



MSCBOT-606

M. Sc. IV Semester **BIOSTATISTICS**

Standard Deviation

Give logical explanation of the relative facts



Analyze the relationship between variables



Search the common points in the data



Now carefully make a generalized statement

Data Collection

Data set

$$\sigma = \sqrt{\frac{(x - \bar{x})^2}{n}}$$

Variables

$$\text{Mean} = \frac{x_1 + x_2 + x_3 + \dots + x_n}{n}$$

**Analysis of
Variance (ANOVA)**

$$\text{Median} = \frac{\left\{ \left(\frac{n}{2}\right)^{\text{th}} \text{ observation} + \left(\frac{n}{2} + 1\right)^{\text{th}} \text{ observation} \right\}}{2}$$

Class intervals	10-20	20-30	30-40	40-50	50-60	60-70	70-80
No. of plants	5	12	15	20	10	4	2

**DEPARTMENT OF BOTANY
SCHOOL OF SCIENCES
UTTARAKHAND OPEN UNIVERSITY**

BIOSTATISTICS



**DEPARTMENT OF BOTANY
SCHOOL OF SCIENCES
UTTARAKHAND OPEN UNIVERSITY**

Phone No. 05946-261122, 261123

Toll free No. 18001804025

Fax No. 05946-264232, E. mail info@uou.ac.in

<http://uou.ac.in>

Expert Committee
Prof. J.C. Ghildiyal

Retired Principal
Govt. PG College, Karnprayag

Prof. G.S. Rajwar

Principal
Government PG College, Augustmuni

Prof. Lalit M. Tewari

Department of Botany
DSB Campus, Kumaun University, Nainital

Dr. Hemant Kandpal

School of Health Science
Uttarakhand Open University, Haldwani

Dr. Pooja Juyal

Department of Botany, School of Sciences
Uttarakhand Open University, Haldwani

Board of Studies
Prof. P.D. Pant

Director, School of Sciences
Uttarakhand Open University, Haldwani

Prof. S.S. Bargali

HOD, Department of Botany
DSB Campus, Kumaun University, Nainital

Prof. Amrita Nigam

School of Sciences
IGNOU, New Delhi

Dr. S.S. Samant

Retd. Director
Himalayan Forest Research Institute (H.P)

Dr. S.N. Ojha

Assistant Professor
Department of Botany
Uttarakhand Open University, Haldwani

Dr. Pooja Juyal

Assistant Professor (AC)
Department of Botany
Uttarakhand Open University, Haldwani

Dr. Kirtika Padalia

Assistant Professor (AC)
Department of Botany
Uttarakhand Open University, Haldwani

Dr. Prabha Dhondiyal

Assistant Professor (AC)
Department of Botany
Uttarakhand Open University, Haldwani

Dr. Pushpesh Joshi

Assistant Professor (AC)
Department of Botany
Uttarakhand Open University, Haldwani

Programme Co-ordinator
Dr. S.N. Ojha

Assistant Professor
Department of Botany, School of Sciences,
Uttarakhand Open University, Haldwani, Nainital

	Unit Written By:	Unit No.
1.	Dr. Manish Tripathi Scientist 'B' GBP-NIHE, Reginal Centres, Kullu, Himachal Pradesh	1 & 2
2.	Dr. Arun Khajuriya Assistant Professor Department of Botany Cluster University of Jammu, Jammu and Kashmir	3 & 4
3.	Dr. S. N. Ojha Assistant Professor Department of Botany Uttarakhand Open University, Haldwani	5
4.	Dr. Sparsh Bhatt Department of Statistics DSB Campus, Kumaun University, Nainital	6,7, & 8

Chief Course Editor

Dr. S.N. Ojha

Assistant Professor

Department of Botany, School of Sciences

Uttarakhand Open University, Haldwani, Nainital

Co- Editors

Dr. Pooja Juyal

Assistant Professor (AC)

Department of Botany

School of Sciences

Uttarakhand Open University, Haldwani

Dr. Kirtika Padalia

Assistant Professor (AC)

Department of Botany

School of Sciences

Uttarakhand Open University, Haldwani

Dr. Prabha Dhondiyal

Assistant Professor (AC)

Department of Botany

School of Sciences

Uttarakhand Open University, Haldwani

Dr. Pushpesh Joshi

Assistant Professor (AC)

Department of Botany

School of Sciences

Uttarakhand Open University, Haldwani

Title	:	Biostatistics
ISBN No.	:	
Copyright	:	Uttarakhand Open University
Edition	:	2023

Published By: Uttarakhand Open University, Haldwani, Nainital-263139

CONTENTS

BLOCK-1- THE DATA: BASIC CONCEPTS		PAGE NO.
Unit-1	Data Collection	1-14
Unit-2	Data Presentation	15-25
BLOCK-2- STATISTICAL METHODS		
Unit-3	Descriptive Statistics	27-66
Unit-4	Statistical Inference and Probability	67-94
Unit-5	Advances Analysis Methods	95-128
BLOCK-3- HYPOTHESIS TESTING AND EXPERIMENTAL DESIGNS		
Unit-6	Testing of Hypothesis	130-170
Unit-7	Analysis of Variance (ANOVA)	171-191
Unit-8	Design of Experiments	192-208

BLOCK-1-THE DATA: BASIC CONCEPTS

UNIT-1- DATA COLLECTION

Contents

- 1.1- Objectives
- 1.2- Introduction
- 1.3- Collection of primary and secondary data
 - 1.3.1- Collection of primary data
 - 1.3.2- Collection of secondary data
- 1.4- Selection of appropriate methods of data collection
- 1.5- Classification and tabulation of data
 - 1.5.1 Classification of data
 - 1.5.2 Tabulation of data
- 1.6- Data interpretation
- 1.7- Glossary
- 1.8- References
- 1.9- Suggested readings
- 1.10- Self assessment Questions
- 1.11- Terminal Questions
 - 1.11.1- Short answer type questions
 - 1.11.2- Long answer type questions

1.1 OBJECTIVES

After reading this unit the students will be familiar to the;

- Collection of primary and secondary data
- Selection of appropriate methods of data collection
- Classification and tabulation of data
- Data interpretation

1.2 INTRODUCTION

Statistics is the study of collecting, organizing, analyzing and interpreting data. In simple words statistics is the science of converting data into information. Statistical analysis of any data helps us to make informed decisions and how these decisions might affect us in the long run. Data are the key in statistics, so now the question arises what are data? Data are the raw information from which statistics are created. If we want to understand for example why people like ‘A’ shoe brand over ‘B’, we need data. Raw data is collected as a part of research, observations and surveys. A data unit is one entity in the population being studied, about which data are collected. A population is any complete group with at least one characteristic in common. It is the complete pool from which a statistical sample is drawn. For example if we want to study the eye color of adult humans in Uttarakhand, the population would be all adult humans in Uttarakhand.

Generally it is not possible to measure/count every unit in a given population. Hence a representative sample is taken to study. A sample is a sub-set of the population, selected to represent all units in a population to be studied. Information acquired from the sampled units is used to conclude the characteristics for the entire population to be studied. Sample must be large enough to provide reliable representation of whole population. Samples should be selected in a random manner so that each unit in the population has equal and independent opportunity of being selected. Random sampling reduces bias and sampling error.

1.3 COLLECTION OF PRIMARY AND SECONDARY DATA

Before collecting the data we should formulate our research problem meaning for which purpose we are going to collect the data. Then we should focus on the type of data we want to collect. Data are of two types: primary data and secondary data.

The **primary data** are original and collected for the first time. The **secondary data** are collected by someone else and have already been passed through the statistical process. The selection of the data type (primary or secondary) and the methods of collecting the data depend on the type of study you are conducting.

There are various methods of data collection mentioned below:

1.3.1 COLLECTION OF PRIMARY DATA

Primary data are collected during the experimental, descriptive researches and surveys. It can be collected via observation or direct communication with respondents or personal interviews. There are various methods of collecting primary data such as observation method, interview method and questionnaires.

1.3.1.1 OBSERVATION METHOD

It is the most commonly used method. It is systematically planned and recorded and subjected to checks and controls on validity and reliability.

Types: The observations are of two types; structured observations and unstructured observations.

- **Structured observations:** When the observation is characterized by a careful definition of the units to be observed, the style of recording the observed information, standardized conditions of observation and the selection of pertinent data of observation. It is used mostly in descriptive studies.
- **Unstructured observations:** These observations take place without the above mentioned characteristics to be thought of in advance. It is used mostly in exploratory study.
- **Participant observations:** It is type of observation in which the observer him/herself becomes a member of a group under observation.
- **Non-participant observations:** In this type the observer is a separate entity from the group under observation.
- **Disguised observations:** In this type the presence of the observer may be unknown to the group under observation.

Advantages: In this method, the information is gathered by the researcher's direct observation without asking from the respondent. Suppose you want to know about the most favorite shoe brands in Haldwani then you have no need to ask the respondent, you may yourself look at the shoe respondents are wearing and record the data. There is no subjective bias. It does not depend on respondent's willingness to participate in the process unlike in the interview or the questionnaire method.

Limitations: It is an expensive method and the information gathered is very limited. Sometimes unforeseen factors may interfere with the observational task. At times, the fact that some people are rarely accessible to direct observation creates obstacle for this method to collect data effectively.

Precautions: Your observation should be performed very carefully. It is very significant that how you are recording the observations. Please ensure the accuracy of observations.

1.3.1.2 INTERVIEW METHOD

This method includes the verbal communication between the interviewer and the respondent. The interviews can be conducted in person or over the telephone.

- **Personal interviews:** These kinds of interviews should be conducted in person with the respondent/s. Personal interviews can be of two types: Structured and unstructured.
- **Structured interviews:** In this method we should use a set of predetermined questions and of highly standardized techniques of recording. We follow a rigid system of asking questions in a prearranged form and order.
- **Unstructured interviews:** There is no predetermined set of questions. We can ask questions in a flexible manner. In this method we have more liberty of questioning, we can ask supplementary questions and according to situation we can even skip some questions.

Advantages: Through this method detailed information can be gathered. It is a very flexible method in comparison to the other methods of data collection. We can collect personal information by this method. In this method the interviewer can control that which respondent will reply. Among all the data collection methods this method collects the most spontaneous reactions of respondents. The language and questions of the interview can be changed according to the educational level of the respondent.

Limitations: It is a very time consuming and expensive method in case the sample size is large and contains geographically distant areas. There is a possibility of bias of interviewer and the respondent. Some high profile respondents might not be easily approachable. This method mainly depends on the credibility of the respondent/s.

Precautions: Interviewers should be honest, sincere, hardworking, impartial and must have necessary practical experience. The interviewers should be monitored so that cheating can be avoided. The interviewer must clear all the doubts of respondent/s about the interview. The interviewers should be friendly, courteous, conversational and unbiased with the respondent/s.

- **Telephone interviews:** In this method we contact the respondents through the telephone. It is a very uncommon method and used when the respondent is out of reach and in industrial surveys.

Advantages: It is more quick and flexible in comparison to other methods. It is a pocket friendly method than personal interview. In this method the rate of response is higher and the responses can be recorded without making the respondent uncomfortable. This method does not require any field staff. In this method more representation and wider distribution of sample is possible.

Limitations: In this method less time is given to respondents for answering in comparison to the interview method. We can only collect information from the respondents who have telephonic facilities. Telephonic interviews cannot have long or detailed questions so the questions have to be short and to the point.

1.3.1.3 QUESTIONNAIRE METHOD

This is one of the most common methods of data collection. It is used by common individuals, researchers, organizations and government institutions. In this method a questionnaire is provided to the respondents to answer the questions and after that the questionnaire is collected back. A questionnaire consists of a number of questions in a definite order. The respondents have to answer the questions by themselves and without any external help. The questionnaires can also be mailed to the respondents.

Advantages: In case of data collection from large area this method is cost effective. This method is unbiased because there is no interviewer to affect the answers whatsoever. This method gives adequate time to respondents to answer.

Limitations: In case of the mailed questionnaires there is low rate of return of the duly filled in questionnaires and sometimes very few questions are answered properly. This method requires a certain level of education and cooperation among the respondents. In case of the mailed questionnaires we cannot know for sure that the answers given by respondents are actually given by them or someone else filled their questionnaires.

Precautions: Before using questionnaire method, we should conduct a pilot survey for testing the questionnaires. Pilot survey is just the rehearsal of the main survey. Through the pilot survey we can understand the flaws in our technique and also can modify the questionnaire according to our respondents.

1.3.2 COLLECTION OF SECONDARY DATA

As we have seen earlier in this chapter that collection of primary data is a tedious task and it involves a lot of money, time and manpower. Hence to save all the trouble sometimes we collect data which are already available in published and unpublished form, called as Secondary data. These secondary data are collected by some organization or individual in the past, so we just go to that source and collect that data to avoid the problems associated with the collection of original data.

We can collect published data from different reports of foreign, central, state and local governments, international bodies, various journals, books, magazines, newspapers, historical documents and other sources of published information. We can collect published data from diaries, letters, unpublished biographies and autobiographies, unpublished research work etc.

We should be very careful in applying the secondary data to our research because that data have not been collected according to our objectives. Secondary data might not be suitable or adequate in the context of our research. Before using the data we should check the origin point of data, source of data, the collection methodology of data, the time of data collection, bias in the compiler and the level of accuracy. Only after getting ensured about all these points we can use the data.

1.4 SELECTION OF APPROPRIATE METHODS OF DATA COLLECTION

Data collection is a crucial step in statistical analysis, as the quality and reliability of the data gathered directly impact the validity of the statistical findings. The selection of appropriate methods of data collection plays a vital role in ensuring accurate and meaningful results. Various factors need to be considered when determining the most suitable methods for collecting data. Here are some key considerations:

Research Objectives: The first step in selecting appropriate data collection methods is to clearly define the research objectives. Different research questions may require different data collection techniques. For example, if the objective is to understand consumer preferences, surveys or interviews may be appropriate. If the objective is to analyze sales trends, collecting sales transaction data might be more suitable.

Nature of the Data: Consider the type of data required for the analysis. Data can be quantitative (numeric) or qualitative (descriptive). Quantitative data often involves measurements, such as age, income, or product ratings. Qualitative data, on the other hand, focuses on subjective information, such as opinions, experiences, or narratives. Depending on the nature of the data, methods like surveys, experiments, observations, or interviews may be used.

Population and Sample: Determine the population of interest and whether it is feasible to collect data from the entire population or a representative sample. If the population is large, a sample may be more practical. Sampling techniques like simple random sampling, stratified sampling, or cluster sampling can be employed. If the population is small, it may be possible to collect data from all individuals or units.

Data Collection Instruments: Choose the appropriate tools or instruments for collecting data. Surveys, questionnaires, interviews, observation checklists, or measurement devices are common instruments. The selection depends on factors such as the type of data, the level of detail required, and the resources available. Ensure that the instruments are reliable, valid, and appropriate for the target population.

Time and Resources: Consider the available time and resources for data collection. Some methods may be more time-consuming and costly than others. For instance, conducting face-to-face interviews can be resource-intensive compared to online surveys. Evaluate the trade-offs between the desired level of accuracy and the practical constraints.

Ethical Considerations: Pay attention to ethical considerations when selecting data collection methods. Ensure that privacy and confidentiality are maintained, informed consent is obtained when necessary, and any potential risks to participants are minimized. Adhere to relevant ethical guidelines and regulations.

Data Quality and Accuracy: Consider the potential for errors and biases associated with different data collection methods. Minimize sources of bias and ensure data quality through

proper training of data collectors, clear instructions, and appropriate validation measures. Statistical techniques such as data cleaning and outlier detection can also be employed to enhance data accuracy.

Feasibility and Practicality: Evaluate the feasibility and practicality of different data collection methods within the given research context. Consider factors such as the availability of participants, accessibility of the target population, logistical requirements, and budget constraints.

In summary, the selection of appropriate methods of data collection in statistics involves considering research objectives, the nature of data, population and sample, data collection instruments, time and resources, ethical considerations, data quality, and feasibility. By carefully assessing these factors, researchers can choose the most suitable data collection methods to ensure reliable and valid statistical analysis.

1.5 CLASSIFICATION AND TABULATION OF DATA

1.5.1 CLASSIFICATION OF DATA

In most of the researches huge amount of data are collected and is called raw data. By using the raw data we cannot predict or suggest anything about the research problem in focus. Hence, the raw data are grouped into different homogenous groups so that we can use it for the statistical analysis. This grouping of data is called as classification of data. Classification of data is the process of arranging data in groups on the basis of common characteristics. The data which exhibit common characteristics are placed in one class and in this manner the complete data get divided into a number of groups or classes. For example human population can be classified into two broad groups on the basis of sex, males and females; or on the basis of literacy, literate and illiterate; or plants can be arranged into various classes according to their heights.

The characteristics are of two types: descriptive and numerical.

Descriptive: These characteristics include the individuals which cannot be numerically measured such as blindness, education, honesty, sex etc.

Numerical: These characteristics include the individuals which can be numerically measured such as age, height, weight, yield etc.

Types of Classification:

Two types of data classifications are available:

- i) **Classification according to attributes:** This system of classification deals with the data which have the descriptive characteristics or qualities or attributes. Descriptive or qualitative characteristics cannot be measured, only their presence or absence in an individual item can be noticed. Data obtained this way on the basis of certain attributes are

known as *statistics of attributes* and their classification is said to be classification according to attributes.

This classification system can further be divided into simple or manifold classification. In simple classification we take only one quality or attribute and divide the data into two classes, one class with items having the considered attribute and the other class with items not having the considered attribute. However, in manifold classification we take more than one attributes simultaneously, and divide that data into a number of classes.

- ii) ***Classification according to class-intervals:*** This system of classification deals with the data having the quantitative or numerical characters which can be measured. Such data are known as *statistics of variables*. For instance, the plant height is the criterion of the classification, so we will make various classes representing a range of heights e.g. 25-50 cm, 50-75 cm, 75-100 cm and above 100 cm. These groups are known as class intervals.

The class intervals have two limits, upper limit and lower limit. The difference between these two limits is called as width of class intervals. The number of observations lying in any class interval is known as frequency of that class interval. The way in which the observations are classified and distributed in the proper class intervals is known as frequency distribution.

1.5.2 TABULATION OF DATA

After the classification, now the data are arranged in a concise and logical order, this process is known as tabulation of data. Technically speaking tabulation is an orderly arrangement of data in a table with columns and rows. Tabulation of data is very significant because it helps in the comparison of data, detection of errors and omissions.

Types of Tabulation:

There are various kinds of tables are present but generally tabulation is of two types: Simple tabulation and complex tabulation.

Simple Tabulation: This type of table provides information related to about one or more groups of independent questions. For instance, the population of plants of an area can be divided into many families. Here are only two columns, one for the different families and the other for the number of genera fall under that particular family. In this table we study only one point which is pertaining to the number of genera falling under various families and because of that reason it is also called as simple table or one-way table.

S. N.	Family	Number of Genera
1	Asteraceae	
2	Brassicaceae	
3	Rosaceae	
4	etc.	
	Total	

- **Complex Tabulation:** This type of table provides information related to more than one group of inter-related questions. Complex tabulation is of three types:

i) **Double table or two-way table:** In this table, the numerical data are classified according to two characteristics and then finally tabulated. For example, the population of genera can be divided according to the families under which they fall and can be further be subdivided according to the type of fruit (e.g. drupe or berry) they produce. This kind of table is called as double table or two-way table.

S. No.	Family	Number of Genera		Total
		Drupe	Berry	
1	Asteraceae			
2	Brassicaceae			
3	Rosaceae			
4	etc.			
	Total			

ii) **Treble table or three-way table:** In this table, the data are classified according to three characteristics and then finally tabulated. For example, the population of genera can be divided according to the families under which they fall and can be further be subdivided according to the type of fruit (e.g. drupe or berry) they produce and finally divided according to the color of petals (e.g. red or yellow) they have possess. This kind of table is called as treble table or three-way table.

S. N.	Family	Number of Genera						Total		
		Drupe			Berry			Red	Yellow	Total
		Red	Yellow	Total	Red	Yellow	Total			
1	Asteraceae									
2	Brassicaceae									
3	Rosaceae									
4	etc.									
	Total									

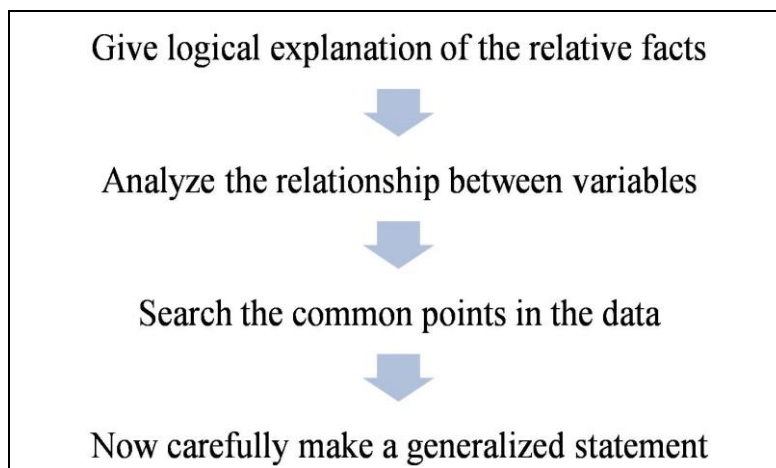
iii) **Manifold table:** In this table, the data are classified according to more than three characteristics and then finally tabulated. For example, the population of genera can be divided according to the families under which they fall and can be further be subdivided according to the type of fruit (e.g. drupe or berry) they produce and again divided according to the color of petals (e.g. red or yellow) they have possess and after that can further be classified according to the type of life cycle (e.g. annual, biennial or perennial) they follow. This kind of table is called as manifold table.

SN	Family		Number of Genera						Total		
			Drupe			Berry			Red	Yellow	Total
			Red	Yellow	Total	Red	Yellow	Total			
1	Asteraceae	Annual									
		Biennial									
		Perennial									
		Total									
2	Brassicaceae	Annual									
		Biennial									
		Perennial									
		Total									
3	Rosaceae	Annual									
		Biennial									
		Perennial									
		Total									
4	etc.										
	Total										

1.6 DATA INTERPRETATION

After collection and tabulation of data the next task is to know what this data is suggesting about the objectives of our study. This is the most significant step of the statistical analysis and should be performed very carefully otherwise the results of the study can flip upside down. The interpretation of the data tells us the relations between the different variables and their effect on the objectives of the study. So now the question is what does interpretation means. Interpretation means making conclusive remarks from the collected data after analyzing it. Interpretation is the basic element in any research because without it we cannot connect to the results of other studies and further cannot predict the future prospects.

Flow chart: How to interpret the data



1.7 GLOSSARY

Analysis: An investigation of the component parts of a whole and their relations in making up the whole.

Attributes: The characters of an entity.

Bias: A partiality that prevents objective consideration of an issue or situation.

Classification: The act of distributing things into classes or categories of the same type.

Data: A collection of facts from which conclusions may be drawn.

Generalization: Reasoning from detailed facts to general principles.

Group: Any number of entities (members) considered as a unit.

Interpretation: A mental representation of the meaning or significance of something.

Observation: The act of making and recording a measurement.

Population: The entire aggregation of items from which samples can be drawn.

Qualitative: Relating to or involving comparisons based in qualities.

Quantitative: Relating to the measurement of quantity.

Questionnaire: A form containing a set of questions; submitted to people to gain statistical information.

Raw Data: The data which are not yet subjected to analysis.

Sample: Items selected at random from a population and used to test hypotheses about the population.

Tabulation: The act of putting information into tabular form.

1.8 REFERENCES

1. Chandel SRS (2013) Handbook of Agricultural Statistics. Achal Prakashan Mandir.
2. Kothari CR (2004) Research Methodology: Methods and Techniques. New Age International (P) Ltd.

1.9 SUGGESTED READINGS

1. Chandel SRS (2013) Handbook of Agricultural Statistics. Achal Prakashan Mandir.
2. Kothari CR (2004) Research Methodology: Methods and Techniques. New Age International (P) Ltd.

1.10 SELF ASSESSMENT QUESTIONS

1. When data are classified according to a single characteristic, it is called:
 - a) Quantitative classification
 - b) Qualitative classification
 - c) Area classification
 - d) Simple classification
2. Classification of data by attributes is called:
 - a) Quantitative classification
 - b) Chronological classification
 - c) Qualitative classification
 - d) Geographical classification
3. Classification is applicable in case of:
 - a) Normal characters
 - b) Quantitative characters
 - c) Qualitative characters
 - d) Both (b) and (c)
4. In classification, the data are arranged according to:
 - a) Similarities
 - b) Differences
 - c) Percentages
 - d) Ratios
5. When an attribute has more than three levels it is called:
 - a) Manifold-division
 - b) Dichotomy
 - c) One-way
 - d) Bivariate
6. The number of classes in a frequency distribution is obtained by dividing the range of variable by the:
 - a) Total frequency
 - b) Class interval
 - c) Mid-point
 - d) Relative frequency

7. The arrangement of data in rows and columns is called:
 - a) Classification
 - b) Tabulation
 - c) Frequency distribution
 - d) Cumulative frequency distribution
8. When the qualitative or quantitative raw data are classified according to one characteristic, the tabulation of different groups is called:
 - a) Dichotomy
 - b) Manifold-division
 - c) Bivariate
 - d) One-way
9. The arrangement of data in rows and columns is called
 - a) Frequency distribution
 - b) Cumulative frequency distribution
 - c) Tabulation
 - d) Classification
10. Questionnaire method is used in collection of:
 - a) Primary Data
 - b) Secondary Data
 - c) Internet Data
 - d) None of these
11. In statistics, collection of related observations is called:
 - a) Data
 - b) Information
 - c) Attribute
 - d) None of these
12. Data gathered through the publication of various journals represent:
 - a) Primary Data
 - b) Secondary Data
 - c) Firsthand Data
 - d) Basic Data

13. The simple classification and manifold classification are types of
- Qualitative classification
 - Quantitative classification
 - Open end classification
 - Tine series classification
14. The complex type of table in which the variables to be studied are subdivided with interrelated characteristics is called as
- Two way table
 - One way table
 - Subparts of table
 - Order level table

Answers Key: 1-d, 2-c, 3-d, 4-a, 5-a, 6-b, 7-b, 8-d, 9-c, 10-a, 11-a, 12-b, 13-a, 14-a

1.11 TERMINAL QUESTIONS

1.11.1 Short answer type questions:

- What is statistics?
- What are data?
- Differentiate between primary and secondary data?
- Name the types of classifications of data?
- What is the meaning of data interpretation?

1.11.2 Long answer type questions:

- Give a detailed account of classifications of data and its types?
- Explain in detail about the tabulation of data and its types?

UNIT-2- DATA PRESENTATION

Contents

- 2.1- Objectives
- 2.2- Introduction
- 2.3- Graphic and non graphic representation
 - 2.3.1- Graphic representation of data
 - 2.3.2- Non graphic representation of data
- 2.4- Results communication
- 2.5- Summary
- 2.6- Glossary
- 2.7- References
- 2.8- Suggested readings
- 2.9- Self assessment Questions
- 2.10- Terminal Questions
 - 2.10.1- Short answer type questions
 - 2.10.2- Long answer type questions

2.1 OBJECTIVES

This unit will give basic knowledge to the students about:

- Data representation
- Graphical and Non-Graphical Data representation
- Result communication

2.2 INTRODUCTION

In the previous unit, we learned how to collect data, classification and tabulation of data. The next step is to represent the data. So what is representation of data? Representation of data is to present the data to the audience so that they can understand and analyze it in a short time. Suppose there is a corporate board meeting going on and somebody is showing the annual work he/she has done through a power point presentation to the senior officials of the company. So, e.g. you have to present the twelve months of sales data in a single slide, how you will do that? Simple answer, if you know a little bit of statistics, through graphs, or chart or bar diagrams. Pictures speak more than words and this is what we are doing here. These graphs etc are placed there to make the huge amount of data easy to understand in short time, and any small point of data should not remain hidden from the policy makers, so that they can make a well informed decision about your performance throughout the year.

Data representation can be broadly categorized into two groups:

1. Graphical representation: Pie Chart, Bar Chart, Line Graphs, Geometrical Diagrams
2. Non graphical representation: Tabular Form, Case Form

2.3 GRAPHIC AND NON GRAPHIC REPRESENTATION

2.3.1 Graphical Representation of the data

Graphical representation is the visual articulation of data using plots and charts. For the ease of analysis the tabulated data can be exhibited in pictorial form through the use of a graph. This is a visual approach to display the data and statistical results. It is very easy to understand the results in graphs than in tables. Graphical representation is a very effective method to represent the data. There are various kinds of graphical representations (e.g. graphs, plots, charts and diagrams) and which are selected on the basis of the nature of the data and the type of statistical results.

Few commonly used graphical representations of data are listed below:

i. Histogram:

A Histogram is a vertical bar chart that shows the distribution of a set of data. It is the most common form of graphical representation of data. It is used to organize and exhibit the data in a

more user-friendly structure. It makes very easy to notice the variations in the data. Suppose we have a data of height of 20 trees so how we will plot the histogram of this data.

To plot a histogram take a graph paper and then put the values of the variable on horizontal axis or X-axis and put the frequencies on the vertical axis or Y-axis. For each class interval a rectangle is drawn with the base equal to the length of the class interval and height according to the frequency of the class interval.

ii. Bar diagram or bar graph or bar chart:

Bar charts are a commonly used and clear way of presenting categorical data or any ungrouped discrete frequency observations. Let us suppose we have the data of mode of transport of a group of students (Table 2.1) and we want to know which is the most and least used mode of transport by them to come to college.

Table 2.1 Mode of transport of a group of students

Student	Mode	Student	Mode	Student	Mode
1	Car	11	Walk	21	Walk
2	Walk	12	Walk	22	Metro
3	Car	13	Metro	23	Car
4	Walk	14	Bus	24	Car
5	Bus	15	Train	25	Car
6	Metro	16	Bike	26	Bus
7	Car	17	Bus	27	Car
8	Bike	18	Bike	28	Walk
9	Walk	19	Bike	29	Car
10	Car	20	Metro	30	Car

So the first thing we will do here is to count the number of times a mode of transport comes (frequency) and make a list of it (Table 2.2), like the one mentioned below. After that, put these frequency values in a graph.

Table 2.2 Frequency of mode of transports

Mode	Frequency
Car	10
Walk	7
Bike	4
Bus	4
Metro	4
Train	1
Total	30

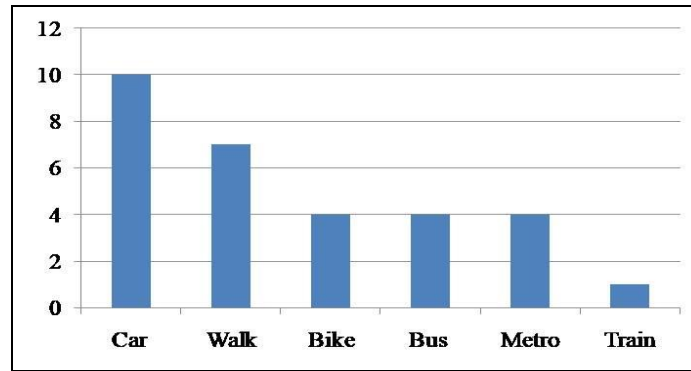


Figure 2.1 Bar graph showing various modes of transport

The horizontal axis of the chart shows the specific categories being compared, and the vertical axis represents a discrete value. In this bar graph (Figure 2.1) we can clearly see that the most popular mode of transport is the car and that the metro, bus and bike are all equally popular. This method provides a simple way of quickly spotting simple patterns of popularity within a discrete data set.

iii. Frequency polygon:

The frequency polygon is very similar to histogram but instead of drawing bars, each class is represented by one point and these points are connected together by straight lines. Frequency polygons are used to plot frequencies of data in different classes and are useful to show patterns and trends within the data. The polygon shown below (Figure 2.2) is based on the data used to draw the above mentioned histogram.

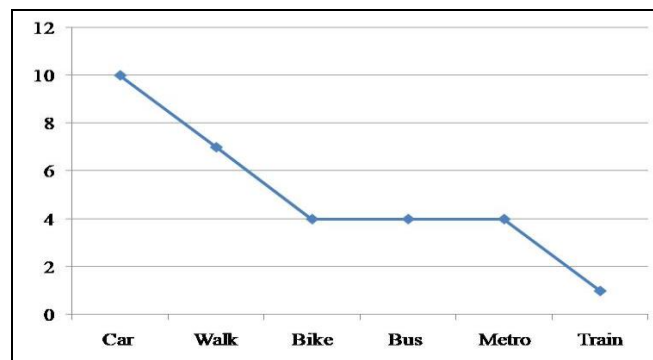


Figure 2.2 Showing the frequency polygon representation of data

iv. Cumulative frequency curve or ogive

The cumulative frequency curve is a method to represent the cumulative frequencies for the classes in a frequency distribution. So what is cumulative frequency, it is the sum of the frequencies accumulated up to the upper boundary of a class in the distribution. It shows how many of values of the data are below certain boundary.

To draw the curve firstly labels the class boundaries on the x (horizontal) and the cumulative frequencies on y (vertical) axes. Plot the cumulative frequency at each upper class boundary with

the height being the corresponding cumulative frequency. Connect the points with segments.

v. Pie chart

Pie chart is just a circle which is divided into segments (Figure 2.3). Each segment represents the of each value. It exhibits data, information and statistics in an easy to read “pie slice” format with varying slice sizes telling how much of one data element exists. It is very helpful to a layman to understand the statistics because anyone can measure the size of the slice of the cake or pie they are receiving whether it is small or large.

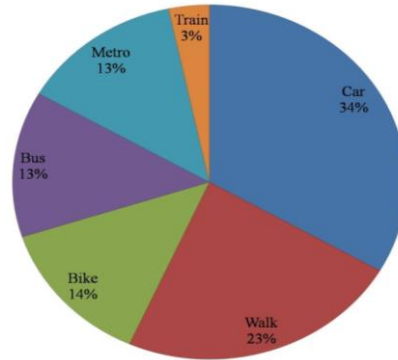


Figure 2.3 Pie-chart showing the graphical representation of the data

vi. Pictogram:

In a pictogram instead of bars or charts symbols or pictures are used. It is used to show huge differences between categories. We just have to decide that what symbol we are using to denote which character and have to mention that. An example is being given here to give you an idea of a pictogram. This pictogram represents the data of consumption of chapattis in a hostel (Figure 2.4). The circle represents the chapatti and each circle represents 100 chapattis.

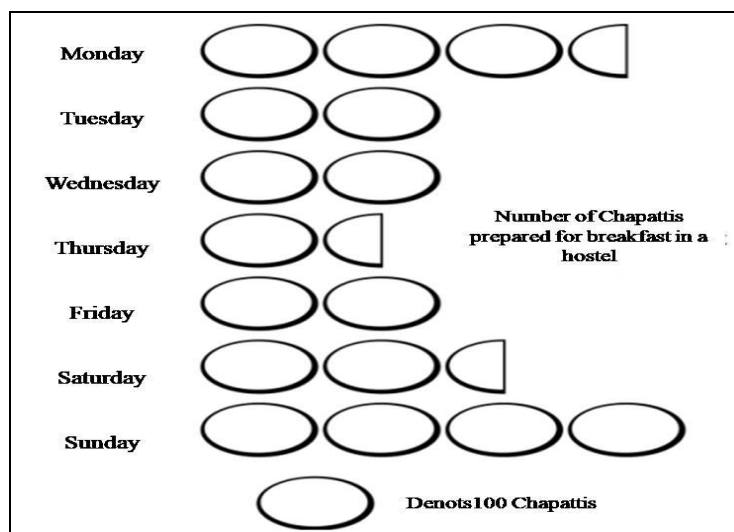


Figure 2.4 Pictogram showing the graphical representation of the data

vii. Stem-leaf diagram

These diagrams are a much convenient and fast method of listing a large group of numbers in order. It is a better way of presenting data than just writing a long list of numbers.

To create the stem-leaf plot the leading values are used to make stem and the trailing values are used in making leaf. For example if the value is ‘23’ then the leading value would be ‘2’ and trailing value would be ‘3’. First make a stem of fixed class intervals like the one made in the graph below regardless of the presence or absence of that class interval value in the data. Now search the value of that particular class interval in the data, if the value is present in the data then put the trailing value in the leaf column. For example look at the first row of the stem column, the value is ‘2’, now look at the first row of data, the value is ‘23’. The value ‘23’ belongs to that row because the leading value of ‘23’ is ‘2’ so the stem value is ‘2’ and leaf value is ‘3’. In case of multiple values of a single class interval present in the data look at the fifth row of the stem column, the value is ‘6’ now search the data. We have five values belonging to this class interval i.e. 62, 62, 63, 65, and 67. Put these values like the graph (Figure 2.5). These diagrams are the only graphical representations that also display all the original data values. One disadvantage of the stem-and-leaf plots is that data must be grouped according to place value. We cannot use different groupings.

Data		Stem	Leaf				
23	71	2	3				
58	71	3					
62	72	4					
62	80	5	8				
63	82	6	2	2	3	5	7
65	82	7	1	1	2		
67	82	8	0	2	2	2	

Figure 2.5 Stem-leaf of the data

viii. Scatter diagram

The scatter diagram is used to show the relationship between two variables and to prove or disprove cause-and-effect relationships. It is used to study the connections between the two sets of data. For example the weight and height of a tree are related: the taller the tree the greater the

weight. One variable is plotted on the horizontal axis and the other is plotted on the vertical axis. The graph below (Figure 2.6) shows the positive correlation between the plant height and weight.

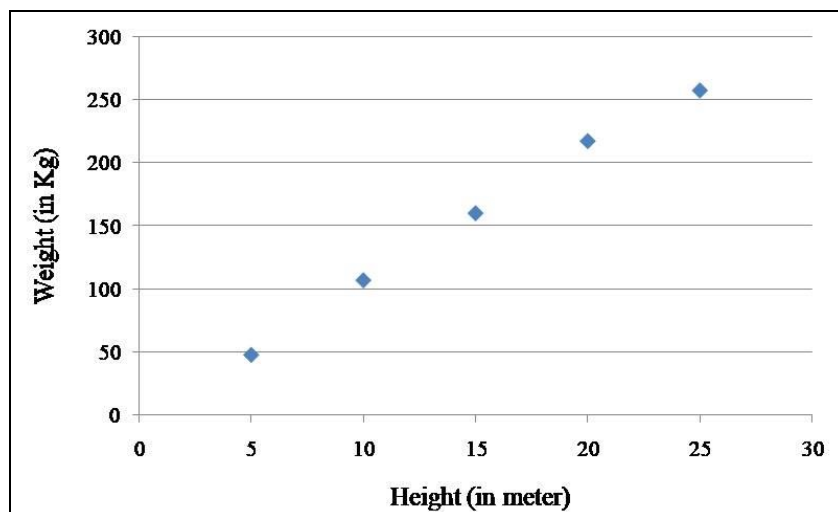


Figure 2.6 Scatter diagram showing the representation of the data

2.3.2 Non-Graphical representation of the data

These are of two types: (a) Tabular Form (b) Case Form

i. Tabular Form: In this method the numerical data are simply put into tables. This is most commonly used non-graphical method of data representation. In this manner we can correlate or measure the two values/variables at a time. To understand, suppose we have data of marks of students of a class (Table 2.3) and we have to arrange data in this manner.

Table 2.3 Marks of students in different subjects

Student	Maximum Marks (100)		
	Botany	Chemistry	Zoology
Student 1	90	50	86
Student 2	65	57	68
Student 3	73	76	94
Student 4	82	79	89
Student 5	63	88	86

ii. Case Form: This is a rarely used method of data representation. In this method, data are represented in the form of paragraphs and follows a rigid protocol to examine limited number of variables. The protocol can be identified according the following cases:

1. Determine all the possible strategies for the present scenario.
2. Determine the optimum suitable strategy.
3. Evaluating the selected strategy in a particular case.

2.4 RESULTS COMMUNICATION

After analyzing all the data we have the results of the study. Now this result is just some words written on a piece of paper until it reaches to a professional community and published as scholarly literature. Broadly there are two ways to communicate the results; conferences and journals. Both these means of communication are peer reviewed. Peer review is the method of evaluation of the authenticity and quality of the work by the experienced professionals of the same field. Blind reviews are considered the best kind of peer reviews because in this type the reviewer is unaware of the researcher's identity. This method is used to minimize biases.

Conference Presentations

This method is generally used as first step towards result communication. It is normally used for the researches which are still in progress. This way of result communication provides the researcher instant feedback from the peers of his/her field of research. In this channel of result communication he/she can have on the spot suggestions regarding their mistakes before the submission of the work as a manuscript in any journal.

Journal Articles

It is the final step of the result communication. Research article published in any scientific peer reviewed journal is the most accepted and reliable source of information. This is so because the submitted research goes through various steps of evaluation of the authenticity and quality of the work. After getting published in any peer reviewed journal the research results are acceptable to everyone and reach a very large readership.

2.5 SUMMARY

In this unit we learned about the various methods of data representation. Data representation is a very important step in any research because the huge chunks of collected data cannot tell anything without the appropriate analysis and presentation. There are two major ways to present data; graphical and non-graphical. Graphical methods include the visual articulation of data using plots and charts. These are of various types namely Histogram, Ogive, Pie-charts, Pictogram etc. The other way of presenting data is non-graphical method. This includes the representation of data in tabular and case forms. The next step in research is the communication of results to the others. The results can be communicated by either presenting them to the live audiences i.e. via conferences or publishing them in the reputed peer reviewed journals. Both data representation and result communication are the key components in any research.

2.6 GLOSSARY

Bar Diagram: A diagram in which the numerical values of variables are represented by the height or length of lines or rectangles of equal width.

Conference: A formal meeting for discussion.

Frequency: The rate at which something occurs over a particular period of time or in a given sample.

Graph: A diagram showing the relation between variable quantities, typically of two variables, each measured along one of a pair of axes at right angles.

Histogram: A diagram consisting of rectangles whose area is proportional to the frequency of a variable and whose width is equal to the class interval.

Journal: A newspaper or magazine that deals with a particular subject or professional activity.

Mode: The value that occurs most frequently in a given set of data.

Ogive: A cumulative frequency graph.

Pictogram: A pictorial symbol for a word or phrase. Pictographs were used as the earliest known form of writing, examples having been discovered in Egypt and Mesopotamia from before 3000 BC.

Pie chart: A type of graph in which a circle is divided into sectors that each represent a proportion of the whole.

Scatter Diagram: A graph in which the values of two variables are plotted along two axes, the pattern of the resulting points revealing any correlation present.

2.7 REFERENCES

1. Chandel SRS (2013) Handbook of Agricultural Statistics. Achal Prakashan Mandir.
2. http://www.mas.ncl.ac.uk->notes_chapter2.pdf
3. Various other internet sources

2.8 SUGGESTED READINGS

3. Chandel SRS (2013) Handbook of Agricultural Statistics. Achal Prakashan Mandir.
4. Kothari CR (2004) Research Methodology: Methods and Techniques. New Age International (P) Ltd.

2.9 SELF ASSESSMENT QUESTIONS

1. The graph of the cumulative frequency distribution is called:
 - a) Histogram
 - b) Frequency polygon
 - c) Pictogram
 - d) Ogive
2. The stem and leaf displaying technique is used to present data in
 - a) Descriptive data analysis

- b) Exploratory data analysis
 - c) Nominal data analysis
 - d) Ordinal data analysis
3. The curve of cumulative frequency is also known as
- a) Ogive
 - b) A-give
 - c) C-give
 - d) B-give
4. If the midpoints of bars on the charts are marked dots are joined by a straight line then this graph is classified as
- a) Class interval polygon
 - b) Paired polygon
 - c) Marked polygon
 - d) Frequency polygon
5. The graphical diagram in which total number of observations are represented in percentages rather than absolute values is classified as
- a) Asymmetrical diagram
 - b) Ungrouped diagram
 - c) Grouped diagram
 - d) Pie diagram
6. Qualitative data can be graphically represented by using a(n)
- a) Histogram
 - b) Frequency polygon
 - c) Ogive
 - d) Bar graph
7. Fifteen percent of the students in Uttarakhand Open University are doing post-graduation in Zoology, 20% in Botany, 35% in Chemistry, and 30% in Physics. The graphical method(s) which can be used to present these data is (are)
- a) A line graph
 - b) Only a bar graph
 - c) Only a pie chart

- d) Both a bar graph and a pie chart
8. The most common graphical presentation of quantitative data is a
- Histogram
 - Bar graph
 - Relative frequency
 - Pie chart
9. A graphical presentation of the relationship between two variables is
- An Ogive
 - A histogram
 - Either an Ogive or a histogram, depending on the type of data
 - A scatter diagram
10. In a scatter diagram, a line that provides an approximation of the relationship between the variables is known as
- Approximation line
 - Trend line
 - Line of zero intercept
 - Line of zero slope

Answers Key: 1-d, 2-b, 3-a, 4-d, 5-d, 6-d, 7-d, 8-a, 9-d, 10-b

2.10 TERMINAL QUESTIONS

2.10.1 Short answer type questions:

- What is data presentation?
- What is a pie chart?
- Differentiate between pictogram and histogram?
- What is the meaning of result communication?
- Explain the stem-leaf method of data presentation?

2.10.2 Long answer type questions:

- Explain the types of graphical data presentation?
- What is the need to represent the data, explain with logical reasoning and appropriate examples?

BLOCK-2- STATISTICAL METHODS

UNIT-3- DESCRIPTIVE STATISTICS

Contents

- 3.1-Objectives
- 3.2-Introduction
- 3.3-Measures of central tendencies
 - 3.3.1-Mean
 - 3.3.2-Median
 - 3.3.3-Mode
 - 3.3.4-Other averages
- 3.4-Measures of dispersion and deviation
 - 3.4.1-Range
 - 3.4.2-Mean deviation
 - 3.4.3-Standard deviation
 - 3.4.4-Standard error
- 3.5-Summary
- 3.6-Glossary
- 3.7-Self Assessment Question
- 3.8-References
- 3.9-Suggested Readings
- 3.10-Terminal Questions

3.1- OBJECTIVE

After reading this unit you will be able to:

1. Understand the importance of biostatistics in the biological expressions.
2. Understand the different measures of central tendency.
3. Understand the different measures of dispersion.
4. Understand how to work out the problems related to measure of central tendency and dispersion for different types of data i.e., ungrouped and grouped data.

3.2- INTRODUCTION

In this unit we are going to study the concepts of biostatistics, first thing came in our mind is what is biostatistics? In general we may define the biostatistics as the application of statistical methods to the solution of biological problems. So second thing came in our mind is what are biological problems, so for this definition basic biological problems are those arising in the basic biological sciences as well as in such applied areas as the health-related sciences and the agricultural sciences.

Before discussing more about biostatistics, we just recall what statistics is? The word Statistics was first time used by “Gattfried Ahenwall”. In general, statistics is the science that deals with the collection, classification, analysis and interpretation of numerical facts or data. According to Webster, “Statistics are the classified facts representing the conditions of the people in a state especially those facts which can be stated in number or in a table of number or in any tubular or classified arrangement” According to Bowley, “Statistics is a numerical statement of facts in any department of enquiry placed in relation to each other. Further, according to Croxton and Cowden, “Statistics may be defined as a science of collection, presentation, analysis and interpretation of numerical data.

The use of statistics can be easily understood with this example, for example, we continuously hear about the UP elections right now, and every now and then you see an opinion poll, what is the statistics by which Yogi Adityanath will beat Akhilesh Yadav in the polls? You have the statistics which is repeatedly taken and you still come up with different companies which hold this polls which come up with different numbers. So, one person might predict that Yogi Adityanath by win by 50% points, another predicts by 35% points. So, this gives us the idea that this process is extremely complicated; it is very easy to come up with a number saying this is the difference that is the difference, but there is a very active science behind this process in order to make sure that the numbers you pop out of the process are accurate.

So, what is biostatistics? So, biostatistics is nothing but application of statistics for the study of living organisms, for human beings, for animals, or for any biological process for that matter. The use of statistics in Biology is known as biostatistics or biometry. So, we can take examples from any field which concerns biology or medicine or bioengineering, biology. For example, we

want to talk about ecology for example, right in forest we want to see by let us say comparing the structure of the different sites how a mixed forest eventually became a pine forest, so on, and so forth and for these kinds of things we have to make exact analysis of specific structural components of a particular component of the forest and their topographic changes, climatic changes, invasion if any in a function of time. In microbiology lab, we want to predict the ability of synthesis drug to study the zone of inhibition to a different microbial strains, so on different bacterial we do the antibacterial assay and then in last came with the observation that how effective the synthesis drug is or how resistance the bacteria is for the drug and this open the aspect of market either to launch the particular drug in the market or not. Last case for example, is in public health you all been aware of the damage that Corona virus is currently producing. So, in the case of corona virus, whenever we go to the airport or any other places and we will see that people coming from different places must undergo a mandatory health checkup to eliminate the possibility that that person is a carrier of for corona virus. These are examples where statistics has been used to make some informed decisions, first to measure, and based on that measure to make come up with some analysis, and then based on analysis to make some predictions as to what should be the corrective step.

If in general we discuss the usefulness of biostatistics, we can say that it helps in presenting large quantity of data in a simple and more classified form which is easily understandable. It gave the methods of comparison of data and makes some predictions or judgments. Further, it finds out the possible relationship between the different variables.

3.3-MEASURES OF CENTRAL TENDENCY

In biology we usually encountered with large set of data's, because whatever our approach is for data collection, we have to collect data at least in replicates so that more reliable conclusion can be drawn. In order to describe the whole mass of unwieldy data a single value is required. Thus, biostatistics provides us the tool to get a single value which can better describe the group of data i.e. central value or an average. Or in other words we can say that, from the collected data the values of the variable tend to concentrate around some central value of observation, that value can be used as the representative value. So this tendency of distribution is known as central tendency. Under measures of central tendency Mean, median and mode are the most popular averages that are studied.

3.3.1-STUDY OF MEAN

• Arithmetic Mean

Arithmetic Mean is the most popular and commonly used measure of central tendency. It is defined as the number obtained by dividing the total values of different items by their number and it's denoted by \bar{X} .

In general if arithmetic mean for ungrouped data or individual observation is, $X_1, X_2, X_3, \dots, X_n$ be 'n' observations for a variable x, the arithmetic mean \bar{x} is given by

$$\bar{X} = \frac{X_1 + X_2 + X_3 + \dots + X_n}{n}$$

To further simplify the writing of a sum, the Greek letter Σ (sigma) is used. The sum $x_1 + x_2 + x_3 + \dots + x_n$ is denoted,

$$\bar{X} = \sum_{i=1}^n X_i / n$$

• **Calculation of arithmetic mean**

1. Series of individual observation
2. Discrete series
3. Continuous series

1. Series of individual observations: The calculation of arithmetic mean for such data which are present in individual series, the calculations is easy. Generally we have to get the total of values and divide this total by number of observation.

Suppose we have five values for stomata's on leaves of a plant is 14 for first leaf, 17 for second leaf, 13, 15, and 19 for third, fourth and fifth leaf respectively.

The arithmetic mean therefore is

$$\frac{14 + 17 + 13 + 15 + 19}{5} = 15.6$$

So the average number of stomata for the plant is 15.6

Symbolically

Leaf's	Stomata's
A	14
B	17
C	13
D	15
E	19
N=5	$\Sigma X = 15.6$

Further, the arithmetic mean for series of individual observation can be calculated by two methods:

- A. Direct method
- B. Assumed mean method or short cut method

A. Direct Method

Example: Calculate the arithmetic mean of the following marks in Hindi obtained by 10 students in a unit test.

Student	A	B	C	D	E	F	G	H	I	J
Marks	15	17	13	16	18	19	14	16	17	18

Work Procedure

Obtained ΣX by adding all the variables and divide the total by number of observation (N) symbolically.

$$\Sigma X = 15 + 17 + 13 + 16 + 18 + 19 + 14 + 16 + 17 + 18$$

$$\Sigma X = 163$$

$$N = 10$$

$$\text{Thus, } \bar{X} = \Sigma X / N,$$

$$\bar{X} = 163 / 10, 16.3$$

The average marks in Hindi is 16.3

B. Assumed mean method or short cut method

Example: Calculate the arithmetic mean of the marks in Hindi obtained by 10 students in a unit test given in illustration by assumed method.

Student	Marks (X)	X-A (d)
A	15	-4
B	17	-2
C	13	-6
D	16	-3
E	18	-1
F	19	0
G	14	-5
H	16	-3
I	17	-2
J	18	-1
N=10		-27

Work procedure:

In this example, first we have to assume a mean, suppose assume mean=19. Calculate the deviation from assumed mean (X-A)=d

Get the total, of deviation from data using the following formula,

$$\text{Mean } \bar{X} = A + \Sigma d / N, \text{ where "A" is assumed mean and "d" is deviation}$$

$$= 19 + (-27/10)$$

$$= 19 - 2.7$$

$$= 16.3 \text{ marks}$$

Thus the average marks in Hindi= 16.3.

Thus from results of both methods we got the same result (16.3 marks).

B. Discrete series:

In case of discrete series, frequency against each variable (observation) is multiplied by the value of the observation. The values, so obtained, are summed up and divided by the total number of frequencies. Symbolically,

$$\bar{X} = \frac{\sum fx}{\sum f}$$

Where, $\sum fx$ = sum of the product of variables and frequencies.

$\sum f$ = sum of frequencies.

Example: The number of seeds produced by 120 plants in garden given in table. Calculate the arithmetic mean.

Seed's (X)	Number of Plants (f)	fX
200	4	800
190	12	2280
180	15	2700
170	37	6290
160	22	3520
150	6	900
140	10	1400
N=106		$\sum fX = 17890$

Work procedure:

Multiply the frequency with the variable X and get the sum of the product ($\sum fx$). Divide $\sum fx$ with the total number of observation $\sum f$ or N

Number of class: 200, 190, 180, 170, 160, 150, 140.

Frequency (f) = 4+12+15+37+22+6+10

$\sum fx$ or N=106

$fx = 800+2280+2700+6290+3520+900+1400$

$\sum fx = 17890$

$$\bar{X} = \frac{\sum fx}{\sum f}$$

So $\bar{X} = 17890/106$

$\bar{X} = 168.77$

Short cut method or assumed mean method: We can use this method for discrete series also, the formula to be used for this method is

Mean $\bar{X} = A + \frac{\sum fd}{N}$

Here, A= assume mean

N= Number of observation

d= deviation of variable taken from assumed mean

Σfd = Sum of the product of frequencies and their respective deviations.

Number of Plants			
Seed's (X)	(f)	X-A = d	fd
200	4	30	120
190	12	20	240
180	15	10	150
170	37	0	0
160	22	-10	-220
150	6	-20	-120
140	10	-30	-300
106			-130

Work procedure:

In this example, first we have to assume a mean, suppose assume mean=170. Calculate the deviation from assumed mean (X-A)=d

Get the total, of deviation from data using the following formula,

Mean \bar{X} = A + $\Sigma fd/N$, where “A” is assumed mean and “d” is deviation

Mean \bar{X} = A + $\Sigma fd/N$

$$= 170 + \frac{-130}{106}$$

$$= 170 + (-1.23)$$

$$= 168.77$$

The average number of seeds for the plant is 168.77 and again the answer we got by both method is same.

C. Continuous series: In this approach of calculating arithmetic mean, class intervals are given. The method of calculating arithmetic mean in case of continuous series is same as that of a discrete series. The only difference is that the mid-points of various class intervals are obtained. We have already known that class intervals may be exclusive or inclusive or of unequal size. The following equation may be used to derived or get the mid-points.

$$\text{Mid-point (m)} = \frac{l_1 + l_2}{2}$$

Here, l_1 = lower limit and l_2 = upper limit.

After obtaining the midpoint we calculate arithmetic mean as we calculated in case of discrete series.

Example: Find out the mean of the following distribution in a class

Marks	4-8	8-12	12-16	16-20
Student	4	8	6	3

Work procedure:

In first step: obtain mid-point (m) of the classes by using following formula

$$\text{Mid-point (m)} = \frac{11+12}{2}$$

In second step: Multiply the frequency (f) with mid-point (m) and get the product (fm)

In third step: Divide the Σfm by the total number of observation (N), i.e., **Mean $\bar{X} = \Sigma fm/N$**

Marks (X)	Number of Student (f)	Mid- point	fm
4-8	4	6	24
8-12	8	10	80
12-16	6	14	84
16-20	3	18	54
N = 21		$\Sigma fm = 242$	

$$\text{Mean } \bar{X} = \Sigma fm/N$$

$$= 242/21$$

$$= 11.52.$$

Thus, the mean mark of the students in class among 21 students was 11.52.

Example: calculate the mean of the following distribution in an experimental garden for fruit production.

Number of fruits	100-120	120-140	140-160	160-180	180-200	200-220
Number of plants	10	20	25	30	15	5

Work procedure:

In first step: obtain mid-point (m) of the classes by using following formula

$$\text{Mid-point (m)} = \frac{11+12}{2}$$

In second step: Multiply the frequency (f) with mid-point (m) and get the product (fm)

In third step: Divide the Σfm by the total number of observation (N), i.e., **Mean $\bar{X} = \Sigma fm/N$**

Number of fruits (X)	Number of plants (f)	Mid-Point (m)	fm
100-120	10	110	1100
120-140	20	130	2600
140-160	25	150	3750
160-180	30	170	5100
180-200	15	190	2850
200-220	5	210	1050
N = 105		$\Sigma fm = 16450$	

$$\text{Mean } \bar{X} = \Sigma fm/N$$

= 16450/105

= 156.67

Thus, the mean number of fruit a tree’s have in the garden is 156.67.

Assumed or Short-cut method

Example: Find out the mean of the following distribution in a class

Marks	4-8	8-12	12-16	16-20
Student	4	8	6	3

Work procedure:

In first step: obtain mid-point (m) of the classes by using following formula

$$\text{Mid-point (m)} = \frac{l_1 + l_2}{2}$$

In second step: Decide an assumed mean from the data A=10.

In third step: calculate the deviation from assumed mean $m - A = d$, and then multiply the deviation (d) with mid-point (m) and get the product (fd)

In fourth step Divide the Σfd by the total number of observation (N), i.e., **Mean $\bar{X} = A + \Sigma fd/N$**

Marks (X)	Student (f)	Mid-point (m)	m-A = d	Fd
4-8	4	6	0	0
8-12	8	10	4	32
12-16	6	14	8	48
16-20	3	18	12	36
N = 21				$\Sigma fd = 116$

$$\begin{aligned} \text{Mean } \bar{X} &= A + \Sigma fd/N \\ &= 6 + 116/21 \\ &= 6 + 5.52 \\ &= 11.52. \end{aligned}$$

Thus, the average mark in a class among 21 students is 11.52.

Some other types of mean’s

1. Geometric Mean:

Geometric mean of N variate value is the Nth root of their product. In algebra, geometric mean is used or calculated in case of geometric progression. Further, like arithmetic mean it also depends on all observations.

Symbolically

$$G.M = (x_1 \cdot x_2 \cdot \dots \cdot x_n)^{1/N} \text{ or } (x_1 f_1 \cdot x_2 f_2 \cdot \dots \cdot x_n f_n)^{1/N}$$

2. Harmonic Mean:

Harmonic Mean is defined as reciprocal of arithmetic mean the reciprocal of individual observations. In algebra, harmonic mean is found out in the case of harmonic progression only, but in statistics this mean is suitable measure of central tendency (data pertains to time, speed and rates). This mean is rigidly defined and calculation is based on all the observations.

Symbolically harmonic Mean is

$$H = \frac{N}{f_1X_1 + f_2X_2 + \dots + f_kX_k}$$

3. Weighted Arithmetic Mean:

In this type of arithmetic mean weights to various items according to their importance has been assigned when we have to calculate the arithmetic mean. For example, we have two commodities, apple and potatoes. We are interested to find the average price of both commodities. Thus for the calculation of arithmetic mean we consider the condition in this way

i.e., we consider P1 and P2 to our commodities and use the formula $\frac{P_1 + P_2}{2}$. In other words, the simple arithmetic mean gives equal weightage to all the items but usually all items don't have equal importance in the series, some may be more important and some may be less important. But, if we want to give importance to the rise in price of potatoes (P2), to do this, we have to use as 'weights' the share of apples in the budget of the consumer (W1) and the share of potatoes in the budget (W2).

Now the arithmetic mean weighted by the shares in the budget given by following formula,

$$\frac{W_1P_1 + W_2P_2}{W_1 + W_2}$$

In general the weighted arithmetic mean is given by the following formula

$$W_1P_1 \frac{W_1P_1 + W_2P_2 + W_3P_3 + \dots + W_nP_n}{W_1 + W_2 + W_3 + \dots + W_n} = \frac{\sum WX}{\sum W}$$

3.3.2- STUDY OF MEDIAN

Median is the positional value that divide the given observations or data into equal half, one part of the observation includes all value that are greater than the median and other part includes the values that are lesser than the observations. Further, by definition, median may defined as the middle-most or the central value of the variable in a set of observations, when the observations, are arranged either in ascending or in descending order of their magnitudes and denoted by capital "M". In comparison to the arithmetic mean median is position average while the arithmetic mean is mathematical average because suppose we have a series of 7 observations i.e., 20,30,40,50,60,70 and 80, so from the observations 50 will be the median and in second condition if we change the observation of the series and we have new observation such as 10,30,40,50,60,70 and 80, but for this series the median again comes to be 50. But, in case of arithmetic mean, change in one item, changes the average value, where as in case of median if items other than central value change the median will not change. Thus, this further confirms that the median is positional value.

Calculation of Median: As we encounter different observation or data's so we have different working formulas for calculation of median.

A. When the data is ungrouped data or for Simple series:

Procedure:

1) Arrange the “n” (number of values) values of the variable in either ascending or descending order of magnitudes.

2) When “n” is odd or the number of observations is odd, the $\frac{n+1}{2}^{th}$ value is the median

Thus, $M = \frac{n+1}{2}^{th}$ term

3) When “n” is even, then there are two value in the middle, so the median can be estimated by finding arithmetic mean of middle two values i.e. adding two values in the middle and dividing by two.

Thus, ‘n’ is even. In this case there are two middle terms $\frac{n}{2}^{th}$ and $(\frac{n}{2} + 1)^{th}$.

So median will be calculate by using this formula

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

Example 1: The average weight of 9 students is: 45, 48, 53, 46, 54, 59, 42, 58, 41, find the median.

Work procedure:

Arrange the observation is ascending order:

41, 42, 45, 46, 48, 53, 54, 58, 59

$$\text{Median} = M = \left(\frac{n}{2} + 1\right)^{th}$$

$$= 9+1/2$$

$$= 5^{th} \text{ value}$$

Now the 5th value in the table is 48

Therefore the media = 48.

Example 2: The average mark of the students in the class test is, 12, 14, 17, 19, 15, 14, 13, 18, 16, 17, 11, 18. Find the median.

Work procedure:

Arrange the data in ascending order

11,12,13,14,14,15,16,17,17,18,18,19

The number of items in this data n = 12, which is even.

Therefore, Median = $M = \frac{n}{2} th$ and $\left(\frac{n}{2} + 1\right) th$

$$M = 12/2 \text{ and } \frac{12}{2} + 1$$

= 6th and 7th terms

Thus, we have to calculate the mean of 6th and 7th term

$$= \frac{15+16}{2}$$

$$= 31/2 = 15.5$$

Median = 15.5.

B. When the data is grouped

1. When the data is in discrete series: In case of discrete series the position of median i.e. $(N+1)/2$ th item can be located through cumulative frequency.

Example: Find the median for the following data obtained for monthly salaries of peoples

No. of Persons	3	6	7	15	11	9	5	4	11	2
Income (in Rs.)	1300	700	1900	2100	2300	2500	6000	4800	3500	4000

Work procedure:

Arrange the data in either ascending or descending order of magnitude.

Calculate the cumulative frequency.

A table is prepared showing the corresponding frequencies and cumulative frequencies.

Now median is calculated by the following formula

$$M = \left(\frac{n+1}{2}\right) th$$

Where $N = \sum f$

Arrange the monthly salary in table

Income (in Rs.)	No. of Persons	cumulative frequency
1300	3	3
700	6	9
1900	7	16
2100	15	31
2300	11	42
2500	9	51
6000	5	56
4800	4	60
3500	11	71
4000	2	73

$$N = \sum f = 73$$

Here $n = 73$

Thus, Median = $M = \left(\frac{n+1}{2}\right)^{th}$ value

$$= \frac{73+1}{2} \text{ th value}$$

$$= 37\text{th value}$$

The above results give us a value 37 and table shows that all items from 31 to 42 have their values 2300.

Since 37th value lies in this interval so its value is 2300

Hence, the Median for given data is 2300.

2. When the series is continuous: In case of continuous series you have to locate the median class where $N/2$ th item [not $(N+1)/2$ th item] lies. The median can then be obtained by using the following formula.

$$M = L + \frac{\frac{N}{2} - cf}{fm} \times i$$

Where, L = lower limit of class in which the median lies.

N = Total no of frequencies, $n = \Sigma f$.

fm = frequency of the class in which the median lies.

C = cumulative frequency of the class proceeding to the median class.

i = width of the class interval of the class in which the median lies.

Example: The following table is given the marks obtained by students in a Pteridophytes class. Find the median.

Marks	10-14	15-19	20-24	25-29	30-34	35-39	40-44	45-49
Number of students	6	12	4	15	10	3	7	2

Work procedure:

- 1) Here data is given in the form of a frequency table with class interval.
- 2) Cumulative frequencies are found out for each value.
- 3) Determine the class where median lies. Median class is that class where Nth item lies
- 4) After ascertaining the Median Class, the following formula can be used to find out exact value of median

$$M = L + \frac{\frac{N}{2} - cf}{fm} \times i$$

Marks	Number of students	Cumulative frequency
10-14	6	6
15-19	12	18

20-24	4	22
25-29	15	37
30-34	10	47
35-39	3	50
40-44	7	57
45-49	3	60
N= Σf = 60		

Here, n=60, so n/2= 30

Thus, median class is 25-29

Lower limit of the median class is 24.5.

Cumulative frequency of the class preceding the median class = 22.

fm=median class = 15.

I= class interval of the median class= 5.

Thus, putting all these values in given formula

$$M = L + \frac{\frac{N}{2} - cf}{fm} \times i$$

$$\begin{aligned} M &= 24.5 + \frac{30-22}{15} \times 5 \\ &= 24.5+2.66 \\ &= 27.16 \end{aligned}$$

So the calculated Median for the given data is M= 27.16.

Thus, this mean half of the class have score equal or less than 27 marks in subject and half of the class have score higher than 27 marks in the subject.

Example: Find the median of the following data, the table is given the marks obtained by students in a Gymnosperm class.

Marks	20-25	25-30	30-35	35-40	40-45	45-50	50-55	55-60
Number of students	14	26	23	33	35	13	8	15

Work procedure:

- 1) Here data is given in the form of a frequency table with class interval.
- 2) Cumulative frequencies are found out for each value.
- 3) Determine the class where median lies. Median class is that class where Nth item lies
- 4) After ascertaining the Median Class, the following formula can be used to find out exact value of median

$$M = L + \frac{\frac{N}{2} - cf}{fm} \times i$$

Marks	Number of students	Cumulative frequency
20-25	14	14
25-30	26	40
30-35	23	63
35-40	33	96
40-45	35	131
45-50	13	144
50-55	8	152
55-60	16	168

Here, $n=168$, so $n/2= 84$

Thus, median class is 35-40

Lower limit of the median class is 35.

Cumulative frequency of the class preceding the median class = 63.

f_m = median class = 33.

i = class interval of the median class = 5.

Thus, putting all these values in given formula

$$M = L + \frac{\frac{N}{2} - cf}{f_m} \times i$$

$$M = 35 + \frac{96 - 63}{33} \times 5$$

$$= 35 + 5$$

$$M = 40.$$

So the calculated Median for the given data is $M= 40$.

Thus, this mean half of the class have score equal or less than 40 marks in subject and half of the class have score higher than 40 marks in the subject.

Advantages (merits) of median

- It is easily understood, easy to locate without any difficulty although it is not as popular as arithmetic mean.
- The value is not affected by magnitude of extreme deviations.
- Median can also be determined graphically to ogives.
- It is very useful in open ended classes or where the extreme classes are ill defined, like "less than 20" or "more than 20".
- Median is very good in case of qualitative data's or when the items are not susceptible to measurements in definite units.
- Median is unaffected by abnormal value.
- Median always remain the same whatsoever methods of computation be applied.

Disadvantages (demerits) of median

- It fails with data having great variation.
- For its calculation data must be arranged either in ascending or descending order of magnitude which proves sometime very tedious.
- It gives same weights to all observations and ignores the extreme values.
- It is not suitable for further algebraic treatment.

3.3.3- STUDY OF MODE

Mode is generally defined as the most common or value in a series which appears most frequently. The word “mode” has been derived from the French word “la Mode” which signifies the most fashionable values of a distribution, because it is repeated the highest number of times in the series. For example in a series 9, 8, 6, 9, 5, 6, 3, 9, 2, 1, 7, 9, 9, 4, 1, 9, 9, 8 we noticed that 9 comes seven times so the mode for the series is 9 or in other words, mode represents the maximum demanding figure. In other words, mode represents that value which is most frequent or typical or predominant. According to Croxton and Cowden, "The mode is that value or point around which the items to be most heavily concentrated." Further, according to Kenny and Reepura “the value of the variable which occurs most frequently in a distribution is called mode”. Mode is some time also known as Norm and denoted by M_o .

Calculation of mode

A. When data is simple or mode in simple series. In case of simple series, the value which appears in maximum number of times is mode of that series. It can be determined by the inspection method, by counting the number of times, the various values repeat themselves and the value which occurs for maximum number of times is the modal value.

Example: Marks of the 10 students are recorded as 26, 22, 37, 30, 45, 40, 37, 30, 37, 26, 30, 37, 45, 37, 22. Calculate the mode of the series.

Work procedure:

- Arrange the data in series and locate the value which occurs maximum number of times.
- Write the number of times the value located in the data against each value.
- The value comes maximum will be the mode of the series.

Marks Obtained	Number of time repeat
22	2
26	2
30	3
37	5
40	1
45	2
$\Sigma N = 15$	

The number 37 occurs for the largest number of times. So 37 is the mode of the above series.

B. Group data

- Discrete series

2. Continuous series

1. Discrete series: For discrete series mode can be calculated by using two approaches i.e., by inspection method (when data is unimodal) or by grouping method (when data is bi or multimodal).

By inspection Method

Example: Protein content of 15 milk samples was recorded as 18, 15, 8, 12, 6, 20, 9, 14, 12, 7, 12, 20, 19, 18, 9.

Work procedure: Arrange the data in ascending order and then convert the observation into frequency distribution table.

Protein %	6	7	8	9	12	14	15	18	19	20
Frequency	1	1	1	2	3	1	1	2	1	2

During the inspection of the above table we find that 12 repeated three times, which is maximum from the given data. Therefore, the mode of the given observation is 12.

Grouping Method: When the discrete series is bimodal or multimodal then grouping method is used to obtain mode for the series.

Example: In an ecology survey data regarding girth of trees has been collected, calculate the mode from the following frequency distribution.

Girth	110	115	120	125	130	135	140	145	150	155	160	165	170	175
frequency	3	6	9	10	13	15	16	16	12	14	10	8	5	4

Work procedure

- 1) Arrange the value in ascending or descending orders.
- 2) Draw a table and arranged data in tabular form.
- 3) In column I put the values of variables
- 4) In column II frequency to be write against their values.
- 5) Sum of 2-2 frequency to be written in IIIrd column.
- 6) Again write sum of 2-2 frequencies in column IVth, ignoring the first frequency.
- 7) Further, repeat the process, but this time sum of 3-3 frequencies have to take and write in the column Vth.
- 8) Again write sum of 3-3 frequencies in column VIth, ignoring the first and second frequency or in other words first two frequencies.
- 9) Repeat according to the need of the data for more grouping i.e., 4-4, 5-5- and 6-6.

10) Detection of maximum frequency done in each column of group and variable noted in analysis table and variable with maximum repetition consider as mode.

Girth	Frequencies					
	Individual	Grouping by twos		Grouping by threes		
	I	II	III	IV	V	VI
110	3	} 9				
115	6					
120	9	} 19	} 15	} 18	} 25	
125	10					
130	13	} 28	} 23	} 38		} 32
135	15					
140	16	} 32	} 31		} 44	} 47
145	16					
150	12	} 26	} 28	} 44	} 42	
155	14					
160	10	} 18	} 24	} 32		} 36
165	8					
170	5	} 9	} 13		} 23	} 17
175	4					

Column	Girth of tree with maximum frequency
I	140, 145
II	140, 145
III	135, 140
IV	140, 145, 150
V	130, 135, 140
VI	135, 140, 145

So from the above results, we have reported that 140 occurs maximum number of time i.e., 6 times.

Therefore mode is 140.

2. Mode of a Continues Series

In case, where we have data in continuous series we have to find out the class in which the mode is situated and the class in which mode is situated in known as modal class. After determining the modal class we calculate the mode by using the following formula.

$$\text{Mode} = L_1 + \left(\frac{D_1}{D_1 + D_2} \right) \times i$$

Where L_1 = Lower limit of modal class.

D_1 = difference between the frequency of the modal class and the frequency of the class preceding the modal class.

D_2 = difference between the frequency of the modal class and the frequency of the class succeeding the modal class.

i = Class interval or width of the class

Or some time, another expression also used for computing the mode is

$$\text{Mode } (M_o) = L_1 + \left(\frac{f_m - f_1}{2f_m - (f_1 + f_2)} \right) \times i$$

Where L_1 = Lower limit of modal class

f_m = Frequency of modal class or the maximum frequency.

f_1 = Frequency of pre-modal class.

f_2 = Frequency of post modal class.

i = Class interval or class width of the class.

Example: Find the mode for the frequency distribution.

Marks	30-40	40-50	50-60	60-70	70-80	80-90	90-100
Number of students	6	9	8	12	17	7	1

Step 1: By inspection we can say, the modal class of above series has 70-80 because this series has highest frequency.

Step 2: Calculate the mode using formula

$$\text{Formula 1, Mode} = L_1 + \left(\frac{D_1}{D_1 + D_2} \right) \times i$$

Where, $D_1 = 17 - 12 = 5$.

$$D_2 = 17 - 7 = 10.$$

$$L_1 = 70$$

$$i = 10$$

$$\text{So according to formula, } 70 + \frac{5}{5+10} \times 10$$

$$= 70 + 3.33$$

So, the calculated mode for the given data is $M_o = 73.33$.

Marks	No. of students
30—40	6
40—50	9
50—60	8
60—70	12
70—80	17
80—90	7
90—100	1

$$\text{Formula 2, } M_o = L_1 + \left(\frac{f_m - f_1}{2f_m - (f_1 + f_2)} \right) \times i$$

Here, L_1 is 70.

$$F_m = 17$$

$$F_1 = 12$$

$$F_2 = 7$$

$$= 70 + \left(\frac{17-12}{2 \times 17 - (12+7)} \right) \times 10$$

$$= 70 + \left(\frac{5}{34-19} \right) \times 10$$

$$= 70 + \frac{5}{15} \times 10$$

$$= 70 + 3.33$$

$$= 73.33$$

So, the calculated mode for the given data is $M_o = 73.33$

Thus, results for both formula's remains same.

Example: Example: Find the mode for the frequency distribution.

Marks	Less than 80	Less than 75	Less than 70	Less than 65	Less than 60	Less than 55	Less than 50
Cumulative frequency	97	95	80	60	30	12	4

Step 1: Arrange the data and calculate the frequency

Marks group	Cumulative frequency	frequency
75-80	97	$97-95 = 2$
70-75	95	$95-80 = 15$
65-70	80	$80-60 = 20$
60-65	60	$60-26 = 34$
55-60	26	$26-9 = 17$
50-55	9	$9-4 = 5$
45-50	4	4

By inspection we can say, the modal class of above series has 60-65 because this series has highest frequency.

Step 2: Calculate the mode using formula

$$\text{Formula 1, Mode} = L_1 + \left(\frac{D_1}{D_1 + D_2} \right) \times i$$

$$\text{Where, } D_1 = 34 - 20 = 14.$$

$$D_2 = 34 - 17 = 17.$$

$$L_1 = 60$$

$$i = 5$$

$$\text{So according to formula, } 60 + \frac{14}{14+17} \times 10$$

$$= 60 + 4.51$$

So, the calculated mode for the given data is $M_o = 64.51$.

In some special case or for moderately skewed or asymmetrical frequency distribution, mode can be calculated by Karl Pearson's empirical formula

$$\text{Mean} - \text{Mode} = 3 (\text{Mean} - \text{Median})$$

$$\text{Mode} = 3 \text{ Media} - 2 \text{ Mean}$$

Merits of mode

- 1) It is easily calculated without involving much mathematical calculation and for simple and discrete series mode can be obtain merely by the inspection method.
- 2) It is not affected by the values of extreme items.
- 3) Mode can be determined in open-ended class without ascertaining the class limits.
- 4) It can be used to describe qualitative phenomenon and can also determine graphically.
- 5) This is the value whose expectation is the greatest in whole of the statistical series.

Disadvantage

- 1) In some case, there may be more than one mode or there is no single mode.
- 2) Mode does not consider all observation of the series.
- 3) Mode is not rigidly defined and its use is very limited.
- 4) Mode is unsuitable for algebraically treatments.
- 5) In multimodal series mode does not prove to be a good representative.

3.3.4- STUDY OF OTHER AVERAGES

Besides, the Mean, Median and Mode we have some other important averages also such as: Quartiles, Deciles and Percentiles.

1. Quartiles: Quartiles are the measures which divide the data into four equal parts; each portion contains equal number of observations. There are three quartiles which are abbreviated as Q_1 , Q_2 and Q_3 . The first Quartile or lower quartile has twenty five percent of the values or items of the distribution below it and seventy five percent of the items are greater than it. The second Quartile coincides with median and fifty percent of items below it and fifty percent of the observations above it. The third Quartile or the upper Quartile has seventy five percent of the items of the distribution below it and twenty five percent of the items above it. Thus, Q_1 and Q_3 denote the two limits within which central fifty percent of the data lies.

Formula's for quartile calculations

$$Q_1 = \text{Size of } \left(\frac{n+1}{4}\right)^{\text{th}} \text{ items for ungrouped.}$$

$$Q_2 = \text{Size of } \left(\frac{N}{4}\right)^{\text{th}} \text{ items for grouped}$$

$$Q_i = L_1 \left(\frac{\left(\frac{n}{4}\right) - c.f.}{f} \times i \right)$$

where L_1 = the lower limit of the class in which particular quartile lies

c.f.= cumulative frequency of class preceding to the particular quartile class

f = frequency of the particular quartile class

i = size of the class interval in the quartile class.

2. Deciles: Deciles is one more measure, in this deciles divides the series or data or items in ten equal parts. There are nine deciles namely D_1 to D_9 . The D_5 is coincides with median or their value consider as median. Further, the values that lie between any two deciles are ten per cent.

Formula's used

D_1 = Size of $\left(\frac{n+1}{4}\right)^{\text{th}}$ items for ungrouped.

D_2 = Size of $\left(\frac{N}{4}\right)^{\text{th}}$ items for grouped

$$Q_i = L_1 \left(\frac{\left(\frac{N}{4}\right) - c.f.}{f} \times i \right)$$

3. Percentiles: In percentiles, the percentiles divides the series in hundred equal parts and there are ninety nine percentiles namely P_1 to P_{99} . The value for each percentile or any percentile is one percent and P_{50} coincides with median.

Formula's used

P_1 = Size of $\left(\frac{n+1}{4}\right)^{\text{th}}$ items for ungrouped.

P_2 = Size of $\left(\frac{N}{4}\right)^{\text{th}}$ items for grouped

$$P_i = L_1 \left(\frac{\left(\frac{N}{4}\right) - c.f.}{f} \times i \right)$$

3.4-MEASURES OF DISPERSION AND DEVIATION

Dispersion is commonly used term, use to explain scatterness, deviation, spread, fluctuation and variability of collected data. In general, averages are representatives of a frequency distribution in a data. But they fail to give a complete picture of the distribution. In other words, measure of central tendency alone is not sufficient to describe a frequency distribution we must have a measure of scatterness or variability of observations. Thus, an average will be more meaningful and more accurate if it is studied in light of measures of dispersion or variation.

Various measure of dispersion:

- i. Range
- ii. semi-inter quartile range
- iii. Mean deviation
- iv. Standard deviation
- v. Variance

Characteristics of a good measure of dispersion

An ideal measure of dispersion is expected to possess the following properties

1. It should be rigidly defined and based on all the observation or items.
2. It should not be affected by extreme items and should be simple to understand and easy to calculate
3. It should lend itself for algebraic manipulation.

3.4.1- STUDY OF RANGE

Range: Range is the simplest measure of dispersion. It is the difference between the highest and the lowest terms of a series of observations. In general, variability being a characteristic of statistical data's either of biological origin or other. Because not a single experiment or sampling without repeats is to be considered as an indicator of normality. For example in a biological characters such as height, weight, hemoglobin, etc. is worked out after measuring the characteristic in large number of healthy persons of the same age and of same area on a globe. A range defines the normal limits of a biological characteristic. Some ranges are given below:

Characteristic	Range
Bilirubin	0-0.6 mg/dL
Glucose (fasting) (plasma or serum)	70-110 mg/dL
Creatinine (serum)	0.6-1.2 mg/dL
Leukocytes	3500-12,000/mm ³
Uric acid (serum)	2.0-7.0 mg/dL
Erythrocytes (RBCs) – Males	4.6-6.2 x 10 ¹² /L 4.6-6.2 million/mm ³
Cholesterol (serum)	< 200 mg/dL
Systolic blood pressure	100–140 mmHg
Diastolic blood pressure	80–90 mmHg

$$\text{Range} = X_H - X_L$$

X_H = Highest value of the observations.

X_L = Lowest value of the observations.

In practice, range does not tell us actual condition of a person, for example some time the leukocytes count of a person may range above 12000 but the person is normal, but some time within the range person may have some symptoms and which require other pathological single sign for observation which describe variability more accurately.

Merits of range

Range is simplest measure of dispersion and it limits the value in upper and lower most value, in other words its rough measure of the dispersal. Further, from data it's easy to calculate and understand.

Demerits of range

It gives rough idea or rough answer and is not based on all observations. Being a rough measure of dispersal, ranges do not provide the authenticity towards the data. Further, it is affected by

change in the sample, in other words is not a satisfactory measure as it is based only on two extreme values.

3.4.2- STUDY OF MEAN DEVIATION

Mean deviation of series is the arithmetic average of the deviations of various items from a measure of central tendency. Although, mathematically, mean deviation is not a logical measure of dispersion more this method is not valid for algebraic expressions.

Calculation of mean deviation:

Like other measures, mean deviation is also calculated for all three types of data

- A. Series of individual observation.
- B. Discrete series.
- C. Continuous series.

A. Series of individual observation:

Example: In a gymnosperms exam the marks obtained by the ten students is given, calculate mean deviation and its coefficient from the data.

Marks	62	68	66	79	87	75	60	89	77	83
-------	----	----	----	----	----	----	----	----	----	----

Work procedure

Arrange the data accordingly into the table and then calculate its arithmetic mean.

Calculate the deviation from mean ignoring signs.

Calculate the mean deviation using formula, $M.D. = \frac{\sum |D|}{N}$

Then calculate the coefficient of mean deviation using the following formula,

Coefficient of M.D = $\frac{M.D}{Mean}$

Marks	Deviation from Mean (ignoring signs) D
62	12.6
68	6.6
66	8.6
79	4.4
87	12.4
75	0.4
60	14.6
89	14.4
77	2.4
83	8.4
$\Sigma X = 746$	$\Sigma D = 84.8$

Step I: Calculate the arithmetic mean

Here we have,

$\Sigma X = 746$ and $N = 10$

Arithmetic mean = $\bar{X} = \Sigma X/N$

$\bar{X} = 746/10$

$= 74.6.$

Step II: Calculate the deviation, ignoring signs

Calculate deviation is $\Sigma |D| = 84.8$

Step III: calculate mean deviation

M.D. = $\frac{\Sigma |D|}{N}$

M.D = $84.8/10, = 8.48$

Thus, M.D is 8.48.

Step IV: Calculate coefficient of Mean deviation

Coefficient of M.D = $\frac{M.D}{Mean}$

$= 8.48/74.6$

$= 0.113$

In conclusion we have, Arithmetic mean = 74.6; Mean deviation = 8.48 and Coefficient of mean deviation from the data = 0.113.

B. Discrete series

Example: In a given data of road accidents, find out the mean deviation and also calculate its coefficient.

No of accidents	0	1	2	3	4	5	6	7	8	9	10	11
Person involved	12	8	15	21	6	15	14	11	2	1	0	2

Work procedure

Arrange the data accordingly into the table and then calculate its cumulative frequency.

Calculate the median using $\frac{N+1}{2}$

Calculate the deviation from median ignoring signs.

Calculate the mean deviation using formula, M.D. = $\frac{\Sigma f |D|}{N}$

Then calculate the coefficient of mean deviation using the following formula,

Coefficient of M.D = $\frac{M.D}{Mean}$

Number of accidents	Person involved	Cumulative frequency c.f.	D	f D
0	12	12	3	36
1	8	20	2	16
2	15	35	1	15
3	21	56	0	0
4	6	62	1	6

5	15	77	2	30
6	14	91	3	42
7	11	102	4	44
8	2	104	5	10
9	1	105	6	6
10	0	105	7	0
11	2	107	8	16
N= 107			$\Sigma f D = 221$	

Step I: Calculate cumulative frequency

Step II: Calculate median using $\frac{N+1}{2}$

$$= 107+1/2$$

= 54th term or item, since for the 54th item or in other words value located at the size of the item in which cumulative frequency falls is 3.

Therefore, Median is 3

Step III: Calculate the deviation using median.

Step IV: Calculate mean deviation

$$M.D. = \frac{\Sigma f |D|}{N}$$

$$M.D = 221/107, = 2.06$$

Thus, M.D is 2.06

Step IV = Calculate coefficient of Mean deviation

$$\text{Coefficient of M.D} = \frac{M.D}{\text{Median}}$$

$$= 2.06/3$$

$$= 0.68$$

In conclusion we have, Median = 3; Mean deviation = 2.06 and Coefficient of mean deviation from the data = 0.68.

C. Continuous series

Example: In a class of 50 students their marks ranges from 0 to 50 the mean deviation from median and its coefficient from the data.

Marks	0-10	10-20	20-30	30-40	40-50
Number of students	3	10	17	12	8

Marks	Number of students	c.f	Mid-point (m)	m-27.35 D	f D
0-10	3	3	5	22.35	67.05

10-20	10	13	15	12.35	123.5
20-30	17	30	25	2.35	39.95
30-40	12	42	35	7.65	91.8
40-50	9	51	45	17.65	158.85
N = 51					481.15

Step I: Calculate median using $\frac{N+1}{2}$

$$= 51+1/2$$

= 26th and thus lies in the class 20-30.

Step II: So by using the following formula we can obtain median

$$M = L + \frac{\frac{N}{2} - c.f.}{f} \times i$$

Here we have,

$$L = 20; N/2 = 25.5, c.f. = 13, f = 17 \text{ and } i = 10$$

$$= 20 + \frac{25.5-13}{17} \times 10$$

$$= 27.35$$

Step III: Calculate the deviation using median.

Step IV: Calculate mean deviation

$$M.D. = \frac{\sum f |D|}{N}$$

$$M.D = 481.15/51, = 9.43$$

Thus, M.D is 9.43

Step IV: Calculate coefficient of Mean deviation

$$\text{Coefficient of M.D} = \frac{M.D}{\text{Median}}$$

$$= 9.43/27.5$$

$$= 0.34$$

In conclusion we have, Median = 27.5; Mean deviation = 9.43 and Coefficient of mean deviation from the data = 0.34.

Merits

1. Easy to work, as this is based on the concept of average so it is easy to calculate and understand.
2. During working or calculation it includes all observations.

Demerits

1. Method is non-algebraic
2. This method is ill defined.

3.4.3- STUDY OF STANDARD DEVIATION

The concept of standard deviation was first used by Karl Person in 1823 and today its most commonly used measure of dispersion in statistics work and frequently used in biological samples and satisfies most of the characteristics of good measure. Standard deviation is denoted

by Greek letter σ (sigma) and abbreviated variously as S.D. or SD. By definition standard of deviation is defined as “Square root of the arithmetic average of the squares of the deviations measured from the mean”. Further, the SD is an index of variability of the original data points and is reported in most studies.

In general it is computed by six general steps

1. Calculate the mean.
2. Find the difference of each observation from the mean.
3. Square the differences of observations from the mean.
4. Add the squared values to get the sum of squares of the deviation.
5. Divide this sum by the number of observations minus one to get mean-squared deviation, called Variance (σ).
6. Find the square root of this variance to get root-mean squared deviation, called standard deviation. Having squared the original, reverse the step of taking square root.

Calculation of standard deviation: For calculation we may counter all three types of data i.e.,

- A. Series of individual observation.
- B. Discrete series.
- C. Continuous series.

A. When data is in series of individual observations: in case when data is in series of individual method we may use both (i) Actual mean method and (ii) assumed mean method.

Example 1: In a class of 10 students, compute the standard deviation of their marks in Botany paper.

45, 70, 65, 84, 72, 68, 91, 59, 77, 89.

Work procedure:

Calculate the actual mean of the observations.

Calculate the deviation (x) of the value (marks) from the mean ($X-\bar{X}$).

Square the deviation and calculate the Σx^2 .

Calculate the Standard deviation using the following formula

$$SD (\sigma) = \sqrt{\frac{\Sigma x^2}{N}}$$

1. By actual mean method

Marks (X)	$x = (X-\bar{X})$	x^2
45	-27	729
70	-2	4
65	-7	49
84	12	144
72	0	0
68	-4	16
91	19	361

59	-13	169
77	5	25
89	17	289
$\Sigma X = 720$		$\Sigma x^2 = 1786$

Step 1: calculate the mean from the observations

Here, $x = (X - \bar{X})$

For the calculation of \bar{X} we know that $\bar{X} = \Sigma X/N$

i.e., $\bar{X} = 720/10$

So the actual mean of the observation is $\bar{X} = 72$

Step II: calculate the standard deviation by using the formula

$$\begin{aligned} \text{SD } (\sigma) &= \sqrt{\frac{\Sigma x^2}{N}} \\ &= \sqrt{\frac{1786}{10}} \\ &= \sqrt{178.6} \end{aligned}$$

So the deviation of marks among the ten student is, $\sigma = 13.36$.

2. By assumed mean method

Marks (X)	$x = (X-68)$	x^2
45	-23	529
70	2	4
65	-3	9
84	16	256
72	4	16
68	0	0
91	23	529
59	-9	81
77	9	81
89	21	441
	$\Sigma x = 40$	$\Sigma x^2 = 1946$

Here we have assumed 68 as mean from the observation.

Formula for standard deviation

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

here, $\Sigma x^2 = 1946$; $\Sigma x = 40$ and $N = 10$

so put all these value in the formula and we get

$$\sigma = \sqrt{\frac{\Sigma 1946^2}{10} - \frac{(40)^2}{10}}$$

$$\sigma = \sqrt{194.6 - 16}$$

So the deviation of marks among the ten student is same by both methods i.e., $\sigma = 13.36$.

Example 2: In a survey of ten villages for graduate students data was recorded ($X= 36, 49, 79, 51, 37, 87, 63, 74, 31, 43$), calculate the Standard deviation from the observation.

1. By actual mean method.

Graduate people (X)	$x = (X - \bar{X})$	x^2
36	-19	361
49	-6	36
79	24	576
51	-4	16
37	-18	324
87	32	1024
63	8	64
74	19	361
31	-24	576
43	-12	144
$\Sigma X = 550$		$\Sigma x^2 = 3482$

Step 1: calculate the mean from the observations

Here, $x = (X - \bar{X})$

For the calculation of \bar{X} we know that $\bar{X} = \Sigma X / N$

i.e., $\bar{X} = 550 / 10$

So the actual mean of the observation is $\bar{X} = 55$

Step II: calculate the standard deviation by using the formula

$$\begin{aligned} \text{SD } (\sigma) &= \sqrt{\frac{\Sigma x^2}{N}} \\ &= \sqrt{\frac{3482}{10}} \\ &= \sqrt{348.2} \end{aligned}$$

So the deviation of marks among the ten student is, $\sigma = 18.66$.

2. By assumed mean method

Graduate people (X)	$x = (X - 51)$	x^2
36	-15	225
49	-2	4
79	28	784
51	0	0
37	-14	196
87	36	1296

63	12	144
74	23	529
31	-20	400
43	-8	64
$\Sigma x = 40$		$\Sigma x^2 = 3642$

Here we have assumed 68 as mean from the observation.

Formula for standard deviation

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

here, $\Sigma x^2 = 3642$; $\Sigma x = 40$ and $N = 10$

so put all these value in the formula and we get

$$\sigma = \sqrt{\frac{\Sigma 3642^2}{10} - \frac{(40)^2}{10}}$$

$$\sigma = \sqrt{364.2 - 16} = \sqrt{348.2}$$

So the deviation of marks among the ten student is same by both methods i.e., $\sigma = 13.36$.

B. For Discrete Series:

Example: In a sample from pond we got fishes of different weight, calculate the standard deviation within the observation.

Weight of fishes (Kg)	1	2	3	4	5
Frequency	13	12	10	6	4

In case of discrete method standard deviation can be calculated by both methods

Actual mean method and assumed mean method

1. Actual mean method

Work procedure:

Calculate the actual mean of the observations.

Calculate the deviation (x) of the value from the mean ($X - \bar{X}$).

Square the deviation and calculate the Σx^2 .

Multiple them by the respective frequencies and make the total i.e., Σfx^2 .

Calculate the Standard deviation using the following formula.

$$SD (\sigma) = \sqrt{\frac{\Sigma fx^2}{N}}$$

Weight (X)	Frequency (f)	fX	x = (X - \bar{X})	X ²	fx ²
1	13	13	-1.46	2.16	28.08
2	12	24	-0.46	0.22	2.64

3	10	30	0.53	0.28	2.8
4	6	24	1.53	2.34	20.34
5	4	20	2.53	6.4	25.6
N = 45		Σfx = 111		Σfx ² = 79.46	

Here, Here, $x = (X - \bar{X})$

For the calculation of \bar{X} we know that $\bar{X} = \Sigma fX/N$

i.e., $\bar{X} = 111/45$

So the actual mean of the observation is $\bar{X} = 2.47$

Step II, calculate the standard deviation by using the formula

$$SD (\sigma) = \sqrt{\frac{\Sigma fx^2}{N}}$$

$$= \sqrt{\frac{79.46}{45}}$$

$$= \sqrt{1.76}$$

So the deviation of weight among the forty five fishes is, $\sigma = 1.328$.

C. Continuous series

Example: calculate the Standard deviation for the marks obtained by the students in their exam

Marks	4-8	8-12	12-16	16-20
Frequency	2	5	8	3

Work procedure

Calculate the actual mean of the observations.

Find the midpoint from the class.

Calculate the deviation (x) of the value (marks) from the mean $(X - \bar{X})$.

Square the deviation and calculate the Σx^2 .

Multiple them by the respective frequencies and make the total i.e., Σfx^2 .

Calculate the Standard deviation using the following formula.

$$SD (\sigma) = \sqrt{\frac{\Sigma fx^2}{N}}$$

Marks	frequency	mid point (m)	fm	$x = (m - \bar{X})$	x^2	fx^2
4-8	2	6	12	-6.67	44.48	89.17
8-12	5	10	50	-2.67	7.12	35.64
12-16	8	14	112	1.33	1.76	14.15
16-20	3	18	54	5.33	28.40	85.22
N = 18		Σfm = 228		Σfx ² = 224.18		

Here $\bar{X} = \Sigma fm/N$

$$= 228/18 \text{ or } = 12.67$$

According to formula

$$\begin{aligned} \text{SD } (\sigma) &= \sqrt{\frac{\sum fx^2}{N}} \\ &= \sqrt{\frac{\sum 224.18}{18}} \\ &= \sqrt{12.45} \end{aligned}$$

Thus, $\sigma = 3.52$

Merits of Standard Deviation

- 1) Standard deviation rigidly defined and helps to summarises the deviation of a large distribution.
- 2) It is one of the most reliable measures of dispersion and used to observe the variations occur among the collected data.
- 3) The higher the value of standard of variation more the data have odd reading or more the data is fluctuating, so more the data in non-reliable in case of laboratory results.
- 4) It helps in finding the suitable size of sample for valid conclusions.

Demerits of Standard Deviation

- 1) Standard deviation includes very length mathematical calculations.
- 2) Standard deviation gives weightage to extreme values.

3.4.4- STUDY OF STANDARD ERROR

The standard error is nothing but the intra differences in the sample measurements of number of samples taken from a single population. We can estimate how much sample means will vary from the standard deviation of this sampling distribution, which we call the standard error (SE). In other words, we can say that the basic difference between the standard deviation and standard error of mean is that the standard deviation measures the extent to which the individual items differ from the central value and the standard error measures the extent to which individual sample mean differ from the population mean. Further, the standard error of the sample mean depends on both the standard deviation and the sample size, by the simple relation $SE = SD/\sqrt{(\text{sample size})}$.

In general, the standard error of mean measure the extent to which the sample mean differ from the population mean and this we have true for standard error of the median, standard deviation, proportion, variance, etc.

Standard Error of sampling measurements on statistics is calculated by using the formula given below:

$$S.E = \frac{\sigma}{\sqrt{N}}$$

Where, S.E = Standard Error of sampling measurement

σ = Standard deviation of the scores obtained from the population.

N = Size of the sample or total number of units in a sample

3.5- SUMMARY

The word Statistics was first time used by “Gattfried Ahenwall”. In general, statistics is the science that deals with the collection, classification, analysis and interpretation of numerical facts or data. The use of statistics in Biology is known as biostatistics or biometry.

Biostatistics may be defined as the application of statistical methods to the solution of biological problems. In biology we usually encountered with large set of data's, because whatever our approach is for data collection, we have to collect data at least in replicates so that more reliable conclusion can be drawn. Or in other words we can say that, from the collected data the values of the variable tend to concentrate around some central value of observation, that value can be used as the representative value. So this tendency of distribution is known as central tendency. Under measures of central tendency Mean, median and mode are the most popular averages that are studied, while Quartiles, Deciles and Percentiles are some other famous measure of central tendency. Besides, the central tendency, dispersion is commonly used term to explain scatterness, deviation, spread, fluctuation and variability of collected data. Range which deals with the highest and lowest terms of series to conclude its result is one of the simplest measures of dispersion. Standard deviation is another example of measure of dispersion and most commonly used measure of dispersion in statistics and frequently used in biological samples and satisfies most of the characteristics of good measure.

According to one worker, an average will be more meaningful and more accurate if it is studied in light of measures of dispersion. Thus both techniques i.e., central tendency and measure of dispersion collectively gives a useful hand to solve the modern day problems or satisfied the collected data statistically.

3.6-GLOSSARY

Analysis -The process of going into the deep of a phenomenon, data-set, thought, etc., and looking at its various components.

Biometrics- The study of measurement and statistical analysis in medicine and biology.

Biostatistics- The application of research study design and statistical analysis to applications in medicine and biology.

Data- A set of observations, generally in numerical format but can be in text format also.

Descriptive statistics- Statistics, such as the mean, the standard deviation, the proportion, and the rate, used to describe attributes of a set of data.

Frequency distribution- In a set of numerical observations, the list of values that occur along with the frequency of their occurrence. It may be set up as a frequency table or as a graph.

Frequency- The number of times a given value of an observation occurs. It is also called counts.

Mean - Mean refers to the average that is used to derive the central tendency of the data in question.

Measure of central tendency- For quantitative data it is observed that there is a tendency of the data to be distributed about a central value.

Measurement error- The amount by which a measurement is incorrect because of problems inherent in the measuring process; also called bias.

Measures of dispersion- Index or summary numbers that describe the spread of observations about the mean.

Median - To find the median, we arrange the observations in order from smallest to largest value. If there are an odd number of observations, the median is the middle value. If there is an even number of observations, the median is the average of the two middle values.

Modal class- The interval (generally from a frequency table or histogram) that contains the highest frequency of observations.

Mode - The mode is a statistical term that refers to the most frequently occurring number found in a set of numbers.

Standard error (SE)- The standard deviation of the sampling distribution of a statistic.

Statistic - A summary measure for any characteristic in the sample or the group actually studied, such as mean, median or standard deviation of a sample, or proportion of subjects found affected in a sample.

Statistical analysis- Subjecting data to the rigours of statistical methods so that the uncertainty levels are either quantified or minimized, or both.

Variable - A characteristic that varies from person to person, or from situation to situation. Platelet count in different persons is variable but number of eyes or number of fingers is not a variable. See quantitative variable, qualitative variable, discrete variable, continuous variable, dependent variable, and independent variable.

Variance -A measure of dispersion or scatteredness of quantitative data obtained as average of the squared deviations from mean.

3.7- SELF- ASSESSMENT QUESTIONS

3.7.1- Short answers

Q1. Write a short note on Biostatistics.

Q2. Write difference between mean, median and mode.

Q6. Height of 7 students (in cm) is given below. If the mean of height of 7 students is 165, what is the value of x?

168 170 X 160 162 164 162

- (a) 170 (b) 165
(c) 160 (d) 169

Q7. Observe the following observations. They are runs scored by a batsman in six matches and arranged in ascending order. If the value of median is equal to 54, then what is the value of x? 25 38 50 X 84 106

- (a) 60 (b) 58
(c) 54 (d) 59

Q8. On his first 5 zoology tests, Aman received the following scores: 72, 86, 92, 63, and 77. What test score must Aman earn on his sixth test so that his average (mean score) for all six tests will be 80

- (a) 90 (b) 86
(c) 95 (d) 80

Q9. Which of the following is NOT a common measure of central tendency?

- (a) Mean (b) Mode
(c) Median (d) Range

Q10. Below are the observations of the marks of a student. What is the value of mode?

84 85 89 92 93 89 87 89 92

- (a) 92 (b) 9
(c) 93 (d) 89

Answers: 1 = d; 2 = c; 3 = d; 4 = a; 5 = d; 6 = d; 7 = b; 8 = a; 9 = d; 10 = d.

3.8- REFERENCES

1. Agresti, A. (1990). *Categorical Data Analysis*. New York: Wiley.
2. Altman, D.G. and Bland, JM. (2005) Standard deviations and standard errors. *BMJ*, 331: 903.
3. Carley, S and Lecky, F. (2003) Statistical consideration for research. *Emerg Med J*, 20: 258-62.
4. Carlin, J.B. and Doyle, L.W. (2001) Basic concepts of statistical reasoning: standard errors and confidence intervals. *J Paediatr Child Health*, 36, 502–505.

5. Cox, D. R. and Oakes, D. (1984). Analysis of Survival Data. New York: Chapman & Hall.
6. Curran-Everett D. (2008) Explorations in statistics: standard deviations and standard errors. Adv Physiol Educ, 32: 203-8.
7. Dodge, Y. (2003) The Oxford Dictionary of Statistical Terms, Oxford: Oxford University Press. ISBN 0-19-920613-9.
8. Everitt, B. S. (2003) The Cambridge Dictionary of Statistics, CUP. ISBN 0-521-81099-x.
9. Freeman, D. H. (1980). Applied Categorical Data Analysis. New York: Marcel Dekker.
10. Livingston, E.H. (2004) The mean and standard deviation: what does it all mean? J Surg Res, 119: 117-23.
11. Mahler, D.L (1967) Elementary statistics for the anesthesiologist. Anesthesiology; 28: 749-59.
12. Nagele, P. (2003) Misuse of standard error of the mean (SEM) when reporting variability of a sample. A critical evaluation of four anaesthesia. Br J Anaesthesiol, 90, 514–516.
13. Reichmann, J. (1961) Use and Abuse of Statistics, London: Methuen. Reprinted 1964–1970 by Pelican. Appendix 8.
14. Rosenbaum, S.H. (2015) Statistical methods in anesthesia. In: Miller’s Anesthesia. 8th ed. Edited by Miller RD, Cohen NH, Eriksson LI, Fleisher LA, Wiener-Kronish JP, Young WL: Philadelphia, Elsevier Inc.
15. Streiner, D. L. (1996) Maintaining standards: differences between the standard deviation and standard error, and when to use each. Can. J. Psychiatry, 41, 498–502.
16. Vysochanskij, D. F. & Petunin, Y. I. (1980) Justification of the 3s rule for unimodal distributions. Theory Probab Math Stat, 21, 25–36.
17. Webster, C. S. & Merry, A. F. (1997) The standard deviation and the standard error of the mean. Anaesthesia, 52, 183.

3.9- SUGGESTED READING

- Agresti, A. (1990). Categorical Data Analysis. New York: Wiley.
- Chap T. Le (2003) Introductory Biostatistics, John Wiley & Sons, Inc., Hoboken, New Jersey
- Khanal, A.B. (2016) Methods in Biostatistics for Medical Students and Research Workers, Jaypee Brothers Medical Publishers.
- Kothari, R. (2004). Research methodology, Methods and techniques. New Age International Limited, New Delhi.

3.10- TERMINAL QUESTIONS

Q1. Calculate the sample mean.

Marks	30-40	40-50	50-60	60-70	70-80
-------	-------	-------	-------	-------	-------

Frequency	15	13	31	33	22
-----------	----	----	----	----	----

Q2. Calculate the sample median for the following data.

Number of student	12	6	8	17	4
Marks	15	20	23	18	27

Q3. Compute the standard deviation of marks taken by 10 students by actual mean and assumed mean method.

Marks	80	45	92	66	76	81	39	77	59	86
-------	----	----	----	----	----	----	----	----	----	----

Q4. Discuss in brief the utility of standard of deviation in biology and write its merits and demerits.

Q5. Calculate the mean, median, mode and standard deviation from the following data.

42	37	33	46	29	30	37	40	21	25	18
----	----	----	----	----	----	----	----	----	----	----

Q6. A basket has 8 apple of weight 120 gm each, 6 apple weight 130 gm each, 5 apples of 150 gm each 3 apples of 160 gm each and 2 apples of 175 gm each. What is the mean weight of an apple?

Q7. The mean wage of 100 laborers working in a factory running two shifts of 60 and 40 workers respectively is Rs. 38. The mean wage of 60 laborers working in the morning shift is Rs. 40. Find the mean wage of 40 laborers working in the evening shift.

UNIT-4- STATISTICAL INFERENCE AND PROBABILITY

4.1-Objectives

4.2-Introduction

4.3-Statistical inference

4.4-Correlation

4.5-Regression

4.5.1-Linear

4.5.2- Multiple

4.6-Probability

4.6.1-The theory

4.6.2-Applications in biology (particularly genetics)

4.7-Probability distribution

4.8-Summary

4.9-Glossary

4.10-Self Assessment Question

4.11-References

4.12-Suggested Readings

4.13-Terminal Questions

4.1- OBJECTIVE

After going through this unit, you will be able to understand:

- A. Covariance between two variables.
- B. Scatter diagram and correlation coefficient.
- C. The concept of regression and its types.
- D. Apply linear regression models to given data and use the regression equation for prediction.
- E. The basic principles of probability.
- F. The use of the principles of probability to solve genetic problems.

4.2- INTRODUCTION

In our previous units we have those statistical measures that use only a single variable or in other words we have discussed the univariate distribution. Now, in this unit we are going to discuss or shall study the problem of describing the degree of simultaneous variation for two or more variables or when we have measures on two variables the joint presentation of the two variables is called a bivariate distribution. In this unit we are going to discuss the descriptive method to explore the joint distribution of the pair of values of two variables.

In research work we come across many activities, which are dependent on each other. In plant physiology or ecology we see a large number of problems involving the use of two or more variables. Identifying these variables and their dependency helps us in resolving the objective of the research and drawing inference from the results. Many times there are problems or situations where two variables seem to move in the same direction (use of growth regulators and plant growth) such as both are increasing or decreasing. At times an increase in one variable is accompanied by a decline in another. We are going to discuss correlation, regression and probability. In probability we are going to discuss the basic concept of probability and its distribution and to show how the various basic probability distributions are constructed and how these probability distributions have immensely useful applications and explain a wide variety of operations.

4.3- STATISTICAL INFERENCE

Statistical inference is a term which deals with the methods of drawing conclusions from the experiment or statement or a sample or a population characteristic on the basis of information contained in a sample drawn from the source. Statistical inference has two important aspects i.e., statistical estimation and hypothesis testing.

Statistical estimation: Statistical estimation may be explained as estimating unknown population parameters from the knowledge of statistical measures based on sample studies. Since in most of the statistical research studies have unknown population parameters and have to be

estimated from a sample. Thus, the random variables (such as \bar{X} and σ_p^2) used to estimate population parameters. These parameters are known as estimators and their values (suppose $\bar{X} = 12$ and $\sigma_p^2 = 31$) are known as estimation.

In general we can calculate the estimation by two methods or estimation may be of two types: point estimation and interval estimation. In point estimation we estimate the value of the sample or experiment or a parameter as a single point or single figure. While in the case of interval estimation we estimate both, the lower and upper limits of the sample within which our mean is suppose to move.

Hypothesis: A hypothesis is a tentative statement about a characteristic of a something. A hypothesis can be an assertion or a claim and consider as principle instrument of claim or in other words hypothesis is a formal question that is created with the intends to resolve. Hypothesis is denoted by symbol “H”. For example, official records for recent years show that the cases of forest fire in Uttarakhand is increase by 13 percent from the last year. Since a claim about the rate of increase in forest fire is made, it could be considered as a hypothesis.

Generally in statistical hypotheses, we often talk about null hypothesis and alternative hypothesis. A null hypothesis denoted by H_0 is the statement that we consider to be true about the hypothesis and put to test by using a test statistic. For example, for the forest fire 13 percent, there are possibility that the null hypothesis that we intend to test is not the forest fire is not equal to 13 percent rather it may be more then 13 or less than 13 percent. If our claim is statically proved to fit or good or the claim is right then the null hypothesis is accepted. But if the results are not as per claim then there is other hypothesis named alternative hypothesis, which holds true in case the null hypothesis is not true. We denote alternative hypothesis by the symbol H_A . Further there are some statistical tests, to test the hypothesis and these are T test, F test, Chi test and Z test.

4.4- CORRELATION

The term correlation indicates the relationship between two variables in which the change in the value of one variable the value of other variable also change. In statics we often encounter situations where data appears as pairs of figures relating to two variables, so the statistical tool which helps us to study the relationships between two or more than two variables is called correlation. Correlation works only with the quantifiable data in which numbers are meaningful and it cannot be used for purely non-quantifiable data (gender, status, colour), in the words of Croxton and Cowden, when the relationship is of a quantitative nature, the appropriate statistical tool for discovering and measuring the relationship and expressing it in a brief formula is known as correlation. In last according to Tuttle, “Correlation is an analysis of the co-variation between two or more variables.

4.4.1- Type of correlation

1. Positive and Negative correlation

2. Simple and Multiple correlation
3. Partial and Total Correlation
4. Linear and Non Linear correlation

Positive and Negative Correlation:

Positive correlation refers to the movement of variables in the one direction. It means if one variable is increasing, the other is also increasing or if one variable is decreasing, the other is also decreasing. Whereas the negative correlation is refers to movement of variables in opposite direction. In other words if the value of one variable increases the value of other variable is accompanied by decreases its value or the vice-versa of the same.

Few examples for positive correlation

- a. Increase in heights and weight of a group of persons is a positive correlation.
- b. The longer your hair grows the more shampoo you needs.
- c. The more petrol you put in your bike, the farther it can go.
- d. Price of commodity and amount of supply.
- e. As the amount of moisture increases in an environment, the growth of mold spores increases.

Few examples for negative correlation-

- a. The more one works, the less free time one has.
- b. As a tadpole gets older, its tail gets smaller.
- c. Demand of a commodity may go down as a result of rise in price.

Simple and Multiple Correlations: Simple correlation refers to the type of correlation in which only two variables are studied or in other words the relationship is confined to two variables. For example, when one studies relationship between the yield of rice and the area or land on which the seeds were sown. Multiple correlations, is a type of correlation in which there are more than two or in other words three or more variables are studied. For example, relationship of yield of rice, soil type, chemical and use of pesticides.

Partial Correlation: This type of correlation refers to the subtype of multiple correlation in which three or more variables but not all variables are consider. For example, the yield of rice is depend on nature of soil, water, fertilizer, type of seed and use of pesticides, but only two or more variable used as rest are assumed to be constant.

Linear and Non-linear curvilinear correlation: When the variation in the values of any two variables have a constant ratio, then the relationship is known as linear correlation, but if the values are not constant then the relationship is non-linear.

4.4.2- Degree of correlation

The degree or intensity of relationship between two variables can be classified into there

- A. **Perfect degree of correlation:** This type of correlation shows the perfect relationship between two variables and with the fluctuation in the value of one the value of other is also changes.
- B. **Limited degree of correlation:** In this type of correlation the variable shows low level of interdependence among each other or in other words there is unequal change in the same or opposite direction.
- C. **Absence of correlation:** In this type of correlation, there is no relationship exists between variables or in other words there is no interdependence between the two variables. Thus, this it indicate that there is no correlation or zero correlation between two variables.

4.4.3- Method of studying correlation

There are three popular methods of studying correlation.

- A. **Scatter diagram**
- B. **Karl Pearson's coefficient of correlation**
- C. **Rank correlation**

A. Scatter diagram

A scatter diagram is a simple but useful technique for visually examining or a method for getting some idea about the presence of correlation, without calculating any numerical value. It is a diagrammatic representation of bivariate data to ascertain the relationship between two variables. In this technique, the values of the two variables are plotted as points on a graph paper.

In a scatter diagram the degree of closeness of the scatter points and their overall direction enable us to examine the relationship. When the plotted points show some trend i.e., upward or downward we can assume that there is some correlation between them. If the trend is upward the correlation is supposed to be positive and if all the points lie on a line and then followed upward trend the correlation is perfect or in other word strong positive correlation (Figure 4.1). But if the points are widely dispersed than the correlation is positive but week or low correlation. Similarly, if the points followed the downward trend then the correlation is negative and further if the points are widely dispersed than the correlation is negative but week or low correlation.

Merits of Scatter diagram

1. This method is the first step in the investigation of correlation for a large data.
2. Scatter method is non-mathematical method to study the nature of given data.
3. This method is easy and attractive.
4. This give us a good idea whether the results have any correlation, if have what type of correlation the two series have by just observing the graph.
5. The results do not effect with the extreme values.

Demerits of Scatter diagram

1. Since the method is non-mathematical, it give a rough idea of how to variables are related.
2. The degree of correlation cannot be calculated by this method.
3. Not suitable when the data have more than two variables.

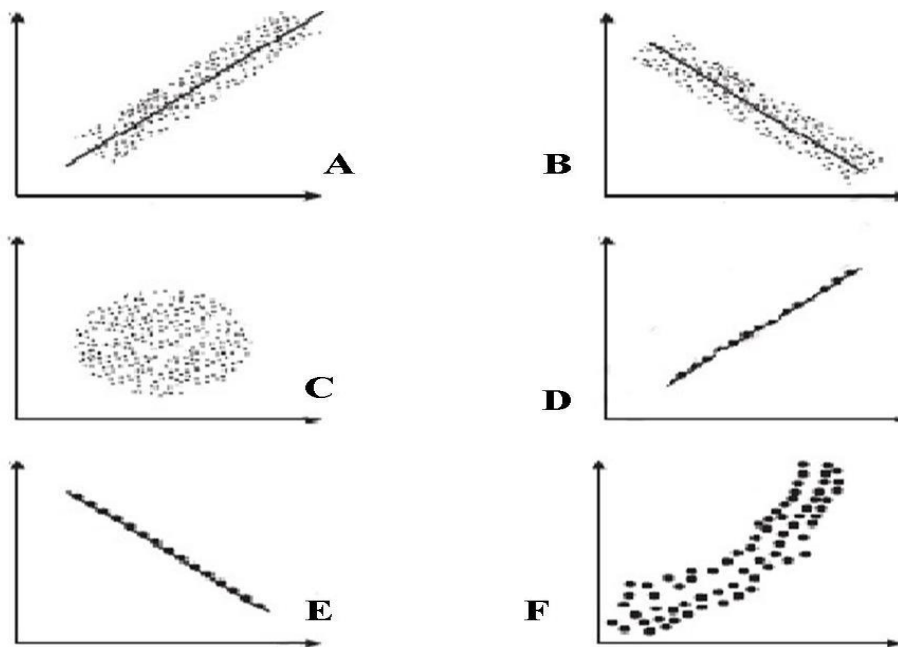


Figure 4.1: a= Positive correlation; b= Negative correlation; c= No correlation; d= Prefect positive correlation; e= Prefect negative correlation; f= Positive non-linear relation.

B. Karl Pearson's coefficient of correlation

This method of calculating correlation is also known as product moment correlation coefficient, simple correlation coefficient or Pearson's coefficient; it is denoted by (r). This method of calculating coefficient of correlation is based on the covariance of the two or more variables in a series.

If the two variables under study are X and Y, the following formula suggested by Karl Pearson can be used for measuring the degree of relationship of correlation

$$r = \frac{\Sigma xy}{N.\sigma_x\sigma_y} \text{ or } r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \times \Sigma y^2}}$$

Where: σ_x = standard deviation of X series

σ_y = standard deviation of Y series

N= number of pairs of observation

r = coefficient of correlation

The coefficient of correlation of two variables is obtained by dividing the sum of the products of the corresponding deviation of the various items of two series from their respective means by the products of their standard deviation and the number of pairs of observation.

Example: Calculate coefficient of correlation from the given data and interpret the results.

Marks in Botany	34	22	16	23	24	21	15	19	26	30
Marks in Zoology	12	23	29	20	17	16	18	16	22	27

Solution:

1. Calculate the arithmetic mean of X and Y series.
2. Find out the deviation of X and Y.
3. Square these deviation and obtained Σx^2 and Σy^2 respectively.
4. Multiply the calculated deviation and find out the total Σxy .
5. Calculate coefficient using the given formula.

Marks in Botany	X-\bar{X}	x²	Marks in Zoology	Y-\bar{Y}	y²	xy
34	11	121	12	-9	81	99
22	-1	1	23	2	4	2
16	-7	49	29	8	64	56
23	0	0	20	-1	1	0
24	1	1	17	-4	16	4
21	-2	4	16	-5	25	10
15	-8	64	18	-3	9	24
19	-4	16	16	-5	25	20
26	3	9	22	1	1	3
30	7	49	27	6	36	42
$\Sigma X = 230$		$\Sigma x^2 = 314$	$\Sigma Y = 200$		$\Sigma y^2 = 262$	$\Sigma xy = 260$

According to formula:

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \times \Sigma y^2}}$$

Arithmetic mean of X and Y series can be calculated as

$$\bar{X} = \frac{\Sigma X}{N} = 230/10 = 23.$$

$$\bar{Y} = \frac{\Sigma Y}{N} = 200/10 = 20.$$

Further we have, $\Sigma xy = 260$, $\Sigma x^2 = 314$ and $\Sigma y^2 = 262$

So by putting all value in given formula we get

$$r = \frac{260}{\sqrt{314 \times 262}}$$

$$r = \frac{260}{286.82}$$

Thus, r= 0.91

Hence, there is high degree of positive correlation.

C. Rank Correlation

Occasionally data are known not to follow the bivariate normal distribution, yet we wish to test for the significance of association between the two variables. One method of analyzing such data is by ranking the variates and calculating a coefficient of rank correlation. Rank correlation also known as Spearman correlation and denoted by “r or ρ (rho)” is a method which is use, when quantification of variables becomes difficult such beauty of nature, colour, stress, leadership ability, knowledge of person etc, and was developed by British psychologist Charles Edward Spearman in 1904. In this method results cannot be measure quantitatively but ranks are allotted to each element either in ascending or descending order. Thus the developed formula of Spearman help in obtaining the correlation coefficient between ranks of *n* individual allotted in two series of ranks.

Formula,
$$r \text{ or } \rho (\text{rho}) = 1 - \frac{6\sum D^2}{n(n^2 - 1)}$$

Where r or ρ (rho)= rank difference between of X and Y variables.

D= difference between the pair of the same individual in the two characteristics.

n= number of pairs.

Example: Calculate the coefficient correlation by rank method of following data:

Marks in Botany	24	19	27	36	30	25
Marks in Zoology	37	29	28	31	33	24

Solution: For the calculation of coefficient of rank correlation, first we have to determine the ranks on the bases of marks obtained in each subject and then tabulate the obtained data.

S. No	Marks in Botany	Rank (R ₁)	Marks in Zoology	Rank (R ₂)	d=R ₁ -R ₂	d ²
1	24	5	37	1	4	16
2	19	6	29	4	2	4
3	27	3	28	5	-2	4

4	25	4	31	3	-1	1
5	30	2	33	2	0	0
6	2536	1	24	6	-5	25
						$\Sigma d^2 = 50$

Thus according to formula: $r \text{ or } \rho (\text{rho}) = 1 - \frac{6\Sigma D^2}{n(n^2-1)}$

$$= 1 - \frac{6 \times 50}{6(36-1)}$$

$$= 1 - \frac{300}{210}$$

$$= 1 - 1.43 = -0.43$$

Since the obtained results showed negative rank correlation (-0.43), this indicates that the student who is best in one subject is worst in the other and vice-versa.

4.4.4- Uses of correlations:

1. Correlation helps to study or draw a result in bivariate data.
2. Correlation analysis helps in deriving precisely the degree and the direction of relationship when two or more variable are involved.
3. Correlation helps in determining the individual difference and finding error in the perditions.
4. The effect of correlation is to reduce the range of uncertainty of our prediction. The prediction based on correlation analysis will be more reliable and near to reality.
5. The measure of coefficient of correlation is a relative measure of change
6. Ecological data and other bivariate data can be analysis by this method.
7. Correlation is important in many areas of measurement and evaluation on education.
8. Correlation help in determining Reliability in given data or condition.

4.5- REGRESSION

In general data collection or sample survey the researcher is asked to relate two or more variables to predict an outcome. For example, in phyto-sociological samplings, how the vegetation varies with altitude for a given area. An example is the association between type of soil, vegetation, and forest type. The statistical methods used to define or describe such relationships are termed correlation and regression analysis. In general correlation analysis provides a quantitative way of measuring the strength of a relationship between two variables, while the regression analysis is used to mathematically describe that relationship, with the ultimate goal being the development of an equation for prediction of one variable from one or more other variables.

Regression analysis is a method, introduced by Fransis Galton for estimating or practicing the unknown value of one variable from known value of another. Like correlation analysis,

regression analysis is also used to describe the relationship between two or more variables. In correlation analysis, correlation shows the degree and direction relationship between two variables, it does not clearly specify as to one variable is the cause and the other effect. "In regression analysis, the relationship between two variables (or more) is expressed by fitting a line or curve to the pairs of data points".

In other words if we take a simple case of regression analysis we have to consider one dependent variable and one independent variable. For example, land holding of a family is related to the crop production for the family, so if we assume that land holding of a family increases the crop production for the family is also increases. From this we may say that, crop production is dependent variable and land holding is the independent variable.

If we are going to denote the same on graph, we denote the dependent variable as Y and the independent variable as X.

4.5.1 Linear regression

Simple linear regression is used to describe and predict a linear relationship between two variables X and Y, one independent and one dependent. Independent variables are characteristics that can be measured directly; these variables are also called predictor or explanatory variables used to predict or to explain the behavior of the dependent variable. Dependent variable is a characteristic whose value depends on the values of independent variables.

Mathematically, the regression model is represented by the following equation:

$$y = \beta_0 \pm \beta_1 x_1 \pm e$$

Where,

x = independent variable.

y = dependent variable.

β_1 = The Slope of the regression line

β_0 = The intercept point of the regression line and the y axis.

e = error.

Or by this equation

$$Y_i = a + bX_i$$

Where again,

X = independent variable.

Y = dependent variable.

i = range from 1 to n

From this expression we can calculate the value of Y if we know the values of parameters.

Suppose $a=2$ and $b=1$, then according to formula or equation we have

$$Y = 2 + 1X.$$

And for given X value, we can calculate the Y, suppose $X=3$, thus

$$Y = 2 + 1(3)$$

$$Y = 2 + 3$$

$Y=5$, thus the value for Y is 5.

The value of a and b in the equations are obtained by solving following two equations

$$\Sigma X = Na + b\Sigma Y$$

$$\Sigma XY = a\Sigma Y + b\Sigma Y^2$$

4.5.2 Multiple regression

Multiple regression, is an extension of simple linear regression. It is used when we want to predict the value of a dependent variable (target or criterion variable) based on the value of two or more independent variables (predictor or explanatory variables). Multiple regression allows us to determine the overall fit (variance explained) of the model and the relative contribution of each of the predictors to the total variance explained.

Mathematically, the multiple regression model is represented by the following equation:

$$Y = \beta_0 \pm \beta_1 X_1 \dots \dots \dots \pm \beta_n X_n \pm u$$

Where: X_1 to X_n = Represent independent variables.

Y = Dependent variable.

β_1 = The regression coefficient of variable x_1

β_2 = The regression coefficient of variable x_2

β_0 = The intercept point of the regression line and the y axis.

u = is the stochastic error term

The coefficient of regression

The coefficient of regression are calculated by the formula

$$\text{Regression coefficient of X on Y} = b_{xy} = r \frac{\sigma_x}{\sigma_y}$$

$$\text{Regression coefficient of Y on X} = b_{yx} = r \frac{\sigma_y}{\sigma_x}$$

Where, r = Coefficient of correlation

σ_x =Standard deviation of X series

σ_y =Standard deviation of Y series

Example 1: Calculate the two regression equation for X on Y and Y on X for the following data.

X	1	4	6	8	10	3	5	7	9
Y	10	12	14	16	18	24	20	15	13

Solution: we know to calculate the regression for X on Y or Y on X we have two equations, so by using both equation we can calculate the regression equations.

X	Y	X²	Y²	XY
1	10	1	100	10
4	12	16	144	48
6	14	36	196	84
8	16	64	256	128
10	18	100	324	180
3	24	9	596	72
6	20	36	400	100
7	17	49	289	119
9	13	81	169	117
$\Sigma X= 54$	$\Sigma Y= 144$	$\Sigma X^2= 392$	$\Sigma Y^2= 2474$	$\Sigma XY= 878$

Here, n=9, for both X and Y

In first case: Regression equation of X on Y: $X = a + bY$

The two normal equations are:

$$\Sigma X = Na + b\Sigma Y$$

$$\Sigma XY = a\Sigma Y + b\Sigma Y^2$$

Substituting the values in above normal equations, we get

$$54 = 9a + 144b \dots\dots\dots(i)$$

$$878 = 144a + 2474b \dots\dots\dots(ii)$$

Let us solve these equations (i) and (ii) by simultaneous equation method

Multiply equation (i) by 16 we get $864 = 144a + 2304b$

Now rewriting these equations:

$$864 = 144a + 2304b$$

$$878 = 144a + 2474b$$

$$(-) = (-) + (-)$$

$$-14 = -170$$

Therefore now we have $-14 = -170b$, this can be rewritten as $170b = 14$

Now, $b = \frac{14}{170} = 0.082$ (rounded off)

Substituting the value of b in equation (i), we get

$$54 = 9a + 144(0.082)$$

$$54 = 9a + 11.85$$

$$9a = 54 - 11.85$$

$$9a = -42.15$$

$$a = \frac{-42.15}{9} = -4.68$$

therefore, $a = -4.68$

Now the regression equation X on Y is:

$$X = -4.77 + 0.082Y$$

In second case: Regression equation for Y on X

$$Y = a + bX$$

The two normal equations are:

$$\Sigma Y = Na + b\Sigma X$$

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

Substituting the values in above normal equations, we get

$$144 = 9a + 54b$$

$$878 = 54a + 392b$$

Let us solve these equations (i) and (ii) by simultaneous equation method

Multiply equation (i) by 6 we get $864 = 54a + 324b$

Now rewriting these equations:

$$864 = 54a + 324b$$

$$878 = 54a + 392b$$

$$(-) = (-) + (-)$$

$$\hline -8 = -68$$

Therefore now we have $-8 = -68b$, this can be rewritten as $68b = 8$

Now, $b = \frac{8}{68} = 0.117$ (rounded off)

Substituting the value of b in equation (i), we get

$$144 = 9a + 54(0.117)$$

$$144 = 9a + 6.318$$

$$9a = 144 - 6.318$$

$$a = 15.29$$

therefore, $a = 15.29$

$$Y = 15.29 + 0.117X$$

So the two equations are

$$X = -4.77 + 0.082Y \quad \dots\dots(i) \qquad Y = 15.29 + 0.117X \quad \dots\dots(ii)$$

Example 2: From the following data obtain the two regression lines:

X	9	8	11	10	12	13	16	14	15
Y	9	1	8	6	7	2	5	3	4

Solution: we know to calculate the regression for X on Y or Y on X we have two equations, so by using both equation we can calculate the regression equations.

X	Y	X²	Y²	XY
9	15	81	81	135
1	8	1	64	8
8	16	64	121	128
6	13	36	100	78
7	12	49	144	84
2	10	4	169	20
5	11	25	256	55
3	14	9	196	42
4	9	16	225	36
ΣX=45	ΣY= 108	ΣX²=285	ΣY²=1356	ΣXY=586

Here, n=9, for both X and Y

In first case: Regression equation of X on Y: $X = a + bY$

The two normal equations are:

$$\Sigma X = Na + b\Sigma Y$$

$$\Sigma XY = a\Sigma Y + b\Sigma Y^2$$

Substituting the values in above normal equations, we get

$$45 = 9a + 108b \dots\dots\dots(i)$$

$$586 = 108a + 1356b \dots\dots\dots(ii)$$

Let us solve these equations (i) and (ii) by simultaneous equation method

Multiply equation (i) by 12 we get $540 = 108a + 1296b$

Now rewriting these equations:

$$540 = 108a + 1296b$$

$$586 = 108a + 1356b$$

$$\begin{array}{r} (-) = (-) + (-) \\ \hline \end{array}$$

$$-46 = -60b$$

Therefore now we have $-46 = -60b$, this can be rewritten as $60b = 46$

Now, $b = \frac{46}{60} = 0.766$ (rounded off)

Substituting the value of b in equation (i), we get

$$45 = 9a + 108(0.766)$$

$$45 = 9a + 82.80$$

$$9a = 45 - 82.80$$

$$9a = -37.80$$

$$a = \frac{-37.80}{9} = -4.20$$

therefore, $a = -4.20$

Now the regression equation X on Y is:

$$X = -4.20 + 0.766Y$$

In second case: Regression equation for Y on X

$$Y = a + bX$$

The two normal equations are:

$$\Sigma Y = Na + b\Sigma X$$

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

Substituting the values in above normal equations, we get

$$108 = 9a + 45b$$

$$586 = 45a + 285b$$

Let us solve these equations (i) and (ii) by simultaneous equation method

Multiply equation (i) by 5 we get $540 = 45a + 225b$

Now rewriting these equations:

$$540 = 45a + 225b$$

$$586 = 45a + 285b$$

$$\begin{array}{r} (-) = (-) + (-) \\ \hline \end{array}$$

$$\begin{array}{r} -46 = \quad \quad -68 \end{array}$$

Therefore now we have $-46 = -60b$, this can be rewritten as $60b = 46$

Now, $b = \frac{46}{60} = 0.766$ (rounded off)

Substituting the value of b in equation (i), we get

$$108 = 9a + 45(0.766)$$

$$108 = 9a + 34.47$$

$$9a = 108 - 34.47$$

$$a = 73.53/9$$

therefore, $a = 8.17$

$$Y = 8.17 + 0.766X$$

So the two equations are

$$X = -4.20 + 0.766Y \quad \dots\dots(i)$$

$$Y = 8.17 + 0.766X \quad \dots\dots(ii)$$

Uses of Regression Analysis

1. It provides estimates of values of the dependent variables from values of independent variables.
2. It is used to obtain a measure of the error involved in using the regression line as a basis for estimation.
3. With the help of regression analysis, we can obtain a measure of degree of association or correlation that exists between the two variables.

4.6 PROBABILITY

4.6.1 Probability theory

Probability theory is concerned with the analysis of phenomena that take place in indeterministic, in other words random circumstances. In day to day life we encounter certain proposition where there is no certainty or there is doubt about the certainty for number of events. The degree of doubt for particular event is variable from situation to situation. For example, what card will turn up in the game, what are the chances of rain, what are the chances of twin pregnancy among the 100 pregnancies, or what are the chances of giving birth to a boy in the pregnancy?. In all these statements there is degree of doubt, so measurement of doubt or degree of doubt is called probability. In other words, Probability may define as the *relative frequency or probable chances of occurrence* with which an event is expected to occur on an average. Probability is usually expressed by the symbol ' p ' and its value ranges from zero (0) to one (1). When $p = 0$, it means there is no chance of an event happening or its occurrence is impossible. If $p = 1$, it means the chances of an event happening are 100%, i.e. it is inevitable.

If the probability of an event happening in a sample is p and that of not happening is denoted by the symbol q ,

$$\text{Then } q = 1 - p \text{ or } p + q = 1$$

A probability calculation allows us to predict the likelihood that an event will occur in the future. The accuracy of this prediction, however, depends to a great extent on the size of the sample.

Laws of Probability

It is very important to have a clear concept of probability as it provides the basis for all the tests of significance. It is estimated usually on the basis of following five laws of probability, normal curve and tables.

1. Addition law of probability
2. Multiplication law of probability
3. Binomial law of probability distribution

1. Addition law

Addition law also called as sum rule of probability, “or” rule of probability, states that, the probability that one of two or more mutually exclusive event will occur is equal to the sum of the individual probabilities of the events.

If one event excludes the probability of occurrence of the other event, the events are called mutually exclusive. For example, getting head excludes the possibility of getting tail on tossing a coin; the birth of a male will excludes the possibility of female baby.

Example, if a green ball can be drawn from a bag in 3 ways and red ball from the same bag in 2 ways, then the number of ways in which a green or a red ball can be drawn from the bag will be $(3+2)$ or 5.

2. Multiplication law

Multiplication law also known as product rule of probability, “and” rule of probability, state that, the chance of two or more independent event occurring together is the product or the probability of the events occurring separately. Example, if a person go from one station to the other by 5 vehicles and can come back by 4 vehicles, the number of ways in which a man can come back is 5×4 or 20.

1. Binomial law of probability distribution

Binomial law of probability distribution is explained by binomial expression $(p + q)^n$,

- Where, n = sample size or number of events
- p = the probability of a ‘success’.
- q = probability of ‘failure’.
- and $p + q = 1$

Theorems on probability

There are two important theorems probability, namely

- A. The addition theorem or the theorem of total probability.
- B. The multiplication theorem or theorem on compound probability.

The addition theorem or the theorem of total probability

If the events are mutually exclusive, then the probability of happening of any one of them is equal to the sum of the probabilities of the happening of the separate events, i.e., in other words if, $B_1, B_2, B_3, B_4, \dots, B_n$ be n events and event E is a subset of the union of $B_1, B_2, B_3, B_4, \dots, B_n$

$$\begin{aligned}
 &E \subset (B_1 \cup B_2 \cup \dots \cup B_n), \text{ then} \\
 &P(E) = P(E \cap B_1) + P(E \cap B_2) + \dots + P(E \cap B_n) \\
 &= P(B_1) \times P(E|B_1) + P(B_2) \times P(E|B_2) + \dots + P(B_n) \times P(E|B_n) \\
 &= \sum [P(B_i) \times P(E|B_i)] \qquad \qquad \qquad i = 1, 2, \dots, n
 \end{aligned}$$

The multiplication theorem (Laplace’s principle)

“The probability of occurrence of several independent events is the product of their separate probabilities”.

Suppose, B is an event which is the joint occurrence of n independent events $B_1, B_2, B_3, B_4, \dots, B_n$.

Such that

$$B = B_1 \text{ and } B_2 \text{ and } B_3 \dots \text{ and } B_n.$$

Then, $P(B) = P(B_1) \cdot P(B_2) \cdot P(B_3) \dots \text{ And } P(B_n)$

Or $P(B) = P_1 \cdot P_2 \cdot P_3 \cdot P_4 \dots P_n$

4.6.2 Application of probability in Genetics

Probability is one of the effective tools in the hands of a doctor's or a geneticist and a genetic counselor to assess the possible occurrence of a trait in a family so that by using the basis of probability a lethal trait or some inheritance diseases can be ruled out. Simply by preparing a pedigree chart of the family and applying the basis of this technique future predication can be made for a particular trait.

The chance that an event will occur in the future is called events probability. The general formula of probability is:

$$\text{Probability} = \frac{\text{Number of times an event occurs}}{\text{Total number of events}}$$

Or in other words, the use of probability in genetic in trend to make a prediction of several inherent disease for example, if a doctor wants to calculate the probability that a couple's future offspring will inherit a disease found on a specific locus than doctor may apply this probability rules.

Probability helps them with breeding of livestock, with weather predictions for farming and with crop yield predictions for the market.

For example, when two heterozygous tall pea plant (Tt) are crossed, the phenotypic ratio of the offspring is 3 tall : 1 dwarf.

From this information we can calculate the probability for either type of offspring.

i.e.,
$$\text{Probability} = \frac{\text{Number of times an event occurs}}{\text{Total number of events}}$$

$$P_{\text{tall}} = 3 \text{ tall} / 3 \text{ tall} + 1 \text{ dwarf}$$

$$= 3/4 = 0.75 \text{ or } 75 \text{ percent}$$

$$P_{\text{dwarf}} = 1 \text{ dwarf} / 3 \text{ tall} + 1 \text{ dwarf}$$

$$= 1/4 = 0.25 \text{ or } 25 \text{ percent}$$

Thus from the results we can conclude that the probability of obtaining a tall plant is 75% and dwarf 25%.

In another example, suppose we have AaBbccDd and AaBbCcdd, what is the probability that the offspring will be of genotype aabbccdd.

If we assume that all gene assort independently according to Mendelian cross, then we can do the calculation as

From Aa × Aa, one fourth of the progeny will be "aa" as per the above solved example.

From $Bb \times Bb$, again one fourth of the progeny will be “bb”.

From $cc \times Cc$, one half of the progeny will be “cc”.

And from $Dd \times dd$, again one half of the progeny will be “dd”.

Thus from the observations, the possible probability for aabbccdd is

$$1/4 \times 1/4 \times 1/2 \times 1/2 = 1/64.$$

Thus from results we have that one out of every 64 will have aabbccdd genome.

Example: Assuming a couple is heterozygous (Aa) for albinism, what is the probability that four children out of six born to them are normal?

Solution:

Let A = allelic for normal skin, while a = allele for albinism

Since couple is heterozygous we have

$$Aa \times Aa$$

We have, AA (Normal), Aa and Aa (heterozygous normal) and aa (albino)

Since the ration of normal to albino is 3 : 1,

Thus, the probability for normal son to born is 3/4 and albino is 1/4

Thus, the probability of 4 children being normal is

$$\frac{3}{4} \times \frac{3}{4} \times \frac{3}{4} \times \frac{1}{4} = \frac{81}{216}$$

Thus the probability is = 0.316.

4.7 PROBABILITY DISTRIBUTION

Probability distribution is an extension to the theory of probability which we have discussed above in section 4.6, and in this section we are going to study the concept of a various probability distribution ((binomial, poisson, and normal), and to show these are constructed. To understand better the meaning of or the concept of probability we have to recall the frequency distribution. For example, we may study some data and classify the data in two columns with class intervals in the first column, and corresponding classes frequencies in the second column. The probability distribution is also a two-column presentation with the values of the random variable in the first column and the corresponding probabilities in the second column. For example we recall, Mendelian law of independent assortment with its phenotypic ration 9:3:3:1, so in this case if we collect for example seed from the field randomly and we want to study two character, so we may produce assumption for its expected value for particular characters. Then if we made a tabular condition, we place the observe value by random sampling in first column and the calculated expected values in the second column. Further, in other word this can be explained

theoretically as “These distributions are obtained by expectations on the basis of theoretical or past experience considerations”.

Type of probability distribution

Probability distributions are broadly classified under two heads:

- (i) Discrete Probability Distribution: In this type of distribution, the probability is allowed to take only a limited number of values for example a the probability of student securing marks out of 20, since we have maximum marks 20 so only 21 possible outcomes may come, thus student may secure any one number from 0 to 20.
- (ii) Continuous Probability Distribution: In this type of distribution, the variable of interest may take on any values within a given range i.e., upper or lower value. This type of random variable which can take an infinite number of values is called a continuous random variable

1. Discrete probability distribution: There are two kinds of distributions in the discrete probability distribution.

(i) Binomial distribution, (ii) Poisson distribution

(i) **Binomial distribution:** It is also known as Bernoulli distribution, given by Swiss Mathematician James Bernoulli in 1654-1705. This distribution is the basic and most common probability distribution. This distribution generally deals with the experiments where there are only two possible outcomes. For example, if seed were sown than the result is either a seed shall germinates or fails to germinate, similarly if coined were tossed it may come head or tail.

Binomial law of probability distribution is applicable on following assumptions:

- (i) A trial results or experiment or procedure in either success or failure of an event.
- (ii) The probability of success “p” remains constant for each trial.
- (iii) The trials are mutually independent and the outcome of one trial is neither affected and nor affects others.

Binomial Probability Formula:

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

Where, $p(r)$ = Probability of r successes in n trials

p = Probability of success

q = Probability of failure = $1-p$.

r = No. of successes desired.

n = No. of trials undertaken

Example: A fair coin is tossed 8 times. What is the probability of obtaining 6 or more heads?

Solution: When a fair coin is tossed, the probabilities of head and tail in case of an unbiased coin are equal, i.e., $p = q = \frac{1}{2}$ or 0.5.

According to formula

The probabilities of obtaining 6 heads is : $p(6) = {}^8C_4 (1/2)^6 (1/2)^{8-6}$

$$\text{Or } p(r) = \frac{8!}{6!(8-6)!} (0.5)^6 (0.5)^2$$

$$p(r) = \frac{8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1}{(6 \times 5 \times 4 \times 3 \times 2 \times 1)(2 \times 1)} (0.015)(0.25)$$

$$p(r) = \frac{40320}{(720)(2)} (0.015)(0.25)$$

$$p(r) = 28 \times 0.015 \times 0.25$$

$$p(r) = 0.105$$

Probability of obtaining 7 head is

$$\text{Or } p(r) = \frac{8!}{7!(8-7)!} (0.5)^7 (0.5)^1$$

$$p(r) = \frac{8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1}{(7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1)(1)} (0.0078)(0.5)$$

$$p(r) = \frac{40320}{(5040)(1)} (0.0078)(0.5)$$

$$p(r) = 8 \times 0.0078 \times 0.5$$

$$p(r) = 0.031$$

Probability of obtaining 8 head is

$$\text{Or } p(r) = \frac{8!}{8!(8-8)!} (0.5)^8 (0.5)^0$$

$$p(r) = \frac{8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1}{(8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1)(2 \times 1)} (0.0039)(1)$$

$$p(r) = \frac{40320}{(40320)(1)} (0.0039)(1)$$

$$p(r) = 1 \times 0.0039 \times 1$$

$$p(r) = 0.0039$$

Therefore the probability of obtaining 6 heads is = $0.105 + 0.031 + 0.0039 = 0.1399$ or = 0.14.

(ii) Poisson distribution: Poisson distribution is named after the developer Simeon Poisson a French mathematician. This equation or distribution is generally used when the trials discussed in the case of binomial distribution is very large. Further, it deals with counting the number of occurrences of a particular event in a specific time interval or region of space.

Formula for Poisson distribution

$$p(r) = \frac{m^r e^{-m}}{r!}$$

Where, $p(r)$ = Probability of successes desired.

$r = 0, 1, 2, 3, 4, \dots, \infty$ (any positive integer)

e = a constant with value: 2.7183 (the base of natural logarithms)

m = The mean of the Poisson Distribution, i.e., np or the average number of occurrences of an event.

Therefore we may say that, Poisson is also a discrete probability distribution and it is the limiting form of the binomial distribution. Where the range of the random variable is $0 \leq r < \infty$ and further it consist of a single parameter “ m ”.

2. Continuous probability distribution: In this type of probability distribution the variable of interest may take any value with the range i.e., upper or lower level.

(i) The Uniform Random Variable or distribution: This type of distribution is the simplest of a few well-known continuous distributions. It may be explained as suppose X is a continuous random variable such that if we take any subinterval of the sample space, then the probability that X belongs to that subinterval is the same as the probability that X belongs to any other subintervals of the same length.

(ii) Normal Distribution: Normal distribution is one of the most versatile and used continuous distributions. The application of this distribution or this distribution is can be applied in different areas like real life situations, biology, psychology, ecology, probability and statistics etc.

During the mid 17th and 18th century mathematicians Abraham De Moivre and Pierre Laplace, while working on various problems in probability found that the distribution corresponding to certain random variables got special property when graphed, the special property was a bell shaped curve, the pattern of the curve they consider as normal curve and distribution as normal distribution. Later this class of distribution was studied extensively by another mathematician Karl Friedrich Gauss in 1809 and 1816 and because of Gauss’s enormous contribution, it is popularly known as 'Gaussian distribution'.

Some important characteristics of Normal Distribution are

1. This is a unimodal concept i.e., the curve has only a single peak.
2. Because of the symmetry of the normal probability distribution (skewness = 0), the mean, median and the mode of the distribution are also at the centre. Thus, for a normal curve, the mean, median and mode are the same value.
3. The two tails of the normal probability distribution extend indefinitely but never touch the horizontal axis.

To define a particular normal probability distribution, we need only two parameters i.e., the mean (μ) and standard deviation (σ). Irrespective of the value of mean (μ) and standard deviation (σ), for a normal distribution, the total area under the curve is considered to be as 1.00.

The area under the normal curve is approximately distributed by its standard deviation as follows:

If $\mu \pm 1\sigma$ covers 68% area, i.e., 34.13% area will lie on either side of μ .

If $\mu \pm 2\sigma$ covers 95.5% area, i.e., 47.75% will lie on either side of μ .

If $\mu \pm 3\sigma$ covers 99.7% area, i.e., 49.85% will lie on either side of μ

4.8 SUMMARY

In sum up from this unit, we came to know the importance of statistical interference and how the interference can be drawn from the given data. In next section we studied the bivariate data and the different tools to study the bivariate data in which first we discussed correlation. The term correlation indicates the relationship between two variables in which the change in the value of one variable the value of other variable also change. Depending upon the observation from sample we have positive or negative correlation, Simple and Multiple correlation, Partial and Total Correlation and Linear and Non Linear correlation. Further, Correlation helps to study or draw a result in bivariate data. Correlation analysis helps in deriving precisely the degree and the direction of relationship when two or more variable are involved. Correlation is important in many areas of measurement and evaluation on education. Correlation help in determining the reliability in given data or condition.

Regression analysis is a method, introduced by Francis Galton for estimating or practicing the unknown value of one variable from known value of another. Simple linear regression is used to describe and predict a linear relationship between two variables X and Y, of which one is independent and other is dependent variable. In extension to simple linear regression we use multiple regression analysis to predict the value of a dependent variable based on the value of two or more independent variables.

In another section we worked out the concept of probability and probability may define as the relative frequency or probable chances of occurrence with which an event is expected to occur on an average. In probability we are going to discuss the basic concept of probability and its distribution and to show how the various basic probability distributions are constructed and how these probability distributions have immensely useful applications and explain a wide variety operations.

4.9 GLOSSARY

Auto-correlation: Similar to correlation in that it described the association or mutual dependence between values of the same variable but at different time periods.

Correlation Coefficient: A number lying between -1 (Perfect negative correlation) and +1 (perfect positive correlation) to quantify the association between two variables.

Correlation: Degree of association between the two variables.

Degrees of Freedom: It refers to the pieces of independent information that are required to compute some characteristic of a given set of observations.

Estimation: It is the method of prediction about parameter values on the basis of sample statistics.

Nominal Variable: Such a variable takes qualitative values and do not have any ordering relationships among them. For example, gender is a nominal variable taking only the qualitative values, male and female; there is no ordering in 'male' and 'female' status.

Parameter: It is a measure of some characteristic of the sample or population.

Random Sampling: random sampling also called as probability sampling, is a procedure where every member of the population has a definite chance or probability of being selected in the sample.

Variable: A characteristic that varies from person to person, or from situation to situation. Platelet count in different persons is variable but number of eyes or number of fingers is not a variable. See quantitative variable, qualitative variable, discrete variable, continuous variable, dependent variable, and independent variable.

Sample: It is the entire collection of units of a specified type in a given place and at a particular point of time.

Scatter Diagram: An ungrouped plot of the two variables, on the X and Y axes.

Statistic: A summary measure for any characteristic in the sample or the group actually studied, such as mean, median or standard deviation of a sample, or proportion of subjects found affected in a sample.

4.10 SELF- ASSESSMENT QUESTIONS

4.10.1- Short answers

- Q1. Write a short note on correlation.
- Q2. Write difference between correlation and regression.
- Q3. Write merits and demerits of coefficient of correlation.
- Q4. Write short note on linear regression.
- Q5. Define probability.

- Q6. Discuss the role of probability in inheritance.
- Q7. Discuss the role of correlation in biological samples.
- Q8. Write short note on merits and utility of probability.
- Q9. Notes on null hypothesis and alternative hypothesis.

4.10.2- Multiple choices

1. The correlation coefficient is used to determine:
 - a. A specific value of the y-variable given a specific value of the x-variable.
 - b. A specific value of the x-variable given a specific value of the y-variable.
 - c. The strength of the relationship between the x and y variables.
 - d. None of these.
2. If there is a very strong correlation between two variables then the correlation coefficient must be
 - a. any value larger than 1.
 - b. much smaller than 0, if the correlation is negative.
 - c. much larger than 0, regardless of whether the correlation is negative or positive.
 - d. None of these alternatives is correct.
3. In regression analysis, the variable that is being predicted is the
 - a. response, or dependent, variable
 - b. independent variable
 - c. intervening variable
 - d. is usually x
4. In a regression analysis if $r^2 = 1$, then
 - a. SSE must also be equal to one
 - b. SSE must be equal to zero
 - c. SSE can be any positive value
 - d. SSE must be negative
5. The coefficient of correlation
 - a. is the square of the coefficient of determination
 - b. is the square root of the coefficient of determination
 - c. is the same as r-square
 - d. can never be negative
6. In regression analysis, the variable that is used to explain the change in the outcome of an experiment, or some natural process, is called
 - a. the x-variable
 - b. the independent variable
 - c. the predictor variable

- d. the explanatory variable
 e. all of the above (a-d) are correct
 f. none are correct
7. If the correlation coefficient is a positive value, then the slope of the regression line
- a. must also be positive
 b. can be either negative or positive
 c. can be zero
 d. can not be zero
8. If the coefficient of determination is 0.81, the correlation coefficient
- a. is 0.6561
 b. could be either + 0.9 or - 0.9
 c. must be positive
 d. must be negative

Answers: 1 =c ; 2 =b ; 3 =a ; 4 =b ; 5 =b ; 6 =e ; 7 =a ; 8 =b.

4.11- REFERENCES

- Agresti, A. (1990). *Categorical Data Analysis*. New York: Wiley.
- Best, J. W. and James V. K. (2004), 'Research in Education', Prentice Hall of India Pvt. Ltd., New Delhi.
- Carley, S. and Lecky, F. (2003) Statistical consideration for research. *Emerg Med J*, 20: 258-62.
- Cox, D. R. and Oakes, D. (1984). *Analysis of Survival Data*. New York: Chapman & Hall.
- Dodge, Y. (2003) *The Oxford Dictionary of Statistical Terms*, Oxford: Oxford University Press. ISBN 0-19-920613-9.
- Everitt, B. S. (2003) *The Cambridge Dictionary of Statistics*, CUP. ISBN 0-521-81099-x.
- Freeman, D. H. (1980). *Applied Categorical Data Analysis*. New York: Marcel Dekker.
- Reichmann, J. (1961) *Use and Abuse of Statistics*, London: Methuen. Reprinted 1964–1970 by Pelican. Appendix 8.
- Vysochanskij, D. F. & Petunin, Y. I. (1980) Justification of the 3s rule for unimodal distributions. *Theory Probab Math Stat*, 21, 25–36.

4.12- SUGGESTED READING

- Agresti, A. (1990). *Categorical Data Analysis*. New York: Wiley.
- Chap T. Le (2003) *Introductory Biostatistics*, John Wiley & Sons, Inc., Hoboken, New Jersey
- Khanal, A.B. (2016) *Methods in Biostatistics for Medical Students and Research Workers*, Jaypee Brothers Medical Publishers.
- Kothari, R. (2004). *Research methodology, Methods and techniques*. New Age International Limited, New Delhi.

4.13- TERMINAL QUESTIONS

1. What are simple linear and multiple regressions? Write the equation for simple linear and multiple regression models.
2. Write a brief essay on statistical estimation.
3. From the following data obtain the two regression lines:

Capital employed	8	5	6	12	15	10	7
Sale	9	7	9	5	2	6	4

4. Calculate the correlation coefficient between the heights of fathers in cm (X) and their sons (Y)

X	171	167	169	177	178	165	172
Y	173	166	167	177	179	164	171

4. Explain the characteristics of a poisson distribution. Give two examples, the distribution of which will conform to the poisson form.

UNIT-5- ADVANCE ANALYSIS METHODS

Contents

- 5.1- Objectives
- 5.2- Introduction
- 5.3- Basic concept and algorithms of cluster analysis
- 5.4- Cluster analysis: what is it?
- 5.5- Different types of clusters
- 5.6- Types of clustering
- 5.7- Types of clustering algorithms
- 5.8- Applications of clustering
- 5.9- Multivariate analysis: Introduction, characteristics and applications
- 5.10- Classification of multivariate techniques
- 5.11- Types of multivariate techniques
- 5.12- Summary
- 5.13- References
- 5.14- Suggested readings
- 5.15- Self assessment Questions
- 5.16- Terminal Questions

5.1 OBJECTIVES

After reading this unit the learners will be familiar to the;

- An introduction to cluster analysis and multivariate analysis
- Define basic concept and algorithms of cluster analysis
- Understand the types of clusters and various clustering methods
- Describe the applications of cluster analysis.
- Explain introduction, characteristics and applications of multivariate analysis.
- Define basic concepts of multivariate analysis.
- Describe the types of multivariate techniques

5.2 INTRODUCTION

A multivariate technique called cluster analysis seeks to categories a sample of subjects (or objects) into various groups according to a variety of measured variables, so that people with similar characteristics are grouped together. There are several techniques for grouping data according to similarity, and they differ greatly from one another. The necessity to divide a sizable data set into groups has led to the application of clustering techniques to a wide range of research challenges. Before beginning the grouping process, it is unknown how many groups there will be and how many observations will be in each cluster. Cluster analysis is useful for sorting and identifying actual groups. This can be applied, for instance, to the study of plant ecology, which allows for the classification of forest vegetation based on groups of species' habitats, habits, and types found there.

Data sets can be analysed using a variety of approaches known as multivariate analysis. Many of these methods are modern and frequently involve the use of highly powerful computing technologies. All statistical techniques that simultaneously assess several measurements on each person or object under study are referred to as such analyses. Therefore, any analysis that simultaneously examines more than two variables can be broadly referred to as multivariate analysis. In order to assist in determining when to employ a specific statistical technique for a specific sort of data, this unit will present a list of these analyses. Additionally, a brief explanation of each technique is given. It is arranged in accordance with the number of data sets to be analysed: one, two, or more. With two data sets, we analyse two scenarios: in the first, one set of data serves as predictors or independent variables, while the second set of data corresponds to measurements or dependent variables; in the second. Different sets of dependent variables are matched by various types of data.

5.3 BASIC CONCEPT AND ALGORITHMS OF CLUSTER ANALYSIS

Data are separated into groups (clusters) using cluster analysis, and these groups can be meaningful, practical, or both. The clusters should reflect the natural structure of the data if the

purpose is to create meaningful groups. Cluster analysis is sometimes only useful as an initial starting point for subsequent tasks, including data summarization. Cluster analysis has long been crucial in a wide range of disciplines, including statistics, biology, pattern recognition, information retrieval, machine learning, and data mining, whether for understanding or utility. Cluster analysis has been used to solve a wide variety of problems in the real world. We offer a few concrete examples, arranged according to whether the clustering is meant to be understood or to be useful.

5.3.1 Clustering for Understanding

- Conceptually significant groups of things with shared traits are crucial to how individuals understand and represent the world. In fact, grouping items into categories and classifying them according to their characteristics are human abilities. For instance, even young infants are able to identify the various items in a picture as being either a building, a car, a person, an animal, a plant, etc. Cluster analysis is the study of methods for automatically discovering classes and clusters are prospective classes in the context of comprehending data. These are a few instances:
- **Biology:** All living creatures are categorised into kingdoms, phyla, classes, orders, families, genera, and species by biologists over the course of many years. Therefore, it may not come as a surprise that a significant portion of the early work in cluster analysis tried to develop a field of mathematical taxonomy that could automatically discover such classification structures. In more recent times, biologists have used clustering to examine the vast volumes of genetic data that are now accessible. To find groupings of genes with related functions, for instance, clustering has been utilised.
- **Business:** Businesses gather a lot of data on their clients, both present and potential. Customers can be divided into a limited number of groups using clustering to facilitate further analysis and marketing efforts.
- **Climate:** It's important to look for trends in the water and atmosphere to understand the Earth's climate. As a result, cluster analysis has been used to identify trends in the atmospheric pressure of the Polar Regions and regions of the ocean that have a significant impact on the climate on land.
- **Information Retrieval:** There are billions of Web pages on the World Wide Web, and a search engine query can produce thousands of pages in response. These search results can be grouped using clustering into a limited number of clusters, each of which represents a different component of the question. For example, a search for "movie" may produce web sites categorised into sections like reviews, trailers, stars, and theatres. A user's examination of the query results is made easier by the hierarchical structure that emerges from the possibility of subdividing each category (cluster) into subcategories (sub-clusters).

- **Psychology and Medicine:** Cluster analysis can be used to find the various subcategories that typically exist for a given illness or condition. For instance, clustering has been applied to categorise various forms of depression. The spatial or temporal distribution of an illness can be examined using cluster analysis to find trends.

5.3.2 Clustering for Utility

The clusters in which individual data objects are found can be abstracted using cluster analysis. Additionally, a cluster prototype—a data object that serves as a representative of the other objects in the cluster—is used by some clustering approaches to describe each cluster. These cluster prototypes can serve as the basis for a variety of data processing or analysis methods. Therefore, cluster analysis is the study of methods for identifying the most representative cluster prototypes in the context of utility.

Summarization: Many data analysis methods, including regression and PCA, have time or space complexity of $O(m^2)$ or more (where m is the number of objects), making them impractical for huge data sets. The approach can be used on a smaller data set that only contains cluster prototypes, as opposed to the complete data set. The results might be comparable to those that would have been achieved if all the data had been available, depending on the type of study, the number of prototypes, and the accuracy with which the prototypes represent the data.

Compression: Data compression is another application for cluster prototypes. Each prototype is given an integer value that serves as its location (index) in the table, which is specifically made up of the prototypes for each cluster. The prototype index linked to each item's cluster serves as a representation of that object. This type of compression is known as vector quantization and is often applied to image, sound, and video data, where (1) many of the data objects are highly similar to one another, (2) some loss of information is acceptable, and (3) a substantial reduction in the data size is desired.

Efficiently finding nearest neighbors: The pairwise distance between all points can be calculated in order to find nearest neighbours. Clusters and their cluster prototypes are frequently considerably easier to find. We can use the cluster prototypes to cut down on the number of distance calculations required to determine an item's closest neighbours if the object and the prototype are relatively close to one another. The objects in the corresponding clusters cannot be nearest neighbours of one another if two cluster prototypes are far apart, according to intuition. Because of this, computing the distance to objects in nearby clusters—where the proximity of two clusters is determined by the distance between their prototypes—is all that is required to identify an object's closest neighbours.

5.4 CLUSTER ANALYSIS: WHAT IS IT?

Cluster analysis groups data objects based on information found only in the data that explains the objects and their associations. The idea behind groups is to have items that are similar to (or

connected to) one another while being distinct from (or unrelated to) those in other groups. Better or more distinct clustering occurs when there is a higher degree of similarity (or homogeneity) within a group and a higher degree of variation between groups.

The term "cluster" is not well defined in many applications. Consider Fig. 5.1, which depicts 20 points and three distinct clustering methods, to get a better understanding of how difficult it is to decide what qualifies as a cluster. Marker shapes reveal cluster membership. The data are divided into two and six portions, respectively, in Figs. 5.1(b) and (d). However, it's possible that the apparent separation of each of the two bigger clusters into three subclusters is really an anomaly of the human visual system. Furthermore, it might not be unreasonable to claim that the points form four clusters, as shown in Fig. 5.1(c). This diagram demonstrates how the definition of a cluster is ambiguous and that the ideal definition relies on the type of data being used and the desired outcomes. The division of data objects into groups using various methods is connected to cluster analysis. For instance, clustering can be seen as a type of classification since it labels objects with class (cluster) labels. However, it only gets these labels from the data.

Cluster analysis is different from classification analysis, whereas in cluster analysis there is no assumption of predefined groups and the number of groups will be determined based on the similarity between the observations (items), in classification analysis the number of groups is known and the goal is to reassign each observation to one predefined group.

Additionally, although segmentation and partitioning are occasionally used as synonyms for clustering, these names are usually employed for methods that fall beyond the conventional parameters of cluster analysis. For instance, the term "partitioning" is frequently used to refer to methods for breaking up graphs into smaller ones and is not always associated with clustering. The term "segmentation" frequently refers to the division of data into groups using straightforward methods; for instance, a picture might be separated into segments based just on pixel intensity and colour, or people can be grouped according to their wealth. However, cluster analysis has certain applications in the partitioning of graphs as well as in the segmentation of images and markets. Many disciplines, including anthropology, geology, chemistry, biology, food science and engineering, have used cluster analysis.

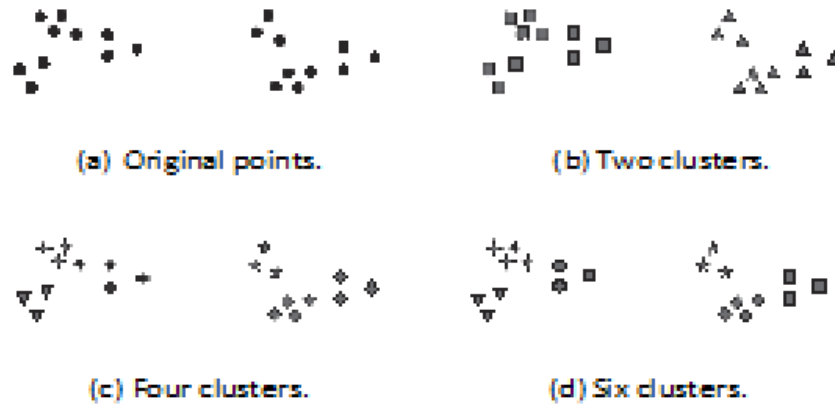


Figure 5.1. Different ways of clustering the same set of points. (Source: <https://www.studocu.com/in/document/>)

5.5 DIFFERENT TYPES OF CLUSTERS

- Finding useful groups of objects (clusters) is the goal of clustering, where usefulness is determined by the objectives of the data analysis. Unsurprisingly, there are a variety of cluster conceptions that are useful in actual practise. We employ two-dimensional points as our data objects in Fig. 5.2 in order to graphically depict the differences between these kinds of clusters. We emphasise that the types of clusters discussed here are equally applicable for data from other sources.
- **Well-Separated:** A cluster is a collection of objects where every object is more similar to every other object in the cluster than it is to any other object. It is occasionally necessary to define a threshold in order to ensure that all of the objects in a cluster are sufficiently similar to (or close to) one another. Only when the data comprises naturally occurring clusters that are spread apart from one another can this idealistic definition of a cluster be satisfied. A good-separated cluster is illustrated in Fig. 5.2(a) as two sets of points in a two-dimensional space. Any two points are separated from each other by a greater distance than any two points within the same group. The shape of well-separated clusters need not be spherical; they can take any form.
- **Prototype-Based:** A cluster is a group of objects where each one is more similar to the prototype that defines it than to the prototype of any other cluster. A cluster's centroid, or the average (mean) of all its points, serves as the model for clusters in data with continuous properties. The prototype, or most representative point of a cluster, is frequently a medoid when a centroid is not meaningful, such as when the data has categorical qualities. Prototype-based clusters are frequently referred to as center-based clusters since they can often be thought of as the most central place for many different forms of data. That such clusters are typically spherical is not surprising. Center-based

clusters are seen in Fig. 5.2(b).

- **Graph-Based:** A cluster can be defined as a connected component, that is, a group of objects that are connected to one another but that have no connections to objects outside the group, if the data is represented as a graph, where the nodes are objects and the links represent connections among objects. Contiguity-based clusters, in which two items are connected only if they are close enough to one another, are a key illustration of graph-based clusters. This suggests that every object in a contiguity-based cluster is closer to another object in the cluster than to any point in another cluster. Such clusters for two-dimensional points are illustrated in Fig. 5.2(c). This concept of a cluster is helpful when clusters are asymmetric or interwoven, but it can be problematic when noise is present because, as shown by the two spherical clusters in Fig. 5.2(c), a sparse point bridge can combine two separate clusters.
- There may be further kinds of graph-based clusters. One such method refers to a cluster as a clique, or a collection of nodes in a graph that are intimately connected to one another. In particular, a cluster is created when a group of items forms a clique if we build connections between them in the order of their proximity to one another. These clusters often have the same globular shape as prototype-based clusters.
- **Density-Based:** A cluster is a collection of objects that is dense in one area and low density in the area around it. Data obtained by introducing noise to the data in Figure 5.2(c) are displayed in certain density-based clusters in Figure 5.2(d). Due to the bridge connecting the two circular clusters fading into the noise, they are not combined as in Fig. 5.2(c). Similar to how it happens in Figure 5.2(c), the curve in Figure 5.2(d) also fades into the background noise and does not form a cluster. When clusters are asymmetric or entangled, noise, and outliers are present, a density-based definition of a cluster is frequently used. The data in Fig. 5.2(d) do not lend themselves to a contiguity-based definition of a cluster, however, as noise tends to generate bridges between clusters.
- **Shared-Property (Conceptual Clusters):** A cluster can also be defined more broadly as a group of items with a common attribute. This definition includes all preceding definitions of a cluster; for instance, items in a center-based cluster all have the characteristic of being located near the same centroid or medoid. New varieties of clusters are also included in the shared-property method, though. Think about the clusters in Figure 5.2(e). There are two interconnected circles (clusters) and a triangular area (cluster) next to a rectangle. In both situations, a clustering algorithm would require a very particular definition of a cluster in order to correctly detect these clusters. Conceptual clustering is the method used to find these clusters. We don't investigate more complex cluster types in this course because doing so would introduce us to the field of pattern recognition.

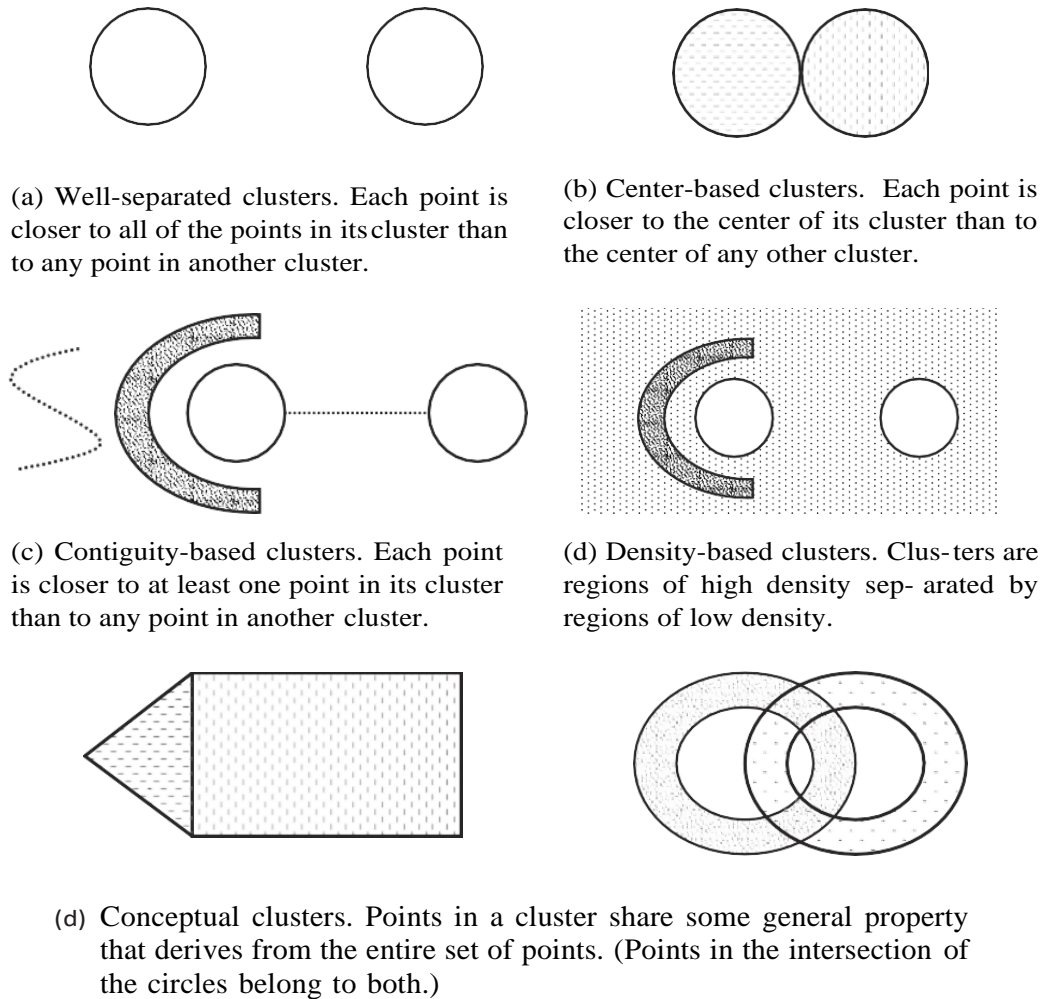


Fig. 5.2. Different types of clusters as illustrated by sets of two-dimensional points. Source:
<https://www.studocu.com/in/document/>

5.6 TYPES OF CLUSTERING

A clustering is a general term for an entire set of clusters. There are two main objectives for a good grouping: (i) similarity between one data point and another, and (ii) difference between those comparable data points and those that unquestionably, heuristically depart from those points.

There are different kinds of clustering algorithms to resolve one or many of these issues, including scalability, characteristics, dimensionality, border shape, noise, and interpretation. There are different types of clustering methods such as:

- Centroid-based / Partitioning (*K-means*) method
- Connectivity-based (*Hierarchical Clustering*) method
- Density-based Method (Model-based methods)

- Distribution-based method
- Fuzzy Clustering method
- Constraint-based Method (Supervised Clustering)

5.6.1 CENTROID-BASED / PARTITIONING (*K-MEANS*) METHOD

One of the most common approaches used by analysts to generate clusters is this one. This is often referred to as the Supervised Clustering technique. The features of the data points are used to divide the clusters in partitioning clustering. For this clustering technique, we must define how many clusters will be formed.

The simplest type of clustering used in data mining is centroid-based. It bases its operation on how closely the data points resemble the selected centre value (Fig. 5.3). The datasets are separated into a predetermined number of clusters, and each cluster is referenced by a vector of values. When compared to the vector value, the input data variable shows no difference and joins the cluster. The most important yet challenging step in the clustering approach is determining the number of clusters in advance.

The distances between the clusters and the characteristic centroids are iteratively measured by these clustering method groups using a variety of distance metrics. These are the Minkowski Distance, the Manhattan Distance, or the Euclidian Distance.

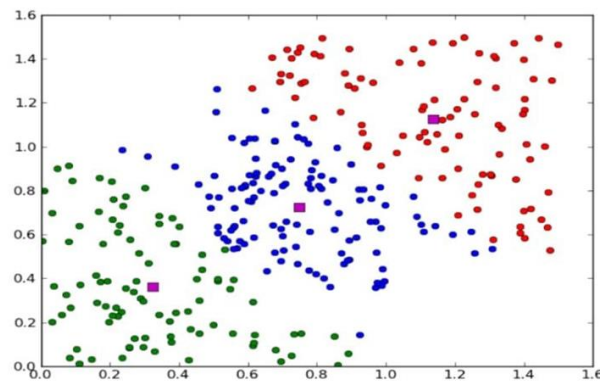


Fig. 5.3 Centroid-based clustering (partitioning method)

(Source: <https://www.analytixlabs.co.in/blog/types-of-clustering-algorithms>)

In order to reassign the data points across groups based on distance, these clustering algorithms employ an iterative procedure. Here are some examples of algorithms that fit this category:

a) **k-Means Algorithm**

This approach can be used to repeatedly create clusters out of the data collection. It is one of the iterative, unsupervised algorithms. This algorithm's primary goal is to locate the clusters while minimising the distance between the cluster and the data set. This process, often known as "Lloyd's algorithm," involves clustering m data sets into, say, k clusters, where each data set is assigned to the closest mean cluster.

Algorithm

1. Specify the number of identical centroids for the data points and the number of clusters (k) to be formed.
2. Each data point is assigned to the cluster with the shortest distance after the distance between it and each centroid is determined.
3. To process all the data points, repeat step one.
4. To determine the cluster's new centroid, the average of the data points in the cluster is computed.
5. Repeat Step 2 as necessary to generate as many clusters as required.

Due to the random selection of the original centroid, the ensuing clusters have a greater impact on them. K-means algorithm complexity is $O(tkn)$ where n is the total data set, k is the number of clusters produced, and t is the number of cluster iterations.

Advantages

- Simple process for implementation.
- Fit for massive databases.
- When compared to the hierarchical technique, spherical clusters yield more dense clusters.

Disadvantages

- An iterative run does not yield equivalent results.
- Inappropriate for clusters of varying sizes and densities.
- Unsuccessful for categorical data and non-linear data set.
- Outliers and noisy data are challenging to manage.
- Because of underlying causes, Euclidean distance measurements can have uneven weights.

b) k-Medoids or PAM (Partitioning Around Medoids)

With the assignment of the cluster's centre differing, the method is similar to that of the K-means clustering algorithm. PAM requires the medoid of the cluster to be an input data point, whereas K-means clustering allows the average of all cluster members to be a non-input data point.

According to this approach, each cluster is represented by an item that is close to the cluster centroid. Until the resulting cluster is enhanced, the procedure of replacing described objects with non-described objects is repeated. The value is forecast using the cost function, which calculates the difference between an object and its corresponding described object in a cluster.

The algorithm is put into practice in two steps:

Build: Initial medoids are innermost objects.

Swap: A function can be replaced by another function until it can no longer be reduced.

Algorithm

1. *Initialise the medoids from the given data set with m random points.*
2. *Determine the closest medoid by distance metrics for each data point.*
3. *Swapping costs are determined for each selected and unselected object and are presented as TCns, where s is a selected object and n is an unselected object.*
4. *If $TCns < 0$, s is replaced by n*
5. *Repeat steps 2 and 3 until the medoids remain unchanged*

Four characteristics to be considered are:

- *Shift-out membership:* Movement of an object from current cluster to another is allowed.
- *Shift-in membership:* Movement of an object from outside to current cluster is allowed.
- *Update the current medoids:* Current medoid can be replaced by a new medoid.
- *No change:* Objects are at their appropriate distances from cluster.

Advantages

- Simple comprehension and application procedure.
- Has a high rate of speed and can converge swiftly.
- Differences between the objects are permitted
- Less susceptible to outliers than k-means.

Disadvantages

- Different clusterings can be produced by initial sets of medoids. As a result, it is advised to execute the method multiple times with various initial sets.
- The clusters that are produced may vary depending on the units of measurement. Different-sized variables can be standardised.

5.6.2 CONNECTIVITY-BASED (HIERARCHICAL CLUSTERING) METHOD

Hierarchical clustering, also known as connectivity-based clustering, it is relies on the concept that, to varying degrees (of relationship), everything is connected to its neighbours. Extensive hierarchical structures are used to depict the clusters, which are spaced apart by the maximum distance needed to connect the cluster's component elements. Dendrograms are used to visualise the clusters, with the X-axis showing the items that do not merge and the Y-axis showing the distance at which clusters merge (Fig. 5.4). Similar data objects are clustered together closely, whereas different data pieces are positioned higher up the hierarchy. Among discrete characteristics relating to the multidimensional scaling, quantitative correlations among data variables or cross-tabulation in some aspects, mapped data items correlate to a Cluster.

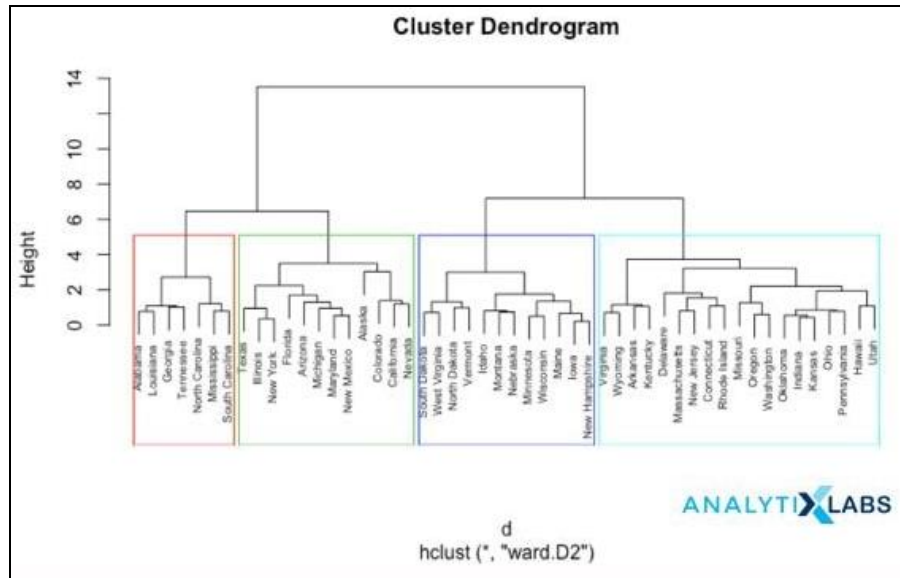


Fig. 5.4 Clusters are represented as dendrograms

(Source: <https://www.analytixlabs.co.in/blog/types-of-clustering-algorithms>)

Using this technique, a hierarchy is created from a group of data objects. We can categorise hierarchical approaches according to how the hierarchical breakdown is produced. Following are the two approaches;

- Agglomerative Approach
- Divisive Approach

a) Agglomerative Approach

Agglomerative strategy is very different from a divisive one. This method treats every single data point as a distinct cluster and then merges each cluster by taking into account how similar (distance) they are to one another until a single, huge cluster is obtained or a certain condition is met.

A number-based criterion (no more clusters after this point), a distance criterion (clusters shouldn't be too far apart to be merged), a variance criterion (increase in the variance of the cluster being merged shouldn't exceed a threshold, Ward Method), or any type of logic can be used.

Algorithm

1. Create N distinct clusters by initialising all n data points.
2. Combine the cluster pairs with the shortest (closest) distance into a single cluster.
3. Determine the pair-wise distance between the currently created cluster and the most-available clusters, which are the clusters at hand.
4. Carry out procedures 2 and 3 once again to combine all data samples into a single, sizable cluster of size N.

Advantages

- Nested clusters are simple to recognise, produce better outcomes, are simple to deploy, and are excellent for automation.
- Lessens the impact of the cluster's starting values on the clustering outcomes.
- Decreases the difficulty of computation in both time and space

Disadvantages

- Previous actions cannot ever be undone.
- The temporal complexity rises as a result of challenges handling various cluster sizes and convex shapes.
- The objective function's minimization is indirect.
- The dendrogram can occasionally make it challenging to determine the precise number of clusters.

b) Divisive Approach

This approach is also referred as the top-down approach (Fig. 5.5). Here, we treat the full sample of data as a single cluster and iteratively divide it into smaller clusters. It continues until all objects in a cluster have been reached or the termination condition is met. For categorical data, the metric can be the GINI coefficient within a cluster, or this termination logic can be based on the minimum sum of squares of error within a cluster. This approach is rigid since merging or splitting operations cannot be undone once they have been completed.

Algorithm

1. Start the process with a single cluster that contains all of the samples.
2. From the cluster containing the largest diameter, choose the largest cluster.
3. Locate the data point in the cluster that was discovered in step 2 that has the least average resemblance to the other components of that cluster.
4. The data samples discovered in step 4 are the first element to be added to the fragment group.
5. Find the component of the original group that shares the most similarities on average with the fragment group.
6. If the average similarity of element obtained in step 5 with the fragment group is greater than its average similarity with the original group then assign the data sample to the fragment group and go to step 5; otherwise do nothing;
7. Continue using steps 2 through 6 until every data point has been divided into its own cluster.

Advantage

- In some cases, it generates more precise hierarchies than a bottom-up approach.

Disadvantages

- Top down approaches require a second flat clustering method, making them

computationally more difficult than bottom up approaches.

- Different distance measures may produce various outcomes when used to measure the separation between clusters.

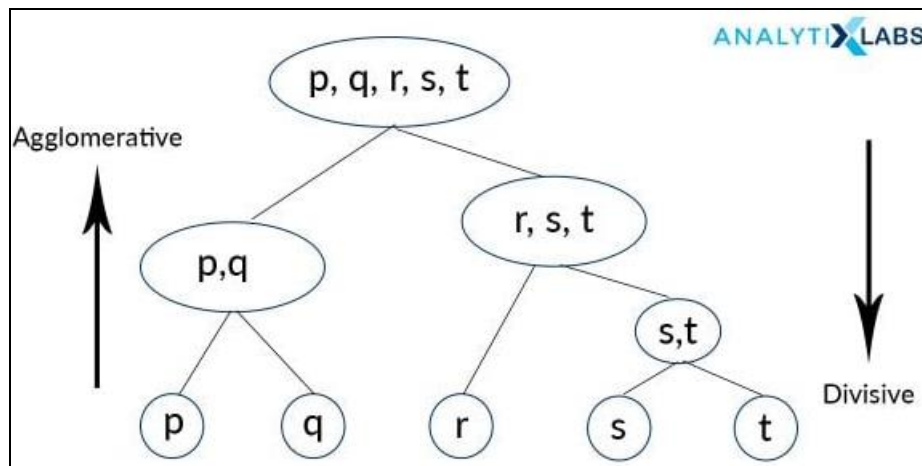


Fig. 5.5 Agglomerative and divisive approaches of hierarchical clustering

(Source: <https://www.analytixlabs.co.in/blog/types-of-clustering-algorithms>)

5.6.3 DENSITY-BASED METHOD (MODEL-BASED METHODS)

The first two approaches are based on a metric of distance (similarity/proximity). The density-based clustering approach prioritises density over distance. Data is organised into regions with high concentrations of data objects that are surrounded by low concentration regions. A maximum set of related data points are classified into the clusters that have created.

The main principle of density-based approaches is that for each point of a cluster, the neighbourhood of a given unit distance has at least a minimum number of points, i.e. the density in the neighbourhood should surpass some threshold. However, this theory is predicated on the notion that the clusters have spherical or uniform forms.

Due to similar density, the clusters created (Fig. 5.6) contain a maximum degree of homogeneity and come in a variety of arbitrary shapes and sizes. This clustering strategy efficiently accounts for noise and outliers in the datasets.

When performing most of the clustering, we take two major assumptions: the data is devoid of any noise and the shape of the cluster so formed is purely geometrical (circular or elliptical).

The fact is, data always has some extent of inconsistency (noise) which cannot be ignored. And we must not confine ourselves to a particular attribute shape. To avoid ignoring any data points, it is preferable to use arbitrary shapes. These are the applications where the value of density-based algorithms has been demonstrated.

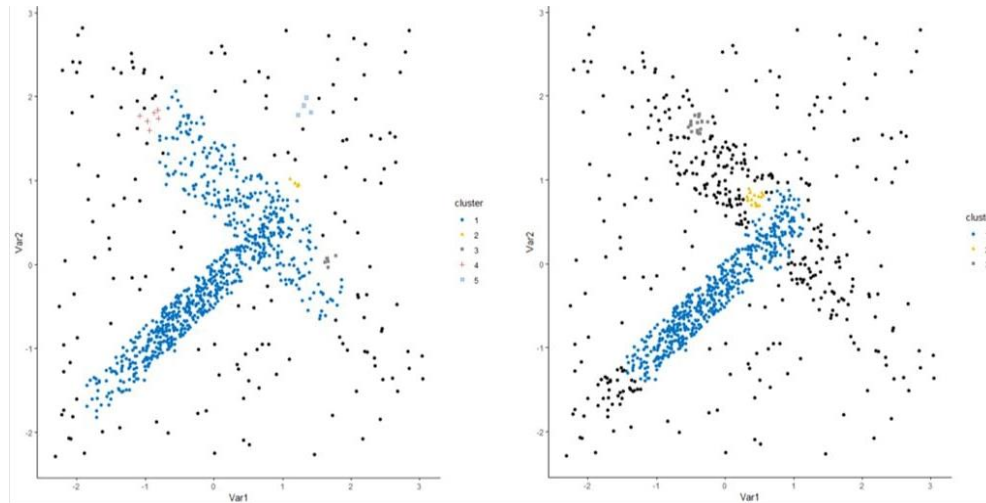


Fig. 5.6 Density-based clustering

(Source: <https://www.analytixlabs.co.in/blog/types-of-clustering-algorithms>)

a) DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

For addressing the noise and clusters with arbitrary shapes, DBSCAN was suggested to use density-reachability and density connectivity. An extremely dense set of data is referred to as a cluster in DBSCAN. DBSCAN takes two factors into account, including:

Eps: The maximum value of radius from its closest neighbours.

MinPts: The Eps is surrounded by data points that should be kept to a minimum, such as Eps-Neighborhood.

To define Eps-Neighborhood it should satisfy the following condition,

$$NEps(q) : \{ p \text{ belongs to } D | (p, q) \leq Eps \}.$$

Following are a few definitions that may help you understand density-based clustering:

- **Core point:** It is a point that falls within the user-specified Eps and MinPts. Additionally, that point is surrounded by dense neighborhood.
- **Border point:** It is a point that is adjacent to a core point; numerous core points might share the same border point, therefore this point lacks a dense neighbourhood.
- **Noise/Outlier:** It is point that does not belongs to cluster.
- **Direct Density Reachable:** A point p is directly Density Reachable from point q with respect to Eps, MinPts if point p belongs to NEps(q) and Core point condition i.e. $|NEps(q)| \geq MinPts$
- **Density Reachable:** A point p is said to Density Reachable from point q with respect to Eps, MinPts if there a chain points such as p_1, p_2, \dots, p_n , $p_1 = q$, $p_n = p$ such that $p_i + 1$ is directly reachable from p_n .

Algorithm

1. To begin forming clusters, think of a random point, let's call it point p.
2. The next step is to identify all sites that, in relation to Eps and MinPts, are densely reachable from point p. To create the cluster, the following criteria is checked.
 - a. If point p is found to be core point, then cluster is obtained.
 - b. If point p is found to be border point, then no points are density reachable from point p and hence visit the next point of database.
3. Once each point has been processed, repeat the process.

Advantages

- It can recognise Outlier.
- It is not necessary to predetermine the number of clusters.

Disadvantages

- Finding clusters is challenging if the data density is constantly changing.
- The user must specify the parameter in advance, and it is not suitable for data of high quality.

5.6.4 DISTRIBUTION-BASED CLUSTERING

The clustering methods that we are familiar with up to this point are based either on proximity (similarity/distance) or composition (density). A family of clustering algorithms considers probability, which is a whole separate measure.

Distribution-based clustering creates and groups data points based on their likely hood of belonging to the same probability distribution (Gaussian, Binomial, etc.) in the data.

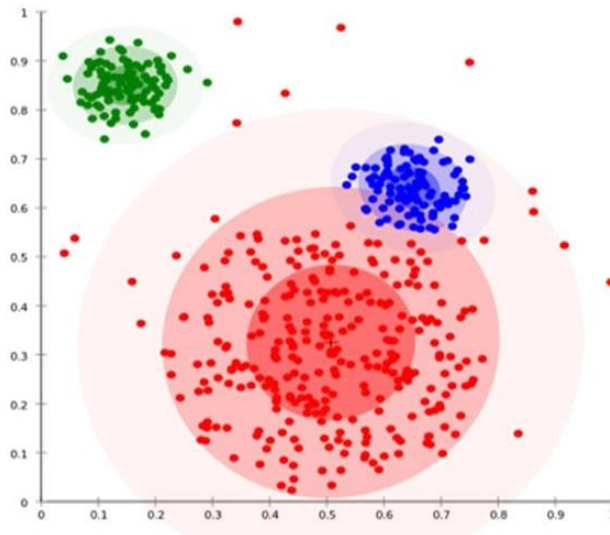


Fig. 5.7 Distribution-based clustering

(Source: <https://www.analytixlabs.co.in/blog/types-of-clustering-algorithms>)

The clustering of the data objects takes place using a probability-based distribution that makes use of statistical distributions. Data objects that have a greater chance of being in the

cluster are included. There is a centre point in each cluster, and the farther a data point is from the central point, the less likely it is to be included in the cluster.

A major drawback of density and boundary-based approaches is in *specifying the clusters a priori to some of the algorithms* and mostly the definition of the shape of the clusters for most of the algorithms. It is necessary to choose at least one tuning or hyper-parameter, and while doing so is straightforward, it could also have unintended consequences.

In terms of flexibility, accuracy, and shape of the clusters created, distribution-based clustering offers a clear advantage over proximity and centroid-based clustering algorithms. However, the main issue is that these clustering approaches only perform well with artificial or simulated data, or with data where the majority of the data points unquestionably belong to a preset distribution; otherwise, the results would overfit.

5.6.5 FUZZY CLUSTERING

Fuzzy clustering simplifies the **partition-based clustering method** by allowing a data object to be a part of more than one cluster. Based on the geographic probabilities, a weighted centroid is used in the process. Initialization, iteration, and termination are the steps, which result in clusters that can be best analysed as probabilistic distributions rather than by assigning labels rigidly.

In order for the process to function, all of the data points connected to each cluster centre are given membership values. The distance between the cluster centre and the data point is used to calculate it. It is more likely that an object will be in a certain cluster if its membership value is closer to the cluster centre.

Values related to membership and cluster centres are reorganised at the end of an iteration. When data points are unclear or partially between cluster centres, fuzzy clustering is used to address the problem. Choosing probability above distance is how you do this.

5.6.6. CONSTRAINT-BASED (SUPERVISED CLUSTERING)

The idea behind the clustering procedure is that the data can be split up into the ideal number of "unknown" groupings. All clustering algorithms' fundamental steps involve discovering those hidden patterns and parallels without human participation or set parameters. We might, however, be required to segment the data depending on certain requirements in some business cases. Here, machine learning approaches for clustering are used in a supervised manner.

A constraint is described as the desirable qualities of the clustering results or a user's expectation of the clusters so created. This can be expressed in terms of a definite number of clusters, the size of the clusters, or significant dimensions (variables) that are necessary for the clustering process.

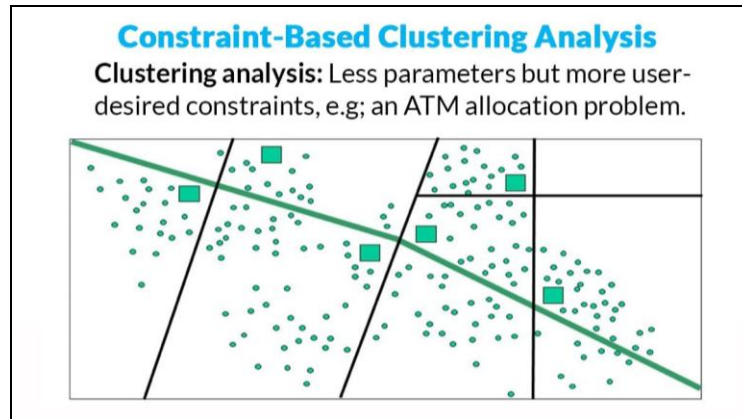


Fig. 5.8 Constraint based clustering (Source: <https://www.analytixlabs.co.in/blog/types-of-clustering-algorithms>)

Constraint-based clustering is frequently achieved using tree-based, classification machine learning algorithms like Decision Trees, Random Forest, Gradient Boosting, etc. Without the use of constraints or clustering labels, a tree is created by splitting. Using the appropriate methods and taking into account the limitations, the leaf nodes of the tree are then joined to create the clusters.

5.7 TYPES OF CLUSTERING ALGORITHMS

Clustering techniques are employed for investigating data, identifying anomalies, locating outliers, or seeing patterns in the data. Similar to neural networks and reinforcement learning, clustering is an unsupervised learning method. The accessible data is noisy, varied, and highly unstructured. Therefore, the choice of method depends on the appearance of the data. Finding useful industrial insights is made possible with the aid of an appropriate clustering method. Let's take a closer look at the various means of learning clustering techniques.

1. K-Means clustering
2. Mini batch K-Means clustering algorithm
3. Mean Shift
4. Divisive Hierarchical Clustering
5. Hierarchical Agglomerative clustering
6. Gaussian Mixture Model
7. DBSCAN

8. OPTICS
9. BIRCH Algorithm

5.7.1 K-MEANS CLUSTERING

A partition-based clustering method called K-Means employs the distance between Euclidean distances between the points as a criterion for cluster creation. Assuming there are 'n' numbers of data objects, K-Means groups them into a predetermined 'k' number of clusters.

Each cluster has a designated cluster centre, and they are all spread out further apart. Each new data point is assigned to the cluster with the nearest cluster centre. Once every data point has been assigned to a cluster, this process is repeated. The cluster centres, or centroids, are updated once all the data points have been examined.

After obtaining these 'k' new centroids, a fresh grouping is created between the points in the same data set and the closest new centroid. The k centroid values and their location may change with each iteration. This cycle repeats until there is no longer any movement of the centroids, or cluster centres. The objective function is to be minimised by the algorithm.

$$J = \sum_{j=1}^k \sum_{i=1}^n \|x_i^{(j)} - c_j\|^2$$

Where, J = Objective function; k = number of clusters; n = number of clusters; $x_i - C_j$ is the chosen distance between cluster center C_j and data point X_i

The Silhouette approach and Elbow method can be used to select the right value of K. For each data point, the distance is calculated using the mean intra-cluster distance and an average of the closest cluster distance. While the Elbow approach calculates the average distance using the sum of the squares of the data points.

Implementation: K-Means clustering algorithm

- Choose 'k' number of clusters and centroids for each cluster.
- Shuffle the data points in the dataset, then initialise the chosen centroid.
- Without replacing any data points, assign the clusters to them.
- Assign the clusters to the data points without replacement.
- Create new centroids by calculating the mean value of the samples.
- Reset the cluster centres until no further changes to the clusters occur.

5.7.2 MEAN SHIFT CLUSTERING ALGORITHM

Mean shift clustering is a nonparametric, straightforward, and adaptable clustering method. It is based on a technique called kernel density estimation, which estimates the essential distribution for a given dataset. The algorithm's fundamental idea is to iteratively allocate the data points to the designated clusters by shifting them in the direction of the peak or highest density of data points. It is applied during the image segmentation process.

Algorithm:

1. Assembling a cluster for each individual data point
2. Calculating the centroids
3. The new centroids' location should be updated
4. Iteratively shifting the data points to areas with higher densities.
5. When the centroids reach a fixed place and stop moving, the process ends.

5.7.3 GAUSSIAN MIXTURE MODEL

A clustering method based on distributions is the Gaussian Mixture Model (GMM). It is predicated on the idea that the data are distributed according to Gaussian distributions. It uses a clustering technique for statistical inference. The probability that a point will be a member of a cluster is inversely proportional to its distance from the distribution; the possibility that a point will be a member of the cluster diminishes as distance from the distribution rises. The GM model trains the dataset and makes a cluster assumption for each object. Data points are later used to generate a scatter plot, with various colours being allocated to each cluster.

Probabilities are determined and assigned to data points in the 'K' number of clusters using GMM. There are three factors for each one: mean, covariance, and mixing probability. GMM uses the **Expectation Maximisation Approach** to calculate these parameters.

The optimisation function starts the Gaussian parameters at random and determines whether the hypothesis fits into the selected cluster. The settings are then updated during the maximisation process to better fit the points into the cluster. The approach attempts to increase the probability that the data sample belongs to the cluster distribution, which indicates that cluster distributions have high peaks (closely related cluster data), and the mixture model captures the dominating pattern data objects by component distribution.

The optimisation function starts the randomly generated Gaussian parameters and verifies that the hypothesis fits into the selected cluster. Once the parameters have been updated to fit the points into the cluster, the maximisation phase is performed. The approach tries to increase the probability that the data sample belongs to the cluster distribution, which indicates that cluster distributions have high peaks (closely related cluster data), and the mixture model captures the dominating pattern data objects by component distribution.

5.7.4 DBSCAN

DBSCAN – Density-Based Spatial Clustering of Applications with Noise recognize distinct groupings in data, The algorithm aims to cluster the data as contiguous regions having high point density. The spots of low density that divide each cluster from the others. Simply said, the cluster includes data points that meet the density condition, which is the minimal number of data objects within a specified radius.

Terms which are used in DBSCAN are following:

- **Core:** Point having least points within the distance from itself
- **Border:** At least one core point exists at a given distance for a given point.
- **Noise:** Having less than m points are present within a particular distance. This location is neither the core nor the edge.
- **minPts:** A minimal number of points is the threshold value at which the cluster is considered to be dense.
- **eps ϵ :** a unit of measurement that is used to place the points in relation to other data points.
- **Reachability:** Density distribution identifying whether a point is reachable from other points if it lies at a distance ϵ from it.
- **Connectivity:** Establishes the location of any point within a given cluster using a transitivity-based chaining technique.

For implementing DBSCAN, we first begin with defining two important parameters – a radius parameter ϵ and a minimum number of points within the radius (m).

5.7.5 BIRCH ALGORITHM

Balanced Iterative Reducing and Clustering using Hierarchies, or BIRCH is a clustering method used for very large datasets. A quick algorithm that performs a single run through the full dataset. By concentrating on highly populated spaces and producing an accurate summary, it is committed to resolving the problems associated with huge dataset clustering.

BIRCH fits in with any provided amount of memory and minimizes the I/O complexity. Only metric attributes—those without categorical variables or those whose value can be represented by explicit coordinates in a Euclidean space—may be processed by the method. The CR tree and the threshold serve as the algorithm's primary inputs.

- *CF tree:* The clustering feature tree is a tree in which each leaf node consists of a sub-cluster. A CF tree contains a pointer to each child node in each entry. The sum of the CF entries in the child nodes constitutes the CF entry.
- *Threshold:* A maximum number of entries in each leaf node

Steps of BIRCH Algorithm

- Step 1- Building the Clustering feature (CF) Tree: Building small and dense regions from the large datasets. Optionally, in phase 2 condensing the CF tree into further small CF.
- Step 2 – Global clustering: Applying the clustering technique to the CF tree's leaf nodes.
- Step 3 – Refining the clusters, if necessary.

5.8 APPLICATIONS OF CLUSTERING

In several disciplines, clustering is used to prepare the data for different appliance learning techniques. Some of the applications of clustering are as follows.

1. **Social network analysis:** Graph Theory and network theory are used to examine both qualitative and quantitative social configurations. In order to learn more about different roles and groupings in the network, it is necessary to cluster individuals and monitor how they interact.

2. **Data processing and feature weighing:** Cluster IDs can be used to represent the data. Storage is reduced, and the feature data is made simpler. You can get the data using the demographics, date, and time.

3. **Image compression:** Clustering reduces the image size without sacrificing image quality, assisting in the compression of the photos for storage.

4. **Market segmentation:** To better understand their target customer, businesses must divide their market into smaller segments. In order to create recommendations that are similar, clustering brings like-minded people together in the same neighbourhood. This aids in the formation of patterns and insights.

5. **Network traffic classification:** Network traffic source characteristics are grouped together in clustering. To categorise the different traffic kinds, clusters are constructed.

Accurate knowledge of traffic sources helps in capacity planning and site traffic growth.

6. **Identifying good or bad content:** By utilising factors like source, keywords, and content, clustering efficiently separates out phoney news and identifies frauds, spam, or rough content.

7. **Life science and Healthcare:** Plant and animal taxonomies are developed through clustering to arrange genes with similar activities. Using medical picture segmentation, it is also employed in the detection of malignant cells.

8. **Retail marketing and sales:** Clustering is a technique used by marketing to analyse consumer purchasing patterns and control the supply chain and suggestions. It groups individuals with like features and likelihood of purchasing. It aids in connecting with the right customer segments and offers efficient promotions.

9. **Regulating streaming services:** Clustering recognising viewers with comparable behaviour and interests. In order to divide users into high and low usage categories, Netflix and other OTT services classify users based on criteria including genre, daily watching time, and total viewing

sessions. In doing so, it makes it easier to display adverts and provide relevant user recommendations.

8. Tagging suggestions using co-occurrence: Understanding the search behavior and giving them tags when searched repeatedly. Taking an input for a data set and maintaining a log each time the keyword was searched and tagged it with a certain word. A similarity measure can be used to cluster the number of times two tags are present together.

5.9 MULTIVARIATE ANALYSIS: INTRODUCTION, CHARACTERISTICS AND APPLICATIONS

You are used to the statistical methods used in univariate (one variable) and bivariate (two variables) data analysis from previous Units. You are probably aware that the majority of real-world events in the social sciences, particularly in economics, cannot be well explained by the statistical methods employed in univariate or bivariate data analysis. Most frequently, more than two factors interact and reinforce one another in such circumstances, making the explanation of such scenarios more difficult.

As a result, over the past 50 years, numerous statisticians have worked to develop a variety of multivariate methodologies. In numerous sectors today, including economics, sociology, psychology, agriculture, anthropology, biology, and medicine, similar methodologies are being used. These methods are employed for assessing social, psychological, medical, and economic data, particularly when the variables pertaining to research studies in these domains are intended to be associated with one another and when rigorous probabilistic models cannot be effectively applied. Due to the development of high speed electronic computers, the usage of multivariate approaches in practise has increased recently.

Multivariate approaches can examine complex data because they are essentially empirical, deal with reality, and are grounded in science. As a result, for actual outcomes in the majority of practical and behavioural inquiries, we typically use multivariate analysis approaches. Multivariate approaches are a tool for data analysis, but they can aid in other kinds of decision-making. Consider the situation of the college entrance exam, when students are given a series of tests and are only admitted if they have good overall results across multiple courses. Despite appearing fair, this system may occasionally be biased in favour of the same subjects with larger standard deviations. Multivariate approaches may be utilised effectively in these circumstances to establish rules regarding who should be admitted to colleges. We may also use a case study from the medical industry. Patients undergo a variety of medical exams, including blood pressure and cholesterol checks. Each of these tests' results has significance on its own, but it's also crucial to take into account correlations between them or results from the same test at various times in order to make the accurate diagnoses and choose the best course of treatment. Multivariate methods can help us in this circumstance.

In context of all of this, we are able to state that "using a suitable multivariate statistical technique is the best strategy of data analysis if the researcher is interested in making probability statements on the basis of sampled multiple measurements." Multivariate analysis is briefly explained in this unit's part. It emphasises that the design and strategy for data gathering for decision-making and problem-solving will be influenced by multivariate analysis methods in addition to the analytical aspects of the research. Although multivariate procedures share many traits with its univariate or bivariate complements, there are a few significant distinctions that emerge when moving from a univariate to a multivariate study.

This unit's section offers a classification of multivariate strategies to show how this shift is made. However, because this course is introductory in nature and does not allow us to investigate all aspects of multivariate analysis, the topic is restricted to a full explanation of multiple regression and correlation as well as analysis of variance. It then provides general suggestions for employing these methods as well as the structural method of formulation, estimation, and result interpretation.

5. 9.1 MULTIVARIATE ANALYSIS

All statistical methods that evaluate several measures on the subjects or objects under study all comes under the category of multivariate analysis. Multivariate is often used by academics to refer to the process of analysing the relationships between or among more than two variables. Some people exclusively use this phrase to refer to issues with multiple variable difficulties where it is assumed that each variable has a multivariate normal distribution. But for a model to be regarded as really multivariate, every variable must be random and connected in such a way that the effects of the various variables on one another cannot be effectively assessed and evaluated individually.

According to some authors, the goal of multivariate analysis is to quantify, clarify, and forecast the strength of the relationships between variates (weighted combinations of variables). As a result, rather than just being about the quantity of variables or observations, the multivariate feature is about the various variates (different combinations of variables). Because we think that understanding multivariate procedures is a crucial first step in recognising multivariate analysis, multivariate analysis will be used to describe both multivariate approaches for the purposes of this Unit.

5.9.2 SOME BASIC CONCEPTS OF MULTIVARIATE ANALYSIS

Although univariate and bivariate statistics are used as a basis for multivariate analysis, the multivariate domain expansion brings new ideas and raises concerns of particular importance. These ideas extend from the fundamental notion of a variate to the specific problems relating to the many measuring scales employed, as well as the statistical problems of significance tests and confidence levels.

The Variate: The variate, a linear combination of variables with empirically determined weights, serves as the fundamental unit of multivariate analysis. While the weights are decided using a multivariate technique to achieve a certain goal, the variables are chosen by the researchers. Mathematically, a variate comprising n weighted variables (X_1 to X_n) can be written as:

$$\text{Variate value} = w_1X_1 + w_2X_2 + w_3X_3 + \dots + w_nX_n$$

The outcome is a single value that combines all of the variables to best meet the goal of the particular multivariate study. For instance, the variate in a multiple regression analysis is chosen to optimise the correlation between the various independent variables and the dependent variable. In discriminant analysis, a variate is built to produce scores for every observation that best distinguish across sets of observations. Each example's variate effectively conveys the multidimensional nature of the analysis. As a result, the variate serves as the main analytical focus in our discussion of approaches.

Measurement Scales: The measuring scale is crucial in establishing whether multivariate procedures are better suited to the data, taking into account both dependent and independent factors.

Measurement Error and Multivariate Measurement: It specifically presupposes that every observation consists of the real value plus some error, which includes random error value and systematic error. Any elements that randomly alter the measurement of the variable throughout the sample are the source of random error. For instance, a person's mood can either enhance or detract from their performance on any given occasion. Any elements that consistently influence how the variable is measured across the sample are considered systemic errors. In addition to exercising extreme caution to minimise these errors, researchers will occasionally decide to create multivariate measurements, also known as summed scales, in which they repeat the observation a certain number of times and take the average, or they will combine similar responses obtained using different variables to create a composite measure commonly known as an indicator.

Since they are a part of the observable variables, measurement error and low dependability have an indirect effect that cannot be detected. Although measurement error is not necessarily the cause of poor results, it is certain to distort the observed associations and make multivariate procedures less effective.

5.10 CLASSIFICATION OF MULTIVARIATE TECHNIQUES

There are currently many different multivariate approaches, which can be simply divided into two main types.

- Dependence methods and
- Interdependence methods.

These types of classification depend upon the question: Are some of the concerned variables dependent upon others? If the answer is 'yes', we have dependence methods; but in case the answer is 'no', we have interdependence methods. Two further questions are relevant for a better comprehension of the nature of multivariate techniques. Assuming that certain variables are dependent, the first thing to ask is how many of them are. The other query determines whether the data is metric or not. This refers to whether the data are qualitative, obtained on a nominal or ordinal scale, or quantitative, collected on an interval or ratio scale. The best course of action to take in each situation depends on the answers to each of these questions.

Thus, there are two distinct types of multivariate techniques: one type for data that contains both dependent and independent variables, and the other type for data that has numerous variables without a link of dependency. The first category includes multiple regression, multiple discriminant, multivariate analysis of variance, and canonical analysis; the second category includes factor, cluster, multidimensional scaling (MDS) (both metric and non-metric), and latent structure analysis.

5.11 TYPES OF MULTIVARIATE TECHNIQUES

According to the classification scheme outlined in the preceding section, multivariate analysis is an ever-expanding set of data analysis techniques that can be applied in a number of different research organizations. The following are some of the most popular and up-and-coming techniques:

- 1) Principal Components and Common Factor Analysis
- 2) Multiple Regression and multiple correlation
- 3) Multiple discriminant analysis and logistic regression
- 4) Canonical correlation analysis
- 5) Multivariate analysis of variance
- 6) Conjoint Analysis
- 7) Cluster Analysis
- 8) Perceptual mapping
- 9) Correspondence analysis
- 10) Structural equation modelling

5.11.1 Principal Components and Common Factor Analysis

Factor analysis, which includes both principal component analysis and factor analysis, is a statistical method that may be used to study the interactions between several variables and to explain such correlations in terms of their shared underlying dimensions (actors). Finding a

means to reduce the number of original variables while preserving as much of their information as possible is the goal of factor reduction. Factor analysis transforms into an impartial foundation for developing summed scales by offering an empirical approximation of the structure of the variables taken into account.

5.11.2 Multiple Regression

We described the ideas of correlation and basic linear regression when talking about the bivariate analysis. One independent variable, x , and one dependent variable, y , make up the regression equation in simple linear regression.

$$Y_e = a + bx$$

There are several independent variables and one dependent variable in multiple regression, and the equation is

$$y = a + b_1x_1 + b_2x_2 + b_3x_3 + \dots + b_nx_n$$

where x_1, x_2, \dots, x_n are the independent variables. If there is a substantial correlation between the independent variables and the dependent variable, it can be calculated using a multiple regression correlation R . Multiple regression analysis is employed when a statistician believes that the dependent variable's variance is caused by a number of independent factors.

The accuracy of predictions for the dependent variable can then be improved over just one independent variable alone using this methodology. When a store manager wants to know whether advertising expenditures and floor space devoted to displays have an impact on product sales, as well as when a sociologist wants to know whether children's television and video game viewing habits are related to their weight, multiple regression analysis can be used.

Similar to basic regression, multiple regression makes the following assumptions:

- 1) The values of the y variable have a normal distribution for any given independent variable value. (This is referred to "normality assumption".)
- 2) For every value of the independent variable, the variances (or standard deviations) for the y variables are the same. (This is known as the equal-variance hypothesis.)
- 3) The dependent variable and the independent variables are correlated linearly. (This is referred to as the linearity assumption.)
- 4) There is no correlation between the independent variables. (This is known as the non-multicollinearity assumption.)
- 5) The values for the y variables are independent. (This is called the independence assumption.)

A correlation coefficient is used in both simple and multiple regression to determine the strength of the relationship between the independent and dependent variables. This multiple correlation coefficient is symbolised by R^2 . The value of R^2 can range from 0 to +1, R^2 can never be negative. Closer the R is to +1, the stronger the relationship; the closer to 0 the weaker the

relationship. The individual correlation coefficient values can be used to calculate the value of R^2 , which accounts for all of the independent variables. The following is the formula for the multiple correlation coefficients when there are two independent variables:

$$R = \sqrt{\frac{r_{yx_1}^2 + r_{yx_2}^2 - 2r_{yx_1} \cdot r_{yx_2} \cdot r_{x_1x_2}}{1 - r_{x_1x_2}^2}}$$

$$R = \frac{\text{Variability explained}}{\text{Total variability}} = \frac{\sum(y_e - \bar{y})^2}{\sum(y - \bar{y})^2}$$

The degree of variance that the regression model can account for is measured by R^2 , which is the coefficient of multiple determinations. The quantity of residual variation, also known as mistake variation, is represented by the equation $1 - R^2$.

The significance of R is evaluated using the F-test. The following is the theory:

$$H_0: \rho = 0 \quad \text{and} \quad H_1: \rho \neq 0$$

Where ρ represents the population correlation coefficient for multiple correlation

The formula for the F-test is

$$F = \frac{R^2/k}{(1 - R^2)/(n - k - 1)}$$

Where n is the number of data groups (x_1, x_2, \dots, y) and k is the number of independent variables

The degrees of freedom are d.f.N. = $n - k$ and d.f.D. = $n - k - 1$.

Adjusted R^2

Since the value of R^2 is dependent on n (the number of data pairs) and k (the number of variables), statisticians also calculate what is called an adjusted R^2 , denoted by R^2_{adj} . This is based on the number of degrees of freedom. The formula for R^2_{adj}

$$R^2_{\text{adj}} = 1 - \frac{(1 - R^2)(n - 1)}{n - k - 1}$$

The adjusted R^2 is less than R and accounts for the possibility that, when n and k are nearly equal, the value of R may be inflated due to sampling error rather than a genuine link between the variables. This happens because the regression equation is derived using all of the variables' random variations in combination with one another. Even when the individual correlation coefficients for each independent variable and the dependent variable were all equal to zero, the multiple correlation coefficient resulting from sampling error can be larger than zero. Because of this, a multiple regression study typically reports both R^2 and R^2_{adj} .

5.11.3 Multiple Discriminant Analysis (MDA) and Logistic Regression

The study of relationships between various factors and a categorical result can be done using multivariate statistical approaches like logistic regression and linear discriminant analysis. Both approaches have been widely used in study, particularly in the fields of medicine and sociology.

MDA is the suitable multivariate technique if the single dependent variable is dichotomous (e.g. male and female) or multi-chotomous (eg. high-medium-low) and therefore nonmetric.

A binary dependent variable and one or more nominal, ordinal interval, or ratio-level independent variables are analysed using logistic regression to describe the data and to explain the relationship between them. with example, how does the probability of developing lung cancer alter (yes/no) with every additional pound of weight and each pack of cigarettes smoked each day?

Or, do factors like participant age, body weight, calorie intake, and fat intake affect the possibility of heart attacks (yes/no)?

5.11.4 CANONICAL CORRELATION ANALYSIS

An approach for calculating the linear dependence between two sets of data is called canonical correlation analysis (CCA). To put it more specifically, CCA identifies a linear combination of variables in one data set and another linear combination of variables in a different data set, with the goal of maximising the correlation between the two linear combinations. A canonical correlation is the resultant maximised correlation that extends the meaning of correlation to more than two variables.

5.11.5 MULTIVARIATE ANALYSIS OF VARIANCE (MANOVA)

A t-test is used to determine if two groups' means are likely to have been taken from the same sampling distribution of means. Using an ANOVA, you may determine whether the means for two or more groups were drawn from the same sampling distribution. Hotelling's T² is the t test's equivalent for multivariate data. The Hotelling's T² test determines if the two mean vectors for the two groups are samples from the same sampling distribution. The goal of MANOVA is to determine if the mean vectors for two or more groups were drawn from the same distribution of samples or not. The likelihood of selecting two random vectors of means from the same hat will be measured by Hotelling's T, just as the likelihood of selecting two random vectors of means from different hats will be measured by MANOVA.

The two main applications of MANOVA are as follows. The first occurs when there are several dependent variables that are correlated and the researcher prefers to conduct a single, comprehensive statistical test on this collection of variables rather than multiple separate tests. Examining how independent variables affect certain patterns of response on the dependent variables is the second, and occasionally more important, goal.

For example, suppose a researcher is interested in the effect of different types of treatments on several types of anxieties: test anxiety, anxiety in reaction to minor life stresses, and so-called free-floating anxiety. The IV is different treatment with three levels (desensitisation, relaxation training, and a waiting-list control). After random assignment of subjects to treatments and a subsequent period of treatment, subjects are measured for test anxiety, stress anxiety, and free floating anxiety. Scores on all three measures for each subject serve as DVs. The three anxiety measures are combined to determine whether they change as a result of treatment using MANOVA.

MANOVA and discriminant analysis are statistically equivalent. Only the emphasis is different between the techniques. The mean differences and statistical significance of group differences are highlighted by MANOVA. Discriminant analysis places a strong emphasis on predicting group membership and the characteristics that distinguish groups.

5.11.6 CONJOINT ANALYSIS

Conjoint Analysis is a method developed since the 1970s that enables organisations to identify and quantify the hidden principles consumers use to compare and contrast various goods and services as well as the values they place on certain aspects or component components of the offer. You can determine the sweet spot or ideal level of features and services that balance the value to the customer against the cost to the business and forecast potential demand or market share in a competitive market environment by understanding precisely how people make decisions and what they value in your products and services.

5.11.7 CLUSTER ANALYSIS

The Cluster analysis is an exploratory technique used to find patterns in the data. Other names for cluster analysis include segmentation analysis and taxonomy analysis. More specifically, it looks for homogeneous groupings of examples, such as observations, participants, and responders. If the grouping of the cases was not previously known, cluster analysis was employed to find the groupings. As an exploratory study, it does not distinguish between dependent and independent factors.

5.11.8 PERPETUAL MAPPING

A perceptual map is a type of visual approach used to illustrate how the typical target market consumer perceives the placement of rival products in the market. In other words, it is a tool that aims to represent in a diagram the consumer's views and understandings. The word "perceptual" derives from the word "perception," which essentially relates to the consumer's comprehension of the competing products and their associated features.

The definitions most frequently employed in analyses are:

"Perceptual maps measure the positioning of products in consumers' minds and display these perceptions on a graph whose axes are formed by product attributes." A perceptual map uses a

visual display to geographically reflect the perceptions and preferences of customers. You'll notice that both definitions emphasise that the data is visually shown on a graph or other display.

5.11.9 CORRESPONDENCE ANALYSIS

Cross tabulations, often known as cross tabs or contingency tables, are shown graphically using the statistical approach known as correspondence analysis. Cross tabulations are necessary whenever it is possible to group events into two or more distinct sets of categories, such as the product and location of purchases in market research or the symptom and course of therapy in medical testing.

5.11.10 STRUCTURAL EQUATION MODELLING (SEM)

A method called structural equation modelling (SEM) enables distinct relationships for every dependant variable in a collection. SEM is the most appropriate and effective estimation method for a number of distinct multiple regression equations evaluated concurrently.

Numerous real-world problems can be solved using multivariate techniques, which highlight their diversity. But in order to apply it effectively, one must have an understanding of each the techniques themselves and the contexts in which they are to be applied. As a result, it is crucial to develop a clear connection between the scenario and the approaches that should be employed in it.

5.12 SUMMARY

Clustering is a fundamental part of data mining and machine learning. It divides the datasets into groups with comparable traits so you can forecast user behaviour more accurately. In this unit, explanations of several clustering methods assist you in forming the ideal groupings of data objects. This strong base of clustered data can support a limitless number of opportunities.

A clustering is a general term for an entire set of clusters. There are two main objectives for a good grouping: (i) similarity between one data point and another, and (ii) difference between those comparable data points and those that unquestionably, heuristically depart from those points.

There are different kinds of clustering algorithms to resolve one or many of these issues, including scalability, characteristics, dimensionality, border shape, noise, and interpretation. There are different types of clustering methods such as Centroid-based / Partitioning (*K-means*); method; Connectivity-based (*Hierarchical Clustering*) method; Density-based Method (Model-based methods); Distribution-based method; Fuzzy Clustering method and Constraint-based Method (Supervised Clustering).

Clustering techniques are employed for investigating data, identifying anomalies, locating outliers, or seeing patterns in the data. Similar to neural networks and reinforcement

learning, clustering is an unverified learning method. The accessible data is noisy, varied, and highly unstructured. Therefore, the choice of method depends on the appearance of the data. Finding useful industrial insights is made possible with the aid of an appropriate clustering method. There are various means of learning clustering techniques such as K-Means clustering; Mini batch K-Means clustering algorithm; Mean Shift; Divisive Hierarchical Clustering; Hierarchical Agglomerative clustering; Gaussian Mixture Model; DBSCAN; OPTICS and BIRCH Algorithm.

In several disciplines, clustering is used to prepare the data for different appliance learning techniques. Some of the applications of clustering are social network analysis; data processing and feature weighing; image compression; market segmentation; network traffic classification; identifying good or bad content; life science and healthcare etc.

The majorities of the time, more than two factors interacts and reinforce one another in real-world situations. Univariate or bivariate data analysis techniques are insufficient to deal with this kind of issue. Data analysis may manage the analysis of several variables by using multivariate approaches. Although univariate and bivariate statistics are used as the basis for multivariate analysis, some new concepts, such as measurement scales, weight assignments, measurement error, and multivariate measurement, are introduced in the multivariate domain. Multiple regression, multiple discriminant analysis, multivariate analysis of variance, factor analysis, cluster analysis, correspondence analysis, latent structure analysis, structural equation modelling, etc. are among the techniques used in multivariate analysis.

The main goal of multivariate techniques is to represent a collection of large amounts of data in a more straightforward manner; in other words, multivariate techniques reduce the number of composite scores from a large number to a manageable number so that they can extract as much information from the raw data collected for a research study as possible. In order to arrange a great amount of complicated information included in the actual data into a simplified visible form, this is the main contribution of these approaches.

5.13 REFERENCES

- <https://www.analytixlabs.co.in/blog/types-of-clustering-algorithms/>.
- [https://www.studocu.com/in/document/lovely-professional-university/corporate-strategy/clustering notes /11981275](https://www.studocu.com/in/document/lovely-professional-university/corporate-strategy/clustering-notes/11981275).
- B.FJ. (1994) Multivariate Statistical Methods-A Primer, Chapman and Hall, London
- Everitt, B.S. and Dunn, G. (2001). Applied Multivariate Data Analysis, Arnold, London.
- Everitt, B.S., Landau, S. and Leese, M. (2001), Cluster Analysis, Fourth edition, Arnold.
- Harris RJ. (1985). A Primer in Multivariate Statistics, Academic Press, New York. Manly. Hair,
- J. F. Jr. (1995) Multivariate Data Analysis, 4th ed. Prentice-Hall.

- Johnson, Richard A.; Wichern, Dean W. (2007). Applied Multivariate Statistical Analysis, 6th edition,. Prentice Hall.
- Hamerly, G. and Elkan, C.(2002) Alternatives to the k-means algorithm that find better clusterings. In Proc. of the 11th Intl. Conf. on Information and Knowledge Management, pages 600–607, McLean, Virginia, ACM Press.
- Hardle, W., Simar, L (2007). Applied Multivariate Statistical Analysis Springer.
- Manly, B.F.J. (2005), Multivariate Statistical Methods: A primer, Third edition, Chapman and Hall.
- Rencher, A.C. (2002), Methods of Multivariate Analysis, Second edition, Wiley.
- Sharma, S. (1996). Applied Multivariate Techniques, University of South California, John Wiley & Sons, Inc.
- Tabachnick B., Fidell, L. (2007). Using Multivariate Statistics, 5th edition Pearson Education. Inc.

5.14 SUGGESTED READINGS

- <https://egyankosh.ac.in/bitstream/123456789/89136/1/Unit-11.pdf>.
- <https://egyankosh.ac.in/handle/123456789/77325>.
- Data Mining: Concepts and Techniques, 3rd Edition, Jiawei Han, Micheline Kamber, Jian Pei, Elsevier, 2012.
- Data Mining Techniques and Applications: An Introduction, Hongbo Du, Cengage Learning, 2013.

5.15 SELF ASSESSMENT QUESTIONS

1. Clustering islearning
 - e) Supervised
 - f) Unsupervised
 - g) Both a & b
 - h) None of Above
2. Which of the following is cluster analysis?
 - a) Grouping similar objects
 - b) Labeled Classification
 - c) Query Results Grouping
 - d) Simple segmentation
3. Which of the following is required by K-means clustering?
 - a) Assign the clusters to the data points without replacement
 - b) Defined distance metric

- c) Initial guess as to cluster centroids
 - d) All of the above
4. What are the Two Types of Hierarchical Clustering Analysis?
- a) Top-down clustering (Divisive)
 - b) Bottom-top clustering (Agglomerative)
 - c) Dendrogram
 - d) Both a & b
5.analysis used to study the interactions between several variables and to explain such correlations in terms of their shared underlying dimensions.
- a) PCA
 - b) MNOVA
 - c) CCA
 - d) MDA

Answers Key: 1-b, 2-a, 3-d, 4-d, 5-a.

5.16 TERMINAL QUESTIONS

5.16.1 Short answer type questions:

- a) Give a brief note on applications of cluster analysis.
- b) Write a short note on types of clusters.
- c) Describe in brief about the types of clustering
- d) Write on some basic concepts of multivariate analysis

5.16.2 Long answer type questions:

- a) What is cluster analysis? Describe its basic concept and algorithms.
- b) Give a general account on various clustering methods along with their description, advantages, disadvantages and algorithms available
- c) Describe the uses of cluster analysis in data mining.
- d) Define the term multivariate analysis. Explain its introduction, characteristics and applications.
- e) Give a detail emphasis on types of multivariate techniques.

BLOCK-3- HYPOTHESIS TESTING AND EXPERIMENTAL DESIGNS

UNIT-6- TESTING AND TESTS OF HYPOTHESIS

Contents

- 6.1- Objectives
- 6.2- Introduction
- 6.3- Testing of Hypothesis
 - 6.3.1- Basic concepts of testing of hypothesis
 - 6.3.2- Procedure for testing of hypothesis/test of significance
- 6.4- Chi-square test
- 6.5- T-test
- 6.6- Z-test
- 6.7- Tukey's Q Test
- 6.8- Summary
- 6.9- Glossary
- 6.10- Self assessment Questions
- 6.11- References
- 6.12- Suggested readings
- 6.13- Terminal Questions

6.1 OBJECTIVES

After reading this chapter, you should be able to:

- Identify the four steps of hypothesis testing.
- Define null hypothesis, alternative hypothesis, level of significance, test statistic, p value, and statistical significance.
- Define Type I error and Type II error, and identify the type of error that researchers control.
- Distinguish between a one-tailed and two-tailed test.
- Define power of a test, Chi-square test, t -test, Z test, Tukeys Q test and its applications

6.2 INTRODUCTION

A researcher or experimenter has always some fixed ideas about certain population(s) vis-à-vis population parameters(s) based on prior experiments/sample surveys or past experience. It is therefore desirable to ascertain whether these ideas or claims are correct or not by collecting information in the form of data through conduct of experiment or survey. In this manner, we come across two types of problems, first is to draw inferences about the population on the basis of sample data and other is to decide whether our sample observations have come from a postulated population or not. In this lesson we would be dealing with the second type of problem. In the ensuing section we would provide concepts and definitions of various terms used in connection with the testing of hypothesis.

6.3 TESTING OF HYPOTHESIS

The inductive inference is based on deciding about the characteristics of the population on the basis of a sample. Such decisions involve an element of risk, the risk of wrong decisions. In this endeavor, modern theory of probability plays an important role in decision making and the branch of statistics which helps us in arriving at the criterion for such decisions is known as testing of hypothesis. The theory of testing of hypothesis was initiated by J. Neyman and E.S. Pearson. Thus, theory of testing of hypothesis employs various statistical techniques to arrive at such decisions on the basis of the sample theory. We first explain some fundamental concepts associated with testing of hypothesis.

6.3.1 Basic Concepts of Testing of Hypothesis

6.3.1.1 Hypothesis

According to Webster Hypothesis is defined as tentative theory or supposition provisionally adopted, explain certain facts and guide in the investigation of others e.g. looking at the cloudy weather, the statement that 'It may rain today' is considered as hypothesis.

6.3.1.2 Statistical hypothesis

A statistical hypothesis is some assumption or statement, which may or may not be true about a population, which we want to test on the basis of evidence from a random sample. It is a definite statement about population parameter. In other words, it is a tentative conclusion logically drawn concerning any parameter of the population. For example, the average fat percentage of milk of Red Sindhi Cow is 5%, the average quantity of milk filled in the pouches by an automatic machine is 500 ml.

6.3.1.3 Null hypothesis

According to Prof. R. A. Fisher 'A hypothesis which is tested for possible rejection under the assumption that it is true is usually called Null Hypothesis' and is denoted by H_0 . The common way of stating a hypothesis is that there is no difference between the two values, namely the population mean and the sample mean. The term 'no difference' means that the difference, if any, is merely due to sampling fluctuations. Thus, if the statistical test shows that the difference is significant, the hypothesis is rejected. A statistical hypothesis which is stated for the purpose of possible acceptance is called Null Hypothesis. To test whether there is any difference between the two populations we shall assume that there is no difference. Similarly, to test whether there is relationship between two variates, we assume there is no relationship. So a hypothesis is an assumption concerning the parameter of the population. The reason is that a hypothesis can be rejected but cannot be proved. Rejection of no difference will mean a difference, while rejection of no relationship will imply a relationship. For example if we want to test that the average milk production of Karan Swiss cows in a lactation is 3000 litres then the null hypothesis may be expressed symbolically as $H_0: \mu = 3000$ litres.

6.3.1.4 Alternative hypothesis

Any hypothesis which is complementary to the null Hypothesis is called an alternative hypothesis. It is usually denoted by H_1 or H_A . For example if we want to test the null hypothesis that the population has a specified mean μ_0 i.e. $H_0: \mu = \mu_0$ then the alternative hypothesis could be

- (i) $H_1: \mu \neq \mu_0$ ($H_0: \mu = \mu_0$)
- (ii) $H_1: \mu > \mu_0$ ($H_0: \mu \leq \mu_0$)
- (iii) $H_1: \mu < \mu_0$ ($H_0: \mu \geq \mu_0$)

The alternative hypothesis in (i) is known as two tailed alternative and the alternatives in (ii) and (iii) are known as right tailed and left tailed alternatives. The setting of alternative hypothesis is very important since it enables us to decide whether to use a single tailed (right or left) or two tailed test. The null hypothesis consists of only a single parameter value and is usually simple while alternative hypothesis is usually composite.

6.3.1.5 Simple and Composite Hypothesis

If the statistical hypothesis completely specifies the population or distribution, it is called a simple hypothesis; otherwise it is called a composite hypothesis. For example, if we consider a

normal population $N(\mu, \sigma^2)$ where σ^2 is known and we want to test the hypothesis $H_0: \mu=25$ against $H_1: \mu=30$. From these hypotheses, we know that μ can take either of the two values, 25 or 30. In this case H_0 and H_1 are both simple. But generally H_1 is composite, i.e., of the form $H_1: \mu \neq 25$, viz, $H_1: \mu < 25$ or $H_1: \mu > 25$. In sampling from a normal population $N(\mu, \sigma^2)$, the hypothesis $H: \mu = \mu_0$ and $\sigma^2 = \sigma_0^2$ is a simple hypothesis because it completely specifies the distribution. On the other hand (i) $\mu = \mu_0$ (σ^2 is not specified) (ii) $\sigma^2 = \sigma_0^2$ (μ is not specified) (iii) $\mu < \mu_0$, $\sigma^2 = \sigma_0^2$ etc. are composite hypothesis.

6.3.1.6 Types of Errors in Testing of Hypothesis

The main objective in sampling theory is to draw a valid inference about the population parameters on the basis of the sample results. In practice we decide to accept or reject a null hypothesis (H_0) after examining a sample from it. As such we are liable to commit errors. The four possible situations that arise in testing of hypothesis are expressed in the following dichotomous table:

Table 6.1

Decision from sample	True Situation	
	Hypothesis is true	Hypothesis is false
Accept the hypothesis	No error	Type II error
Reject the hypothesis	Type I error	No error

In testing hypothesis, there are two possible types of errors which can be made. The error of rejection of a hypothesis H_0 when H_0 is true is known as Type I error and error of acceptance of a hypothesis H_0 when H_0 is false is known as type II error. When setting up an experiment to test a hypothesis it is desirable to minimize the probabilities of making these errors. But practically it is not possible to minimize both these errors simultaneously. In practice, in most decision making problems, it is more risky to accept a wrong hypothesis than to reject a correct one. These two types of errors can be better understood with an example where a patient is given a medicine to cure some disease and his condition is examined for some time. It is just possible that the medicine has a positive effect but it is considered that it has no effect or adverse effect. Therefore it is the Type I error. On the other hand if the medicine has an adverse effect but it is considered to have had a positive effect, it is called Type II error. Now let us consider the implications of these two types of error. If type I error is committed, the patient will be given another medicine, which may or may not be effective. But if type II error is committed i.e., the medicine is continued inspite of an adverse effect, the patient may develop some other complications or may even die. This clearly shows that the type II error is much more serious than the type I error. Hence in drawing inference about the null hypothesis, generally type II error is minimized even at the risk of committing type I error which is usually chosen as 5 per cent or 1 per cent.

Probability of committing type I error and type II error are denoted by α and β and are called size of type I and type II error respectively. In Industrial Quality Control, while inspecting the quality

of a manufactured lot, the Type I error and type II error amounts to rejecting a good lot and accepting a bad lot respectively. Hence $\alpha = P(\text{Rejecting a good lot})$ and $\beta = P(\text{Accepting a bad lot})$. The sizes of type I and type II errors are also known as producer's risk and consumer's risk respectively. The value of $(1-\beta)$ is known as the power of the test.

6.3.1.7 Level of Significance

It is the amount of risk of the type I error which a researcher is ready to tolerate in making a decision about H_0 . In other words, it is the maximum size of type I error, which we are prepared to tolerate is called the level of significance. The level of significance denoted by α is conventionally chosen as 0.05 or 0.01. The level of 0.01 is chosen for high precision and the level 0.05 for moderate precision. Sometimes this level of risk is further brought down in medical statistics where the efficiency of life saving drug on the patient is tested. If we adopt 5% level of significance, it means that on 5 out of 100 occasions, we are likely to reject a correct H_0 . In other words this implies that we are 95% confident that our decision to reject H_0 is correct. That is, we want to make the significance level as small as possible in order to protect the null hypothesis and to prevent, as far as possible, the investigator from inadvertently making false claims. Level of significance is always fixed in advance before collecting the sample information.

6.3.1.8 P-Value Concept

Another approach followed in testing of hypothesis is to find out the P-value at which H_0 is significant i.e., to find the smallest level α at which H_0 is rejected. In this situation, it is not inferred whether H_0 is accepted or rejected at a level of 0.05 or 0.01 or any other level. But the researcher only gives the smallest level α at which H_0 is rejected. This facilitates an individual to decide for himself as to how much significant the research results are. This approach avoids the imposition of a fixed level of significance. About the acceptance or rejection of H_0 , the experimenter can himself decide the level of α by comparing it with the P-value. The criterion for this is that if the P-value is less than or equal to α , reject H_0 otherwise accept H_0 .

6.3.1.9 Degrees of Freedom

For a given set of conditions, the number of degrees of freedom is the maximum number of variables which can freely be designed (i.e., calculated or assumed) before the rest of the variates are completely determined. In other words, it is the total number of variates minus the number of independent relationships existing among them. It is also known as the number of independent variates that make up the Statistic. In general, degree of freedom is the total number of observations (n) minus the number of independent linear constraints (k) i.e. $n-k$.

6.3.1.10 Critical Region

The total area under a standard curve is equal to one representing probability distribution. In test of hypothesis the level of significance is set up in order to know the probability of making type I error of rejecting the hypothesis which is true. A statistic is required to be used to test the null

hypothesis H_0 . This test is assumed to follow some known distribution. In a test, the area under the probability density curve is divided into two regions, viz, the region of acceptance and the region of rejection. If the value of test statistics lies in the region of rejection, the H_0 will be rejected. The region of rejection is also known as a critical region. The critical region is always on the tail of the distribution curve. It may be on both sides of the tails or on one side of the tail depending upon alternative hypothesis H_1 .

6.3.1.10.1 One tailed test

A test of any statistical hypothesis where the alternative hypothesis is one tailed (right-tailed or left- tailed) is called a one tailed test. For example, a test for testing the mean of a population $H_0: \mu = \mu_0$ against the alternative hypothesis $H_1: \mu > \mu_0$ (Right tailed) or $H_1: \mu < \mu_0$ (Left tailed) is a single tailed test. If the critical region is represented by only one tail, the test is called one-tailed test or one sided test. In right tailed test ($H_1: \mu > \mu_0$) the critical region lies entirely on the right tail of the sampling distribution of \bar{x} , as shown in Fig. 6.2 , while for the left tail test ($H_1: \mu < \mu_0$), the critical region is entirely in the left tail of the distribution of \bar{x} , as shown in Fig. 6.1.

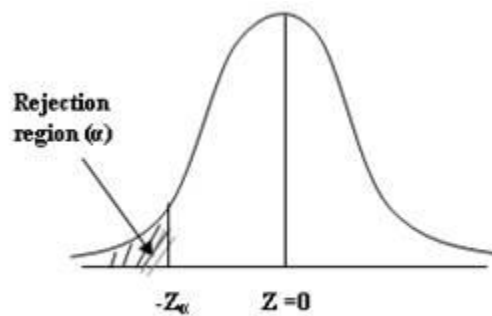


Fig. 6.1 Left tailed Test

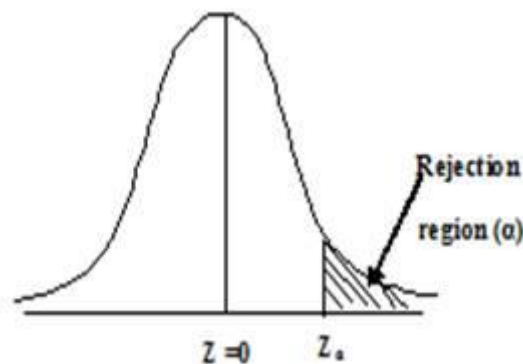


Fig. 6.2 Right tailed Test

6.3.1.10.2 Two tailed test

A test of statistical hypothesis where the alternative hypothesis is two sided such as: $H_0: \mu = \mu_0$ against the alternative hypothesis $H_1: \mu \neq \mu_0$ ($\mu > \mu_0$ or $\mu < \mu_0$) is known as a two tailed test and

in such a case the critical region is given by the portion of the area lying on both the tails of the probability curve of the test statistic as shown in Fig. 6.3

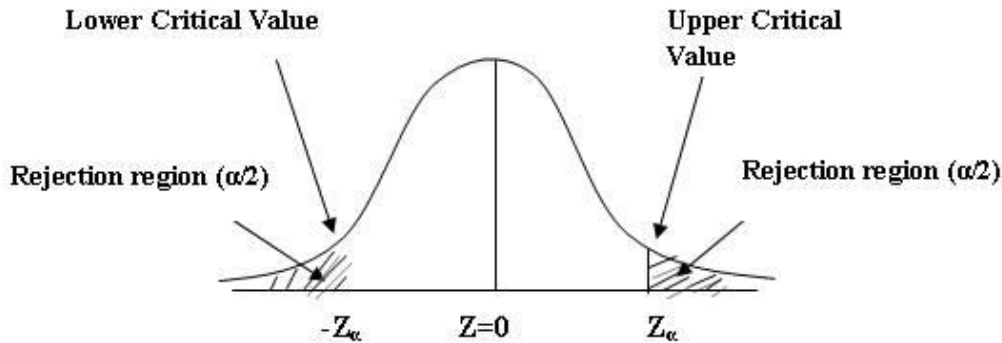


Fig. 6.3 Two Tailed Test

In a particular problem, whether one-tailed or two-tailed test is to be applied depends entirely on the nature of the alternative hypothesis.

6.3.1.11 Critical Values or Significant Values

The value of a test statistic which separates the critical (or rejection) region and the acceptance region is called the critical value or significant value. It depends upon

- } the level of significance and
- } the alternative hypothesis, whether it is two tailed or single tailed.

The critical value of the test statistic at α level of significance for a two tailed test is given by Z_α where Z_α is determined by the equation $P(|Z| > Z_\alpha)$, where Z_α is the value so that the total area of the critical region on both tails is α . Since normal probability curve is symmetric curve, we get $P[Z > Z_\alpha] = \alpha/2$ i.e., the area of each tail is $\alpha/2$. Thus Z_α is the value such that area to the right of Z_α is $\alpha/2$ and to the left of $-Z_\alpha$ is $\alpha/2$ as shown in fig. 6.3 .In case of single tail alternative, the critical value of Z_α is determined so that total area to the right of it (for right tailed test) is α (as shown in fig. 6.2) and for left tailed test the total area to the left of $-Z_\alpha$ is α (as shown in Fig. 6.1).Thus the significant or critical value of Z for a single tailed test (left or right) at level of significance ' α ' is same as the critical value of Z for a two tailed test at level of significance 2α .

Table 6.2 Critical values of (Z_α) of Z

Critical values (Z_α)	Level of significance		
	1%	5%	10%
Two tailed test	$ Z_\alpha = 2.58$	$ Z_\alpha = 1.96$	$ Z_\alpha = 1.645$
Right tailed test	$Z_\alpha = 2.33$	$Z_\alpha = 1.645$	$Z_\alpha = 1.28$
Left tailed test	$Z_\alpha = -2.33$	$Z_\alpha = -1.645$	$Z_\alpha = -1.28$

6.3.2 Procedure for Testing of Hypothesis/Test of Significance

The tests of significance which are dealt hereafter pertain to parametric tests. A statistical test is defined as a procedure governed by certain rules, which leads to take a decision about the hypothesis for its acceptance or rejection on the basis of sample observations. Test of significance enables us to decide on the basis of sample results if

- (i) Deviation between observed sample statistic and the hypothetical parameter value or
- (ii) The deviation between two sample statistics is significant.

Test of significance is a procedure of either accepting or rejecting a Null hypothesis. The tests are usually called tests of significance since here we test whether the difference between the sample values and the population values or between the values given by two samples are so large that they signify evidence against the hypothesis or these differences are small enough to be accounted for as due to fluctuations of sampling, i.e. they may be regarded as due only to the fact that we are dealing with a sample and not with the whole population. Statistical tests play an important role in biological sciences, dairy industry, social sciences and agricultural sciences etc. The use of these tests is made clear through a number of practical examples:

1. An automatic machine is filling 500 ml. of milk in the pouch. Now to make sure whether the claim is correct or not one has to take a random sample of the filled in pouches and note the actual quantity of milk in the pouches. From these sample observations it would be decided whether the automatic machine is filling the right quantity of milk in the pouches. This is done by performing test of significance.
2. There is a process A which produced certain items. It is considered that a new process B is better than process A. Both the processes are put under operation and then items produced by them are sampled and observations are taken on them. A statistical based test on sample observations will help the investigator to decide whether the process B is better than A or not.
3. Psychologists are often interested in knowing whether the level of IQ of a group of students is up to a certain standard or not. In this case some students are selected and an intelligence test is conducted. The scores obtained by them are subjected to certain statistical test and a decision is made whether their IQ is up to the standard or not.

There is no end to such types of practical problems where statistical tests can be applied. Here one important point may be noted. Whatever conclusions are drawn about the population (s), they are always subjected to some error.

6.3.2.1 Steps of Testing of Hypothesis

Various steps in test of hypothesis / significance are as follows:

- (i) Set up the Null hypothesis H_0 .

- (ii) Set up the alternative hypothesis H_1 . This will decide whether to go for single tailed test or two tailed test.
- (iii) Choose the appropriate level of significance depending upon the reliability of the estimates and permissible risk. This is to be decided before sample is drawn.

$$Z = \frac{t - E(t)}{S.E.(t)}$$

- (iv) Compute the test statistic .
- (v) Compare the computed value of Z in previous step with the significant value Z_α at a given level of significance.

Conclusion :

- a) If $|Z| < Z_\alpha$ i.e. if calculated value of Z (test statistic) is less than Z_α , we say it is not significant, null hypothesis is accepted at level of significance α .
- b) If $|Z| > Z_\alpha$ i.e. if calculated value of Z (test statistic) is greater than Z_α , we say it is significant and null hypothesis is rejected at level of significance α .

6.4 CHI-SQUARE TEST

6.4.1 Introduction

The Chi-square (χ^2) test represents a useful method of comparing experimentally obtained results with those to be expected theoretically on some hypothesis. Thus Chi-square is a measure of actual divergence of the observed and expected frequencies. It is very obvious that the importance of such a measure would be very great in sampling studies where we have invariably to study the divergence between theory and fact. Chi-square as we have seen is a measure of divergence between the expected and observed frequencies and as such if there is no difference between expected and observed frequencies the value of Chi-square is 0. If there is a difference between the observed and the expected frequencies then the value of Chi-square would be more than 0. That is, the larger the Chi-square the greater the probability of a real divergence of experimentally observed from expected results. If the calculated value of chi-square is very small as compared to its table value it indicates that the divergence between actual and expected frequencies is very little and consequently the fit is good. If, on the other hand, the calculated value of chi-square is very big as compared to its table value it indicates that the divergence between expected and observed frequencies is very great and consequently the fit is poor.

The equation for chi-square (χ^2) is stated as follows:

$$\chi^2 = \sum \left[\frac{(f_o - f_e)^2}{f_e} \right]$$

in which f_o = frequency of occurrence of observed or experimentally determined facts

f_e = expected frequency of occurrence on some hypothesis.

Thus chi-square is the sum of the values obtained by dividing the square of the difference between observed and expected frequencies by the expected frequencies in each case. In other words the differences between observed and expected frequencies are squared and divided by the expected number in each case, and the sum of these quotients is χ^2 .

Several illustrations of the chi-square test will clarify the discussion given above. The differences of f_o and f_e are written always + ve.

6.4.2 Applications of Chi Square Test

6.4.2.1. Testing the divergence of observed results from those expected on the hypothesis of equal probability (null hypothesis):

Example 1:

Ninety-six subjects are asked to express their attitude towards the proposition “Should AIDS education be integrated in the curriculum of Higher secondary stage” by marking F (favourable), I (indifferent) or U (unfavourable).

It was observed that 48 marked ‘F’, 24 ‘I’ and 24 ‘U’:

- (i) Test whether the observed results diverge significantly from the results to be expected if there are no preferences in the group.
- (ii) Test the hypothesis that “there is no difference between preferences in the group”.
- (iii) Interpret the findings.

Solution:

Following steps may be followed for the computation of χ^2 and drawing the conclusions:

Step 1:

Compute the expected frequencies (f_e) corresponding to the observed frequencies in each case under some theory or hypothesis.

In our example the theory is of equal probability (null hypothesis). In the second row the distribution of answers to be expected on the null hypothesis is selected equally.

	Favourable	Indifferent	Unfavourable	
Observed (f_o)	48	24	24	96
Expected (f_e)	32	32	32	96
$(f_o - f_e)$	16	8	8	
$(f_o - f_e)^2$	256	64	64	
$\frac{(f_o - f_e)^2}{f_e}$	8	2	2	

$$\chi^2 = \sum \left[\frac{(f_o - f_e)^2}{f_e} \right] = 12 \quad df = 2 \quad P \text{ is less than } .01$$

Step 2:

Compute the deviations ($f_o - f_e$) for each frequency. Each of these differences is squared and divided by its f_e ($256/32$, $64/32$ and $64/32$).

Step 3:

Add these values to compute:

$$\chi^2 = \sum \left[\frac{(f_o - f_e)^2}{f_e} \right]$$

From Table E
Tabulated χ^2 with 2 df at
.01 level = 9.21

$$\left(\frac{256}{32} + \frac{64}{32} + \frac{64}{32} \right) \text{ to give } \chi^2 = 8 + 2 + 2 = 12$$

Step 4:

The degrees of freedom in the table is calculated from the formula $df = (r - 1)(c - 1)$ to be $(3 - 1)(2 - 1)$ or 2.

Step 5:

Look up the calculated (critical) values of χ^2 for 2 df at certain level of significance, usually 5% or 1%.

With $df = 2$, the χ^2 value to be significant at .01 level is 9.21 (Table E). The obtained χ^2 value of $12 > 9.21$.

- i. Hence the marked divergence is significant.
- ii. The null hypothesis is rejected.
- iii. We conclude that our group really favours the proposition.

We reject the “equal answer” hypothesis and conclude that our group favours the proposition.

6.4.3.2. Testing the divergence of observed results from those expected on the hypothesis of a normal distribution:

The hypothesis, instead of being equally probable, may follow the normal distribution. An example illustrates how this hypothesis may be tested by chi-square.

Example 3:

Two hundred salesmen have been classified into three groups very good, satisfactory, and poor—by consensus of sales managers.

Does this distribution of rating differ significantly from that to be expected if selling ability is normally distributed in our population of salesmen?

	Good	Satisfactory	Poor	Total
Observed (f_o)	76	96	28	200
Expected (f_e)	32	136	32	200
$(f_o - f_e)$	44	40	4	
$(f_o - f_e)^2$	1936	1600	16	
$\frac{(f_o - f_e)^2}{f_e}$	60.50	11.76	0.50	

$$\chi^2 = \sum \left[\frac{(f_o - f_e)^2}{f_e} \right] = 60.50 + 11.76 + .50 = 72.76$$

We set up the hypothesis that selling ability is normally distributed. The normal curve extends from -3σ to $+3\sigma$. If the selling ability is normally distributed the base line can be divided into three equal segments, i.e.

Rating	σ range between	% in Table A	% of 200 or (f_e)
Good	+ 3.00 σ and + 1.00 σ	16%	32
Satisfactory	+ 1.00 σ and - 1.00 σ	68%	136
Poor	- 1.00 σ and - 3.00 σ	16%	32
		100%	200

(+ 1 σ to + 3 σ), (- 1 σ to + 1 σ) and (- 3 σ to - 1 σ) representing good, satisfactory and poor salesmen respectively. By referring Table A we find that 16% of cases lie between + 1 σ and +3 σ , 68% in between - 1 σ and + 1 σ and 16% in between - 3 σ and - 1 σ . In case of our problem 16% of 200 = 32 and 68% of 200 = 136.

df= 2. P is less than .01

The calculated $\chi^2 = 72.76$

The calculated χ^2 of 72.76 > 9.21. Hence P is less than .01.

From Table E
Tabulated χ^2 for 2df at
.01 level = 9.21

∴ The discrepancy between observed frequencies and expected frequencies is quite significant. On this ground the hypothesis of a normal distribution of selling ability in this group must be rejected. Hence we conclude that the distribution of ratings differ from that to be expected.

6.4.3.3. Chi-square test when our expectations are based on predetermined results:

Example 4:

In an experiment on breeding of peas a researcher obtained the following data:

The theory predicts the proportion of beans, in four groups A, B, C and D should be 9: 3: 3: 1. In an experiment among 1,600 beans, the numbers in four groups were 882, 313, 287 and 118. Does the experiment result support the genetic theory? (Test at .05 level).

Solution:

We set up the null hypothesis that there is no significant difference between the experimental values and the theory. In other words there is good correspondence between theory and experiment, i.e., the theory supports the experiment.

Category	Expected frequency (f_e)
A	$\frac{9}{16} \times 1600 = 900$
B	$\frac{3}{16} \times 1600 = 300$
C	$\frac{3}{16} \times 1600 = 300$
D	$\frac{1}{16} \times 1600 = 100$

$9 + 3 + 3 + 1 = 16$

Computation of χ^2

	A	B	C	D
Observed frequency f_o	882	313	287	118
Expected frequency f_e	900	300	300	100

$(f_o - f_e)$	18	13	13	18
$(f_o - f_e)^2$	324	169	169	324
$\frac{(f_o - f_e)^2}{f_e}$.360	.563	.563	3.240

$$\Sigma \left[\frac{(f_o - f_e)^2}{f_e} \right] = .360 + .563 + .563 + 3.240 = 4.726$$

$df = 3$ P is near about .20

The calculated $\chi^2 = 4.726$

From Table E
Tabulated χ^2 for 3df
at .05 level = 7.81

Since the calculated χ^2 value of $4.726 < 7.81$, it is not significant. Hence null hypothesis may be accepted at .05 level of significance and we may conclude that the experimental results support the genetic theory.

6.4.3.4. The Chi-square test when table entries are small:

When table entries are small and when table is 2 x 2 fold, i.e., $df = 1$, χ^2 is subject to considerable error unless a correction for continuity (called Yates' Correction) is made.

Example 5:

Forty rats were offered opportunity to choose between two routes. It was found that 13 chose lighted routes (i.e., routes with more illumination) and 27 chose dark routes.

- (i) Test the hypothesis that illumination makes no difference in the rats' preference for routes (Test at .05 level).
- (ii) Test whether the rats have a preference towards dark routes.

Solution:

If illumination makes no difference in preference for routes i.e., if H_0 be true, the proportionate preference would be 1/2 for each route (i.e., 20).

In our example we are to subtract .5 from each ($f_o - f_e$) difference for the following reason:

In 2×2 fold tables, especially when entries are small, the χ^2 curve is not continuous. Hence, the deviation of 27 from 20 must be written as 6.5 ($26.5 - 20$) instead of 7 ($27 - 20$), as 26.5 is the lower limit of 27 in a continuous series. In like manner the deviation of 13 from 20 must be taken from the upper limit of 13, namely, 13.5.

The data can be tabulated as follows:

	Dark routes	Lighted routes	Total
Observed frequencies f_o	27	13	40
Expected frequencies f_e	20	20	40

$(f_o - f_e)$	7	7
$[(f_o - f_e) - .5]$	6.5	6.5
$[(f_o - f_e) - .5]^2$	42.25	42.25
$\frac{[(f_o - f_e) - .5]^2}{f_e}$	2.11	2.11

$\therefore \chi^2 = 2.11 + 2.11 = 4.22.$

When the expected entries in 2×2 fold table are the same as in our problem the formula for chi-square may be written in a somewhat shorter form as follows:

$$\chi^2 = \frac{2[(f_o - f_e) - .5]^2}{f_e}$$

$$= \frac{2(6.5)^2}{20} = \frac{2 \times 42.25}{20} = 4.22$$

$df = 1$ P is .043 (by interpolation)

Calculated $\chi^2 = 4.22$

From Table E ... (56)
The tabulated value of χ^2 for 1 df at .05 level = 3.841.

- (i) The critical value of χ^2 at .05 level is 3.841. The obtained χ^2 of 4.22 is more than 3.841. Hence the null hypothesis is rejected at .05 level. Apparently light or dark is a factor in the rats' choice for routes.

(ii) In our example we have to make a one-tailed test. Entering table E we find that χ^2 of 4.22 has a $P = .043$ (by interpolation).

$\therefore P/2 = .0215$ or 2%. In other words there are 2 chances in 100 that such a divergence would occur.

Hence we mark the divergence to be significant at 02 level.

Therefore, we conclude that the rats have a preference for dark routes.

6.4.3.5. The Chi-square test of independence in contingency tables:

Sometimes we may encounter situations which require us to test whether there is any relationship (or association) between two variables or attributes. In other words χ^2 can be made when we wish to investigate the relationship between traits or attributes which can be classified into two or more categories.

For example, we may be required to test whether the eye-colour of father is associated with the eye-colour of sons, whether the socio-economic status of the family is associated with the preference of different brands of a commodity, whether the education of couple and family size are related, whether a particular vaccine has a controlling effect on a particular disease etc.

To make a test we prepare a contingency table and to calculate f_e (expected frequency) for each cell of the contingency table and then compute χ^2 by using formula:

$$\chi^2 = \sum \left[\frac{(f_o - f_e)^2}{f_e} \right]$$

Null hypothesis:

χ^2 is calculated with an assumption that the two attributes are independent of each other, i.e. there is no relationship between the two attributes.

The calculation of expected frequency of a cell is as follows:

$$f_e \text{ of a cell} = \frac{\text{Row Total} \times \text{Column Total}}{\text{Grand total}}$$

Example 6:

In a certain sample of 2,000 families' 1,400 families are consumers of tea where 1236 are Hindu families and 164 are non-Hindu.

And 600 families are not consumers of tea where 564 are Hindu families and 36 are non-Hindu. Use χ^2 – test and state whether there is any significant difference between consumption of tea among Hindu and non-Hindu families.

Solution:

The above data can be arranged in the form of a 2 x 2 contingency table as given below:

	Hindu	Non-Hindu	Total
Families consuming tea	(I) 1236	(II) 164	1400
Families not consuming tea	(III) 564	(IV) 36	600
Grand Total	1800	200	2000

We set up the null hypothesis (H_0) that the two attributes viz., ‘consumption of tea’ and the ‘community’ are independent. In other words, there is no significant difference between the consumption of tea among Hindu and non-Hindu families.

Calculation of (f_e):

	Hindu	Non-Hindu	Total
Families consuming tea	(I) $\frac{1800 \times 1400}{2000}$ = 1260	(II) $\frac{200 \times 1400}{2000}$ = 140	1400
Families not consuming tea	(III) $\frac{1800 \times 600}{2000}$ = 540	(IV) $\frac{200 \times 600}{2000}$ = 60	600
Total	1800	200	2000

$$f_e \text{ of each cell} = \frac{\text{Row Total} \times \text{Column Total}}{\text{Grand total}}$$

Calculation of χ^2

Cells	f_o	f_e	$(f_o - f_e)$	$(f_o - f_e)^2$	$\frac{(f_o - f_e)^2}{f_e}$
I	1236	1260	24	576	0.4571
II	164	140	24	576	4.1143
III	564	540	24	576	1.0667
IV	36	60	24	576	9.6000

$(f_o - f_e)$ is written disregarding sign. $\chi^2 = 15.2381$

$$df = (2 - 1)(2 - 1) = 1$$

P is less than .01

$$\text{Calculated } \chi^2 = 15.2381$$

From Table E Tabulated value of χ^2 for 1 df at .05 level = 3.841 .01 level = 6.635

Since the calculated value of χ^2 , viz., 15.24 is much greater than the tabulated value of χ^2 at .01 level of significance; the value of χ^2 is highly significant and null hypothesis is rejected.

Hence we conclude that the two communities (Hindu and Non-Hindus) differ significantly as regards the consumption of tea among them.

Example 7:

The table given below shows the data obtained during an epidemic of cholera.

	Attacked	Non Attacked	Total
Inoculated	31	469	500
Not Inoculated	185	1315	1500
Total	216	1784	2000

Test the effectiveness of inoculation in preventing the attack of cholera.

Solution:

We set up the null hypothesis (H₀) that the two attributes viz., inoculation and absence of attack from cholera are not associated. These two attributes in the given table are independent.

	Attacked	Non Attacked	Total
Inoculated	(I) 31	(II) 469	500
Not Inoculated	(III) 185	(IV) 1315	1500
Total	216	1784	2000

Basing on our hypothesis we can calculate the expected frequencies as follows:

Calculation of (f_e):

	Attacked	Not Attacked	Total
Inoculated	(I) $\frac{500 \times 216}{2000} = 54$	(II) $\frac{500 \times 1784}{2000} = 446$	500
Not Inoculated	(III) $\frac{1500 \times 216}{2000} = 162$	(IV) $\frac{1500 \times 1784}{2000} = 1338$	1500
Total			2000

$$f_e \text{ of each cell} = \frac{\text{Row Total} \times \text{Column Total}}{\text{Grand Total}}$$

Calculation of χ^2

Cells	f _o	f _e	f _o - f _e	(f _o - f _e) ²	$\frac{(f_o - f_e)^2}{f_e}$
I	31	54	23	529	9.796
II	469	446	23	529	1.186
III	185	162	23	529	3.265
IV	1315	1338	23	529	0.395

$$\chi^2 = \sum \left[\frac{(f_o - f_e)^2}{f_e} \right] = 9.796 + 1.186 + 3.265 + 0.395 = 14.642$$

df = (2 - 1) (2 - 1) = 1. P is less than .01

Calculated $\chi^2 = 14.64$

From Table E
Tabulated value of χ for 1
df at .05 level = 3.841

The five percent value of χ^2 for 1 df is 3.841, which is much less than the calculated value of χ^2 . So in the light of this, conclusion is evident that the hypothesis is incorrect and inoculation and absence of attack from cholera are associated.

6.5. T-TEST

6.5.1 Introduction

The various tests of significance discussed in the previous lesson were related to large samples. The large sample theory was based on the application of Normal deviate test. However if sample

size n is small ($n < 30$), the distribution of the various statistics, e.g., $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$, are far from normality and as such Normal deviate test cannot be applied if n is small. Hence to deal with small samples, new techniques and tests of significance known as exact sample tests were developed which were pioneered by W. S. Gosset (1908) who wrote under the pen name of Student and later on developed and extended by Professor R. A. Fisher (1926). From practical point of view, a sample is small if its size is less than 30. In this lesson we shall discuss Student's t-test. In exact sample tests, the basic assumption is that the population(s) from which sample(s) are drawn is (are) normal i.e., the parent population(s) is (are) normally distributed and sample(s) is (are) random and independent of each other. The exact sample tests can be used even for large samples but large sample theory cannot be used for small samples.

6.5.2 Student's T

Definition

Let X_i ($i=1,2,\dots,n$) be a random sample of size n drawn from a normal population with mean μ and variance σ^2 , then student's t is defined by the statistic.

$$t = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

$$\text{where } \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \text{ and } S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

where S^2 is an unbiased estimate of the population variance σ^2 , and it follows student's t distribution with $(n-1)$ degrees of freedom.

Therefore $(n-1) S^2 = n s^2$

6.5.3 Applications of T-Test

The t-test has number of applications in statistics which are discussed in following sections

⌋ t-test for significance of single mean, population variance being unknown

- } t-test for the significance of the difference between two means, the population variances being equal
- } t-test for significance of an observed sample correlation coefficient.

6.5. 3.1 t-Test for single mean

Suppose we want to test

- (i) If the given normal population has a specified value of the population mean μ_0 .
- (ii) If the sample mean differs from the specified value μ_0 of the population mean.
- (iii) If a random sample of size n viz., X_i ($i=1,2,\dots, n$) has been drawn from a normal population with specified mean μ_0 .

Basically all the above three problems are same with corresponding null hypothesis H_0 as follows

- (i) $\mu = \mu_0$ i.e., the population mean is μ_0
- (ii) There is no difference between the sample mean \bar{x} and the population means μ .
- (iii) The given sample has been drawn from the population with mean μ_0 .

The test statistic is given by

$$t = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

$$\text{where } \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \text{ and } S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

follows student's t distribution with $(n-1)$ degrees of freedom. If calculated $|t| >$ tabulated value of t at 5 percent level of significance viz., $t_{0.05; (n-1)}$ d.f. then H_0 is rejected at 5 per cent level of significance which implies that there is a significant difference between sample mean and population mean or the sample has not been drawn from the population having specified mean $\mu = \mu_0$. If calculated $|t| <$ tabulated value of t at 5 percent level of significance viz., $t_{0.05; (n-1)}$ d.f. then H_0 is accepted. This is explained with the help of following illustrations.

Example .1: A random sample of 9 values from a normal population showed a mean of 41.5 and the sum of squares of deviations from the mean equal to 72. Test whether the assumption of mean 44.5 in the population is reasonable.

Solution: In this problem $n=9$ $\mu=44.5$, $\bar{X}=41.5$ and $\sum_{i=1}^9 (X_i - \bar{X})^2 = 72$

$H_0: \mu=44.5$ i.e., population mean is 44.5

$H_1: \mu \neq 44.5$

Applying t-test

$$t = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 = \frac{1}{8}(72) = 9$$

$$t = \frac{41.5 - 44.5}{3/\sqrt{9}} = -3$$

Tabulated value of t at 5% level of significance and 8 d.f. = 2.306. Since the calculated value of |t| is greater than tabulated value 2.306, hence it is significant. We reject null hypothesis and conclude that the population mean is not equal to 44.5.

Example 2: An automatic machine was expected to fill 250 ml of flavored milk in the pouches. A random sample of pouches was taken and the actual content of milk was weighed. Weight of flavored milk (in ml.) is

253, 251, 248, 251, 252, 250, 249, 254, 247, 249, 248, 255, 245, 246, 254.

Do you consider that the average quantity of flavored milk in the sample is the same as that of adjusted value?

Solution: In this problem $n=15$ $\mu=250$ ml.

$H_0: \mu=250$ ml i.e., automatic machine on an average fills 250 ml milk in each pouch

$H_1: \mu \neq 250$

Prepare the following table

Table 6.3

X_i	$X_i - \bar{X}$	$(X_i - \bar{X})^2$
253	2.8667	8.2178
251	0.8667	0.7511
248	-2.1333	4.5511
251	0.8667	0.7511
252	1.8667	3.4844
250	-0.1333	0.0178
249	-1.1333	1.2844
254	3.8667	14.9511
247	-3.1333	9.8178
249	-1.1333	1.2844
248	-2.1333	4.5511
255	4.8667	23.6844
245	-5.1333	26.3511
246	-4.1333	17.0844
254	3.8667	14.9511
3752	0.0000	131.7333

$$\bar{X} = \frac{3752}{15} = 250.1333, \quad S^2 = \frac{131.7333}{14} = 9.4095$$

Applying t-test

$$t = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} = \frac{250.1333 - 250}{0.7920} = 0.1683$$

Tabulated value of t at 5% level of significance for 14 d.f. is 2.15. Since the calculated value of |t| is less than tabulated value 2.15, hence it is not significant. We accept null hypothesis and conclude that the on an average automatic machine fills 250 ml. of flavored milk in pouches.

6.5.3.2 t-Test for difference of means

Suppose we want to test if two independent samples $X_i (i=1,2,\dots,n_1)$ and $Y_j (j=1,2,\dots,n_2)$ of sizes n_1 and n_2 have been drawn from two normal populations with means μ_1 and μ_2 respectively. Under the Null hypothesis $H_0: \mu_1 = \mu_2$ i.e., that the samples have been drawn from the populations having same mean .

$$H_1 \mu_1 \neq \mu_0$$

The t- statistic is given by

$$t = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

which follows t distribution with $(n_1 + n_2 - 2)$

$$\text{where } \bar{X} = \frac{1}{n_1} \sum_{i=1}^{n_1} X_i \quad \text{and} \quad \bar{Y} = \frac{1}{n_2} \sum_{j=1}^{n_2} Y_j$$

$$S^2 = \frac{1}{n_1 + n_2 - 2} \left[\sum_{i=1}^{n_1} (X_i - \bar{X})^2 + \sum_{j=1}^{n_2} (Y_j - \bar{Y})^2 \right]$$

is an unbiased estimate of the common population variance σ^2 based on both the samples. By comparing the computed value of t with the tabulated value of t for $(n_1 + n_2 - 2)$ d.f. and at desired level of significance, we reject or retain null hypothesis H_0

6.5.3.2.1 Assumptions for difference of means test

- (i) Parent populations from which the samples have been drawn are normally distributed.
- (ii) The two samples are random and independent of each other.
- (iii) The population variances are equal $\sigma_1^2 = \sigma_2^2 = \sigma^2$ but unknown.

Thus before applying t-test for testing the equality means, it is theoretical desirable to test the equality of population variances by applying F-test. If the hypothesis $H_0: \sigma_1^2 = \sigma_2^2$ is rejected then we cannot apply t-test and in such situations Behren's d test is applied. This procedure is explained with the help of following illustrations.

Example 3 : The prices of ghee were compared in two cities. For this purpose ten shops were selected at random in each city. The following table gives per kg. prices of ghee in two cities:

City A	361	363	356	364	359	360	362	361	358	357
City B	368	369	370	366	367	365	371	372	366	367

Test whether the average price of ghee is of the same order in two cities.

Solution:

Null hypothesis $H_0: \mu_A = \mu_B$ i.e., average price of ghee is of same order in cities A and B.

$H_1: \mu_A \neq \mu_B$

Prepare the following table:

Table 6.4

City A			City B		
X_i	$X_i - \bar{X}$	$(X_i - \bar{X})^2$	Y_j	$Y_j - \bar{Y}$	$(Y_j - \bar{Y})^2$
361	0.9	0.81	368	-0.1	0.01
363	2.9	8.41	369	0.9	0.81
356	-4.1	16.81	370	1.9	3.61
364	3.9	15.21	366	-2.1	4.41
359	-1.1	1.21	367	-1.1	1.21
360	-0.1	0.01	365	-3.1	9.61
362	1.9	3.61	371	2.9	8.41
361	0.9	0.81	372	3.9	15.21
358	-2.1	4.41	366	-2.1	4.41
357	-3.1	9.61	367	-1.1	1.21
3601		60.9	3681		48.9

and calculate,

$$\bar{X} = \frac{3601}{10} = 360.1 \text{ and } \bar{Y} = \frac{3681}{10} = 368.1$$

$$s^2 = \frac{1}{18}[60.9 + 48.9] = \frac{109.8}{18} = 6.1$$

$$t = \frac{(\bar{X} - \bar{Y})}{s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{360.1 - 368.1}{1.1045} = -7.2428$$

Tabulated value of t at 5% level of significance and 18 d.f. (for two tail) is 2.10. Since the calculated value of |t| is more than tabulated value (2.10), hence it is significant. We reject null hypothesis at 5 percent level of significance and conclude that average prices of ghee in both the cities are different.

6.5.3.3 Paired t-test

Let us now consider the case when

- (i) Sample sizes are equal i.e., $n_1 = n_2 = n$ and
- (ii) The samples are not independent but the sample observations are paired together i.e., the pair of observations (X_i, Y_i) $i=1,2,\dots,n$ corresponds to the same i^{th} sample unit. The problem is to test if the sample means differ significantly or not.

For example suppose we want to test the efficacy of a particular drug say for inducing sleep or controlling blood pressure or blood sugar among the patients or if we want to test the difference between two analysts or machines with regard to detection of mean fat percentage in milk. Let X_i and Y_i ($i=1,2,\dots,n$) be the readings of fat percentage of i^{th} milk sample, detected by two machines A and B respectively. Here instead of applying the difference of the means test discussed in previous section, we apply paired t-test.

Here we consider the difference $d_i = X_i - Y_i$ ($i=1,2,\dots,n$)

Under the Null hypothesis H_0 difference in fat percent in milk by both the machines is due to fluctuations of sampling i.e., $H_0: \mu_d = 0$

against $H_1: \mu_d \neq 0$

then the test statistic

$$t = \frac{\bar{d}}{S/\sqrt{n}}$$

follows t distribution with $(n-1)$ degrees of freedom

$$\text{where } \bar{d} = \frac{1}{n} \sum_{i=1}^n d_i$$

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (d_i - \bar{d})^2 = \frac{1}{n-1} \left[\sum_{i=1}^n d_i^2 - \frac{(\sum_{i=1}^n d_i)^2}{n} \right]$$

Different examples of paired t test are:

1. A sample of boys was given a test mathematics. They were given a month's extra coaching and a second test was held at the end of it? Do the marks give evidence that the students have been benefitted by the extra coaching?
2. A sample of patients was examined to know whether a drug tends to reduce the blood pressure. The data give the blood pressure readings before the drug was given and also after it was given. The question is to examine whether the drug is effective in controlling blood pressure.
3. It is desired to test the adoption of a new technology by the farmers. A group of farmers is taken where the knowledge level score is measured before the new technology is infused and after infusion of technology, the knowledge level score is again measured. Does the difference in technology level scores provide the evidence that the farmers have been benefitted by the adoption of new technology?

This procedure is explained with the help of following illustrations.

Example 4: Ten B.Tech. (Dairy Tech.) second year students were selected for a training on quality control on the basis of marks obtained in an examination conducted for this purpose . After one month training they were given a test and marks were recorded out of 50.

Student	A	B	C	D	E	F	G	H	I	J
Before training	25	20	35	15	42	28	26	44	35	48
After training	26	20	34	13	43	40	29	41	36	46

Test whether there is any change in performance after the training.

Solution:

In this problem, the marks obtained by the students before training (X) and after training (Y) are not independent but paired together, hence we shall apply paired t test. Null Hypothesis H_0 : $\mu_X = \mu_Y$ or H_0 : $\mu_d = 0$ i.e., mean scores before training and after training are same. In other words, the training has no impact on students' performance against H_1 : $\mu_d \neq 0$.

Prepare the following table

Table 6.5

Before training (X _i)	After training(Y _i)	d _i = X _i - Y _i	d _i ²
25	26	-1	1
20	20	0	0
35	34	1	1
15	13	2	4
42	43	-1	1
28	40	-12	144
26	29	-3	9
44	41	3	9
35	36	-1	1
48	46	2	4
Total		$\sum d_i = -10$	$\sum d_i^2 = 174$

and calculate

$$\bar{d} = \frac{1}{n} \sum_{i=1}^n d_i = -1$$

$$s^2 = \frac{1}{n-1} \left[\sum_{i=1}^n d_i^2 - \frac{(\sum_{i=1}^n d_i)^2}{n} \right] = \frac{1}{9} \left[174 - \frac{(-10)^2}{10} \right] = 18.2222$$

$$t = \frac{\bar{d}}{s/\sqrt{n}} = \frac{-1}{1.3498} = -0.7408$$

Tabulated value of t at 5% level of significance and 9 d.f. (for two tail) is 2.262. Since the calculated value of |t| is less than tabulated value 2.262, hence it is not significant. We accept null hypothesis and conclude that students have not been benefited from the training.

6.5.3.4 t-Test for significance of an observed sample correlation coefficient

Let a random sample (x_i, y_i) (i=1,2---,n) of size n has been drawn from a bivariate normal distribution and let r be the observed sample correlation coefficient . In order to test whether sample correlation coefficient r is significant or there is no correlation between the variables in the population. Prof. R. A. Fisher proved that under the null hypothesis Ho: ρ=0 i.e. the population correlation coefficient is zero. The statistic

$$t = \frac{r}{\sqrt{1-r^2}} \times \sqrt{n-2} \sim t_{n-2}$$

follows student's t distribution with $(n-2)$ d.f., n being the sample size.

6.6 Z-TEST

6.6.1 Introduction

In the previous lesson we encountered a problem to decide whether our sample observations have come from a postulated population or not. On the basis of sample observations, a test is performed to decide whether the postulated hypothesis is accepted or rejected and this involves certain amount of risk. The amount of risk is termed as a level of significance. When the hypothesis is rejected, we consider it as a significant result and when a reverse situation is encountered, we consider it as a non-significant result. We have seen that for large values of n , the number of trials, almost all the distributions e.g., Binomial, Poisson etc. are very closely approximated by Normal distribution and in this case we apply Normal Deviate test (Z-test). In cases where the population variance (s) is/are known, we use Z-test. The distribution of Z is always normal with mean zero and variance one. In this lesson we shall be studying the problem relating to test of significance for large samples only. In statistics a sample is said to be large if its size exceeds 30.

6.6.2 Test of Significance for Large Samples

In cases where the population variance(s) is/are known, we use Z-test. Moreover when the sample size is large, sample variance approaches population variance and is deemed to be almost equal to population variance. In this way, the population variance is known even if we have sample data and hence the normal deviate test is applicable. The distribution of Z is always normal with mean zero and variance one. Thus, if $X \sim N(\mu, \sigma^2)$

$$\text{then, } Z = \frac{X - \mu}{\sqrt{v(X)}} = \frac{X - E(X)}{\sigma} \sim N(0,1)$$

From normal probability tables, we have

$P[-3 \leq Z \leq 3] = P[|Z| \leq 3] = 0.9973 \Rightarrow P[|Z| \leq 3] = 1 - P[|Z| > 3] = 0.0027$. Thus, the value of $Z=3$ is regarded as critical or significant value at all levels of significance. Thus if $|Z| \leq 3$, H_0 is always rejected. If $|Z| < 3$, we test its significance at certain level of significance usually at 5% and sometimes at 1% level of significance. Also $P[|Z| > 1.96] = 0.05$ and $P[|Z| > 2.58] = 0.01$. Thus, significant values of Z at 5% and 1% level of significance are 1.96 and 2.58 respectively. If $|Z| > 1.96$, H_0 is rejected at 5% level of significance if $|Z| < 1.96$, H_0 may be retained at 5% level of significance. Similarly $|Z| > 2.58$, H_0 is rejected at 1% level of significance and if $|Z| < 2.58$, H_0 is retained at 1% level of significance. In the following sections we shall discuss the large sample (normal) tests for attributes and variables.

6.6.3 Applications of Z-Test

6.6.3.1 Test for single proportion

If the observations on various items or objects are categorized into two classes c_1 and c_2 (binomial population), viz. defective or not defective item, we often want to test the hypothesis, whether the proportion of items in a particular class, viz., defective items is P_0 or not. For example, the management of a dairy plant is interested in knowing that whether the population of leaked pouches filled by automatic milk filling machine is one percent. Thus for binomial population, the hypothesis we want to test is whether the sample proportion is representative of the Population proportion $P = P_0$ against $H_1: P \neq P_0$ or $H_1: P > P_0$ or $H_1: P < P_0$ can be tested by Z-test where P is the actual proportion of items in the population belonging to class c_1 . Proportions are mostly based on large samples and hence Z-test is applied.

If X is the number of successes in n independent trials with constant probability P of success for each trial then $E(X) = nP$ and $V(X) = nPQ$ where $Q = 1 - P$. It is known that for large n , the Binomial distribution tends to Normal distribution. Hence, for large n , $X \sim N(nP, nPQ)$. Therefore, Z statistic for single proportion is given by

$$Z = \frac{X - E(X)}{SE(X)} = \frac{X - E(X)}{\sqrt{V(X)}}$$

$$Z = \frac{(X - nP)}{\sqrt{nPQ}} \sim N(0,1)$$

and we can apply a normal deviate test.

If in a sample of size n , X be the number of persons possessing the given attributes then observed proportion of successes $\frac{X}{n} = p$

$$E(p) = E\left(\frac{X}{n}\right) = \frac{1}{n} E(X) = \frac{1}{n} nP = P$$

$$V(p) = V\left(\frac{X}{n}\right) = \frac{1}{n^2} V(X) = \frac{1}{n^2} nPQ = \frac{PQ}{n}$$

$$S.E.(p) = \sqrt{\frac{PQ}{n}}$$

Since X and consequently X/n is asymptotically normal for large n , the normal deviate test for the proportion of success becomes.

$$Z = \frac{p - E(p)}{SE(p)} = \frac{p - P}{\sqrt{\frac{PQ}{n}}} \sim N(0,1)$$

Example 1. In a large consignment of baby food packets, a random sample of 100 packets revealed that 5 packets were leaking. Test whether the sample comes from the population (large consignment) containing 3 percent leaked packets.

Solution: In this example $n=100$, $X=5$, $P=0.03$, $p = \frac{x}{n} = \frac{5}{100} = 0.05$

$H_0: P = 0.03$.i.e., the proportion of the leaked pouches in the population is 3 per cent

$H_1: P \neq 0.03$.

Here, we shall use standard normal deviate (Z) test for single proportion as under

$$Z = \frac{p - P}{\sqrt{\frac{PQ}{n}}} = \frac{0.05 - 0.03}{\sqrt{\frac{(0.03)(0.97)}{100}}} = \frac{0.02}{0.01706} = 1.17$$

Since calculated value of Z statistic is less than 1.96 therefore H_0 is not rejected at 5% level of significance which implies that the sample is representative of the population (large consignment) of packets containing 3% leaked packets.

6.6.3.2 Test of Significance for difference of proportions

If we have two populations and each item of a population belong to either of the two classes C_1 and C_2 . A person is often interested to know whether the proportion of items in class C_1 in both the populations is same or not that is we want to test the hypothesis.

$H_0: P_1 = P_2$ against $H_1: P_1 \neq P_2$ or $P_1 > P_2$ or $P_1 < P_2$ where P_1 and P_2 are the proportions of items in the two populations belonging to class C_1 .

Let X_1, X_2 be the number of items belonging to class C_1 in random samples of sizes n_1 and n_2 from the two populations respectively. Then the sample proportion

$$p_1 = \frac{X_1}{n_1}, p_2 = \frac{X_2}{n_2}$$

If P_1 and P_2 are the proportions then $E(p_1) = P_1, E(p_2) = P_2$

$$V(p_1) = \frac{P_1 Q_1}{n_1}, V(p_2) = \frac{P_2 Q_2}{n_2}$$

Since for the large sample, p_1 and p_2 are asymptotically normally distributed, $(p_1 - p_2)$ is also normally distributed. Therefore, the Z-statistic for difference between two proportions is given by

$$Z = \frac{(p_1 - p_2) - E(p_1 - p_2)}{\sqrt{V(p_1 - p_2)}} \sim N(0,1)$$

Since, $E(p_1 - p_2) = E(p_1) - E(p_2) = P_1 - P_2 = 0$

$$V(p_1 - p_2) = V(p_1) + V(p_2) = \frac{P_1 Q_1}{n_1} + \frac{P_2 Q_2}{n_2}$$

$$Z = \frac{(p_1 - p_2)}{\sqrt{V(p_1 - p_2)}} = \frac{(p_1 - p_2)}{\sqrt{\frac{P_1 Q_1}{n_1} + \frac{P_2 Q_2}{n_2}}}$$

Since $P_1 = P_2 = P$ and $Q_1 = Q_2 = Q$, therefore

$$Z = \frac{p_1 - p_2}{\sqrt{PQ \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

If the population proportion P_1 and P_2 are given to be distinctly different that is $P_1 \neq P_2$, then

$$Z = \frac{(p_1 - p_2) - (P_1 - P_2)}{\sqrt{\frac{P_1 Q_1}{n_1} + \frac{P_2 Q_2}{n_2}}}$$

In general P , the common population proportion (under H_0) is not known, then an unbiased estimate of population proportion, P , based on both the samples is used and is given by

$$\hat{p} = \frac{X_1 + X_2}{n_1 + n_2} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$$

Example 2. Before an increase in excise duty on tea, 400 people out of a sample of 500 persons were found to be tea drinkers. After an increase in excise duty, 400 people were observed to be tea drinkers in a sample of 600 people. Test whether there is a significant change in the number of tea drinkers after increase in excise duty on tea.

Solution: In this example $X_1 = 400$, $n_1 = 500$, $X_2 = 400$, $n_2 = 600$

$H_0: P_1 = P_2$ i.e., there is no change in the number of tea drinkers after increase in excise duty on tea

$H_1: P_1 \neq P_2$

Here we shall use standard normal deviate (Z) test for difference of proportions as under:

In our example $p_1 = 400/500 = 0.8$, $p_2 = 400/600 = 0.6667$
 $q_1 = 1 - p_1 = 0.2$, $q_2 = 1 - p_2 = 0.333$

$$Z = \frac{(p_1 - p_2)}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}} = \frac{0.8 - 0.6667}{\sqrt{\frac{(0.8)(0.2)}{500} + \frac{(0.6667)(0.333)}{600}}} = \frac{0.1333}{0.0263} = 5.07$$

Since calculated value of Z statistic is greater than 3, therefore H_0 is rejected at all levels of significance which implies that there is a significant change in the number of tea drinkers after increase in excise duty on tea. It is further observed that the number of tea drinkers have significantly declined after increase in excise duty on tea which is due to decrease in the value of p_2 (0.667) from the value of p_1 (0.8).

Example 3. A machine turns out 16 imperfect articles in a sample of 500. After overhauling it turns 3 imperfect articles in a batch of 100. Has the machine improved after overhauling?

Solution : We are given $n_1 = 500$ and $n_2 = 100$

$p_1 =$ Proportions of defective items before overhauling of machine $= 16/500 = 0.032$

$p_2 =$ Proportions of defective items after overhauling of machine $= 3/100 = 0.03$

$$\begin{aligned}
 &H_0: P_1=P_2 \text{ i.e. the machine has not improved after overhauling.} \\
 &H_1: P_1>P_2 \\
 Z &= \frac{(P_1 - P_2)}{\sqrt{\frac{P_1 Q_1}{n_1} + \frac{P_2 Q_2}{n_2}}} = \frac{0.032 - 0.03}{\sqrt{\frac{(0.032)(0.968)}{500} + \frac{(0.3)(0.7)}{100}}} = \frac{0.002}{0.01878} = 0.106
 \end{aligned}$$

Since $Z < 1.645$ (Right tailed test), it is not significant at 5% level of significance. Hence we may accept the null hypothesis and conclude that the machine has not improved after overhauling.

6.6.3.3 Test for significance of single mean

We have seen that if X_i ($i=1, 2, \dots, \dots, n$) is a random sample of size n from a normal population with mean μ and variance σ^2 , then the sample mean \bar{X} is distributed normally with mean μ and variance σ^2/n i.e., $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$. Thus for large samples normal variate corresponding to \bar{X} , is

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

In test of significance for a single mean we deal the following situations

- 1) To test if the mean of the population has a specified value (μ_0) and null hypothesis in this case will be $H_0: \mu = \mu_0$ i.e., the population has a specified mean value.
- 2) To test whether the sample mean differs significantly from the hypothetical value of population mean with null hypothesis as there is no difference between sample mean \bar{X} and population mean (μ).
- 3) To test if the given random sample has been drawn from a population with specified mean μ_0 and variance σ^2 with null hypothesis the sample has been drawn from a normal population with specified mean μ_0 and variance σ^2

In all the above three situations the test statistic is given by

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

If $|Z| < 1.96$, H_0 is not rejected at 5% level of significance which implies that there is no significant difference between sample mean and population mean and whatever difference is there, it exists due to fluctuation of sampling.

$|Z| > 1.96$, H_0 is rejected at 5% level of significance which implies that there is a significant difference between sample mean and population mean. The above situations are illustrated by following examples:

Example 4. A random sample of 100 students gave a mean weight of 64 kg with a standard deviation of 16 kg. Test the hypothesis that the mean weight in the population is 60 kg.

Solution: In this example, $n=100$, $\mu = 60$ kg., $\bar{X} = 64$ kg., $\sigma = 16$

$H_0: \mu = 60$ kg, i.e. the mean weight in the population is 60 kg.

We shall use standard normal deviate (z) test for single mean as under:

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{64 - 60}{16/\sqrt{100}} = 2.5$$

Since calculated value of Z statistic is more than 1.96, it is significant at 5% level of significance. Therefore, H_0 is rejected at all levels of significance which implies that mean weight of population is not 60 kg.

Example 5. A sample of 50 cows in a herd has average lactation yield 1290 litres. Test whether the sample has been drawn from the population having herd average lactation yield of 1350 litres with a standard deviation of 65 litres.

Solution: In this example, $n=50$, $\mu=1350$ litres, $\bar{X}=1290$, $\sigma=65$

$H_0: \mu = 1350$ litres i.e., the mean lactation milk yield of the cows in the population is 1350

$H_1: \mu \neq 1350$ litres

We shall use standard normal deviate (Z) test for single mean as under:

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{1290 - 1350}{65/\sqrt{50}} = -6.53 \Rightarrow |Z| = 6.53$$

Since calculated value of Z statistic is more than 3, it is significant at all levels of significance. Therefore, H_0 is rejected at all levels of significance which implies that the sample has not been drawn from the population having mean lactation milk yield as 1350 litres or there is a significant difference between sample mean and population mean.

6.6.3.4 Test of significance for difference of means

Let \bar{X}_1 be the mean of a sample of size n_1 drawn from a population with mean μ_1 and variance σ_1^2 and let \bar{X}_2 be the mean of an independent sample of size n_2 drawn from another population with mean μ_2 and variance σ_2^2 . Since sample sizes are large.

$$\bar{X}_1 \sim N\left(\mu_1, \frac{\sigma_1^2}{n_1}\right) \text{ and } \bar{X}_2 \sim N\left(\mu_2, \frac{\sigma_2^2}{n_2}\right)$$

Also $(\bar{X}_1 - \bar{X}_2)$, being the difference in means of two independent normal variates is also a normal variate. The standard normal variate corresponding to $\bar{X}_1 - \bar{X}_2$, is given by

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - E(\bar{X}_1 - \bar{X}_2)}{\sqrt{v(\bar{X}_1 - \bar{X}_2)}} \sim N(0,1)$$

Under the null hypothesis $H_0: \mu_1 = \mu_2$ i.e., the two population means are equal, we get $E(\bar{X}_1 - \bar{X}_2) = E(\bar{X}_1) - E(\bar{X}_2) = \mu_1 - \mu_2 = 0$

$$V(\bar{X}_1 - \bar{X}_2) = V(\bar{X}_1) + V(\bar{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

The covariance terms vanish, since the sample means \bar{X}_1 and \bar{X}_2 are independent.

Thus under $H_0: \mu_1 = \mu_2$, the Z statistic is given by

$$Z = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

Here σ_1^2 and σ_2^2 are assumed to be known. If they are unknown then their estimates provided by corresponding sample variances s_1^2 and s_2^2 respectively are used, i.e., $\hat{\sigma}_1^2 = s_1^2$ and $\hat{\sigma}_2^2 = s_2^2$, thus, in this case the test statistic becomes

$$Z = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim N(0,1)$$

Remarks: If we want to test whether the two independent samples have come from the same population i.e., if $\sigma_1^2 = \sigma_2^2 = \sigma^2$ (with common S.D. σ), then under $H_0: \mu_1 = \mu_2$

$$Z = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \sim N(0,1)$$

If the common variance σ^2 is not known, then we use its estimate based on both the samples which is given by

$$\hat{\sigma}^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2}$$

Example 6. In a certain factory there are two independent processes manufacturing the same item. The average weight in a sample of 100 items produced from one process is found to be 50g with a standard deviation of 5g while the corresponding figures in a sample of 75 items from the other process are 52g and 6g respectively. Is the difference between two means significant?

Solution: In this example, $n_1 = 100, \bar{X}_1 = 50 \text{ g}, s_1 = 5 \text{ g}, n_2 = 75, \bar{X}_2 = 52 \text{ g}, s_2 = 6 \text{ g}$.

Let μ_1 and μ_2 be the population mean of the weight of items manufactured by two independent processes.

$H_0: \mu_1 = \mu_2$, i.e., mean weights of the items manufactured by two independent processes in the population is same.

$H_0: \mu_1 \neq \mu_2$

Here, we shall use standard normal deviate test (Z-test) for calculating difference between two means as under

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{50 - 52}{\sqrt{\frac{25}{100} + \frac{36}{75}}} = \frac{-2}{0.8544} = -2.34 \Rightarrow |Z| = 2.34$$

Since calculated value of Z statistic is more than 1.96, therefore, H_0 is rejected at 5% level of significance which implies that there is a significant difference between mean weights of the items obtained from two manufacturing processes.

6.7 TUKEY'S Q TEST

A post-hoc test is needed after we complete an ANOVA in order to determine which groups differ from each other. Do not conduct a post-hoc test unless you found an effect (rejected the null) in the ANOVA problem. If you fail to reject the null, then there are no differences to find.

For the Tukey's post-hoc test we will first find the differences between the means of all of our groups. We will compare this difference score to a critical value to see if the difference is significant. The critical value in this case is the HSD (honestly significant difference) and it must be computed. It is the point when a mean difference becomes honestly significantly different.

$$HSD = q \sqrt{\frac{MSE_{within}}{n}}$$

Note that "q" is a table value, and n is the number of values we are dealing with in each group (not total n). The Mean Square value is from the ANOVA you already computed. To find "q" or the studentized range statistic, refer to the q table. On the table 'k' or the number of groups is found along the top, and degrees of freedom within is down the side. Cross index the row and column to find the value you need to put in the formula above.

Example

This example is a ANOVA problem where $MS_{within} = 1.43$. Here I show the groups, but have computed the average or mean of each group.

Therapy A	Therapy B	Therapy C
5	3	1
2	3	0
5	0	1
4	2	2
2	2	1

$$\bar{X}_1 = 3.6$$

$$\bar{X}_2 = 2$$

$$\bar{X}_3 = 1$$

The first step is to compute all possible differences between means:

$$\bar{X}_1 - \bar{X}_2 = 3.6 - 2 = 1.6 \quad \bar{X}_1 - \bar{X}_3 = 3.6 - 1 = 2.6 \quad \bar{X}_2 - \bar{X}_3 = 2 - 1 = -1$$

We will only be concerned with the absolute difference, so, you can ignore any negative signs. Next we compute HSD.

$$HSD = q \sqrt{\frac{MSE_{within}}{n}} = 2.77 \sqrt{\frac{1.43}{5}} = 1.99$$

Now we will compare the difference scores we computed with the HSD value. If the difference is larger than the HSD, then we say the difference is significant.

Groups 1 and 2 do not differ

Groups 1 and 3 differ

Groups 2 and 3 differ

6.8 SUMMARY

6.8.1 Hypothesis Testing

Hypothesis testing is an act in statistics whereby an analyst tests an assumption regarding a population parameter. The methodology employed by the analyst depends on the nature of the data used and the reason for the analysis.

Steps of Hypothesis Testing

All hypotheses are tested using a four-step process:

1. The first step is for the analyst to state the two hypotheses so that only one can be right.
2. The next step is to formulate an analysis plan, which outlines how the data will be evaluated.
3. The third step is to carry out the plan and physically analyze the sample data.
4. The fourth and final step is to analyze the results and either reject the null hypothesis, or state that the null hypothesis is plausible, given the data.

6.8.2 Chi-square Test

Chi-square tests are often used in hypothesis testing. The chi-square statistic compares the size any discrepancies between the expected results and the actual results, given the size of the sample and the number of variables in the relationship. For these tests, degrees of freedom are utilized to determine if a certain null hypothesis can be rejected based on the total number of variables and samples within the experiment. As with any statistic, the larger the sample size, the more reliable the results.

Applications of Chi Square Test:

- Testing the divergence of observed results from those expected on the hypothesis of equal probability (null hypothesis)
- Testing the divergence of observed results from those expected on the hypothesis of a normal distribution
- Chi-square test when our expectations are based on predetermined results
- The Chi-square test of independence in contingency tables

6.8.3 T-Test

t- test is a test of significance for the mean of a population in case of small samples with ($n < 30$) when the population variance is usually unknown.

Applications of t-test:

- t-test for significance of single mean, population variance being unknown
- t-test for the significance of the difference between two means, the population variances being equal
- t-test for significance of an observed sample correlation coefficient.

6.8.4 Z-Test

Z-test is a test of significance used in case of large samples ($n > 30$) when the population variance is known.

Applications of Z test:

- Test for single proportion
- Test of Significance for difference of proportions
- Test for significance of single mean
- Test of significance for difference of means

6.8.5 Tukey,S Q-Test:

This is a post hoc test used when the ANOVA is found to be significant.

6.9 GLOSSARY

Hypothesis: Any statement about the parameters of a population.

Null hypothesis: The *null hypothesis* is a clear statement about the relationship between two (or more) statistical objects. These objects may be measurements, distributions, or categories. Typically, the null hypothesis, as the name implies, states that *there is no relationship*.

Alternative hypothesis: Once the null hypothesis has been stated, it is easy to construct the *alternative hypothesis*. It is essentially the statement that the null hypothesis is false. In our example, the alternative hypothesis would be that the means of the two populations are not equal.

Significance: The *significance* level is a measure of the statistical strength of the hypothesis test. It is often characterized as the probability of incorrectly concluding that the null hypothesis is false. The significance level is something that you should specify up front. In applications, the significance level is typically one of three values: 10%, 5%, or 1%. A 1% significance level

represents the strongest test of the three. For this reason, 1% is a *higher* significance level than 10%.

Power: Related to significance, the *power* of a test measures the probability of correctly concluding that the null hypothesis is true. Power is not something that you can choose. It is determined by several factors, including the significance level you select and the size of the difference between the things you are trying to compare. Unfortunately, significance and power are inversely related. Increasing significance decreases power. This makes it difficult to design experiments that have both very high significance and power.

Test statistic: The *test statistic* is a single measure that captures the statistical nature of the relationship between observations you are dealing with. The test statistic depends fundamentally on the number of observations that are being evaluated. It differs from situation to situation.

Distribution of the test statistic: The whole notion of hypothesis rests on the ability to specify (exactly or approximately) the distribution that the test statistic follows. In the case of this example, the difference between the means will be approximately normally distributed (assuming there are a relatively large number of observations).

One-tailed vs. two-tailed tests: Depending on the situation, you may want (or need) to employ a *one-* or *two-tailed test*. These tails refer to the right and left tails of the distribution of the test statistic. A two-tailed test allows for the possibility that the test statistic is either very large or very small (negative is small). A one-tailed test allows for only one of these possibilities.

In an example where the null hypothesis states that the two population means are equal, you need to allow for the possibility that either one could be larger than the other. The test statistic could be either positive or negative. So, you employ a two-tailed test.

The null hypothesis might have been slightly different, namely that the mean of population 1 is larger than the mean of population 2. In that case, you don't need to account statistically for the situation where the first mean is smaller than the second. So, you would employ a one-tailed test.

Critical value: The *critical value* in a hypothesis test is based on two things: the distribution of the test statistic and the significance level. The critical value(s) refer to the point in the test statistic distribution that give the tails of the distribution an area (meaning probability) exactly equal to the significance level that was chosen.

Decision: Your *decision* to reject or accept the null hypothesis is based on comparing the test statistic to the critical value. If the test statistic exceeds the critical value, you should reject the null hypothesis. In this case, you would say that the difference between the two population means is significant. Otherwise, you accept the null hypothesis.

P-value: The *p-value* of a hypothesis test gives you another way to evaluate the null hypothesis. The p-value represents the highest significance level at which your particular test statistic would justify rejecting the null hypothesis. For example, if you have chosen a significance level of 5%,

and the p-value turns out to be .03 (or 3%), you would be justified in rejecting the null hypothesis.

6.10 SELF ASSESSMENT QUESTIONS

6.10.1 Fill In The Blanks:

1. The hypothesis which is under test for possible rejection is the _____ hypothesis.
2. There can be only _____ types of errors in taking a decision about H_0 .
3. A null hypothesis is rejected if the value of the test statistic lies in the _____ region.
4. Students t test is applicable in case of _____ samples.
5. Paired t test is applicable only when the samples are _____.
6. The _____ test is used to find the association between attributes.
7. The degrees of freedom for Chi Square test for a 3 x 3 contingency table is _____.
8. The _____ test is used to test the difference between proportions.
9. Power of a test is given by _____.
10. The level of significance is the probability of committing _____ error.

6.10.2 Multiple Choice Questions:

1. A hypothesis maybe classified as
 - a) simple
 - b) null
 - c) composite
 - d) All of the above
2. If the null hypothesis is false then which of the following is accepted?
 - a) Null Hypothesis
 - b) Positive Hypothesis
 - c) Negative Hypothesis
 - d) Alternative Hypothesis.
3. The rejection probability of Null Hypothesis when it is true is called as?
 - a) Level of Confidence
 - b) Level of Significance
 - c) Level of Margin
 - d) Level of Rejection
4. Consider a hypothesis H_0 where $\phi_0 = 5$ against H_1 where $\phi_1 > 5$. The test is?
 - a) One tailed
 - b) Two tailed

- c) Cross tailed
d) None of the above
5. Consider a hypothesis where H_0 where $\phi_0 = 23$ against H_1 where $\phi_1 \neq 23$. The test is?
a) One tailed
b) Two tailed
c) Cross tailed
d) None of the above
6. Which of the following is defined as the rule or formula to test a Null Hypothesis?
a) Test statistic
b) Population statistic
c) Null statistic
d) None of the Above
7. A test of difference between the observed and expected frequencies is
a) t-test
b) Chi-Square test
c) Z-test
d) None of the above
8. The degrees of freedom for t test for difference of means is
a) n_1
b) n_2
c) $(n_1 + n_2)/2$
d) n_1+n_2-2
9. The test used in case of large samples is
a) t-test
b) Chi-square test
c) Z-test
d) None of the above
10. The test used when ANOVA is found to be significant is
a) t-test
b) Tukey's Q test
c) Z-test
d) None of the above

6.10.3 Answers

Fill in the blanks:

- (1) Null,(2)Two, (3) critical/rejection,(4) Small,(5) Related/dependent,(6) Chi Square,(7) Four,
(8)Z, (9) $1-\beta$, (10)Type I

Multiple Choice Questions

(1)d,(2) d, (3)b, (4) a, (5)b, (6)a, (7)b, (8)d, (9)c,(10)b

6.11 REFERENCES:

- <https://en.wikipedia.org/wiki/Statistics>
- <http://www.fao.org>
- <https://online.stat.psu.edu/stat200>
- <https://nptel.ac.in>
- <https://swayam.gov.in>

6.12 SUGGESTED READINGS

1. Fundamentals of Statistics Vol-I: Goon Gupta Dasgupta
2. Fundamentals of Mathematical Statistics: SC Gupta & VK Kapoor
3. Mathematical Statistics: Kapoor & Saxena
4. Mathematical Statistics: OP Gupta & BD Gupta
5. Fundamentals of Statistics Vol-II: Goon Gupta Dasgupta
6. Fundamentals of Applied Statistics: SC Gupta & VK Kapoor
7. Programmed Statistics: BL Agarwal
8. Basic Statistics: B.L Agarwal

6.13 TERMINAL QUESTIONS:

1. Write notes on the following clearly mentioning their role in testing
 - a) Null and Alternative Hypothesis
 - b) Simple and Composite Hypothesis
 - c) Errors in Hypothesis Testing
 - d) Test Statistic
2. Give an overview of the steps involved in Hypothesis Testing.
3. Write a note on Chi Square Test and it's uses.
4. Write a note on t Test and it's uses.
5. Write a note on Z Test and it's uses.
6. A breeder claims that his variety of cotton contains, at the most, 40 % lint in seed cotton. 18 samples of 100 gms were taken and after ginning the following quantity of lint was found in the samples. Perform a t-test at 1% l.o.s to check the breeders claim.

36.3, 37.0, 36.6, 37.5, 37.5, 37.9, 37.8, 36.9, 36.7, 38.5, 37.9, 38.8, 37.5, 37.1, 37.0, 36.3, 36.7, 35.7 (Ans: $t=-14.89$, null hypothesis rejected)

7. The following table gives the average solar radiation on an inclined and horizontal surface at a particular place. Test whether the solar radiation varies among the surfaces at 5% l.o.s.

Month	Horizontal surface	Inclined surface
1	363	536
2	404	474
3	518	556
4	521	549
5	613	479
6	587	422
7	365	315
8	412	414
9	469	505
10	468	552
11	371	492
12	330	507

(Ans: $t=-0.94$, null hypothesis not rejected)

8. The following table gives the PI of 11 patients during and after seizure

During seizure	After seizure
0.45	0.60
0.54	0.65
0.48	0.63
0.62	0.78
0.48	0.63
0.60	0.80
0.45	0.69
0.46	0.62
0.35	0.68
0.40	0.50
0.44	0.57

Check whether there is a significant increase in the PI after seizure at 5 % l.o.s

(Ans: $t=8.72$, null hypothesis rejected)

9. Following table provides the number of teachers according to the time devoted to public activities by rank

Time devoted	Rank		
	Professor	Lecturer	Reader
Large	25	13	9
Some	62	53	49
None	12	34	43

Check whether the time devoted is independent of rank at 5 % l.o.s.

(Ans: $\chi^2 = 27.70$, null hypothesis rejected)

10. The number of automobile accidents per week in a certain community was as follows:

12, 8, 20, 2, 14, 10, 15, 6, 9, 4.

Are these frequencies in agreement with the belief that accident conditions were the same during this 10-week period?

(Ans: $\chi^2 = 26.6$, null hypothesis rejected)

UNIT-7- ANALYSIS OF VARIANCE (ANOVA)

Contents

- 7.1- Objectives
- 7.2- Introduction
- 7.3- Analysis of Variance'
- 7.4- One way analysis of variance
- 7.5- Two way classification
- 7.6- Summary
- 7.7- Self assessment Questions
- 7.8- References
- 7.9- Suggested readings
- 7.10- Terminal Questions

7.1 OBJECTIVES

After reading this unit you will be able to understand

1. ANOVA and its assumptions
2. One way and Two way ANOVA- its procedure and uses
3. Identify the information in the ANOVA table.
4. Interpret the results from ANOVA output.
5. Perform multiple comparisons and interpret the results, when appropriate.

7.2 INTRODUCTION

The t-test enables us to test the significance of the difference between two sample means but if we have a number of means and we need to test the hypothesis that the means are homogenous or that there is no difference among the means, then the technique known as Analysis of variance developed by Professor R. A. Fisher in 1920 is useful. Initially the technique was used in agricultural experiments but now days it is widely used in almost all the branches of agricultural and animal sciences. This technique is used to test whether the differences between the means of three or more populations is significant or not. By using the technique of analysis of variance, we can test whether moisture contents of paneer or khoa prepared by different methods or batches differ significantly or not. Analysis of variance thus enables us to test on the basis of sample observations whether the means of three or more populations are significantly different or not. Thus basic purpose of the analysis of variance is to test the homogeneity of several means and the technique consists in splitting up the total variation into component variation due to independent factors where each of the components give us the estimate of population variation. In other words, in this technique, the total sum of squares is decomposed into sum of squares due to independent factors and the remaining is attributed to random causes or commonly called due to error.

7.3 ANALYSIS OF VARIANCE

The term ‘Analysis of Variance’ was introduced by Prof. R.A. Fisher in 1920’s to deal with problem in the analysis of agronomical data. Variation is inherent in nature. The total variation in any set of numerical data is due to a number of causes which may be classified as:

- (i) Assignable causes, and (ii) Chance causes.

The variation due to assignable causes can be detected and measured whereas the variation due to chance causes is beyond the control of human being and cannot be accounted for separately.

7.3.1 Definition

According to Prof. R. A. Fisher, Analysis of Variance (ANOVA) is the 'Separation of variance ascribable to one group of causes from the variance ascribable to other group.' Thus, ANOVA consists in the estimation of the amount of variation due to each of the independent factors (causes) and the remaining due to chance factor (causes), the later being known as experimental error or simply error. The technique of the Analysis of variance consisting in splitting up the total variation into component variation due to independent factors where each of the components gives us the estimate of the population variance. The total sum of squares is broken up into sum of squares due to independent factors and the remaining is attributed to random causes or commonly called due to error. Consider, for instance, an industrial problem such as the following. A factory produces components, many machines being at work on the same operation. The process is not purely mechanical, the machine operators having an influence on the quality of the output. Moreover it is thought that on certain days of the week (e.g. Monday) the output is found to be of poorer quality than on other days (e.g. Friday). The quality therefore depends on at least three factors, the machine, the operator and the day of the week. There may be other factors in operation and some of the factors mentioned may have no significant effect. It will be possible by the technique of analysis of variance whether any of the above factors, or some combinations of these has an appreciable effect on the quality and also to estimate the contribution made by each factor to the overall variability in the production or quality of product. Thus the purpose of the analysis is to establish relations of cause and effect.

7.3.2 Assumptions in analysis of variance

For the validity of the F-test in ANOVA, the following assumptions are made:

- (i) The samples are drawn from the population randomly and independently.
- (ii) The data are quantitative in nature and are normally distributed. Parent population from which observations are taken is normal.
- (iii) Various treatments and environmental effects are additive in nature.
- (iv) The population from where the samples have been drawn should have equal variance σ^2 . This is known as **Homoscedasticity** and can be tested by Bartlett's test.

7.4 ONE WAY ANALYSIS OF VARIANCE

The simplest type of analysis of variance is known as one way analysis of variance, in which only one source of variation or factor of interest is controlled and its effect on the elementary units is observed. It is an extension of three or more samples of the t-test procedure for use with two independent samples. In other words t-test for use with two independent samples is a special case of one-way analysis of variance. In typical situation one way classification refers to the comparison of means of several univariate normal populations, having the same unknown variance σ^2 , on the basis of random samples selected from each population. The population means are denoted by $\mu_1, \mu_2, \dots, \mu_k$, if there are k populations. The one way analysis of variance is designed to test the null hypothesis:

$H_0 : \mu_1 = \mu_2 = \dots = \mu_k$ i.e. the arithmetic means of the population from which the k samples have been randomly drawn are equal to one another.

Let us suppose that N observations X_{ij} , ($i = 1, 2, \dots, k ; j = 1, 2, \dots, n_i$) of a random variable X are grouped on some basis, into k classes (T_1, T_2, \dots, T_k) of sizes n_1, n_2, \dots, n_k respectively $N = \sum_{i=1}^k n_i$, as exhibited below:

Table 7.4.1

Treatment					Means	Total
T_1	X_{11}	X_{12}	X_{1n_1}	$\bar{X}_{1\cdot}$	$T_{1\cdot}$
T_2	X_{21}	X_{22}	X_{2n_2}	$\bar{X}_{2\cdot}$	$T_{2\cdot}$
	\cdot	\cdot		\cdot	\cdot	\cdot
	\cdot	\cdot		\cdot	\cdot	\cdot
T_i	X_{i1}	X_{i2}	X_{in_i}	$\bar{X}_{i\cdot}$	$T_{i\cdot}$
	\cdot	\cdot	\cdot		\cdot	\cdot
	\cdot	\cdot	\cdot		\cdot	\cdot
T_k	X_{k1}	X_{k2}	X_{kn_k}	$\bar{X}_{k\cdot}$	$T_{k\cdot}$
						G

The total variation in the observations X_{ij} can be split into the following two components:

- (i) The variation between the classes or the variation due to different bases of classification, commonly known as treatments.
- (ii) The variation within the classes, i.e., the inherent variation of the random variable within the observations of class

The first type of variation is due to assignable causes which can be detected and controlled by human being and the second type of variation is due to chance causes which are beyond the control of human being.

The main object of analysis of variance technique is to examine if there is significant difference between the class means in view of the inherent variability within the separate classes.

In particular, let us consider the effect of k brands of yoghurt on price of yoghurt of N shops / retail stores (of same type) divided into k brands/classes of sizes n_1, n_2, \dots

, n_k respectively, $N = \sum_{i=1}^k n_i$.

Here the sources of variations are

- (i) Effect of the brands
- (ii) Error 'e' produced by numerous causes of such magnitude that they are not detected and identified with the knowledge that we have and they together produce a variation of random nature obeying Gaussian (Normal) law of errors.

7.4.1 Mathematical model

The linear mathematical model will be

$$X_{ij} = \mu_{ij} + e_{ij}$$

$$X_{ij} = \mu + \alpha_i + e_{ij} \quad (i=1,2,\dots,k) \quad (j=1,2,\dots,n_i)$$

where X_{ij} is the value of the variate in the j^{th} observation ($j=1,2,\dots,n_i$) belonging to i^{th} class ($i=1,2,\dots,k$)

μ is the general mean effect

α_i is the effect due to i^{th} class where $\alpha_i = \mu_i - \mu$

e_{ij} is random error which is assumed to be independently and normally distributed with mean zero and variance σ_e^2 .

Let the mean of k populations be $\mu_1, \mu_2, \dots, \mu_k$ then our aim is to test null hypothesis

$H_0 : \mu_1 = \mu_2 = \dots = \mu_k = \mu$ which reduces to $H_0 : \alpha_1 = \alpha_2 = \dots = \alpha_k = 0$.

$H_1 : \text{At least one pair of } \mu_i \text{'s is not equal.}$

7.4.2 Calculation of different sum of squares

a) **Total Sum of Squares (TSS)** =

$$\begin{aligned} \sum_i \sum_j (X_{ij} - \bar{X})^2 &= \sum_i \sum_j X_{ij}^2 - \frac{(\sum_i \sum_j X_{ij})^2}{N} \\ &= \sum_i \sum_j X_{ij}^2 - \frac{(G)^2}{N} \end{aligned}$$

where G is the grand total of all the observations and $N = n_1 + n_2 + \dots + n_k$,

The expression $\sum_i \sum_j X_{ij}^2$, i.e., sum of squares of all the observations is known as Raw Sum of

Squares (R.S.S.) and the expression $\frac{(G)^2}{N}$ is called Correction Factor (C.F.)

b) **Sum of Squares Among Classes (SSC)**: To find the SSC, divide the squares of sum of each class by their class size or number of observations in each class and find their sum and thereafter, Subtract the correction factor from this sum i.e.,

$$SSC = \left[\frac{T_1^2}{n_1} + \frac{T_2^2}{n_2} + \dots + \frac{T_k^2}{n_k} \right] - \frac{(G)^2}{N} = \sum_i \frac{T_i^2}{n_i} - C.F.$$

where T_i is the total of the observations pertaining to the i^{th} class.

c) **Sum of Squares within classes (SSE)**: It is obtained by subtracting sum of squares among the classes from the total sum of squares i.e., $SSE = TSS - SSC$.

This sum of squares is also called error sum of squares denoted by SSE.

d) **Mean Sum of Squares (M.S.S.)**: It is obtained by dividing sum of squares by their respective degrees of freedom.

e) **Analysis of Variance Table**

The results of the above calculations are presented in a table called Analysis of Variance or ANOVA table 7.4.2 as follows:

Table 7.4.2

Source of variation	Degree of Freedom (d.f.)	Sum of Squares (S.S.)	Mean Sum of Squares (M.S.S.)	F-Ratio
Among Classes	k-1	SSC	$S_C^2 = \frac{SSC}{k-1}$	$S_C^2/S_E^2 \sim F(k-1, N-k)$
Within Classes (Error)	N-k	SSE	$S_E^2 = \frac{SSE}{N-k}$	
Total	N-1	TSS		

If the calculated value of F is greater than the tabulated value of $F_{\alpha}(k-1, N-k)$, where α denotes the level of significance, the hypothesis H_0 , is rejected and can be inferred that the class effects are significantly different from one another.

Standard Error

- a) The estimated standard error of any class/treatment mean, say i^{th} treatment/class mean, is given by

$$SE_d = \sqrt{\frac{S_E^2}{n_i}}$$

Where S_E^2 is the mean sum of squares within samples or MSS(Error)

- b) The estimated standard error of the difference i^{th} and j^{th} treatment mean, is

$$SE_d = \sqrt{S_E^2 \left(\frac{1}{n_i} + \frac{1}{n_j} \right)}$$

Where n_i and n_j are the number of observations for i^{th} and j^{th} treatment/class

- c) If $n_i = n_j = n$ then S.E. of difference of means is

$$SE_d = \sqrt{\frac{2 S_E^2}{n}}$$

d) The Critical Difference (C.D.) or Least Significant Difference (L.S.D.) can be calculated as $C.D. = SE_d \times t_{\alpha, (N-k)}$ where α is level of significance and $(N-k)$ is the d.f. for error.

The treatment means are $\bar{X}_i = \frac{T_i}{n_i} \quad i=1,2,\dots,k$. These can be compared with the help of critical difference. Any two treatments means are said to differ significantly if the difference is larger than the critical difference (CD). The procedure of one way ANOVA is illustrated through the following example:

Example 1: The following table 7.4.3 gives the moisture contents of Paneer prepared by four methods Manual(M_1), Mechanical with pressure 10 pound/inch² (M_2); with pressure 12 pound/inch² (M_3) and pressure 15 pound/inch² (M_4).

Table 7.4.3

Methods			
M ₁	M ₂	M ₃	M ₄
50.3	54.1	57.5	52.3
52.2	53.7	56.3	53.2
52.5	55.5	55.8	53.6
51.7	54.6	56.9	53.4
52.6		55.8	53.8
		59.6	

Analyze the data to find whether the mean moisture content in paneer is different prepared by different methods.

Solution:

H₀ : μ₁=μ₂=μ₃=μ₄ i.e., the mean moisture content in paneer prepared by different methods is same.

H₁ : Mean moisture content in paneer prepared by at least two methods are not equal.

Prepare the following table 7.4.4 to calculate sum of squares due to different components:

Table 7.4.4

Methods	M ₁	M ₂	M ₃	M ₄	Total
Total (<i>T_i</i>)	259.30	217.90	341.90	266.30	G=1085.40
No. of observations (<i>n_i</i>)	5	4	6	5	20
Mean	51.8600	54.4750	56.9833	53.2600	

$$\text{Correction Factor (CF)} = \frac{(G)^2}{N} = \frac{(1085.4)^2}{20} = 58904.66$$

$$\text{Total Sum of Squares (TSS)} = \sum_i \sum_j X_{ij}^2 - CF$$

$$= (50.3)^2 + (52.2)^2 + \dots + (53.4)^2 + (53.8)^2 - 58904.66$$

$$= 59000.2200 - 58904.66 = 95.5620$$

Sum of Squares among Classes (SSC) or Sum of Squares between Methods =

$$\sum_i \frac{T_i^2}{n_i} - CF = \frac{(259.30)^2}{5} + \frac{(217.9)^2}{4} + \frac{(341.9)^2}{6} + \frac{(266.30)^2}{5} - 58904.66$$

$$= 58983.14 - 58904.66 = 78.4822$$

Sum of Squares within classes (SSE) or Sum of squares due to error:
 SSE=TSS-SSC= 95.5620-78.48217=17.0798

Prepare the following analysis of variance table 7.4.5:

Table 7.4.5 ANOVA Table

Source of variation	Degree of Freedom (d.f.)	Sum of Squares (S.S.)	Mean Sum of Squares (M.S.S.)	F-Ratio
Among Methods	4-1=3	78.4822	$S_c^2 = \frac{78.4822}{3}$ =26.1607	$F = \frac{26.1607}{1.0675}$ =24.5068
Within Methods (Error)	20-4=16	17.0798	$S_E^2 = \frac{17.0798}{16}$ =1.0675	
Total	20-1=19	95.5620		

From Fisher and Yate’s tables, F value for 3 and 16 d.f. at 5% level of significance is 3.2389 Since the observed value of F in the analysis of variance table is greater than the 5 % tabulated F value, it can be inferred that mean moisture content in paneer prepared by different methods differ significantly from one another.

Calculation of critical differences for comparison among various pairs of methods of preparing paneer

Table 7.4.6

Methods	M ₃	M ₂	M ₄	M ₁
Mean	56.9833	54.4750	53.2600	51.8600
No. of observations	6	4	5	5

C.D.(for comparing mean moisture content prepared by Method 3 and Method 2)

$$= \sqrt{1.0675 \left(\frac{1}{6} + \frac{1}{4} \right)} \times t_{5\%, 16 \text{ d.f.}}$$

$$= 0.6669 \times 2.12 = 1.4138$$

C.D. (for comparing mean moisture content prepared by Method 2 and Method 4)

$$= \sqrt{1.0675 \left(\frac{1}{4} + \frac{1}{5} \right)} \times t_{5\%, 16 \text{ d.f.}}$$

$$= 0.6931 \times 2.12 = 1.4693$$

C.D. (for comparing mean moisture content prepared by Method 4 and Method 1)

$$= \sqrt{1.0675 \left(\frac{1}{5} + \frac{1}{5} \right)} \times t_{5\%, 16 \text{ d.f.}}$$

$$= 0.6534 \times 2.12 = 1.3853$$

Conclusion

It can be concluded the moisture content of paneer prepared by different methods was found to be significantly different from each other. The mean moisture content was found to be maximum in method $M_3(56.9833)$ followed by method $M_2(54.4750)$ which is significantly different from each other. The next mean moisture contents was found for method $M_4(53.26)$ followed by method $M_1(51.86)$ which is significantly different from each other.

7.5 TWO WAY CLASSIFICATION

7.5.1 Introduction

In one way classification analysis of variance explained in the previous lesson the treatments constitute different levels of a single factor which is controlled in the experiment. There are, however, many situations in which the response variable of interest may be affected by more than one factor. For example milk yield of cow may be affected by differences in treatments i.e. feeds fed as well as differences in breed of the cows, moisture contents of butter prepared by churning cream may be affected with different levels of fat and churning speed etc. When two independent factors might have an effect on the response variable of interest, it is possible to design the test so that an analysis of variance can be used to test the effect of the two factors simultaneously. Such a test is called two factor analysis of variance. In a two way classification the data are classified according to two different criteria or factors. The procedure for analysis of variance is somewhat different than the one followed earlier while dealing with problems of one-way classification.

7.5.2 Two Way Classification

Let us plan the experiment in such a way so as to study the effect of two factors at a time in the same experiment. For each factor there will be a number of classes or levels. Let us consider the case when there are two factors which may affect the variate values be operators and machines. Suppose the N observations are classified into p categories (or classes) O_1, O_2, \dots, O_p according to Factor A (Operator) and into q categories M_1, M_2, \dots, M_q according to factor B (Machine) having pq combinations $A_i B_j$ ($O_i M_j$) $i=1,2,\dots,p$; $j=1,2,\dots,q$; often called cells. This scheme of classification according to two factors is called two way classifications and analysis is called two

way analysis of variance. The number of observations in each cells may be equal or different, but we shall consider the case of one observation per cell so that $N=pq$. i.e., total number of cells is N .

Let X_{ij} be the observation on the i^{th} level of Operator (O_i) and j^{th} level of Machine (M_j) $i=1,2,---,p$; $j=1,2,---,q$;

$$T_i = \sum_{j=1}^q X_{ij} = \text{Total or Sum of the observatios for } i^{th} \text{ operator } O_i$$

$$\bar{X}_i = \frac{1}{q} \sum_{j=1}^q X_{ij} = \frac{T_i}{q} = \text{mean of the observations for } i^{th} \text{ operator } O_i$$

$$T_j = \sum_{i=1}^p X_{ij} = \text{Sum of the observatios for } i^{th} \text{ machine } M_j$$

$$\bar{X}_j = \frac{1}{p} \sum_{i=1}^p X_{ij} = \frac{T_j}{p} = \text{mean of the observatios for } j^{th} \text{ machine } M_j$$

$$G \text{ or } T_{..} = \sum_{i=1}^p \sum_{j=1}^q X_{ij} = \sum_{i=1}^p T_i = \sum_{j=1}^q T_j = \text{Sum of all the observations or grand total}$$

$$\bar{X}_{..} = \frac{1}{N} \sum_{i=1}^p \sum_{j=1}^q X_{ij} = \frac{T_{..}}{N} = \text{mean of all the observations or overall mean}$$

These N observations, the marginal totals and their means can be represented in the tabular form as follows:

Table 7.5.1

Operators	Machines				Total	Mean
	M_1	M_2	M_j	M_q		
O_1	X_{11}	X_{12}	M_{1j}	X_{1q}	$T_{1.}$	\bar{X}_1
O_2	X_{21}	X_{22}	M_{2j}	X_{2q}	$T_{2.}$	\bar{X}_2
..
O_i	X_{i1}	X_{i2}	M_{ij}	X_{iq}	$T_{i.}$	\bar{X}_i
..
O_p	X_{p1}	X_{p2}	M_{pj}	X_{pq}	$T_{p.}$	\bar{X}_p
Total	$T_{.1}$	$T_{.2}$	$T_{.j}$	$T_{.q}$	$G = T_{..}$	
Mean	\bar{X}_1	\bar{X}_2	\bar{X}_j	\bar{X}_q		$\bar{X}_{..}$

7.5.2.1 Assumptions

- i. The observations are independent random variables having normal distributions with mean μ_{ij} and common but unknown variance σ^2 . Under this assumption model for this problem may be taken as $X_{ij} = \mu_{ij} + e_{ij}$. Where e_{ij} vary from observation to observation and are independent random variable values having normal distributions with mean zero and variance $\sigma^2 \Rightarrow E(X_{ij}) = \mu_{ij}$.
- ii. The observations in the p rows are independent random samples of size q from p normal populations having mean $\mu_1, \mu_2, \dots, \mu_p$ and a common variance σ^2 .
- iii. The observations in the q columns are independent random samples of size p from q normal populations with mean $\mu_1, \mu_2, \dots, \mu_q$ and a common variance σ^2 .
- iv. The effects are additive.

Here μ_i ($i=1,2,\dots,p$) are called fixed effect due to factor operators O_i ; μ_j ($j=1,2,\dots, q$) are fixed effect due to the factor machines M_j .

7.5.2.2 Mathematical model

Here the mathematical model can be written as

$$X_{ij} = \mu_{ij} + e_{ij}$$

$$X_{ij} = \mu + \alpha_i + \beta_j + e_{ij}$$

i) μ is the general mean effect given by $\mu = \sum \mu_{ij} / N$.

ii) α_i ($i=1, 2, \dots, p$) is the effect due to i^{th} operator

$$\text{where } \alpha_i = \mu_i - \mu; \mu_i = \frac{1}{p} \sum_{j=1}^q \mu_{ij} (i = 1, 2, \dots, p) \text{ Obviously } \sum_{i=1}^p \alpha_i = 0$$

iii) β_j ($j=1, 2, \dots, q$) is the effect due to j^{th} machine

$$\text{where } \beta_j = \mu_j - \mu; \mu_j = \frac{1}{p} \sum_{i=1}^p \mu_{ij} (j = 1, 2, \dots, q) \text{ Obviously } \sum_{j=1}^q \beta_j = 0$$

iv) e_{ij} 's are independently normally distributed with mean zero and variance σ_e^2 i.e. $e_{ij} \sim N(0, \sigma_e^2)$

$$\sum_i \alpha_i = \sum_j \beta_j = 0$$

7.5.2.3 Null hypothesis

We set up the null hypothesis, that the operators and machines are homogeneous. In other words, the null hypothesis for operators and machines are respectively:

$$H_{01}: \mu_1 = \mu_2 = \dots = \mu_p.$$

or

$$\alpha_1 = \alpha_2 = \dots = \alpha_p = 0$$

i. e. mean output obtained from different operators is same.

$$H_{02}: \mu_{.1} = \mu_{.2} = \dots = \mu_{.q}$$

or

$$\beta_1 = \beta_2 = \dots = \beta_q = 0$$

i. e. mean output obtained from different machines is same.

Against the corresponding hypothesis

H_{11} : at least two of the means μ'_i 's are not equal

H_{12} : at least two of the means μ'_j 's are not equal

7.5.2.4 Computations of different sum of squares

$$\begin{aligned} \text{a) Total Sum of Squares (TSS)} &= \sum_i \sum_j (X_{ij} - \bar{X}_{..})^2 = \sum_i \sum_j X_{ij}^2 - \frac{(\sum_i \sum_j X_{ij})^2}{N} \\ &= \sum_i \sum_j X_{ij}^2 - \frac{(G)^2}{N} \end{aligned}$$

where G is the grand total of all the observations and $N = pq$. The expression $\sum_i \sum_j X_{ij}^2$ i.e., sum of squares of all the observations is known as Raw Sum of Squares (R.S.S.) and the expression $(G)^2/N$ is called Correction Factor (CF)

b) Sum of Squares due to factor A (Operators) denoted by SSA

To find the sum of squares due to factor A (SSA) i.e., sum of squares among the rows (SSR) divide the squares of sum of each row by the number of observations in respective rows and find their sum and thereafter, subtract the correction factor from this sum i.e.,

$$\text{SSA(SSR)} = \left[\frac{T_1^2}{q} + \frac{T_2^2}{q} + \dots + \frac{T_k^2}{q} \right] - \frac{(G)^2}{N} = \sum_i \frac{T_i^2}{q} - \text{CF}$$

where T_i is the total of the observations pertaining to the i^{th} row.

c) Sum of Squares due to factor B (Machines) denoted by SSB

To find the sum of squares due to factor B (SSB) i.e. sum of squares among the columns (SSC) divide the squares of sum of each column by number of observations in respective columns and find their sum and thereafter, subtract the correction factor from this sum i.e.,

$$\text{SSB(SSC)} = \left[\frac{T_{.1}^2}{p} + \frac{T_{.2}^2}{p} + \dots + \frac{T_{.k}^2}{p} \right] - \frac{(G)^2}{N} = \sum_j \frac{T_{.j}^2}{p} - \text{CF}$$

where $T_{.j}$ is the total of the observations pertaining to the j^{th} column.

d) **Sum of Squares due to residuals or error denoted by SSE**

The sum of squares of the residuals is obtained by subtracting sum of squares due to Factor A (SSA) and sum of squares due to factor B (SSB) from the total sum of squares (TSS) i.e., $SSE=TSS-SSA-SSB$.

This sum of squares is also called error sum of squares denoted by SSE.

Prepare the analysis of variance table as follows:

Table 7.5.2 ANOVA Table

Source of variation	d.f.	S.S.	M.S.S.	F-Ratio
Among levels of factor A (Operators)	$(p - 1)$	SSA	$S_A^2 = \frac{SSA}{p - 1}$	$F_1 = \frac{S_A^2}{S_E^2}$
Among levels of factor B (Machines)	$(q-1)$	SSB	$S_B^2 = \frac{SSB}{q - 1}$	$F_2 = \frac{S_B^2}{S_E^2}$
Error	$(p - 1)(q - 1)$	SSE	$S_E^2 = \frac{SSE}{(p - 1)(q - 1)}$	
Total	$pq - 1$			

Interpretation

By comparing the values of F_1 and F_2 with the tabulated value of F for respective d.f. and at α level of significance , the null hypothesis of the homogeneity of various factor A (Operators) and various factor B (Machines) may be rejected or accepted at the desired level of significance .

Standard error

- a) The estimated standard error of the difference between means of factor A i.e., between means of two operators is


$$SE_d = \sqrt{\frac{2S_E^2}{q}}$$

- b) The estimated standard error of the difference between means of factor B i.e., between means of two machines is

$$SE_d = \sqrt{\frac{2S_E^2}{p}}$$

- c) The Critical Difference (C.D.) or Least Significant Difference (L.S.D.) can be calculated as

C.D. = $SE_d \times t_{\alpha, (p-1)(q-1)}$ where SE_d is the S.E. of difference between two means, α is level of significance and $(p-1)(q-1)$ is the d.f. for error .

The treatment means are $\bar{X}_i = \frac{T_i}{q} \forall, i = 1, 2, \dots, p$ and $\bar{X}_j = \frac{T_j}{p} \forall, j = 1, 2, \dots, q$ 

These can be compared with the help of critical difference. Any two treatments means are said to differ significantly if the difference is larger than the critical difference (CD). The procedure of two way ANOVA is illustrated through the following example:

Example 1: The average particle size of dried ice-cream mix spray powder dried by varying in-let temperature and automiser speed was measured in an experiment with 6 in-let temperatures and 4 automiser speed. The results obtained from the experiment are given below:

Table 7.5.3

Automiser Speed	In-let Temperatures					
	T ₁	T ₂	T ₃	T ₄	T ₅	T ₆
S ₁	35.7	39.0	42.1	25.1	29.9	27.3
S ₂	32.9	33.6	37.7	24.0	23.2	24.3
S ₃	35.6	32.5	37.4	21.0	24.9	23.1
S ₄	30.7	35.8	40.1	26.3	28.3	26.4

Analyze the data and discuss whether there is any significant difference between in-let temperature and automiser speed on particle size of ice-cream mix powder?

Solution:

$H_{0A} : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6$ i.e., the mean particle size of ice-cream mix powder at different in-let temperature is same.

$H_{1A} : \text{At least two of the means } \mu'_i \text{ s are not equal}$

$H_{0B} : \mu_{.1} = \mu_{.2} = \mu_{.3} = \mu_{.4}$ i.e., the mean particle size of ice-cream mix powder at different automiser speeds is same.

$H_{1B} : \text{At least two of the means } \mu'_j \text{ s are not equal}$

Prepare the following two way table:

Table 7.5.4 Calculation of Treatments totals, means and the grand total

Automiser Speed	In-let Temperatures						Total	Mean
	T ₁	T ₂	T ₃	T ₄	T ₅	T ₆		
S ₁	35.7	39.0	42.1	25.1	29.9	27.3	T _{1.} =199.1	$\bar{X}_1=33.1833$
S ₂	32.9	33.6	37.7	24.0	23.2	24.3	T _{2.} =175.7	$\bar{X}_2=29.2833$
S ₃	35.6	32.5	37.4	21.0	24.9	23.1	T _{3.} =174.5	$\bar{X}_3=29.0833$
S ₄	30.7	35.8	40.1	26.3	28.3	26.4	T _{4.} =187.6	$\bar{X}_4=31.2667$
Total	T _{.1} = 134.9	T _{.2} = 140.9	T _{.3} = 157.3	T _{.4} = 96.4	T _{.5} = 106.3	T _{.6} = 101.1	G=736.9	
Mean	$\bar{X}_{.1} =$ 33.73	$\bar{X}_{.2} =$ 35.23	$\bar{X}_{.3} =$ 39.33	$\bar{X}_{.4} =$ 24.10	$\bar{X}_{.5} =$ 26.58	$\bar{X}_{.6} =$ 25.28		

$$\text{Correction factor} = (G)^2 / N = (736.9)^2 / 24 = 22625.9004$$

$$\begin{aligned} \text{Total Sum of Squares (TSS)} &= \sum_i \sum_j X_{ij}^2 - \frac{(G)^2}{N} \\ &= (35.7)^2 + (39.0)^2 + \dots + (28.3)^2 + (26.4)^2 - 22625.9004 \\ &= 23513.27 - 22625.9004 = 887.3696 \end{aligned}$$

a) Sum of Squares due to factor A (Speed)

$$\begin{aligned} \text{SSA (SS due to speed)} &= \sum_i \frac{T_i^2}{q} - \text{C.F.} \\ &= \frac{(199.1)^2 + (175.7)^2 + (174.5)^2 + (187.6)^2}{6} - 22625.9004 \\ &= 22692.5517 - 22625.9004 = 66.6512 \end{aligned}$$

b) Sum of Squares due to factor B (In-let temperature)

$$\begin{aligned} \text{SSB (SS due to inlet temperature)} &= \sum_j \frac{T_j^2}{p} - \text{C.F.} \\ &= \frac{(134.9)^2 + (140.9)^2 + (157.3)^2 + (96.4)^2 + (106.3)^2 + (101.1)^2}{4} - 22625.9004 \\ &= 23401.9925 - 22625.9004 = 776.0921 \end{aligned}$$

c) Sum of Squares due to residuals (SSE)

$$SSE = TSS - SSA - SSB = 887.3696 - 66.6512 - 776.0921 = 44.62625$$

Prepare the following ANOVA table:

Table 7.5.5 ANOVA Table

Source of variation	d.f.	S.S.	M.S.S.	F-Ratio
Among levels of factor A (Speed)	(4-1)=3	66.65125	$S_A^2 = \frac{66.6512}{3}$ =22.2171	$F_1 = \frac{22.2171}{2.9751}$ =7.4677
Among levels of factor B (Temperature)	(6-1)=5	776.0921	$S_B^2 = \frac{776.0921}{5}$ =155.2184	$F_2 = \frac{155.2184}{2.9751}$ =52.1728
Error	(4-1)(6-1) =15	44.62625	$S_E^2 = \frac{44.62625}{15}$ =2.9751	
Total	24 - 1=23	887.3696		

From Fisher and Yate's tables, tabulated F values for 3 and 15 d.f. and for 5 and 15 d.f. at 5% level of significance are 3.2874 and 2.9013 respectively. Since the observed values of F for factor A (automiser speed) and factor B (in-let temperature) in the analysis of variance table are greater than the respective 5% tabulated F value, F_1 and F_2 are significant at 5% level of significance. Hence both the null hypothesis H_{0A} and H_{0B} are rejected at 5% level of significance.

Critical difference

C.D. (For comparison of different speed)

$$= \sqrt{\frac{2 \times 2.9751}{6}} \times t_{0.05, 15 \text{ d.f.}} = 0.9958 \times 2.131 = 2.1221$$

C.D. (For comparison of different in-let temperature)

$$= \sqrt{\frac{2 \times 2.9751}{4}} \times t_{0.05, 15 \text{ d.f.}} = 1.2196 \times 2.131 = 2.5991$$

Conclusion

It can be concluded that mean particle size of ice-cream mix powder differ significantly at various levels of in-let temperature as well as at various automiser speed levels. The mean particle size of ice-cream mix powder was found maximum at different auto miser speed S_1 (33.1833) which is at par with speed S_4 (31.2667). Similar argument holds for speed S_4 and S_2 as well as for the S_2 and S_3 speeds. Similarly the mean particle size of ice-cream mix powder was found maximum in temperature T_3 (39.33) followed by temperature T_2 (35.23) and

$T_1(33.73)$, both are statistically at par with each other. Similar argument holds for temperature T_5 , T_6 and T_4 .

7.6 SUMMARY

ANOVA

Analysis of variance enables us to test on the basis of sample observations whether the means of three or more populations are significantly different or not. Thus basic purpose of the analysis of variance is to test the homogeneity of several means and the technique consists in splitting up the total variation into component variation due to independent factors where each of the components give us the estimate of population variation. In other words, in this technique, the total sum of squares is decomposed into sum of squares due to independent factors and the remaining is attributed to random causes or commonly called due to error.

ONE WAY ANOVA

A one-way ANOVA is a type of statistical test that compares the variance in the group means within a sample whilst considering only one independent variable or factor. It is a hypothesis-based test, meaning that it aims to evaluate multiple mutually exclusive theories about our data. A one-way ANOVA compares three or more than three categorical groups to establish whether there is a difference between them. Within each group there should be three or more observations and the means of the samples are compared.

In a one-way ANOVA there are two possible hypotheses.

- The null hypothesis (H_0) is that there is no difference between the groups and equality between means.
- The alternative hypothesis (H_1) is that there is a difference between the means and groups.

The various sum of squares are calculated and then the results are presented in the ANOVA table,

TWO WAY ANOVA

In the two-way ANOVA each sample is defined in two ways, and resultingly put into two categorical groups. The two-way ANOVA therefore examines the effect of two factors on a dependent variable and also examines whether the two factors affect each other to influence the continuous variable.

Because the two-way ANOVA consider the effect of two categorical factors, and the effect of the categorical factors on each other, there are two pairs of null or alternative hypotheses for the two-way ANOVA

- H_0 : The means of categorical factor 1 are equal
- H_1 : The mean of at least one category of factor 1 is different
- H_0 : The means of categorical factor 2 are equal
- H_1 : The means of at least one category of factor 2 is different

The various sum of squares are calculated and then the results are presented in the ANOVA table,

7.7 SELF ASSESSMENT QUESTIONS:

7.7.1 Fill in the blanks:

1. The term ANOVA was given by _____.
2. The full form of ANOVA is _____.
3. ANOVA is defined as “Separation of _____ ascribable due to.....causes”.
4. ANOVA is used when we want to compare the means of more than _____ groups.
5. One of the assumptions of ANOVA implies that the groups being compared should be _____.
6. In one way ANOVA, the number of degrees of freedom for k Treatment/ Factors is _____.
7. In one way ANOVA, if the calculated F exceeds the tabulated F, then the null hypothesis is _____.
8. In two way ANOVA, there are _____ categorical factors that are compared.
9. In two way ANOVA, _____ null hypothesis are tested.
10. In one way ANOVA, p categories of factor A and q categories of factor B the error sum of squares is _____.

7.7.2 Multiple choice questions:

- 1) Analysis of variance is a statistical method of comparing the _____ of several populations.
 - a) Means
 - b) Variances
 - c) Standard Deviations
 - d) None of The Above
- 2) The _____ sum of squares measures the variability of the observed values around their respective treatment means.
 - a) Error
 - b) Total
 - c) Treatment
 - d) Interaction
- 3) Which of the following is an assumption of one-way ANOVA comparing samples from three or more experimental treatments?
 - a) The samples associated with each population are randomly selected and are independent from all other samples
 - b) The response variable within each of the k populations have equal variances
 - c) All the response variables within the k populations follow a normal distributions

- d) All of the above
- 4) When the k population means are truly different from each other, it is likely that the average error deviation:
- a) is relatively small compared to the average treatment deviations
 - b) is about equal to the average treatment deviation
 - c) is relatively large compared to the average treatment deviations
 - d) none of the above
- 5) The _____.sum of squares measures the variability of the sample treatment means around the overall mean.
- a) Error
 - b) Interaction
 - c) Total
 - d) Treatment
- 6) Which test is used by ANOVA
- a) Z test
 - b) t test
 - c) F test
 - d) None of the above
- 7) In One Way ANOVA the number of factors considered is
- a) One
 - b) Two
 - c) three
 - d) none of the above
- 8) In one way ANOVA with 5 treatments, the degree of freedom for treatment sum of squares is
- a) 5
 - b) 4
 - c) 6
 - d) None of the above
- 9) In two way ANOVA the total variation is divided into _____ parts
- a) 2
 - b) 5
 - c) 6
 - d) None of the above
10. In two way ANOVA with factor 1 having 5 categories and factor 2 having 4 categories the error degree of freedom is
- a) 12
 - b) 15

- c) 19
- d) None of the above

7.7.3 Answers

Fill in the blanks:

(1) Fisher,(2)Analysis of Variance, (3)variation,(4)two,(5)independent,(6)k-1,(7)rejected, (8)two, (9)two, (10)(p-1)(q-1)

Multiple Choice Questions

(1)a,(2)a, (3)d, (4)a, (5)d, (6)c, (7)a, (8)b, (9)a,(10)a

7.8 REFERENCES:

- <https://en.wikipedia.org/wiki/Statistics>
- <http://www.fao.org>
- <https://online.stat.psu.edu/stat200>
- <https://nptel.ac.in>
- <https://swayam.gov.in>

7.9 SUGGESTED READINGS:

1. Fundamentals of Statistics Vol-I: Goon Gupta Dasgupta
2. Fundamentals of Mathematical Statistics: SC Gupta & VK Kapoor
3. Mathematical Statistics: Kapoor & Saxena
4. Mathematical Statistics: OP Gupta & BD Gupta
5. Fundamentals of Statistics Vol-II: Goon Gupta Dasgupta
6. Fundamentals of Applied Statistics: SC Gupta & VK Kapoor
7. Programmed Statistics: BL Agarwal
8. Basic Statistics: B.L Agarwal

7.10 TERMINAL QUESTIONS:

1. Define ANOVA. Give the assumptions of ANOVA.
2. Write a note on One way ANOVA clearly mentioning the mathematical model, hypothesis, various sum of squares and the ANOVA table.
3. Write a note on Two way ANOVA clearly mentioning the mathematical model, hypothesis, various sum of squares and the ANOVA table.
4. From the data given below find out if the means of the various samples differ significantly among themselves at 5% l.o.s

Sample 1	Sample 2	Sample 3	Sample 4
9	13	19	14
11	12	13	10
13	10	17	13
9	15	7	17
8	5	9	16

(Ans: $F=1.28$, non significant)

5. To study the performance of three detergents at three different temperatures the following whiteness readings were obtained

Water temperature	Detergent A	Detergent B	Detergent C
Cold	57	55	67
Warm	49	52	68
Hot	54	46	58

Perform a two way ANOVA at 5 % l.o.s.

(Ans: $F_{\text{Detergent}}= 9.845$, Significant and $F_{\text{temper}}=2.381$, Non Significant).

UNIT-8- EXPERIMENTAL DESIGNS AND THEIR ANALYSIS

Contents

- 8.1- Objectives
- 8.2- Introduction
- 8.3- Design of experiment
- 8.4- Basic principles of experimental design
- 8.5- The types of experimental design
- 8.6- Randomized block design
- 8.7- Completely Randomized Block Design
- 8.8- Split-plot design
- 8.9- Complete and incomplete block designs
- 8.10- Augmented designs
- 8.11- Grid and honeycomb designs
- 8.12- Summary
- 8.13- Self assessment Questions
- 8.14- References
- 8.15- Suggested readings
- 8.16- Terminal Questions

8.1 OBJECTIVES

After reading this unit you will be able to understand

1. Basic definitions and concepts of Design of Experiment.
2. Principles of Design of Experiment.
3. Types of Designs used in Statistical Analysis
4. Conditions under which different Designs are used

8.2 INTRODUCTION

Design of experiment means how to design an experiment in the sense that how the observations or measurements should be obtained to answer a query in a valid, efficient and economical way. The designing of the experiment and the analysis of obtained data are inseparable. If the experiment is designed properly keeping in mind the question, then the data generated is valid and proper analysis of data provides the valid statistical inferences. If the experiment is not well designed, the validity of the statistical inferences is questionable and may be invalid.

It is important to understand first the basic terminologies used in the experimental design.

Experimental unit:

For conducting an experiment, the experimental material is divided into smaller parts and each part is referred to as an experimental unit. The experimental unit is randomly assigned to treatment is the experimental unit. The phrase “randomly assigned” is very important in this definition.

Experiment: A way of getting an answer to a question which the experimenter wants to know.

Treatment: Different objects or procedures which are to be compared in an experiment are called treatments.

Sampling unit: The object that is measured in an experiment is called the sampling unit. This may be different from the experimental unit.

Factor: A factor is a variable defining a categorization. A factor can be fixed or random in nature. A factor is termed as a fixed factor if all the levels of interest are included in the experiment.

A factor is termed as a random factor if all the levels of interest are not included in the experiment and those that are can be considered to be randomly chosen from all the levels of interest.

Replication: It is the repetition of the experimental situation by replicating the experimental unit

Experimental error: The unexplained random part of the variation in any experiment is termed as experimental error. An estimate of experimental error can be obtained by replication.

Treatment design: A treatment design is the manner in which the levels of treatments are arranged in an experiment.

Suppose some varieties of fish food is to be investigated on some species of fishes. The food is placed in the water tanks containing the fishes. The response is the increase in the weight of fish. The experimental unit is the tank, as the treatment is applied to the tank, not to the fish. Note that if the experimenter had taken the fish in hand and placed the food in the mouth of fish, then the fish would have been the experimental unit as long as each of the fish got an independent scoop of food.

8.3 DESIGN OF EXPERIMENT:

One of the main objectives of designing an experiment is how to verify the hypothesis in an efficient and economical way. In the contest of the null hypothesis of equality of several means of normal populations having the same variances, the analysis of variance technique can be used. Note that such techniques are based on certain statistical assumptions. If these assumptions are violated, the outcome of the test of a hypothesis then may also be faulty and the analysis of data may be meaningless. So the main question is how to obtain the data such that the assumptions are met and the data is readily available for the application of tools like analysis of variance. The designing of such a mechanism to obtain such data is achieved by the design of the experiment. After obtaining the sufficient experimental unit, the treatments are allocated to the experimental units in a random fashion. Design of experiment provides a method by which the treatments are placed at random on the experimental units in such a way that the responses are estimated with the utmost precision possible.

8.4 BASIC PRINCIPLES OF EXPERIMENTAL DESIGN:

There are three basic principles of design which were developed by Sir Ronald A. Fisher.

- (i) Randomization
- (ii) Replication
- (iii) Local control

i) Randomization

The principle of randomization involves the allocation of treatment to experimental units at random to avoid any bias in the experiment resulting from the influence of some extraneous unknown factor that may affect the experiment. In the development of analysis of variance, we assume that the errors are random and independent. In turn, the observations also become random. The principle of randomization ensures this.

The random assignment of experimental units to treatments results in the following outcomes.

- a) It eliminates systematic bias.
- b) It is needed to obtain a representative sample from the population.

- c) It helps in distributing the unknown variation due to confounded variables throughout the experiment and breaks the confounding influence.

Randomization forms a basis of a valid experiment but replication is also needed for the validity of the experiment.

If the randomization process is such that every experimental unit has an equal chance of receiving each treatment, it is called **complete randomization**.

ii) Replication:

In the replication principle, any treatment is repeated a number of times to obtain a valid and more reliable estimate than which is possible with one observation only. Replication provides an efficient way of increasing the precision of an experiment. The precision increases with the increase in the number of observations. Replication provides more observations when the same treatment is used, so it increases precision. For example, if the variance of x is σ^2 than variance of the sample mean \bar{x} based on n

$$\text{observation is } \frac{\sigma^2}{n}. \text{ So as } n \text{ increases, } \text{Var}(\bar{x}) \text{ decreases.}$$

iii) Local control (error control)

The replication is used with local control to reduce the experimental error. For example, if the experimental units are divided into different groups such that they are homogeneous within the blocks, then the variation among the blocks is eliminated and ideally, the error component will contain the variation due to the treatments only. This will, in turn, increase the efficiency.

8.5 THE TYPES OF EXPERIMENTAL DESIGN:

The types of experimental research design are determined by the way the researcher assigns subjects to different conditions and groups. They are of 3 types, namely; pre-experimental, quasi-experimental, and true experimental research.

8.5.1 Pre-experimental Research Design

In pre-experimental research design, either a group or various dependent groups are observed for the effect of the application of an independent variable which is presumed to cause change. It is the simplest form of experimental research design and is treated with no control group.

Although very practical, experimental research is lacking in several areas of the true-experimental criteria. The pre-experimental research design is further divided into three types

a) One-shot Case Study Research Design: In this type of experimental study, only one dependent group or variable is considered. The study is carried out after some treatment which was presumed to cause change, making it a posttest study.

b) One-group Pretest-posttest Research Design: This research design combines both posttest and pretest study by carrying out a test on a single group before the treatment is administered and after the treatment is administered. With the former being administered at the beginning of treatment and later at the end.

c) Static-group Comparison: In a static-group comparison study, 2 or more groups are placed under observation, where only one of the groups is subjected to some treatment while the other groups are held static. All the groups are post-tested, and the observed differences between the groups are assumed to be a result of the treatment.

8.5.2 Quasi-experimental Research Design

The word "quasi" means partial, half, or pseudo. Therefore, the quasi-experimental research bearing a resemblance to the true experimental research, but not the same. In quasi-experiments, the participants are not randomly assigned, and as such, they are used in settings where randomization is difficult or impossible.

This is very common in educational research, where administrators are unwilling to allow the random selection of students for experimental samples.

Some examples of quasi-experimental research design include; the time series, no equivalent control group design, and the counterbalanced design.

8.5.3 True Experimental Research Design

The true experimental research design relies on statistical analysis to approve or disprove a hypothesis. It is the most accurate type of experimental design and may be carried out with or without a pretest on at least 2 randomly assigned dependent subjects.

The true experimental research design must contain a control group, a variable that can be manipulated by the researcher, and the distribution must be random. The classification of true experimental design includes:

- **The posttest-only Control Group Design:** In this design, subjects are randomly selected and assigned to the 2 groups (control and experimental), and only the experimental group is treated. After close observation, both groups are post-tested, and a conclusion is drawn from the difference between these groups.
- **The pretest-posttest Control Group Design:** For this control group design, subjects are randomly assigned to the 2 groups, both are presented, but only the experimental group is treated. After close observation, both groups are post-tested to measure the degree of change in each group.
- **Solomon four-group Design:** This is the combination of the pretest-only and the pretest-posttest control groups. In this case, the randomly selected subjects are placed into 4 groups.

The first two of these groups are tested using the posttest-only method, while the other two are tested using the pretest-posttest method.

8.6 RANDOMIZED BLOCK DESIGN

If a large number of treatments are to be compared, then a large number of experimental units are required. This will increase the variation among the responses and CRD may not be appropriate to use. In such a case when the experimental material is not homogeneous and there are v treatments to be compared, then it may be possible to

- Group the experimental material into blocks of sizes v units.
- Blocks are constructed such that the experimental units within a block are relatively homogeneous and resemble to each other more closely than the units in the different blocks.
- If there are b such blocks, we say that the blocks are at b levels. Similarly, if there are v treatments, we say that the treatments are at v levels. The responses from the b levels of blocks and v levels of treatments can be arranged in a two-way layout. The observed data set is arranged as follows:

	Blocks							Block Totals
		1	2	...	i	...	b	
Treatments	1	y_{11}	y_{21}	...	y_{i1}	...	y_{b1}	$B_1 = y_{o1}$
	2	y_{12}	y_{22}	...	y_{i2}	...	y_{b2}	$B_2 = y_{o2}$

	J	y_{1j}	y_{2j}	...	y_{ij}	...	y_{bj}	$B_j = y_{oj}$

V	y_{1v}	y_{2v}	...	y_{iv}	...	y_{bv}	$B_b = y_{ob}$	
Treatment Totals	$T_1 = y_{1o}$	$T_2 = y_{2o}$...	$T_i = y_{io}$...	y_{vo}	Grand Total G = y_{oo}	

Layout:

A two-way layout is called a randomized block design (RBD) or a randomized complete block design (RCB) if, within each block, the v treatments are randomly assigned to v experimental units such that each of the $v!$ ways of assigning the treatments to the units has the same

probability of being adopted in the experiment and the assignment in different blocks are statistically independent.

The RBD utilizes the principles of design - randomization, replication and local control - in the following way:

1. Randomization:

- Number the v treatments $1, 2, \dots, v$.
- Number the units in each block as $1, 2, \dots, v$.
- Randomly allocate the v treatments to v experimental units in each block.

2. Replication

Since each treatment is appearing in each block, so every treatment will appear in all the blocks. So each treatment can be considered as if replicated the number of times as the number of blocks. Thus in RBD, the number of blocks and the number of replications are same.

3. Local control

Local control is adopted in RBD in the following way:

- First form the homogeneous blocks of the experimental units.
- Then allocate each treatment randomly in each block.

The error variance now will be smaller because of homogeneous blocks and some variance will be parted away from the error variance due to the difference among the blocks.

Example: Suppose there are 7 treatments T_1, T_2, \dots, T_7 in 4 blocks. So one possible layout of the assignment of 7 treatments to 4 different blocks in an RBD is as follows

Block 1	T_2	T_7	T_3	T_5	T_1	T_4	T_6
Block 2	T_1	T_6	T_7	T_4	T_5	T_3	T_2
Block 3	T_7	T_5	T_1	T_6	T_4	T_2	T_3
Block 4	T_4	T_1	T_5	T_6	T_2	T_7	T_3

Analysis

There are two factors which affect the outcome – treatment effect and block effect. So the set-up of two-way analysis of variance discussed in unit 7 is to be used.

8.7 COMPLETELY RBD

The most common block design is the Completely Randomized Block Design. In this design all the treatments are applied randomly to plots within each block. Here the number of units within each block is equal to the the number of treatments.

8.7.1 Layout:

Example: Suppose there are 4 treatments T_1, T_2, \dots, T_4 in 4 blocks. So one possible layout of the assignment of 4 treatments to 4 different blocks in an RBD is as follows

Block 1	T_2	T_1	T_3	T_4
Block 2	T_1	T_4	T_2	T_3
Block 3	T_4	T_3	T_1	T_2
Block 4	T_3	T_2	T_4	T_1

8.7.2 Analysis

There are two factors which affect the outcome – treatment effect and block effect . So the set-up of two-way analysis of variance discussed in unit 7 is to be used.

8.8 SPLIT-PLOT DESIGN

In field experiments certain factors may require larger plots than for others. For example, experiments on irrigation, tillage, etc requires larger areas. On the other hand experiments on fertilizers, etc may not require larger areas. To accommodate factors which require different sizes of experimental plots in the same experiment, split plot design has been evolved.

In this design, larger plots are taken for the factor which requires larger plots. Next each of the larger plots is split into smaller plots to accommodate the other factor. The different treatments are allotted at random to their respective plots. Such arrangement is called split plot design.

In split plot design the larger plots are called main plots and smaller plots within the larger plots are called as sub plots. The factor levels allotted to the main plots are main plot treatments and the factor levels allotted to sub plots are called as sub plot treatments.

Layout and analysis of variance table

First the main plot treatment and sub plot treatment are usually decided based on the needed precision. The factor for which greater precision is required is assigned to the sub plots.

The replication is then divided into number of main plots equivalent to main plot treatments. Each main plot is divided into subplots depending on the number of sub plot treatments. The main plot treatments are allocated at random to the main plots as in the case of RBD. Within

each main plot the sub plot treatments are allocated at random as in the case of RBD. Thus randomization is done in two stages. The same procedure is followed for all the replications independently.

The analysis of variance will have two parts, which correspond to the main plots and sub-plots. For the main plot analysis, replication X main plot treatments table is formed. From this two-way table sum of squares for replication, main plot treatments and error (a) are computed. For the analysis of sub-plot treatments, main plot X sub-plot treatments table is formed. From this table the sums of squares for sub-plot treatments and interaction between main plot and sub-plot treatments are computed. Error (b) sum of squares is found out by residual method. The analysis of variance table for a split plot design with m main plot treatments and s sub-plot treatments is given below.

Analysis of variance for split plot with factor A with m levels in main plots and factor B with s levels in sub-plots will be as follows:

Sources of Variation	d.f.	SS	MS	F
Replication	$r-1$	RSS	RMS	RMS/EMS (a)
A	$m-1$	ASS	AMS	AMS/EMS (a)
Error (a)	$(r-1)(m-1)$	ESS (a)	EMS (a)	
B	$s-1$	BSS	BMS	BMS/EMS (b)
AB	$(m-1)(s-1)$	ABSS	ABMS	ABMS/EMS (b)
Error (b)	$M(r-1)(s-1)$	ESS (b)	EMS (b)	
Total	$rms-1$	TSS		

Analysis

Arrange the results as follows

Treatment Combination	Replication				Total
	R1	R2	R3	...	
A0B0	a0b0	a0b0	a0b0	...	T00
A0B1	a0b1	a0b1	a0b1	...	T01
A0B2	a0b2	a0b2	a0b2	...	T02
Sub Total	A01	A02	A03	...	T0
A1B0	a1b0	a1b0	a1b0	...	T10
A1B1	a1b1	a1b1	a1b1	...	T11
A1B2	a1b2	a1b2	a1b2	...	T12
Sub Total	A11	A12	A13	...	T1
.
.
.
Total	R1	R2	R3	...	G.T

$$\text{Compute CF} = \frac{(G T)^2}{r \times m \times s}$$

$$\text{TSS} = [(a_0b_0)^2 + (a_0b_1)^2 + (a_0b_2)^2 + \dots] - \text{CF}$$

Form A x R Table and calculate RSS, ASS and Error (a) SS

Treatment	Replication				Total
	R1	R2	R3	...	
A0	A01	A02	A03	...	T0
A1	A11	A12	A13	...	T1
A2	A21	A22	A23	...	T2
.
.
.
Total	R1	R2	R3	...	GT

$$\text{RSS} = \left(\frac{R1^2 + R2^2 + R3^2 + \dots}{m \cdot s} \right) - \text{CF}$$

$$\text{ASS} = \left(\frac{T0^2 + T1^2 + T2^2 + \dots}{r \cdot s} \right) - \text{CF}$$

$$\text{A x R table SS} = \left(\frac{A01^2 + A02^2 + A03^2 + \dots}{b} \right) - \text{CF}$$

Error (a) SS = A x R TSS - RSS - ASS.

Form A x B Table and calculate BSS, Ax B SSS and Error (b) SS

Treatment	Replication				Total
	B0	B1	B2	...	
A0	T00	T01	T02	...	T0
A1	T10	T11	T12	...	T1
A2	T20	T21	T22	...	T2
.
.
.
Total	C0	C1	C2	...	GT

$$\text{BSS} = \left(\frac{C0^2 + C1^2 + C2^2 + \dots}{r \cdot m} \right) - \text{CF}$$

$$A \times B \text{ table SS} = \left(\frac{T_0^2 + T_1^2 + T_3^2 + \dots}{r} \right) - CF$$

ABSS= A x B Table SS – ASS- ABSS

Error (b) SS= Table SS-ASS-BSS-ABSS –Error (a) SS.

Then complete the ANOVA table.

8.9 COMPLETE AND INCOMPLETE BLOCK DESIGNS:

In most of the experiments, the available experimental units are grouped into blocks having more or less identical characteristics to remove the blocking effect from the experimental error. Such design is termed as **block designs**.

The number of experimental units in a block is called the **block size**.

If size of block = number of treatments and each treatment in each block is randomly allocated, then it is a **full replication** and the design is called a **complete block design**.

In case, the number of treatments is so large that a full replication in each block makes it too heterogeneous with respect to the characteristic under study, then smaller but homogeneous blocks can be used. In such a case, the blocks do not contain a full replicate of the treatments. Experimental designs with blocks containing an incomplete replication of the treatments are called **incomplete block designs**.

8.10 AUGMENTED DESIGNS

In agricultural experiments often the existing practices on check varieties called control treatments are compared with new varieties or germplasms collected through exotic or domestic collections called test treatments. In some cases experimental material for test treatments is limited and it is not possible to replicate them in the design. However, adequate material is available for replicating control treatment in the design. Augmented designs are useful for these experimental situations.

An augmented design is any standard design in control treatments augmented with additional (new or test) treatments in complete or incomplete blocks in one-way heterogeneity setting. In general, these experiments are conducted using an augmented randomized complete block design, where the test treatments are replicated once in the design and control treatments appear exactly once in each block

8.10.1 Layout

In experimental material huge quantity of data to the layout of randomization augmented design. Some symbols used to make easy to understand. u (number of control treatment), w (number of

test treatment) and b (number of blocks in experiment).

For data are put to online to make a design layout by SAS or layout can be made manually. First the user enters the design parameter in, and replication of control treatment(s) to maximize the efficiency per observation is automatically computed.

In the hypothetical example, in that location are four check varieties (control treatment) denoted C1, C2, C3, C4 and 20 germplasm lines (test treatment) are denoted T1, T2, T3,....., T20 computed complete randomized block design.

Block 1: (T6, T2, C4, C1, T12, C3, T8, C2)

Block 2: (C3, C1, T4, T14, C2, T17, C4, T16)

Block 3: (C2, C3, T10, C1, C4, T13, T20, T15)

Block 4: (C2, C3, T9, C4, T7, C1, T5, T19)

Block 5: (T11, C1, C3, C4, T18, T3, T1, C2)

8.10.2 Analysis of data

Analysis is done by online software SPAD (statistical package for augment designs) software or web service. A partitioning of the degrees of freedom in an analysis of variance (ANOVA) table for this design is:

Source of variation - Degrees of freedom

Block - 4

Genotype - 23

Check - 3

New - 19

Check versus new - 1

B x check - 12

Correction for mean - 1

Total - 39

The user enters data in specified format for analysis through augment block design. The online software is available for analysis. For this treatment are renumbering as 1,2,....., u, u+1,....., u+w. Data should contain three columns first for block number, second for treatment and observed value in the third block. The data value separated by a TAB. If data already prepared in Excel sheet in columns than only paste the data online to get analysed.

<http://www.iasri.res.in/SpadWeb/>

The analysis data gives ANOVA for Adjusted means of treatments and ANOVA for Adujusted

means of block. It gives values of R-square, coefficient of variation, Root mean sum of square of error and General mean. It also gives all possible paired treatment comparisons and significance levels ($\text{Prob} > F$) and Critical difference between two controls, two test treatment (in a same block and different block separately) and in between test treatment and control treatment at 5% and 1% separately if it's possible to calculate. SPAD is also offered two more type contrast analysis. The GBD test for test vs controls and user defined contrast analysis.

8.10.3 Interpreted the values of analysis

In the analysis of variance treatment means with at least one word same are at par or not significantly different. In case treatment effect is at par than no pairwise analysis generated. There are not useful to compute multiple comparison test when treatment effects are at par. From the table, highest value (>0.05) in paired treatment comparison shows significantly different from paired value and lower value (<0.05) in paired treatment comparison shows at par from paired value.

The treatment divided into Tests, Controls and Tests vs Controls, the tests and Tests vs Controls are significantly different If $\text{Prob} > 0.05$ instead of at par if $\text{Prob} < 0.05$. This exhibit that test treatments are performing differently and there is a significant difference between test and control treatments.

8.11 GRID AND HONEYCOMB DESIGNS

Both Grid Method and Honeycomb Designs fall under the field plot techniques.

The Grid Method also known as stratified mass selection was first introduced by C.O. Gardner who used it to determine the yield of maize. In this method the whole experimental area is divided into smaller size subplots known as Grids. These subplots are of equal shapes and sizes. We then select a fixed number of plants from each subplot. This method increases the selection effectiveness as environmental effects are low when selections are made from smaller subplots.

Fasoulas developed the "Honeycomb Breeding Method" for crop breeding, in which nil-competition is the first and inviolable principle. He also constructed the "Honeycomb Selection Designs", where each plant lies in the center of a circle surrounded by six equidistant plants.

In principle, the procedure has three key characteristics: (1) the absence of inter-genotypic competition to allow recognition of genotypes of the productive ideotype within a progeny line, (2) an equal share of plenty of inputs for each progeny line, and (3) comparable conditions to evaluate and select progeny lines and individual plants objectively.

Owing to their systematic entry arrangement, locating each plant in the centre of a circular replicate/ring to ensure increased local control and allocating the plants of each entry in a moving triangular grid spread across the whole field for an effective sampling of soil heterogeneity, honeycomb designs objectively evaluate sister-lines and apply single-plant selection under a pattern of ultra-low planting density.

8.12 SUMMARY:

Design of Experiment:

Design of experiment means how to design an experiment in the sense that how the observations or measurements should be obtained to answer a query in a valid, efficient and economical way.

Treatment:

Different objects or procedures which are to be compared in an experiment are called treatments.

Factor:

A factor is a variable defining a categorization. A factor can be fixed or random in nature. A factor is termed as a fixed factor if all the levels of interest are included in the experiment.

A factor is termed as a random factor if all the levels of interest are not included in the experiment and those that are can be considered to be randomly chosen from all the levels of interest.

Basic Principles of experimental design:

There are three basic principles of design which were developed by Sir Ronald A. Fisher.

- (i) Randomization
- (ii) Replication
- (iii) Local control

Randomized Block Design:

A two-way layout is called a randomized block design (RBD) or a randomized complete block design (RCB) if, within each block, the v treatments are randomly assigned to v experimental units such that each of the $v!$ ways of assigning the treatments to the units has the same probability of being adopted in the experiment and the assignment in different blocks are statistically independent

Completely RBD:

The most common block design is the Completely Randomized Block Design. In this design all the treatments are applied randomly to plots within each block. Here the number of units within each block is equal to the the number of treatments.

Augmented Designs:

An augmented design is any standard design in control treatments augmented with additional (new or test) treatments in complete or incomplete blocks in one-way heterogeneity setting. In general, these experiments are conducted using an augmented randomized complete block design, where the test treatments are replicated once in the design and control treatments appear exactly once in each block

Complete and incomplete block designs:

If size of block = number of treatments and each treatment in each block is randomly allocated, then it is a **full replication** and the design is called a **complete block design**

Experimental designs with blocks containing an incomplete replication of the treatments are called **incomplete block designs**.

8.13 SELF ASSESSMENT QUESTIONS:

8.13.1 Fill in the blanks:

1. The smallest possible division of the experimental material is known as _____.
2. Different objects or procedures which are to be compared in an experiment are called _____.
3. A _____ is a variable defining a categorization.
4. The unexplained random part of the variation in any experiment is termed as _____.
5. A treatment design is the manner in which the levels of _____ are arranged in an experiment.
6. The experimental units within a block are relatively _____.
7. Each treatment occurs _____ in a block in a completely RBD.
8. To accommodate factors which require different sizes of experimental plots in the same experiment, _____ design has been evolved.
9. If the units are subjected to two way variation then _____ is preferred.
10. Allocation of treatments at random to units within each block is known as _____.

8.13.2 Multiple Choice questions:

1. The repetition of the treatments in an experiment is known as
 - a) Randomization
 - b) Local Control
 - c) Replication
 - d) None of the above
2. A factor is termed as a _____ factor if all the levels of interest are included in the experiment.
 - a) Fixed
 - b) Random
 - c) Mixed
 - d) None of the above
3. A factor is termed as a _____ factor if all the levels of interest are not included in the experiment and those that are can be considered to be randomly chosen from all the levels of interest.
 - a) Fixed
 - b) Random
 - c) Mixed
 - d) None of the above

4. In _____ research design, either a group or various dependent groups are observed for the effect of the application of an independent variable which is presumed to cause change.
- pre-experimental
 - quasi experimental
 - true experimental
 - None of the above
5. _____ designs are used in settings where randomization is difficult or impossible.
- pre-experimental
 - quasi experimental
 - true experimental
 - None of the above
6. The _____ research design relies on statistical analysis to approve or disprove a hypothesis.
- pre-experimental
 - quasi experimental
 - true experimental
 - None of the above
7. Randomized Block uses _____ principles of design of experiment
- One
 - Two
 - Three
 - Four
8. The analysis of RBD is similar to that of _____ way ANOVA
- One
 - Two
 - Three
 - None of the above
9. The larger plots in a split plot design are known as
- Main
 - Sub
 - Intermediate
 - None of the above
10. The smaller plots in a split plot design are known as
- Main
 - Sub
 - Intermediate
 - None of the above

8.13.3 Answers:**Fill in the blanks:**

(1) experimental unit (2) treatments (3) factors (4) experimental error (5) treatments
(6) homogeneous (7) once (8) Split Plot (9) Randomized Block
Design (10) Randomization

MCQ: (1) c (2) a (3) b (4) a (5) b (6) c (7) c (8) b (9) a (10) b

8.14 REFERENCES:

- <https://en.wikipedia.org/wiki/Statistics>
- <http://www.fao.org>
- <https://online.stat.psu.edu/stat200>
- <https://nptel.ac.in>
- <https://swayam.gov.in>
- <http://www.iasri.res.in>

8.15 SUGGESTED READINGS:

9. Fundamentals of Statistics Vol-I: Goon Gupta Dasgupta
10. Fundamentals of Mathematical Statistics: SC Gupta & VK Kapoor
11. Mathematical Statistics: Kapoor & Saxena
12. Mathematical Statistics: OP Gupta & BD Gupta
13. Fundamentals of Statistics Vol-II: Goon Gupta Dasgupta
14. Fundamentals of Applied Statistics: SC Gupta & VK Kapoor
15. Programmed Statistics: BL Agarwal
16. Basic Statistics: B.L Agarwal

8.16 TERMINAL QUESTIONS:

1. Define Design of experiment. Discuss the various types of Designs used.
2. Give the principles of Design of Experiment
3. What do you understand by Replication? Explain it's role in Design of Experiment.
4. What do you understand by Randomization? Explain it's role in Design of Experiment.
5. What do you understand by Local Control? Explain it's role in Design of Experiment.
6. Write an essay on Randomized Block Design explaining its main features layout and uses.
7. Write an essay on Completely Randomized Block Design explaining its main features layout and uses.
8. Write an essay on Split Plot Design explaining its main features layout and uses.
9. Write a short note on Augmented Designs.
10. Briefly discussed Grid and Honeycomb designs.



UTTARAKHAND OPEN UNIVERSITY

**Teenpani Bypass Road, Behind
Transport Nagar,
Haldwani- 263139, Nainital
(Uttarakhand)**

**Phone: 05946-261122, 261123; Fax No.
05946-264232**

**Website: www.uou.ac.in; e-mail:
info@uou.ac.in**