{th} power of deviations of set of n observations taken from any arbitrary constant A . That is, if x_1, x_2, \dots, x_n are the set of n observations, and A be any constant, then the r^{th} order raw moments or r^{th} order moments about A , is given by

$$\mu'_r = \frac{1}{n} \sum_{i=1}^n (x_i - A)^r, \quad r = 1, 2, \dots$$

Note: The first order raw moment about origin zero is the ‘mean’ of the distribution. i.e., when $r = 1$, first order raw moment about zero is

$$\mu'_1 = \frac{1}{n} \sum_{i=1}^n (x_i - 0) = \frac{1}{n} \sum_{i=1}^n x_i, \text{ the mean.}$$

Note: When $r=1$, $\mu'_1 = \frac{1}{n} \sum_{i=1}^n (x_i - A) = \frac{1}{n} \sum_{i=1}^n x_i - A = \bar{x} - A \Rightarrow \bar{x} = A + \mu'_1$.

4.4 Relation Between r^{th} order Central and Raw Moments

Consider the r^{th} order central moments

$$\begin{aligned} \mu_r &= \frac{1}{N} \sum_{i=1}^n f_i (x_i - \bar{x})^r, \quad r = 1, 2, \dots \\ &= \frac{1}{N} \sum_{i=1}^n f_i (x_i - A + A - \bar{x})^r \\ &= \frac{1}{N} \sum_{i=1}^n f_i (d_i + A - \bar{x})^r, \text{ where } d_i = x_i - A. \\ &= \frac{1}{N} \sum_{i=1}^n f_i (d_i - \mu'_1)^r, \text{ since } \bar{x} = A + \mu'_1 \end{aligned}$$

Thus we have,

$$\begin{aligned} \mu_r &= \frac{1}{N} \sum_{i=1}^n f_i \left\{ d_i^r - r C_1 d_i^{r-1} \mu'_1 + r C_2 d_i^{r-2} \mu_1'^2 - r C_3 d_i^{r-3} \mu_1'^3 + \dots + (-1)^r \mu_1'^r \right\} \\ &= \frac{1}{N} \sum_{i=1}^n f_i d_i^r - \frac{1}{N} \sum_{i=1}^n f_i r C_1 d_i^{r-1} \mu'_1 + \frac{1}{N} \sum_{i=1}^n f_i r C_2 d_i^{r-2} \mu_1'^2 + \dots + (-1)^r \mu_1'^r \frac{1}{N} \sum_{i=1}^n f_i \\ &= \mu'_r - r C_1 \frac{1}{N} \sum_{i=1}^n f_i d_i^{r-1} \mu'_1 + r C_2 \frac{1}{N} \sum_{i=1}^n f_i d_i^{r-2} \mu_1'^2 + \dots + (-1)^r \mu_1'^r, \text{ where } N = \sum_{i=1}^n f_i. \\ &= \mu'_r - r C_1 \mu'_{r-1} \mu'_1 + r C_2 \mu'_{r-2} \mu_1'^2 + \dots + (-1)^r \mu_1'^r \end{aligned}$$

In particular, the first four moments about mean are

When $r = 1$, we have

$$\mu_1 = 0$$

When $r = 2$, we have

$$\mu_2 = \mu'_2 - 2_{C_1} \mu'_{2-1} \mu'_1 + 2_{C_2} \mu'_{2-2} \mu_1'^2$$

$$\mu_2 = \mu'_2 - 2\mu_1'^2 + \mu_1'^2, \text{ where } \mu'_0 = 1.$$

$$\mu_2 = \mu'_2 - \mu_1'^2$$

When $r = 3$, we have

$$\mu_3 = \mu'_3 - 3_{C_1} \mu'_{3-1} \mu_1' + 3_{C_2} \mu'_{3-2} \mu_1'^2 - 3_{C_3} \mu'_{3-3} \mu_1'^3$$

$$\mu_3 = \mu'_3 - 3\mu_2' \mu_1' + 3\mu_1'^3 - \mu_1'^3, \text{ where } \mu'_0 = 1.$$

$$\mu_3 = \mu'_3 - 3\mu_2' \mu_1' + 2\mu_1'^3$$

When $r = 4$, we have

$$\mu_4 = \mu'_4 - 4_{C_1} \mu'_{4-1} \mu_1' + 4_{C_2} \mu'_{4-2} \mu_1'^2 - 4_{C_3} \mu'_{4-3} \mu_1'^3 + 4_{C_4} \mu'_{4-4} \mu_1'^4$$

$$\mu_4 = \mu'_4 - 4\mu_3' \mu_1' + 6\mu_2' \mu_1'^2 - 4\mu_1'^4 + \mu_1'^4, \text{ where } \mu'_0 = 1.$$

$$\mu_4 = \mu'_4 - 4\mu_3' \mu_1' + 6\mu_2' \mu_1'^2 - 3\mu_1'^4$$

Above formulae are enable us to find the moments about mean, once a constant and moments about any point or constant are known.

4.5 Property of Change of Origin and Change of Scale on Moments

The r^{th} order central moments(moments about mean) is independent of change of origin but not independent of change of scale. That is, if $u_i = (x_i - A)/h$, where A , the origin and h , the scale are two positive constants, then

$$\mu_r(x) = h\mu_r(u) = h^r \times \frac{1}{N} \sum_{i=1}^n f_i (u_i - \bar{u})^r .$$

Proof: Let x_1, x_2, \dots, x_n are the set of n observations with mean(\bar{x}), then the standard deviation is given by

$$\mu_r = \frac{1}{N} \sum_{i=1}^n f_i (x_i - \bar{x})^r, r = 1, 2, \dots \quad (1)$$

Let u_i be a new variable such that $u_i = (x_i - A)/h$, where A , the origin and h , the scale are two positive constants. Then,

$$x_i = A + hu_i \quad (2)$$

Taking sum over $i = 1, 2, \dots, n$ on both sides (2) and dividing by n , we get

$$\sum_{i=1}^n x_i / n = A \sum_{i=1}^n (1) / n + h \sum_{i=1}^n u_i / n$$

$$\bar{x} = nA / n + h\bar{u}$$

$$\Rightarrow \bar{x} = A + h\bar{u} \quad (3)$$

Therefore from equations (2) and (3), is given by

$$x_i - \bar{x} = h(u_i - \bar{u}) \quad (4)$$

Taking r th power on both sides of (4), and then multiplying by f_i and taking sum over $i = 1, 2, \dots, n$, we get

$$\sum_{i=1}^n f_i (x_i - \bar{x})^r = h^r \sum_{i=1}^n f_i (u_i - \bar{u})^r \quad (5)$$

Dividing equation (5) throughout by ' N ', we get

$$\frac{\sum_{i=1}^n f_i (x_i - \bar{x})^r}{N} = h^r \times \frac{\sum_{i=1}^n f_i (u_i - \bar{u})^r}{N}$$

$$\Rightarrow \mu_r(x) = h^r \mu_r(u) \quad (6)$$

Where, $\mu_r(x) = \frac{1}{N} \sum_{i=1}^n f_i (x_i - \bar{x})^r$, $r = 1, 2, \dots$ and $\mu_r(u) = \frac{1}{N} \sum_{i=1}^n f_i (u_i - \bar{u})^r$, for $r = 1, 2, \dots$

Equation(6) is independent of ' A ', the origin but not independent of change of scale(h). Hence moments about mean are independent of change of origin but not independent of change of scale.

4.6 Karl-Pearson's β and γ Coefficients

Karl-Pearson's defined some coefficients which are based on first four central moments and they are given by

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} \text{ and } \beta_2 = \frac{\mu_4}{\mu_2^2} \text{ and } \gamma_1 = \sqrt{\beta_1} \text{ and } \gamma_2 = \beta_2 - 3$$

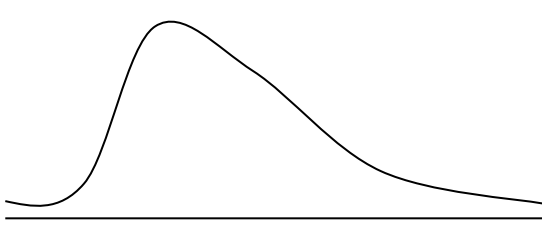
These β and γ coefficients' are independent of units of measurements. Distribution is negatively skewed if μ_3 is negative and positively skewed if μ_3 is positive.

4.7 Skewness

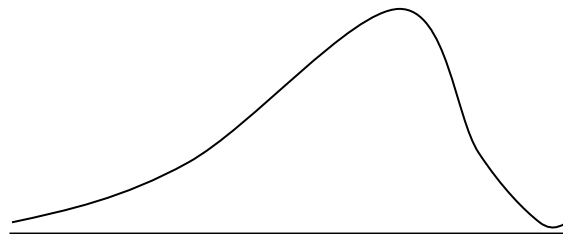
Skewness indicates '*lack of symmetry or asymmetry*' of a distribution. For a frequency distribution, if mean is more than median and median is more than mode (Mean > Median > Mode), then the distribution is said to be 'positively skewed' if mean is less than median less than mode i.e., Mean < Median < Mode, then the distribution is said to be negatively skewed, and if Mean = Median = Mode, i.e., if mean, median and mode coincide then the distribution is said to be symmetrically skewed or in simple, it is called a symmetric distribution. Thus for a symmetric distribution, $\beta_1 = 0$, and $\beta_2 = 3$.

Note: that for any symmetric distribution, all odd ordered moments about mean are zero. That is, $\mu_{2r+1} = 0$, $r = 0, 1, 2, 3, \dots$. In particular, $\mu_1 = 0$, $\mu_3 = 0$, etc.

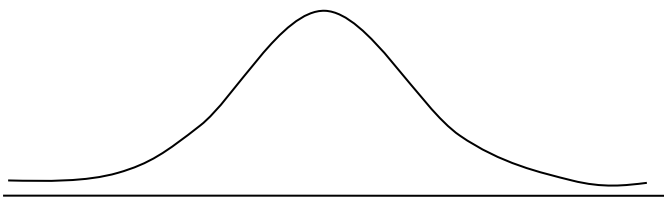
Following are the curves represent the nature of distribution of observations



Positively skewed distribution
(mean > median > mode)



Negatively skewed distribution
(mean < median < mode)



Symmetric distribution ($\bar{X} = M = Z$).

4.8 Measurement of skewness

There are mainly two types of measures to measure skewness of the distribution. Namely, they are

- a. Absolute measures
- b. Relative measures

4.8.1 Absolute measures of Skewness

Absolute measures of skewness are defined due to Karl-Pearson, and these are defined as follows

- a) $S_{KP} = \bar{X} - M$
- b) $S_{KP} = \bar{X} - Z$, where \bar{X} is the mean, M, the median and Z is the Mode of the given data.

Absolute measures of skewness due to Bowley is defined by

- c) $S_{KB} = (Q_3 - M) - (M - Q_1)$.

Absolute measures have units of measurements and these are not ideal for comparing two or more sets of data.

4.8.2 Relative Measures of Skewness

Relative measures of skewness are defined due to Karl-Pearson, and are generally known as coefficient of skewness. Thus, Karl-Pearson's coefficient of skewness is given by

- a) $S_{KP} = \frac{\bar{X} - Z}{\sigma}$, when mode is uniquely defined.
- b) $S_{KP} = \frac{3(\bar{X} - M)}{\sigma}$, when mode is ill defined, where \bar{X} is the mean, M, the median and Z is the Mode of the given data.
- c) Bowley's coefficient of skewness is defined by

$$S_B = \frac{Q_3 + Q_1 - 2M}{Q_3 - Q_1} = \frac{Q_3 + Q_1 - 2Q_2}{Q_3 - Q_1} = \frac{(Q_3 - M) - (M - Q_1)}{(Q_3 - M) + (M - Q_1)},$$

since median (M) = Q_2 .

This is also known as quartile coefficient of skewness. It is very useful when the data is highly skewed, data is very much affected by sampling fluctuations, and when frequency distribution has open end classes or of unequal class width.

Relative measures are independent of units of measurements and thus good for comparative study of two or more sets of data.

4.9 Limits of Karl-Pearson's coefficient of skewness

Karl-Pearson's coefficient of skewness lies between -3 and +3. That is, $-3 < S_{KP} < 3$.

Proof: Consider,

$$\begin{aligned} |\bar{X} - M| &= \left| \frac{\sum_{i=1}^n x_i}{n} - M \right| \\ &= \left| \frac{1}{n} \left(\sum_{i=1}^n x_i - M \right) \right| \leq \frac{1}{n} \left(\sum_{i=1}^n |x_i - M| \right) \leq \frac{1}{n} \left(\sum_{i=1}^n |x_i - \bar{x}| \right) \end{aligned} \quad (1)$$

(\because the sum of the absolute deviations is minimum when taken about median)

Therefore, using equation(1), we get

$$|S_{KP}|^2 = \left| \frac{3(\bar{X} - M)}{\sigma} \right|^2 \leq \frac{\left(3 \frac{1}{n} \sum_{i=1}^n |x_i - \bar{x}| \right)^2}{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\left(3 \sum_{i=1}^n |x_i - \bar{x}| \right)^2}{n \sum_{i=1}^n (x_i - \bar{x})^2} \quad (2)$$

By using Cauchy –Schwartz inequality we have

$$\left(\sum_{i=1}^n a_i b_i \right)^2 \leq \sum_{i=1}^n a_i^2 \sum_{i=1}^n b_i^2$$

Letting $b_i = 1$, for $i = 1, 2, \dots, n$ we have

$$\left(\sum_{i=1}^n a_i \right)^2 \leq n \sum_{i=1}^n a_i^2, \quad \left(\because \sum_{i=1}^n (1) = n \right)$$

Implies,

$$\frac{\left(\sum_{i=1}^n a_i \right)^2}{n \sum_{i=1}^n a_i^2} \leq 1 \quad (3)$$

Thus, on using equations (2) and (3), we get

$$|S_{KP}|^2 \leq 3^2 \Rightarrow |S_{KP}| \leq 3$$

$$\Leftrightarrow -3 \leq S_{KP} \leq 3$$

That is the limits or the range of Karl-Pearson's coefficient of skewness is (-3 and +3).

Remark: The above limits are rarely attained in practical cases. If $S_{KP} = 0$, if $\bar{x} = M$.

4.10 Limits of Bowley's coefficient of skewness

Bowley's coefficient of skewness lies between -1 and +1. That is, $-1 < S_B < 1$. i.e. show that for any two positive constants, a and b, $-1 < S_B < 1$.

Proof: Consider any two positive constants, a and b, such that

$$|a - b| \leq |a + b| \Rightarrow \frac{|a - b|}{|a + b|} \leq 1, \quad (1)$$

We know that $(Q_3 - M)$ and $(M - Q_1)$ are both non-negative. Thus, letting, $a = (Q_3 - M)$ and $b = (M - Q_1)$ in equation(1), we get

$$\frac{|(Q_3 - M) - (M - Q_1)|}{|(Q_3 - M) + (M - Q_1)|} \leq 1$$

$$\Rightarrow |S_B| \leq 1$$

$$\Leftrightarrow -1 \leq S_B \leq 1$$

Thus, limits of Bowley's coefficient of skewness lies between -1 and +1.

Note 1: Suppose, $S_B = +1$, if $M = Q_1 = 0 \Rightarrow M = Q_1$, $S_B = -1$, if $Q_3 - M = 0 \Rightarrow Q_3 = M$, and $S_B = 0$, if $Q_3 - M = M - Q_1 \Rightarrow Q_3 = -Q_1$.

Note 2: Sometimes it may happen that one of the coefficients of skewness give positive value while the other gives negative skewness.

Note 3: In Bowley's coefficient of skewness, the distribution factor of variations is eliminated by dividing the absolute measure of skewness i.e., $(Q_3 - M) = (M - Q_1)$, by the measure of dispersion $Q_3 - Q_1$, quartile range.

Note 4: the main drawback of Bowley's coefficient of skewness is that it is based only on the central 50% part of the data and ignores the remaining 50% of the data towards the extremes.

Note 4: Coefficient of skewness based on moments is

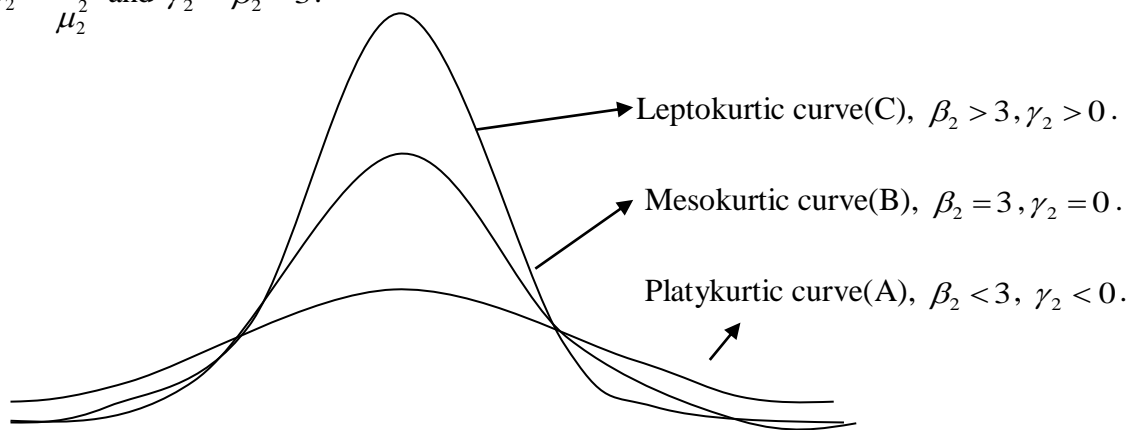
$$S_{KM} = \frac{\sqrt{\beta_1}(\beta_2 + 3)}{2(5\beta_2 - 6\beta_1 - 9)}$$

Note that, $SKM = 0$, if $\beta_1 = 0$, or $\beta_2 = -3$. But $\beta_2 = \frac{\mu_4}{\mu_2^2} > 0$ always, and therefore $SKM = 0$, if $\beta_1 = 0$.

4.11 Kurtosis

Kurtosis indicates the ‘flatness or peakedness’ of a distribution. Prof. Karl-Pearson suggested a measure to represent the frequency distribution in terms of curve called ‘convexity of the frequency curve’. Measure of Central values, dispersion and skewness do not give complete picture about the nature of the frequency distribution as will be clear from the kurtosis curves, in which all the three curves say A, B, and C are symmetric about the mean ‘ μ ’, and have the same range. Peakedness of the curve is measured by the Pearson’s coefficient β_2 or γ_2 , is given by

$$\beta_2 = \frac{\mu_4}{\mu_2^2} \text{ and } \gamma_2 = \beta_2 - 3.$$



In the above, if the curve(A) is more flat then it is said to be ‘Platy Kurtic curve’, and in this case $\beta_2 < 3$, i.e., $\gamma_2 < 0$. If the curve of the type(B), which is neither flat nor peak is called the mesokurtic curve or normal curve. And in this case $\beta_2 = 3$, i.e., $\gamma_2 = 0$. And If the curve of the type(C), which is more peaked then the curve is called the ‘Leptokurtic’ curve and in this case $\beta_2 > 3$, i.e., $\gamma_2 > 0$.

Objective Questions

1. For a symmetric distribution mean, median and mode are
 - a. Equal
 - b. mean > median > mode
 - c. mean < median < mode
 - d. None
2. For a distribution, mean > median > mode is called
 - a. Negatively skewed
 - b. positively skewed
 - c. symmetric
 - d. all
3. For a distribution, mean < median < mode is called
 - a. Negatively skewed
 - b. positively skewed
 - c. symmetric
 - d. all
4. Bowley’s coefficient of skewness lies between
 - a. (-1, 0)
 - b. (-1, 1)
 - c. (1, 3)
 - d. (0, 1)
5. Karl-Pearson’s coefficient of skewness lies between
 - a. (-1, 0)
 - b. (-1, 1)
 - c. (-3, 3)
 - d. (0, 3)

Exercise

1. A survey was conducted by a group of students as a part of their environment awareness programme. During the survey, they have collected the following data regarding the number of plants in a locality containing 50 houses. Find Karl-Pearson's coefficient of skewness from the following data

Number of plants	0-2	2-4	4-6	6-8	8-10	10-12	12-14
No. of houses	1	5	9	15	6	9	5

2. A physician examined and recorded the heartbeats(per minute) of 30 pregnant women in his hospital. Find Bowley's coefficient of skewness from the following data

Heartbeats/minute	65-69	70-74	75-79	80-84	85-89	90-94
No. of pregnant women	2	5	11	8	3	1

3. Find the first two moments about mean for the following data

Marks	30	35	40	45	50	55
No. of students	5	16	27	15	8	3

BLOCK – II
(PROBABILITY AND RANDOM VARIABLES)

UNIT 5: INTRODUCTION TO PROBABILITY THEORY

UNIT 6: RANDOM VARIABLE AND PROBABILITY DISTRIBUTIONS

UNIT 7: MATHEMATICAL EXPECTATION OF A RANDOM VARIABLE

UNIT 8: CENTRAL LIMIT THEOREM

UNIT 5

INTRODUCTION TO PROBABILITY THEORY

5.1 Objective

The main objective of probability theory is to understand the concept of likelihood or chance or the possibility of occurrence of a random event.

5.2 Introduction

So far we have studied descriptive statistics, which are used to describe and summarize the data, especially raw data from a research sample. Thus descriptive statistics pertains to a small group that is, simply choose a group you are interested in, record data about the group, and then use summary statistics and graphs to describe the group properties and characteristics. That is, there is no uncertainty involved in it. But in real life situations a number of cases that exists which are uncertain or probabilistic. For example, in a coin tossing experiment, hardly 50% chance of getting head or a tail; a tuberculosis patient admitted to hospital may survive or die is not 50:50, a 40% of body burnt patient may or may not survive, likewise, when a fair dice is thrown, either the face with number 1, or 2, or . . . , or 6 will appear up with a chance of $1/6$ each, and in an another example, a 4th stage or the last stage bladder malignancy (type of cancer) patient may survive for 10 more years is may be almost 1 in 1000 such patients, and so on. etc., which are probabilistic in nature. That is the certainty of happening of these events is not sure. Thus probability is a measure related to the study of occurrence of such random events, i.e., rate at which a random event occurs

The literal meaning of the word ‘probability’ is either a ‘chance or possibility or likelihood’ of occurrence of a random event or a trial.

Definition: *Probability is defined as a measure of finding the degree of occurrence of an event. In other words, probability is defined as the chance of occurrence of an event.*

The concept of probability theory is based on set theory. ‘A set is a collection of objects, which are the elements of the set. If S is a set and x is an element of S , we write $x \in S$. If x is not an element of S , we write $x \notin S$. A set can have no elements, in which case it is called the empty set, denoted by ϕ (Dimitri P. Bertsekas and John N. Tsitsiklis)’.

“If the set S contains a finite number of elements, say x_1, x_2, \dots, x_n , we write it as a list of the elements, in braces: $S = \{x_1, x_2, \dots, x_n\}$. For example, the set of all possible outcomes of a die roll is $\{1, 2, 3, 4, 5, 6\}$, and the set of possible outcomes of a coin toss is $\{H, T\}$, where H stands for “heads” and T stands for “tails” (Dimitri P. Bertsekas and John N. Tsitsiklis)”.

5.3 Some Basics on Algebra of sets

Sets under operations of union, intersection, and complement satisfy various identities (laws) which are given below:

- i. Null set: $\phi = \{\}$.
- ii. Compliment of a set: Compliment of a set A is A^C or A' or \bar{A} .
- iii. Union of two sets: $A \cup B$.
- iv. Law of complement: $A \cup A^C = S$; $A \cap A^C = \phi$.

- v. Intersection of two sets: $A \cap B$.
- vi. Associate law: $(A \cup B) \cup C = A \cup (B \cup C)$
- vii. Identity Laws: $A \cup \phi = A$; $A \cup S = S$; $A \cap S = A$; $A \cap \phi = \phi$;
- viii. De Morgan's Laws: $(A \cup B)^c = A^c \cap B^c$; $(A \cap B)^c = A^c \cup B^c$

5.4 Some Terminologies

Experiment or trial: An experiment is a procedure to get possible outcomes. For eg. Coin tossing, dies throwing, picking an Ace card from a well shuffled pack of cards, etc., are experiments.

Deterministic Experiment: Suppose an experiment is repeated several times under identical conditions and the outcome obtained remain same then that experiment is called Deterministic experiment. For eg. lab experiments such as Physics, Chemistry etc., are in general deterministic. Say, for eg. $2H_2 + O_2 \rightarrow 2H_2O$, at room temperature ($23^{\circ}C$)

Random Experiment: Suppose an experiment is repeated several times under identical, or homogeneous conditions, the outcome obtained is not same in all trials, then that experiment is called random experiment. For eg. Coin tossing, dies throwing, picking an Ace or a King card from a well shuffled pack of cards, sales, purchase etc., are random experiments.

Outcome: It is the result or output of a random experiment. For eg., in coin tossing experiment Head (H) and Tail (T), are the two outcomes. In dies throwing experiment, 1,2,3,4,5, and 6 are the outcomes.

Sample space: It is the set containing all possible outcomes of a trial or random experiment. It is denoted by S or Ω . For eg. In coin tossing experiment, the sample space S is, $S = \{H, T\}$

Event: An event is a set containing few or all possible outcomes of a random experiment. Events are denoted by capital letters A, B, C etc. or, A_1, A_2, \dots, A_n .

For eg., Let A be the event of getting an 'even number' when a dies is thrown once, then $A = \{2,4,6\}$; Suppose the event $B = \{\text{getting 'prime number' when a dies is thrown once}\}$, then $B = \{2,3,5\}$ and etc.

5.5 Types of Events

Simple event: An event contains single outcome is called simple event. For eg., let event $A = \{\text{getting 6 in a single throw of dies}\}$, then $A = \{6\}$; Let event $B = \{\text{getting both heads when two coins are tossed}\}$, then $B = \{HH\}$ and etc.

Sure event: An event contains all possible outcomes of a random experiment is called sure event. It is equal to sample space (S). For eg. Let event $A = \{\text{getting odd or even numbers in single throw of dies}\}$, then $A = \{1, 2, 3, 4, 5, 6\} = S$; Let event $B = \{\text{getting at most one head or both heads when coin is tossed once}\}$, then $B = \{TT, TH, HT, HH\} = S$.

Null event: An event does contain any of the outcomes is called null event. It is denoted by ϕ and is given by $\phi = \{\}$.

Union of events: Union of two events say A and B is the occurrence of either event A or B or both A and B . It is denoted by $A \cup B$.

For eg. When a dies is thrown once, events A and B are defined as $A = \{\text{getting even No.}\} = \{2,4,6\}$; $B = \{\text{getting multiple of 3}\} = \{3,6\}$, then $A \cup B = \{2,3,4, 6\}$.

Intersection of events: Intersection of two or more events is the simultaneous occurrence of both or all events. It is denoted by $A \cap B$ or AB ; $A \cap B \cap C$ or ABC .

For eg. When a dies is thrown once, events A and B are defined as $A = \{\text{getting even No.}\} = \{2,4,6\}$; $B = \{\text{getting multiple of 3}\} = \{3,6\}$, then $A \cap B = \{6\}$. Suppose event $C = \{\text{getting No. more than 4}\} = \{5, 6\}$, then $A \cap B \cap C = \{6\}$.

Mutually exclusive events: Two or more events are said to be mutually exclusive if their intersection should be equal to null set or event. For eg., when a dies is thrown once, events A and B are such that $A = \{\text{getting even No.}\} = \{2, 4, 6\}$; $B = \{\text{getting odd no.}\} = \{1, 3, 5\}$, then $A \cap B = \phi = \{\}$, the null event.

5.6 Definitions of Probability

5.6.1 Classical or Mathematical or A priori or uniformity definition of probability

Statement: Let there be n equally likely, mutually exclusive, and exhaustive outcomes for a random experiment, out of which m ($1 \leq m \leq n$) outcomes are favourable to the happening of an event A . Then the probability of an event A is denoted as $P(A)$ and is given by

$$P(A) = \frac{\text{Number of favourable outcomes to } A}{\text{All possible outcomes of a Random Experiment}} = \frac{m}{n} = \frac{n(A)}{n(S)}$$

Where, $n(A)$ denote number of outcomes favourable to the happening of A , $n(S)$ denote number of outcomes of the sample space S .

Example 5.1: A coin is tossed two times. What is the probability of getting i) head both the times, ii) exactly one head, iii) at least one head, and iv) at most one head?

Solution: given coin is tossed two times, then the sample space: $S = \{HH, HT, TH, TT\}$, i.e., $n = 4$, possible outcomes. Now, let the events A and B be

i. $A = \{\text{head both the times}\} = \{HH\}$, implies $m = 1$ possibility, therefore,
 $P(A) = m/n = 1/4$.

ii. $B = \{\text{exactly one head}\} = \{HT, TH\}$, implies $m = 2$ possibilities, therefore
 $P(B) = m/n = 2/4 = 1/2$.

iii. $C = \{\text{at least one head}\} = \{\text{more than or equal to one head}\} = \{\geq 1\} = \{HT, TH, HH\}$, implies
 $m=3$, therefore $P(C) = m/n = 3/4$.

iv. $D = \{\text{at most one head}\} = \{\text{less than or equal to one head}\} = \{\leq 1 \text{ head}\}$

$\Rightarrow D = \{TH, HT, TT\}$, implies $m=3$, therefore $P(D) = m/n = 3/4$.

Note: It is known that in a pack playing cards there will be four suits of which 13 are diamonds (♠), 13 are clubs (♣), 13 are spades (♠), and 13 are hearts (♥), thus total 52 cards in a pack. Among this, 26 are red (diamond and heart), and 26 are of black (clubs and spades) coloured cards. Out of these 52 cards, there will be 4-Kings, 4-Queens, 4-Aces, 4-jacks, and each four of 2, 3, . . ., 10 in a pack.

Example 5.2. A card is selected at random from a pack of well shuffled playing cards. What is the probability of getting i) a king ii) an Ace iii) a king or a queen iii) a king or a red card?

Solution: It is known that in a pack there will be 52 cards. One card can be selected in ${}^{52}C_1 = 52$ ways, i.e, $n = n(S) = 52$. Now we define events as

- i. $A = \{\text{getting a king card}\} = \{{}^4C_1 \Rightarrow m = 4\} \Rightarrow P(A) = m/n = 4/52 = 1/13$.
- ii. $B = \{\text{getting an Ace}\} = \{{}^4C_1 \Rightarrow m = 4\} \Rightarrow P(B) = m/n = 4/52 = 1/13$.
- iii. $C = \{\text{getting a king or a queen}\} = \{{}^4C_1 + {}^4C_1 \Rightarrow m = 8\} \Rightarrow P(C) = m/n = 8/52 = 2/13$.
- iv. $D = \{\text{getting a king or a red card}\} = \{{}^4C_1 + {}^{13}C_1 - {}^1C_1 \Rightarrow m = 16\} \Rightarrow P(D) = m/n = 16/52 = 4/13$.

Example 5.3. A dice are thrown once, what is the probability of getting i) an even number, ii) an odd number, iii) an even or odd number iv) a prime number, v) an even or multiple of 3.

Solution: A die is thrown, then the sample space $S = \{1, 2, 3, 4, 5, 6\} \Rightarrow n = 6$. Then the events

- i. $A = \{\text{getting an even Number}\} = \{2, 4, 6\} \Rightarrow m = 3$. Therefore $P(A) = m/n = 3/6 = 1/2$.
- ii. $B = \{\text{getting an odd Number}\} = \{1, 3, 5\} \Rightarrow m = 3$. Therefore $P(B) = m/n = 3/6 = 1/2$.
- iii. $C = \{\text{getting an even or odd Number}\} = \{(2, 4, 6), \text{ or } (1, 3, 5)\} \Rightarrow m = 6$. Implies, $P(C) = m/n = 6/6 = 1$.
- iv. $D = \{\text{getting a prime number}\} = \{2, 3, 5\} \Rightarrow m = 3$. Therefore $P(D) = m/n = 3/6 = 1/2$.
- v. $E = \{\text{an even or multiple of 3}\} = \{2, 4, 6 \text{ or } 3, 6\} \Rightarrow m = 4$. Therefore $P(E) = m/n = 4/6 = 2/3$.

Limitations: Classical definition of probability has the following limitations:

- i. All the outcomes of a random experiment are equally likely or equally probable. For example,
 - a) probability of a candidate will pass in certain test is not 50%, since the two possible outcomes say, 'pass and fail' are the two exhaustive, mutually exclusive outcomes but not equally likely.
 - b) If a person jumps from a running bus, then his probability of survival or death will not be 50%, though survival and deaths are the two exhaustive, mutually exclusive outcomes but not equally probable.
- ii. If the exhaustive number of outcomes of a random experiment is infinite or unknown.

5.6.2 Empirical or Statistical/Posterior definition of Probability (Richard Von Mises): *If an experiment is performed repeatedly under essentially homogeneous and identical conditions, then the limiting value of the ration of the number of times the event occurs to the number of trials , as the number of trials becomes indefinitely large, is called the probability of happening of the event, it being assumed that the limit is finite and unique. Symbolically, if n trials an event A happens m times , then the probability of happening of A denoted as P(A), is given by*

$$P(A) = \lim_{n \rightarrow \infty} \frac{m}{n}$$

Note: Empirical probability is the probability of occurrence of an event based on the results obtained for a random experiment which is conducted actually for several times.

Example 5.4. A symmetric die is rolled 180 times, what is the probability of getting exactly one 6? Also find the number of times that 6 appears only once.

Solution: We know that when a symmetric die is rolled once, P{getting exactly one 6}=1/6, since it is rolled 180 times, and each throw is independent of the other throws, therefore we have P{getting exactly one 6 in 180 tosses} = (1/6)¹⁸⁰, and the number of times the exactly one 6 appears in 180 throws is 180x(1/6)=30 times.

5.6.2a Sampling with replacement: Here repetitions are allowed. Items ‘r (≤ n)’ are to be drawn out of ‘n’, things in n^r ways. i.e., n^r samples of size r each using with replacement. For example,

- a. In ‘r’, tossing’s of a fair coin results in 2^r possible outcomes.
- b. In ‘r’, tossing’s of a fair die results in 6^r possible outcomes.

5.6.2b Sampling without replacement: Here an item once chosen is not replaced before the next draw is made, so that repetitions are not permitted.

- a. If ordered samples of size ‘r (≤ n)’ are drawn from n things, then there are ⁿP_r samples without replacement. Symbolically, ⁿP_r = n!/(n-r)!
- b. If order is not considered then the samples of size ‘r (≤ n)’ are drawn from n things in ⁿC_r, ways without replacement. Symbolically,

$${}^n C_r = \binom{n}{r} = n! / [r!(n - r)!].$$

- The number of ways in which the population of n elements can be partitioned in to k subpopulations of sizes r₁, r₂, ..., r_k = n, 0 ≤ r_i ≤ n, is given by

$$\binom{n}{r_1, r_2, \dots, r_k} = \frac{n!}{r_1! r_2! \dots r_k!},$$

which is known as multinomial coefficients.

Example 5.5. The birthdays of r students of a class form a sample of size r from the 365 days of a certain year. Then the probability that all r birthdays are different is ³⁶⁵P_r/(365)^r.

Example 5.6. An urn contains 3 red, 4 green and 5 blue balls. A sample of size 6 is selected at random without replacement. Then, the probability that the sample contains 2 red, 3 green and one blue ball is

$$\frac{\binom{3}{2}\binom{4}{3}\binom{5}{1}}{\binom{12}{6}} = \frac{{}^3C_2 \times {}^4C_3 \times {}^5C_1}{{}^{12}C_6}.$$

Example 5.7. A purse contains five Rs.100/- notes, ten Rs.200/- green and four Rs.500/-notes. Five notes are selected at random without replacement. Then, the probability that the sample contains two Rs.100/- notes, two Rs.200/- green and one Rs.500/-notes is

$$\frac{\binom{5}{2}\binom{10}{2}\binom{4}{1}}{\binom{19}{5}} = \frac{{}^5C_2 \times {}^{10}C_2 \times {}^4C_1}{{}^{19}C_5}.$$

5.6.3 Axiomatic Approach of Probability

Statement: Let S be a sample space and A be an event such that $A \subseteq S$. Let a set function P is called a probability of occurrence if it satisfies the following conditions:

- i. $P(A) \geq 0$
- ii. $P(S) = 1$
- iii. Let A and B be any two disjoint events defined on S , then $P(A \cup B) = P(A) + P(B)$. In

general, A_1, A_2, \dots, A_k are k disjoint sets, then $P\left(\sum_{i=1}^k A_i\right) = \sum_{i=1}^k P(A_i)$.

5.7. Some Theorems on Probability of events

This section provides some simple theorems on probability of events which will help to evaluate probabilities of occurrence some events.

Theorem 1. Probability of impossible event (or null event) is zero. i.e., $P(\phi) = 0$.

Proof: let S be the sample space of a random experiment and ϕ , the null event or the impossible event. Then,

$$S \cup \phi = S \Rightarrow P(S \cup \phi) = P(S)$$

Since S and ϕ are mutually exclusive, and therefore

$$P(S \cup \phi) = P(S) + P(\phi) = P(S)$$

Since $P(S) = 1$, we have

$$\Rightarrow 1 + P(\phi) = 1 \Rightarrow P(\phi) = 0. \text{ Hence proved.}$$

Theorem 2. Probability of the complimentary event A^c is given by

$$P(A^c) = 1 - P(A)$$

Proof: Since A and A^c are mutually exclusive events, we have

$$P(A \cup A^c) = P(S) = 1$$

$$\Rightarrow P(A) + P(A^c) = 1$$

$$\Rightarrow P(A^c) = 1 - P(A). \text{ Hence proved.}$$

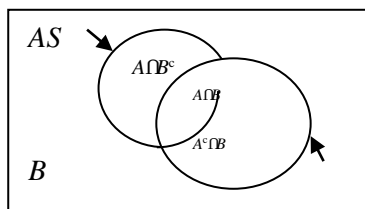
Cor. It is known that $P(A) = 1 - P(A^c) \leq 1$, always ($\because P(A^c) \geq 0$, by axiom 1). Further, since $P(A) \geq 0$, (by axiom 1), and therefore, $0 \leq P(A) \leq 1$.

Implies that probability of an event lies between 0, and 1, or the limits of probability of an event is (0, 1).

Theorem 3. Probability For any two events A and B we have

$$i) \quad P(A^c \cap B) = P(B) - P(A \cap B) \quad ii) \quad P(A \cap B^c) = P(A) - P(A \cap B).$$

Proof: From the Venn diagram, we have $B = (A \cap B^c) \cup (A \cap B)$.



Since $A \cap B$ and $A \cap B^c$ are two disjoint events, we have by axiom 3,

$$P(B) = P\{(A \cap B^c) \cup (A \cap B)\} \\ = P(A \cap B^c) + P(A \cap B)$$

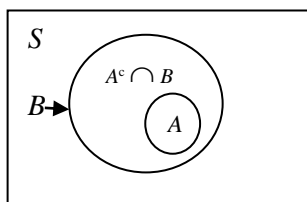
Implies, $P(A \cap B^c) = P(B) - P(A \cap B)$.

Similarly we can prove (ii) (proof left to the exerciser).

Theorem 4. If $A \subset B$, then

$$i) \quad P(A^c \cap B) = P(B) - P(A) \quad ii) \quad P(A) \leq P(B).$$

Proof: i) From the Venn diagram, $B = A \cup (A^c \cap B)$



Since A and $A^c \cap B$ are two disjoint events, we have by Axiom 3,

$$P(B) = P\{A \cup (A^c \cap B)\} \\ = P(A) + P(A^c \cap B)$$

Implies, $P(A^c \cap B) = P(B) - P(A)$.

Since $P(A^c \cap B) \geq 0$, always, $\Rightarrow P(B) - P(A) \geq 0 \Rightarrow P(B) \geq P(A)$.

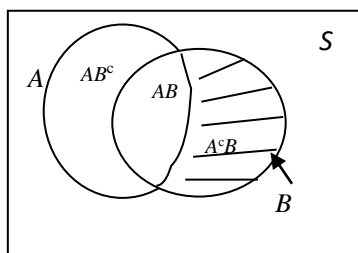
Thus, when $A \subset B$, then $P(A) \leq P(B)$.

Cor. 2. When $B \subset A$, then $P(B) \leq P(A)$ and $P(A \cap B^c) = P(A) - P(B)$. (proof left to the exerciser).

5.7.1 Addition Theorem of Probability for any two events

Theorem 5: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Proof: From the Venn diagram, $A \cup B = A \cup (A^c \cap B)$, it can be observed that, the events A and $A^c \cap B$ are mutually disjoint events, and therefore by taking probability on both sides we get



$$P(A \cup B) = P\{A \cup (A^c \cap B)\} \\ = P(A) + P(A^c \cap B)$$

By theorem 3-i, we have

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

Example 5.8. A die is thrown once. What is the probability of getting i) an even number or a prime number, ii) an odd number or a multiple of 3

Solution: Given that die is thrown once. Then the sample space, S is

$$S = \{1, 2, 3, 4, 5, 6\} \Rightarrow n = 6.$$

Let A and B be two events such that,

$$i) A = \{\text{getting an even number}\} = \{2, 4, 6\} \Rightarrow m_1 = 3 \text{ and}$$

$$B = \{\text{getting a prime number}\} = \{2, 3, 5\} \Rightarrow m_2 = 3$$

Then, $P(A) = m_1/n = 3/6 = 0.5$; and $P(B) = m_2/n = 3/6 = 0.5$

$$A \cap B = \{2\} \Rightarrow m_3 = 1 \Rightarrow P(A \cap B) = m_3/n = 1/6$$

Thus by addition theorem, we have

$$P(\text{getting an even number or a prime number}) = P(A \cup B) = P(A) + P(B) - P(A \cap B) \\ = 3/6 + 3/6 - 1/6 = 5/6$$

$$\text{Or} \quad P(A \cup B) = 0.5 + 0.5 - 0.17 = 0.83$$

ii) Let C and D be two events such that,

$$C = \{\text{getting an odd number}\} = \{1, 3, 5\} \Rightarrow m_1 = 3, \text{ and}$$

$$D = \{\text{getting a multiple of 3}\} = \{3, 6\} \Rightarrow m_2 = 2$$

Then, $P(C) = m_1/n = 3/6 = 0.5$, and $P(D) = m_2/n = 2/6 = 0.33$

$$C \cap D = \{3\} \Rightarrow m_3 = 1 \Rightarrow P(C \cap D) = m_3/n = 1/6$$

Thus, by addition theorem, we have

$$P(\text{getting an odd number or a multiple of 3}) = P(C \cup D) = P(C) + P(D) - P(C \cap D) \\ = 3/6 + 2/6 - 1/6 = 4/6$$

$$\text{Or} \quad P(C \cup D) = 0.5 + 0.33 - 0.17 = .66$$

Cor. 1. If A and B are mutually exclusive (mutually disjoint), then

$$A \cap B = \phi \Rightarrow P(A \cap B) = P(\phi) = 0, \text{ implies } P(A \cup B) = P(A) + P(B).$$

Cor. 2. If A , B and C are three non mutually exclusive events then

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(B \cap C) - P(A \cap C) + P(A \cap B \cap C).$$

Proof: Consider the LHS of above

$$P(A \cup B \cup C) = P[(A \cup B) \cup C] \\ = P(A \cup B) + P(C) - P[(A \cup B) \cap C]$$

Using theorem 5, we have

$$P(A \cup B \cup C) = P(A) + P(B) - P(A \cap B) + P(C) - P[(A \cap C) \cup (B \cap C)] \\ = P(A) + P(B) + P(C) - P(A \cap B) - \{P(B \cap C) + P(A \cap C) - P(A \cap B \cap C)\}, \text{ (by theorem 5).}$$

Thus we have,

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(B \cap C) - P(A \cap C) + P(A \cap B \cap C).$$

Hence proved.

Remark. Generalisation of Addition Theorem of probability for any k events

Theorem 6: For k events A_1, A_2, \dots, A_k , we have

$$P\left(\bigcup_{i=1}^k A_i\right) = \sum_{i=1}^k P(A_i) - \sum_{1 \leq i < j \leq k} P(A_i \cap A_j) + \sum_{1 \leq i < j < l \leq k} P(A_i \cap A_j \cap A_l) + \dots \\ \dots + (-1)^{k-1} P(A_1 \cap A_2 \cap A_3 \cap \dots \cap A_k)$$

5.8 Dependent Events: Two or more events are said to be dependent then the occurrence of one event affects the occurrence of other events.

For example,

- i) The chance of withdrawing an amount (say, in Rupees (Rs.)) for the second time from a savings bank account depends on the first withdrawn amount.
- ii) The probability of drawing a ball from a box for the second time is dependent up on first drawn ball(s), etc.

5.8.1 Conditional Probability and Bayes' Theorem

5.8.1a Conditional Probability: Let A and B be any two events, then the conditional probability of an event A when B has already or happened is symbolically denoted as $P(A/B)$ and is given by

$$P(A/B) = \frac{P(A \cap B)}{P(B)}, \text{ if } P(B) > 0.$$

Similarly, the conditional probability of an event B when A has already happened is denoted as $P(B/A)$, $P(A) > 0$, and is given by

$$P(B/A) = \frac{P(A \cap B)}{P(A)}, \text{ if } P(A) > 0.$$

Note: $P(A/B)$ denotes the probability of occurrence event A for given B , when event B has happened already. Similarly, $P(B/A)$ denotes the probability of occurrence event B , for given A , when event A has happened already.

Theorem 7. Multiplication (or Compound) Probability theorem

Statement: The probability of simultaneous occurrence of any two events A and B , is given by

$$P(A \cap B) = P(B) \cdot P(A/B), \text{ if } P(B) > 0. \\ = P(A) \cdot P(B/A), \text{ if } P(A) > 0.$$

Proof: By definition of conditional probability, when event B , has happened already is given by

$$P(A/B) = \frac{P(A \cap B)}{P(B)}, \text{ if } P(B) > 0.$$

$$\Rightarrow P(A \cap B) = P(B) \cdot P(A/B), \text{ if } P(B) > 0.$$

Similarly, when event A , has happened already, then we have

$$P(B/A) = \frac{P(A \cap B)}{P(A)}, \text{ if } P(A) > 0.$$

$$\Rightarrow P(A \cap B) = P(A) \cdot P(B/A), \text{ if } P(A) > 0.$$

Hence proved.

Theorem 8: For any three events A , B and C

$$P(A \cup B / C) = P(A/C) + P(B/C) - P(A \cap B / C)$$

Proof : By definition of conditional probability, we have

$$P(A \cup B / C) = \frac{P[(A \cap C) \cup (B \cap C)]}{P(C)}$$

Since

$$P(A \cup B) = P(A) + P(B) - P(A \cap B),$$

we have therefore

$$\begin{aligned} P(A \cup B / C) &= \frac{P(A \cap C) + P(B \cap C) - P(A \cap B \cap C)}{P(C)} \\ &= \frac{P(A \cap C)}{P(C)} + \frac{P(B \cap C)}{P(C)} - \frac{P(A \cap B \cap C)}{P(C)} \end{aligned}$$

$$\Rightarrow P(A \cup B / C) = P(A/C) + P(B/C) - P(A \cap B / C)$$

Hence proved.

5.8.1b Examples on Conditional and multiplication theorem of Probability

Example 5.9. A box contains 4 yellow and 6 white tennis balls. Two draws are made and in each draw a ball is drawn at random. What is the probability that

- both draw gives yellow balls when first drawn ball is replaced before the second draw is made? Secondly, first drawn ball is not replaced before the second draw is made?
- first draw gives a yellow and second draw gives a white ball when first drawn ball is replaced before the second draw is made? Secondly, first drawn ball is not replaced before the second draw is made?

Solution: Total number of balls = 10, of which 4 yellow, 6 white. Now let, the events A and B be

A = First draw gives yellow ball

B = Second draw gives yellow ball

Here as per the given condition, second draw depends on first draw. Thus, we have

$$P[\text{Both draw gives Yellow balls}] = P(A \cap B) = P(A)P(B|A)$$

- $P[\text{both draw gives yellow balls when first drawn ball is replaced before the second draw is made}] = P(A \cap B) = P(A)P(B|A)$
 $= ({}^4C_1/{}^{10}C_1)({}^4C_1/{}^{10}C_1) = 16/100 = 0.16.$
 - $P[\text{both draw gives yellow balls when first drawn ball is not replaced before the second draw is made}] = P(A \cap B) = P(A)P(B|A)$
 $= ({}^4C_1/{}^{10}C_1)({}^3C_1/{}^9C_1) = 12/90 = 2/15 = 0.1333.$
- $P[\text{first draw gives yellow ball and second gives white ball when first drawn ball is replaced before the second draw is made}]$
 $= P(A \cap B) = P(A)P(B|A)$
 $= ({}^4C_1/{}^{10}C_1)({}^6C_1/{}^{10}C_1) = 24/100 = 0.24.$

ii. P [first draw gives yellow ball and second gives white ball when first drawn ball is not replaced before the second draw is made]

$$\begin{aligned} &= P(A \cap B) = P(A)P(B|A) \\ &= ({}^4C_1/{}^{10}C_1)({}^6C_1/{}^9C_1) = 24/90 = 0.2667. \end{aligned}$$

5.8.2 Bayes' Theorem(Thomas Bayes).

In previous section we have discussed about the conditional probability, and it indicates the likelihood of an outcome occurring, based on a previous outcome having occurred in similar circumstances. Often we come across the analysis with initial or prior probability estimates. In such a scenario one could apply Bayes' theorem, which is based on the prior knowledge of conditions that might be related to the event. For example, if the risk of developing health problems is known to increase with age, Bayes' theorem allows the risk to an individual of a known age to be assessed more accurately by conditioning it relative to their age, rather than assuming that the individual is typical of the population as a whole.

Statement: If E_1, E_2, \dots, E_n are mutually disjoint events with $P(E_i) > 0$, for all $i = 1, 2, \dots, n$, then for any arbitrary event A , which is a subset of $\bigcup_{i=1}^n E_i$, such that $P(A) > 0$, we have

$$P(E_i | A) = \frac{P(E_i)P(A/E_i)}{\sum_{i=1}^n P(E_i)P(A/E_i)} = \frac{P(E_i)P(A/E_i)}{P(A)}, i = 1, 2, \dots, n$$

Proof: Since $A \subset \bigcup_{i=1}^n E_i$, we have $A = A \cap \left(\bigcup_{i=1}^n E_i\right) = \bigcup_{i=1}^n (A \cap E_i)$, (by distributive law)

Since $(A \cap E_i) \subset E_i$, $i = 1, 2, \dots, n$; and are mutually disjoint events, we have by addition theorem of probability

$$P(A) = P\left\{\bigcup_{i=1}^n (A \cap E_i)\right\} = \sum_{i=1}^n P(A \cap E_i)$$

Since by multiplication theorem of probability, we have

$$P(A) = \sum_{i=1}^n P(E_i)P(A/E_i), \tag{1}$$

Also, since $P(A) > 0$, we have by multiplication theorem of probability

$$P(A \cap E_i) = P(A)P(E_i/A) \tag{2}$$

Therefore, from (1) and (2), we have

$$P(E_i/A) = \frac{P(A \cap E_i)}{P(A)} = \frac{P(E_i)P(A/E_i)}{\sum_{i=1}^n P(E_i)P(A/E_i)}.$$

Note: $P(E_i) > 0$, for all $i = 1, 2, \dots, n$ are known as 'prior' probabilities because they exist before we gain any information from the experiment itself

The probabilities $P(A/E_i)$, $i = 1, 2, \dots, n$ are called likelihoods because they indicate how likely the event A under consideration to occur given each and every a prior probability.

The probability $P(E_i/A)$, $i = 1, 2, \dots, n$ are called ‘posterior probability’, because they are determined after the results of the experiment are known.

Example 5.10: A box I contains 3 red and 4 yellow balls and another box II contains 4 red and 5 yellow balls. One ball is drawn from one of the boxes, and is found to be red. Find the probability that it was drawn from box I.

Solution: Let us define the events

E_1 : Box I is selected

E_2 : Box II is selected

A: Drawing a red ball

Then we have, $P(E_1) = 1/2$, $P(E_2) = 1/2$, $P(A|E_1) = {}^3C_1/{}^7C_1 = 3/7$ and $P(A|E_2) = {}^4C_1/{}^9C_1 = 4/9$

Now, our aim is to find the probability of getting red ball from the I-box is $P(E_1|A)$, and is given by the Bayes’ theorem as

$$P(E_1/A) = \frac{P(E_1)P(A/E_1)}{\sum_{i=1}^{n=2} P(E_i)P(A/E_i)} = \frac{\frac{1}{2} \times \frac{3}{7}}{\frac{1}{2} \times \frac{3}{7} + \frac{1}{2} \times \frac{4}{9}} = 0.4909$$

Example 5.11: A boy is known to speak the truth 4 out of 5 times. He throws a die and reports that the number obtained is a six. What is the probability that the number obtained is actually six?

Solution: Let the events be

E_1 : Six is obtained on the die

E_2 : Non-Six obtained on the die

A: Boy reports that the number obtained is Six

Then, $P(E_1) = 1/6$, $P(E_2) = 5/6$ and $P(A|E_1) = 4/5$ = Probability that Boy reports six and it is actually six.

And $P(A|E_2)$ = probability that boy reports six and it is not actually six = $1/5$

Now our aim is to find $P(E_1|A)$ = probability that the number obtained is actually six when he reported it as six is given by Bayes’ theorem as

$$P(E_1/A) = \frac{P(E_1)P(A/E_1)}{\sum_{i=1}^{n=2} P(E_i)P(A/E_i)} = \frac{\frac{1}{6} \times \frac{4}{5}}{\frac{1}{6} \times \frac{4}{5} + \frac{5}{6} \times \frac{1}{5}} = \frac{4}{9} = 0.4444$$

Example 5.12. A company has two machines to produce a particular product. Machine A produces 45% of the products and machine B produces 55%. The defective rate for machine A is 8% and is 10% for machine B. If a defective item is observed, what is the probability that it was from machine A.

Solution: Let the events be

E_1 = machine A produces items

E_2 = machine B produces items

A = It is defective

$$P[E_1] = 0.45, P[E_2] = 0.55, P[A|E_1] = 0.08, P[A|E_2] = 0.10$$

Then by Bayes' theorem, probability that defective item was from machine A is given by

$$P(E_1/A) = \frac{P(E_1)P(A/E_1)}{\sum_{i=1}^{n=2} P(E_i)P(A/E_i)} = \frac{0.45 \times 0.08}{0.45 \times 0.08 + 0.55 \times 0.1} = \frac{0.036}{0.091} = 0.395$$

Example 5.13. A police radar gun is 98% accurate, that is it indicates that a car is speeding when the car is actually is with probability 0.98 and indicates that the car is not speeding when it is not with probability 0.98. Your teenager speeds 75% of the times. If she comes home and tells you that she got the ticket, what is the probability that she was speeding?

Solution: Let the events be

E_1 = car was speeding

E_2 = car was not speeding

A = she got the ticket

$$P[E_1] = 0.75, P[E_2] = 0.25, P[A|E_1] = 0.98, P[A|E_2] = 0.02$$

Then by Bayes' theorem, probability that she was speeding is given by

$$P(E_1/A) = \frac{P(E_1)P(A/E_1)}{\sum_{i=1}^{n=2} P(E_i)P(A/E_i)} = \frac{0.75 \times 0.98}{0.75 \times 0.98 + 0.25 \times 0.02} = \frac{0.735}{0.740} = 0.9932$$

Example 5.14. A doctor is to visit a patient and from past experience it is known that the probability that he will come by train, bus, or scooter are respectively 3/10, 1/5, and 1/10, the probability that he will come by some other means of transport being therefore 2/5. If he comes by a train, the probability that he will be late is 1/4, if by bus is 1/3 and if by scooter is 1/12. If he uses some other means of transport it can be assumed that he will not be late. Then

- what is the chance he will be late?
- when it is known that he arrived late, what is the probability that he comes by train?

Solution: Let the events be

E_1 : Doctor comes by train;

E_2 : Doctor comes by bus

E_3 : Doctor comes by scooter;

E_4 : Doctor comes by some other transport

A : He arrived late

- $P[\text{he will be late}] = P(A) = P[AE_1 \text{ or } AE_2 \text{ or } AE_3 \text{ or } AE_4]$

Since AE_1, \dots, AE_4 are mutually exclusive, we have

$$\begin{aligned} P(A) &= P[AE_1] + P[AE_2] + P[AE_3] + P[AE_4] \\ &= P[E_1]P[A/E_1] + P[E_2]P[A/E_2] + P[E_3]P[A/E_3] + P[E_4]P[A/E_4] \\ &= 18/120 = 0.15. \end{aligned}$$

- By Bayes' theorem

$$P[\text{he comes by train when it is known that he arrived late}] =$$

$$P(E_1/A) = \frac{P(E_1)P(A/E_1)}{\sum_{i=1}^{n=4} P(E_i)P(A/E_i)} = \frac{\frac{3}{10} \cdot \frac{1}{4}}{\frac{3}{10} \cdot \frac{1}{4} + \frac{1}{5} \cdot \frac{1}{3} + \frac{1}{10} \cdot \frac{1}{12} + \frac{2}{5} \cdot 0} = \frac{\frac{3}{40}}{\frac{18}{120}} = 0.5$$

5.9 Independence of Events

Let $A, B \in S$, the sample space with $P(B) > 0$, then by the multiplication rule we have

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

In many experiments, the information provided by B , does not affect the occurrence of event A , that is, $P(A|B) = P(A)$, and thus we have

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

Definition: Two or more events are said to be independent, if the occurrence of one event does not affect in any way the occurrence of the other events. Symbolically, if A_1, A_2, \dots, A_n are n independent events, then

$$P(A_1 \cap A_2 \cap \dots \cap A_n) = P(A_1) \times P(A_2) \times \dots \times P(A_n)$$

In particular, if two events A and B are two independent events then we have

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

For example,

- i) When two coins are tossed simultaneously, getting “Head(H)” on the first coin is independent of getting “head(H)” on the second coin,
- ii) Chance of hitting the target for the first time is independent of chance of hitting the same target for the second time,
- iii) Probability that of solving a mathematical problem is independent of solving another problem in mathematics,
- iv) Probability or the chance of getting male child for the first time is independent of getting male child for the second time, etc.

Note: If A and B are independent events, then $P(A \cap B) = P(A) \cdot P(B)$, which implies

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

and

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{P(A) \cdot P(B)}{P(A)} = P(B).$$

5.9.1. Theorems on independent events

Theorem : If A and B are independent events, then

- i) A and B^c , ii) A^c and B , and iii) A^c and B^c are also independent.

Proof: Since A and B are independent implies $P(A \cap B) = P(A) \cdot P(B)$ (*)

$$\begin{aligned} i. \quad P(A \cap B^c) &= P(A) - P(A \cap B) \Rightarrow P(A) - P(A) \cdot P(B) && \text{(from *)} \\ &= P(A)[1 - P(B)] = P(A) \cdot P(B^c) \end{aligned}$$

implies A and B^c are independent.

$$\begin{aligned} ii. \quad P(A^c \cap B) &= P(B) - P(A \cap B) \Rightarrow P(B) - P(A) \cdot P(B) && \text{(from *)} \\ &= P(B)[1 - P(A)] = P(B) \cdot P(A^c) \end{aligned}$$

implies A^c and B are independent.

$$\begin{aligned}
 \text{iii. } P(A^c \cap B^c) &= P(\overline{A \cup B}) = 1 - P(A \cup B) \Rightarrow 1 - P(A) - P(B) + P(A \cap B) \\
 &= 1 - P(A) - P(B) + P(A) \cdot P(B) \\
 &= [1 - P(A)] - P(B)[1 - P(A)] \\
 &= [1 - P(A)] \cdot [1 - P(B)] = P(A^c) \cdot P(B^c)
 \end{aligned}$$

implies, A^c and B^c are independent.

5.9.2. Pair-wise independence of events: Let A_1, A_2, \dots, A_n be n events defined on same sample space S , such that $P(A_i) > 0$; $i=1,2,\dots,n$. These events are said to be pair-wise independent if every pair of two events is independent.

Definition: The events A_1, A_2, \dots, A_n are said to be pairwise independent if and only if $P(A_i \cap A_j) = P(A_i) \cdot P(A_j)$, for $i \neq j = 1, 2, \dots, n$

In particular, if the events A_1, A_2, A_3 are pair-wise independent if and only if

$$\begin{aligned}
 P(A_1 \cap A_2) &= P(A_1) \cdot P(A_2) \\
 P(A_1 \cap A_3) &= P(A_1) \cdot P(A_3) \\
 P(A_2 \cap A_3) &= P(A_2) \cdot P(A_3)
 \end{aligned}$$

5.9.3 Mutually Independent Events. Let U be a family of events from S , the sample space. We say that the events U are pairwise independent if and only if, for every pair of distinct events $A, B \in U$,

$$P(AB) = P(A)P(B)$$

In other words, the n events A_1, A_2, \dots, A_n in a sample space S are said to be mutually independent if

$$P(A_{i_1} \cap A_{i_2} \cap \dots \cap A_{i_r}) = P(A_{i_1}) \cdot P(A_{i_2}) \cdots P(A_{i_r}), \text{ for } r = 1, 2, \dots, n$$

i.e., the n events are said to be mutually independent if they are independent by pairs, and by triplets, and by quadruples, and so on.

5.10 Examples on independence of events

Remark: Some rules

$$\text{i. } P(x: \text{at least one}) = P(x \geq 1) = 1 - P(x < 1) = 1 - P(x = 0) = 1 - P(\text{None}).$$

Thus in general,

$$P(x: \text{at least } k) = P(x \geq k) = 1 - P(x < k)$$

$$\text{ii. } P(x: \text{at most one}) = P(x \leq 1) = P(x = 0 \text{ or } 1), \text{ if } x = 0, 1, 2, \dots, k.$$

Thus in general,

$$P(x: \text{at most } k) = P(x \leq m) = P(x = 0 \text{ or } 1 \text{ or } 2 \text{ or } \dots \text{ or } m).$$

Example 5.15. Two sportsmen A and B hitting a certain target with probability $1/2$, and $3/8$ respectively. If a chance is given to them to hit a particular target what is the probability that i.

the target is hit? ii. At the most one of them hits the target? iii. A and B hit it? iv. None hit the target?

Solution: Given that, A hits the target with probability $1/2$, i.e., $P(A) = 1/2$, implies, $P(A') = 1 - P(A) = 1 - 1/2 = 1/2$ similarly, $P(B) = 2/5 \Rightarrow P(B') = 1 - P(B) = 1 - 2/5 = 3/5$

$$\begin{aligned} \text{i. } P(\text{the target is hit}) &= P(x:\text{at least one hits the target}) \\ &= P(x \geq 1) = 1 - P(x < 1) = 1 - P(x = 0) = 1 - P(\text{none hits the target} = A' B') \end{aligned}$$

Since events A, and B are independent implies A' & B' are also independent. Thus $P(\text{the target is hit}) = 1 - P(\text{none hits the target}) = 1 - P(A' B')$

$$\begin{aligned} &= 1 - P(A')P(B') \\ &= 1 - (1/2 * 3/5) \\ &= 1 - 3/10 = 7/10 = 0.7 \end{aligned}$$

Alternatively,

$$P(\text{the target is hit}) = P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Since events A, and B are independent, $P(A \cap B) = P(A)P(B)$ and thus we have

$$P(\text{the target is hit}) = P(A \cup B) = P(A) + P(B) - P(A)P(B) = 1/2 + 2/5 - (1/2)(2/5) = 0.7$$

$$\begin{aligned} \text{ii. } P(\text{At the most one of them hits the target}) &= P(x \leq 1) = P(x = 0 \text{ or } 1) \\ \Rightarrow P(\text{None or 1 hits it}) &= P(A' B' \text{ or } A' B \text{ or } A B') \\ &= P(A' B') + P(A' B) + P(A B') \\ &= P(A')P(B') + P(A')P(B) + P(A)P(B'), \text{ (since A \& B are independent)} \end{aligned}$$

$$\Rightarrow P(\text{None or 1 hits it}) = 1/2 * 3/5 + 1/2 * 2/5 + 1/2 * 3/5 = 13/10$$

$$\begin{aligned} \text{iii. } P(A \text{ and } B \text{ hit the target}) &= P(AB) \\ &= P(A)P(B), (\because \text{events A, and B are independent}) \\ &= 1/2 * 2/5 = 1/5 \end{aligned}$$

$$\begin{aligned} \text{iv. } P(\text{None hit it}) &= P(A' B') = P(A')P(B'), \text{ (since A, \& B are independent} \Rightarrow A' \& B' \text{ are also independent)} \\ \Rightarrow P(\text{None hit it}) &= 1/2 * 3/5 = 3/10 \end{aligned}$$

Example 5.16: The odds against the wife who is 45 years old survives till she is 75 is 7:4 and her husband now 50 years who survives till he is 80 is 5:4. Find the probability that

- i. Both will be alive
- ii. At least one of them will be alive
- iii. One of them will be alive, for 30 more years

Solution: let the events A and B be

A: wife will be alive for 30 more years

B: Husband will be alive for 30 more years

Then $P(A) = 4/(7+4) = 4/11$ and $P(B) = 4/(5+4) = 4/9$. Therefore,

$$P(A') = 1 - P(A) = 7/11, \text{ and } P(B') = 1 - P(B) = 5/9.$$

We assume that the survival of an individual is independent of the other, and hence

i. $P(\text{Both will be alive}) = P(A \cap B) = P(A)P(B) = 4/11 * 4/9 = 16/99$, (since A and B are independent).

ii. $P(\text{at least one them will be alive}) = 1 - P(\text{None alive})$
 $= 1 - P(A' B')$

A and B are independent $\Rightarrow A'$ and B' are also independent, implies
 $= 1 - P(A')P(B') = 1 - (7/11 * 5/9)$
 $= 1 - 35/99 = 64/99$.

iii. $P(\text{One of them will be alive}) = P(A' B \text{ or } A B') = P(A' B) + P(A B')$

Since A and B are independent $\Rightarrow A' \& B$ and $A \& B'$ are also independent, implies
 $= P(A')P(B) + P(A)P(B')$,
 $= 4/11 * 5/9 + 7/11 * 4/9 = 48/99$.

Example 5.17: An electric circuit of a system contains three components, each work independently with probability 0.85, 0.8 and 0.9. System works if all components work, then what is the probability that the system works?

Solution: Let the events be

A : component 1 works; B : component 2 works, and C : component 3 works. Then

$P(A) = 0.85$, implies, $P(A') = 1 - P(A) = 1 - 0.85 = 0.15$

similarly, $P(B) = 0.8 \Rightarrow P(B') = 1 - P(B) = 1 - 0.8 = 0.2$ and $P(C') = 0.10$

$P(\text{the system works}) = P(A \cap B \cap C) = 1 - P(A' B' C')$.

Since all the 3 components work independently, i.e., A, B, C are independent and therefore $A', B' \& C'$ are also independent, therefore

$P(\text{the system works}) = 1 - P(A')P(B')P(C') = 1 - (0.15 * 0.20 * 0.10) = 1 - 0.003 = 0.997$.

Example 5.18: Three students A, B and C , whose chances of solving a problem in Mathematics are $1/2, 3/5$ and $2/3$ respectively. If a problem is given to them and suppose all of them try independently then what is the probability that

- i. the problem will be solved?
- ii. A and B solve it?
- iii. None of them solve it?

Solution: Let the events be

A : Student A solves the problem; B : Student B solves the problem and C : Student C solves the problem. Then,

$P(A) = 1/2$, implies, $P(A') = 1 - P(A) = 1 - 1/2 = 1/2$

similarly, $P(B) = 3/5 \Rightarrow P(B') = 1 - P(B) = 1 - 3/5 = 2/5$ and $P(C') = 1/3$

Now,

- i. $P(\text{the problem is solved}) = P(A \cup B \cup C) = P(\text{at least one solves it})$
 $= 1 - P(\text{none solve it}) = 1 - P(A' B' C')$.

Since A, B and C are independent, implies $A', B' \& C'$ are also independent, thus we have,

$P(\text{the problem is solved}) = 1 - P(A')P(B')P(C') = 1 - (1/2 * 2/5 * 1/3) = 1 - 1/15 = 14/15 = 0.933.$

ii. $P(A \text{ and } B \text{ solve it}) = P(ABC') = P(A) * P(B) * P(C')$
 $= 1/2 * 3/5 * 1/3 = 1/10 = 0.1,$ (since A, B and C are independent)

iii. $P(\text{none solve it}) = P(A'B'C') = P(A')P(B')P(C') = 1/2 * 2/5 * 1/3 = 1/15 = 0.037.$

Objective Questions

- Limits of probability of an event is
 a) $(-1, 0),$ b) $(0, 1)$ c) $(-1, 1)$ d. None
- Probability of the sample space is
 a) 0 b) 1 c) -1 d) ± 1
- For any two mutually exclusive events A and $B,$
 a) $A \cap B = S$ b) $A \cap B = A$ c) $A \cap B = \phi$ d) $A \cup B = S$
- If $P(A) = 0.8, P(B) = 0.4$ and $P(A \cap B) = 0.5,$ then $P(A \cup B)$ is
 a) 0.6 b) 0.75 c) 0.8 d) 0.7
- If $P(A) = 0.9, P(A \cap B) = 0.6,$ and $P(A \cup B) = 0.4,$ then $P(B)$ is
 a) 0.5 b) 0.7 c) 0.1 d) 0.2

Exercise

- A box has 9 tickets marked with numbers 1,2, 3,...,9. Two tickets are drawn at random from the box. Find the probability that both the numbers drawn are even or odd.
- In Bangalore city 25% people read the news paper X and 40% read the news paper Y and 15% read people read both the news papers. Find the probability that a randomly selected person read at least one of these news papers.
- Four cards are drawn from a pack of playing cards. What is the probability that i. All are diamonds, ii. One card of each suit, iii. There are two spades and two hearts.
- What is the chance that a leap year selected will contain 53 Sundays?
- Two chess players A and B play 10 games of chess of which 6 are won by A and 3 are won by B, and one in a tie. They agree to play tournament of three games. Find the probability that i. A win all the three games, ii. Two game end in a tie, iii. A and B win alternatively iv. B wins at least one game
- Two fair dice are rolled once. Find the probability that the sum of the numbers obtained is i. 7 or 9 ii. Less than 10, iii. Sum is divisible by 3 and 4, iv. The sum of the numbers exceeds to 5, and v. Both numbers obtained are even.
- If the letters of the word "REGULATIONS", be arranged at random, what is the chance that there will be exactly 4 letters between R and E
- What is the probability that four S's come consecutively in the word "MISSISSIPPI"?
- If the letters of the word "MATHEMATICS" be arranged at random, find the probability that all 'vowels' come together.
- A box contains 10 floppy disks, 3 of which are defective, 3 disks are drawn at random from this box without replacement. Events A and b are defined as follows:

A: At least 2 of the floppy disks that are drawn are defective

B: At least 1 of the floppy disks is defective

Then what is i) $P(A)$? ii) $P(B)$? iii. $P(A \cap B)$? and iv) are A&B independent?

11. A problem in statistics is given to 3 students A, B, and C whose chance of solving it are $\frac{1}{2}$, $\frac{3}{4}$ and $\frac{1}{4}$ respectively. What is the probability that the problem will be solved?
12. A, B, and C are independent witnesses of an event which is known to have occurred. A speaks truth 3 times, out of 4, B 4 times out of 5, C 5 times out of 6. What is the probability that the occurrence will be reported truthfully by majority of 3 witnesses?
13. A can solve 80% of the problems given in a text book; B can solve 70% of the problems of the text book. If a problem is given to both and they try to solve it separately, find the probability that the a) problem is solved, b) problem is not solved.
14. Four persons A, B, C and D are able to hit the a target 8, 4, 5, and 5 times respectively with 10 shots each. If each of them fires at the target once, what is the probability i) that the target is hit? ii) any two of them hit it iii) A, B, C hit it but not D?
15. The odds favouring the survival of a man aged 60 for 20 more years are 2 to 6, and that of a women aged 55 for 20 more years are 3 to 5. What is the probability that
a) A man aged 60 & his wife aged 55 will survive for 20 more years?
b) at least one of them will survive?

UNIT 6

RANDOM VARIABLE AND PROBABILITY DISTRIBUTIONS

6.1 Objectives

The aim is to introduce the concept of random variable (r.v.) and probability distribution of an r.v. Concept of random variable technique is a mapping from the sample space (S) to the real line and hence to introduce the induced probability measure.

6.2 Introduction

In the previous chapter we have discussed about the assignment and computation of probabilities of occurrence of events. But in several real experiments we may be more interested in how many times an event has occurred or happened, not just in knowing which outcome has occurred, i.e., the number(s) associated with them. For example, when n fair coins are tossed simultaneously, when a pair of dies is rolled, here one may be interested in knowing the number of times the head appears, the sum of the points obtained on dies respectively. Thus, we associate a real number with each outcome of the experiment. That is, we consider a function whose domain is the set of possible outcomes, and whose range is a subset of the set of real and such a function is called a random variable.

That is more precisely, consider a coin tossing experiment. Suppose a coin is tossed once, and let X : denote the number of head occurs, then the variable X takes real values say, ‘1’ if head(H) occurs, and ‘0’(say), when tail(T) appears. Symbolically,

$$X = \begin{cases} 1, & \text{if head}(H) \text{ appears} \\ 0, & \text{if tail}(T) \text{ appears} \end{cases}$$

Here, the sample space(S) = $\{H, T\} \sim \{1, 0\}$, where ‘1’ for Head and ‘0’ for Tail. Thus, X is a random variable, i.e. X is a real valued function defined on the sample space S , which takes us from the sample space S to a space of real numbers $\mathfrak{R} = \{x; x = 0, 1\}$.

Definition: *The random variable(r.v.) is a real valued function defined on sample space(S)of a random experiment.*

Note: More rigorously, in the probability space, the triplet(S, Ω, P), where S is the sample space, Ω is the σ -field of subsets in S , and P is a probability function on Ω , so that the random variable is then a function $X(w)$ with domain S and range $(-\infty, \infty)$ such that for every real number a , the event $[w; X(w) \leq a] \in \Omega$, where w indicate the outcome of a random experiment.

6.3 Types of random variable

“Let S be the sample space associated with a given random experiment. A real valued function $X(w)$ defined on S and taking values in $\mathfrak{R}(-\infty, \infty)$ is called a *one dimensional random variable*. If the function values are ordered pairs of real numbers(i.e., vectors in two space) the function is called a *two dimensional random variable*. In general, an n -dimensional random variable is simply a function whose domain in S , and whose range is a collection of n -tuples of real numbers(vectors in n -space)”.

Further we define types of random variables as

- i. Discrete random variable
- ii. Continuous random variable

6.3.1 Discrete random variable: A random variable if it assumes finite (say' $x: 0, 1, 2, \dots, n (< \infty)$) or finitely large number (say' $x: 0, 1, 2, \dots, \infty$) of distinct or dissimilar values, then it is said to be discrete or it is a real valued function defined on a discrete sample space.

For eg. Number of accidents occurring in a city in a day; number of defective electric lamps in a lot of 100 such lamps, number of absentees in a class of 50 students, number of television(TV) sets sold at a super market, etc., are discrete.

6.3.2 Continuous random variable: A random variable if it assumes infinitely large number of values (both integral and fractional) within some specified domain, then it would be called as continuous random variable. For eg., weight, height, age, length, temperature, etc., are continuous. Symbolically, $\{x: a < x \leq b \text{ or } a \leq x < b \text{ etc., where } a, b \in \mathfrak{R}, \text{ real line}\}$, then X is a continuous random variable.

6.4 Distribution function(D.F.) or cumulative distribution function of an R.V.

Let X be random variable. The function F defined for all real values of x by

$$F(x) = P(X \leq x) = P\{w : X(w) \leq x\}, -\infty < x < \infty,$$

is called the distribution function (d.f.) or cumulative distribution function of the random variable X .

Note 1. If F is the d.f. of one dimensional r.v. X , then i) $0 \leq F(x) \leq 1$. ii). $F(x) \leq F(y)$ if $x < y$. That is, all d.f.'s are monotonically non-decreasing and lie between(0, 1).

Note 2:
$$F(x) = P[X \leq x] = \begin{cases} \sum_{i=a}^x p(i), & a \in \mathfrak{R}, \text{ when } X \text{ is discrete} \\ \int_{-\infty}^x f(t)dt, & \text{when } X, \text{ is continuous} \end{cases}$$

Properties of Distribution Function(d.f.)

Here we provide the properties of distribution function, which are common to all distribution functions.

- i. If f is the d.f. of the r.v. X and if $a < b$, then $P(a < X \leq b) = F(b) - F(a)$
[Hint: $P(a < X \leq b) + P(X \leq a) = P(X \leq b)$].
- ii. $P(a \leq X \leq b) = P(X = a) + F(b) - F(a)$
- iii. If f is the d.f. of one dimensional r.v. X , then
$$F(-\infty) = \lim_{x \rightarrow -\infty} F(x) = 0 \text{ and } F(\infty) = \lim_{x \rightarrow \infty} F(x) = 1.$$

Example 6.1: When a coin is tossed once, the sample space $S = \{H, T\}$, and X be defined by $X(H) = 1$ and $X(T) = 0$. If p assigns equal probability to $\{H\}$ and $\{T\}$, then $P\{X = 0\} = 1/2 = P\{X = 1\}$, and the distribution function is

$$F(x) = \begin{cases} 0, & \text{if } x < 0 \\ \frac{1}{2}, & \text{if } 0 \leq x < 1 \\ 1, & \text{if } x \geq 1 \end{cases}$$

Example 6.2: Let X be a continuous r.v. with probability function $f(x) = \begin{cases} 2x, & 0 < x < 1 \\ 0, & \text{otherwise} \end{cases}$,

Then, $P[X \leq x] = \int_{-\infty}^x f(t)dt = \int_0^x f(t)dt = \int_0^x 2t dt = 2 \frac{t^2}{2} \Big|_0^x = x^2$, for $0 < x < 1$,

and the distribution function F is then

$$F(x) = \begin{cases} 0, & \text{if } x < 0 \\ x^2, & \text{if } 0 \leq x < 1 \\ 1, & \text{if } x \geq 1 \end{cases}$$

6.5 Probability distribution of a random variable

Probability distribution of a random variable is defined as a real valued function such that every value of $X = x$ is associated with the probability of occurrence of an outcome of a random experiment such that the total probability should be equal to 1.

For eg., when a coin is tossed twice, the sample space $S = \{HH, HT, TH, TT\}$ contains four outcomes. Here, X : denote the number of head occurs, then we have the variable X takes values x : 0, 1, 2 such that ‘0’ for “No head = $\{TT\}$ ”, ‘1’ for ‘one head = $\{HT, TH\}$ ’, and ‘2’ for ‘two heads = $\{HH\}$ ’, with respective probability 1/4, 1/2 and 1/4. Symbolically, we denote

$$\begin{array}{l} X=x: \quad 0 \quad 1 \quad 2 \\ P(X=x): 1/4 \quad 1/2 \quad 1/4 \end{array}$$

And an another example is that suppose a dice is rolled once, then X be the number obtained on the dice is a random variable (discrete) such that its probability distribution is given by

$$\begin{array}{l} X=x: \quad 1 \quad 2 \quad 3 \quad 4 \quad 5 \quad 6 \\ P(X=x): 1/6 \quad 1/6 \quad 1/6 \quad 1/6 \quad 1/6 \quad 1/6 \end{array}$$

6.5.1 Probability mass function (pmf): A probability function $P(X = x)$ of a discrete random variable say X is said to be probability mass function (pmf) if

$$p(x) = \begin{cases} P(X = x_i) = p_i, & \text{if } x = x_i \\ 0, & \text{if } x \neq x_i; i = 1, 2, \dots \end{cases}$$

is called the probability mass function of the r.v. X .

Properties of $p(x)$

- i. $p(x) \geq 0$, for all $x \in \mathcal{R}$,
- ii. $\sum_{\forall x} p(x) = 1$, for all $x \in \mathcal{R}$.

6.5.2 Discrete distribution function (d.f. or cdf): Let X be a discrete r.v. having the pmf $p(x)$ at countable number of points x_1, x_2, \dots and number $p_i \geq 0$; $\sum_{\forall i} p_i = 1$, such that $F(x) = \sum_{i: x_i \leq x} p_i$, then

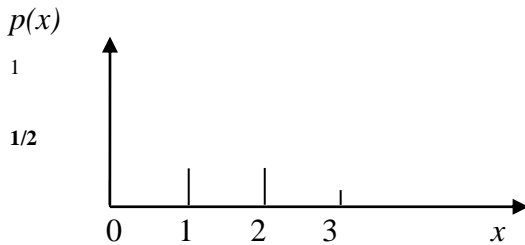
$F(x)$ has a “step function” having jump p_i at i , and being constant between each pair of integers.

For example: When a coin is tossed thrice, the sample space $S = \{HHH, HHT, HTH, THH, HTT, THT, TTH, TTT\}$, and X be defined by

X : denote the number of head appears, and then the probability distribution is

$X:$	0	1	2	3
$P(x):$	1/8	3/8	3/8	1/8

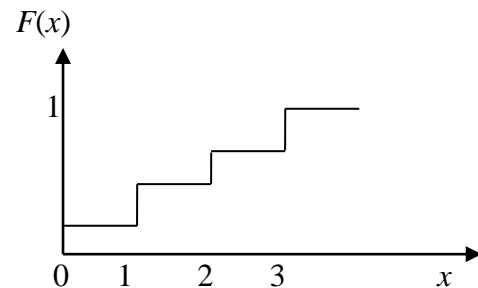
And the probability function (pmf) curve is



and the distribution function is

$$F(x) = \begin{cases} 0, & \text{if } x < 0 \\ 1/8, & \text{if } 0 \leq x < 1 \\ 4/8, & \text{if } 1 \leq x < 2 \\ 7/8, & \text{if } 2 \leq x < 3 \\ 1, & \text{if } x \geq 3 \end{cases}$$

and the curve of $F(x)$ is:



Example 6.3. If $p(x) = \begin{cases} \frac{x}{15}, & x = 1, 2, 3, 4, 5 \\ 0, & \text{elsewhere} \end{cases}$

Find i. $P\{X = 1 \text{ or } 2\}$, and ii. $P\{0.5 < X < 2.5 \mid X > 1\}$

Solution: We have,

i. $P\{X=1 \text{ or } 2\} = P\{X = 1\} + P\{X = 2\} = \frac{1}{15} + \frac{2}{15} = \frac{3}{15} = 0.2$

ii.
$$P\{0.5 < X < 2.5 \mid X > 1\} = \frac{P\{(0.5 < X < 2.5) \cap X > 1\}}{P\{X > 1\}}$$

$$= \frac{P\{(x = 1 \text{ or } 2) \cap X > 1\}}{1 - P\{X \leq 1\}}$$

$$= \frac{P\{x = 2\}}{1 - P\{X = 1\}} = \frac{2/15}{1 - 1/15} = \frac{1}{7}$$

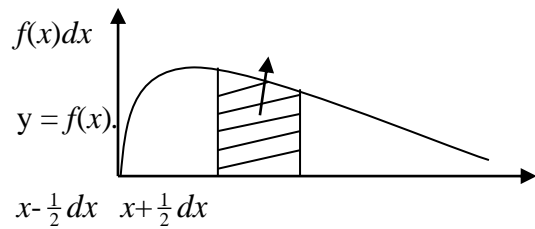
6.5.3 Probability density function(pdf): A probability function $f(X = x)$ of a continuous random variable say X , is said to be probability density function(pdf), if it satisfies the properties:

- i. $f(x) \geq 0$, for all $x \in \mathfrak{R}$, and
- ii. $\int_x f(x) dx = 1, \forall x \in \mathfrak{R}$, (or) $\int_{-\infty}^{\infty} f(x) dx = 1$
- iii. The probability $P(A)$ given by $\int_A f(x) dx$, is well defined for any event A .

Note: More rigorously, the pdf $f(x)$ of the random variable X is defined as

$$f_x(x) = \lim_{\delta x \rightarrow 0} \frac{P(x \leq X \leq x + \delta x)}{\delta x}, \text{ where } \delta x \text{ is the}$$

small increment in x ; i.e., pdf $f(x)$ of the random variable X is a continuous function of x , such that $f(x)dx$ represents the probability that X falls in the infinitesimal interval $(X, X + dx)$ or $f(x)dx$ represents the *area bounded* by the curve $y = f(x)$.



Remark: When the random variable X is discrete, the probability at a point c , or $P(X = c)$ is not zero for some fixed c . However when X is a continuous random variable the probability at a point is always zero, i.e., $P(X = c) = 0$, for all c . This implies that $P(A) = 0$, does not imply that the event A is null or impossible event. The property of continuous r.v., viz.,

$P(X = c) = 0$, for all c , leads to the following important result:

$$P(a \leq X \leq b) = P(a \leq X < b) = P(a < X \leq b) = P(a < X < b),$$

i.e., in the case of continuous r.v., it does not matter whether we include the end points of the interval from (a, b) , however, this result is not true in general for discrete r.v.

Probability distribution function of a continuous random variable

Let X be a continuous r.v. having the pdf $f(x)$, then the function $F(x)$ defined by

$$F(x) = P[X \leq x] = \int_{-\infty}^x f(t) dt, -\infty < x < \infty$$

is called the distribution function(df) or the cumulative df of the random variable X .

Properties of Distribution Function:

- j. $0 \leq F(x) \leq 1, -\infty < x < \infty$.
- ii. From analysis(Reimann integral), we know that

$$F'(x) = \frac{d}{dx} F(x) = f(x) \geq 0$$

$\Rightarrow F(x)$ is non-decreasing function of x

$$\text{iii. } F(-\infty) = \lim_{x \rightarrow -\infty} F(x) = \lim_{x \rightarrow -\infty} \int_{-\infty}^x f(x)dx = \int_{-\infty}^{-\infty} f(x)dx = 0$$

$$F(+\infty) = \lim_{x \rightarrow \infty} F(x) = \lim_{x \rightarrow \infty} \int_{-\infty}^x f(x)dx = \int_{-\infty}^{+\infty} f(x)dx = 1$$

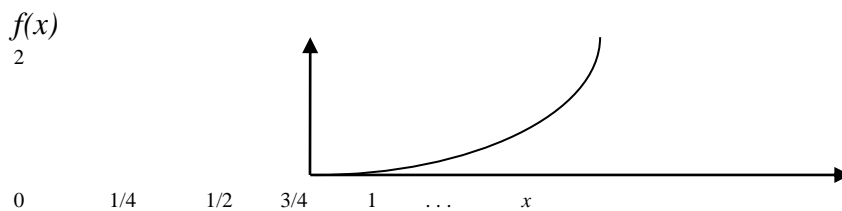
iv. $F(x)$ is right continuous function of x .

Example 6.4: Let X be a continuous r.v. with probability function $f(x) = \begin{cases} 2x, & 0 < x < 1 \\ 0, & \text{otherwise} \end{cases}$,

Then, the probability distribution function is given by

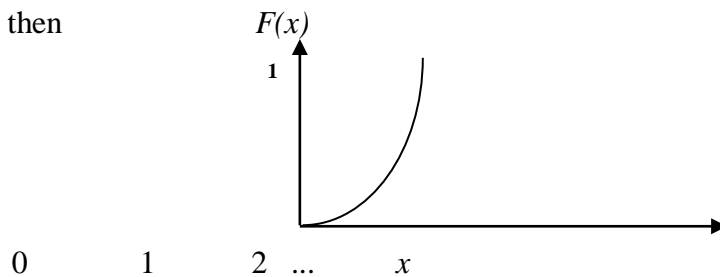
$$P[X \leq x] = \int_{-\infty}^x f(t)dt = \int_0^x f(t)dt = \int_0^x 2t dt = 2 \frac{t^2}{2} \Big|_0^x = x^2, \text{ for } 0 < x < 1$$

And the curve for pdf is given by



and the distribution function F is then

$$F(x) = \begin{cases} 0, & \text{if } x < 0 \\ x^2, & \text{if } 0 \leq x < 1 \\ 1, & \text{if } x \geq 1 \end{cases}$$



6.5.4 Various measures of central tendency, dispersion, skewness and kurtosis in terms of random variable

Let $f(x)$ be the pdf of a continuous random variable X , such that X is defined in (a, b) , then we have,

i. Arithmetic mean = mean(X) = $\int_a^b xf(x)dx$

ii. Geometric mean(G) is given by $\log G = \int_a^b \log x \cdot f(x)dx$

iii. Harmonic mean(H) is obtained by: $\frac{1}{H} = \int_a^b \frac{1}{x} f(x)dx$

iv. Median(M) is given by solving $\int_a^M f(x)dx = \frac{1}{2} = \int_M^b f(x)dx$, as median divides the total area in to two equal parts.

v. Mode(Z) is obtained by solving $f'(x) = 0$, and $f''(x) < 0$, provided it lies between [a, b], since mode is the value of x , for which $f(x)$ is maximum.

vi. Quartiles($Q_r, r = 1, 2, 3$), Deciles($D_r, r = 1, 2, \dots, 9$) and Percentiles($P_r, r = 1, 2, \dots, 99$) are obtained by

$$\int_a^{Q_r} f(x)dx = r \frac{1}{4}, \text{ where } r = 1, 2, 3 \text{ for quartiles}$$

$$\int_a^{D_r} f(x)dx = r \frac{1}{10}, \text{ where } r = 1, 2, \dots, 9 \text{ for deciles}$$

$$\text{And } \int_a^{P_r} f(x)dx = r \frac{1}{100}, \text{ where } r = 1, 2, \dots, 99 \text{ for deciles}$$

vii. Moments about origin μ'_r (about origin) = $\int_a^b x^r f(x)dx$,

$$\text{in particular, mean } \mu'_1 = \int_a^b xf(x)dx, \mu'_2 = \int_a^b x^2 f(x)dx, \text{ and}$$

$$\mu'_r \text{ (about any point A)} = \int_a^b (x - A)^r f(x)dx$$

$$\mu_r \text{ (about mean)} = \int_a^b (x - \text{mean})^r f(x)dx$$

viii. Mean deviation(M.D.) about mean μ'_1 is given by

$$\text{M.D. (about mean)} = \int_a^b |x - \text{mean}| f(x)dx$$

6.5.5 Important remark: Above measures can even be applied to discrete random variable with pmf $p(x)$, for which, we need to replace integral (\int) sign by summation (Σ) sign over the given range of the variable X , in the above formulae i to viii.

Example 6.5: A continuous random variable X has a pdf $f(x) = \begin{cases} kx^2, & \text{if } 0 \leq x \leq 1 \\ 0, & \text{elsewhere} \end{cases}$,

find i. k , ii. mean, iii. median, iv. mode, v. $P(X > 1/2)$, and vi. $P(0.25 < X < 0.75)$.

Solution:

i. To find k , we have

$$\int_0^1 f(x) dx = 1 \Rightarrow \int_0^1 kx^2 dx = 1 \Rightarrow k \int_0^1 x^2 dx = k \left. \frac{x^3}{3} \right|_0^1 = 1 \Rightarrow k \left(\frac{1}{3} - 0 \right) = 1 \Rightarrow k = 3$$

$$\Rightarrow f(x) = \begin{cases} 3x^2, & \text{if } 0 \leq x \leq 1 \\ 0, & \text{elsewhere} \end{cases}$$

$$\text{ii. Arithmetic mean} = \int_0^1 xf(x) dx \Rightarrow \int_0^1 x \cdot 3x^2 dx \Rightarrow 3 \int_0^1 x^3 dx = 3 \left. \frac{x^4}{4} \right|_0^1 \Rightarrow 3 \left(\frac{1}{4} - 0 \right) = 3/4$$

iii. To find median, we have

$$\int_0^M f(x) dx = \frac{1}{2} = \int_M^1 f(x) dx$$

$$\text{Consider, } \int_0^M f(x) dx = \frac{1}{2} \Rightarrow \int_0^M 3x^2 dx = \frac{1}{2}$$

$$\Rightarrow 3 \int_0^M x^2 dx = 3 \left. \frac{x^3}{3} \right|_0^M = \frac{1}{2} \Rightarrow M^3 = \frac{1}{2} \Rightarrow M = (1/2)^{1/3} = 0.7937$$

or

$$\int_M^1 f(x) dx = \frac{1}{2} \Rightarrow \int_M^1 3x^2 dx = \left. \frac{3x^3}{3} \right|_M^1 = \frac{1}{2} \Rightarrow (1 - M^3) = \frac{1}{2} \Rightarrow M = (1/2)^{1/3} = 0.7937$$

$$\Rightarrow \text{median}(M) = 0.7937$$

iv. To find Mode(Z), we have

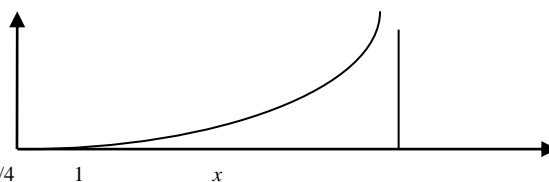
$$f'(x) = 0 \Rightarrow \frac{d}{dx} f(x) = \frac{d}{dx} (3x^2) = 6x = 0 \Rightarrow x = 0$$

$$\text{and } f''(x) < 0 \Rightarrow \frac{d^2}{dx^2} f(x) = \frac{d}{dx} \left(\frac{d(3x^2)}{dx} \right) = 6 > 0, \text{ which contradicts that } f(x) \text{ is maximum.}$$

Therefore, calculus method fails. Hence, we use graphical method to find mode

f(x)

3



Thus, from the graph it is observed that f(x) attains maximum at x = 1, hence Mode(Z) = 1.

$$\text{v. } P(X > 1/2) = \int_{1/2}^1 f(x) dx = \int_{1/2}^1 3x^2 dx = x^3 \Big|_{1/2}^1 = (1 - 1/8) = 7/8$$

$$\text{and, vi) } P(0.25 < X < 0.75) = \int_{0.25}^{0.75} f(x) dx = \int_{0.25}^{0.75} 3x^2 dx = x^3 \Big|_{0.25}^{0.75} = (0.75^3 - 0.25^3)$$

$$= 0.421875 - 0.015625 = 0.40625 = 13/32$$

Example 6.6: Find i. k , and ii. compute $P(X < 1.25)$, for x a continuous r.v., with pdf $f(x)$ defined by

$$f(x) = \begin{cases} kx, & 0 \leq x \leq 1 \\ k, & 1 \leq x \leq 2 \\ -kx + 3k, & 2 \leq x \leq 3 \\ 0, & \text{elsewhere} \end{cases}$$

Solution: i. To determine the value of k , we have

$$\int_{-\infty}^{\infty} f(x) dx = 1 \Rightarrow \int_0^1 f(x) dx + \int_1^2 f(x) dx + \int_2^3 f(x) dx = 1$$

$$\Rightarrow \int_0^1 kx dx + \int_1^2 k dx + \int_2^3 (-kx + 3k) dx = 1$$

$$\Rightarrow k \left. \frac{x^2}{2} \right|_0^1 + kx \Big|_1^2 - k \left. \frac{x^2}{2} \right|_2^3 + 3kx \Big|_2^3 = 1$$

$$\Rightarrow \frac{k}{2}(1-0) + k(2-1) - \frac{k}{2}(9-4) + 3k(3-2) = 1$$

On simplification, $k = 1/2 \Rightarrow f(x) = \begin{cases} x/2, & 0 \leq x \leq 1 \\ 1/2, & 1 \leq x \leq 2 \\ -x/2 + 3/2, & 2 \leq x \leq 3 \\ 0, & \text{elsewhere} \end{cases}$

ii. $P(X < 1.25) = \int_{-\infty}^{1.25} f(x) dx \Rightarrow \int_0^1 f(x) dx + \int_1^{1.25} f(x) dx$

$$\Rightarrow \int_0^1 f(x) dx + \int_1^{1.25} f(x) dx = \int_0^1 \frac{x}{2} dx + \int_1^{1.25} \frac{1}{2} dx = \left. \frac{x^2}{4} \right|_0^1 + \left. \frac{x}{2} \right|_1^{1.25} = \frac{1}{4} + \frac{1}{8} = \frac{3}{8}$$

6.6 Two dimensional random variables

Here we deal with two dimensional random variable defined on the same sample space. For example, one may be interested in getting the information about height and weight of each individual from a certain organisation. To describe such experiments mathematically we introduce the study of two random variables.

Definition: Let X and Y be two random variables defined on the same sample space S , then the function (X, Y) that assigns a point in \mathbb{R}^2 (i.e. $\mathbb{R} \times \mathbb{R}$), is called a two dimensional random variable.

Note: A two-dimensional r.v. is said to be discrete if it takes at most countable number of points in \mathbb{R}^2 (i.e. $\mathbb{R} \times \mathbb{R}$).

Note: Two random variables X and Y are said to be jointly distributed if they are defined on the same probability space.

6.7 Two dimensional or joint distribution function

Definition: The distribution function F , of the two dimensional random variable (X, Y) is a real valued function, defined for all real x and y by the relation:

$$F_{XY}(x, y) = P(X \leq x, Y \leq y).$$

Properties of joint distribution function

i. $F(-\infty, y) = F(x, -\infty) = 0; F(+\infty, +\infty) = 1.$

ii. If the density function $f(x, y)$ is continuous at (x, y) , then $\frac{\partial^2 F(x, y)}{\partial x \partial y} = f(x, y).$

6.8 Marginal and conditional distribution functions

Here we determine the marginal and conditional distribution functions with respect to joint distribution function $F_{XY}(x, y)$.

With the notion of joint distribution function $F_{XY}(x, y)$, we obtain the individual distribution functions $F_X(x)$ and $F_Y(y)$ called *marginal probability functions defined by*,

$$F_X(x) = P(X \leq x) = P(X \leq x, Y \leq \infty) = \lim_{y \rightarrow \infty} F_{XY}(x, y) = F_{XY}(x, \infty)$$

Similarly,

$$F_Y(y) = P(Y \leq y) = P(X < \infty, Y \leq y) = \lim_{x \rightarrow \infty} F_{XY}(x, y) = F_{XY}(\infty, y)$$

Where, $F_X(x)$ and $F_Y(y)$ are the marginal distribution functions of X and Y respectively.

Note: If (X, Y) discrete then

$$F_X(x) = \sum_y P(X \leq x, Y = y) \text{ and } F_Y(y) = \sum_x P(X = x, Y \leq y)$$

And when (X, Y) continuous then

$$F_X(x) = \int_{-\infty}^x \left(\int_{-\infty}^{\infty} f_{XY}(x, y) dy \right) dx \text{ and } F_Y(y) = \int_{-\infty}^y \left(\int_{-\infty}^{\infty} f_{XY}(x, y) dx \right) dy$$

With the notion of joint distribution function $F_{XY}(x, y)$, we obtain the distribution functions $F_{X|Y}(x|y)$ and $F_{Y|X}(y|x)$ called *conditional distribution functions for given values of a variable defined by*,

$$F_{X|Y}(x|y) = P(X \leq x, Y = y), \text{ and } F_{Y|X}(y|x) = P(Y \leq y, X = x)$$

Where $F_{X|Y}(x|y)$ is called conditional distribution of X , for given values of Y , and $F_{Y|X}(y|x)$ is called conditional distribution of Y , for given values of X .

6.9 Two dimensional or Joint probability mass function

If (X, Y) is a two dimensional discrete random variable, then the joint discrete probability function of X, Y also called the joint probability mass function of X, Y denoted by $p_{xy}(x, y)$ is defined as

$$p_{xy}(x_i, y_j) = \begin{cases} P(X = x_i, Y = y_j), & \text{for a value of } (x_i, y_j) \text{ of } (X, Y) \\ 0, & \text{elsewhere} \end{cases}$$

such that, $\sum_x \sum_y p_{xy}(x_i, y_j) = 1, \forall (x_i, y_j)$ and $p_{xy}(x_i, y_j) \geq 0.$

6.10 Two dimensional or Joint probability density function

b) Similarly, when X and Y are continuous random variables, then the conditional probability function of X for given value of $Y=y$ is given by

$$f(x/y) = \frac{f_{xy}(x,y)}{f_y(y)}, \quad -\infty \leq x_i \leq \infty; \forall y$$

and the conditional probability function of Y for given value of $X=x$ is given by

$$f(y/x) = \frac{f_{xy}(x,y)}{f_x(x)}, \quad -\infty \leq y \leq \infty; \forall x$$

6.14 Independence of random variables: Let (X, Y) be a two dimensional random variable with joint pdf $f_{xy}(x, y)$. Then X and Y are said to be independent if

$$f_{xy}(x, y) = f_x(x) \times f_y(y).$$

Example 6.7. The joint probability mass function of the two discrete random variables X and Y is given by

$$p(x, y) = \frac{2x+y}{27}, \quad x = 0, 1, 2; y = 0, 1, 2$$

Obtain i. marginal probability function of X and Y ; ii. Conditional distribution function of X given $Y = 1$ and Y given $X = 2$. iii. $P(1 \leq X \leq 2)$, $P(|X| < 1)$

Solution: i. To obtain marginal probability function of X and Y , we have

$$\begin{aligned} p_x(x) &= \sum_y p_{xy}(x_i, y_j), \quad \forall x_i, y_j \\ &= \sum_{y=0}^2 \frac{2x+y}{27} = \frac{1}{27} [(2x+0) + (2x+1) + (2x+2)] = \frac{6x+3}{27} = \frac{2x+1}{9}, \quad x = 0, 1, 2. \end{aligned}$$

$$\begin{aligned} p_y(y) &= \sum_x p_{xy}(x_i, y_j), \quad \forall y_j \\ &= \sum_{x=0}^2 \frac{2x+y}{27} = \frac{1}{27} [(0+y) + (2 \cdot 1 + y) + (2 \cdot 2 + y)] = \frac{6+3y}{27} = \frac{2+y}{9}, \quad y = 0, 1, 2. \end{aligned}$$

ii. Now to find the conditional distribution of X given $Y = 1$, we have

$$\begin{aligned} p(x/y) &= \frac{p_{xy}(x, y)}{p(Y=1)}, \quad \forall x \\ &= \frac{\binom{2x+y}{27}}{\binom{2+y}{9}} = \frac{1}{3} \left(\frac{2x+y}{2+y} \right) \Bigg|_{y=1} = \frac{1}{3} \times \frac{2x+1}{2+1} = \frac{2x+1}{9}, \quad x = 0, 1, 2 \end{aligned}$$

And, to find the conditional distribution of Y given $X = 2$, we have

$$\begin{aligned} p(y/x) &= \frac{p_{xy}(x, y)}{p(X=2)}, \quad \forall y \\ &= \frac{\binom{2x+y}{27}}{\binom{2x+1}{9}} = \frac{1}{3} \left(\frac{2 \times 2 + y}{2 \times 2 + 1} \Bigg|_{x=2} \right) = \frac{1}{3} \cdot \frac{4+y}{5} = \frac{4+y}{15}, \quad y = 0, 1, 2 \end{aligned}$$

$$\text{iii. } P(1 \leq X \leq 2) = P(X=1) + P(X=2) = \frac{2 \times 1 + 1}{9} + \frac{2 \times 2 + 1}{9} = \frac{8}{9},$$

and, $P(|X|<1) = P(-1<X<1) = P(X=0) = 1/9$.

Example 6.8. The joint probability mass function of the two discrete random variables X and Y is given by

$$p(x, y) = \frac{x^2 + y}{32}, \quad x = 0, 1, 2, 3; \quad y = 0, 1.$$

Obtain i. marginal probability function of X and Y ; ii. Conditional distribution function of X given $Y=1$ and Y given $X=2$. (left as exercise)

Example 6.9: If X and Y are two continuous random variables having joint density function:

$$f(x, y) = \begin{cases} \frac{6-x-y}{8}, & 0 \leq x < 2, 2 \leq y < 4 \\ 0, & \text{elsewhere} \end{cases}$$

Obtain i. marginal probability function of X and Y ; ii. Conditional distribution function of X given $Y=1$ and Y given $X=2$ iii. Also, find $P[X < 1 \cap Y < 1]$, $P[X+Y < 3]$ and $P[X < 1/Y < 3]$

Solution: i. To find the marginal probability function of X and Y , we have

$$\begin{aligned} f_x(x) &= \int_y f(x, y) dy = \int_2^4 \frac{6-x-y}{8} dy = \frac{1}{8} \left(6y - xy - \frac{y^2}{2} \right) \Big|_2^4 \\ &= \frac{1}{8} [(24 - 4x - 8) - (12 - 2x - 2)] = \frac{-2x + 6}{8} = \frac{3-x}{4}, \quad 0 \leq x < 2 \end{aligned}$$

and,

$$\begin{aligned} f_y(y) &= \int_x f(x, y) dx = \int_0^2 \frac{6-x-y}{8} dx = \frac{1}{8} \left(6x - \frac{x^2}{2} - yx \right) \Big|_0^2 \\ &= \frac{1}{8} [12 - 2 - 2y] = \frac{10-2y}{8} = \frac{5-y}{4}, \quad 2 \leq y < 4 \end{aligned}$$

ii. Conditional distribution function of X given $Y=1$ is given by

$$\begin{aligned} f(x/y) &= \frac{f_{xy}(x, y)}{f_y(y)}, \quad -\infty \leq x_i \leq \infty; \forall y \\ &= \frac{(6-x-y)/8}{(5-y)/4} = \frac{6-x-y}{2(5-y)} \Big|_{y=1} = \frac{5-x}{8}, \quad \text{for } 0 \leq x < 2 \end{aligned}$$

and, conditional distribution function of Y given $X=2$.

$$\begin{aligned} f(y/x) &= \frac{f_{xy}(x, y)}{f_x(x)}, \quad -\infty \leq y \leq \infty; \forall x \\ &= \frac{(6-x-y)/8}{(3-x)/4} = \frac{6-x-y}{2(3-x)} \Big|_{x=2} = \frac{4-y}{2}, \quad \text{for } 2 \leq y < 4 \end{aligned}$$

$$\text{iii. a. } P[X < 1 \cap Y < 3] = \int_0^1 \int_0^3 f(x, y) dx dy = \int_0^1 \int_0^3 \frac{6-x-y}{8} dx dy = \frac{3}{8}.$$

$$\text{b. } P[X+Y < 3] = \int_0^1 \int_0^{3-x} \frac{6-x-y}{8} dx dy = \frac{5}{24}$$

$$\text{c. } P[X < 1 | Y < 3] = \frac{P(X < 1 \cap Y < 3)}{P(Y < 3)} = \frac{3/8}{5/8} = \frac{3}{5}, \quad \{\because P(Y < 3) = \int_0^3 \frac{5-y}{4} dy = 5/8\}.$$

Example 6.10. The joint pdf of a two dimensional r.v. (X, Y) is defined by

$$f(x, y) = \begin{cases} 2, & \text{if } 0 < x < 1, 0 < y < x \\ 0, & \text{elsewhere} \end{cases}$$

Find i. marginal density function of X and Y , ii. conditional density of Y given $X=x$, and iii. verify the independence of X and Y .

Solution: i. To find the marginal density function of X and Y , we have

$$f_x(x) = \int_y f(x, y) dy = \int_0^x 2 dy = 2x, 0 < x < 1$$

and

$$f_y(y) = \int_x f(x, y) dx = \int_y^1 2 dx = 2(1-y), 0 < y < 1$$

ii. To find conditional density of Y given $X=x$, we have

$$f(y/x) = \frac{f_{xy}(x, y)}{f_x(x)} = \frac{2}{2x} = \frac{1}{x}, 0 < y < x$$

iii. To verify the independence of X and Y , we have

$$f_x(x) \times f_y(y) = 4x(1-y) \neq f(x, y) \Rightarrow X \text{ and } Y \text{ are not independent.}$$

Exercise

1. Verify whether the following function is a probability mass function? If so, find $P(X > 0.25)$, and $P(0.25 < X \leq 0.75)$.

$$f(x) = x, \quad x = \frac{1}{16}, \frac{3}{16}, \frac{1}{4}, \frac{1}{2}$$

Verify whether the function $f(x) = x, 0 < x < 1$ is a probability density function? If yes, find $P(X > 0.25)$, and $P(0.2 < X \leq 0.5)$.

2. Let X be an rv, with pdf $f(x) = cx, 0 < x < 1$ find c .
3. Let X be an rv, with pdf $f(x) = cx^2, 0 < x < 1$ find c .
4. The joint pdf of abivariate rv (X, Y) is $f(x, y) = 4xy, 0 < x < 1; 0 < y < 1$, find the marginal probability function of X and Y . Also, find the conditional probability of X given $Y=y$ and $X=x$.

5. The joint pdf of abivariate rv (X, Y) is $f(x, y) = cxy$, $0 < x < 1; 4 < y < 5$, find c , then find marginal probability function of X and Y . Also, find the conditional probability of X given $Y=y$ and $Y=y$ given $X=x$.

UNIT 7

MATHEMATICAL EXPECTATION OF A RANDOM VARIABLE

7.1 Objective. Here we study the expectation, variance and other moments in terms of random variables. Also, study the expectation of sum of two or more random variables, difference of random variables and other properties.

7.2 Introduction

We may be interested in talking about a value ‘average’ i.e., average income, average expenditure, average profit, average winnings etc., in all ‘the value average’ is a random phenomenon which is also termed as expected value or mathematical expectation. Here we study this concept in detail.

Definition: Let X be a discrete random variable with probability mass function $p(x)$, then the mathematical expectation (expected value) of X is given by

$$E(X) = \sum_x xp(x), \forall x.$$

Let X be a continuous random variable with probability density function $f(x)$, then the mathematical expectation (expected value) of X is given by

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx,$$

Provided, right hand integral is absolutely convergent, that is, $\int_{-\infty}^{\infty} |xf(x)|dx = \int_{-\infty}^{\infty} |x|f(x)dx < \infty$, i.e., converges to a finite value.

7.3 Variance of a random variable: The variance of a random variable say X , is given by

$$V(X) = E\{X - E(X)\}^2 = E(X^2) - \{E(X)\}^2.$$

Note. Variance of an r.v. can also be denoted as $Var(X)$ or by Greek letter σ^2 .

7.4 Expected value of function of a random variable

Consider a r.v. X , with pdf(or pmf) $f(x)$ and distribution function $F(x)$. Let $g(\cdot)$ be a function such that $g(X)$ is a r.v. and $E[g(x)]$ exists(i.e., defined), then,

$$E(g(X)) = \begin{cases} \int_{-\infty}^{\infty} g(x)f(x)dx, & \text{when } X \text{ is continuous} \\ \sum_x g(x)f(x), & \text{when } X \text{ is discrete} \end{cases}$$

Note. In particular, when X is a continuous r.v., and if $g(X) = X^r$, $r > 0$, then

$$E(X^r) = \int_{-\infty}^{\infty} x^r f(x)dx = \mu', r^{\text{th}} \text{ moment about origin}$$

7.5 Some properties Expectation

Property 1. Expectation of a constant is constant. i.e., $E(a) = a$, a being the constant.

Proof: By definition of expectation of a r.v.,

$$E(X) = \sum_x xp(x).$$

Now letting $X = a$, we have

$$\Rightarrow E(a) = \sum_x ap(x) = a \sum_x p(x) = a \cdot 1 = a. (\because \sum_x p(x) = 1, \forall x)$$

Note. Proof can be extended to continuous r.v., provided $E(X)$ exists.

Property 2. $E(aX) = aE(X)$

Proof: It's trivial by(i), provided $E(X)$ exists.

Property 3. $E(aX \pm b) = aE(X) \pm b$.

Proof: It's trivial by(i), provided $E(X)$ exists.

Property 4. Additive Property(Addition theorem of Expectation)

Statement. If X and Y are the two r.v.'s, then $E(X+Y) = E(X) + E(Y)$, provided all expectations exist.

Proof: Let X and Y be two continuous r.v.'s, with joint pdf $f(x, y)$ and marginal pdf's $f(x)$ and $f(y)$, respectively. Then by definition,

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx, \quad \text{and} \quad E(Y) = \int_{-\infty}^{\infty} yf(y)dy \quad (i)$$

$$\text{And, } E(X+Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x+y)f(x,y)dxdy$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xf(x,y)dxdy + \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} yf(x,y)dxdy$$

$$= \int_{-\infty}^{\infty} x \left\{ \int_{-\infty}^{\infty} f(x,y)dy \right\} dx + \int_{-\infty}^{\infty} y \left\{ \int_{-\infty}^{\infty} f(x,y)dx \right\} dy$$

$$= \int_{-\infty}^{\infty} xf(x)dx + \int_{-\infty}^{\infty} yf(y)dy, \quad (\because \text{ by definition of marginal pdf.})$$

Using (i), we have

$$E(X+Y) = E(X) + E(Y)$$

Note 1. Above result can be extended even for discrete r.v.'s just by replacing integral(\int) sign by summation(Σ) sign.

$$[\text{Hint. } E(X) = \sum_x xp(x), \forall x, E(Y) = \sum_y yp(y), \forall y, E(X+Y) = \sum_x \sum_y (x+y)p(x,y)]$$

Note 2. Above result can be extended for n r.v.'s as given below.

7.5.1 Generalised Addition theorem of Expectation

Statement. Let X_1, X_2, \dots, X_n be n random variables, then mathematical expectation of sum of these n r.v.'s is equal to sum of their expectations provided all expectations exist. Symbolically,

$$E(X_1 + X_2 + \dots + X_n) = E(X_1) + E(X_2) + \dots + E(X_n),$$

Or
$$E\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n E(X_i) \quad (1)$$

provided all $E(X_i)$ exists.

Proof: Consider two random variables, X_1 and X_2 , we have

$$E(X_1 + X_2) = E(X_1) + E(X_2), \quad (2)$$

Implies result (1) is true for $n = 2$.

Suppose, result(1) is true for $n = k$, then for $n = k+1$, we have

$$\begin{aligned} E\left(\sum_{i=1}^{k+1} X_i\right) &= E\left(\sum_{i=1}^k X_i + X_{k+1}\right) = E\left(\sum_{i=1}^k X_i\right) + E(X_{k+1}), [\text{by (2)}] \\ &= \sum_{i=1}^k E(X_i) + E(X_{k+1}) \\ &= \sum_{i=1}^{k+1} E(X_i), \end{aligned}$$

implies, result (1) is true for $n = k+1$. Hence, if (1) is true for $n = k+1$, it is also true for $n = k$. Thus, by mathematical induction, result (1) is true for all positive integer values of n .

Property 5. Multiplicative property(Multiplication theorem of Expectation)

Statement: If X and Y are two independent r.v.'s, then $E(XY) = E(X) E(Y)$, provided all expectations exist.

Proof: Let X and Y be two continuous independent r.v.'s, with joint pdf $f(x, y)$ and marginal pdf's $f(x)$ and $f(y)$, respectively. Then by definition,

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx, \quad \text{and} \quad E(Y) = \int_{-\infty}^{\infty} yf(y)dy \quad (i)$$

And,
$$E(XY) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy f(x, y) dx dy,$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy f(x)f(y) dx dy, \quad [\because X \text{ and } Y \text{ are independent, } f(x, y) = f(x)f(y)]$$

$$= \left(\int_{-\infty}^{\infty} xf(x)dx\right) \left(\int_{-\infty}^{\infty} yf(y)dy\right)$$

Using (i), we have

$$E(XY) = E(X) \cdot E(Y)$$

Hence proved.

Note. Above result can be extended for n independent r.v.'s

7.5.2 Generalisation Multiplication theorem of Expectation of n r.v.'s

Statement. Let X_1, X_2, \dots, X_n be n independent random variables, then mathematical expectation of product of these n r.v.'s is equal to product of their expectations, provided all expectations exist. Symbolically,

$$E(X_1 X_2 \cdots X_n) = E(X_1)E(X_2) \cdots E(X_n) = \prod_{i=1}^n E(X_i),$$

provided all $E(X_i)$ exists.

[Hint. Proof By mathematical induction property, result holds].

Property 6. Expectation of a Linear combination of Random Variables

Let X_1, X_2, \dots, X_n be any n random variables and if a_1, a_2, \dots, a_n are any n constantans, then

$$E\left(\sum_{i=1}^n a_i X_i\right) = \sum_{i=1}^n a_i E(X_i),$$

provided all the expectation exist.

[*Proof is trivial from property 2, 4, and generalised addition theorem]

7.6 Properties of variance

Property 1. Variance of a constant is zero, i.e., $V(a) = 0$, a being any constant.

Proof: By definition,

$$V(X) = E(X^2) - \{E(X)\}^2$$

Letting, $X = a$, then

$$V(a) = E(a^2) - \{E(a)\}^2 = a^2 - \{a\}^2 = 0.$$

Propert 2. $V(aX) = a^2 V(X)$

By definition,

$$\begin{aligned} V(aX) &= E(a^2 X^2) - \{E(aX)\}^2 \\ &= a^2 E(X^2) - \{aE(X)\}^2 \\ &= a^2 [E(X^2) - \{E(X)\}^2] = a^2 V(X). \end{aligned}$$

Property 3. $V(AX + b) = a^2 V(X)$

Proof is obvious by property 1 and 2.

7.6.1 Covariance

If X and Y are two random variables, then the covariance between them is defined as

$$\text{Cov}(X, Y) = E[\{X - E(X)\}\{Y - E(Y)\}] = E(XY) - E(X)E(Y)$$

If X and Y are independent then $E(XY) = E(X)E(Y)$, implies $\text{Cov}(X, Y) = 0$.

7.6.2 Properties of Covariance

- i. $\text{Cov}(aX, bY) = ab \text{cov}(X, Y)$
- ii. $\text{Cov}(X+a, Y+b) = \text{Cov}(X, Y)$

$$\text{iii. } \text{Cov} \left(\frac{X - \bar{X}}{\sigma_x}, \frac{Y - \bar{Y}}{\sigma_y} \right) = \frac{1}{\sigma_x \sigma_y} \text{Cov}(X, Y) = r_{xy}, \text{ the correlation coefficient.}$$

$$\text{iv. } \text{Cov}(aX+b, cY+d) = ac \text{Cov}(X, Y)$$

$$\text{v. } \text{Cov}(aX+bY, cX+dY) = ac\sigma_x^2 + bd\sigma_y^2 + (ad+bc)\text{cov}(X, Y)$$

7.6.3. Variance of a linear combination of r.v.'s

Let X_1, X_2, \dots, X_n be any n random variables and if a_1, a_2, \dots, a_n are any n constants, then

$$V\left(\sum_{i=1}^n a_i X_i\right) = \sum_{i=1}^n a_i^2 V(X_i) + 2 \sum_{i=1}^n \sum_{\substack{j=1 \\ i < j}}^n a_i a_j \text{cov}(X_i, X_j)$$

Proof. Let $U = a_1 X_1 + a_2 X_2 + \dots + a_n X_n$, then

$$E(U) = a_1 E(X_1) + a_2 E(X_2) + \dots + a_n E(X_n)$$

$$U - E(U) = a_1 [X_1 - E(X_1)] + a_2 [X_2 - E(X_2)] + \dots + a_n [X_n - E(X_n)]$$

Squaring and taking expectations on both sides, we get

$$\begin{aligned} E\{U - E(U)\}^2 &= a_1^2 E[X_1 - E(X_1)]^2 + a_2^2 E[X_2 - E(X_2)]^2 + \dots + a_n^2 E[X_n - E(X_n)]^2 \\ &\quad + 2 \sum_{i=1}^n \sum_{\substack{j=1 \\ i < j}}^n a_i a_j E\{[X_i - E(X_i)][X_j - E(X_j)]\} \end{aligned}$$

$$V(U) = a_1^2 V(X_1) + a_2^2 V(X_2) + \dots + a_n^2 V(X_n) + 2 \sum_{i=1}^n \sum_{\substack{j=1 \\ i < j}}^n a_i a_j \text{cov}(X_i, X_j)$$

$$\Rightarrow V\left(\sum_{i=1}^n a_i X_i\right) = \sum_{i=1}^n a_i^2 V(X_i) + 2 \sum_{i=1}^n \sum_{\substack{j=1 \\ i < j}}^n a_i a_j \text{cov}(X_i, X_j)$$

Remark: In the above result, if all $a_i = 1$, then

$$V\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n V(X_i) + 2 \sum_{i=1}^n \sum_{\substack{j=1 \\ i < j}}^n \text{cov}(X_i, X_j)$$

In particular, if $n=2$, and let $a_1 = 1, a_2 = 1$, then

$$V(X_1 + X_2) = V(X_1) + V(X_2) + 2\text{cov}(X_1, X_2)$$

let $a_1 = 1, a_2 = -1$, then

$$V(X_1 - X_2) = V(X_1) + V(X_2) - 2\text{cov}(X_1, X_2)$$

Thus we have,

$$V(X_1 \pm X_2) = V(X_1) + V(X_2) \pm 2\text{cov}(X_1, X_2)$$

If X and Y are independent, then $\text{cov}(X_1, X_2) = 0$, which implies

$$V(X_1 \pm X_2) = V(X_1) + V(X_2)$$

Example 7.6. Let X be a random variable with probability distribution defined by

$x:$	-3	6	9
$p(x):$	1/6	1/2	1/3

Determine i. $E(X)$, ii. $E(3X)$, iii. $E(-2X+5)$, vi. $V(X)$, v. $V(2X+3)$.

Solution. For the above problem, we have

- i. $E(X) = \sum_{i=1}^n xp(x) = 11/2.$
- ii. $E(3X) = 3E(X) = 3(11/2) = 33/2.$
- iii. $E(-2X+5) = -2E(X) + 5 = -6.$
- iv. $V(X) = E(X^2) - \{E(X)\}^2 = \sum_x x^2 p(x) - (11/2)^2 = 65/4.$
- v. $V(2X+3) = 2^2 V(X) = 4(65/4) = 65.$

Example 7.7. A continuous random variable X has a pdf:

$$f(x) = \begin{cases} 3x^2, & \text{if } 0 \leq x \leq 1 \\ 0, & \text{elsewhere} \end{cases}, \quad \text{find i. } E(X), \quad \text{ii. } E(X^2 + 2) \quad \text{iii. } V(X)$$

Solution. By definition-

- i. $E(X) = \int_0^1 xf(x)dx = \int_0^1 3x^3 dx = 3/4.$
- ii. $E(X^2) = \int_0^1 x^2 f(x)dx = \int_0^1 3x^4 dx = 3/5,$
 $\Rightarrow E(X^2 + 2) = 3/5 + 2 = 13/2.$
- iii. $V(X) = E(X^2) - \{E(X)\}^2 = (3/5) - (9/16) = 3/80.$

7.7. Cauchy Schwartz inequality: If X and Y are random variables taking real values then

$$\{E(XY)\}^2 \leq E(X)^2 E(Y)^2$$

[see proof in Fund. Math. Stat. by S.C. Gupta and V.K. Kapoor, P-6.22]

Note: In particular, by letting $X = |X - E(X)| = |X - \mu_x|$, and $Y = 1$, in the above inequality, we get

$$\{E|X - \mu_x|\}^2 \leq E|X - \mu_x|^2$$

\Rightarrow (mean deviation about mean)² \leq Variance(X)

\Rightarrow $M.D. \leq S.D.$, where S.D. denote the standard deviation.

7.8. Jensen's inequality: If g is a continuous and convex function on the interval I , and X is a random variable whose values are in I with probability 1, then

$$E(g(X)) \geq g\{E(X)\}, \quad \text{provided the expectation exist.}$$

[see proof in Fund. Math. Stat. by S.C. Gupta and V.K. Kapoor, P-6.23]

Cor. If g is a continuous and concave function on the interval I , then

$$E(g(X)) \leq g\{E(X)\}, \quad \text{provided the expectation exist.}$$

[see proof in Fund. Math. Stat. by S.C. Gupta and V.K. Kapoor, P-6.23]

Remark 1. If $E(X^2)$ exist then $E(X^2) \geq \{E(X)\}^2$,

Since $g(X) = X^2$, is convex function of X as $g''(X) = 2 > 0$.

2. If $X > 0$, i.e., X assumes only positive values and $E(X)$ and $E(1/X)$ exist then

$$E\left(\frac{1}{X}\right) \geq \frac{1}{E(X)},$$

Since $g(X) = 1/X$, is convex function of X as $g''(X) = \frac{2}{X^3} > 0$, for $X > 0$.

Example 7.11 For any two variates X and Y , show that

$$\{E(X+Y)^2\}^{1/2} \leq [\{E(X^2)\}^{1/2} + \{E(Y^2)\}^{1/2}] \quad (*)$$

Solution. Squaring both sides of(*), we have

$$\{E(X+Y)^2\} \leq E(X^2) + E(Y^2) + 2\sqrt{E(X^2)E(Y^2)}$$

$$\Rightarrow E(XY) \leq \sqrt{E(X^2)E(Y^2)}$$

$$\Rightarrow \{E(XY)\}^2 \leq E(X^2)E(Y^2),$$

which is a Cauchy –Schwartz inequality.

7.9. Mathematical Expectation of a two dimensional random variable

The mathematical expectation of a function $g(x, y)$ of two –dimensional r.v. (X, Y) with pdf $f(x, y)$ is given by

$$\begin{aligned} E[g(X, Y)] &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f(x, y) dx dy, \text{ if } X \text{ and } Y \text{ are continuous} \\ &= \sum_x \sum_y g(x, y) P\{X = x \cap Y = y\}, \text{ if } X \text{ and } Y \text{ are discrete} \end{aligned}$$

provided the expectation exist.

Imp. Note: In particular, if $g(x, y) = XY$, then

$$\begin{aligned} E[XY] &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xyf(x, y) dx dy, \text{ if } X \text{ and } Y \text{ are continuous} \\ &= \sum_x \sum_y xyP\{X = x \cap Y = y\}, \text{ if } X \text{ and } Y \text{ are discrete.} \end{aligned}$$

7.9.1 Conditional Expectation and Conditional Variance

The conditional expectation of X for given $Y=y$ is given by

$$\begin{aligned} E[X/Y=y] &= \frac{\int_{-\infty}^{\infty} x f(x, y) dx}{f(y)}, \text{ if } X \text{ and } Y \text{ are continuous} \\ &= \sum_x x P\{X = x/Y = y\} = \sum_x x \frac{p(x, y)}{p(y)}, \text{ if } X \text{ and } Y \text{ are discrete.} \end{aligned}$$

provided expectation exist.

$$E[Y|X=x] = \frac{\int_{-\infty}^{\infty} yf(x, y)dy}{f(x)}, \text{ if } X \text{ and } Y \text{ are continuous}$$

$$= \sum_y yP(Y=y|X=x) = \sum_y y \frac{p(x, y)}{p(x)}, \text{ if } X \text{ and } Y \text{ are discrete.}$$

provided expectation exist.

7.9.2 Conditional Variance

The conditional variance of X for given $Y=y$ is given by

$$V[X|Y=y] = E\{X^2|Y=y\} - \{E(X|Y=y)\}^2.$$

Similarly, conditional Variance of Y for given $X=x$, is given by

$$V[Y|X=x] = E\{Y^2|X=x\} - \{E(Y|X=x)\}^2.$$

Example 8.12. Two dimensional random variable (X, Y) with joint pdf

$$f(x, y) = \begin{cases} 2-x-y, & \text{if } 0 \leq x \leq 1, 0 \leq y \leq 1 \\ 0, & \text{elsewhere} \end{cases}$$

Find i. marginal density of X and Y ii. Conditional density of X given $Y=y$ and Y for given $X=x$;
 iii. $E(X)$ iv. $E(X|Y=1/2)$ v. $E(Y|X=3/4)$ vi. $V(Y|X=3/4)$.

Solution. i. marginal density of X :

$$f(x) = \int_0^1 f(x, y)dy = \int_0^1 (2-x-y)dy = \frac{3}{2} - x, 0 \leq x \leq 1$$

$$\text{And, } f(y) = \int_0^1 f(x, y)dx = \int_0^1 (2-x-y)dx = \frac{3}{2} - y, 0 \leq y \leq 1.$$

ii. Conditional density of X given $Y=y$ is

$$f(x/y) = \frac{f(x, y)}{f(y)} = \frac{2-x-y}{(3/2)-y}, \text{ for } 0 < (x, y) < 1$$

And,

$$f(y/x) = \frac{f(x, y)}{f(x)} = \frac{2-x-y}{(3/2)-x}, \text{ for } 0 < (x, y) < 1$$

$$\text{iii. } E(X) = \int_0^1 xf(x)dx = \int_0^1 x\left(\frac{3}{2}-x\right)dx = \frac{5}{12}$$

$$\text{iv. } E(Y|X=3/4) = \frac{\int_{-\infty}^{\infty} yf(x, y)dy}{f(x)} = \frac{\int_0^1 y(2-x-y)dy}{\frac{3}{2}-x} = \frac{\int_0^1 y(2-(3/4)-y)dy}{3/4} = 7/18$$

$$\text{vi. } V(Y|X=3/4) = E\{Y^2|X=3/4\} - \{E(Y|X=3/4)\}^2 = 23/324$$

(hint $E\{Y^2 / X = x\} = \frac{\int_{-\infty}^{\infty} y^2 f(x, y) dy}{f(x)} = 2/9$)

7.10 Moment Generating Function(MGF): Let X be a random variable, with probability function $f(x)$, then the moment generation function[$M_x(t)$] of X , is given by

$$M_x(t) = E(e^{tX}) = \begin{cases} \int_{-\infty}^{\infty} e^{tx} f(x) dx, & \text{when } X \text{ is a continuous r.v.} \\ \sum_x e^{tx} f(x), & \text{when } X \text{ is a discrete r.v.} \end{cases}$$

The summation or summation being extended to the entire range of x , 't', being the real value parameter and it's being assumed that the right hand side of $M_x(t)$ is absolutely convergent for some $h > 0$, such that $-h < t < h$.

Properties of MGF

a. $\frac{d^r}{dt^r} M_x(t) |_{t=0} = \mu'_r, r = 1, 2, \dots,$

In particular when $r = 1$, then $\frac{d}{dt} M_x(t) |_{t=0} = \mu'_1 = E(X)$, the mean of the r.v. X

When $r = 2$, then $\frac{d^2}{dt^2} M_x(t) |_{t=0} = \mu'_2 = E(X^2)$

Thus, variance $V(X) = \frac{d^2}{dt^2} M_x(t) |_{t=0} - \left(\frac{d}{dt} M_x(t) |_{t=0} \right)^2 = \mu'_2 - (\mu'_1)^2 = \mu_2$

b. $M_{cx}(t) = M_x(ct)$, c being a constant

c. Let $U = (X-a)/h$, where a and h are constants, $h \neq 0$, then $M_U(t) = Ee^{-(X-a)/h} = e^{-at/h} M_x(t/h)$

d. Let X_1, X_2, \dots, X_n be n random variables then the MGF of sum $Y = \sum_{i=1}^n X_i, i = 1, 2, \dots, n$ is given by

$$\begin{aligned} M_Y(t) &= Ee^{-tY} = Ee^{-t \sum_{i=1}^n X_i} = Ee^{-tX_1} \times Ee^{-tX_2} \times \dots \times Ee^{-tX_n} \\ &= M_{X_1}(t) \times M_{X_2}(t) \dots M_{X_n}(t) = \prod_{i=1}^n M_{X_i}(t) \end{aligned}$$

Exercise

1. Verify whether the following function is a probability mass function? If so, find mean and variance. Also obtain the MGF.

$$f(x) = x, x = \frac{1}{16}, \frac{3}{16}, \frac{1}{4}, \frac{1}{2}$$

2. Verify whether the function $f(x) = x$, $0 < x < 1$ is a probability density function? If yes, find MGF and hence determine mean and variance from it.
3. Let X be an rv, with pdf $f(x) = 2x$, $0 < x < 1$ find $E(X)$, $V(X)$, and $E(X+2)$
4. Let X be an rv, with pdf $f(x) = 3x^2$, $0 < x < 1$ find $E(X)$, $V(2X)$, and $E(3X-2)$. Also, find MGF.
5. The joint pdf of abivariate rv (X, Y) is $f(x, y) = 4xy$, $0 < x < 1; 0 < y < 1$, find $E(X|Y=1/2)$
6. The joint pdf of abivariate rv (X, Y) is $f(x, y) = 4xy/9$, $0 < x < 1; 4 < y < 5$, find $E(Y|X=1/4)$, and $V(Y|X=1/4)$.

UNIT 8

CENTRAL LIMIT THEOREM

8.1 Objective: Here our aim is to find the approximate distribution of the sum of random variables when sample size $n(n \rightarrow \infty)$ is large.

8.2 Introduction

It is one of the most important results in the theory of probability. It states that under certain very general conditions, the sum(S_n , say) of a large number of n ($n \rightarrow \infty$) random variables is approximately distributed as normal. Note that X_i 's can be either discrete or continuous or mixed random variables. That is CLT states that the distribution or the CDF (cumulative distribution function) of the sum(S_n , say) converges in distribution to the normal random variable when n is very large.

Here, we confined to state the statements of few theorems only.

8.3 Central Limit Theorem: Here we state central limit theorem for independent and i.i.d.(independent and identically distributed) random variables.

8.3.1 CLT for independent r.v.'s [Laplace - Liapounoff]: Let X_1, X_2, \dots, X_n be n independent random variables with $E(X_i) = \mu_i$ and $V(X_i) = \sigma_i^2$, then the sum $S_n = \sum X_i$ is asymptotically normal

with mean $\mu = \sum_{i=1}^n \mu_i$ and $\sigma^2 = \sum_{i=1}^n \sigma_i^2$.

This theorem was first stated by Laplace(1812) and rigorous proof under fairly general conditions was given by Liapounoff(1901).

8.3.2 CLT for independent and identically distributed(iid) r.v.'s [Lindeberg-Levy-Theorem]:

Let X_1, X_2, \dots, X_n be n independent and identically distributed(i.i.d.) random variables with $E(X_i) = \mu_1$ and $V(X_i) = \sigma_1^2$, then the sum $S_n = \sum X_i$ is asymptotically normal with mean $\mu = n\mu_1$ and variance $\sigma^2 = n\sigma_1^2$.

8.3.3 CLT for iid r.v.'s [De-Moivre's Laplace- Theorem].

If
$$X_i = \begin{cases} 1, & \text{with probability } p \\ 0, & \text{with probability } q = 1 - p \end{cases}$$

Then, the distribution of the random variable $S_n = X_1 + X_2 + \dots + X_n = \sum X_i$ is asymptotically normal as $n \rightarrow \infty$.

Proof: Since X_i 's are distributed as Bernoulli r.v.'s, we have by definition of MGF,

$$M_{X_i}(t) = E(e^{tx}) = (q + pe^t) \quad (1)$$

Then the sum $S_n = X_1 + X_2 + \dots + X_n = \sum X_i$, is distributed as binomial (n, p). Therefore,

$$M_{S_n}(t) = E(e^{ts}) = E\left(e^{t \sum_{i=1}^n X_i}\right) = \prod_{i=1}^n M_{X_i}(t) = (q + pe^t)^n \quad (2)$$

by uniqueness theorem of mgf's .

Therefore, $E(S_n) = np = \mu$ (say), and $V(S_n) = npq = \sigma^2$, (say)

Let $Z_n = \frac{S_n - E(S_n)}{\sqrt{\text{Var}(S_n)}} = \frac{S_n - np}{\sqrt{npq}}$, then

$$\begin{aligned} M_{Z_n}(t) &= E(e^{tZ_n}) = E\left(e^{\left(\frac{S_n - \mu}{\sigma}\right)t}\right) \\ &= e^{-\mu t / \sigma} M_{S_n}(t / \sigma) = e^{-npt / \sqrt{npq}} (q + pe^{t / \sqrt{npq}})^n \\ &= \left(1 + \frac{t^2}{2n} + O(n^{-3/2})\right)^n \end{aligned}$$

Where, $O(n^{-3/2})$ denote the terms containing $n^{3/2}$ and higher powers of n in the denominator.

Then as $n \rightarrow \infty$, we get

$$\lim_{n \rightarrow \infty} M_{Z_n}(t) = \lim_{n \rightarrow \infty} \left(1 + \frac{t^2}{2n} + O(n^{-3/2})\right)^n = \lim_{n \rightarrow \infty} \left(1 + \frac{t^2}{2n}\right)^n = e^{t^2/2},$$

which is the MGF of a standard normal variate. Hence by uniqueness theorem of mgf's,

$Z_n = \frac{S_n - np}{\sqrt{npq}}$, is asymptotically $N(0, 1)$. Which implies, the sum $S_n = X_1 + X_2 + \dots + X_n = \sum X_i$ is asymptotically normal $N(\mu = np, \sigma^2 = npq)$.

Example 2. Let X_1, X_2, \dots, X_n be i.i.d. $P(\lambda)$ rv's. Show that the sum $S_n = \sum_{i=1}^n X_i$ is distributed asymptotically normal.

Solution: Given $X_i \sim P(\lambda)$, then $S_n = \sum_{i=1}^n X_i \sim P(n\lambda = m, \text{ say}) \Rightarrow E S_n = m$ and $V(S_n) = m$

Let $Z_n = \frac{S_n - E(S_n)}{\sqrt{\text{Var}(S_n)}} = \frac{S_n - m}{\sqrt{m}}$, then

Then by definition of MGF, we have

$$\begin{aligned} M_{Z_n}(t) &= E(e^{tZ_n}) = E\left(e^{\left(\frac{S_n - \mu}{\sigma}\right)t}\right) \\ &= e^{-\mu t / \sigma} M_{S_n}(t / \sigma) = e^{-mt / \sqrt{m}} e^{m(e^{t/\sqrt{m}} - 1)} \\ &= e^{-t\sqrt{m}} e^{m(e^{t/\sqrt{m}} - 1)} \end{aligned}$$

Taking log on both sides we get

$$\log M_{Z_n}(t) = -t\sqrt{m} + m(e^{t/\sqrt{m}} - 1)$$

$$= -t\sqrt{m} + m\left(\frac{t}{\sqrt{m}} + \frac{t^2}{2m} + O(m^{-3/2})\right)$$

Where, $O(m^{-3/2})$ denote the terms containing $m^{-3/2}$ and higher powers of n or m in the denominator. Then as $n \rightarrow \infty \Leftrightarrow m \rightarrow \infty$, we get

$$\lim_{m \rightarrow \infty} M_Z(t) = \lim_{m \rightarrow \infty} \left(\frac{t^2}{2} + O(m^{-3/2}) \right) = \lim_{m \rightarrow \infty} \left(\frac{t^2}{2} \right) = e^{t^2/2},$$

which is the MGF of a standard normal variate. Hence by uniqueness theorem of mgf's,

$Z_n = \frac{S_n - m}{\sqrt{m}}$, is asymptotically $N(0, 1)$. Which implies, the sum $S_n = X_1 + X_2 + \dots + X_n = \sum X_i$ is asymptotically normal $N(\mu = m, \sigma^2 = m)$.

Note: The important applications of central limit theorem in real life are

- laboratory measurement errors are generally modelled by normal random variable
- In communication and signal processing, Gaussian (normal) distribution is frequently used to model Gaussian noise(error).
- In finance, the percentage changes in the prices of some assets are sometimes modelled by normal distribution

8.4 Some Practical Examples

Example. A bank teller serves customer standing in the queue one by one. Suppose that the service time X_i , for customer i has mean $E(X_i) = 2$ minutes, and $V(X_i) = 1$ minute², assume that the different bank customers are independent. Let Y be the total time the bank teller spends servicing 50 customers. find $P[90 < Y < 110]$.

Solution. Let Y be the total time the bank teller spends servicing 50 customers

$$\text{Then } Y = X_1 + X_2 + \dots + X_n$$

Where $n=50$, $E(X_i) = \mu = 2$, $V(X_i) = 1 = \sigma^2$, for all $i = 1, 2, \dots, n$

Therefore, $E(Y) = n\mu = 50 \times 2 = 100$, $V(Y) = n\sigma^2 = 50 \times 1 = 50$

$$\text{Let } Z = \frac{Y - \mu}{\sigma} = \frac{Y - 100}{\sqrt{50}} \sim N(0, 1)$$

Now, to find $P[90 < Y < 110]$

$$\begin{aligned} \text{We have } P[90 < Y < 110] &= P\left(\frac{90 - \mu}{\sigma} < \frac{Y - \mu}{\sigma} < \frac{110 - \mu}{\sigma}\right) \\ &= P\left(\frac{90 - 100}{\sqrt{50}} < Z < \frac{110 - 100}{\sqrt{50}}\right) \\ &= P(-1.41 < Z < 1.41) = \phi(1.41) - \phi(-1.41) = 0.9207 - [1 - \phi(1.41)] \\ &= 0.9207 - (1 - 0.9207) = 0.8414 \end{aligned}$$

Important Remark: Continuity Correction - In the above problem, it is noticed that the approximation is not so good. Part of the error is due to the fact that Y is a discrete random

variable, and we have used continuous distribution to determine $P[90 < Y < 100]$. Here we use better approximation called continuity correction. Since Y can take integer values only, we write

$$\begin{aligned} P[90 < Y < 100] &= P[89.5 < Y < 100.5] = P\left(\frac{89.5 - \mu}{\sigma} < \frac{Y - \mu}{\sigma} < \frac{100.5 - \mu}{\sigma}\right) \\ &= P\left(\frac{89.5 - 100}{\sqrt{50}} < Z < \frac{100.5 - 100}{\sqrt{50}}\right) \\ &= P(-1.48 < Z < 1.48) = \phi(1.48) - \phi(-1.48) = 0.9306 - [1 - \phi(1.41)] \\ &= 0.9306 - (1 - 0.9306) = 0.8612 \end{aligned}$$

It indicates that the continuity correction will significantly improve (here, 2% increase) the probability of occurrence. So, continuity correction to be useful especially when Bernoulli or binomial distribution is used, that to find the probability of occurrence between any two values.

Example. In a communication system each data consists of 1000 bits. Due to the noise, each bit may be received in error with probability 0.1. It is assumed bit errors occur independently. Find the probability that there are more than 120 errors in a certain data packet.

Solution. Let X_i be an indicator random variable for the i^{th} bit in the packet. That is,

$$X_i = \begin{cases} 1, & \text{if the } i^{\text{th}} \text{ bit is received in error} \\ 0, & \text{otherwise} \end{cases}$$

Then $X_i \sim \text{Bernoulli}(1, p = 0.1)$, where X_i 's are iid.

Let Y be the total number of bit errors in the packet.

Then $Y = X_1 + X_2 + \dots + X_n$

where, $n = 1000$, $E(X_i) = p = \mu = 0.1$, $V(X_i) = pq = \sigma^2 = 0.09$, for all $i = 1, 2, \dots, n$

Therefore, $E(Y) = n\mu = 1000 \times 0.1 = 100$, $V(Y) = n\sigma^2 = npq = 1000 \times 0.1 \times 0.9 = 90$

Let $Z = \frac{Y - \mu}{\sigma} = \frac{Y - 100}{\sqrt{90}} \sim N(0, 1)$

Now, to find i. $P[Y > 120]$

$$\begin{aligned} \text{We have } P[Y > 120] &= P\left(\frac{Y - \mu}{\sigma} > \frac{120 - \mu}{\sigma}\right) \\ &= P\left(Z > \frac{120 - 100}{\sqrt{90}}\right) \\ &= P(Z > 2.11) = 1 - \phi(2.11) = 1 - 0.9826 = 0.0174 \end{aligned}$$

Questions

- The approximate distribution of the sum of random variables for large n is
 - Normal
 - standard normal
 - binomial
 - Poisson
- In a communication system each data consists of 1000 bits. Due to the noise, each bit may be received in error with probability 0.01. It is assumed bit errors occur independently. Find the probability that there are more than 8 errors in a certain data packet.

3. In a production system each batch consists of 10000 units. Due to the disturbance in electricity or by mishandling, each batch may be received defectives with probability 0.2. It is assumed batch errors occur independently. Find the probability that there are more than 5 defectives in a certain batch box.

BLOCK –III
(PROBABILITY & SAMPLING DISTRIBUTIONS AND ESTIMATION)

UNIT 9: STANDARAD DISCRETE PROBABILITY DISTRIBUTIONS

UNIT 10: STANDARAD CONTINUOUS PROBABILITY DISTRIBUTIONS

UNIT 11: SAMPLING DISTRIBUTIONS

UNIT 12: POINT AND INTERVAL ESTIMATION

UNIT 9

STANDARAD DISCRETE PROBABILITY DISTRIBUTIONS

9.1 Objectives:

After studying this chapter, we are able to know the probability mass functions of some standard discrete distributions and some features, moment generating functions and examples over various discrete distributions. Also this will help us to learn about the possible applications of these distributions in the analysis of data.

9.2 Introduction:

In this chapter we will study some probability distribution that figures most useful in statistical theory and application. The purpose of this chapter is to show the types of situation in which these distribution can be applied. Some of the standard univariate discrete distributions are uniform, Bernoulli, binomial, Poisson, negative binomial, geometric and hyper geometric distribution.

9.3 Bernoulli distribution:

Bernoulli distribution was discovered by James Bernoulli. This is a discrete probability distribution. It is a distribution of number of successes on a single Bernoulli trial. If a trial results in to success or failure with the probability of success remains constant throughout an experiment when it is repeated for any number of times is called Bernoulli experiment (or trial). If for this experiment, a random variable X is defined such that it takes value 1 when success occurs and 0 if failure occurs, then X follows Bernoulli distribution with parameter ' p '. i.e., $X \sim B(1, p)$.

The Bernoulli distribution with parameter p can be written as follows:

X	0	1
$P(x)$	$p^0 q^{1-0} = q$	$p^1 q^{1-1} = p$

Since Sum of all the probabilities is equal to one, therefore Bernoulli distribution is a probability distribution.

Definition: if X is a discrete random variable with probability mass function

$$p(x) = \begin{cases} p^x q^{1-x}, & x = 0, 1; 0 < p < 1; q = 1 - p \\ 0, & \text{otherwise} \end{cases}$$
 then the distribution of X is called Bernoulli distribution.

Features of Bernoulli distribution:

1. p is the parameter of Bernoulli distribution.
2. The range of Bernoulli distribution is $x=0, 1$.
3. For Bernoulli distribution, mean = p , variance = pq and $SD = \sqrt{pq}$.
4. For Bernoulli distribution, mean > variance.
5. The moment generating distribution of $B(1, p)$ is $M_X(t) = q + pe^t$.

Examples on Bernoulli distribution:

1. Observe the new born baby and determine if the baby is a male or a female.

2. A contractor makes a certain tender for a contract; the outcome may be success or failure.
3. Inspect an item from production line and observe if it is defective or non-defective.

9.4 Binomial distribution:

Binomial distribution was discovered by James Bernoulli (1654-1705) in the year 1700 and was first published posthumously in 1713. Binomial distribution has n-independent Bernoulli trials. Here ‘n’ is finite and fixed. Each trials results either in a success or failure. The trials are mutually exclusive and exhaustive. The probability of success say, ‘p’ remains same for each trial.

The probability of x successes and consequently (n-x) failures in n-independent trials, in a specific order (say) SSFFSFFFFS (where S=success and F=failure) is given by multiplication probability theorem by the expression:

$$\begin{aligned}
 P(SSFFSFFFFS) &= P(S) * P(S) * P(F) * \dots * P(F) * P(S) * P(S) \\
 &= p * p * q * \dots * q * p * p \\
 &= (p.p.p\dots p) * (q.q\dots q) \\
 &\text{(here x times of p and (n - x) times of q appears)} \\
 &= p^x q^{n-x}.
 \end{aligned}$$

But x successes in n trials can occur in $\binom{n}{x}$ ways and the probability for each of these ways is same, viz., $p^x q^{n-x}$. Hence by addition theorem of probability the p.m.f can be written as $\binom{n}{x} p^x q^{n-x}$.

The Binomial distribution with parameters n and p can be written as follows:

X	0	1	2	n
P(x)	$\binom{n}{0} p^0 q^{n-0} = q^n$	$\binom{n}{1} p^1 q^{n-1}$	$\binom{n}{2} p^2 q^{n-2}$	$\binom{n}{n} p^n q^{n-n} = p^n$

There are (n+1) probability terms in a binomial distribution. The successive probability terms are the successive terms in the binomial expansion of $(q + p)^n$. Sum of all the probabilities are equal to one, because binomial distribution is a probability distribution.

Definition: A random variable X is said to follow a Binomial distribution if it assumes only non-negative values and its p.m.f is given by

$$P(x) = \begin{cases} \binom{n}{x} p^x q^{n-x}, & x = 0,1,2, \dots, n; 0 < p < 1; q = 1 - p \\ 0, & \text{otherwise} \end{cases}$$

Here n and p are known as the parameter s of the distribution.

Remark: the assignment of probability is permissible, because

$$\sum_{x=0}^n P(X = x) = \sum_{x=0}^n \binom{n}{x} p^x q^{n-x} = (q + p)^n = 1$$

9.5 Mean and variance of Binomial distribution:

To derive mean and variance of Binomial distribution, we use the definition of Expectation as

$$\begin{aligned}
\mu = E(X) &= \sum_{x=0}^n xP(x) = \sum_{x=0}^n x \binom{n}{x} p^x q^{n-x} \\
&= 0 + \sum_{x=1}^n x \binom{n}{x} p^x q^{n-x} \\
&= \sum_{x=1}^n x \left(\frac{n(n-1)!}{(n-x)! x(x-1)!} \right) p^x q^{n-x} \\
&= n \sum_{x=1}^n \binom{n-1}{x-1} p^{x-1+1} q^{n-x} \\
&= np \sum_{x=1}^n \binom{n-1}{x-1} p^{x-1} q^{n-x} \\
&= np \sum_{x=1}^n \binom{n-1}{x-1} p^{x-1} q^{n-x} \\
&= np(q+p)^{n-1}
\end{aligned}$$

[∵ W. K. T successive terms of binomial distribution for $\binom{n}{x} p^x q^{n-x}$ is $(q+p)^n = 1$]

$\mu = np$, which is the mean of binomial distribution.

Consider, $E(X^2) = E(X^2 - X + X) = E(X(X-1) + X) = E(X(X-1)) + E(X)$

$= \sum_{x=0}^n x(x-1) \binom{n}{x} p^x q^{n-x} + E(X)$

$$= 0 + 0 + \sum_{x=2}^n x(x-1) \left(\frac{n(n-1)(n-2)!}{(n-x)! x(x-1)(x-2)!} \right) p^{x-2+2} q^{n-x} + E(X)$$

$$= n(n-1)p^2 \sum_{x=2}^n \binom{n-2}{x-2} p^{x-2} q^{n-x} + E(X)$$

$$= n(n-1)p^2(q+p)^{n-2} + E(X) = n(n-1)p^2(1) + np$$

[∵ W. K. T successive terms of binomial distribution for $\binom{n}{x} p^x q^{n-x}$ is $(q+p)^n = 1$]

$$\therefore V(X) = \sigma^2 = E(X^2) - (E(X))^2$$

$$= n(n-1)p^2 + np - n^2p^2$$

$$= np(1-p)$$

$= npq$, which is the variance of binomial distribution.

Example on Binomial distribution:

1. Number of heads obtained on tossing 4 coins.
2. Number of bombs hitting a target among 3 bombs which are aimed at it.

Features of Binomial distribution:

1. The parameters of Binomial distribution are n and p .
2. The range of Binomial distribution is $x = 0, 1, 2, \dots, n$.

3. The mean and variance of Binomial distribution is mean = np and variance = npq .
4. The moment generating function of Binomial distribution is $M_X(t) = [1 + p(e^t - 1)]^n$.
5. The relationship between mean and variance of Binomial distribution is mean > variance (because $0 < p < 1$ and $0 < q < 1$)
6. If $p = q = 0.5$, then Binomial distribution is symmetrical (i.e., $\beta_1 = 0$)
7. The recurrence relation of binomial distribution in terms of probability is $P(x) = \frac{n-x+1}{x} \times \frac{p}{q} \times P(x-1)$, where $x = 1, 2, \dots, n$

Note: 1. Additive property of Binomial distribution: Let $X \sim B(n_1, p_1)$ and $Y \sim B(n_2, p_2)$ be independent random variables then sum of two independent Binomial variate is not a binomial variate.

2. If $p_1 = p_2 = p$, then $X + Y \sim B(n_1 + n_2, p)$ which means Binomial distribution possess additive property if $p_1 = p_2$.

Some Examples:

Example 1. In a college, 70% of the students are boys. In a random sample of 3 students, find the probability of getting i) two boys, ii) at least one boy.

Solution: Here X denotes number of boys selected at random of 3 students.

Then $X \sim B(n, p)$ where $n = 3$, $p = 0.7$, $q = 1 - p = 1 - 0.7 = 0.3$

WKT, the p.m.f of Binomial distribution is

$$P(x) = \binom{n}{x} p^x q^{n-x}, x = 0, 1, 2, \dots, n; 0 < p < 1; q = 1 - p$$

$$= \binom{3}{x} (0.7)^x (0.3)^{3-x}, x = 0, 1, 2, 3.$$

$$\text{i) } P(\text{two boys}) = P(X=2) = \binom{3}{2} (0.7)^2 (0.3)^{3-2} = 3 \times 0.49 \times 0.3$$

$$= 0.441$$

$$\text{ii) } P(\text{at least one boy}) = P(X \geq 1) = P(X = 1 \text{ or } 2 \text{ or } 3)$$

$$= P(X = 1) + P(X = 2) + P(X = 3)$$

$$= \binom{3}{1} (0.7)^1 (0.3)^{3-1} + \binom{3}{2} (0.7)^2 (0.3)^{3-2} + \binom{3}{3} (0.7)^3 (0.3)^{3-3}$$

$$= 0.189 + 0.441 + 0.343 = 0.973$$

Alternatively, $P(\text{at least one boy}) = P(X \geq 1) = 1 - P(X < 1)$

$$= 1 - P(X = 0)$$

$$= 1 - \left[\binom{3}{0} (0.7)^0 (0.3)^{3-0} \right]$$

$$= 1 - 0.027 = 0.973$$

Example 2. In a garden, there are 200 trees. Out of which 50 are orange trees. Among them, if 20 samples of 4 trees each are selected, in how many samples will you expect i)

exactly 2 orange tree, ii) at the most one orange tree.

Solution: Here X denotes number of orange trees selected at random of 4 trees.

Then $X \sim B(n, p)$ where $n=4$, $p=50/200=0.25$, $q=1-p=1-0.25=0.75$, $N=20$

WKT, the p.m.f of Binomial distribution is

$$P(x) = \binom{n}{x} p^x q^{n-x}, x = 0, 1, 2, \dots, n; 0 < p < 1; q = 1 - p$$

$$= \binom{4}{x} (0.25)^x (0.75)^{4-x}, x = 0, 1, 2, 3, 4.$$

$$\begin{aligned} \text{i) } P(\text{two orange tree}) &= P(X=2) = \binom{4}{2} (0.25)^2 (0.75)^{4-2} \\ &= 6 \times 0.0625 \times 0.5625 \\ &= 0.2109 \end{aligned}$$

Therefore, number of samples in which there are 2 exactly two orange trees = $N \times P(X = 2) = 20 \times 0.2109 = 4.218 \cong 4$.

$$\begin{aligned} \text{ii) } P(\text{at most one orange tree}) &= P(X \leq 1) \\ &= P(X = 0 \text{ or } 1) \\ &= P(X = 0) + P(X = 1) \\ &= \binom{4}{0} (0.25)^0 (0.75)^{4-0} + \binom{4}{1} (0.25)^1 (0.75)^{4-1} \\ &= 1 \times 1 \times 0.3164 + 4 \times 0.25 \times 0.4218 \\ &= 0.7382 \end{aligned}$$

Therefore, number of samples in which there are 2 at most one orange tree = $N \times P(X \leq 1) = 20 \times 0.7382 = 14.764 \cong 15$.

Example 3. Five coins are tossed and the number of heads are noted when an experiment is repeated 128 times and the following data is obtained. Fit a binomial distribution assuming a coin is unbiased.

No. of heads	0	1	2	3	4	5
Frequency	10	26	35	28	20	9

Solution: $X \sim B(n, p)$ where $n=5$, $p=0.5$ (unbiased coin), $q = 1 - p = 1 - 0.5 = 0.5$, $N=128$.

WKT, the p.m.f of Binomial distribution is

$$P(x) = \binom{n}{x} p^x q^{n-x}, x = 0, 1, 2, \dots, n; 0 < p < 1; q = 1 - p$$

$$= \binom{5}{x} (0.5)^x (0.5)^{5-x}, x = 0, 1, 2, \dots, 5.$$

To find the expected frequencies: $E_x = N \cdot P(x)$ we need to fit the binomial distribution which is as follows:

$$\text{At } x=0, \quad P(X=0) = \binom{5}{0} (0.5)^0 (0.5)^{5-0} = 0.03125$$

$$E_0 = N \cdot P(0) = 128 \cdot 0.03125 = 4$$

By using recurrence relation for expected frequencies

$$E_x = \frac{n-x+1}{x} \times \frac{p}{q} \times E(x-1), \text{ where } x=1, 2, \dots, n$$

$$E_1 = \frac{5-1+1}{1} \times 1 \times 4 = 20 ; \quad E_2 = \frac{5-2+1}{2} \times 1 \times 20 = 40$$

$$E_3 = \frac{5-3+1}{3} \times 1 \times 40 = 40 ; \quad E_4 = \frac{5-4+1}{4} \times 1 \times 40 = 20$$

$$E_5 = \frac{5-5+1}{5} \times 1 \times 20 = 4$$

Thus, the expected frequencies can be written in the below table:

No. of heads	0	1	2	3	4	5	Total
Frequency	10	26	35	28	20	9	N=128
Expected frequency	4	20	40	40	20	4	N=128

9.6 Poisson distribution

Poisson distribution was discovered by the French mathematician and physicist Denis Poisson (1781-1840) who published it in 1837. Poisson distribution is a limiting case of Binomial distribution under the following conditions:

- n , the number of trials is infinitely large. i.e., $n \rightarrow \infty$
- p , the constant probability of success for each trial is indefinitely small. i.e., $p \rightarrow 0$
- $np = \lambda$, (say) is finite.

9.7 Poisson distribution is a limiting case of Binomial distribution:

Proof: consider the p.m.f of Binomial distribution

$$P(x) = \binom{n}{x} p^x q^{n-x}, x = 0, 1, 2, \dots, n; 0 < p < 1; q = 1 - p$$

$$= \binom{n}{x} p^x (1 - p)^{n-x}$$

$$= \binom{n}{x} \left(\frac{p}{1-p}\right)^x (1 - p)^n$$

$$= \frac{n(n-1)(n-2)\dots(n-x+1)}{x!} \left(\frac{p}{1-p}\right)^x (1 - p)^n$$

Consider $np = \lambda \Rightarrow p = \frac{\lambda}{n} \rightarrow (*)$

$$= \frac{n(n-1)(n-2)\dots(n-x+1)}{x!} \left(\frac{\frac{\lambda}{n}}{1-\frac{\lambda}{n}}\right)^x \left(1 - \frac{\lambda}{n}\right)^n \quad [\text{from } (*)]$$

$$= \frac{\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\dots\left(1-\frac{(x-1)}{n}\right)}{x! \left(1-\frac{\lambda}{n}\right)^x} (\lambda)^x \left(1 - \frac{\lambda}{n}\right)^n$$

Take limits on both sides as $n \rightarrow \infty$ we get

$$\lim_{n \rightarrow \infty} P(x) = \frac{e^{-\lambda} \lambda^x}{x!}, x = 0, 1, 2, \dots$$

Note. $\lim_{n \rightarrow \infty} \left(1 - \frac{\lambda}{n}\right)^n = e^{-\lambda}$, and $\lim_{n \rightarrow \infty} \left(1 - \frac{\lambda}{n}\right)^\alpha = 1$, if α is not a function of n .

Definition: A random variable X is said to follow a Poisson distribution if it assumes only non-negative values and its p.m.f is given by

$$p(x) = \begin{cases} \frac{e^{-\lambda}\lambda^x}{x!}, & x = 0,1,2, \dots; \lambda \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

Here λ is known as the parameter of the distribution.

Note. Poisson process measures the number of occurrence of an outcome of a discrete random variable in a predetermined time interval, for which an average number of occurrences is known.

9.8 Mean and variance of Poisson distribution:

To derive mean and variance of Poisson distribution, we use the definition of Expectation as

$$\begin{aligned} \mu = E(X) &= \sum_{x=0}^{\infty} xP(x) = \sum_{x=0}^{\infty} x \frac{e^{-\lambda}\lambda^x}{x!} \\ &= e^{-\lambda} \left[0 \times \frac{\lambda^0}{0!} + 1 \times \frac{\lambda^1}{1!} + 2 \times \frac{\lambda^2}{2!} + \dots \right] \\ &= e^{-\lambda} \left[\lambda + \lambda^2 + \frac{\lambda^3}{2!} + \dots \right] \\ &= \lambda e^{-\lambda} \left[1 + \lambda + \frac{\lambda^2}{2!} + \dots \right] \\ &= \lambda e^{-\lambda} e^{\lambda} \end{aligned}$$

= λ , which is the mean of Poisson distribution.

Consider, $E(X^2) = E(X^2 - X + X) = E(X(X - 1) + X) = E(X(X - 1)) + E(X)$

$$\begin{aligned} &= \sum_{x=0}^{\infty} x(x-1) \frac{e^{-\lambda}\lambda^x}{x!} + E(X) \\ &= 0 + 0 + \sum_{x=2}^{\infty} x(x-1) \frac{e^{-\lambda}\lambda^x}{x(x-1)(x-2)!} + E(X) \\ &= e^{-\lambda} \sum_{x=2}^{\infty} \frac{\lambda^x}{(x-2)!} + E(X) \\ &= e^{-\lambda} \left[\frac{\lambda^2}{1} + \frac{\lambda^3}{1!} + \frac{\lambda^4}{2!} \dots \right] + E(X) \\ &= \lambda^2 e^{-\lambda} \left[1 + \lambda + \frac{\lambda^2}{2} \dots \right] + E(X) \\ &= \lambda^2 e^{-\lambda} e^{\lambda} + E(X) \\ &= \lambda^2 e^{-\lambda} e^{\lambda} + \lambda \\ &= \lambda^2 + \lambda \\ \therefore V(X) = \sigma^2 &= E(X^2) - (E(X))^2 \\ &= \lambda^2 + \lambda - \lambda^2 \end{aligned}$$

= λ , which is the variance of Poisson distribution.

Examples:

1. Number of accidents per week in a city.
2. Number of defective items in a box.
3. Number of patients visits the doctor per day between 6pm to 8pm.

Features of Poisson distribution:

1. The parameter of Poisson distribution is λ .
2. The range of Poisson distribution is $x=0, 1, 2, \dots, \infty$
3. The mean and variance of Poisson distribution is mean= λ and variance= λ .
4. The moment generating function of Poisson distribution is $M_x(t) = e^{\lambda(e^t-1)}$.
5. The relationship between mean and variance of Poisson distribution is mean = variance.
6. When λ is large, the Poisson distribution tends to normal distribution.
7. The recurrence relation of Poisson distribution in terms of probability is

$$P(x) = \frac{\lambda}{x} \times P(x-1), \text{ where } x=1,2,\dots$$

Note. Additive property of Poisson distribution: Let $X \sim P(\lambda_1)$ and $Y \sim P(\lambda_2)$ be independent random variables then sum of two independent Poisson variate is also a Poisson variate (i.e., $X+Y$ is also a Poisson variate with parameter $\lambda_1 + \lambda_2$). More elaborately, if $X_i, i=1, 2, \dots, n$ are independent Poisson variates with parameter $\lambda_i, i=1, 2, \dots, n$ respectively, then $\sum_{i=1}^n X_i$ is also a Poisson variate with parameter $\sum_{i=1}^n \lambda_i$.

Some Examples:

Example 1. A typist makes 3 mistakes per page on an average. Find the probability that a page typed by him has i) 1 mistake, ii) at the most 2 mistakes.

Solution: Here X : Number of typing mistakes per page.

Then $X \sim P(\lambda)$, where $\lambda = \text{average typing mistakes} = 3$.

WKT, the p.m.f of Poisson distribution is

$$P(x) = \frac{e^{-\lambda} \lambda^x}{x!}, x = 0, 1, 2, \dots; \lambda \geq 0$$

$$= \frac{e^{-3} 3^x}{x!}, x = 0, 1, 2, \dots$$

$$\text{i) } P(1 \text{ mistake}) = P(X=1) = \frac{e^{-3} 3^1}{1!}$$

$$= \frac{0.049 \times 3}{1} = 0.147$$

$$\text{ii) } P(\text{at the most 2 mistakes}) = P(X \leq 2)$$

$$= P(X = 0 \text{ or } 1 \text{ or } 2)$$

$$= P(X = 0) + P(X = 1) + P(X = 2)$$

$$= \frac{e^{-3} 3^0}{0!} + \frac{e^{-3} 3^1}{1!} + \frac{e^{-3} 3^2}{2!}$$

$$= 0.049 + 0.147 + 0.2205$$

$$= 0.4165$$

Example 2. On an average, the number of defective items in a box is 4. If there are 100 such boxes, in how many of them would you expect i) 2 defective items, ii) at least 1 defective item.

Solution: Here X: Number of defective item per box. Then $X \sim P(\lambda)$, where $\lambda =$ average number of defective items $= 4$.

WKT, the p.m.f of Poisson distribution is

$$P(x) = \frac{e^{-\lambda} \lambda^x}{x!}, x = 0, 1, 2, \dots; \lambda \geq 0$$

$$= \frac{e^{-4} 4^x}{x!}, x = 0, 1, 2, \dots$$

$$\begin{aligned} \text{i) } P(2 \text{ defective items}) &= P(X=2) = \frac{e^{-4} 4^2}{2!} \\ &= \frac{0.0183 \times 16}{2} \\ &= 0.1464 \end{aligned}$$

Therefore, expected number of boxes having two defective items =

$$N \times P(X = 2) = 100 \times 0.1464 = 14.64 \cong 15 \text{ boxes.}$$

$$\begin{aligned} \text{ii) } P(\text{at least 1 defective item}) &= P(X \geq 1) \\ &= 1 - P(X < 1) = 1 - P(X = 0) \\ &= 1 - \left[\frac{e^{-4} 4^0}{0!} \right] \\ &= 1 - 0.0183 = 0.9817 \end{aligned}$$

Therefore, expected number of boxes having at least 1 defective item =

$$N \times P(X \geq 1) = 100 \times 0.9817 = 98.17 \cong 98 \text{ boxes.}$$

Exercise

- For a Bernoulli distribution with parameter $p=0.4$. Write the p.m.f and hence find its mean and variance.
- The probability of hitting the target is $\frac{1}{4}$. If 3 arrows are aimed at the tree, find the probability that i) 2 arrows hit the tree, ii) at least one arrow hit the tree.
- The incidence of an occupational disease in a factory is such that the workers have 30% chance of suffering from it. What is the probability that out of 5 workers 3 or more contract the disease?
- The following data relates to the number of defective items in a sample of 5 for 500 samples taken during a week.

No. Of defective items	0	1	2	3	4	5
No. Of samples	160	188	120	20	10	2

- If has been found that on an average 3 patients visits a particular doctor during an hour. What is the probability that during a particular hour i) no patients visit the doctor, ii) more than 2 patients visits the doctor.
- On an average a box contains 2 defective items. Find the probability that a randomly selected box has i) no defective items, ii) at the most 2 defective items.

7. Fit a Poisson distribution to the following data and hence find the expected frequencies.

X	0	1	2	3	4	Total
F	211	90	20	4	0	325

UNIT 10

STANDARD CONTINUOUS PROBABILITY DISTRIBUTIONS

10.1 Objective

After studying this chapter, we are able to know the probability density functions of some standard continuous distributions and some properties, moment generating functions of these continuous distributions. Also, this will help us to learn about the possible applications of these distributions in the analysis of data.

10.2 Introduction

Since continuous random variables such as height, weight, income, etc can take large number of both integer and non-integer values. The sum of probability to each of these values is no longer sum to 1. Unlike discrete random variable, continuous random variables do not have probability distribution function specifying the exact probabilities of their specific values. Instead, a distribution determines probabilities that the random variable fall into a specified interval of values such function called as probability density function. Here, area under the probability distribution is equal to one and $P(a < X < b)$ represents the area under the probability density function curve between the values a and b .

Some of the standard univariate continuous distributions are uniform, Normal, exponential, Gamma and beta distribution.

10.3 Continuous uniform distribution:

Definition: A random variable X is said to have continuous uniform distribution over an interval

(a, b) , the its p.d.f is given by $f(x) = \begin{cases} \frac{1}{b-a}, & \text{if } a < x < b; -\infty < a < b < \infty \\ 0, & \text{otherwise} \end{cases}$

It is also called as rectangular distribution.

- The cumulative distribution function of continuous uniform distribution is given by

$$F(x) = \begin{cases} 0, & x \leq a \\ \frac{x-a}{b-a}, & a < x < b \\ 1, & x \geq b \end{cases}$$

- This distribution is also called as constant distribution because the probability is constant $\left(\frac{1}{b-a}\right)$ at every point of the interval (a, b) and is independent of values of the variable may take within the interval.
- This distribution is useful when the probability of occurrences of an event is constant and all possible values of the continuous variable are assumed equally likely.

Features of continuous uniform distribution:

1. The parameters of continuous uniform distribution is a and b .
2. The mean and variance of continuous uniform distribution is mean = $\frac{b+a}{2}$ and variance = $\frac{(b-a)^2}{12}$.

3. The moment generating function of continuous uniform distribution is $M_x(t) = \frac{e^{bt} - e^{at}}{t(b-a)}$, $t \neq 0$.

4. The mean deviation from mean of continuous uniform distribution is $MD(\bar{X}) = \frac{b-a}{4}$.

10.4 Normal distribution

It is the most useful theoretical distribution for continuous variables. Many statistical data concerning problems are displayed in the form of normal distribution. It is the corner stone of modern statistics. Historically normal distribution is associated with the names of De-Morvie, Pierre Laplace and Karl F. Gauss. In 1809 Gauss derived this distribution (also known as Gaussian distribution) as a model for measurement of errors, which is called ‘normal law of error’. The frequency distribution of values of the random variable observed in nature which follows this pattern approximately bell shaped. Thus, such distribution of measurements is called a normal distribution.

Definition: A continuous random variable X is said to be a normal distribution with probability density function is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}; \text{ where } -\infty < x < \infty; -\infty < \mu < \infty; \sigma > 0.$$

Here μ and σ^2 are the parameters of Normal distribution. It is denoted as $X \sim N(\mu, \sigma^2)$.

Examples:

1. Heights of a group of persons in a locality.
2. Weights of mangoes grown in tree.
3. Marks scored by students in an examination.

10.5 Mean and Variance of Normal distribution:

W.K.T the p.d.f of normal distribution is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}; \text{ where } -\infty < x < \infty; -\infty < \mu < \infty; \sigma > 0.$$

$$\begin{aligned} \text{Then mean} = E(X) &= \int_{-\infty}^{\infty} x \cdot f(x) dx = \int_{-\infty}^{\infty} x \cdot \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} dx \\ &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} x \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx \end{aligned}$$

By substitution method, put $t = \frac{x-\mu}{\sigma} \Rightarrow t\sigma = x - \mu$
 $\Rightarrow \sigma dt = dx$

When $x = -\infty \Rightarrow t = -\infty$ and $x = \infty \Rightarrow t = \infty$

On substituting to the above integral we get,

$$E(X) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} (\mu + \sigma t) \cdot e^{-\frac{1}{2}t^2} \sigma dt$$

$$\begin{aligned}
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \mu \cdot e^{-\frac{1}{2}t^2} dt + \frac{\sigma}{\sqrt{2\pi}} \int_{-\infty}^{\infty} t \cdot e^{-\frac{1}{2}t^2} dt \\
&= \frac{2\mu}{\sqrt{2\pi}} \int_0^{\infty} e^{-\frac{1}{2}t^2} dt + 0
\end{aligned}$$

(Since first integral is an even function and second integral is odd function)

$$= \frac{2\mu}{\sqrt{2\pi}} \sqrt{\frac{\pi}{2}} + 0 \quad (\text{From the below remark 4, 5 and 6})$$

= μ , which is the mean of normal distribution.

$$\begin{aligned}
\text{Variance} = V(X) &= E(X - \mu)^2 = \int_{-\infty}^{\infty} (x - \mu)^2 \cdot f(x) dx \\
&= \int_{-\infty}^{\infty} (x - \mu)^2 \cdot \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} dx
\end{aligned}$$

By substitution, the above integral reduces to

$$\begin{aligned}
&= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} t^2 \sigma^2 \cdot e^{-\frac{1}{2}t^2} \sigma dt \\
&= \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} t^2 \cdot e^{-\frac{1}{2}t^2} dt \\
&= \frac{\sigma^2}{\sqrt{2\pi}} 2 \int_0^{\infty} t^2 \cdot e^{-\frac{1}{2}t^2} dt
\end{aligned}$$

(Since this integral is an even function)

$$= \frac{2\sigma^2}{\sqrt{2\pi}} \sqrt{\frac{\pi}{2}} \quad (\text{From the below remark 4 and 5})$$

= σ^2 , which is the variance of normal distribution.

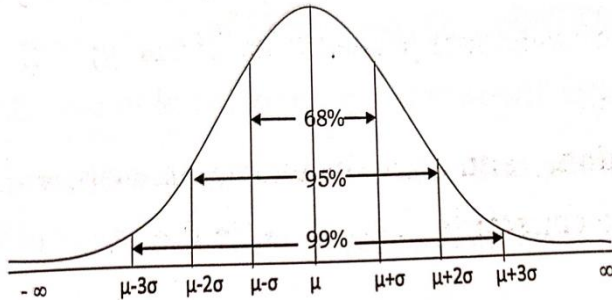
10.6 Properties of Normal distribution:

1. The normal curve is bell shaped.
2. The normal curve is symmetrical about the mean (i.e., $\beta_1=0$)
3. Here, mean=median=mode= μ .
4. The normal distribution has unimodal.
5. The normal distribution is mesokurtic (i.e., $\beta_2=3$)
6. For a normal distribution, standard deviation = σ , quartile deviation = $\frac{2}{3}\sigma$, and mean deviation = $\frac{4}{5}\sigma$.
7. The normal curve has points of inflexion (i.e., changes in curvature) at $\mu-\sigma$ and $\mu+\sigma$.
8. For normal distribution, the odd order moments are equal to zero i.e., $\mu_1 = \mu_3 = \mu_5 = \dots = 0$ and even order moments are constants i.e., $\mu_{2r} = 1 \times 3 \times 5 \times \dots \times (2r - 1)\sigma^{2r}$; $r = 1, 2, 3, \dots$
9. The quartile Q_1 and Q_3 are equidistant from median and it is given by $Q_1 = \mu - 0.6745\sigma$ and $Q_3 = \mu + 0.6745\sigma$.
10. The total area under normal curve is equal to 1. So that area to the right of the ordinate at the mean and left of the ordinate at the mean is 0.5.

Area property of normal distribution:

- $P(\mu-\sigma < x < \mu+\sigma) = 0.6826 = 68.26\%$
- $P(\mu-2\sigma < x < \mu+2\sigma) = 0.9544 = 95.44\%$
- $P(\mu-3\sigma < x < \mu+3\sigma) = 0.9974 = 99.74\%$

Though normal curve extends to $-\infty$ and ∞ , yet hardly 0.3% of the area lies beyond the limits $\mu-3\sigma$ and $\mu+3\sigma$.



10.7 Standard normal distribution:

Definition: if Z is a normal variate with mean $\mu=0$ and S.D. $\sigma=1$, then Z is called a standard normal variate and the distribution is called standard normal distribution. Its p.d.f is given by

$$f(z) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{z^2}{2}}; \text{ where } -\infty < z < \infty.$$

Note.

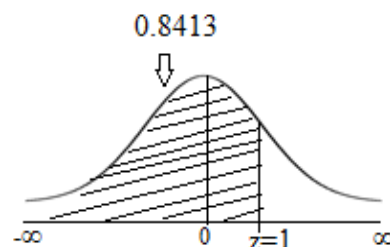
- If X is a normal variate with mean μ and S.D. σ , then $z = \frac{x-\mu}{\sigma} \sim N(0,1)$ is a standard normal variate. i.e., if $X \sim N(\mu, \sigma^2)$ then $Z \sim N(0,1)$.
- For a standard normal distribution mean=0, variance=1 and S.D. =1.
- The curve is bell shaped.
- It is symmetrical about $Z=0$. i.e., mean=median=mode=0.
- Total area under the standard normal curve is equal to one.

Example: If X is a normal variate with mean 60 and S.D.4, find the probability that i) $X \leq 64$, ii) $X \geq 62$, iii) $55 < X < 65$.

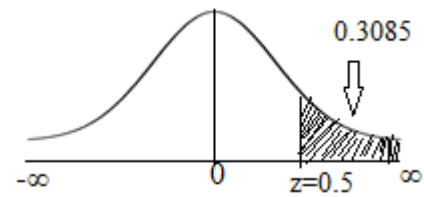
Solution: given $X \sim N(\mu, \sigma^2)$ with $\mu=60$, $\sigma=4$

$$\text{Then, } z = \frac{x-\mu}{\sigma} = \frac{x-60}{4} \sim N(0,1)$$

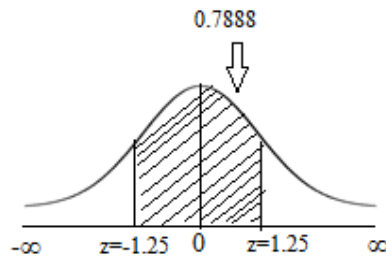
i)
$$\begin{aligned} P(X \leq 64) &= P\left(\frac{x-60}{4} \leq \frac{64-60}{4}\right) \\ &= P(z \leq 1) \\ &= \text{area from } -\infty \text{ to } 1 \\ &= 0.8413 \text{ [from normal table]} \end{aligned}$$



$$\begin{aligned}
\text{ii)} \quad P(X \geq 62) &= P\left(\frac{x-60}{4} \geq \frac{62-60}{4}\right) \\
&= P(z \geq 0.5) \\
&= \text{area from } 0.5 \text{ to } \infty \\
&= (\text{area from } -\infty \text{ to } \infty) - (\text{area from } -\infty \text{ to } 0.5) \\
&= 1 - 0.6915 \text{ [from normal table]} \\
&= 0.3085
\end{aligned}$$



$$\begin{aligned}
\text{iii)} \quad P(55 < X < 65) &= P\left(\frac{55-60}{4} < \frac{x-60}{4} < \frac{65-60}{4}\right) \\
&= P(-1.25 < z < 1.25) \\
&= \text{area from } -1.25 \text{ to } 1.25 \\
&= (\text{area from } -\infty \text{ to } 1.25) - (\text{area from } -\infty \text{ to } -1.25) \\
&= 0.8944 - 0.1056 \text{ [from normal table]} \\
&= 0.7888
\end{aligned}$$



Example: If Z is a standard normal variate and $P(Z < k) = 0.25$, find the value of k.

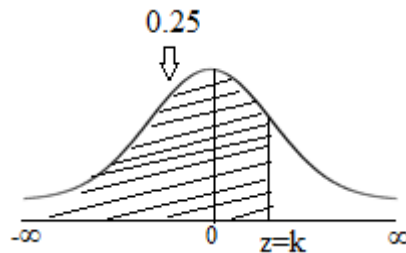
Solution:

Given $P(Z < k) = 0.25$

From normal table we have to find the ordinates where the probability value 0.25 coincide.

Since, 0.25 coincide to the ordinate -1.96

Therefore, $k = -1.96$ and $P(Z < -1.96) = 0.25$.



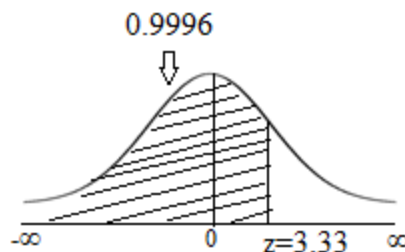
Example: Monthly income of employees follows normal distribution with mean Rs. 20,000 and S.D Rs.600. find the percentage of employees with monthly income i) less than Rs.22000, ii) lies between Rs. 16000 and 21000.

Solution: Here, X: monthly income of employees.

Then $X \sim N(\mu, \sigma^2)$ with $\mu = 20000$, $\sigma = 600$

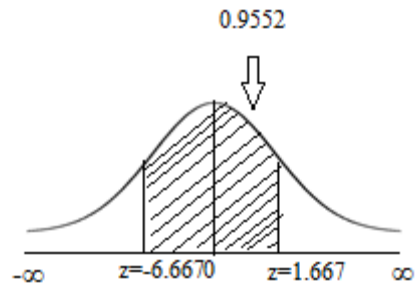
Then, $z = \frac{x-\mu}{\sigma} = \frac{x-20000}{600} \sim N(0,1)$

$$\begin{aligned}
\text{i)} \quad P(X < 22000) &= P\left(\frac{x-20000}{600} < \frac{22000-20000}{600}\right) \\
&= P(z < 3.33) \\
&= \text{area from } -\infty \text{ to } 3.33 \\
&= 0.9996 \text{ [from normal table]}
\end{aligned}$$



Therefore, the percentage of employees with monthly income less than Rs.22000 is $100 * P(X < 22000) = 100 * 0.9996 = 99.96\%$.

$$\begin{aligned}
& \text{ii)} \quad P(16000 < X < 21000) \\
& = P\left(\frac{16000 - 20000}{600} < \frac{x - 20000}{600} < \frac{21000 - 20000}{600}\right) \\
& = P(-6.667 < z < 1.667) \\
& = \text{area from } -6.667 \text{ to } 1.667 \\
& = (\text{area from } -\infty \text{ to } 1.667) - (\text{area from } -\infty \text{ to } -6.667) \\
& = 0.9522 - 0.000 \text{ [from normal table]} \\
& = 0.9552
\end{aligned}$$



Therefore, the percentage of employees with monthly income lies between Rs. 16000 and 21000 is $100 * P(16000 < X < 21000) = 100 * 0.9552 = 95.52\%$.

10.8 Exponential distribution:

Definition: A random variable X is said to have an exponential distribution with parameter $\theta > 0$,

its p.d.f is given by $f(x) = \begin{cases} \theta e^{-\theta x}, & \theta > 0, 0 < x < \infty \\ 0, & \text{otherwise} \end{cases}$

Here θ is the parameter of exponential distribution. it is denoted as $X \sim \exp(\theta)$.

- The cumulative distribution function is given by

$$F(x) = \begin{cases} 1 - e^{-\theta x}, & x \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

- Exponential distribution is closely related with the Poisson distribution. For example, if the Poisson random variable represents the number of arrivals per unit time at a service window, the exponential random variable will represent the time between two successive arrivals.
- The p.d.f of exponential distribution can be also written as

$$f(x) = \begin{cases} \frac{1}{\theta} e^{-x/\theta}, & \theta > 0, 0 < x < \infty \\ 0, & \text{otherwise} \end{cases}$$

Here the mean of the exponential distribution is θ and variance is θ^2 .

Features of exponential distribution:

- The mean and variance of exponential distribution is $\text{mean} = \frac{1}{\theta}$ and $\text{variance} = \frac{1}{\theta^2}$.
- The moment generating function of exponential distribution is $M_x(t) = \left(1 - \frac{t}{\theta}\right)^{-1}$, $\theta > t$.

The relationship between mean and variance of exponential distribution is

if $0 < \theta < 1 \Rightarrow \text{variance} > \text{mean}$

if $\theta = 1 \Rightarrow \text{variance} = \text{mean}$

if $\theta > 1 \Rightarrow \text{variance} < \text{mean}$

10.9 Mean and variance of Exponential distribution:

W.K.T the p.d.f of exponential distribution is $f(x) = \begin{cases} \theta e^{-\theta x}, & \theta > 0, 0 < x < \infty \\ 0, & \text{otherwise} \end{cases}$

Then mean = $E(X) = \int_{-\infty}^{\infty} x \cdot f(x) dx = \int_0^{\infty} x \cdot \theta e^{-\theta x} dx$

By applying integration by parts, the above integral can be written as

$$\begin{aligned} &= \theta \left[x \times \int_0^{\infty} e^{-\theta x} dx - \int_0^{\infty} \left(\int_0^{\infty} e^{-\theta x} \right) dx \times \frac{d}{dx}(x) \right] \\ &= \theta \left[\left(x \times \left(\frac{e^{-\theta x}}{-\theta} \right) \right)_{x=0}^{x=\infty} - \int_0^{\infty} \left(\frac{e^{-\theta x}}{-\theta} \right) dx \times (1) \right] \\ &= \theta \left[\left(\left(\infty \times \left(\frac{e^{-\theta \infty}}{-\theta} \right) \right) - \left(0 \times \left(\frac{e^{-\theta 0}}{-\theta} \right) \right) \right) + \frac{1}{\theta} \left(\frac{e^{-\theta x}}{-\theta} \right)_{x=0}^{x=\infty} \right] \\ &= \theta \left[(0 - 0) - \frac{1}{\theta} \left(\left(\frac{e^{-\theta \infty}}{\theta} \right) - \left(\frac{e^{-\theta 0}}{\theta} \right) \right) \right] \\ &= \theta \left[-\frac{1}{\theta} \left(0 - \frac{1}{\theta} \right) \right] = \theta \left(\frac{1}{\theta^2} \right) \end{aligned}$$

$= \frac{1}{\theta}$, which is the mean of exponential distribution.

Consider $E(X^2) = \int_{-\infty}^{\infty} x^2 \cdot f(x) dx = \int_0^{\infty} x^2 \cdot \theta e^{-\theta x} dx$

By applying integration by parts, the above integral can be written as

$$\begin{aligned} &= \theta \left[x^2 \times \int_0^{\infty} e^{-\theta x} dx - \int_0^{\infty} \left(\int_0^{\infty} e^{-\theta x} \right) dx \times \frac{d}{dx}(x^2) \right] \\ &= \theta \left[\left(x^2 \times \left(\frac{e^{-\theta x}}{-\theta} \right) \right)_{x=0}^{x=\infty} - \int_0^{\infty} \left(\frac{e^{-\theta x}}{-\theta} \right) dx \times (2x) \right] \\ &= \theta \left[\left(\left(\infty^2 \times \left(\frac{e^{-\theta \infty}}{-\theta} \right) \right) - \left(0^2 \times \left(\frac{e^{-\theta 0}}{-\theta} \right) \right) \right) + \frac{2}{\theta} \int_0^{\infty} x \times e^{-\theta x} dx \right] \\ &= \theta \left[(0 - 0) + \frac{2}{\theta} \left(\left(x \times \left(\frac{e^{-\theta x}}{-\theta} \right) \right)_{x=0}^{x=\infty} - \int_0^{\infty} \left(\frac{e^{-\theta x}}{-\theta} \right) dx \times (1) \right) \right] \\ &= \theta \left[\frac{2}{\theta} \left(\left(\infty \times \left(\frac{e^{-\theta \infty}}{-\theta} \right) \right) - \left(0 \times \left(\frac{e^{-\theta 0}}{-\theta} \right) \right) \right) + \frac{1}{\theta} \left(\frac{e^{-\theta x}}{-\theta} \right)_{x=0}^{x=\infty} \right] \\ &= \theta \left[\frac{2}{\theta} \left((0 - 0) - \frac{1}{\theta^2} (e^{-\theta \infty} - e^{-\theta 0}) \right) \right] \\ &= \theta \left[\frac{2}{\theta} \left(-\frac{1}{\theta^2} (0 - 1) \right) \right] = \theta \left[\frac{2}{\theta} \left(\frac{1}{\theta^2} \right) \right] = \frac{2}{\theta^2}. \end{aligned}$$

$$\text{Hence, } V(X) = E(X^2) - (E(X))^2 = \frac{2}{\theta^2} - \left(\frac{1}{\theta} \right)^2$$

$= \frac{1}{\theta^2}$, which is the variance of exponential distribution.

Example: If X is an exponential distribution with parameter 3.5, find variance, $P(X > 1)$, and $P(X \leq 4)$.

Solution: Given $X \sim \exp(\theta)$, where $\theta = 3.5$

W.K.T the p.d.f of exponential distribution is

$$f(x) = \begin{cases} 3.5e^{-3.5x}, & \theta > 0, 0 < x < \infty \\ 0, & \text{otherwise} \end{cases}$$

$$\text{WKT, variance} = \frac{1}{\theta^2} = \frac{1}{3.5^2} = 0.0816$$

$$\begin{aligned} P(X > 1) &= \int_1^{\infty} 3.5e^{-3.5x} dx = 3.5 \left(\frac{e^{-3.5x}}{-3.5} \right)_{x=1}^{x=\infty} = -(e^{-3.5(\infty)} - e^{-3.5(1)}) \\ &= -(0 - e^{-3.5}) = e^{-3.5} = 0.0302 \end{aligned}$$

$$\begin{aligned} P(X \leq 4) &= \int_0^4 3.5e^{-3.5x} dx = 3.5 \left(\frac{e^{-3.5x}}{-3.5} \right)_{x=0}^{x=4} = -(e^{-3.5(4)} - e^{-3.5(0)}) \\ &= -(e^{-14} - 1) = -0.000000831 + 1 = 0.999 \end{aligned}$$

Example: the monthly income of a group of 5000 persons were assumed to be exponential with mean Rs.800. how many persons have income i) between Rs.600 & Rs.1000, ii) less than Rs.600.

Solution: Given $X \sim \exp(\theta)$, where $\theta = 1/\text{mean} = 1/800 = 0.00125$.

W.K.T the p.d.f of exponential distribution is $f(x) = \begin{cases} \theta e^{-\theta x}, & \theta > 0, 0 < x < \infty \\ 0, & \text{otherwise} \end{cases}$

$$= \begin{cases} 0.00125e^{-0.00125x}, & \theta > 0, 0 < x < \infty \\ 0, & \text{otherwise} \end{cases}$$

i) $P(\text{between Rs. 600 \& Rs. 1000}) = P(600 < X < 1000)$

$$\begin{aligned} &= \int_{600}^{1000} 0.00125e^{-0.00125x} dx \\ &= 0.00125 \left(\frac{e^{-0.00125x}}{-0.00125} \right)_{x=600}^{x=1000} \\ &= -(e^{-0.00125(1000)} - e^{-0.00125(600)}) \\ &= -(0.2865 - 0.4723) \\ &= 0.1858 \end{aligned}$$

Therefore, the number of persons having income between Rs.600 & Rs.1000 is $N * P(600 < X < 1000) = 5000 * 0.1858 = 929$ persons.

ii) $P(\text{less than Rs. 600}) = P(X < 600)$

$$\begin{aligned} &= \int_0^{600} 0.00125e^{-0.00125x} dx \\ &= 0.00125 \left(\frac{e^{-0.00125x}}{-0.00125} \right)_{x=0}^{x=600} \end{aligned}$$

$$\begin{aligned}
&= -(e^{-0.00125(600)} - e^{-0.00125(0)}) \\
&= -(0.4723-1) \\
&= 0.5277
\end{aligned}$$

Therefore, the number of persons having income less than Rs.600 is $N \cdot P(X < 600) = 5000 \cdot 0.5277 = 2638.5 \approx 2639$ persons.

Exercise

- The mean and S.D of a normal distribution is 15 and 4 respectively. Find the upper and lower quartiles.
- If Z is a standard normal variate and $P(Z > k) = 0.1$, find the value of k .
- Heights of 300 children are normally distributed with mean 120cms and variance 4cms^2 . Find the number of children having heights i) greater than 116cms, ii) between 115cms and 120cms, iii) less than 118cms.
- The weekly wages of workers are normally distributed with mean Rs.2500 and S.D. Rs.400. find the probability of workers whose weekly wages will be i) more than Rs.3000, ii) less than Rs.3500, iii) between Rs.2000 and Rs.3000.
- The mileage(in thousands of miles), which car owners get with a certain kind of tyres is a random variable having p.d.f $f(x) = \begin{cases} 0.00026e^{-0.00026x}, & \theta > 0, 0 < x < \infty \\ 0, & \text{otherwise} \end{cases}$
Find the probability that one of these tyres will last i) at least 2000miles, ii) between 800 & 12000miles.
- If X is an exponential distribution with parameter 2.8, find variance, $P(X > 0.5)$, $P(0 < X \leq 3)$.

UNIT 11

SAMPLING DISTRIBUTIONS

11.1 Objectives:

After studying this chapter, we can develop the concepts of a sampling distribution that helps to understand the methods and underlying thinking of statistical inference.

11.2 Introduction

As we have studied several methods to calculate parameters such as mean and standard deviation of the population of interest. These values were used to describe the characteristics of the population. If a population is very large and the description of its characteristics is not possible by the census method, then to use at the statistical inference, sample of a given size are drawn repeatedly from the population and a particular 'statistic' is computed for each sample and the computed value is likely to vary from sample to sample. Thus, it is possible to construct frequency table for various values of statistic. The distribution of values of a sample statistic is called a sampling distribution. Here samples are drawn based on simple random sampling, therefore sample statistic is random variable.

Sampling distribution: The sampling distribution of a statistic is the distribution of the statistic for all possible samples from the same population of a given size.

Standard error of an estimate: the standard deviation of the sampling distribution of a statistic is called standard error (S.E).

- The following table represents standard error for some of the statistic:

Statistic	Standard error
Sample mean (\bar{X})	$SE(\bar{X}) = \sigma / \sqrt{n}$
Difference of means ($\bar{X}_1 - \bar{X}_2$)	$SE(\bar{X}_1 - \bar{X}_2) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
Sample proportion(p)	$SE(p) = \frac{PQ}{\sqrt{n}}$
Difference of proportion(p_1-p_2) if $p_1 \neq p_2$	$SE(p_1 - p_2) = \sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}$
Difference of proportion(p_1-p_2) if $p_1 = p_2 = p$	$SE(p_1 - p_2) = \sqrt{PQ \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$

- **Uses of standard error:**

1. Standard error is used to decide the efficiency and consistency of the statistic as an estimator.
2. It is used to obtain the confidence intervals of an estimate.
3. It is used in the testing of hypothesis.

• **Types of sampling distributions:**

1. Chi-square distribution
2. Student's t-distribution
3. F-distribution

11.3 Chi-Square Distribution

The square of a standard normal variate is known as chi-square variate with 1 degree of freedom (d.f). Thus, if $X \sim N(\mu, \sigma^2)$, then $Z = \frac{X-\mu}{\sigma} \sim N(0,1)$ and $Z^2 = \left(\frac{X-\mu}{\sigma}\right)^2$ is a chi-square variate with 1 d.f.

In general, if X_i , ($i=1,2,3,\dots,n$) are n independent normal variates with mean μ_i and variance σ_i^2 ($i=1,2,3,\dots,n$), then

$$\chi^2 = \sum_{i=1}^n \left(\frac{X_i - \mu_i}{\sigma_i}\right)^2 \text{ is a chi-square variate with } n \text{ d.f.}$$

Definition: If X follows chi-square variate with n d.f then the p.d.f of chi-square distribution is given by $f(x) = \frac{1}{2^{n/2} \Gamma(\frac{n}{2})} e^{-x/2} \cdot x^{\frac{n}{2}-1}$, $0 \leq x < \infty$. Then the distribution of X is called chi-square distribution with n d.f.

Note: degrees of freedom (d.f): the number of independent variates which makes up the statistic is known as degrees of freedom.

11.3.1 Features of chi-square distribution:

1. n is the parameter of chi-square distribution.
2. The range of chi-square distribution is $0 \leq \chi^2 < \infty$.
3. For a chi-square distribution, mean= n , variance= $2n$ and $SD=\sqrt{2n}$.
4. Mode of chi-square distribution is $mode = \begin{cases} n-2, & \text{for } n > 2 \\ 0, & \text{for } n \leq 2 \end{cases}$.
5. Chi-square distribution is positively skewed distribution ($\beta_1 > 0$).
6. Chi-square distribution is leptokurtic ($\beta_2 > 3$).
7. The total area under chi-square curve is equal to one.
8. When $n \rightarrow \infty$, chi-square variate tends to standard normal variate.

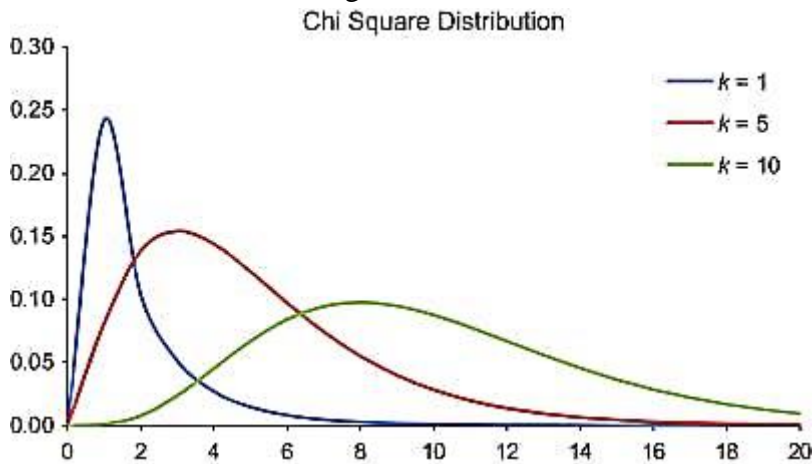
11.3.2 Applications of Chi-square distribution:

It has many uses in the field of testing of hypothesis. Some of them are:

1. To test the population variance.
2. To test the goodness of fit.
3. To test the independence of attributes.
4. To test the homogeneity of independent estimates of the population variance.
5. To test the homogeneity of independent estimates of the population correlation coefficients.

Chi-square probability curve:

Curves for different K- degrees of freedom are as follows:



11.4 Student's t-distribution:

Student's t- distribution is also derived from the normal distribution. The distribution is introduced by W.S.Gossett in 1908. The t-distribution describes the standardized distances of sample means from the population mean when the population standard deviation is not known and the observation come from the normally distributed population.

Definition: Let X_i , ($i=1,2,3,\dots,n$) be a random sample of size n from normal population. Then the student's t-statistic is defined as $t = \frac{\bar{X}-\mu}{S/\sqrt{n}}$ follows student's t-distribution with $(n-1)$ d.f with p.d.f is given by

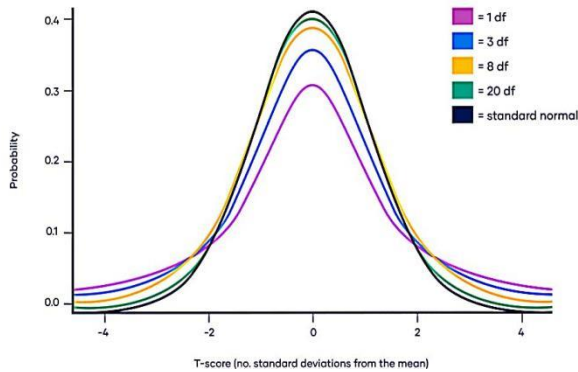
$$f(t) = \frac{1}{\sqrt{n-1}B\left(\frac{1}{2}, \frac{n-1}{2}\right)} \frac{1}{\left(1+\frac{t^2}{n-1}\right)^{n/2}}, -\infty < t < \infty.$$

11.4.1 Features of chi-square distribution:

1. n is the parameter of t-distribution.
2. The range of t-distribution is $-\infty < t < \infty$.
3. The t-curve is a bell-shaped curve.
4. The t-distribution is symmetric about $t=0$ ($\beta_1 > 0$).
5. For a t-distribution, mean=median=mode=0.
6. Variance of t-distribution is given by variance = $\frac{n}{n-2}$, for $n > 2$.
7. t-curve is asymptotic to X-axis.
8. The t-distribution is leptokurtic ($\beta_2 > 3$).
9. When $n \rightarrow \infty$, t-variate tends to standard normal variate.

11.4.2 The probability curve of student's t-distribution:

Curves for different degrees of freedom are as follows:



11.4.3 Applications of student's t-distribution:

1. To test the sample mean differs significantly from population mean.
2. To test the significance difference between two sample means for independent samples.
3. To test the significance difference between two sample means for dependent or paired samples.
4. To test the significance of an observed correlation coefficient and sample regression coefficient.

11.5 Snedecor's F-distribution:

The F- distribution arises from inferential statistics concerning population variances. More specifically, we use an F-distribution when we are studying the ratio of the variances of two normally distributed populations. It is used to construct confidence interval and testing of hypothesis about population variances. It is also used in one factor analysis of variance (ANOVA) and it is concerned with comparing the variation between several groups and variation within each group.

Definition: F is defined as the ratio of two independent chi-square variate divided by their respective d.f i.e. $F = \frac{X/m}{Y/n}$, where X and Y are independent chi-square variate and it follows Snedecor's F-distribution with (m, n) d.f with probability density function is given by:

$$f(F) = \frac{\left(\frac{m}{n}\right)^{\frac{m}{2}}}{B\left(\frac{m}{2}, \frac{n}{2}\right)} \frac{F^{\frac{m}{2}-1}}{\left(1 + \frac{m}{n}F\right)^{\frac{m+n}{2}}}, 0 \leq F < \infty$$

11.5.1 Features of F-distribution:

1. m and n are the parameters of F-distribution.
2. The range of F-distribution is $0 \leq F < \infty$.
3. For a F-distribution, mean = $\frac{n}{n-2}$, $n > 2$, variance = $\frac{2n^2(m+n-2)}{m(n-2)^2(n-4)}$, $n > 4$ and SD =

$$\sqrt{\frac{2n^2(m+n-2)}{m(n-2)^2(n-4)}}$$

4. The mode of F-distribution is $\text{mode} = \frac{n(m-2)}{m(n+2)}$
5. The total area under F-distribution curve is unity.
6. The reciprocal property of F-distribution is

$$F_{\frac{\alpha}{2}, (m-1, n-1)} \times F_{1-\frac{\alpha}{2}, (n-1, m-1)} = 1$$

$$\Rightarrow F_{\frac{\alpha}{2}, (m-1, n-1)} = \frac{1}{F_{1-\frac{\alpha}{2}, (n-1, m-1)}}$$

Assumptions:

1. Independent random samples are drawn from each of two normally distributed populations.
2. The amount of variability in the two populations is same and can be measured by a common variance σ^2 , i. e., $\sigma_1^2 = \sigma_2^2 = \sigma^2$.

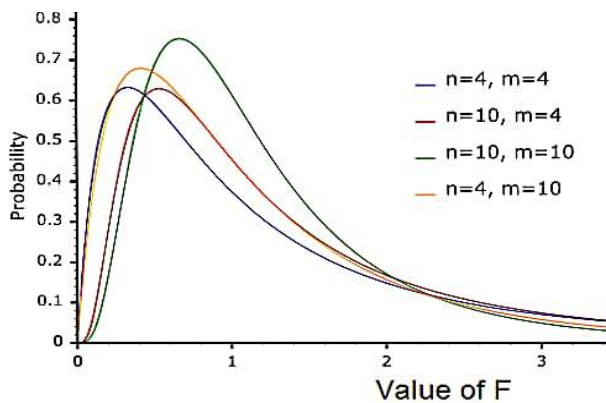
11.5.2 Applications of F-distribution:

It has the following applications in statistical theory.

1. To test the equality of two population variances.
2. To test the significance of an observed multiple correlation coefficient.
3. To test the linearity of regression.
4. To test the equality of several means.

11.5.3 The probability curve of F-distribution:

Curves for different degrees of freedom are as follows:



Exercise

1. What is sampling distribution? Explain standard error in sampling distribution.
2. Write a note on chi-square distribution.
3. Mention some applications of t-distribution.
4. Lists the features of F-distribution.
5. Given $\sigma^2 = 9\text{cm}^2$ and $n=36$, calculate standard error of sample mean.
6. If $P=0.05$ and $n=60$, then find S.E(p).
7. Sizes of two samples are 50 and 100. Population standard deviations are 20 and 10. Compute S.E. $(\bar{X}_1 - \bar{X}_2)$.

8. Write the uses of standard error.

UNIT 12
POINT AND INTERVAL ESTIMATION

12.1 Objective

After studying this chapter, we can able to understand one of the important branches of statistical inference and one can able to implement estimation technique to estimate true population parameter.

12.2 Introduction

Statistical inference is theory of making decisions about the population parameter from the analysis of a sample drawn from that population. It has two branches: Estimation and testing of hypothesis.

Estimation is the method of obtaining the most likely value of the population parameter using statistic.

Any statistic (T) which is used to estimate the population parameter is called an **estimator** and the specific value of the estimator is called an **estimate**.

There are two types of estimation: point estimation and interval estimation.

Definition of Point estimation: A single value is used to estimate an unknown population parameter is called point estimation.

For example: average height of group of 100 students is 165cms.

12.3 Properties (or characteristics) of estimator:

The following are some criteria that should be satisfied by a good estimator:

1. Unbiasedness
2. Consistency
3. Efficiency
4. Sufficiency

12.3.1 UNBIASEDNESS:

An estimator $T_n = T(x_1, x_2, \dots, x_n)$ is said to be unbiased estimator of $\gamma(\theta)$ if $E(T_n) = \gamma(\theta)$, for all $\theta \in \Theta$.

Remark: if $E(T_n) > \theta$, T_n is said to be positively biased and if $E(T_n) < \theta$, then T_n is said to be negatively biased. The amount of bias $b(\theta)$ is given by $b(\theta) = E(T_n) - \gamma(\theta)$, $\theta \in \Theta$.

Example: X_1, X_2, \dots, X_n be a random sample from a normal population $N(\mu, 1)$. Show that $t = \frac{1}{n} \sum_{i=1}^n x_i^2$ is an unbiased estimator of $\mu^2 + 1$.

Solution: Given $E(X_i) = \mu$, $V(X_i) = 1$, for all $i = 1, 2, 3, \dots, n$

Now, $E(X_i) = V(X_i) + (E(X_i))^2 = 1 + \mu^2$

Therefore, $E(t) = E\left(\frac{1}{n} \sum_{i=1}^n x_i^2\right) = \frac{1}{n} \sum E(X_i^2) = \frac{1}{n} \sum (1 + \mu^2) = \frac{n(1 + \mu^2)}{n} = (1 + \mu^2)$.

Hence t is an unbiased estimator of $\mu^2 + 1$.

12.3.2 Consistency

An estimator $T_n = t(x_1, x_2, \dots, x_n)$ based on a random sample of size n , is said to be consistent estimator of $\gamma(\theta)$, $\theta \in \Theta$, if T_n converges to $\gamma(\theta)$ in probability i.e., $T_n \xrightarrow{p} \gamma(\theta)$, as $n \rightarrow \infty$

OR

If T_n is said to be consistent estimator of $\gamma(\theta)$ if for every $\varepsilon > 0$, $\eta > 0$, there exists a positive integer $n \geq m(\varepsilon, \eta)$ such that $P(|T_n - \gamma(\theta)| < \varepsilon) \rightarrow 1$ as $n \rightarrow \infty$

OR $P(|T_n - \gamma(\theta)| \geq \varepsilon) \rightarrow 0$ as $n \rightarrow \infty$

$\Rightarrow P(|T_n - \gamma(\theta)| < \varepsilon) > 1 - \eta$ for all $n \geq m$, where m is some very large value of n .

Theorem 1: Invariance property of consistent estimator: if T_n is a consistent estimator of $\gamma(\theta)$ and $\psi(\gamma(\theta))$ is a continuous function of $\gamma(\theta)$, then $\psi(T_n)$ is a consistent estimator of $\psi(\gamma(\theta))$.

Theorem 2: Sufficient conditions for consistency: Let $\{T_n\}$ be a sequence of estimators such that for all $\theta \in \Theta$,

i. $E_\theta(T_n) \rightarrow \gamma(\theta)$, as $n \rightarrow \infty$

ii. $V_\theta(T_n) \rightarrow 0$, as $n \rightarrow \infty$

Then T_n is a consistent estimator of $\gamma(\theta)$.

Example: let X_1, X_2, \dots, X_n be a random sample from $N(\mu, \sigma^2)$. Show that sample mean \bar{X} is unbiased estimator and consistent estimator of μ .

Solution: Given $X_i \sim N(\mu, \sigma^2)$ then $E(X_i) = \mu$ and $V(X_i) = \sigma^2$

Let $T_n = \bar{X}$ then $E(T_n) = E(\bar{X}) = \frac{1}{n} \sum E(X_i) = \frac{1}{n} n\mu = \mu$

Therefore, \bar{X} is an unbiased estimator of μ .

Also, $V(T_n) = V(\bar{X}) = \frac{1}{n^2} \sum V(X_i) = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n} \rightarrow 0$ as $n \rightarrow \infty$

Therefore, \bar{X} is a consistent estimator of μ .

12.3.3 Most efficient estimator:

Let for large samples two consistent estimators say T_1 and T_2 be both distributed asymptotically normal. If there exist one say T_1 whose sampling variance is less than that of the other say T_2 then T_1 is called the most efficient estimator.

i.e., $V(T_1) < V(T_2)$ then T_1 is more efficient than T_2 or T_1 is called the most efficient estimator.

12.3.4 Efficiency: If T_1 is most efficient estimator with variance σ_1^2 and T_2 is any other estimator with variance σ_2^2 then the efficiency E of T_2 is defined as $E = \frac{\sigma_1^2}{\sigma_2^2} < 1$, always.

Example: Let X_1, X_2 be a random sample from a $N(\mu, \sigma^2)$. Find the efficiency of $T = \frac{1}{3}(X_1 + 2X_2)$ relative to $\bar{X} = \frac{1}{3} \sum_{i=1}^2 X_i$. Which is relatively more efficient?

Solution: Given $X_i \sim N(\mu, \sigma^2)$

Then $E(X_i) = \mu$ and $V(X_i) = \sigma^2$

Consider $V(T) = \frac{1}{3^2} (V(X_1) + 2^2 V(X_2)) = \frac{1}{9} (\sigma^2 + 4\sigma^2) = \frac{5\sigma^2}{9} = 0.555\sigma^2$

and $V(\bar{X}) = \frac{1}{3^2} \sum_{i=1}^2 V(X_i) = \frac{1}{9} (2\sigma^2) = \frac{2\sigma^2}{9} = 0.222\sigma^2$

Thus $E = \frac{V(\bar{X})}{V(T)} = \frac{0.222\sigma^2}{0.555\sigma^2} = 0.4 < 1$

Since $E < 1$, therefore, \bar{X} is more efficient than T.

12.3.5 SUFFICIENCY

An estimator is said to be sufficient for a parameter, if it contains all the information in the sample regarding the parameter.

Definition (Sufficient estimator): if $T = t(x_1, x_2, \dots, x_n)$ is an estimator of a parameter θ , based on a sample x_1, x_2, \dots, x_n of size n from the population with density $f(x, \theta)$ such that the conditional distribution of x_1, x_2, \dots, x_n given T , is independent of θ , then T is sufficient estimator of θ .

Example: let x_1, x_2, \dots, x_n be a random sample from a Bernoulli population with parameter 'p', $0 < p < 1$, i.e., $x_i = \begin{cases} 1, & \text{with probability } p \\ 0, & \text{with probability } q = 1 - p \end{cases}$. Find sufficient estimator of p.

Solution: $x_i, i=1,2,3,\dots,n$ follows Bernoulli distribution with parameter p then the p.m.f of Bernoulli distribution is $f(x) = p^x q^{1-x}, x = 0,1; 0 < p < 1; q = 1 - p$

Let $T = t(x_1, x_2, \dots, x_n) = x_1 + x_2 + \dots + x_n \sim B(n, p)$

This implies $P(T = k) = \binom{n}{k} p^k q^{n-k}, k = 0, 1, 2, \dots, n$.

The conditional distribution of (x_1, x_2, \dots, x_n) given T is

$$P(x_1 \cap x_2 \cap \dots \cap x_n | T) = \frac{P(x_1 \cap x_2 \cap \dots \cap x_n \cap T)}{P(T = k)} = \frac{p^k q^{n-k}}{\binom{n}{k} p^k q^{n-k}} = \frac{1}{\binom{n}{k}}$$

Since this does not depend on parameter 'p'

Therefore $T = \sum_{i=1}^n x_i$ is sufficient for 'p'.

Theorem: Factorization theorem (Neymann): The necessary and sufficient condition for a distribution to find sufficient statistic is provided by the factorization theorem due to Neymann.

Statement: $T = t(x)$ is sufficient for θ if and only if the joint density function L (say), of the sample values can be expressed in the form: $L = g_\theta[t(x)] \cdot h(x) \rightarrow (*)$ where $g_\theta[t(x)]$ depends on θ and x only through the value of $t(x)$ and $h(x)$ is independent of θ .

Remark: invariance property of sufficient estimator: if T is a sufficient estimator for the parameter θ and if $\Psi(T)$ is a one-to-one function of T , then $\Psi(T)$ is sufficient for $\Psi(\theta)$.

Example: let x_1, x_2, \dots, x_n be a random sample from $N(\mu, \sigma^2)$ population. Find the sufficient estimators for μ and σ^2 .

Solution: consider $\theta = (\mu, \sigma^2); -\infty < \mu < \infty; \sigma^2 > 0$.

$$\begin{aligned} \text{Then } L &= \prod_{i=1}^n f(x_i, \theta) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left\{\frac{-1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2\right\} \\ &= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left\{\frac{-1}{2\sigma^2} \left(\sum_{i=1}^n x_i^2 - 2\mu \sum_{i=1}^n x_i + n\mu^2\right)\right\} \\ &= g_\theta[t(x)].h(x) \\ \text{where } g_\theta[t(x)] &= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left\{\frac{-1}{2\sigma^2} (t_2(x) - 2\mu t_1(x) + n\mu^2)\right\} \end{aligned}$$

$$t(x) = \{t_1(x), t_2(x)\} = (\sum_{i=1}^n x_i, \sum_{i=1}^n x_i^2) \text{ \& } h(x) = 1$$

Thus $t_1(x) = \sum_{i=1}^n x_i$ is the sufficient for μ and $t_2(x) = \sum_{i=1}^n x_i^2$ is the sufficient for σ^2 .

12.3.6 Drawback of Point estimation:

Since no information is available regarding the reliability (closeness to the actual population parameter) of point estimation, therefore probability that a single sample statistic actually equals the population parameter is very small. For this reason, point estimates are rarely used alone to estimate population parameters. Hence, it is better to know the width of values within which the population parameters are expected to fall so that reliability of the estimate can be measured.

12.4 Method of estimation:

So far, we have studied the rules and requisites of a good estimator. Now we shall briefly discuss some of the important methods for obtaining estimators such as:

- Method of moments
- Method of maximum likelihood estimation.

12.4.1 Method of moments: Let $f(X_i; \theta_1, \theta_2, \dots, \theta_k)$ be the density function of the parent population with n parameters $\theta_1, \theta_2, \dots, \theta_k$. If μ'_r denotes the r^{th} moment about origin then $\mu'_r = E(X^r) = \int_{-\infty}^{\infty} x^r f(x; \theta_1, \theta_2, \dots, \theta_k) dx, r = 1, 2, \dots, k \rightarrow (1)$

Note. Let $X_i, i=1, 2, \dots, n$ be a random sample of size n from the given population. The method of moments consists in solving k -equations (1) for $\theta_1, \theta_2, \dots, \theta_k$ in terms of $\mu'_1, \mu'_2, \dots, \mu'_k$ and then replacing these moments (μ'_r) by the sample moments (m'_r). i.e., $\hat{\theta}_i = \theta_i(\mu'_1, \mu'_2, \dots, \mu'_k) = \theta_i(m'_1, m'_2, \dots, m'_k), \text{ for all } i = 1, 2, \dots, k.$

Example: obtain the moment estimator of Poisson distribution with parameter ' λ '.

Solution: W.K.T the p.m.f of Poisson distribution is

$$f(x) = \frac{e^{-\lambda} \lambda^x}{x!}, x = 0, 1, 2, \dots; \lambda \geq 0,$$

and the mean of Poisson distribution is λ . Therefore, $\mu'_1 = E(X) = \lambda$

Thus, $m'_1 = \bar{X}$ which is the sample mean, implies that $\hat{\lambda} = \bar{X}$ is the moment estimator of λ .

Example: Obtain the moment estimator of Normal distribution with parameter μ and σ^2 .

Solution: if $X \sim N(\mu, \sigma^2)$ then

$$\mu'_1 = E(X) = \mu \text{ and } \mu'_2 = \sigma^2 = E(X - \hat{\mu})^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \hat{\mu})^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

Therefore, moment estimator of μ is $\hat{\mu} = \mu'_1 = \bar{X}$, the sample mean and

Moment estimator of σ^2 is $\hat{\sigma}^2 = s^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$, the sample variance.

Example: obtain the moment estimator of Exponential distribution with parameter ' θ '.

Solution: W.K.T the p.d.f of Exponential distribution is $f(x) = \theta e^{-\theta x}, x > 0; \theta > 0$ and the mean of Exponential distribution is $1/\theta$.

Therefore, $\mu'_1 = E(X) = \frac{1}{\theta} \Rightarrow \bar{X} = \frac{1}{\theta}$

Thus, $\mu'_1 = \bar{X} = \frac{1}{\theta}$ which is the sample mean.

Which implies that $\hat{\theta} = \frac{1}{\bar{X}}$ is the moment estimator of θ .

12.4.2 Maximum Likelihood Estimation(MLE):

Definition of likelihood function: let x_1, x_2, \dots, x_n be the value of a random sample from a Bernoulli population with parameter θ . The likelihood function of the sample is given by $L(\theta; x_1, x_2, \dots, x_n) = L(\theta; X) = f(x_1, x_2, \dots, x_n; \theta)$ for values of θ within the given domain.

If X_i 's are independent, then $L(\theta) = \prod_{i=1}^n f(x_i, \theta)$.

Definition of Maximum likelihood estimation: An estimator $\hat{\theta}$ is said to be MLE of unknown parameter θ then it should maximize the likelihood function $L(\theta; X)$ i.e., $\hat{\theta} = \text{Sup}_{\theta \in \Theta} L(\theta; x_i) = L(\hat{\theta})$ i.e., $L > L(\theta)$.

Note. If the likelihood function $L(\theta; X)$ is differentiable with respect to θ , then one can use usual differentiation method i.e., $\frac{\partial \log L}{\partial \theta} = 0 \rightarrow (1)$ and $\frac{\partial^2 \log L}{\partial \theta^2} < 0 \rightarrow (2)$ then from (1) we get the value of θ .

- If $L(\theta; X)$ is not differentiable with respect to θ then we use ordered sample to estimate θ or indicator function.

12.4.3 Properties of MLE:

1. MLE's are always consistent but need not be unbiased estimator.
2. MLE need not be unique.
3. If MLE exists, it is most efficient estimator in the class of such estimators.
4. If a sufficient estimator exists, it is a function of MLE.

Example: If $X \sim N(\mu, \sigma^2)$, then find MLE of μ and σ^2 .

Solution: Given $X \sim N(\mu, \sigma^2)$ then the p.d.f of Normal distribution is $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}, -\infty < x < \infty; -\infty < \mu < \infty; \sigma > 0$

Then the likelihood function is $L = \prod_{i=1}^n f(x_i) = \prod_{i=1}^n \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2\right]$

$$= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2\right]$$

By taking logarithm on both sides, we get

$$\log L = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2$$

The simultaneous estimation of μ and σ^2 are

$$\frac{\partial \log L}{\partial \mu} = 0 \Rightarrow \frac{\partial \left[-\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2\right]}{\partial \mu} = 0$$

$$\Rightarrow \hat{\mu} = \frac{\sum_{i=1}^n x_i}{n} = \bar{X}, \text{ the sample mean.}$$

$$\text{and } \frac{\partial \log L}{\partial \sigma^2} = 0 \Rightarrow \frac{\partial \left[-\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2\right]}{\partial \sigma^2} = 0$$

$$\Rightarrow \hat{\sigma}^2 = \frac{\sum_{i=1}^n (x_i - \hat{\mu})^2}{n} = \frac{\sum_{i=1}^n (x_i - \bar{X})^2}{n} = s^2, \text{ the sample variance.}$$

Therefore the MLE of μ and σ^2 is $\hat{\mu} = \bar{X}$ and $\hat{\sigma}^2 = s^2$.

Example: If $X \sim P(\lambda)$, then find MLE of λ .

Solution: Given $X \sim P(\lambda)$ then the p.m.f of Poisson distribution is

$$f(x) = \frac{e^{-\lambda} \lambda^x}{x!}, x = 0, 1, 2, \dots; \lambda > 0.$$

Then the likelihood function is $L = \prod_{i=1}^n f(x_i) = \prod_{i=1}^n \frac{e^{-\lambda} \lambda^{x_i}}{x_i!} = \frac{(e^{-\lambda})^n \lambda^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!}$

By taking logarithm on both sides, we get

$$\ln L = -n\lambda + \sum_{i=1}^n x_i \ln(\lambda) - \ln\left(\prod_{i=1}^n x_i!\right)$$

$$\frac{\partial \log L}{\partial \lambda} = 0 \Rightarrow -n + \frac{1}{\lambda} \sum_{i=1}^n x_i - 0 = 0$$

$$\Rightarrow \hat{\lambda} = \frac{\sum_{i=1}^n x_i}{n} = \bar{X}, \text{ which is the MLE of } \lambda.$$

Example: Let X follows exponential distribution with mean θ , then find MLE of θ .

Solution: W.K.T the p.d.f of Exponential distribution is $f(x) = \frac{1}{\theta} e^{-x/\theta}, x > 0; \theta > 0$.

Then the likelihood function is $L = \prod_{i=1}^n f(x_i) = \prod_{i=1}^n \frac{1}{\theta} e^{-x_i/\theta}$

By taking logarithm on both sides, we get

$$\ln L = -n \ln \theta - \frac{\sum_{i=1}^n x_i}{\theta}$$

$$\frac{\partial \log L}{\partial \theta} = 0 \Rightarrow -\frac{n}{\theta} + \frac{\sum_{i=1}^n x_i}{\theta^2} = 0$$

$\Rightarrow \hat{\theta} = \frac{\sum_{i=1}^n x_i}{n} = \bar{X}$, which is the MLE of θ .

Exercise

- Let x_1, x_2, \dots, x_n be a random sample of size n drawn from a population with mean μ and variance σ^2 . Obtain an unbiased estimator of μ^2 .
- Let X_1, X_2, X_3 & X_4 be independent random variables such that $E(X_i) = \mu$ and $V(X_i) = \sigma^2$ for $i = 1, 2, 3, 4$. if $Y = \frac{X_1 + X_2 + X_3 + X_4}{4}$ and $Z = \frac{X_1 + 2X_2 + X_3 - X_4}{4}$. Examine whether Y and Z are unbiased estimators of μ ? What is the efficiency of Y relative to Z ?
- If $X_1, X_2, X_3, \dots, X_n$ is a random sample obtained from the density function $f(x, \theta) = \begin{cases} 1, & \theta < x < \theta + 1 \\ 0, & \text{otherwise} \end{cases}$. Show that the sample mean \bar{X} is an unbiased and consistent estimator of $\theta + \frac{1}{2}$.
- If the number of weekly accidents on a mile stretch of a particular road follows a Poisson distribution with λ . Then find the moment estimator and MLE of parameter λ on the basis of the following data.

No. of accidents	0	1	2	3	4	5	6
Frequency	10	15	14	9	6	2	1

- If a random sampling from a normal distribution with mean μ and variance σ^2 . Find the MME and MLE for μ and σ^2 from the following data: 3, 8, 16, 12, 10, 4, 5, 1.

12.5 Interval Estimation

After studying this part, we can be able to understand the concept of confidence interval and we are able to compute and interpret confidence interval for various measures.

One of the main objectives of statistics is to draw inferences about a population from the analysis of a sample drawn from that population. Two important problems in statistical inference are:

- Estimation
- Testing of hypothesis

Some basic definitions:

- Population:** The totality of units under consideration is called population. It may be finite or infinite population.
- Sample:** A part or a portion of population is called sample.
- Parameter:** A statistical constant of the population is called parameter.
- Statistic:** Any function of the random sample x_1, x_2, \dots, x_n that are being observed, say $T_n(X) = T_n(x_1, x_2, \dots, x_n)$ is called a statistic.
- Notations of different population parameter and sample statistic:**

Population parameter	Sample statistic
Population mean - μ	Sample mean - \bar{x}
Population variance - σ^2	Sample variance - S^2
Population standard deviation - σ	Sample standard deviation - S
Population proportion- P	Sample proportion - p
Population size - N	Sample size - n

- **Parameter space:** The set of all possible values of population parameter is called the parameter space. It is denoted by Θ .
For example, if $X \sim N(\mu, \sigma^2)$, then the parameter space is $\Theta = \{(\mu, \sigma^2): -\infty < \mu < \infty, \sigma^2 \geq 0\}$.
- **Sample space:** The set of all possible values of samples are called sample space. It is denoted by S .
- **Estimation:** It is a method of obtaining the most likely values of population parameter using statistic.
- **Estimator:** If a statistic is used to estimate an unknown parameter of the distribution, then it is called an estimator.
- **Estimate:** A particular value of an estimator, say $T_n(X)$ is called an estimate of θ .
- The statistic, say $T_n(X)$ whose distribution concentrates as closely as possible near the true value of the parameter may be regarded as the best estimate.
- There are two types of estimation: 1) point estimation, 2) interval estimation.
- **Point estimation:** If a single value is used to estimate population parameter, then the estimation is called point estimation.
- **Interval estimation:** If an interval $[c_1, c_2]$ is used to estimate the population parameter, then it is called interval estimation. It is also called as confidence interval.

Example: the set of 80 students will get first class with mean marks between 60 to 70.

- Let $t = t(x_1, x_2, \dots, x_n)$, a function of sample value be an estimate of population parameter θ , with the sampling distribution given by $g(t, \theta)$. Here we make some reasonable probability statement about unknown parameter θ in the population by the technique of confidence interval.
- Let us determine two constants say c_1 and c_2 with ' α ' level of significance (either 5% or 1%) such that:

$$P(c_1 < \theta < c_2) = 1 - \alpha \rightarrow (1)$$

The quantities c_1 and c_2 , so determined are known as confidence limits.

- **Confidence interval:** An interval $[c_1, c_2]$ within which the unknown value of the population parameter is expected to lie, is called the confidence interval.
- **Confidence limits:** A boundary values c_1 and c_2 of the confidence interval are known as confidence limits. It is also called as fiducial limits or probable limits.
- **Confidence coefficient:** The probability that an interval $[c_1, c_2]$ within which the unknown value of the population parameter is expected to lie, is called the confidence coefficient. It is denoted by $(1-\alpha)$.

- **Level of significance:** It is the maximum size of rejecting the null hypothesis when it is actually true. It is denoted by α .

Thus, if we take $\alpha=0.05$ (or 0.01), we shall get 95% (or 99%) confidence limits.

To find c_1 and c_2 : let T_1 and T_2 be two statistic such that $P(T_1 < \theta < T_2) = 1 - \alpha$ where c_1 and c_2 is considered as two statistic T_1 and T_2 .

12.5.1 Confidence interval for mean (when variance is known):

Consider the interval estimate of μ . If a sample is selected from a normal population with mean μ and standard deviation σ for n is sufficiently large (or for large sample). Then

$$Z = \frac{\bar{X} - \mu}{\sigma} \sim N(0,1) \text{ and } P\left(-Z_{\frac{\alpha}{2}} < Z < Z_{\frac{\alpha}{2}}\right) = 1 - \alpha \rightarrow (*)$$

By substituting Z in (*) and on simplification, we get 100 (1- α) % confidence interval for unknown population mean (μ), that is

$$P\left(\bar{X} - Z_{\frac{\alpha}{2}}\left(\frac{\sigma}{\sqrt{n}}\right) < \mu < \bar{X} + Z_{\frac{\alpha}{2}}\left(\frac{\sigma}{\sqrt{n}}\right)\right) = 1 - \alpha$$

And $\bar{X} \pm Z_{\frac{\alpha}{2}}\left(\frac{\sigma}{\sqrt{n}}\right)$ are 100(1- α) % confidence limits for population mean(μ).

Example: If a random sample of size $n=64$ from a normal population with the variance $\sigma^2 = 185$ has the mean $\bar{X} = 64.3$. Construct a 95% confidence interval for the population mean μ .

Solution: Given $n=64$, $\sigma^2 = 185$, $\bar{X} = 64.3$, $\alpha=0.05$

We know that 100 (1- α) % confidence interval for population mean(μ) is

$$P\left(\bar{X} - Z_{\frac{\alpha}{2}}\left(\frac{\sigma}{\sqrt{n}}\right) < \mu < \bar{X} + Z_{\frac{\alpha}{2}}\left(\frac{\sigma}{\sqrt{n}}\right)\right) = 1 - \alpha$$

For $\alpha=0.05$, the critical value is $Z_{\frac{\alpha}{2}} = Z_{0.025} = 1.96$ (from standard normal table)

Lower limit= $\bar{X} - Z_{\frac{\alpha}{2}}\left(\frac{\sigma}{\sqrt{n}}\right) = 64.3 - 1.96\left(\frac{13.601}{\sqrt{64}}\right) = 60.967$ and

Upper limit= $\bar{X} + Z_{\frac{\alpha}{2}}\left(\frac{\sigma}{\sqrt{n}}\right) = 64.3 + 1.96\left(\frac{13.601}{\sqrt{64}}\right) = 67.632$

Therefore, 95% confidence interval for the population mean μ is (60.967, 67.632).

12.5.2 Confidence interval for difference of means (for large samples):

If \bar{X}_1 and \bar{X}_2 are the means of independent random samples of size n_1 and n_2 from normal population having the means μ_1 and μ_2 and the variances σ_1^2 and σ_2^2 . Then $\bar{X}_1 - \bar{X}_2$ is a random variable having a normal distribution with the mean $\mu = \mu_1 - \mu_2$ and variance $\sigma^2 = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$.

It follows that $Z = \frac{|\bar{X}_1 - \bar{X}_2| - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$ and $P\left(-Z_{\frac{\alpha}{2}} < Z < Z_{\frac{\alpha}{2}}\right) = 1 - \alpha \rightarrow (*)$

By substituting Z in (*) and on simplification, we get 100 (1- α) % confidence interval for difference of means, that is

$$P\left(|\bar{X}_1 - \bar{X}_2| - Z_{\frac{\alpha}{2}}\left(\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right) < \mu_1 - \mu_2 < |\bar{X}_1 - \bar{X}_2| + Z_{\frac{\alpha}{2}}\left(\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)\right) = 1 - \alpha$$

And $|\bar{X}_1 - \bar{X}_2| \pm Z_{\frac{\alpha}{2}}\left(\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$ are $100(1-\alpha)$ % confidence limits for difference of means.

Example: Following data refers to mean daily wages of workers of two factory A and B. construct 95% confidence limits for mean daily wages of workers.

Factory	No. of workers	Mean daily wages (Rs.)	S.D (Rs.)
A	200	195	20
B	450	200	30

Solution: Given $\bar{X}_1 = 195$, $\bar{X}_2 = 200$, $n_1 = 200$, $n_2 = 450$, $\sigma_1 = 20$, $\sigma_2 = 30$, $\alpha = 0.05$

We know that $100(1-\alpha)$ % confidence interval for difference of means, that is

$$P\left(|\bar{X}_1 - \bar{X}_2| - Z_{\frac{\alpha}{2}}\left(\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right) < \mu_1 - \mu_2 < |\bar{X}_1 - \bar{X}_2| + Z_{\frac{\alpha}{2}}\left(\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)\right) = 1 - \alpha$$

For $\alpha = 0.05$, the critical value is $Z_{\frac{\alpha}{2}} = Z_{0.025} = 1.96$ (from standard normal table)

$$\text{Lower limit} = |\bar{X}_1 - \bar{X}_2| - Z_{\frac{\alpha}{2}}\left(\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right) = |195 - 200| - 1.96\left(\sqrt{\frac{20^2}{200} + \frac{30^2}{450}}\right) = 1.08$$

$$\text{Upper limit} = |\bar{X}_1 - \bar{X}_2| + Z_{\frac{\alpha}{2}}\left(\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right) = |195 - 200| + 1.96\left(\sqrt{\frac{20^2}{200} + \frac{30^2}{450}}\right) = 8.92$$

Therefore, 95% confidence limits for difference of means is 1.08 and 8.92.

12.5.3 Confidence interval for mean (when variance is not known):

In order to construct an appropriate $100(1-\alpha)$ % confidence interval for μ when σ is unknown but $n \geq 30$, we replace σ by the value of the sample standard deviation(s) and proceed large sample case. However, when we are dealing with a sample from a normal population and $n < 30$, a $100(1-\alpha)$ % confidence interval for μ can be constructed by making use of the fact that the random variable,

$t = \frac{\bar{X} - \mu}{s/\sqrt{n-1}} \sim t_{(n-1)}$, where t follows t-distribution with (n-1) degrees of freedom and

$$P\left(-t_{\frac{\alpha}{2},(n-1)} < t < t_{\frac{\alpha}{2},(n-1)}\right) = 1 - \alpha \quad \rightarrow (*)$$

By substituting t in (*) and on simplification, we get $100(1-\alpha)$ % confidence interval for unknown population mean (μ) for small samples, that is

$$P\left(\bar{X} - t_{\frac{\alpha}{2},(n-1)}\left(\frac{s}{\sqrt{n-1}}\right) < \mu < \bar{X} + t_{\frac{\alpha}{2},(n-1)}\left(\frac{s}{\sqrt{n-1}}\right)\right) = 1 - \alpha$$

And $\bar{X} \pm t_{\frac{\alpha}{2},(n-1)}\left(\frac{s}{\sqrt{n-1}}\right)$ are $100(1-\alpha)$ % confidence limits for population mean (μ) for small samples.

Example: A paint manufacturer wants to determine the average drying time of a new interior wall paint. If for 12 test areas of equal size he obtained a mean drying time of 66.3 minutes and a S.D of 8.4 minutes, construct a 95% confidence interval for the true mean μ .

Solution: Given $n=12$ (small sample), $s = 8.4$, $\bar{X} = 66.3$, $\alpha=0.05$

We know that 100 (1- α) % confidence interval for mean(μ) is

$$P\left(\bar{X} - t_{\frac{\alpha}{2},(n-1)}\left(\frac{s}{\sqrt{n-1}}\right) < \mu < \bar{X} + t_{\frac{\alpha}{2},(n-1)}\left(\frac{s}{\sqrt{n-1}}\right)\right) = 1 - \alpha$$

For $\alpha=0.05$, the critical value is $t_{\frac{\alpha}{2},(n-1)} = t_{0.025,11} = 2.23$ (from t-distribution table)

Lower limit = $\bar{X} - t_{\frac{\alpha}{2},(n-1)}\left(\frac{s}{\sqrt{n-1}}\right) = 66.3 - 2.23\left(\frac{8.4}{\sqrt{11}}\right) = 60.652$ and

Upper limit = $\bar{X} + t_{\frac{\alpha}{2},(n-1)}\left(\frac{s}{\sqrt{n-1}}\right) = 66.3 + 2.23\left(\frac{8.4}{\sqrt{11}}\right) = 71.947$

Therefore, 95% confidence interval for the population mean μ is (60.652, 71.947).

12.5.4 Confidence interval for difference of means (for small samples, when variance is not known):

The procedure of estimating the difference between two means when σ_1^2 and σ_2^2 are unknown and sample sizes are small. If $\sigma_1^2 = \sigma_2^2 = \sigma^2$, a point estimate of the unknown common variances.

Pooled estimator is denoted by S_p^2 , we write $S_p^2 = \frac{n_1s_1^2 + n_2s_2^2}{n_1 + n_2 - 2}$. Then the test statistic is

$t = \frac{|\bar{X}_1 - \bar{X}_2| - (\mu_1 - \mu_2)}{\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t_{(n_1 + n_2 - 2)}$, where t follows t-distribution with $(n_1 + n_2 - 2)$ degrees of

freedom and $P\left(-t_{\frac{\alpha}{2},(n_1+n_2-2)} < t < t_{\frac{\alpha}{2},(n_1+n_2-2)}\right) = 1 - \alpha \rightarrow (*)$

By substituting t in (*) and on simplification, we get 100 (1- α) % confidence interval for difference of means for small samples, that is

$$P\left(|(\bar{X}_1 - \bar{X}_2)| - t_{\frac{\alpha}{2},(n_1+n_2-2)}\left(\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}\right) < \mu_1 - \mu_2 < |(\bar{X}_1 - \bar{X}_2)| + t_{\frac{\alpha}{2},(n_1+n_2-2)}\left(\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}\right)\right) = 1 - \alpha$$

And $|(\bar{X}_1 - \bar{X}_2)| \pm t_{\frac{\alpha}{2},(n_1+n_2-2)}\left(\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}\right)$ are 100(1- α) % confidence limits for difference of means for small samples.

Example: A study has been made to compare the nicotine contents of two brands of cigarettes. Ten cigarettes of brand A has an average nicotine content of 3.1mg with a S.D of 0.5mg, while eight cigarettes of brand B has an average nicotine content of 2.7mg with a S.D of 0.7mg, assuming that the two sets of data are random samples from normal populations with equal variances. Construct a 95% confidence interval for the true difference in the average nicotine content of the two brands of cigarettes.

Solution: Given $\bar{X}_1 = 3.1$, $\bar{X}_2 = 2.7$, $n_1 = 10$, $n_2 = 8$, $s_1 = 0.5$, $s_2 = 0.7$, $\alpha = 0.05$

We know that 100 (1- α) % confidence interval for difference of means for small samples, that is

$$P\left(|(\bar{X}_1 - \bar{X}_2)| - t_{\frac{\alpha}{2}, (n_1+n_2-2)} \left(\sqrt{S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}\right) < \mu_1 - \mu_2\right. \\ \left. < |(\bar{X}_1 - \bar{X}_2)| + t_{\frac{\alpha}{2}, (n_1+n_2-2)} \left(\sqrt{S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}\right)\right) = 1 - \alpha$$

For $\alpha = 0.05$, the critical value is $t_{\frac{\alpha}{2}, (n_1+n_2-2)} = t_{0.025, 16} = 2.120$ (from t-distribution table) and

$$S_p^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2} = \frac{10(0.25) + 8(0.49)}{10 + 8 - 2} = 0.401$$

$$\text{Lower limit} = |(\bar{X}_1 - \bar{X}_2)| - t_{\frac{\alpha}{2}, (n_1+n_2-2)} \left(\sqrt{S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}\right) \\ = |(3.1 - 2.7)| - 2.120 \left(\sqrt{0.401 \left(\frac{1}{10} + \frac{1}{8}\right)}\right) = -0.236, \text{ and}$$

$$\text{Upper limit} = |(\bar{X}_1 - \bar{X}_2)| + t_{\frac{\alpha}{2}, (n_1+n_2-2)} \left(\sqrt{S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}\right) \\ = |(3.1 - 2.7)| + 2.120 \left(\sqrt{0.401 \left(\frac{1}{10} + \frac{1}{8}\right)}\right) = 1.036$$

Therefore, 95% confidence interval for difference of means for small sample is (-0.236, 1.036).

12.5.5 Confidence interval for population proportion(P):

A point estimate of the proportion 'p' in a binomial experiment is given by the statistic $p = x/n$, where x =number of successes in n -trials. By central limit theorem, for sufficiently large, \hat{p} is approximately normally distributed with mean $\mu = E(\hat{p}) = E\left(\frac{X}{n}\right) = \frac{nP}{n} = P$ and variance $\sigma^2 = V\left(\frac{X}{n}\right) = \frac{nPQ}{n^2} = \frac{PQ}{n}$. Therefore the test statistic is

$$Z = \frac{\hat{p} - P}{\sqrt{\hat{p}\hat{q}/n}} \sim N(0,1) \text{ and } P\left(-Z_{\frac{\alpha}{2}} < Z < Z_{\frac{\alpha}{2}}\right) = 1 - \alpha \rightarrow (*)$$

By substituting Z in (*) and on simplification, we get 100 (1- α) % confidence interval for unknown population proportion(P), that is

$$P\left(\hat{p} - Z_{\frac{\alpha}{2}} \left(\sqrt{\hat{p}\hat{q}/n}\right) < P < \hat{p} + Z_{\frac{\alpha}{2}} \left(\sqrt{\hat{p}\hat{q}/n}\right)\right) = 1 - \alpha$$

And $\hat{p} \pm Z_{\frac{\alpha}{2}} \left(\sqrt{\hat{p}\hat{q}/n}\right)$ are 100(1- α) % confidence limits for population proportion(P).

Example: Let p equal to the proportion of Indians who select jogging as one of their recreational activities. If 1497 out of a random sample of 5757 selected jogging, find an approximate 95% confidence interval for p .

Solution: Given $n=5757$, $x=1497$, $p = \frac{x}{n} = \frac{1497}{5757} = 0.26 = \hat{p}$, $\hat{q} = 1 - 0.26 = 0.74$, $\alpha=0.05$

We know that 100 (1- α) % confidence interval for population proportion(P) is

$$P\left(\hat{p} - Z_{\frac{\alpha}{2}}\left(\sqrt{\frac{\hat{p}\hat{q}}{n}}\right) < P < \hat{p} + Z_{\frac{\alpha}{2}}\left(\sqrt{\frac{\hat{p}\hat{q}}{n}}\right)\right) = 1 - \alpha$$

For $\alpha=0.05$, the critical value is $Z_{\frac{\alpha}{2}} = Z_{0.025} = 1.96$ (from standard normal table)

$$\text{Lower limit} = \hat{p} - Z_{\frac{\alpha}{2}}\left(\sqrt{\frac{\hat{p}\hat{q}}{n}}\right) = 0.26 - 1.96\left(\sqrt{\frac{0.26*0.74}{5757}}\right) = 0.248 \text{ and}$$

$$\text{Upper limit} = \hat{p} + Z_{\frac{\alpha}{2}}\left(\sqrt{\frac{\hat{p}\hat{q}}{n}}\right) = 0.26 + 1.96\left(\sqrt{\frac{0.26*0.74}{5757}}\right) = 0.271$$

Therefore, 95% confidence interval for the population proportion is (0.248, 0.271).

12.5.6 Confidence interval for difference of proportions:

Given a random sample of size n from a normal population, we can obtain 100 (1- α) % confidence interval for P_1-P_2 can be established by considering the sampling distribution of p_1-p_2 . P_1 and P_2 are approximately normally distributed with mean $\mu=P_1-P_2$ and variance $\sigma^2 = \frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}$ respectively. By choosing independent samples from the two populations, P_1 and P_2 will be independent. Therefore, the test statistic is

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - (P_1 - P_2)}{\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}} \sim N(0,1) \text{ where } \hat{p}_1 = \frac{x_1}{n_1}, \hat{p}_2 = \frac{x_2}{n_2}, \hat{q}_1 = 1 - \hat{p}_1, \hat{q}_2 = 1 - \hat{p}_2 \text{ and } P\left(-Z_{\frac{\alpha}{2}} <$$

$$Z < Z_{\frac{\alpha}{2}}\right) = 1 - \alpha \rightarrow (*)$$

By substituting Z in (*) and on simplification, we get 100 (1- α) % confidence interval for difference of proportions, that is

$$P\left\{\left|\hat{p}_1 - \hat{p}_2\right| - Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}} < (P_1 - P_2) < \left|\hat{p}_1 - \hat{p}_2\right| + Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}\right\} = 1 - \alpha$$

And $\left|\hat{p}_1 - \hat{p}_2\right| \pm Z_{\frac{\alpha}{2}}\left(\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}\right)$ are 100(1- α) % confidence limits for difference of proportions.

Example: A survey of 436 workers showed that 192 of them said that it was seriously unethical to monitor employee email. When 121 senior-level bosses were surveyed, 40 said that it is seriously unethical to monitor employee email (based on data from a Gallup poll). Construct 95% confidence interval of the difference between two population proportion.

Solution: Given $n_1=436$, $n_2=121$, $x_1=192$, $x_2=40$, $\hat{p}_1 = \frac{x_1}{n_1} = \frac{192}{436} = 0.440$, $\hat{p}_2 = \frac{x_2}{n_2} = \frac{40}{121} =$

0.331 , $\hat{q}_1 = 1 - \hat{p}_1 = 1 - 0.440 = 0.560$, $\hat{q}_2 = 1 - \hat{p}_2 = 1 - 0.331 = 0.669$

We know that 100 (1- α) % confidence interval for difference of proportions is

$$P\left((\hat{p}_1 - \hat{p}_2) - Z_{\frac{\alpha}{2}}\left(\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}\right) < (P_1 - P_2) < (\hat{p}_1 - \hat{p}_2) + Z_{\frac{\alpha}{2}}\left(\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}\right)\right) = 1 - \alpha$$

For $\alpha=0.05$, the critical value is $Z_{\frac{\alpha}{2}} = Z_{0.025} = 1.96$ (from standard normal table)

$$\text{Lower limit} = (\hat{p}_1 - \hat{p}_2) - Z_{\frac{\alpha}{2}}\left(\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}\right) = ((0.440 - 0.331) - 1.96\left(\sqrt{\frac{0.440*0.560}{436} + \frac{0.331*0.669}{121}}\right) = 0.032 \text{ and}$$

$$\text{Upper limit} = (\hat{p}_1 - \hat{p}_2) + Z_{\frac{\alpha}{2}}\left(\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}\right) = ((0.440 - 0.331) + 1.96\left(\sqrt{\frac{0.440*0.560}{436} + \frac{0.331*0.669}{121}}\right) = 0.186$$

Therefore, 95% confidence interval of difference of two population proportion is (0.032, 0.186).

12.5.7 Confidence interval for population variance(σ^2):

If a sample size of n is drawn from a normal population with variance σ^2 and the sample variance s^2 is computed, we obtain a value of the statistic S^2 . This computed variance will be used as a point estimate of σ^2 . Hence the statistic S^2 is called an estimator of σ^2 . An interval estimate of σ^2 can be established by using the test statistic $\chi^2 = \frac{ns^2}{\sigma^2} \sim \chi^2_{(n-1)}$, where χ^2 follows chi-square distribution with $(n-1)$ degrees of freedom and

$$P\left(\chi^2_{1-\frac{\alpha}{2}} < \chi^2 < \chi^2_{\frac{\alpha}{2}}\right) = 1 - \alpha \quad \rightarrow (*)$$

where $\chi^2_{1-\frac{\alpha}{2}}$ and $\chi^2_{\frac{\alpha}{2}}$ are the table values of chi-square distribution with $(n-1)$ degrees of freedom. By substituting χ^2 in (*) and on simplification, we get 100 (1- α) % confidence interval for population variance, that is

$$P\left(\frac{ns^2}{\chi^2_{\frac{\alpha}{2}}} < \sigma^2 < \frac{ns^2}{\chi^2_{1-\frac{\alpha}{2}}}\right) = 1 - \alpha$$

And $\frac{ns^2}{\chi^2_{\frac{\alpha}{2}}}$, $\frac{ns^2}{\chi^2_{1-\frac{\alpha}{2}}}$ are 100(1- α) % confidence limits for population variance.

Example: In 16 tests runs the gasoline consumption of an experimental engine had a standard deviation of 2.2 gallons. construct a 99% confidence interval for σ^2 measuring the true variability of gasoline consumption of the engine.

Solution: Given $n=16$, $s=2.2$, $s^2=4.840$, $\alpha=0.01$

We know that 100 (1- α) % confidence interval for population variance is

$$P\left(\frac{ns^2}{\chi^2_{\frac{\alpha}{2},(n-1)}} < \sigma^2 < \frac{ns^2}{\chi^2_{1-\frac{\alpha}{2},(n-1)}}\right) = 1 - \alpha$$

From chi-square table, $\chi^2_{1-\frac{\alpha}{2},(n-1)} = \chi^2_{0.995,15} = 4.601$ and $\chi^2_{\frac{\alpha}{2},(n-1)} = \chi^2_{0.005,15} = 32.801$

$$\text{Lower limit} = \frac{ns^2}{\chi^2_{\frac{\alpha}{2},(n-1)}} = \frac{16 \times 4.840}{32.801} = 2.360$$

$$\text{Upper limit} = \frac{ns^2}{\chi^2_{1-\frac{\alpha}{2},(n-1)}} = \frac{16 \times 4.840}{4.601} = 16.831$$

Therefore, 99% confidence interval for population variance is (2.360, 16.831).

12.5.8 Confidence interval for the ratio of two population variances:

A point estimate of the ration of two population variances $\frac{\sigma_1^2}{\sigma_2^2}$ is given by the ratio $\frac{s_1^2}{s_2^2}$ of the sample variances. If σ_1^2 and σ_2^2 are the variances of normal distribution, we can establish an interval estimate of $\frac{\sigma_1^2}{\sigma_2^2}$ by suing the test statistic

$$F = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} = \frac{s_1^2\sigma_2^2}{s_2^2\sigma_1^2} \text{ and } P(F_{1-\frac{\alpha}{2},(n_1-1), n_2-1} < F < F_{\frac{\alpha}{2},(n_1-1), n_2-1}) = 1 - \alpha \rightarrow (*), \quad \text{where}$$

$F_{1-\frac{\alpha}{2},(n_1-1), n_2-1}$ and $F_{\frac{\alpha}{2},(n_1-1), n_2-1}$ are the table values of F-distribution with (n_1-1, n_2-1) degrees of freedom. By substituting F in (*) and on simplification, we get 100 (1- α) % confidence interval of the ratio of two population variances, that is

$$P\left(\frac{s_1^2}{s_2^2 F_{\frac{\alpha}{2},(n_1-1), n_2-1}} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{s_1^2}{s_2^2 F_{1-\frac{\alpha}{2},(n_1-1), n_2-1}}\right) = 1 - \alpha$$

And $\frac{s_1^2}{s_2^2 F_{\frac{\alpha}{2},(n_1-1), n_2-1}}$, $\frac{s_1^2}{s_2^2 F_{1-\frac{\alpha}{2},(n_1-1), n_2-1}}$ are 100(1- α) % confidence limits for ratio of two population variances.

Remark: we have the following reciprocal relation between the upper and lower ' α ' significance points of F-distribution:

$$F_{\frac{\alpha}{2},(n_1-1), n_2-1} \times F_{1-\frac{\alpha}{2},(n_2-1), n_1-1} = 1 \Rightarrow F_{\frac{\alpha}{2},(n_1-1), n_2-1} = \frac{1}{F_{1-\frac{\alpha}{2},(n_2-1), n_1-1}}$$

Example: construct 90% confidence interval in the variability of amount of fill in 475gm and 850gm cornflakes boxes.

Sample	Size	Sample variance
850gm boxes	4	40.917
475gm boxes	15	16.714

Solution: $n_1=4, n_2=15, s_1^2 = 40.917, s_2^2 = 16.714, \alpha=0.1$

We know that 100 (1- α) % confidence interval for the ratio of two population variances is given by

$$P\left(\frac{s_1^2}{s_2^2 F_{\frac{\alpha}{2}}(n_1-1, n_2-1)} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{s_1^2}{s_2^2 F_{1-\frac{\alpha}{2}}(n_1-1, n_2-1)}\right) = 1 - \alpha$$

From F-table, $F_{\frac{\alpha}{2}}(n_1-1, n_2-1) = F_{0.05,(3,14)} = 3.3439$ and $F_{1-\frac{\alpha}{2}}(n_1-1, n_2-1) = \frac{1}{F_{\frac{\alpha}{2}}(n_2-1, n_1-1)} = \frac{1}{F_{0.05,(14,3)}} = \frac{1}{3.3439} = 0.229$

$$\text{Lower limit} = \frac{s_1^2}{s_2^2 F_{\frac{\alpha}{2}}(n_1-1, n_2-1)} = \frac{40.917}{16.714 \times 3.3439} = 0.732$$

$$\text{Upper limit} = \frac{s_1^2}{s_2^2 F_{1-\frac{\alpha}{2}}(n_1-1, n_2-1)} = \frac{40.917}{16.714 \times 0.229} = 10.690$$

Therefore, 90% confidence interval for the ratio of two population variances is (0.732, 10.690).

Estimating sample size:

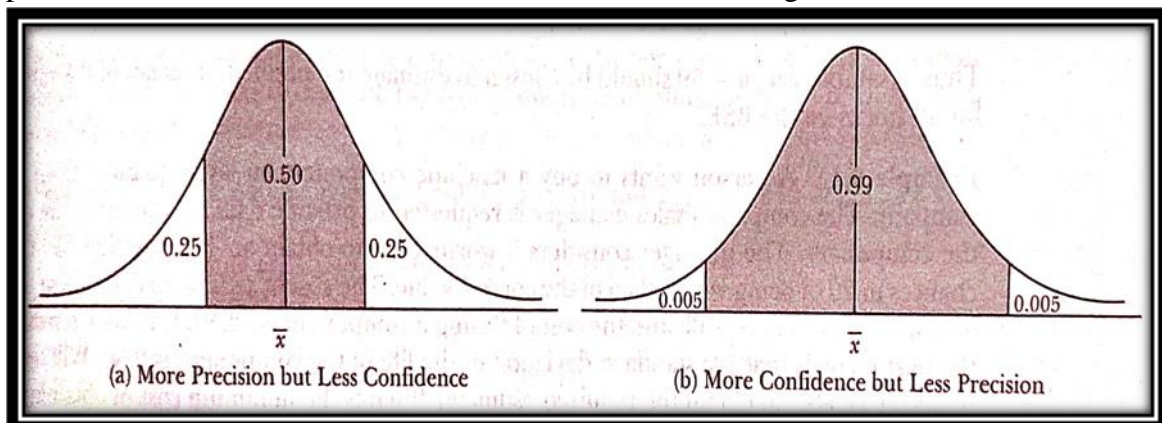
For statistical inference based on sample statistic, estimating suitable sample size is essential. The standard error of sampling distribution of sample statistic is inversely proportional to the sample size n. Also, width of the confidence interval can be decreased by increasing the sample size.

Precision of confidence interval:

The true population parameter value is determined by the width of the confidence interval. If the width of the confidence interval is narrow, then estimate will be more precise and vice-versa. The width of confidence interval is influenced by

- Specified level of confidence
- Sample size
- Population standard deviation

To gain more precision or confidence, or both, the sample size needs to be increased provided variability in the population is less. However, if the sample size cannot be increased, then it may increase the cost of sampling. Hence with the same sample size, the desired level of precision can be gained only by decreasing the confidence level so that estimate may become close to the true population parameter. The difference between precision and confidence levels are illustrated below in the figure.



Exercise

1. A random sample of height of 60 students from large population of students in a university having S.D. of 0.70ft has an average height of 5.5ft. Find 95% confidence interval for the average height of all students of the university.
2. The mean number of production of which for 80 villagers and 100 villagers taken as a sample from a locality are 400pounds and 380pounds respectively. The S.D's of production of these samples are 20 and 30pounds respectively. Obtain 99% C.I for difference of production.
3. A sample of 25 students is found to have average weight of 50 kg with a S.D. of 6 kg. Set up 99% C.I for the average weight of the population.
4. Two samples gave the following results.

Sample size	12	8
Sample mean	15	14
Standard deviation	6	4

Find the confidence interval for the difference of 2 means at 95% confidence coefficient.

5. The SD of heights of 18 male students chosen at random in a school is 3.4. Find 95% and 99% confidence limits of the SD of all males students at the school.
6. Obtain 90% confidence limits for the ratio of two variance from the following:

Data	Size of sample	mean	S.D.
Sample A	80	60	4
Sample B	100	62	6

BLOCK – IV **(TESTS OFSIGNIFICANCE)**

UNIT 13: INTRODUTION TO TESTING OF HYPOTHESIS

UNIT 14: LARGE SAMPLE TESTS

UNIT 15: SMALL SAMPLE TESTS - I

UNIT 16: SMALL SAMPLE TESTS - II

UNIT 13

INTRODUCTION TO TESTING OF HYPOTHESIS

13.1 Objective

After studying this chapter, we will be able to explain why hypothesis testing is important with the help of some basic definitions involved in testing of hypothesis.

13.2 Introduction

Testing of hypothesis is one of the important branches of statistical inference. It helps in decision making about population parameter and tests the validity of the claim (assertion, belief, assumption or statement), also called hypothesis. It refers to the process of evaluating whether a statistical result or observation is meaningful or due to chance.

The goal of significance testing is to make inferences about a population based on sample data. It involves comparing the observed data to a null hypothesis, which represents a default assumption or no effect scenario. By assessing the likelihood of observing the data under the null hypothesis, we can determine if there is sufficient evidence to reject or fail to reject the null hypothesis in favour of an alternative hypothesis.

A very important aspect of the sampling theory is the study of the tests of significance, which enable us to decide on the basis of the sample results, if

- ❖ The deviation between the observed sample statistic and the hypothetical parameter value.
Or
- ❖ The deviation between the two independent sample statistics; is significant or might be attributed to chance or the fluctuations of sampling.

Significance testing helps researchers to draw conclusions from data and make informed or valid decisions. However, it's essential to consider the limitations and assumptions of the chosen test and interpret the results in the context of the specific study or problem.

13.3 Basic concepts of Testing of Hypothesis:

- **Statistical hypothesis:** A statistical statement regarding population parameter which we want to verify on the basis of information available from a sample. It is denoted by H.
For example, H: the average marks scored by large group of students is 80[i.e., $\mu=80$]
- Statistical hypothesis may be simple or composite hypothesis.
- **Simple hypothesis:** If the statistical hypothesis specifies population parameter completely, then it is called simple hypothesis.
For example, if $X \sim N(\mu, \sigma^2)$, then H: $\mu = \mu_0, \sigma^2 = \sigma_0^2$.
- **Composite hypothesis:** If the statistical hypothesis does not specify population parameter completely, then it is called composite hypothesis.
For example, if $X \sim N(\mu, \sigma^2)$, then H: $\mu = \mu_0$, unknown σ^2 .
- There are two types of statistical hypothesis:
 1. Null hypothesis and
 2. Alternative hypothesis.
- **Null hypothesis (H_0):** it is the hypothesis which is being tested for possible rejection under the assumption that it is true. It is a definite statement about the population parameter.
For example, to test the effectiveness of training, the null hypothesis is
 H_0 : the training is not effective.
- **Alternative hypothesis (H_1):** It is complementary to null hypothesis. It is the hypothesis which is accepted when the null hypothesis is rejected.
For example, $H_0: \mu = \mu_0$ against $H_1: \mu \neq \mu_0$ (for two tailed test)
or $H_1: \mu < \mu_0$ (for left tailed test)
or $H_1: \mu > \mu_0$ (for right tailed test)
- **Errors in sampling:**

The main objective in sampling theory is to draw a valid conclusion about population parameter on the basis of sample results. In practice, we decide to accept or reject a lot after examining a sample from it. As we are likely to commit the following two types of errors:

		Decision from sample	
		Reject H_0	Accept H_0
True state	H_0 true	Wrong (type I error)	Correct
	H_0 false	Correct	Wrong (type II error)

1. **Type I error:** the error that occur by rejecting the null hypothesis when it is actually true. It is also called as First kind error.
2. **Type II error:** the error that occurs by accepting the null hypothesis when it is actually not true. It is also called as Second kind error.

Note: once we commit type II error, then there is no chance of making correct decision. Therefore, type II error is more significant error than type I error. So, one fixes type I error probability as it minimizes probability of type II error.

- **Size of the test:** The probability of rejecting the null hypothesis when it is actually true. It is denoted by ' α '.

$$\alpha = P(\text{type I error}) = P(\text{reject } H_0 | H_0 \text{ is true})$$

Here α is also called **size of Type I error**.

- Similarly, **size of Type II error** is defined as the probability of accepting the null hypothesis when it is actually not true. It is denoted by ' β '.

$$\beta = P(\text{type II error}) = P(\text{accept } H_0 | H_0 \text{ is not true})$$

- **Power of the test:** The probability of rejecting the null hypothesis when it is actually not true. It is denoted by ' $1-\beta$ '.

$$1 - \beta = P(\text{reject } H_0 | H_0 \text{ is not true}) = 1 - P(\text{type II error})$$

- **Example:** let X follows Exponential distribution with parameter θ . Obtain sizes of type I and type II error and power of the test if an observation takes at random exceeds 5 then it leads to reject H_0 such that $H_0: \theta=3$ against $H_1: \theta=4$.

Solution: $X \sim \text{exp}(\theta)$, where θ is the parameter.

Then the p.d.f of exponential distribution is $f(x) = \begin{cases} \theta e^{-\theta x}, & \theta > 0, 0 < x < \infty \\ 0, & \text{otherwise} \end{cases}$

Here the rejection region = $W = \{x: x > 5\}$

And the acceptance region is $\bar{W} = \{x: x \leq 5\}$

Size of type I error = $\alpha = P(\text{Type I error}) = P(\text{reject } H_0 | H_0: \theta=3)$

$$= P(x > 5 | H_0: \theta=3)$$

$$= \int_5^{\infty} 3e^{-3x} dx$$

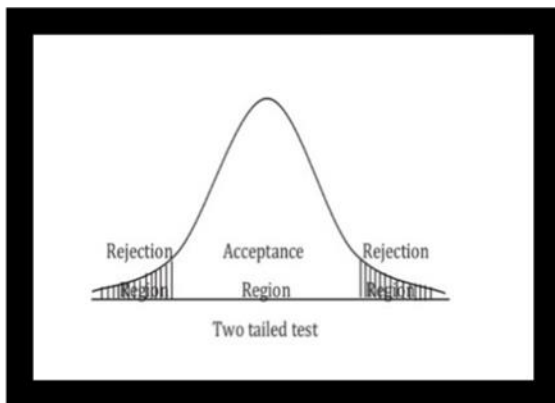
$$\begin{aligned}
&= 3 \left(\frac{e^{-3x}}{-3} \right)_{x=5}^{x=\infty} \\
&= -(e^{-3\infty} - e^{-3(5)}) \\
&= 0.000000305
\end{aligned}$$

Size of type II error = β = P(Type II error) = P(accept H_0 | $H_1: \theta=4$)

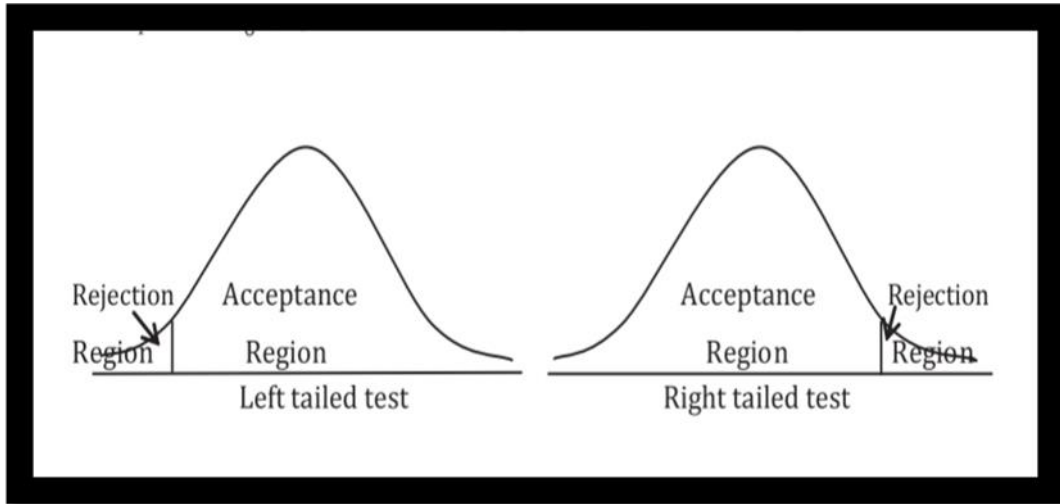
$$\begin{aligned}
&= P(x \leq 5 | H_1: \theta=4) \\
&= \int_0^5 4e^{-4x} dx \\
&= 4 \left(\frac{e^{-4x}}{-4} \right)_{x=0}^{x=5} \\
&= -(e^{-3(5)} - e^{-3(0)}) \\
&= 0.9999
\end{aligned}$$

Power of the test = $1 - \beta = 1 - 0.9999 = 0.00001$

- **Critical region:** A region in the sample space which leads to reject the null hypothesis is termed as critical region or rejection region.
- **Acceptance region:** A region in the sample space which leads to accept the null hypothesis is termed as acceptance region.
- **Critical value:** A value which separates critical region and acceptance region is called critical value or significance value or table value. It depends upon α level of significance and alternative hypothesis.
- **Level of significance:** it is the maximum probability of rejecting the null hypothesis when it is actually true. It is denoted by α .
- **Two tailed test:** it is a test of statistical hypothesis where rejection region is located at both the tails of the normal curve. This is used when H_1 is of the type not equal (\neq).



- **One tailed test:** it is a test of statistical hypothesis where rejection region is located at only one tail of the normal curve. It may be left tailed test or right tailed test. This is used when H_1 is of the type less than ($<$) or more than ($>$).



- Test statistic: The testing of hypothesis is conducted for the distribution of statistic is called test statistic.

$$\text{Test statistic} = \frac{\text{sample statistic} - \text{hypothetical value}}{\text{standard error of statistic}}$$

Exercise problems:

1. Define simple and composite hypothesis with an example.
2. Differentiate Null and alternative hypothesis.
3. Explain error in sampling.
4. Explain one tailed and two tailed test.
5. If $X \geq 2$, is the critical region for testing $H_0: \theta=1$ against $H_1: \theta=2$, on the basis of the single observation from the population $f(x) = \begin{cases} \theta e^{-\theta x}, & \theta > 0, 0 < x < \infty \\ 0, & \text{otherwise} \end{cases}$. Obtain the values of sizes of type I and type II errors and power of the test.
6. Let p be the probability that a coin will fall head in a single toss in order to test $H_0: p=1/3$ against $H_1: p = 3/4$. The coin is tossed 5 times and H_0 is rejected if more than 4 heads are obtained. Find the value of α , β and $1-\beta$.
7. Given the frequency function $f(x) = \begin{cases} \frac{1}{\theta}, & 0 \leq x \leq \theta \\ 0, & \text{otherwise} \end{cases}$ and to test $H_0: \theta=1$ against $H_1: \theta=2.5$ by means of a single observed value of x with the interval $0.5 \leq x$ as the critical region. Find the value of α , β and $1-\beta$.

UNIT 14

LARGE SAMPLE TESTS

14.1 Objectives

After completion of this Unit, you should

- ❖ Know the meaning of large sample test, assumptions and the applications.
- ❖ Know how to test the population mean, equality of means of two populations.

14.2 INTRODUCTION

Large sample test is a statistical test procedure used to test a hypothesis about a population parameter when the sample size is large ($n \geq 30$). Since large samples approach normal distribution; therefore, the tests are based on normal distribution. This leads to the use of standard normal distribution tables or z-scores to test hypotheses about the population mean or proportion.

In other words, large sample tests provide a practical and efficient way to make statistical inferences when dealing with sufficiently large sample sizes. They have wide range of applications in testing of hypothesis, estimation of confidence interval, and comparative analyses in various fields, including social sciences, economics, healthcare, and quality control.

14.3 General test procedure for Large Sample Test

Step 1: Constructing the null hypothesis (H_0).

Step 2: Constructing the alternative hypothesis (H_1). It helps to decide whether to use two tailed or one (left or right) tailed test.

Step 3: computation of test statistic; under H_0

$$Z = \frac{\text{Relevant statistic} - \text{Hypothetical value}}{\text{Standard Error}} \sim N(0, 1)$$

Step 4: Depending on the alternative hypothesis (H_1) and level of significance (α), the critical (Table) value i.e., $\pm Z_{\alpha/2}$ (for two tailed) or Z_{α} (for one tailed) is chosen.

Step 5: If the calculated value of the test statistic (Z) lies in the acceptance region, then we

Do not reject H_0 . Otherwise we reject H_0 .

i.e., for two tailed test, if $-Z_{\alpha/2} \leq Z \leq +Z_{\alpha/2}$, then we do not reject H_0 .

i.e., for left tailed test, if $Z \geq -Z_{\alpha}$ then we do not reject H_0 .

i.e., for Right tailed test, if $Z \leq Z_{\alpha}$ then we do not reject H_0 .

Otherwise H_0 is rejected.

Applications of Large Sample Test:

Applications of Large sample test are:

1. It is used to test for population mean.
2. It is used to test for equality of means of two populations.
3. It is used to test for population proportion.
4. It is used to test for equality of proportions of two populations.
5. It is used in the construction of confidence intervals.
6. It is used to estimate a population parameter and to determine a range of values within which the true parameter is likely to fall with a specified level of confidence.
7. It is used in Quality Control to assess whether a production process is operating within specified limits.
8. It is used in Regression analysis to test the significance of regression coefficients or to compare the performance of different regression models.

14.4 Large sample test procedure to test for population mean

Here we test the characteristic of the population mean μ , from the large sample ($n \geq 30$) which is drawn from that population. And the test procedure is as follows.

Step 1: H_0 : The population mean. i.e., $\mu = \mu_0$.

Step 2: H_1 : The population mean. i.e., $\mu \neq \mu_0$. (Two tailed test)

OR

The population mean. i.e., $\mu < \mu_0$. (Left tailed test)

OR

The population mean. i.e., $\mu > \mu_0$. (Right tailed test)

Step 3: computation of test statistic; under H_0

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$$

where \bar{X} is the sample mean, σ is the population standard deviation, when σ is unknown, then replace it by sample standard deviation's' and n is the sample size.

Step 4: Depending on the alternative hypothesis (H_1) and level of significance (α), the critical (Table) value i.e., $\pm Z_{\alpha/2}$ (for two tailed) or Z_α (for one tailed) is chosen.

Step 5: If the calculated value of the test statistic (Z) lies in the acceptance region, then we Do not reject H_0 . Otherwise we reject H_0 .

i.e., for two tailed test, if $-Z_{\alpha/2} \leq Z \leq +Z_{\alpha/2}$, then we do not reject H_0 .

i.e., for left tailed test, if $Z \geq -Z_\alpha$ then we do not reject H_0 .

i.e., for Right tailed test, if $Z \leq Z_\alpha$ then we do not reject H_0 .

Otherwise H_0 is rejected.

Remark: If the level of significance is not specified in the problem then by default we use 5% level of significance.

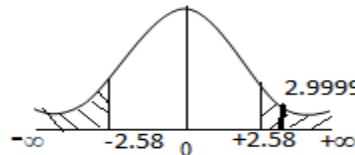
Example 1: A sample of 144 students is chosen from a university. The average height of these students is 160 cm and the standard deviation is 8 cm. At 1% level of significance can we assume that the average height of these university students is 158 cm?

Solution: Given: $n = 144$, $\bar{X} = 160$, $s = 8$, $\mu = 158$ and $\alpha = 1\%$.

H_0 : The average height of the university students is 158 cm. i.e., $\mu = 158$ cm

H_1 : The average height of the university students is not equal to 158 cm. i.e., $\mu \neq 158$ cm. (Two tailed test)

$$\begin{aligned} \text{Test statistic: } Z &= \frac{\bar{X} - \mu}{s / \sqrt{n}} \sim N(0,1) \\ &= \frac{160 - 158}{8 / \sqrt{144}} \\ Z &= 2.9999 \end{aligned}$$



Depending on the alternative hypothesis (H_1) and at 1% level of significance, the critical values are $[-2.58, +2.58]$.

Since, Z value lies in the rejection region, therefore we reject H_0 .

Conclusion: The average height of the university students is not equal to 158 cm.

i.e., $\mu \neq 158$ cm.

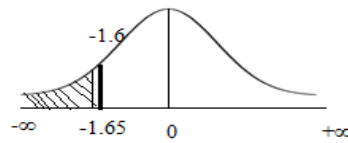
Example 2: A random sample of 400 tins of ghee has mean weight 5.96 kg and the standard deviation of 0.5kg. Test at 5% level of significance that the average weight of tins of ghee is less than 6 kg?

Solution: Given: $n = 400$, $\bar{X} = 5.96$, $s = 0.5$, $\mu = 6$ and $\alpha = 5\%$.

H_0 : The average weight of tins of ghee is 6 kg i.e., $\mu = 6$ kg

H_1 : The average weight of tins of ghee is less than 6 kg i.e., $\mu < 6$ kg (Left tailed test)

Test statistic: $Z = \frac{\bar{X} - \mu}{s/\sqrt{n}} \sim N(0,1)$
 $= \frac{5.96 - 6}{0.5/\sqrt{400}}$
 $Z = -1.6$



Depending on the alternative hypothesis (H_1) and at 5% level of significance, the critical value is -1.65.

Since, Z value lies in the acceptance region; therefore, we do not reject H_0 .

Conclusion: The average weight of tins of ghee is 6 kg i.e., $\mu = 6$ kg.

14.5 Large sample test procedure to test for equality of means of two populations

Here we test the equality of means of two populations, from two large samples, which are drawn either from that population or from two different populations. And the test procedure is as follows.

Step 1: H_0 : The population means are equal $\mu_1 = \mu_2$.

Step 2: H_1 : The population means are not equal $\mu_1 \neq \mu_2$. (Two tailed test), OR

The mean of first population is less than the mean of the second population $\mu_1 < \mu_2$. (Left tailed test)

OR

The mean of first population is more than the mean of the second population $\mu_1 > \mu_2$. (Right tailed test)

Step 3: computation of test statistic; under H_0

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

Where, \bar{X}_1 and \bar{X}_2 are sample means; σ_1 and σ_2 are population standard deviations, when they are unknown, then they are replaced by sample standard deviation 's₁' and 's₂'; n_1 and n_2 are the sample sizes.

Step 4: Depending on the alternative hypothesis (H_1) and level of significance (α), the critical (Table) value i.e., $\pm Z_{\alpha/2}$ (for two tailed) or Z_α (for one tailed) is chosen.

Step 5: If the calculated value of the test statistic (Z) lies in the acceptance region, then we Do not reject H_0 . Otherwise we reject H_0 .

i.e., for two tailed test, if $-Z_{\alpha/2} \leq Z \leq +Z_{\alpha/2}$, then we do not reject H_0 .

i.e., for left tailed test, if $Z \geq -Z_\alpha$ then we do not reject H_0 .

i.e., for Right tailed test, if $Z \leq Z_\alpha$ then we do not reject H_0 .

Otherwise H_0 is rejected.

Example 3: The mean I.Q of 150 randomly selected boys of a college is 95 and that of 125 randomly selected girls of that college is 93. Standard deviations of their I.Q are 12 and 10 respectively. Test whether there is a significant difference between the average I.Q of boys and girls at 1% level of significance.

Solution: Given: $n_1 = 150, \bar{X}_1 = 95, s_1 = 12, n_2 = 125, \bar{X}_2 = 93, s_2 = 10$ and $\alpha = 1\%$.

H_0 : There is no significant difference between the average I.Q of boys and girls
i.e., $\mu_1 = \mu_2$.

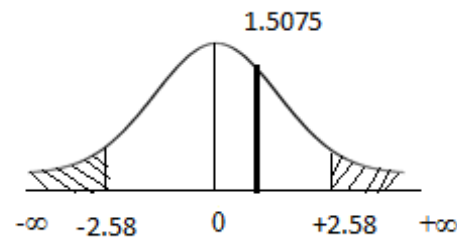
H_1 : There is a significant difference between the average I.Q of boys and girls
i.e., $\mu_1 \neq \mu_2$. (Two tailed test)

Test statistic; under H_0

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim N(0,1)$$

$$Z = \frac{95 - 93}{\sqrt{\frac{12^2}{150} + \frac{10^2}{125}}}$$

$$Z = 1.5075$$



At 1% level of significance, the critical values are $[-2.58, +2.58]$.

Since, Z value lies in the acceptance region; therefore, we do not reject H_0 .

Conclusion: There is no significant difference between the average I.Q. of boys and girls
i.e., $\mu_1 = \mu_2$.

Example 4: The mean and variance of heights of a sample of 100 randomly selected Biharis are 175 cm and 9 cm^2 respectively. The mean and variance of heights of a sample of 80 randomly selected Rajasthanis are 173 cm and 16 cm^2 respectively. Test whether that the mean height of Biharis is taller than Rajasthanis at 5% level of significance?

Given: $n_1 = 100, \bar{X}_1 = 175, s_1^2 = 9, n_2 = 80, \bar{X}_2 = 173, s_2^2 = 16$ and $\alpha = 5\%$.

H_0 : The mean height of Biharis is equal to the mean height of Rajasthanis
i.e., $\mu_1 = \mu_2$.

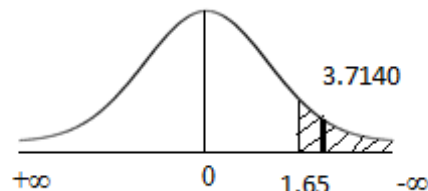
H_1 : The mean height of Biharis is taller than the mean height of Rajasthanis
i.e., $\mu_1 > \mu_2$. (Right tailed test)

Test statistic; under H_0

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim N(0,1)$$

$$Z = \frac{175 - 173}{\sqrt{\frac{9}{100} + \frac{16}{80}}}$$

$$Z = 3.7140.$$



Depending on the alternative hypothesis (H_1) and at 5% level of significance, the critical value is 1.65.

Since, Z value lies in the rejection region, therefore we reject H_0 .

Conclusion: The mean height of Biharis is taller than the mean height of Rajasthanis
i.e., $\mu_1 > \mu_2$ (Right tailed test).

14.6 Large sample test (Z - test) procedure to test for population proportion:

Step 1: H_0 : The population proportion. i.e., $P = P_0$.

Step 2: H_1 : The population proportion. i.e., $P \neq P_0$. (Two tailed test)

OR

The population proportion. i.e., $P < P_0$. (Left tailed test)

OR

The population proportion. i.e., $P > P_0$. (Right tailed test)

Step 3: computation of test statistic; under H_0

$$Z = \frac{p - P_0}{\sqrt{\frac{P_0 Q_0}{n}}} \sim N(0,1)$$

Where, p is sample proportion, i.e., $p = \frac{x}{n}$

Here, x is the number of items possessing an attribute and n is the number of items in the sample.

P_0 is the population proportion value which is to be tested.

$$Q_0 = 1 - P_0$$

Depending on the alternative hypothesis (H_1) and level of significance (α), the critical (Table) value i.e., $\pm Z_{\alpha/2}$ (for two tailed) or Z_α (for one tailed) is chosen.

Step 5: If the calculated value of the test statistic (Z) lies in the acceptance region, then we

Do not reject H_0 . Otherwise we reject H_0 .

i.e., for two tailed test, if $-Z_{\alpha/2} \leq Z \leq +Z_{\alpha/2}$, then we do not reject H_0 .

i.e., for left tailed test, if $Z \geq -Z_\alpha$ then we do not reject H_0 .

i.e., for Right tailed test, if $Z \leq Z_\alpha$ then we do not reject H_0 .

Otherwise H_0 is rejected.

Example 7: The mobile manufacturer states that less than 3% of the mobiles he provided to a certain mobile store are defects. A sample of 500 mobiles revealed that 10 were defects. Test his statement at 1% level of significance.

Solution: Given: $n = 500$, $p = \frac{x}{n} = \frac{10}{500} = 0.02$, $P_0 = 0.03$, $Q_0 = 1 - P_0 = 0.97$ and $\alpha = 1\%$.

H_0 : 3% of the mobiles are defects i.e., $P = 3\%$.

H_1 : less than 3% mobiles are defects. i.e., $P < 3\%$. (Left tailed test)

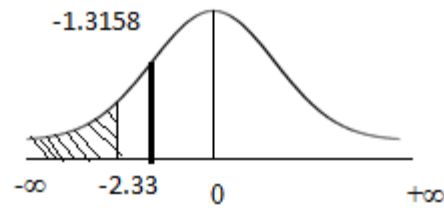
Computation of test statistic;

Under H_0 , test statistic is

$$Z = \frac{p - P_0}{\sqrt{\frac{P_0 Q_0}{n}}} \sim N(0,1)$$

$$= \frac{0.02 - 0.03}{\sqrt{\frac{(0.03)(0.97)}{500}}}$$

$$Z = -1.3158$$



Depending on the alternative hypothesis (H_1) and at 1% level of significance, the critical value is -2.33. Since, Z value lies in the Acceptance region; therefore, we do not reject H_0 .

Conclusion: 3% of the mobiles are defects i.e., $P = 3\%$.

14.7 Large sample test (Z - test) for equality of proportions of two populations:

Step 1: H_0 : The population proportions are equal. i.e., $P_1 = P_2$.

Step 2: H_1 : The population proportions are not equal. i.e., $P_1 \neq P_2$. (Two tailed test)

OR

The first population proportion is less than the second population proportion. i.e., $P_1 < P_2$. (Left tailed test)

OR

The first population proportion is more than the second population proportion. i.e., $P_1 > P_2$. (Right tailed test)

Step 3: computation of test statistic; under H_0

$$Z = \frac{p_1 - p_2}{\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0,1)$$

Where, p_1 and p_2 are the sample proportions from first and second samples respectively,

$$\text{i.e., } p_1 = \frac{x_1}{n_1}, p_2 = \frac{x_2}{n_2} \text{ and } P = \frac{x_1 + x_2}{n_1 + n_2} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}; Q = 1 - P$$

Here, x_1 and x_2 are the number of items possessing an attribute from first and second samples; n_1 and n_2 are the number of items in the sample one and two respectively.

Step 4: Depending on the alternative hypothesis (H_1) and level of significance (α), the critical (Table) value i.e., $\pm Z_{\alpha/2}$ (for two tailed) or Z_α (for one tailed) is chosen.

Step 5: If the calculated value of the test statistic (Z) lies in the acceptance region, then we

Do not reject H_0 . Otherwise we reject H_0 .

i.e., for two tailed test, if $-Z_{\alpha/2} \leq Z \leq +Z_{\alpha/2}$, then we do not reject H_0 .

i.e., for left tailed test, if $Z \geq -Z_\alpha$ then we do not reject H_0 .

i.e., for Right tailed test, if $Z \leq Z_\alpha$ then we do not reject H_0 .

Otherwise H_0 is rejected.

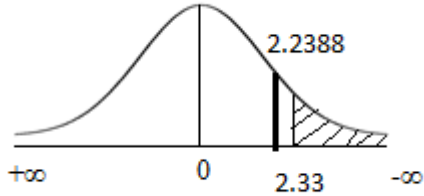
Example 8: In a city, out of 500 students who took M.C.A examination, 460 are passed and out of 400 students who took M.B.A examination, 350 are passed. At 1% level of significance, can we conclude that M.C.A students have performed better than M.B.A students?

Solution: Given: $n_1 = 500$, $p_1 = \frac{x_1}{n_1} = \frac{460}{500} = 0.92$, $p_2 = \frac{x_2}{n_2} = \frac{350}{400} = 0.875$,

$$P = \frac{x_1+x_2}{n_1+n_2} = \frac{460+350}{500+400} = 0.9 \quad Q = 1 - P = 1 - 0.9 = 0.1 \quad \text{and } \alpha = 1\%.$$

H_0 : Performance of M.C.A students and M.B.A students is same. i.e., $P_1 = P_2$.

H_1 : Performance of M.C.A students is better than M.B.A students $P_1 > P_2$. (Right tailed test)

<p>Computation of test statistic; under H_0,</p> $Z = \frac{p_1 - p_2}{\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0,1)$ $= \frac{0.92 - 0.875}{\sqrt{(0.9)(0.1)\left(\frac{1}{500} + \frac{1}{400}\right)}}$ <p>$\Rightarrow Z = 2.2388$</p>	
---	--

Depending on the alternative hypothesis (H_1) and at 1% level of significance, the critical value is 2.33.

Since, Z value lies in the Acceptance region; therefore, we do not reject H_0 .

Conclusion: Performance of M.C.A. students and M.B.A students is same. i.e., $P_1 = P_2$.

Exercise:

1. A sample of 400 students is taken from a University. If the mean and standard deviation of their weights are 55 kg and 3 kg respectively, test at 5% level of significance that the average weight of the university students is 50 kg?
2. A random sample of 65 kids is taken from a kindergarten. The average height of the kids is 103 cm and standard deviation is 5 cm. Can we assume that the average height of the kindergarten kids is less than 105cm?
3. A company manufactures car tyres. Their average life is 40,000 kilometres and standard deviation 5,000 kilometres. A change in the production process is believed to result in a better product. A test of sample of 100 new tyres has mean life of 41,000 kilometres. Can you conclude at 5% level of significance that the new product gives better result?
4. For the following data, test whether means differ significantly.

Sample	Size	Mean	Standard deviation
A	100	55	8
B	50	53	6

5. A random sample of 150 workers from South Karnataka shows that their mean wage is Rs. 215 per day with standard deviation Rs.20. A random sample of 200 workers from North Karnataka shows that their mean wage is Rs. 230 per day with standard deviation Rs.30. Test at 5% level of significance that, mean wages of South Karnataka is less than mean wages of North Karnataka.
6. In a random sample of 1000 persons from a city, 450 are female. Can we conclude that male and female are in the equal ratio in the city?
7. The manufacturer of a surgical instruments claims that less than 3% of the instruments he supplied to a certain hospital are faulty. A sample of 300 instruments revealed that 10 were faulty. Test his claim at 1% level of significance.
8. For the following data, test whether the difference between the proportions of the populations from which the two samples drawn is significant at 1% level of significance.

Sample	Size	Proportion
I	200	0.03
II	300	0.01

9. 500 students are randomly selected from town A, 75% of the students passed. 300 students are randomly selected from town B, 68% of the students passed. Can we conclude that the performance of town A students are better than the performance of town B students at 5 % level of significance?

UNIT 15

SMALL SAMPLE TESTS – I

(Tests for Mean(s), paired t-test, Correlation coefficient)

15.1 Objectives

After completion of this Unit, you should

- ❖ Know the meaning of small sample test, assumptions and the applications.
- ❖ Know how to test the significance of mean of a population, the difference between the means of two populations using two small samples (Independent samples).
- ❖ Know how to test the significance of difference between the means of two populations using two small samples (dependent samples)-paired t-test.

- ❖ Know how to test the significance of correlation coefficient.

15.2 Introduction

Small sample test is a statistical test procedure used to analyse data when the sample size is relatively small i.e., $n < 30$. It helps to determine if there is a significant difference between two or more groups or if there is a significant correlation between two variables.

While we deal with small sample sizes, it is necessary to consider the assumptions of the statistical test being used. For instance, t- test assumes that the data is normally distributed; violating this assumption can lead to inaccurate results.

In other words, small sample tests are useful tools for analysing data when the sample size is small. However, it is necessary to choose the appropriate statistical test and consider the assumptions of the test to ensure accurate results.

15.3 Applications of Small Sample Test

Applications of small sample test are:

1. It is used to test the significance of the mean of a population using small sample.
2. It is used to test the difference between the means of two populations using two small samples (Independent samples).
3. It is used to test the difference between the means of two populations using paired observations (dependent samples).
4. It is used as fundamental component of ANOVA (Analysis Of Variance), which is used to compare means across multiple groups.
5. It is used in Regression analysis to test the significance of regression coefficients.
6. It is used to construct confidence intervals for small samples.

Remark: t- test is performed when the variance of the population is unknown.

Degrees of freedom: it is the number of independent observations and is denoted by d.f. If there are 'n' observations then, degrees of freedom = $(n-c)$, here 'c' is the number of independent constraints.

15.4 Small sample test procedure to test the significance of mean of a population:

Step 1: H_0 : The population mean. i.e., $\mu = \mu_0$.

Step 2: H_1 : The population mean. i.e., $\mu \neq \mu_0$. (Two tailed test)

OR

The population mean. i.e., $\mu < \mu_0$. (Left tailed test)

OR

The population mean. i.e., $\mu > \mu_0$. (Right tailed test)

Step 3: computation of test statistic; under H_0

$$t = \frac{\bar{X} - \mu}{s / \sqrt{n-1}} \sim t \text{ with } (n-1) \text{ d.f.}$$

Where, \bar{X} is the sample mean.

μ is the hypothetical mean of the population.

n is the sample size.

s is the standard deviation of the sample = $\sqrt{\frac{\sum(X_i - \bar{X})^2}{n}} = \sqrt{\frac{\sum X_i^2}{n} - \left(\frac{\sum X_i}{n}\right)^2}$

Step 4: Depending on the alternative hypothesis (H_1), degrees of freedom and level of significance (α), the critical

(Table) value i.e., $\pm t_{\alpha/2}$ (for two tailed) or t_{α} (for one tailed) is chosen.

Step 5: If the calculated value of the test statistic(t) lies in the acceptance region, then we

Do not reject H_0 . Otherwise we reject H_0 .

i.e., for two tailed test, if $-t_{\alpha/2} \leq t \leq +t_{\alpha/2}$, then we do not reject H_0 .

i.e., for left tailed test, if $t \geq -t_{\alpha}$ then we do not reject H_0 .

i.e., for Right tailed test, if $t \leq t_{\alpha}$ then we do not reject H_0 .

Otherwise H_0 is rejected.

Example 1: The length of 10 samples of woollen taken from a population has mean length of 58 cm and standard deviation 5 cm. Test whether the mean length of the population can be taken as 60cm at 5% level of significance?

Solution: Given: $n = 10$, $\bar{X} = 58$, $s = 5$, $\mu = 60$ and $\alpha = 5\%$.

H_0 : The mean length of population is 60 cm i.e., $\mu = 60$ cm.

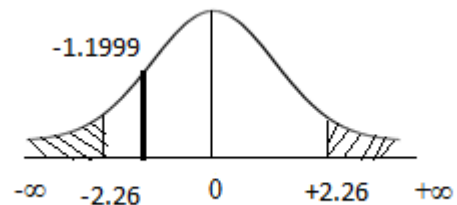
H_1 : The mean length of population is not equal to 60 cm. i.e., $\mu \neq 60$. (Two tailed test)

Under H_0 ,

$$t = \frac{\bar{X} - \mu}{s / \sqrt{n-1}} \sim t \text{ with } (n-1) \text{ d.f.}$$

$$= \frac{58 - 60}{5 / \sqrt{10-1}}$$

$$t = -1.1999$$



Depending on the alternative hypothesis (H_1) degrees of freedom ($n-1 = 10-1 = 9$) and level of significance ($\alpha = 5\%$), the critical values are $[-2.26, +2.26]$.

Since, t value lies in the acceptance region; therefore, we do not reject H_0 .

Conclusion: The mean length of population is 60 cm i.e., $\mu = 60$ cm.

Example 2: The mean weekly sales of the Gudbud in an ice-cream parlour were 156.3. After advertising campaign, the mean weekly sales in 22 parlours for a typical week increased to 163.7 and showed standard deviation of 15.2. Was the advertisement campaign successful?

Solution: Given: $n = 22$, $\bar{X} = 163.7$, $s = 15.2$, $\mu = 156.3$ and $\alpha = 5\%$.

H_0 : Advertisement campaign is not successful i.e., $\mu = 156.3$.

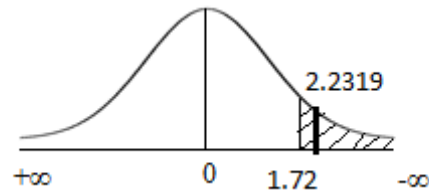
H_1 : Advertisement campaign is successful i.e., $\mu > 156.3$. (Right tailed test)

Under H_0 ,

$$t = \frac{\bar{X} - \mu}{s / \sqrt{n-1}} \sim t \text{ with } (n-1) \text{ d.f.}$$

$$= \frac{163.7 - 156.3}{15.2 / \sqrt{22-1}}$$

$$t = 2.2319$$



Depending on the alternative hypothesis (H_1) degrees of freedom ($n-1 = 22-1 = 21$) and level of significance ($\alpha = 5\%$), the critical value is 1.72.

Since, t value lies in the rejection region; therefore, we reject H_0 .

Conclusion: Advertisement campaign is successful i.e., $\mu > 156.3$.

15.5 Small sample test procedure to test the difference between the means of two populations using two small samples (Independent samples)

Step 1: H_0 : The population means are equal i.e., $\mu_1 = \mu_2$.

Step 2: H_1 : The population means are not equal i.e., $\mu_1 \neq \mu_2$. (Two tailed test)

OR

The mean of first population is less than the mean of the second population. i.e., $\mu_1 < \mu_2$. (Left tailed test)

OR

The mean of first population is more than the mean of the second population. i.e., $\mu_1 > \mu_2$. (Right tailed test)

Step 3: computation of test statistic; under H_0

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{s_c^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \sim t \text{ with } (n_1 + n_2 - 2) \text{ d.f.}$$

Where, \bar{X}_1 and \bar{X}_2 are the sample means. μ_1 and μ_2 are the hypothetical means of the population, and n_1 and n_2 are the sample sizes.

$$s_c^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2} = \frac{\sum (X_1 - \bar{X}_1)^2 + \sum (X_2 - \bar{X}_2)^2}{n_1 + n_2 - 2}$$

Step 4: Depending on the alternative hypothesis (H_1), degrees of freedom and level of significance (α), the critical (Table) value i.e., $\pm t_{\alpha/2}$ (for two tailed) or t_{α} (for one tailed) is chosen.

Step 5: If the calculated value of the test statistic (t) lies in the acceptance region, then we do not reject H_0 . Otherwise we reject H_0 .

i.e., for two tailed test, if $-t_{\alpha/2} \leq t \leq +t_{\alpha/2}$, then we do not reject H_0 .

i.e., for left tailed test, if $t \geq -t_{\alpha}$ then we do not reject H_0 .

i.e., for Right tailed test, if $t \leq t_{\alpha}$ then we do not reject H_0 .

Otherwise H_0 is rejected.

Example 3: Two different types of drugs P and Q were tried on certain patients for increasing weight. 5 persons were given drug P and 7 persons were given drug Q. The increase in weight in pounds is given below.

Drug P	8	12	13	9	3		
Drug Q	10	8	12	15	6	8	11

Do the two drugs differ significantly with regard to their effect in increasing weight? Test at 5 % level of significance.

Solution:

Let ' x_1 ' be the weight of persons using drug P and ' x_2 ' be the weight of persons using drug Q.

H_0 : Two drugs P and Q do not differ significantly in increasing weight. i.e., $\mu_1 = \mu_2$.

H_1 : Two drugs P and Q differ significantly in increasing weight. i.e., $\mu_1 \neq \mu_2$ (Two tailed test)

Variance calculation for two samples is given in the following table:

x_1	$x_1 - \bar{x}_1$	$(x_1 - \bar{x}_1)^2$	x_2	$x_2 - \bar{x}_2$	$(x_2 - \bar{x}_2)^2$
8	-1	1	10	0	0
12	3	9	8	-2	4
13	4	16	12	2	4
9	0	0	15	5	25
3	-6	36	6	-4	16
			8	-2	4
			11	1	1
$\sum x_1 = 45$		$\sum (x_1 - \bar{x}_1)^2 = 62$	$\sum x_2 = 70$		$\sum (x_2 - \bar{x}_2)^2 = 54$

$$\bar{x}_1 = \frac{\sum x_1}{n_1} = \frac{45}{5} = 9$$

$$\bar{x}_2 = \frac{\sum x_2}{n_2} = \frac{70}{7} = 10$$

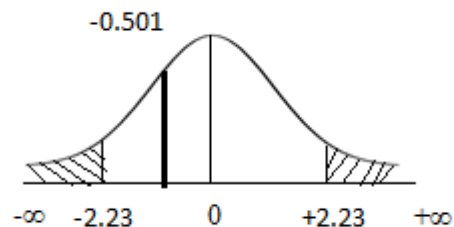
$$s_c^2 = \frac{\sum (x_1 - \bar{x}_1)^2 + \sum (x_2 - \bar{x}_2)^2}{n_1 + n_2 - 2} = \frac{62 + 54}{5 + 7 - 2} = \frac{116}{10} = 11.6$$

Under H_0 , the test statistic is

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{s_c^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \sim t \text{ with } (n_1 + n_2 - 2)$$

d.f.

$$= \frac{9 - 10}{\sqrt{11.6 \left(\frac{1}{5} + \frac{1}{7} \right)}}$$



$$t = -0.501$$

Depending on the alternative hypothesis (H_1) degrees of freedom ($n_1+n_2-2= 10$) and level of significance ($\alpha = 5\%$), the critical value are $[-2.23, +2.23]$.

Since, t value lies in the acceptance region; therefore, we do not reject H_0 .

Conclusion: Two drugs P and Q do not differ significantly in increasing weight. i.e., $\mu_1 = \mu_2$.

Example 4: Mean and standard deviation of heights of residents of two cities gave the following results.

	City X	City Y
Sample	10	12
Mean (cm)	170.5	173.5
Standard deviation (cm)	4	5

Can you conclude at 5% level of significance that the population of city X on an average is shorter than city Y?

Solution: $n_1 = 10, n_2 = 12, \bar{X}_1 = 170.5, s_1 = 4, \bar{X}_2 = 173.5, s_2 = 5$ and $\alpha = 5\%$.

H_0 : The mean height of population of city X and city Y are same i.e., $\mu_1 = \mu_2$.

H_1 : The mean height of population of city X is less than city Y are same i.e., $\mu_1 < \mu_2$. (Left tailed test)

To compute combined sample variance, we have,

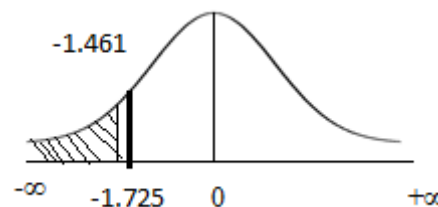
$$S_c^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2} = \frac{10 \times (4)^2 + 12 \times (5)^2}{10 + 12 - 2} = 23$$

Under H_0 , the test statistic is

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{S_c^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \sim t \text{ with } (n_1 + n_2 - 2) \text{ d.f.}$$

$$= \frac{170.5 - 173.5}{\sqrt{23 \left(\frac{1}{10} + \frac{1}{12} \right)}}$$

$$t = -1.461$$



Depending on the alternative hypothesis (H_1) degrees of freedom ($n_1+n_2-2= 20$) and level of significance ($\alpha = 5\%$), the critical value is -1.725 .

Since, t value lies in the acceptance region; therefore, we do not reject H_0 .

Conclusion: The mean height of population of city X and city Y are same i.e., $\mu_1 = \mu_2$.

15.6 Small sample test (t- test) procedure to test the difference between the means of two populations using paired observations (dependent samples):

Step 1: H_0 : The population means are equal i.e., $\mu_1 = \mu_2$.

Step 2: H_1 : The population means are not equal i.e., $\mu_1 \neq \mu_2$. (Two tailed test)

OR

The mean of first population is less than the mean of the second population. i.e., $\mu_1 < \mu_2$. (Left tailed test)

OR

The mean of first population is more than the mean of the second population. i.e., $\mu_1 > \mu_2$. (Right tailed test)

Step 3: computation of test statistic; under H_0

$$t = \frac{\bar{d}}{s_d / \sqrt{n-1}} \sim t \text{ with } (n-1) \text{ d.f.}$$

Where, \bar{d} is the mean of the difference between paired observation.

s_d is the standard deviation of difference of samples.

$$\text{i.e., } s_d = \sqrt{\frac{\sum d^2}{n} - \left(\frac{\sum d}{n}\right)^2} \quad \text{and } d = x_1 - x_2$$

Step 4: Depending on the alternative hypothesis (H_1), degrees of freedom and level of significance (α), the critical (Table) value i.e., $\pm t_{\alpha/2}$ (for two tailed) or t_{α} (for one tailed) is chosen.

Step 5: If the calculated value of the test statistic (t) lies in the acceptance region, then we do not reject H_0 . Otherwise we reject H_0 .

i.e., for two tailed test, if $-t_{\alpha/2} \leq t \leq +t_{\alpha/2}$, then we do not reject H_0 .

i.e., for left tailed test, if $t \geq -t_{\alpha}$ then we do not reject H_0 .

i.e., for Right tailed test, if $t \leq t_{\alpha}$ then we do not reject H_0 .

Otherwise H_0 is rejected.

Example 1: The following data represents the blood pressure of 5 persons before and after performing yoga.

Persons	P	Q	R	S	T
Blood pressure before yoga	90	90	100	88	99
Blood pressure after yoga	88	90	95	90	96

Can we conclude at 5% level of significance that yoga reduces blood pressure?

Solution:

Let ' x_1 ' be the blood pressure before yoga and ' x_2 ' be the blood pressure after yoga.

H_0 : Yoga doesn't reduce blood pressure. i.e., $\mu_1 = \mu_2$.

H_1 : Yoga reduces blood pressure. i.e., $\mu_1 > \mu_2$. (Right tailed test)

Computation of mean and standard deviation of difference of samples is given in the table below:

x_1	x_2	$d = x_1 - x_2$	d^2
90	88	2	4
90	90	0	0
100	95	5	25
88	90	-2	4
99	96	3	9

		$\sum d = 8$	$\sum d^2 = 42$
--	--	--------------	-----------------

$$\bar{d} = \frac{\sum d}{n} = \frac{8}{5} = 1.6$$

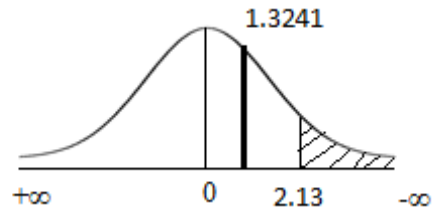
$$s_d = \sqrt{\frac{\sum d^2}{n} - \left(\frac{\sum d}{n}\right)^2} = \sqrt{\frac{42}{5} - \left(\frac{8}{5}\right)^2} = 2.4166$$

Under H_0 , the test statistic is

$$t = \frac{\bar{d}}{s_d / \sqrt{n-1}} \sim t \text{ with } (n-1) \text{ d.f.}$$

$$t = \frac{1.6}{2.4166 / \sqrt{5-1}}$$

$$t = 1.3241$$



Depending on the alternative hypothesis (H_1) degrees of freedom ($n - 1 = 4$) and level of significance ($\alpha = 5\%$), the critical value is 2.13.

Since, t value lies in the acceptance region; therefore, we do not reject H_0 .

Conclusion: Yoga doesn't reduce blood pressure. i.e., $\mu_1 = \mu_2$.

Example 2: The following are the data regarding I.Q of 5 students before and after training:

Student	A	B	C	D	E
Before training	121	126	119	137	122
after training	131	124	121	133	126

Solution: Let ' x_1 ' be the I.Q. of students before training and ' x_2 ' be the I.Q. of students after training.

H_0 : Training doesn't improve the I.Q. of the students. i.e., $\mu_1 = \mu_2$

H_1 : Training improves the I.Q. of the students i.e., $\mu_1 < \mu_2$. (Left tailed test)

Computation of mean and standard deviation of difference of samples is given below:

x_1	x_2	$d = x_1 - x_2$	d^2
121	131	-10	100
126	124	2	4
119	121	-2	4
137	133	4	16
122	126	-4	16
		$\sum d = -10$	$\sum d^2 = 140$

$$\bar{d} = \frac{\sum d}{n} = \frac{-10}{5} = -2$$

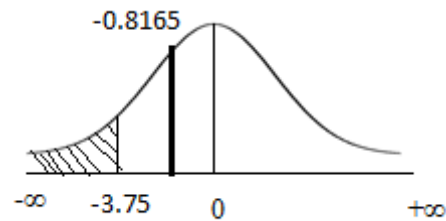
$$s_d = \sqrt{\frac{\sum d^2}{n} - \left(\frac{\sum d}{n}\right)^2} = \sqrt{\frac{140}{5} - \left(\frac{-10}{5}\right)^2} = 4.899$$

Under H_0 , the test statistic is

$$t = \frac{\bar{d}}{s_d / \sqrt{n-1}} \sim t \text{ with } (n-1) \text{ d.f.}$$

$$t = \frac{-2}{4.899 / \sqrt{5-1}}$$

$$t = -0.8165$$



Depending on the alternative hypothesis (H_1) degrees of freedom ($n - 1 = 4$) and level of significance ($\alpha = 5\%$), the critical value is -3.75 .

Since, t value lies in the acceptance region; therefore, we do not reject H_0 .

Conclusion: Training doesn't improve the I.Q. of the students. i.e., $\mu_1 = \mu_2$.

15.7 Small sample (test t- test) procedure to test the significance of an observed sample correlation coefficient

Step 1: H_0 : The population correlation coefficient is zero i.e., $\rho = 0$

Step 2: H_1 : The population correlation coefficient is not equal to zero i.e., $\rho \neq 0$.

Step 3: computation of test statistic; under H_0

$$t = \frac{r \sqrt{(n-2)}}{(\sqrt{1-r^2})} \sim t \text{ with } (n-2) \text{ d.f.}$$

Where, r is sample correlation coefficient which is computed by using Spearman's rank correlation coefficient method.

Step 4: Depending on the alternative hypothesis (H_1), degrees of freedom and level of significance (α), the critical (Table) value i.e., $\pm t_{\alpha/2}$ (for two tailed) is chosen.

Step 5: If the calculated value of the test statistic (t) lies in the acceptance region, then we do not reject H_0 . Otherwise we reject H_0 .

i.e., for two tailed test, if $-t_{\alpha/2} \leq t \leq +t_{\alpha/2}$, then we do not reject H_0 .

Remark: The formula for Spearman's rank correlation coefficient is given by:

1. If the observations are not repeated, then Spearman's rank correlation coefficient is

$$r = 1 - \frac{6 \sum d^2}{n^3 - n}$$

2. If the observations are repeated, then Spearman's rank correlation coefficient is

$$r = 1 - \frac{6 \left[\sum d^2 + \left(\frac{m_i^3 - m_i}{12} \right) \right]}{n^3 - n}$$

Where, d is the difference between the ranks ($R_x - R_y$) of paired observations, n is the number of paired observations and m_i is the number of times an observation is repeated; $i=1,2,3,\dots$

Example 3: A random sample of 25 pairs of observations from a normal population with correlation coefficient 0.7. Is this significant of correlation in the population?

Solution: Given: $n = 25$, $r = 0.7$ and $\alpha = 5\%$

H_0 : The population correlation coefficient is zero i.e., $\rho = 0$

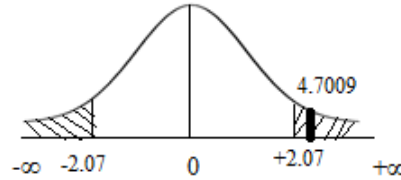
H_1 : The population correlation coefficient is not equal to zero i.e., $\rho \neq 0$.

Under H_0 , the test statistic is

$$t = \frac{r \sqrt{(n-2)}}{(\sqrt{1-r^2})} \sim t \text{ with } (n-2) \text{ d.f.}$$

$$t = \frac{0.7 \sqrt{(25-2)}}{(\sqrt{1-0.7^2})}$$

$$t = 4.7009$$



Depending on the alternative hypothesis (H_1) degrees of freedom ($n - 2 = 23$) and level of significance ($\alpha = 5\%$), the critical values are $[-2.07, +2.07]$.

Since, t value lies in the rejection region; therefore, we reject H_0 .

Conclusion: The population correlation coefficient is not equal to zero i.e., $\rho \neq 0$.

Example 4: The following are the marks obtained by 10 students in science and social tests.

Marks in science	35	37	38	42	44	46	51	54	55	56
Marks in social	40	32	39	42	41	31	50	52	46	55

calculate the rank correlation coefficient for the data given above and test whether the correlation coefficient differs significantly or not?

Solution: H_0 : The correlation coefficient between the marks obtained by 10 students in science and social tests does not differ significantly. i.e., $\rho = 0$

H_1 : The correlation coefficient between the marks obtained by 10 students in science and social tests differs significantly i.e., $\rho \neq 0$.

Here, we need to compute the value of r using Spearman's rank correlation coefficient and is as follows:

Marks in science (x)	35	37	38	42	44	46	51	54	55	56	
Marks in social(y)	40	32	39	42	41	31	50	52	46	55	
R_x	10	9	8	7	6	5	4	3	2	1	
R_y	7	9	8	5	6	10	3	2	4	1	
$d = R_x - R_y$	3	0	0	2	0	-5	1	1	-2	0	
d^2	9	0	0	4	0	25	1	1	4	0	$\sum d^2 = 44$

$$r = 1 - \frac{6\sum d^2}{n^3 - n}$$

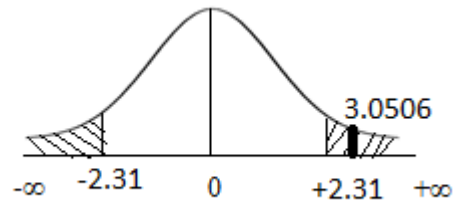
$$= 1 - \frac{6 \times 44}{10^3 - 10}$$

$$r = 0.7333$$

Under H_0 , the test statistic is

$$t = \frac{r\sqrt{(n-2)}}{(\sqrt{1-r^2})} \sim t \text{ with } (n-2) \text{ d.f.}$$

$$t = \frac{0.7333\sqrt{(10-2)}}{(\sqrt{1-0.7333^2})} = 3.0506$$



Depending on the alternative hypothesis (H_1) degrees of freedom ($n - 2 = 8$) and level of significance ($\alpha = 5\%$), the critical values are $[-2.31, +2.31]$.

Since, t value lies in the rejection region; therefore, we reject H_0 .

Conclusion: The correlation coefficient between the marks obtained by 10 students in science and social tests differs significantly i.e., $\rho \neq 0$.

Exercise:

1. A random sample of size 15 taken from a population has a sample mean of 30 and standard deviation 4. Test the hypothesis that the population mean is 32.
2. The mean weekly sales of donuts were 150. After an advertising campaign the mean weekly sale in 23 shops for a typical week increased to 175 with standard deviation 10. Is this evidence indicating that the advertising campaign successful?
3. A fertilizer mixing machine is set to give 15kg of potassium for every bag of fertilizer. Then 20 such bags are examined. The weight of potassium in each bag (in kg) are: 13,14,13,15,15,14,16,15,17,13,16,12,15,15,14,17,15,16,15,14.
Test at 1% level of significance that there is any reason to believe that the machine is defective?
4. Examine whether the means differ significantly for the following data given below:

	Sample I	Sample II
Sample size	10	8
Mean	67.2	62.3
Standard deviation	4.14	4.22

5. A group of 7 persons were given a diet plan A and they weigh 39,43,55,58,65,69 and 68 kg. Another group of 5 persons from the same locality were given a diet plan B who weighs 40,41,47,58 and 60kg. Can you test whether the diet plan B decreases the weight significantly?

6. Two new types of rations are fed to sheep. A sample of ten sheep is fed with Type X ration and another sample of ten sheep is fed with type Y ration, the gains in weight are listed below (in pounds):

Type X	88	89	93	87	85	93	99	85	91	90	86
Type Y	85	87	91	89	95	90	89	83	87	84	92

At 1% level of significance, test whether Type X ration is better than Type Y ration?

7. Two laboratories P and Q carry out independent estimates of pistachio (in grams) content in cassata ice – cream made by a factory. A sample is taken from each batch, halved, and the separate halves sent to the two laboratories. The pistachio content in cassata ice – cream is obtained by the laboratories is recorded below:

Batch no	1	2	3	4	5	6	7	8	9	10
Lab P	7	8	7	6	5	6	7	9	7	5
Lab Q	9	8	9	6	7	9	8	7	6	6

Is there a significant difference between the means of pistachio content obtained by the two laboratories P and Q?

8. Following is the data regarding the I.Q. of 7 students before and after meditation:

Student	P	Q	R	S	T	U	V
I.Q. before meditation	119	121	117	114	123	118	125
I.Q. after meditation	123	118	121	128	125	117	129

Test at 1% level of significance that the meditation improves the I.Q of students?

9. Ten students were given intensive coaching and tests were conducted before and after coaching. The scores of the tests are given below. Test at 1% level of significance that the scores after coaching show an improvement?

Student	A	B	C	D	E	F	G	H	I	J
Marks before coaching	50	45	50	25	37	46	62	49	70	81
Marks after coaching	65	51	37	39	51	27	36	45	68	79

10. A random sample of 29 pairs of observations from a bivariate normal population with correlation coefficient 0.5. Is this significant of correlation in the population?

11. A coefficient of correlation of 0.2 is derived from a random sample of 23pairs of observations. Is this value of r significant?

12. The ranks of same 15 students in tests in mathematics(x) and statistics(y) were as follows:

Ranks in x	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Ranks in y	1	10	3	4	4	7	2	6	8	11	15	9	14	12	13

calculate the rank correlation coefficient for the data given above and test whether the correlation coefficient differs significantly or not?

13. calculate the rank correlation coefficient for the data given below and test whether the correlation coefficient differs significantly or not at 1% level of significance?

X	18	28	35	44	35	26	37	48
Y	83	51	34	34	34	28	46	47

UNIT 16

SAMLL SAMPLE TESTS -2

(Tests for variance(s), goodness of fit, attributes)

16.1 Objective

After completion of this unit, you should know how to perform

- ❖ the significance test for single variance and the equality/ratio of two variances.
- ❖ significance test for goodness of fit
- ❖ significance test for independence of attributes.

16.2 Introduction

While we deal with small sample sizes($n \leq 30$), it is necessary to consider some assumptions of the statistical test being used. For instance, F- test assumes that the data is normally distributed; violating this assumption can lead to erroneous results.

In other words, small sample tests are useful tools for analysing data when the sample size is small. However, it is necessary to choose the appropriate statistical test and consider the assumptions of the test to ensure accurate results.

16.3 Chi-square test

The chi-square test is a statistical hypothesis test which is used to determine whether there is any significant association between two or more qualitative variables (attributes). It helps in determining whether the observed frequencies of a sample differ significantly from the expected frequencies.

The test compares the observed frequencies in each category to the frequencies that would be expected if there were no association between the variables.

It's important to note that the chi-square test has certain assumptions, such as the independence of observations and expected frequencies being reasonably large. Violations of these assumptions can affect the validity of the test results. Additionally, there are variations of the chi-square test, such as the Fisher's exact test for small sample sizes, which can be used in specific situations.

In other words, the chi-square test is an important tool for analysing categorical data and detecting associations between variables in a wide range of disciplines.

Applications of Chi-Square Test:

Applications of Chi-Square Test are

1. It is used to test whether the population has a given variance.
2. It is used to test Goodness of fit of a theoretical distribution to an observed distribution.
3. It is used to test independence of attributes in a $m \times n$ contingency table.
4. It is used for homogeneity test i.e., the chi-square test is used to compare the distributions of two or more populations.
5. In genetics research, the chi-square test is used to analyse observed and expected genotype frequencies.
6. It is used to combine various probabilities obtained from independent experiments to give a single test of significance.

16.4. Chi-Square Test procedure to test for single variance:

Step 1: H_0 : The population Variance. i.e., $\sigma^2 = \sigma_0^2$

Step 2: H_1 : The population variance. i.e., $\sigma^2 \neq \sigma_0^2$ (Two tailed test)

OR

The population variance. i.e., $\sigma^2 < \sigma_0^2$. (Left tailed test)

OR

The population variance. i.e., $\sigma^2 > \sigma_0^2$ (Right tailed test)

Step 3: computation of test statistic; under H_0

$$\chi^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\sigma_0^2} = \frac{ns^2}{\sigma_0^2} \sim \chi^2 \text{ with } (n-1) \text{ d.f.}$$

Where, 's' is the standard deviation of the sample.

Step 4: Depending on the alternative hypothesis (H_1), degrees of freedom and level of significance (α), the critical (Table) value i.e., $\chi^2_{(1-\frac{\alpha}{2})}$ and $\chi^2_{\alpha/2}$ for two tailed test or for one tailed test χ^2_{α} (right tailed) and $\chi^2_{(1-\alpha)}$ (left tailed) is chosen.

Step 5: If the calculated value of the test statistic (χ^2) lies in the acceptance region, then we do not reject H_0 . Otherwise we reject H_0 .

i.e., for two tailed test, if $\chi^2_{(1-\frac{\alpha}{2})} \leq \chi^2 \leq \chi^2_{\alpha/2}$, then we do not reject H_0 .

i.e., for left tailed test, if $\chi^2 \geq \chi^2_{(1-\alpha)}$ then we do not reject H_0 .

i.e., for Right tailed test, if $\chi^2 \leq \chi^2_{\alpha}$ then we do not reject H_0 , otherwise H_0 is rejected.

Example 1: A normal variate has a variance 5. Twenty sample observations of the variate have variance 3. Test at 1% level of significance whether the population variance is 5?

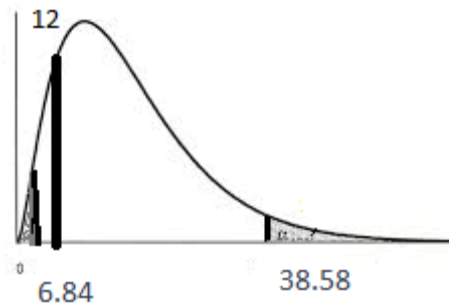
Solution: Given: $n = 20$, $\sigma^2 = 5$, $s^2 = 3$ and $\alpha = 1\%$.

H_0 : The population Variance is 5. i.e., $\sigma^2 = 5$

H_1 : The population variance is not equal to 5. i.e., $\sigma^2 \neq 5$. (Two tailed test)

Under H_0 , the test statistic is

$$\begin{aligned} \chi^2 &= \frac{ns^2}{\sigma_0^2} \sim \chi^2 \text{ with } (n-1) \text{ d.f.} \\ &= \frac{20 \times 3}{5} \\ \chi^2 &= 12 \end{aligned}$$



Depending on the alternative hypothesis (H_1), degrees of freedom ($n - 1 = 19$) and level of significance ($\alpha = 1\%$), the critical values are [6.84, 38.58]

Since, χ^2 value lies in the acceptance region; therefore, we do not reject H_0 .

Conclusion: The population Variance is 5. i.e., $\sigma^2 = 5$

Example 2: The standard deviation of production of sugarcane is assumed to be 12.8 tons. A sample of 20 acres showed that the standard deviation 10.6 tons. Test at 1% level of significance whether the standard deviation of production of sugarcane is less than 12.8 tons.

Solution: Given: $n = 20$, $\sigma = 12.8$, $s = 10.6$ and $\alpha = 1\%$.

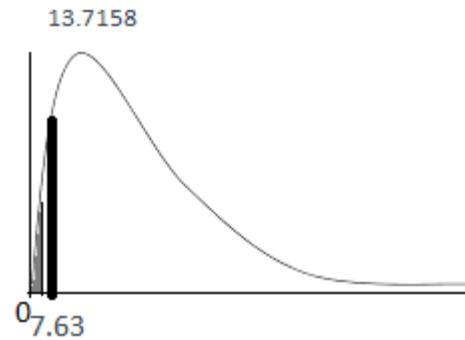
H_0 : Standard deviation of production of sugarcane is 12.8 tons. i.e., $H_0: \sigma = 12.8$ tons.

H_1 : Standard deviation of production of sugarcane is less than 12.8 tons.

i.e. $H_1: \sigma < 12.8$ tons. (left tailed test)

Under H_0 , the test statistic is

$$\begin{aligned}\chi^2 &= \frac{ns^2}{\sigma_0^2} \sim \chi^2 \text{ with } (n-1) \text{ d.f.} \\ &= \frac{20 \times (10.6)^2}{(12.8)^2} \\ \chi^2 &= 13.7158\end{aligned}$$



Depending on the alternative hypothesis (H_1), degrees of freedom ($n - 1 = 19$) and level of significance ($\alpha = 1\%$), the critical value is 7.63

Since, χ^2 value lies in the acceptance region; therefore, we do not reject H_0 .

Conclusion: Standard deviation of production of sugarcane is 12.8 tons. i.e., $\sigma = 12.8$ tons.

Example 3: Following are the points scored by five players in a basketball match: 5, 13, 1, 7, 9

Test whether the population variance is more than 10 at 5% level of significance?

Solution: Given: $n = 5$, $\sigma^2 = 10$, and $\alpha = 5\%$.

Let 'x' denotes the points scored by five players in a basketball match.

Computation of sample variance is given in the table below

X	x^2
5	25
13	169
1	1
7	49
9	81
$\sum x = 35$	$\sum x^2 = 325$

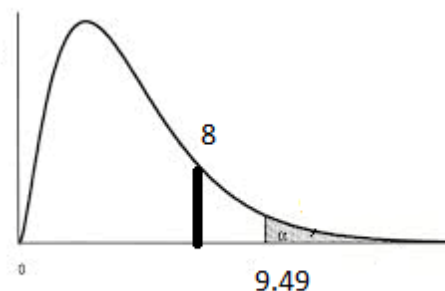
$$s^2 = \frac{\sum x^2}{n} - \left(\frac{\sum x}{n}\right)^2 = \frac{325}{5} - \left(\frac{35}{5}\right)^2 = 16$$

H_0 : The population Variance is 10. i.e., $\sigma^2 = 10$

H_1 : The population variance is not equal to 5. i.e., $\sigma^2 > 10$. (Right tailed test)

Under H_0 , the test statistic is

$$\begin{aligned}\chi^2 &= \frac{ns^2}{\sigma_0^2} \sim \chi^2 \text{ with } (n-1) \text{ d.f.} \\ &= \frac{5 \times 16}{10} \\ \chi^2 &= 8\end{aligned}$$



Depending on the alternative hypothesis (H_1), degrees of freedom ($n - 1 = 4$) and level of significance ($\alpha = 5\%$), the critical value is 9.49

Since, χ^2 value lies in the acceptance region; therefore, we do not reject H_0 .

Conclusion: The population Variance is 10. i.e., $\sigma^2 = 10$

Example 4: Following are the production of paddy in 8 different years

6 9 13 7 14 12 3 and 8 tons.

Test the hypothesis that the standard deviation is more than 3 tons?

Solution: Given: $n = 8$, $\sigma = 3$, and $\alpha = 5\%$.

Let 'x' denotes the production of paddy in 8 different years

H_0 : The population standard deviation is 3. i.e., $\sigma = 2.2$

H_1 : The population standard deviation is more than 3. i.e., $\sigma > 2.2$. (Right tailed test)

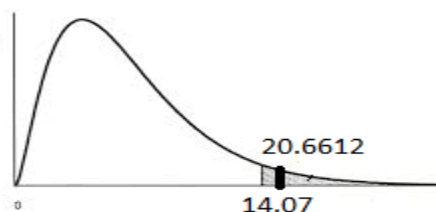
Here,

$$\bar{x} = \frac{\sum x}{n} = \frac{72}{8} = 9$$

X	$x - \bar{x}$	$(x - \bar{x})^2$
6	-3	9
9	0	0
13	4	16
7	-2	4
14	5	25
12	3	9
3	-6	36
8	-1	1
$\sum x = 72$		$\sum (x - \bar{x})^2 = 100$

Under H_0 , the test statistic is

$$\begin{aligned} \chi^2 &= \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\sigma_0^2} \sim \chi^2 \text{ with } (n-1) \text{ d.f.} \\ &= \frac{100}{(2.2)^2} \\ \chi^2 &= 20.6612 \end{aligned}$$



Depending on the alternative hypothesis (H_1), degrees of freedom ($n - 1 = 7$) and level of significance ($\alpha = 5\%$), the critical value is 14.07

Since, χ^2 value lies in the rejection region; therefore, we reject H_0 .

Conclusion: The population standard deviation is more than 3. i.e., $\sigma > 2.2$.

16.5 Chi-Square Test for goodness of fit:

This test helps us to analyse how well the theoretical distributions such as Uniform, Binomial, Poisson, Normal, etc., fit empirical distribution, i.e, those obtained from the sample data. The quantity χ^2 describes the magnitude of the discrepancy between theory and observation.

Conditions for applying Chi-Square Test for goodness of fit:

1. The total frequency should be reasonably large.
2. Theoretical frequency should be greater than or equal to 5. If any theoretical frequency is less than 5, then it should be pooled with the adjacent frequency.
3. If any parameter is estimated from the observed distribution, corresponding to every such estimation, one degree of freedom should be reduced.

15.8.1 Chi-Square Test procedure to test for Goodness of fit:

Step 1: H_0 : The theoretical distribution is a good fit to the observed frequency distribution.

Step 2: H_1 : The theoretical distribution is not a good fit to the observed frequency distribution.

Step 3: computation of test statistic; under H_0

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} = \sum \frac{O_i^2}{E_i} - N \sim \chi^2 \text{ with } (n-c) \text{ d.f.}$$

Where, ‘n’ is the number of terms in the χ^2 after pooling the expected frequencies.

‘c’ is the number of independent constraints, O_i is the observed frequency in the data given and E_i is the expected frequency.

Step 4: Depending on the alternative hypothesis (H_1), degrees of freedom and level of significance (α), the critical (Table) value i.e., χ^2_{α} (right tail) is chosen.

Step 5: If the calculated value of the test statistic (χ^2) lies in the acceptance region, then we do not reject H_0 . Otherwise we reject H_0 .

Remark: This test is always Right tailed test, i.e., if $\chi^2 \leq \chi^2_{\alpha}$, then we do not reject H_0 .

Otherwise H_0 is rejected.

Example 5: A Human Resource manager is interested to determine whether the absenteeism is uniformly distributed throughout the week in production domain of his company. So, he collects the data from past year records which is shown in the table given below:

Day of the week	Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
No. Of absentees:	7	8	11	12	5	13	14

Test whether the absence is uniformly distributed throughout the week at 1% level of significance?

Solution: H_0 : Absence is uniformly distributed throughout the week.

H_1 : Absence is not uniformly distributed throughout the week.

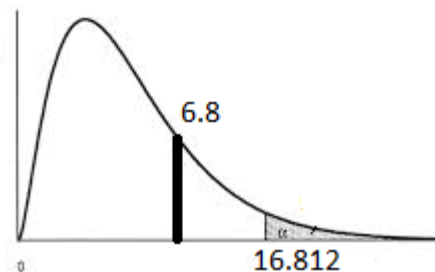
On the basis of H_0 the expected frequencies are $\frac{70}{7} = 10$ absentees for all the days.

To apply χ^2 test we need the following table:

O_i	E_i	$O_i - E_i$	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
7	10	-3	9	0.9
8	10	-2	4	0.4
11	10	1	1	0.1
12	10	2	4	0.4
5	10	5	25	2.5
13	10	3	9	0.9
14	10	4	16	1.6
70	70			$\sum \frac{(O_i - E_i)^2}{E_i} = 6.8$

Under H_0 , the test statistic is:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} = 6.8$$



Depending on the alternative hypothesis (H_1), degrees of freedom ($n-c = 7-1 = 6$) and level of significance ($\alpha = 1\%$), the critical (Table) value is 16.812

Since, χ^2 value lies in the acceptance region; therefore, we do not reject H_0 .

Conclusion: Absence is uniformly distributed throughout the week.

Example 6: Records of 800 families about the number of male births in a family of four children are listed below:

Male births	0	1	2	3	4
No. Of families	32	178	290	236	64

Test the hypothesis that the male and female births are equally likely at 5 % level of significance.

Solution: H_0 : Male and female births are equally likely.

H_1 : Male and female births are not equally likely.

On the basis of hypothesis, we consider $p = 0.5$ and $q = 0.5$; which follows binomial distribution with parameters $n = 4$ and $p = 0.5$.

To find the expected frequencies we need to fit the binomial distribution which is as follows:

The probability mass function is:

$$P(x) = \binom{n}{x} p^x q^{n-x}, x = 0, 1, 2, \dots, n$$

$$P(x) = \binom{4}{x} \left(\frac{1}{2}\right)^x \left(\frac{1}{2}\right)^{4-x}, x = 0,1,2,3,4$$

On simplification,

$$P(x) = \binom{4}{x} \left(\frac{1}{2}\right)^x \left(\frac{1}{2}\right)^{4-x} = \frac{\binom{4}{x}}{2^4} = \frac{\binom{4}{x}}{16}$$

$$P(0) = \frac{\binom{4}{0}}{16} = \frac{1}{16}$$

$$E(0) = N \cdot P(0) = 800 \times \frac{1}{16} = 50$$

By using recurrence relation for expected frequencies

$$E(x) = \frac{n-x+1}{x} \times \frac{p}{q} E(x-1)$$

$$E(1) = \frac{4-1+1}{1} \times 50 = 200; E(2) = \frac{4-2+1}{2} \times 200 = 300$$

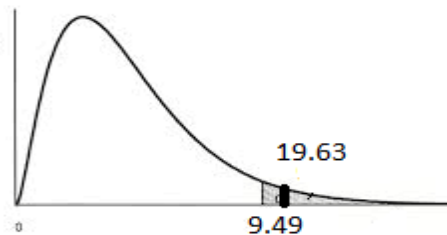
$$E(3) = \frac{4-3+1}{3} \times 300 = 200; E(4) = \frac{4-4+1}{4} \times 200 = 50$$

To apply χ^2 test we need the following table:

O_i	E_i	$O_i - E_i$	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
32	50	-18	324	6.48
178	200	-22	484	2.42
290	300	-10	100	0.33
236	200	36	1296	6.48
64	50	14	196	3.92
800	800			$\sum \frac{(O_i - E_i)^2}{E_i} = 19.63$

Under H_0 , the test statistic is:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} = 19.63$$



Depending on the alternative hypothesis (H_1), degrees of freedom ($n-c = 5-1 = 4$) and level of significance ($\alpha = 5\%$), the critical (Table) value is 9.49

Since, χ^2 value lies in the rejection region; therefore, we reject H_0 .

Conclusion: Male and female births are not equally likely.

Example 7: A book has 700 pages. The number pages with misprints are recorded as follows:

No. Of misprints	0	1	2	3	4	5
No. of	616	70	10	2	1	1

pages						
-------	--	--	--	--	--	--

Fit a Poisson distribution to the data given and test the goodness of fit.

Solution: Null Hypothesis H_0 : Poisson distribution is a good fit.

Alternative hypothesis H_1 : Poisson distribution is not a good fit.

To find the expected number of misprints in each page of the book is as follows:

Mistakes (x)	No. Of pages (f)	Fx
0	616	0
1	70	70
2	10	20
3	2	6
4	1	4
5	1	5
	N = 700	$\sum fx = 105$

$$\text{Expected mistakes } \lambda = \bar{x} = \frac{\sum fx}{N} = \frac{105}{700} = 0.15$$

Calculations of expected frequencies for misprints from 0 to 5 are as follows:

Here $e^{-\lambda} = 0.8607$; implies,

$$p(0) = \frac{e^{-0.15} 0.15^0}{0!} = 0.8607$$

$$E(0) = N.p(0) = 700 \times 0.8607 = 602.5$$

By using recurrence relation for expected frequencies we have,

$$E(x) = \frac{\lambda}{x} E(x-1); \text{ where } x=1,2,\dots$$

$$E(1) = \frac{0.15}{1} \times (602.5) = 90.38; \quad E(2) = \frac{0.15}{2} \times (90.38) = 6.78$$

$$E(3) = \frac{0.15}{3} \times (6.78) = 0.34; \quad E(4) = \frac{0.15}{4} \times (0.34) = 0.013$$

$$E(5) = \frac{0.15}{5} \times (0.013) = 0$$

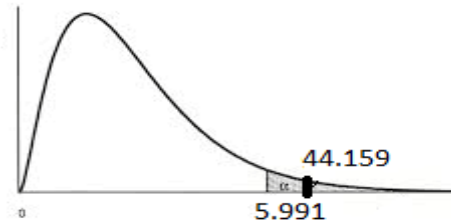
To apply χ^2 test we need the following table:

O_i	E_i	$O_i - E_i$	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
616	602.5	13.50	182.25	0.302
70	90.38	-20.38	415.34	4.595
10	6.78	3.22	10.37	1.529
2	0.34			
1	0.013	0.353	3.65	13.32
4				37.733

1 -	0 -			
				$\sum \frac{(O_i - E_i)^2}{E_i} = 44.159$

Under H_0 , the test statistic is:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} = 44.159$$



Depending on the alternative hypothesis (H_1), degrees of freedom ($n-c = 6-1-3=2$) and level of significance ($\alpha = 5\%$), the critical (Table) value is 5.991

Since, χ^2 value lies in the rejection region; therefore, we reject H_0 .

Conclusion: Poisson distribution is not a good fit.

16.6 Chi-Square Test for independence of attributes:

The chi-square test for independence of attributes is a statistical test used to determine whether there is a relationship between two categorical variables. It is also known as the chi-square test of association or the chi-square test of independence.

The test is based on the principle that if there is no association between two categorical variables, then the distribution of frequencies within each category of one variable should be independent of the distribution of frequencies within each category of the other variable.

Some of the examples are:

- ❖ Success in examination and number of hours studied are independently distributed or not.
- ❖ Is there any association between heights of father and son?
- ❖ Is there any association between marriage and happiness?
- ❖ Is there any association between gender and smoking habits etc.,

Test procedure to test for independence of attributes:

Step 1: H_0 : The two attributes "A" and "B" are independent and identically distributed.

Step 2: H_1 : The two attributes "A" and "B" are not independent and identically distributed.

Step 3: computation of test statistic;

For 2 x 2 contingency table i.e.,

Attribute B	Attribute A		Total
	A	B	a+b
C	D	c+d	
Total	a +c	b+d	a+b+c+d = N

Under H_0 , the test statistic is

$$\chi^2 = \frac{N(ad-bc)^2}{(a+b)(a+c)(b+d)(c+d)} \sim \chi^2 \text{ with (1) d.f.}$$

Step 4: Depending on the alternative hypothesis (H_1), degrees of freedom and level of significance (α), the critical (Table) value i.e., χ^2_{α} (right tail) is chosen.

Step 5: If the calculated value of the test statistic (χ^2) lies in the acceptance region, then we do not reject H_0 . Otherwise we reject H_0 .

Remark: This test is always Right tailed test, i.e., if $\chi^2 \leq \chi^2_{\alpha}$, then we do not reject H_0 . Otherwise H_0 is rejected. The test statistic of Chi- square test for independence of attributes for “m x n” contingency table is:

Under H_0 ,

$$\chi^2 = \sum_{i=1}^m \sum_{j=1}^n \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2 \text{ with } (m-1)(n-1) \text{ d.f}$$

$$\text{Where, } E_{ij} = \frac{R_i \times C_j}{N}$$

R_i is the total of cell frequencies of i^{th} row, and C_j is the total of cell frequencies of j^{th} column.

16.6.1 Yate's correction:

In a 2 x 2 contingency table, the number of d.f is 1. If any one of the theoretical cell frequencies is less than 5, then use of pooling method for χ^2 - test results in χ^2 with 0 (zero) d.f, which is meaningless.

In this case we apply a correction due to F.Yates (1934), which is usually known as Yate's correction for continuity". Thus if any one of the expected frequency is below 5, then we need correction in 2 x 2 contingency table due to "Yates".

Here we add 0.5 to the cell frequency which is less than 5 and then adjusting for the remaining cell frequencies accordingly. Then χ^2 - test is applied without pooling method.

χ^2 test statistic now becomes,

$$\chi^2 = \frac{N \left[|ad - bc| - \frac{N}{2} \right]^2}{(a+c)(b+d)(a+b)(c+d)} \sim \chi^2 \text{ with (1) d.f}$$

Example 8: From the following data test whether 'education' and 'employment' are independent at 1% level of significance.

Education	Employment	
	Employed	Unemployed
Educated	30	10
Uneducated	20	40

Solution:

H₀: Education and Employment are independent.

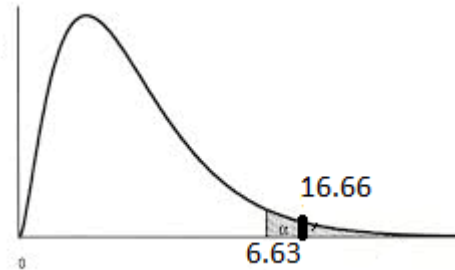
H₁: Education and Employment are not independent.

Under H₀, the test statistic is:

$$\chi^2 = \frac{N(ad-bc)^2}{(a+b)(a+c)(b+d)(c+d)} \sim \chi^2 \text{ with (1) d.f.}$$

$$\chi^2 = \frac{100(30 \times 40 - 10 \times 20)^2}{(40) \times (60) \times (50) \times (50)}$$

$$\chi^2 = 16.66$$



Depending on the alternative hypothesis (H₁), degrees of freedom (1 d.f) and level of significance ($\alpha = 1\%$), the critical (Table) value is 6.63

Since, χ^2 value lies in the rejection region; therefore, we reject H₀.

Conclusion: Education and Employment are not independent.

Example 8: A certain drug is claimed to be effective in curing cold. In an experiment, 164 people with cold, half of them were given the drug and rest of them were treated with sugar pills. The patients' reaction to the treatment are recorded in the following table. Test the hypothesis that the drug is not better than the sugar pills for curing cold.

Type	Effect of drugs		
	Helped	Harmed	No effect
Drug	52	10	20
Sugar pills	44	12	26

Solution: H₀: The drug is not better than the sugar pills for curing cold.

H₁: The drug is better than the sugar pills for curing cold.

Type	Effect of drugs			Total
	Helped	Harmed	No effect	
Drug	52	10	20	82
Sugar pills	44	12	26	82
Total	96	22	46	N = 164

Here the frequencies are arranged in 2 x 3 contingency table. Hence the degrees of freedom is $(m - 1)(n - 1) = (2-1)(3-1) = 2$ d.f

Under H₀, the test statistic is:

$$\chi^2 = \sum_{i=1}^m \sum_{j=1}^n \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2 \text{ with } (m - 1)(n - 1) \text{ d.f}$$

To compute expected frequencies, we have,

$$E_{ij} = \frac{R_i \times C_j}{N}$$

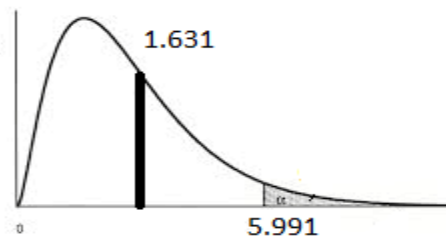
$$E(52) = \frac{82 \times 96}{164} = 48$$

Similarly, $E(44) = 48$; $E(10) = 11$; $E(12) = 11$; $E(20) = 23$; $E(26) = 23$

To apply χ^2 test we need the following table:

Type	Effect of drugs	O_i	E_i	$O_i - E_i$	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
Drug	Helped	52	48	4	16	0.3333
	Harmed	10	11	1	1	0.0909
	No effect	20	23	3	9	0.3913
Sugar Pills	Helped	44	48	4	16	0.3333
	Harmed	12	11	1	1	0.0909
	No effect	26	23	3	9	0.3913
Total		164				$\sum \frac{(O_i - E_i)^2}{E_i} = 1.631$

$$\chi^2 = \sum_{i=1}^m \sum_{j=1}^n \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = 1.631$$



Depending on the alternative hypothesis (H_1), degrees of freedom (2 d.f) and level of significance ($\alpha = 5\%$), the critical (Table) value is 5.991

Since, χ^2 value lies in the acceptance region; therefore, we do not reject H_0 .

Conclusion: The drug is not better than the sugar pills for curing cold.

16.5 Test significance based on F – distribution:

To test the significance of the F distribution, we would typically use an F-test. The F-test is a statistical test that compares the variances of two or more samples to determine if they are significantly different from each other.

Applications of F-distribution:

It has the following applications in statistical theory.

5. To test the equality of two population variances.
6. To test the significance of an observed multiple correlation coefficient.
7. To test the linearity of regression.
8. To test the equality of several means.

Remark:

The reciprocal property of F-distribution is

$$F_{\frac{\alpha}{2},(m-1, n-1)} \times F_{1-\frac{\alpha}{2},(n-1, m-1)} = 1$$

$$\Rightarrow F_{\frac{\alpha}{2},(m-1, n-1)} = \frac{1}{F_{1-\frac{\alpha}{2},(n-1, m-1)}}$$

16.5.1 F – test for equality of two population variances:

Step 1: H_0 : The population variances are equal i. e., $\sigma_1^2 = \sigma_2^2$

Step 2: H_1 : The population variances are not equal i. e., $\sigma_1^2 \neq \sigma_2^2$

Step 3: computation of test statistic; under H_0

$F = \frac{s_1^2}{s_2^2} \sim$ Snedecor's F-distribution with (n_1-1, n_2-1) d.f

Where, $s_1^2 = \frac{1}{n_1-1} \sum_{i=1}^{n_1} (x_i - \bar{x})^2$ and

$$s_2^2 = \frac{1}{n_2-1} \sum_{j=1}^{n_2} (x_j - \bar{x})^2$$

Step 4: Depending on the alternative hypothesis (H_1), degrees of freedom (n_1-1, n_2-1) and level of significance (α), the critical (Table) value i.e., $F_{(1-\frac{\alpha}{2})}$ and $F_{\alpha/2}$ for two tailed test or for one tailed test F_{α} (right tailed) and $F_{(1-\alpha)}$ (left tailed) is chosen.

Step 5: If the calculated value of the test statistic (χ^2) lies in the acceptance region, then we do not reject H_0 . Otherwise we reject H_0 .

i.e., for two tailed test, if $F_{(1-\frac{\alpha}{2})} \leq F \leq F_{\alpha/2}$, then we do not reject H_0 .

i.e., for left tailed test, if $F \geq F_{(1-\alpha)}$ then we do not reject H_0 .

i.e., for Right tailed test, if $F \leq F_{\alpha}$ then we do not reject H_0 .

Otherwise H_0 is rejected.

Example 5: The following are the marks of students of two class X and Y. Test whether the variances of marks of both the classes differ significantly at 2% level of significance.

Class X	80.51	80.46	80.75	80.50	80.36	80.32	82.6	83.4
Class y	85.1	80.28	84.6	83.5	89.5	85.6	80.27	-

Solution: Given: $n_1 = 8, n_2 = 7$ and $\alpha = 2\%$.

Here $\bar{x} = 81.1125$ and $\bar{y} = 84.1214$

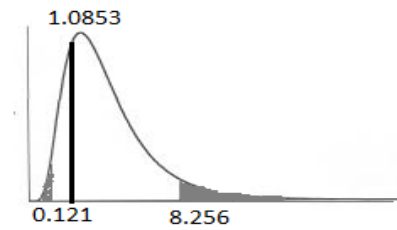
H_0 : The variances of marks of both the classes doesn't differ significantly i. e., $\sigma_1^2 = \sigma_2^2$

H_1 : The variances of marks of both the classes differ significantly i. e., $\sigma_1^2 \neq \sigma_2^2$

Under H_0 , the test statistic is:

$$F = \frac{s_1^2}{s_2^2} \sim \text{Snedecor's F-distribution}$$

with (n_1-1, n_2-1) d.f



Where, $s_1^2 = \frac{1}{n_1-1} \sum_{i=1}^{n_1} (x_i - \bar{x})^2 = 7585.781033$

$$s_2^2 = \frac{1}{n_2-1} \sum_{j=1}^{n_2} (y_j - \bar{y})^2 = 6989.16287$$

$$F = \frac{s_1^2}{s_2^2} = \frac{7585.781033}{6989.16287} = 1.0853$$

Depending on the alternative hypothesis (H_1), degrees of freedom ($n_1-1, n_2-1 = (7,6)$) and level of significance (α), the critical (Table) values are i.e., $F_{(1-\frac{\alpha}{2})} = 0.121$ and $F_{\alpha/2} = 8.26$.

Since, F value lies in the acceptance region; therefore, we do not reject H_0 .

Conclusion: The variance of marks of both the classes doesn't differ significantly. i.e., $\sigma_1^2 = \sigma_2^2$.

Example 6: An experiment was conducted on two groups of plants to compare the growths. Both the plants were given same amount of water and sunlight with the variances in terms of carbon dioxide in normal and enriched air.

The following table gives growth of plants for group1(presence of normal air) and group2 (presence of enriched air). Test whether the variance of growth of plants differs significantly across the group at 1% level of significance.

Plants with normal air (x): 4.67 4.21 2.18 3.91 4.07 5.24 2.94 4.71 4.04
5.79 3.80 4.38

Plants with enriched air(y): 5.04 4.52 6.18 7.01 4.36 1.81 6.22 5.7

Solution: Given $n_1 = 12, n_2 = 8$ and $\alpha = 1\%$. Here $\bar{x} = 4.1616$ and $\bar{y} = 5.105$

H_0 : The variances of growth of plants doesn't differ significantly i.e., $\sigma_1^2 = \sigma_2^2$

H_1 : The variances of growth of plants differ significantly i.e., $\sigma_1^2 \neq \sigma_2^2$

Under H_0 , the test statistic is:

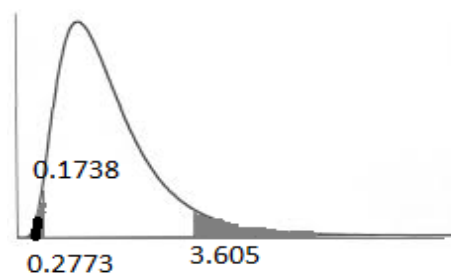
$$F = \frac{s_1^2}{s_2^2} \sim \text{Snedecor's F-distribution with } (n_1-$$

1, $n_2-1)$ d.f

Where, $s_1^2 = \frac{1}{n_1-1} \sum_{i=1}^{n_1} (x_i - \bar{x})^2 = 0.4504$

$$s_2^2 = \frac{1}{n_2-1} \sum_{j=1}^{n_2} (y_j - \bar{y})^2 = 2.5914$$

$$F = \frac{s_1^2}{s_2^2}$$



$$= \frac{0.4504}{2.5914} = 0.1738$$

Depending on the alternative hypothesis (H_1), degrees of freedom ($n_1-1, n_2-1 = (11,7)$) and level of significance (α), the critical (Table) values are i.e., $F_{(1-\frac{\alpha}{2})} = 0.2773$ and $F_{\alpha/2} = 3.605$.

Since, F value lies in the rejection region; therefore, we reject H_0 .

Conclusion: The variances of growth of plants differ significantly i.e., $\sigma_1^2 \neq \sigma_2^2$

Exercise:

1. A random sample of size 20 is taken from a population gives the sample standard deviation 7.5. Test the hypothesis that the population standard deviation is 9 at 1% level of significance.
2. Weights in pounds of 10 sheep are as follows:
85, 92, 87, 101, 98, 113, 115, 87, 93, 110
Can we conclude that the variance of the distribution of weights of sheep is lesser than 50 pounds?
3. The inner diameter of 9 ball bearings was 20.1, 20.35, 20.6, 20.65, 20.32, 20.11, 20.22, 20.48 and 20.23 millimetres. Test the hypothesis that the standard deviation is more than 3 millimetres at 5% level of significance.
4. A normal variate has standard deviation 3. Fifteen sample observations of the variate have standard deviation 4. Test at 1% level of significance whether the population standard deviation is 3.
5. In 150 throws of a single die, the following distribution of faces were obtained.

Faces	1	2	3	4	5	6	Total
Frequency	40	35	18	20	22	15	150

Test at 1% level of significance that the die is unbiased.

6. Demand for a particular product in a market was found to vary from day- to - day. In a sample study the following information was obtained:

Days	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
No. Of products demanded	1234	1122	1100	1200	1232	1412

7. Fit a binomial distribution to the following data and test the goodness of fit.

X	0	1	2	3	4	5	6	7	8	9	10	total
F	2	5	6	10	15	20	18	12	10	9	3	110

8. When the first proof of 395 pages of a book of 1500 pages were read, the distribution of printing mistakes was found to be as follows:

No of mistakes per page	0	1	2	3	4	5	6
No. Of pages	270	75	35	8	4	2	1

Fit a Poisson distribution to the above data and test the goodness of fit at 5% level of significance.

9. The following is the data regarding family condition and examination result of 78 students. Test whether family conditions and results are independent.

Family Conditions	Examination	Result
	Pass	Fail
Good	20	18
Bad	15	25

10. An opinion poll was conducted to find the reaction to a proposed civic reform in 100 members of each of two political parties. The information is tabulated as follows:

	Favourable	Unfavourable	Total
Party X	42	58	100
Party Y	60	40	100

Test whether political parties and the reaction to a proposed civic reform are independent at 1% level of significance.

11. A food service manager for a baseball park wants to know if there is a relationship between gender and the preferred condiment on the hot dog. The following table summarises the result. Test the hypothesis at 5% level of significance.

Gender	Condiment				Total
		Ketchup	Mustard	Relish	
Male		25	19	8	52
Female		15	23	10	48
Total		40	42	18	100

12. A movie producer is bringing out new movie in order to put this in advertising campaign, he wants to determine whether the movie will appeal to a particular age group or whether the movie will appeal equally to all age group.

The producer takes a random sample from person's attending the preview of the movie and obtained the following results:

Taste	Age group (in years)				
		Below 20	20-40	40-60	60 and above
Like		28	146	78	48
Dislike		54	22	42	22
Indifferent		20	10	10	20

Test the hypothesis at 1% level of significance.

14. The following is the data regarding the weights (in grams) of strawberries in 2 boxes. Assume that the weight follows normal distribution. Test whether the variance of strawberries in Box1 is more than that of Box2 at 1 % level of significance.

Box1: 21.7 21 21.2 20.7 20.4 21.9 20.2 21.6 20.6
Box2: 21.5 20.5 20.3 20.1 20 20.4 20.3

15. The following data gives 45 ceramic strength measurements for two batches of material with the following summary statistics:

	No. of observations	Mean	Standard Deviation
Batch 1	24	518.3371	50.2311
Batch 2	21	499.8912	49.5010

Test whether the variances for the two batches differs significantly at 5% level of significance.

16. Two random samples from normal distribution gives the following information:

Sample	Sample size	Sample variance
X	13	6.32
Y	9	4.68

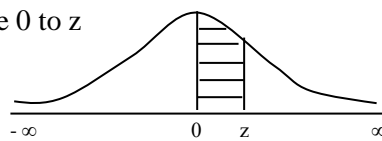
Test whether the variance of sample X is less than that of sample Y at 10 % level of significance.

17. Two random samples from normal distribution gives the following information:

Stripe type	Sample size	Sample variance
A	15	2.1
B	20	1.6

Test whether the variance of strip A and strip B differs significantly at 5 % level of significance.

Table 1: Standard Normal Distribution Table 0 to z



z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
0.7	.2580	.2611	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830

1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990
3.1	.4990	.4991	.4991	.4991	.4992	.4992	.4992	.4992	.4993	.4993
3.2	.4993	.4993	.4994	.4994	.4994	.4994	.4994	.4995	.4995	.4995
3.3	.4995	.4995	.4995	.4996	.4996	.4996	.4996	.4996	.4996	.4997
3.4	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4998
3.5	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998

Table 2: Student's t-distribution table

df	Level of significance(α)							
	0.1	0.05	0.025	0.02	0.01	0.005	0.0025	0.001
1	3.078	6.314	12.71	15.89	31.82	63.66	127.3	318.3
2	1.886	2.92	4.303	4.849	6.965	9.925	14.09	22.33
3	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21
4	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173
5	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893
6	1.44	1.943	2.447	2.612	3.143	3.707	4.317	5.208
7	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785
8	1.397	1.86	2.306	2.449	2.896	3.355	3.833	4.501
9	1.383	1.833	2.262	2.398	2.821	3.25	3.69	4.297
10	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144
11	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025
12	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.93
13	1.35	1.771	2.16	2.282	2.65	3.012	3.372	3.852
14	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787
15	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733

16	1.337	1.746	2.12	2.235	2.583	2.921	3.252	3.686
17	1.333	1.74	2.11	2.224	2.567	2.898	3.222	3.646
18	1.33	1.734	2.101	2.214	2.552	2.878	3.197	3.611
19	1.328	1.729	2.093	2.205	2.539	2.861	3.174	3.579
20	1.325	1.725	2.086	2.197	2.528	2.845	3.153	3.552
21	1.323	1.721	2.08	2.189	2.518	2.831	3.135	3.527
22	1.321	1.717	2.074	2.183	2.508	2.819	3.119	3.505
23	1.319	1.714	2.069	2.177	2.5	2.807	3.104	3.485
24	1.318	1.711	2.064	2.172	2.492	2.797	3.091	3.467
25	1.316	1.708	2.06	2.167	2.485	2.787	3.078	3.45
26	1.315	1.706	2.056	2.162	2.479	2.779	3.067	3.435
27	1.314	1.703	2.052	2.15	2.473	2.771	3.057	3.421
28	1.313	1.701	2.048	2.154	2.467	2.763	3.047	3.408
29	1.311	1.699	2.045	2.15	2.462	2.756	3.038	3.396
30	1.31	1.697	2.042	2.147	2.457	2.75	3.03	3.385
40	1.303	1.684	2.021	2.123	2.423	2.704	2.971	3.307
50	1.295	1.676	2.009	2.109	2.403	2.678	2.937	3.261
60	1.296	1.671	2	2.099	2.39	2.66	2.915	3.232
80	1.292	1.664	1.99	2.088	2.374	2.639	2.887	3.195
100	1.29	1.66	1.984	2.081	2.364	2.626	2.871	3.174
1000	1.282	1.646	1.962	2.056	2.33	2.581	2.813	3.098
inf.	1.282	1.64	1.96	2.054	2.326	2.576	2.807	3.091

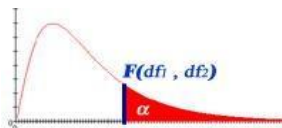
Critical values of Chi-Square distribution

Level of significance(α)

df	0.995	0.99	0.975	0.95	0.05	0.025	0.01	0.005
1	0.04393	0.03157	0.03982	0.02393	3.841	5.024	6.635	7.879
2	0.01	0.0201	0.0506	0.103	5.991	7.378	9.21	10.597
3	0.0717	0.115	0.216	0.352	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	9.488	11.143	13.277	14.86
5	0.412	0.554	0.831	1.145	11.07	12.832	15.086	16.75
6	0.676	0.872	1.237	1.635	12.592	14.449	16.812	18.548
7	0.989	1.239	1.69	2.167	14.067	16.013	18.475	20.278
8	1.344	1.646	2.18	2.733	15.307	17.535	20.09	21.955
9	1.735	2.088	2.7	3.325	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.94	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	19.675	21.92	24.725	26.757
12	3.074	3.571	4.404	5.226	21.026	23.337	26.217	28.3

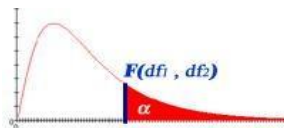
13	3.565	4.107	5.009	5.892	22.362	24.736	27.688	29.819
14	4.075	4.66	5.629	6.571	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	26.296	28.845	32	34.267
17	5.697	6.408	7.564	8.672	27.587	30.191	33.409	35.718
18	6.297	7.015	8.231	9.39	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	30.144	32.852	36.191	38.582
20	7.434	8.26	9.591	10.851	31.41	34.17	37.566	39.997
21	8.034	8.897	10.283	11.591	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	33.924	36.781	40.289	42.796
23	9.26	10.196	11.689	13.091	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	36.415	39.364	42.98	45.558
25	10.52	11.524	13.12	14.611	37.652	40.646	44.314	46.928
26	11.16	12.198	13.844	15.379	38.885	41.923	45.642	48.29
27	11.808	12.879	14.573	16.151	40.113	43.194	46.963	49.645
28	12.461	13.565	15.308	16.928	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	43.773	46.979	50.892	53.672

F Table for $\alpha = 0.05$



/	df ₁ =1	2	3	4	5	6	7	8	9	10	15	20	30	40	60	120	∞		
df₂=1	161.4476	199.5215	215.7073	224.5832	230.1619	233.9862	236.7684	238.8827	240.5433	241.8817	243.9062	245.9499	248.0131	249.0518	250.0951	251.1432	252.1957	253.2529	254.3144
2	18.5128	19	19.1643	19.2468	19.2964	19.3295	19.3532	19.371	19.3848	19.3959	19.4125	19.4291	19.4458	19.4541	19.4624	19.4707	19.4791	19.4874	19.4957
3	10.128	9.5521	9.2766	9.1172	9.0135	8.9406	8.8867	8.8452	8.8123	8.7855	8.7446	8.7029	8.6602	8.6385	8.6166	8.5944	8.572	8.5494	8.5264
4	7.7086	6.9443	6.5914	6.3882	6.2561	6.1631	6.0942	6.041	5.9988	5.9644	5.9117	5.8578	5.8025	5.7744	5.7459	5.717	5.6877	5.6581	5.6281
5	6.6079	5.7861	5.4095	5.1922	5.0503	4.9503	4.8759	4.8183	4.7725	4.7351	4.6777	4.6188	4.5581	4.5272	4.4957	4.4638	4.4314	4.3985	4.365
6	5.9874	5.1433	4.7571	4.5337	4.3874	4.2839	4.2067	4.1468	4.099	4.06	3.9999	3.9381	3.8742	3.8415	3.8082	3.7743	3.7398	3.7047	3.6689
7	5.5914	4.7374	4.3468	4.1203	3.9715	3.866	3.787	3.7257	3.6767	3.6365	3.5747	3.5107	3.4445	3.4105	3.3758	3.3404	3.3043	3.2674	3.2298
8	5.3177	4.459	4.0662	3.8379	3.6875	3.5806	3.5005	3.4381	3.3881	3.3472	3.2839	3.2184	3.1503	3.1152	3.0794	3.0428	3.0053	2.9669	2.9276
9	5.1174	4.2565	3.8625	3.6331	3.4817	3.3738	3.2927	3.2296	3.1789	3.1373	3.0729	3.0061	2.9365	2.9005	2.8637	2.8259	2.7872	2.7475	2.7067
10	4.9646	4.1028	3.7083	3.478	3.3258	3.2172	3.1355	3.0717	3.0204	2.9782	2.913	2.845	2.774	2.7372	2.6996	2.6609	2.6211	2.5801	2.5379
11	4.8443	3.9823	3.5874	3.3567	3.2039	3.0946	3.0123	2.948	2.8962	2.8536	2.7876	2.7186	2.6464	2.609	2.5705	2.5309	2.4901	2.448	2.4045
12	4.7472	3.8853	3.4903	3.2592	3.1059	2.9961	2.9134	2.8486	2.7964	2.7534	2.6866	2.6169	2.5436	2.5055	2.4663	2.4259	2.3842	2.341	2.2962

13	4.6672	3.8056	3.4105	3.1791	3.0254	2.9153	2.8321	2.7669	2.7144	2.671	2.6037	2.5331	2.4589	2.4202	2.3803	2.3392	2.2966	2.2524	2.2064
14	4.6001	3.7389	3.3439	3.1122	2.9582	2.8477	2.7642	2.6987	2.6458	2.6022	2.5342	2.463	2.3879	2.3487	2.3082	2.2664	2.2229	2.1778	2.1307
15	4.5431	3.6823	3.2874	3.0556	2.9013	2.7905	2.7066	2.6408	2.5876	2.5437	2.4753	2.4034	2.3275	2.2878	2.2468	2.2043	2.1601	2.1141	2.0658
16	4.494	3.6337	3.2389	3.0069	2.8524	2.7413	2.6572	2.5911	2.5377	2.4935	2.4247	2.3522	2.2756	2.2354	2.1938	2.1507	2.1058	2.0589	2.0096
17	4.4513	3.5915	3.1968	2.9647	2.81	2.6987	2.6143	2.548	2.4943	2.4499	2.3807	2.3077	2.2304	2.1898	2.1477	2.104	2.0584	2.0107	1.9604
18	4.4139	3.5546	3.1599	2.9277	2.7729	2.6613	2.5767	2.5102	2.4563	2.4117	2.3421	2.2686	2.1906	2.1497	2.1071	2.0629	2.0166	1.9681	1.9168
19	4.3807	3.5219	3.1274	2.8951	2.7401	2.6283	2.5435	2.4768	2.4227	2.3779	2.308	2.2341	2.1555	2.1141	2.0712	2.0264	1.9795	1.9302	1.878
20	4.3512	3.4928	3.0984	2.8661	2.7109	2.599	2.514	2.4471	2.3928	2.3479	2.2776	2.2033	2.1242	2.0825	2.0391	1.9938	1.9464	1.8963	1.8432
21	4.3248	3.4668	3.0725	2.8401	2.6848	2.5727	2.4876	2.4205	2.366	2.321	2.2504	2.1757	2.096	2.054	2.0102	1.9645	1.9165	1.8657	1.8117
22	4.3009	3.4434	3.0491	2.8167	2.6613	2.5491	2.4638	2.3965	2.3419	2.2967	2.2258	2.1508	2.0707	2.0283	1.9842	1.938	1.8894	1.838	1.7831
23	4.2793	3.4221	3.028	2.7955	2.64	2.5277	2.4422	2.3748	2.3201	2.2747	2.2036	2.1282	2.0476	2.005	1.9605	1.9139	1.8648	1.8128	1.757
24	4.2597	3.4028	3.0088	2.7763	2.6207	2.5082	2.4226	2.3551	2.3002	2.2547	2.1834	2.1077	2.0267	1.9838	1.939	1.892	1.8424	1.7896	1.733
25	4.2417	3.3852	2.9912	2.7587	2.603	2.4904	2.4047	2.3371	2.2821	2.2365	2.1649	2.0889	2.0075	1.9643	1.9192	1.8718	1.8217	1.7684	1.711
26	4.2252	3.369	2.9752	2.7426	2.5868	2.4741	2.3883	2.3205	2.2655	2.2197	2.1479	2.0716	1.9898	1.9464	1.901	1.8533	1.8027	1.7488	1.6906
27	4.21	3.3541	2.9604	2.7278	2.5719	2.4591	2.3732	2.3053	2.2501	2.2043	2.1323	2.0558	1.9736	1.9299	1.8842	1.8361	1.7851	1.7306	1.6717
28	4.196	3.3404	2.9467	2.7141	2.5581	2.4453	2.3593	2.2913	2.236	2.19	2.1179	2.0411	1.9586	1.9147	1.8687	1.8203	1.7689	1.7138	1.6541
29	4.183	3.3277	2.934	2.7014	2.5454	2.4324	2.3463	2.2783	2.2229	2.1768	2.1045	2.0275	1.9446	1.9005	1.8543	1.8055	1.7537	1.6981	1.6376
30	4.1709	3.3158	2.9223	2.6896	2.5336	2.4205	2.3343	2.2662	2.2107	2.1646	2.0921	2.0148	1.9317	1.8874	1.8409	1.7918	1.7396	1.6835	1.6223
40	4.0847	3.2317	2.8387	2.606	2.4495	2.3359	2.249	2.1802	2.124	2.0772	2.0035	1.9245	1.8389	1.7929	1.7444	1.6928	1.6373	1.5766	1.5089
60	4.0012	3.1504	2.7581	2.5252	2.3683	2.2541	2.1665	2.097	2.0401	1.9926	1.9174	1.8364	1.748	1.7001	1.6491	1.5943	1.5343	1.4673	1.3893
120	3.9201	3.0718	2.6802	2.4472	2.2899	2.175	2.0868	2.0164	1.9588	1.9105	1.8337	1.7505	1.6587	1.6084	1.5543	1.4952	1.429	1.3519	1.2539
∞	3.8415	2.9957	2.6049	2.3719	2.2141	2.0986	2.0096	1.9384	1.8799	1.8307	1.7522	1.6664	1.5705	1.5173	1.4591	1.394	1.318	1.2214	1



F Table for $\alpha = 0.025$

$df_1 \backslash df_2$	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	647.789	799.5	864.163	899.5833	921.8479	937.1111	948.2169	956.6562	963.2846	968.6274	976.7079	984.8668	993.1028	997.2492	1001.414	1005.598	1009.8	1014.02	1018.258
2	38.5063	39.000	39.1655	39.2484	39.2982	39.3315	39.3552	39.373	39.3869	39.398	39.4146	39.4313	39.4479	39.4562	39.465	39.473	39.481	39.49	39.498
3	17.4434	16.044	15.4392	15.101	14.8848	14.7347	14.6244	14.5399	14.4731	14.4189	14.3366	14.2527	14.1674	14.1241	14.081	14.037	13.992	13.947	13.902
4	12.2179	10.649	9.9792	9.6045	9.3645	9.1973	9.0741	8.9796	8.9047	8.8439	8.7512	8.6565	8.5599	8.5109	8.461	8.411	8.36	8.309	8.257
5	10.007	8.4336	7.7636	7.3879	7.1464	6.9777	6.8531	6.7572	6.6811	6.6192	6.5245	6.4277	6.3286	6.278	6.227	6.175	6.123	6.069	6.015
6	8.8131	7.2599	6.5988	6.2272	5.9876	5.8198	5.6955	5.5996	5.5234	5.4613	5.3662	5.2687	5.1684	5.1172	5.065	5.012	4.959	4.904	4.849
7	8.0727	6.5415	5.8898	5.5226	5.2852	5.1186	4.9949	4.8993	4.8232	4.7611	4.6658	4.5678	4.4667	4.415	4.362	4.309	4.254	4.199	4.142
8	7.5709	6.0595	5.416	5.0526	4.8173	4.6517	4.5286	4.4333	4.3572	4.2951	4.1997	4.1012	3.9995	3.9472	3.894	3.84	3.784	3.728	3.67
9	7.2093	5.7147	5.0781	4.7181	4.4844	4.3197	4.197	4.102	4.026	3.9639	3.8682	3.7694	3.6669	3.6142	3.56	3.505	3.449	3.392	3.333
10	6.9367	5.4564	4.8256	4.4683	4.2361	4.0721	3.9498	3.8549	3.779	3.7168	3.6209	3.5217	3.4185	3.3654	3.311	3.255	3.198	3.14	3.08
11	6.7241	5.2559	4.63	4.2751	4.044	3.8807	3.7586	3.6638	3.5879	3.5257	3.4296	3.3299	3.2261	3.1725	3.118	3.061	3.004	2.944	2.883
12	6.5538	5.0959	4.4742	4.1212	3.8911	3.7283	3.6065	3.5118	3.4358	3.3736	3.2773	3.1772	3.0728	3.0187	2.963	2.906	2.848	2.787	2.725
13	6.4143	4.9653	4.3472	3.9959	3.7667	3.6043	3.4827	3.388	3.312	3.2497	3.1532	3.0527	2.9477	2.8932	2.837	2.78	2.72	2.659	2.595

14	6.2979	4.8567	4.2417	3.8919	3.6634	3.5014	3.3799	3.2853	3.2093	3.1469	3.0502	2.9493	2.8437	2.7888	2.732	2.674	2.614	2.552	2.487
15	6.1995	4.765	4.1528	3.8043	3.5764	3.4147	3.2934	3.1987	3.1227	3.0602	2.9633	2.8621	2.7559	2.7006	2.644	2.585	2.524	2.461	2.395
16	6.1151	4.6867	4.0768	3.7294	3.5021	3.3406	3.2194	3.1248	3.0488	2.9862	2.889	2.7875	2.6808	2.6252	2.568	2.509	2.447	2.383	2.316
17	6.042	4.6189	4.0112	3.6648	3.4379	3.2767	3.1556	3.061	2.9849	2.9222	2.8249	2.723	2.6158	2.5598	2.502	2.442	2.38	2.315	2.247
18	5.9781	4.5597	3.9539	3.6083	3.382	3.2209	3.0999	3.0053	2.9291	2.8664	2.7689	2.6667	2.559	2.5027	2.445	2.384	2.321	2.256	2.187
19	5.9216	4.5075	3.9034	3.5587	3.3327	3.1718	3.0509	2.9563	2.8801	2.8172	2.7196	2.6171	2.5089	2.4523	2.394	2.333	2.27	2.203	2.133
20	5.8715	4.4613	3.8587	3.5147	3.2891	3.1283	3.0074	2.9128	2.8365	2.7737	2.6758	2.5731	2.4645	2.4076	2.349	2.287	2.223	2.156	2.085
21	5.8266	4.4199	3.8188	3.4754	3.2501	3.0895	2.9686	2.874	2.7977	2.7348	2.6368	2.5338	2.4247	2.3675	2.308	2.246	2.182	2.114	2.042
22	5.7863	4.3828	3.7829	3.4401	3.2151	3.0546	2.9338	2.8392	2.7628	2.6998	2.6017	2.4984	2.389	2.3315	2.272	2.21	2.145	2.076	2.003
23	5.7498	4.3492	3.7505	3.4083	3.1835	3.0232	2.9023	2.8077	2.7313	2.6682	2.5699	2.4665	2.3567	2.2989	2.239	2.176	2.111	2.041	1.968
24	5.7166	4.3187	3.7211	3.3794	3.1548	2.9946	2.8738	2.7791	2.7027	2.6396	2.5411	2.4374	2.3273	2.2693	2.209	2.146	2.08	2.01	1.935
25	5.6864	4.2909	3.6943	3.353	3.1287	2.9685	2.8478	2.7531	2.6766	2.6135	2.5149	2.411	2.3005	2.2422	2.182	2.118	2.052	1.981	1.906
26	5.6586	4.2655	3.6697	3.3289	3.1048	2.9447	2.824	2.7293	2.6528	2.5896	2.4908	2.3867	2.2759	2.2174	2.157	2.093	2.026	1.954	1.878
27	5.6331	4.2421	3.6472	3.3067	3.0828	2.9228	2.8021	2.7074	2.6309	2.5676	2.4688	2.3644	2.2533	2.1946	2.133	2.069	2.002	1.93	1.853
28	5.6096	4.2205	3.6264	3.2863	3.0626	2.9027	2.782	2.6872	2.6106	2.5473	2.4484	2.3438	2.2324	2.1735	2.112	2.048	1.98	1.907	1.829
29	5.5878	4.2006	3.6072	3.2674	3.0438	2.884	2.7633	2.6686	2.5919	2.5286	2.4295	2.3248	2.2131	2.154	2.092	2.028	1.959	1.886	1.807
30	5.5675	4.1821	3.5894	3.2499	3.0265	2.8667	2.746	2.6513	2.5746	2.5112	2.412	2.3072	2.1952	2.1359	2.074	2.009	1.94	1.866	1.787
40	5.4239	4.051	3.4633	3.1261	2.9037	2.7444	2.6238	2.5289	2.4519	2.3882	2.2882	2.1819	2.0677	2.0069	1.943	1.875	1.803	1.724	1.637
60	5.2856	3.9253	3.3425	3.0077	2.7863	2.6274	2.5068	2.4117	2.3344	2.2702	2.1692	2.0613	1.9445	1.8817	1.815	1.744	1.667	1.581	1.482
120	5.1523	3.8046	3.2269	2.8943	2.674	2.5154	2.3948	2.2994	2.2217	2.157	2.0548	1.945	1.8249	1.7597	1.69	1.614	1.53	1.433	1.31
∞	5.0239	3.6889	3.1161	2.7858	2.5665	2.4082	2.2875	2.1918	2.1136	2.0483	1.9447	1.8326	1.7085	1.6402	1.566	1.484	1.388	1.268	1

s

Reference

1. S.C. Gupta and V.K. Kapoor: Fundamentals of Mathematical Statistics, 11th Edn.2013, Sultan Chand Publications, New Delhi.
2. Miller, Irwin and Miller, Marylees (2006): John E. Freund's Mathematical Statistics with Applications, (7th Edn.), Pearson Education, Asia.
3. Hogg, Tannis and Rao: Probability and Statistical inference, 7th Edn. - 2008, Pearson.
4. V.K Rohatgi: An introduction to probability and Statistics, 1st Edn, 2015, Jhon Wiley & Sons, Inc.
5. Gilles Cazalais. Typeset with LATEX on April 20, 2006.

