CENTRE FOR DISTANCE & ONLINE EDUCATION

UNIVERSITY OF JAMMU
JAMMU

SELF LEARNING MATERIAL

M.A. EDUCATION
SEMESTER-III

Subject : Medhods of Data Analysis **Unit : I – IV**

Course No. : 303 **Lesson No. : 1 – 8**

Dr. Anuradha Goswami

Course Co-ordinator

http:/www.distanceeducationju.in

Printed and Published on behalf of the Centre for Distance & Online Education, University of Jammu, Jammu by the Director, CD & OE, University of Jammu, Jammu

Course Contributors :

- **Dr. Ankita Sharma**
  Department of Statistics,
  University of Jammu,
  Jammu, J&K 180006

- **Dr. Bhagwati Devi**
  Department of Statistics,
  Central University of Jharkhand,
  Ranchi, Jharkhand, 835205

- **Dr. Apurba Vishal Dabgotra**
  Department of Statistics,
  University of Jammu, Jammu
  J&K 180006

- **Dr. Kiranpeep Kour**
  Division of Statistics and Computer,
  Science, Faculty of Science
  SKUAST-J J&K, 180009

Format Editing :

Dr. Anuradha Goswami
CD&OE, University of Jammu
Jammu, J&K, 180006

Editing and Proof Reading :

Dr. Ankita Sharma
Department of Statistics,
University of Jammu, Jammu, J&K
180006

Printed at : Quick Offset, Delhi

MASTER'S DEGREE PROGRAMME IN EDUCATION (M.A. EDUCATION)
COICE BASED CREDIT SYSTEM

SEMESTER III

Syllabus for the Examination to be held in December 2024, 2025 and 2026

Course No. PSEDTC303 **Title: Methods of Data Analysis**
Analysis Credits : 4

Maximum Marks:	**100**
Minor Test-I :	**10**
Minor Test-II :	**10**
Internal Assessment Assignment	**10**
Major Test :	**70**

Learning Outcomes:

1. Students will understand the concepts and methods used in Statistical analysis of test scores.
2. Students will gain idea about the concept of qualitative and qualitative data.
3. Students will understand the concept of and also analyze the used of Inferential and Descriptive Statistics.
4. Students will develop the skill for computations through statistical techniques.

Course Contents:

Unit-I

(a) Nature of Educational Data: Quantative and Qualitative; Descriptive and Inferential Statistics, Scales of Measurement

(b) Raw scores and Frequency Distribution Graphical Representation of Frequency Distribution – Polygon and Histogram– Differences between the two methods, Cumulative Frequency Curve, – Pie-Chart, ogive and smoothed frequency curve

Unit II

Measures of Central Tendency – Concepts and Calculation of Mean, Median and Mode, Properties of mean, when to employ mean, median and mode, Merits and Demerits

Measures of Variability: Concept and Calculations of Quartile Deviation, Standard Deviation, Interpretation of standard deviation, Percentile and Percentile Rank through ogive also, Merit and Demerits

Unit III

Normal Distribution Curve: Characteristics of Normal Probability Curve. Meaning and uses of standard scores. Concept of skewness and Kurtosis

i) Determination of the percent of cases/number of scores falling within the given limits.
ii) Determination of the limits, which include given percent of cases.
iii) Determination of raw scores from the given percent of cases.

Measures of Relationship: Concept, uses and computations of correlations by Pearson Product Moment Method and first order partial correlation.

Unit IV

a) Concept of levels of significance.
b) Types of Errors (Type I and Type II error), One-tailed and Two-tailed tests.
b) Significance of Statistics: Concept of Standard Error, Estimating Confidence Limits of Mean (Small and Large Sample).
c) Analysis of Variance: Meaning. Assumptions and uses with computations up to one-way classification only.

Modes of Transaction: Lecture-cum-discussion method, Problem solving method.

Note for paper setting:

There shall be two tests & one Assignment as part of Minor Evaluation & one major test at the end of semester in each semester. The students shall be continuously evaluated during the conduct of each course the basis of their performance as follows:

Thory	Syllabus to be covered in the examination	Time allotted for the examination	% weightage (marks)
Minor Test-I	Unit I & Unit II	Sixty Minutes	10 Marks
Minor Test-II	Unit III & Unit IV	Sixty Minutes	10 Marks 10 Mark (two questions of 5 marks each)
Major Test	Unit I to IV	Three Hours	70 Marks

Essential Readings

1. Aggrawal, T.P. (2009) Statistical Methods, Sterling Publishers Private Limited, New Delhi
2. Carter, David Clark. (2004) Quantitative Psychological Research, Psychology Press, East Sussex, New York
3. Cohen, Louis, et. al (2011) Research methods in education, Routledge, New York

4. Garrett, Henry. E. (1981) Statistics in psychology and education, Vakils Feffer and Simons Ltd. Bombay
5. Koul, Lokesh. (2011) Methodology of Educational Research, Vikas Publishing House Pvt. Ltd. New Delhi

Suggested Reading

1. Mangal, S.K. (2007) Statistics in psychology and education, Prentice Hall of India Pvt. Ltd., New Delhi

Note for Paper Setters (Major Test):

The question paper will contain long and short answer type questions. There will be total of eight long answer type questions (two questions from each unit with internal choice) and the candidates will be required to answer one question from each unit. Each long answer type question will carry 15 marks. Question No. 1 will be compulsory and shall have 04 short answer type questions (100 words per question). Short answer type questions will be from all the units. Each short answer type question will carry 2.5 marks.

M.A. EDUCATION
TITLE : MEDHODS OF DATA ANALYSIS

TABLE OF CONTENTS

LESSON : 1

THE POWER OF STATISTICS: ENHANCING EDUCATION THROUGH STATISTICAL TECHNIQUES

Structure

1.1 INTRODUCTION

To a layperson, “statistics” means numerical information presented in quantitative terms. This data can relate to various topics, activities, phenomena, or geographic areas, and it has no limits in its reference or scope. On a macro level, statistics may include figures related to gross national product and the roles of agriculture, manufacturing, and services in GDP (Gross Domestic Product). On a micro level, companies of all sizes generate extensive statistics about their operations. Their annual reports typically include data on sales, production, expenses, inventory, and capital investments. This data is often gathered using scientific survey techniques, and unless it is regularly updated, it usually represents a one-time effort, which limits its usefulness.

For students, statistics is a field of study similar to economics, mathematics, chemistry, and physics. It is a discipline that scientifically investigates data and is commonly referred to as the science of data. To manage data effectively, statistics has developed specific methods for collecting, presenting, summarizing, and analyzing it. Therefore, statistics can be defined in two main ways: as a collection of numerical information and as a scientific discipline dedicated to data analysis.

Plural Sense: In the plural sense, statistics refers to numerical data related to a collective of individuals. Examples include the runs scored by a batsman in various matches, the income and expenses of people living in a specific area, and the demand and supply of a commodity over designated time periods.

Singular Sense: In the singular sense, statistics is described as the science of collecting, organizing, presenting, analyzing, and interpreting numerical data.

Origin of Statistics

In the 9th century, the Islamic mathematician Al-Kindi was the first to apply statistics to decode encrypted messages, creating the earliest code-breaking algorithm based on frequency analysis at the House of Wisdom in Baghdad. He wrote a book called Manuscript on Deciphering Cryptographic Messages, which contained in-depth discussions on statistics, including methods for cryptanalysis, different forms of encipherment, the analysis of specific encipherments, and the statistical examination of letters and letter combinations in Arabic.

In the early 11th century, Al-Biruni highlighted the significance of repeated experimentation in his scientific method. He addressed the need to recognize and minimize both systematic errors and observational biases, such as those resulting from small instruments or human error. Al- Biruni argued that if instruments generate errors due to their imperfections, multiple observations should be taken and qualitatively analyzed to arrive at a “common-sense single value for the constant sought,” whether this be an arithmetic mean or a “reliable estimate.”

1.2 LEARNING OBJECTIVES:

After going through this lesson, student will be able to:

- Understand the concept of qualitative and quantitative data
- Importance and need of Statistics in Education
- Describe the types of variables and their measurement
- Understand the statistical principles behind test design, reliability, and validity.
- Learn to analyze test scores and performance data to assess student achievement and areas for improvement.

1.3 MODERN HISTORY

The term "statistics" originates from the Latin word "status" and the Italian word "statista," both of which mean "political state" or government. Shakespeare used the word "statist" in his play *Hamlet* (1602). Historically, statistics were mainly used by rulers who needed information about land, agriculture, commerce, and population to assess their military strength, wealth, taxation, and other governmental functions.

Gottfried Achenwall introduced the term "statistik" at a German university in 1749, referring to the political science of various countries. In 1771, W. Hooper incorporated the term "statistics" in his translation of Baron B.F. Bieford's *Elements of Universal Erudition*, defining it as the science that describes the political arrangements of modern states around the world. While there is a significant gap between historical and modern statistics, the principles of early statistics continue to inform contemporary practices.

During the 18th century, English writers began to use the word "statistics" in their works, leading to its gradual development over the following centuries. Substantial progress was made by the end of the 19th century. At the beginning of the 20th century, William S. Gosset developed methods for making decisions based on small data sets. Throughout the 20th century, many statisticians were active in creating new methods, theories, and applications for statistics. Today, the widespread availability of electronic computers plays a crucial role in the modern evolution of statistics.

1.4 DEFINITIONS OF STATISTICS:

- *"Statistics comprises the collection, tabulation, presentation and analysis of an aggregate of the facts, collected in methodical manner, without bias and related to predetermined purpose."* – **Sutcliffe**
- **According to Prof. A.L. Bowley:** *"Statistics may be called the science of counting."*
- **According to Boddington:** "*Statistics is the science of estimates and probabilities."*

- "*Statistics are numerical statements of facts in any department of enquiry placed in relation to each other."*– **A.L. Bowley**
- "*Statistics may rightly be called the science of averages*." –**Bowleg**
- "*Statistics may be defined as the collection, presentation, analysis, and interpretation of numerical data.*" – **Croxton and Cowden**

1.5 NEED OF STATISTICS IN EDUCATION:

Education is a key priority on the national agenda, with several important goals and targets that must be achieved for it to effectively support national development. Attaining these goals

requires careful planning and the implementation of impactful programs and initiatives. Evidence-based planning and management in education are essential, not only to justify increased investments in the social sector but also to boost India's competitiveness in the global economy. As a result, establishing a reliable and comprehensive statistical foundation is crucial for effective planning and policymaking. Given that educational planning is recognized as an integral part of socio-economic planning, a strong statistical base in education is necessary. This foundation is especially significant as India increasingly recognizes the vital role of education in socio-economic development. Timely, relevant, and reliable information on education at all levels—national, state, district, sub-district, and school—is critical for effective educational planning, administration, monitoring, and evaluation**.**

Educational statistics have become increasingly essential due to the rapid structural and systemic changes in India's social and economic sectors. The successful implementation of government plans and initiatives relies significantly on a robust information base that encompasses both quantitative and qualitative data at international, national, and sub-national levels. Moreover, socio-economic planning requires aligning strategic national development goals outlined in various sector plans and establishing a long-term development trajectory for the country. Consequently, a comprehensive and objective-oriented database across all sectors of the Indian economy, including education, is crucial for placing the nation on a strategic development path.

A dedicated database for the education sector would greatly enhance educational planning and provide valuable insights for areas such as manpower, labor markets, demographics, and health. Educational statistics are vital for both short-term planning and long-term strategic initiatives. Thus, a strong information base regarding education is essential for effective educational planning and overall economic development in the country. Therefore, it is important to prioritize long-term considerations when reforming statistical information systems.

Four primary purposes highlight the importance of educational statistics: (a) to develop sound policies and

effective plans, (b) for efficient administration and management, (c) to support research, and (d) for the dissemination of information. A reliable and detailed statistical base is critical for facilitating proper policymaking, planning, management, and research.

1.6 IMPORTANCE AND USES OF STATISTICS IN EDUCATION:

1. **Group Comparison:** A class's achievements are not uniform across subjects. It has been discovered that one class is progressing quicker in one subject while another is progressing in another. Even different portions of the same class do not progress in the same manner.
2. **Individual Comparison:** Statistics aids in the individual comparison of students of varying ages, talents, and IQ levels. Statistics explain why students who are similar in every other way do not attain the same level of success in one topic.
3. **Educational and Vocational Guidance**: Every student differs from others in terms of intellectual capacity, interests, attitude, and mental abilities. Students are offered educational and vocational counselling in order to make the most use of their qualities, and the process of guidance is based only on statistics
4. **Educational Experiments and Research**: The goals, curricula, and methods of teaching change as the location, line, and conditions change. Without the use of statistics, research and experimentation cannot become accurate and valid.
5. **Essential for Professional Efficiency:** The responsibility of the teacher does not cease when he teaches a certain subject in the classroom. His responsibilities include instructing the students, acquiring the appropriate level of knowledge for himself, and analyzing the achievement of behaviour modification.
6. **Basis of Scientific Approach to Problems:** Statistics forms the basis of scientific approach to problems of Educational Psychology.

1.7 CHECK YOUR PROGRESS

1. Statistics helps in the_____of students' academic performance and allows for the evaluation of teaching methods.
2. In education, statistics is essential for_____decision-making by administrators and policymakers.
3. By using statistics, educational institutions can improve_____allocation based on student needs and performance.
4. Educational_____relies on statistical methods to gather and analyze data to understand trends, student behavior, and the effectiveness of teaching methods.

5. Statistical analysis helps identify_____in education, such as differences in performance based on socioeconomic status, gender, or ethnicity.
6. The primary goal of using statistics in education is to improve_____outcomes and enhance the quality of teaching.

1.8 QUALITATIVE AND QUANTITATIVE VARIABLES

In education, data can be categorized into two main types: qualitative and quantitative. Each type provides unique insights and serves different purposes in understanding and improving educational practices and outcomes.

Qualitative variables are those that express a qualitative attribute such as hair colour, eye colour, religion, favourite movie, gender, and so on. Qualitative data refers to non-numerical information that provides insights into concepts, thoughts, and experiences. The values of a qualitative variable do not imply a numerical ordering. Values of the variable "religion" differ qualitatively; no ordering of religions is implied. Qualitative variables are sometimes referred to as categorical variables.

Quantitative variables are those variables that are measured in terms of numbers. Some examples of quantitative variables are height, weight, and shoe size.

To study the effect of diet, the independent variable was type of supplement: none, strawberry, blueberry, and spinach. The variable "type of supplement" is a qualitative variable; there is nothing quantitative about it. In contrast, the dependent variable "memory test" is a quantitative variable since memory performance was measured on a quantitative scale (number correct).

In practice, integrating both qualitative and quantitative data provides a more complete perspective. For instance, while quantitative data might indicate that a specific teaching method leads to higher test scores, qualitative data can shed light on how students experience and perceive that method or strategy. This combination offers a deeper understanding of the strategy's effectiveness and identifies areas for improvement. By using both types of data, educators can make more informed decisions and better address the varied needs of their students.

Examples of Quantitative and Qualitative data

Quantitative data

- **Exam Scores**: A teacher recorded the scores of 30 students on a math test.
- **Height Measurement**: A gym measures the heights (in cm) of 15 participants before and after a training program.

- **Daily Steps**: A fitness app tracks the number of steps taken by users over a week

Qualitative data

- **Favorite Movies**: A survey asks 50 people about their favorite movie genres
- **Customer Feedback**: A restaurant collects customer reviews on their service quality.
- **Brand Preferences**: In a focus group, participants discuss their favorite brands of sports shoes.

1.9 DISCRETE AND CONTINUOUS VARIABLES

Variables such as number of children in a household are called discrete variables since the possible scores are discrete points on the scale. For example, a household could have three children or six children, but not 4.53 children. Other variables such as "time to respond to a question" are continuous variables since the scale is continuous and not made up of discrete steps. The response time could be 1.64 seconds, or it could be 1.64237123922121 seconds. Of course, the practicalities of measurement preclude most measured variables from being truly continuous.

1.10 DESCRIPTIVE AND INFERENTIAL STATISTICS

In simple terms, **Descriptive Statistics** is used to summarize and present data, often through charts or graphs. Descriptive statistics are very important because if we simply presented our raw data it would be hard to visualize what the data was showing, especially if there was a lot of it. Descriptive statistics therefore enables us to present the data in a more meaningful way, which allows simpler interpretation of the data. Descriptive statistics describes the important characteristics/ properties of the data using the measures the central tendency like mean/ median/mode and the measures of dispersion like range, standard deviation, variance etc.

subsequently, data can be summarized and represented in an accurate way using charts, tables and graphs.

On the other hand, in **Inferential Statistics**, data is used from the sample and conclusions or inferences are made about the larger population from which the sample is drawn. The goal of the inferential statistics is to draw conclusions from a sample and generalize them to the population. It determines the probability of the characteristics of the sample using probability theory

For example, if we survey 100 people about their preference for shopping at Cosmos Shopping Mall, we might create a bar chart showing the number of people who said "yes" or "no" (which is descriptive statistics). Using inferential statistics, we can then use this sample data to draw conclusions about whether most people in the general population like shopping at Cosmos.

1.11 MEASUREMENT SCALES

Before we can conduct a statistical analysis, we need to measure our dependent variable. Exactly how the measurement is carried out depends on the type of variable involved in the analysis. Different types are measured differently. To measure the time taken to respond to a stimulus, you might use a stop watch. Stop watches are of no use, of course, when it comes to measuring someone's attitude towards a political candidate. A rating scale is more appropriate in this case (with labels like "very favorable," "somewhat favorable," etc.). For a dependent variable such as "favorite color," you can simply note the color-word (like "red") that the subject offers.

Measurement procedures can vary greatly, but they can generally be classified into a few fundamental categories. Each category, known as "scale types" or simply "scales," encompasses procedures that share important characteristics, which are described in this section.

Nominal

Nominal scales involve categorizing or naming responses without implying any order. Examples of variables measured using a nominal scale include gender, handedness, favorite color, and religion. The crucial aspect of nominal scales is that they do not indicate any ranking among the responses. For example, when classifying people by their favorite color, there is no inherent hierarchy where green is considered "better" or "more important" than blue; the responses are merely categorized. Nominal scales represent the most basic level of measurement.

Ordinal

A researcher aiming to measure consumer satisfaction with microwave ovens might ask respondents to indicate their feelings using the options "very dissatisfied," "somewhat dissatisfied," "somewhat satisfied," or "very satisfied." These options are ordered from least to most satisfied, which distinguishes ordinal scales from nominal scales. Unlike nominal scales, ordinal scales allow for meaningful comparisons of satisfaction levels between individuals. For example, if one person chooses "somewhat satisfied" and another selects "somewhat dissatisfied," it is valid to assert that the first person is more satisfied, as their response falls later in the scale.

However, ordinal scales do not capture certain important information that may be present in other scales. Specifically, the differences between levels of an ordinal scale cannot be assumed to be consistent. For instance, in our satisfaction scale, the difference between "very dissatisfied" and "somewhat dissatisfied" is likely not the same as the difference between "somewhat dissatisfied" and "somewhat satisfied." Our measurement method does not allow us to determine whether these differences reflect the same level of psychological satisfaction.

Statisticians often express this by noting that the differences between adjacent scale values do not necessarily indicate equal intervals on the underlying scale—in this case, the actual feeling of satisfaction that we aim to measure.

Now, imagine if the researcher had asked consumers to indicate their level of satisfaction by choosing

a number from one to four. Would the difference between a response of one and two necessarily reflect the same difference in satisfaction as that between two and three? The answer is no. Changing the response format to numbers does not change the scale's meaning; we still cannot assume that the mental transition from 1 to 2 is equivalent to the transition from 3 to 4.

Interval scales

Interval scales are numerical scales in which intervals have the same interpretation throughout. As an example, consider the Fahrenheit scale of temperature. The difference between 30 degrees and 40 degrees represents the same temperature difference as the difference between 80 degrees and 90 degrees. This is because each 10-degree interval has the same physical meaning (in terms of the kinetic energy of molecules).

Interval scales are not perfect, however. In particular, they do not have a true zero point even if one of the scaled values happens to carry the name "zero." The Fahrenheit scale illustrates the issue. Zero degrees Fahrenheit does not represent the complete absence of temperature (the absence of any molecular kinetic energy). In reality, the label "zero" is applied to its temperature for quite accidental reasons connected to the history of temperature measurement. Since an interval scale has no true zero point, it does not make sense to compute ratios of temperatures. For example, there is no sense in which the ratio of 40 to 20 degrees Fahrenheit is the same as the ratio of 100 to 50 degrees; no interesting physical property is preserved across the two ratios. After all, if the "zero" label were applied at the temperature that Fahrenheit happens to label as 10 degrees, the two ratios would instead be 30 to 10 and 90 to 40, no longer the same! For this reason, it does not make sense to say that 80 degrees is "twice as hot" as 40 degrees. Such a claim would depend on an arbitrary decision about where to "start" the temperature scale, namely, what temperature to call zero (whereas the claim is intended to make a more fundamental assertion about the underlying physical reality).

Ratio Scales

The ratio scale of measurement is the most informative type of scale. It is an interval scale that also features a zero point indicating the complete absence of the quantity being measured. You can think of a ratio scale as integrating the three previous scales into one. Like a nominal scale, it provides a name or category for each object (with numbers functioning as labels). Like an ordinal scale, the objects are arranged in a specific order based on their numerical values. Like an interval scale, the difference between two points on the scale has the same meaning everywhere. Additionally, the same ratio between two values is consistently interpreted.

For example, the Fahrenheit temperature scale has an arbitrary zero point, so it is not a ratio scale. In contrast, the Kelvin scale has an absolute zero, making it a ratio scale. This means that if one temperature

is twice as high as another on the Kelvin scale, it corresponds to having twice the kinetic energy.

Another example of a ratio scale is the amount of money you have at the moment, such as 25 cents or 55 cents. Money is measured on a ratio scale because, in addition to possessing the characteristics of an interval scale, it has a true zero point: having zero money signifies the absence of any money. Because of this true zero, it is meaningful to say that someone with 50 cents has twice as much money as someone with 25 cents.

1.12 Score

Let us explore the concept "score", a common term, frequently used in educational research, statistics and school settings. Score refers to the numerical description of the performance of any test, for example, the score secured by a student in social science term-end examination. There are two types of scores.

Raw Score:

The **raw score** is the unadjusted, original score obtained by an individual on a test or assessment. It is the sum of the number of correct answers, points, or responses given. For example: If a test has 50 questions and a student answers 40 correctly, the raw score would be 40.

Characteristics:

- It does not take into account difficulty, comparison with other test-takers, or any other adjustments.
- It is often used as the base for further statistical or comparative analysis.

Derived Score:

A **derived score** is a score that is calculated or adjusted from the raw score using statistical methods to compare or rank performance. It often includes transformations like percentiles, z- scores, T-scores, or standard scores. For example: Percentile Rank: If a student's raw score is in the 90th percentile, it means the student performed better than 90% of the test-takers.

Z-Score: Standardized score representing how far a raw score is from the mean in terms of standard deviations.

T-Score: A type of derived score with a mean of 50 and a standard deviation of 10, often used in psychological testing.

Characteristics:

- Derived scores provide context and allow for comparisons across different tests, populations, or over time.
- They account for factors like test difficulty or group performance.

In summary, raw scores are the basic results of a test, while derived scores provide a more nuanced interpretation, often making scores more comparable and understandable across different contexts.

1.13 CHECK YOUR PROGRESS

Question 1: Define quantitative data and qualitative data. Provide two examples of each type.

Question 2: A researcher collects data on the number of books read by students in a month and their favorite genres. Identify which data type each represents.

Question 3: Explain the difference between descriptive statistics and inferential statistics. Provide an example of each.

1.14. LET US SUM UP:

Educational data is essential for analyzing and improving educational systems. It includes both quantitative and qualitative data, each serving distinct functions. Quantitative data, such as test scores, grades, and enrollment figures, consists of numerical values that can be measured and analyzed statistically. In contrast, qualitative data is non-numerical and includes descriptions, such as open-ended survey answers or observations of student behavior, which provide a deeper understanding of experiences and social interactions. Descriptive statistics help summarize and interpret the main features of a dataset, offering insights through measures like averages, percentages, and standard deviations, which reveal patterns and trends in educational outcomes. On the other hand, inferential statistics go beyond summarization to draw conclusions or make predictions about a larger population based on a sample. Methods such as hypothesis testing and regression analysis are used to assess relationships, evaluate theories, and generalize results, which are crucial for assessing the effectiveness of educational interventions and policies. Scores, representing numerical evaluations of student performance, are a fundamental part of educational data, reflecting achievement and progress. By combining both quantitative and qualitative data with descriptive and inferential statistics, educators and policymakers can make well-informed decisions to improve teaching strategies and learning outcomes.

1.15. KEY POINTS/GLOSSARY:

1. Quantitative Data: Data that can be measured and expressed numerically.

i) Discrete: Countable values (e.g., number of students)

ii) Continuous: Measurable values (e.g., height, weight).

2. Qualitative Data: Data that describes characteristics or qualities that cannot be measured numerically. Like Nominal: Categories without a specific order (e.g., colors, names) and Ordinal: Categories with a defined order (e.g., rankings).

3. Descriptive Statistics: Summarizes and describes the features of a dataset.

4. Inferential Statistics: Uses a random sample of data to make inferences or generalizations about a larger population. Purpose: To draw conclusions beyond the immediate data alone**.**

1.16. SELF-ASSESSMENT QUESTIONS.

Question 1: Imagine you are a teacher who tracks the attendance and grades of your students over a semester. How could you use this data to identify students who may need additional support?

Question 2: As a teacher, you collect data on your students' attendance and grades to evaluate their progress. What type of data are you collecting?

A) Only qualitative data
B) Only quantitative data
C) Both qualitative and quantitative data
D) Neither qualitative nor quantitative data

Question 3. Which of the following is an example of quantitative data in education?

A) A student's written feedback on the course
B) A teacher's assessment of student behavior
C) The number of hours a student spends on homework
D) The descriptions of a student's classroom interactions

Question 4. In which scenario would qualitative data be more useful than quantitative data?

A) Measuring the average test score of a class
B) Evaluating student satisfaction with a new teaching method
C) Counting the number of students absent on a given day
D) Calculating the total number of assignments submitted

Question 5. You are calculating the mean score of students in a class. What does the mean represent?

A) The most common score in the class
B) The middle score in the class
C) The average score of all students
D) The spread of scores around the average

1.17. LESSON END EXERCISE

Question 1: List and explain the four scales of measurement in statistics. Provide an example for each scale**.**

Question 2: Classify the following variables as nominal, ordinal, interval, or ratio:

a. Temperature in Celsius

b. Ranking of movies (1st, 2nd, 3rd)

c. Types of fruits (e.g., apples, bananas, oranges)

d. Weight of students in kilograms

Question 3: A survey collects customer satisfaction ratings using a Likert scale ranging from 1 (very dissatisfied) to 5 (very satisfied).

a). What type of measurement scale does this represent?

b). Describe how you would analyze this data. Include examples of both descriptive and inferential statistics that could be used in your analysis.

Question 4: How can statistical analysis help educators assess and improve student learning outcomes?

Question 5: In what ways does the use of statistics contribute to evidence-based decision- making in education policy?

Question 6: Why is it important for teachers and administrators to understand and apply basic statistical concepts in curriculum development and assessment?

1.18. SUGGESTED READINGS

- Gupta, S. C., & Kapoor, V. K. (2020). Fundamentals of mathematical statistics. Sultan Chand & Sons.
- Tufte, E. R. (2001). The Visual Display of Quantitative Information (2nd ed.) (p.178). Cheshire, CT: Graphics Press.
- Goon, A.M., Gupta, M.K. and Dasgupta, B. (1968). Fundamental of Statistics, Vol I, World Press, Kolkata.
- Goon, A.M., Gupta, M.K. and Dasgupta, B. (1968). Fundamental of Statistics, Vol II, World Press, Kolkata.
- Gupta, S.C. and Kapoor, V.K. (2020). Fundamentals of Mathematical Statistics, 12th Ed., Sultan Chand and Sons.
- Moore, D.S. (2009). The Basic Practice of Statistics. 5th Ed., W H Freeman.

LESSON : 2

UNDERSTANDING GRAPHICAL REPRESENTATIONS: VISUALIZING DISCRETE AND CONTINUOUS DATA

Structure

2.1 INTRODUCTION

Graphical representation of data is a powerful tool for presenting complex information in a visually accessible way. By using charts, graphs, and diagrams, it transforms raw data into clear, understandable patterns and trends, allowing for easier analysis and interpretation. Whether through bar graphs, line charts, pie charts, or scatter plots, graphical representations help highlight relationships between variables, reveal patterns, and provide insights that may not be immediately obvious in raw numerical form. This visual approach is essential for effective communication, making it easier to convey key findings, support decision- making, and facilitate a deeper understanding of the data.

After classifying and presenting the data in a tabular form, one can present the data in a pictorial form. As the adage goes "*A picture is worth a thousand words*", just one diagram is enough to represent a given data more effectively than thousand words. Diagrams and graphs are one of the most convincing and appealing ways in which statistical results may be presented. It is easy to understand diagrams even for ordinary people.

2.2 LEARNING OBJECTIVES

After going through this lesson, student will be able to:

- Different methods of presentation of statistical data such as tables, graphs, and diagrams
- Learn how to use bar graphs to compare different categories or groups of data.
- Develop the skill to visualize the distribution of data, especially for frequency or interval data.
- Understand how scatter plots are used to show relationships between two variables.
- Learn to interpret quartiles, medians, and outliers in a dataset

2.3. TYPES OF GRAPHICAL REPRESENTATIONS

After having an understanding about the concept of graphical representation Let us now discuss various types of graphical representation. It is based on the type of data available for the presentation. Graphical representation of Data representations for continuous and discrete variable.

1) Discrete variable in which nominal data are obtained. This type of ungrouped data is presented by the following graphical presentations:

   i. Bar Graph or Bar Diagram

   ii. Circle or Pie Graphs/Diagram

   iii. Line Graph

2) Continuous variable is one in which frequency distribution tables are prepared or data can be organized into class intervals. This type of grouped data is used in the following graphical presentation:

   i. Histogram or Column diagram

   ii. Frequency Polygon

   iii. Cumulative Frequency Graph

   iv. Cumulative Frequency Percentage Curve or Ogive

2.4 DIAGRAM

A diagram is a visual form for presentation of statistical data, highlighting their basic facts and relationship. In other words, we can say, diagrams are pictorial presentation of quantitative data. The data are represented by using geometric figures.

Advantages

i. They are attractive and impressive.

ii. They make data simple and understandable.

iii. They helps in making comparisons more intuitive and clear.

iv. They save time and labour, in terms of grasping the information.

v. They have universal utility.

vi. They give more information.

vii. They have a great memorizing effect.

2.5 GENERAL RULES FOR CONSTRUCTING DIAGRAMS

- **Choice of a Diagram:** A wide variety of diagrams exist to represent statistical data. Selecting the right diagram requires careful consideration of the data's nature, the study's objectives, and the intended audience. This task is not straightforward; it necessitates significant skill, knowledge, and insight. Choosing the wrong diagram can lead to misleading, incorrect, or deceptive conclusions about the phenomenon. Thus, it is vital to be meticulous in the diagram selection process.
- **Selection of Scale:** The scale of the diagram plays an important role in reading and understanding the figures. Same data represented using differing scales may lead to different conclusions. Therefore, it is important to choose scale cautiously. Unfortunately, there is no hard and fast rule for selection of scale. One should take into account the size of the paper and the observations as guiding factors. The dimensions chosen should effectively showcase the key features of the data. The scales used on both the horizontal and vertical axes should be clearly labelled, with values ideally in even numbers

or multiples of 5 or 10. For comparative analysis of two or more diagrams, maintaining the same scale across all diagrams facilitates valid conclusions.

- **"Proportion between Width and Height":** For a diagram to look attractive and to convey the information clearly, it is important that the width and height should have a good proportion. It should not be too narrow or too wide. There is not strict rule for this also. However, Lutz in his book Graphic Presentation has suggested the 'root two' rule, viz. the ratio 1:2 or 1:1.414 between height and width, respectively.
- **Title, Number and Footnotes:** Similar to a well-designed statistical table, each diagram should feature a suitable title that clearly indicates what it represents. The title should be brief, self-explanatory, clear, and unambiguous, and it should be positioned at the top or bottom of the diagram.

  The diagram should also contain a number along with the title (for e.g. Figure 1: Title of the diagram). It makes it easy to refer to the diagram and provides better readability and comparability in a report. The numbers could be 1, 2, 3,… or chapter/section wise, like third figure of chapter 2 may be numbered as Figure 2.3.

  In case there is any information that is required to be displayed and cannot be covered in the title or anywhere else, it should be provided as a footnote. It is written on the left-hand bottom of the diagram.
- **Source Note:** Including the source of the information at the bottom of the diagram is a good practice. This is essential because the reliability of the information can vary greatly depending on the source, particularly for an informed audience.
- **Index or Legend:** When a diagram contains information on two or more series of observations, a brief indexing of different colors, shading, etc. should be provided. It represents what do the different colors, shading or patterns represent in the diagram. The index, also called legend, can be placed at any of the positions (top, bottom, left, right) depending on its visibility.
- **Neatness:** It is always essential to have a neat, clean and attractive diagram. It conveys the information to the viewer clearly and draws good attention of the reader.
- **Simplicity:** A complex diagram would be the last thing one will want to see, especially when the audience is a layman. Too much information in one diagram can confuse the audience and can make it difficult to grasp. This will defy the purpose of the diagrams. Therefore, it is always recommended to keep the diagram as simple as possible.

2.6 GRAPHICAL REPRESENTATION OF DISCREATE DATA

There exists a large variety of diagrams to present statistical data. However, we shall discuss some of them here. A broad classification of the types of diagrams as given below:

In one- and two-dimensional diagram, the statistical data is represented by means of the shapes in which only one- and two-dimension changes, for example, lines and bars, etc. where height is used to read the data value and the width is not considered. We shall discuss the following diagrams in this category:

1. Line Diagram
2. Bar Diagram
3. Simple Bar
4. Multiple Bar
5. Sub-divided Bar
6. Pie Diagram

2.6.1 LINE DIAGRAM

Line diagram is used in case where there are many items to be shown on one characteristic or series. Such diagram is prepared by drawing a vertical line for each item according to the

scale. The distance between lines is kept uniform. Line diagram makes comparison easy, however, it is less attractive.

Table 1: The below graph shows the annual profit percentage earned by the company during the years 1998 to 2003. The diagrammatic representation of this data using line diagram is given in Figure 1.

Table 1: The annual profit percentage earned by the company during the years 1998 to 2003

Year	1998	1999	2000	2001	2002	2003
Profit (%).	43	49	49	59	62	78

Figure1: Represent the annual profit percentage earned by the company during the years 1998 to 2003

2.6.2 BAR DIAGRAM

Simple Bar Diagrams are used to represent only one variable over time, space etc. It can be drawn either on horizontal or vertical base. The horizontal base (vertical bars) is preferred when the data is collected over time or the categories have small names which could be clearly displayed on horizontal axis. However, when the names of the categories are not small or it is difficult to accommodate them on horizontal axis, one can use vertical base (also known as horizontal bars)

The bars must be of uniform width and intervening space between bars must be equal. The scale is determined on the basis of the highest value in the series. One can use colours or patterns to make the diagram more attractive.

Table 2 provides the number of elected female candidates in parliamentary elections in India from 1998 to 2019. The diagrammatic representation of this data using line diagram is given in Figure 1.

Table 2: Number of elected Female Candidates in Parliamentary Elections in India

Year	1998	1999	2004	2009	2019
Female M.P.	43	49	49	59	78

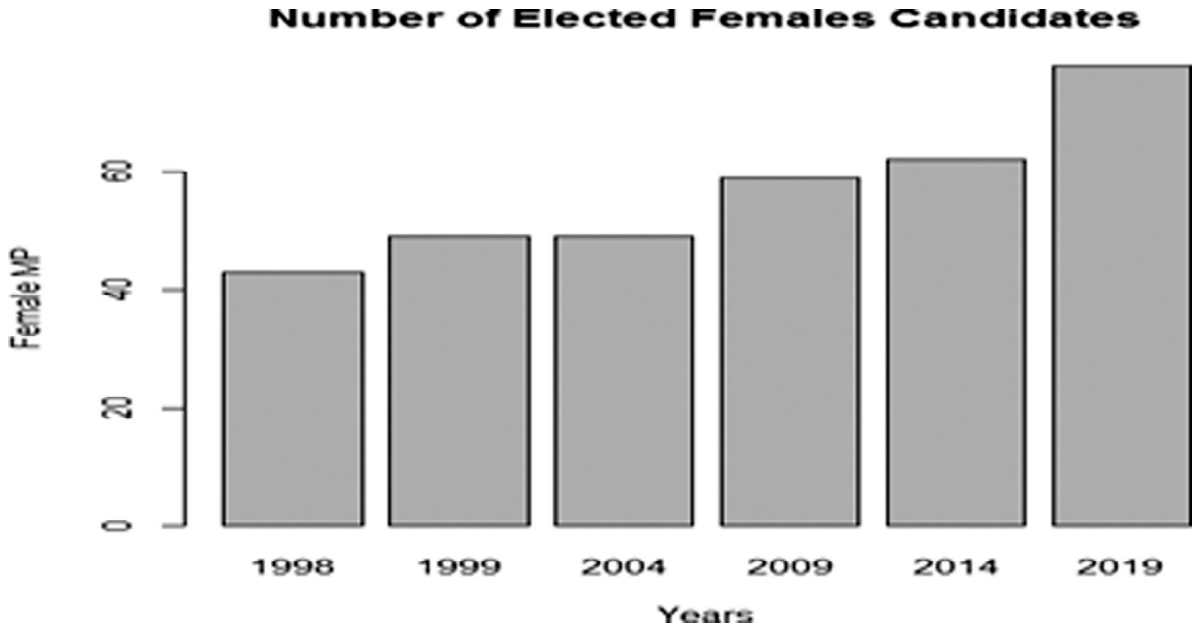


Figure 2: Number of elected Female Candidates in Parliamentary Elections

Multiple Bar Diagram

Multiple bar diagram is used for comparing two or more sets of statistical data. Bars are constructed side by side to represent the set of values for comparison. In order to distinguish bars, they may be either differently

coloured or there should be different types of crossings or dotting, etc. An index is also prepared to identify the meaning of different colours or dotting.

Example 1: A researcher conducted a study on students' perception towards various modes of teaching -learning. He/she administered a questionnaire to 400 higher secondary school students. Following data received on students' perception on various mode of learning.

Modes of Teaching Learning	Face to Face	Distance	Online	Blended
Rural Area Students	120	37	15	34
Urban Area Students	72	36	36	50

Figure 3: Learning modes used by Students in Rural and Urban Areas

A multiple bar diagram can be used to compare modes of learning used by students of Rural area and Urban areas. On the basis of this graph, we see that face to face mode of learning is more preferred by rural area students as compared to urban area students. Whereas online mode is more preferred by urban area students as compared to rural area students.

Subdivided Bar Diagram

In these diagrams each bar represents not only the whole of magnitude but also the various components of which it is composed of. Component Part Chart makes it possible to compare components and also to compare the components with the total. In a sub-divided bar diagram, the bar is sub-divided into various parts in proportion to the values of the components and the whole bar represent the total. Such diagrams are also called Component Bar diagrams. The sub-divisions are distinguished by different colours or crossings or dotting.

Table 3: Number of students in a university

Year	Art	Science	Law
2004-05	18000	9000	4000
2005-06	20000	10000	5000
2006-07	26000	9000	7000
2007-08	31000	9500	7500

Table 3 contains the faculty-wise number of students enrolled in a university in different years. It is clear that the total of Art, Science and Law students will give us the total number of students in that university (assuming it has these three faculties only). It could be of interest to know the composition of how many students enrolled for different faculties along with total number of students enrolled in each year. This

can be represented using sub-divided diagrams. For construction of subdivided diagrams, the values for each category (faculty) are to be stacked. In order to construct the diagram, we should find the cumulative values for the categories as given in Table 4.

Table 4: Number of students in a university (Calculation for Sub-divided Diagram)

Year		Art	Science	Law
2004-05	Value	18000	9000	4000
	Cumulative	18000	27000	31000
2005-06	Value	20000	10000	5000
	Cumulative	20000	30000	35000
2006-07	Value	26000	9000	7000
	Cumulative	26000	35000	42000
2007-08	Value	31000	9500	7500
	Cumulative	31000	40500	48000

Using the cumulative values in Table 4, the subdivided bar diagram can be constructed as given in Figure 4. From Figure 4, we can observe that the total number of enrolments has increased from 2004-05 to 2007-08. Each year, most of the students take admission in Art faculty, followed by science. There are least number of students in Law as compared to the other two faculties.

2.6.3 PIE DIAGRAM OR ANGULAR DIAGRAM

It is one of the most commonly used diagrams. It is used to represent the total magnitude and its various components. It is an alternative to subdivided bar diagram where instead of bars, circle is used. It is known as an angular or pie diagram

A pie diagram is a circular graph which represents the total value with its components. The area of circle represents the total value and the different sectors of the circle represent the different parts. The circle is divided into sectors by radii and the areas of the sectors are proportional to the angle at the centre. It is generally used for comparing the relation between various components of a values and between components and the total value. In pie diagram, the data are expressed as percentages. Each component is expressed as percentage of the total value. A pie diagram is also known as angular diagram.

Method of Construction:

The Surface area of a circle is known to cover 2ð radians or 360 degrees. The data to be represented through a circle diagram may therefore be represented through 360 degrees, part or sections of a circle. The total frequencies or value is equated to 360 and then the angle corresponding to component parts

are calculated. After determining these angles, the required sectors in the circle are drawn. Different shades or colours of design or different types of cross hatchings are used to distinguish the various sectors of the circle.

Working Rule:

Step I: Start by entering the data into a table.

Step II: Calculate the total by summing all the values in the table

Step III: For each value, divide by the total and multiply by 100 to obtain the percentage.

Step IV: To find the degrees for each "pie sector," use the full circle of 360 degrees.

Step V: Draw a circle and use a protractor to measure the degrees for each sector.

Example 2: 120 Students of a college asked about their favourite sports. The details of these sports are as under.

Favourite Sports	**No. of Students**
Football	6
Hockey	30
Cricket	48
Basketball	12
Badmintion	24

Represent the above data through a pie diagram.

Solution:

Favourite Sports	**No. of Students**	**Angle of the Circle**
Football	6	6\120 X 360 =18º
Hockey	30	30\120 X 360 =90º
Cricket	48	48\120 X 360 = 144º
Basketball	12	12\120 X 360 = 36º
Badmintion	24	24\120 X 360 = 72º
Total	120	360º

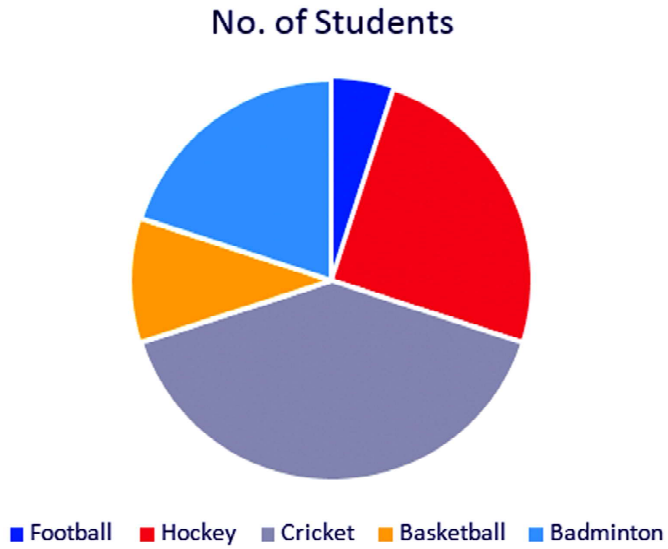


Figure 5: **Pie diagram: Represent the Favorite sports of the students.**

Example 3: Draw the suitable diagram for the following data:

Item	Steel	Bricks	Timber	Labour	Cement	Miscellaneous
Expenditure	20%	18%	10%	15%	25%	12%

Construct the suitable diagram.

Solution

Expenditure	**Central Angles**
20%	$\frac{20}{100} X\ 360 = 72°$
18%	$\frac{18}{100} X\ 360 = 64.8°$
10%	$\frac{10}{100} X\ 360 = 36.0°$
15%	$\frac{15}{100} X\ 360 = 54.0°$
25%	$\frac{25}{100} X\ 360 = 90°$
12%	$\frac{12}{100} X\ 360 = 43.2°$

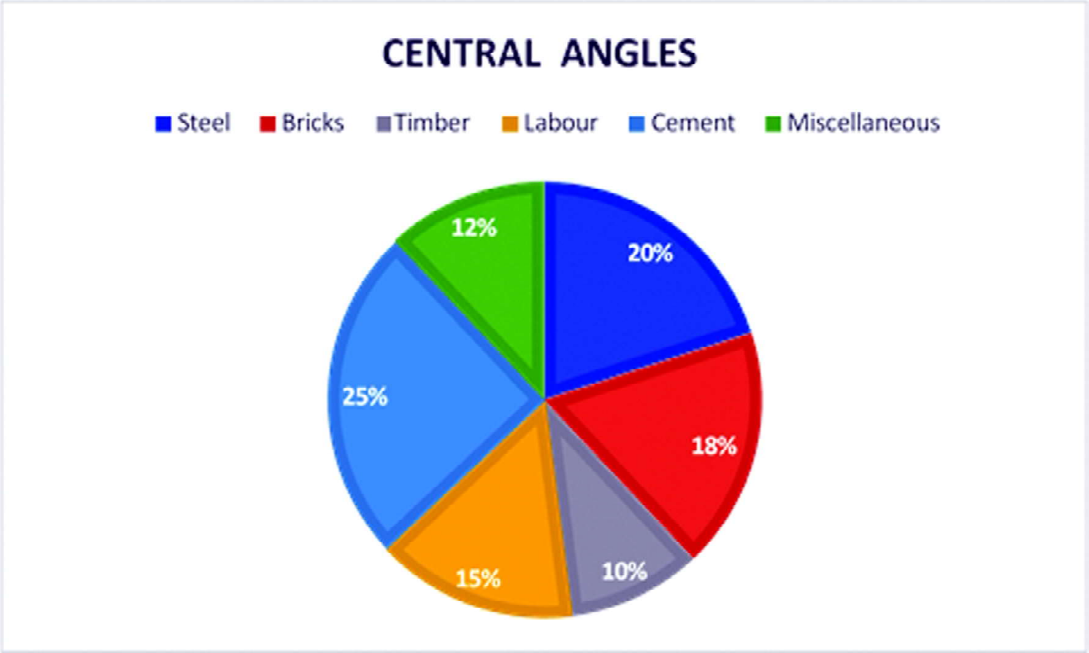


Figure 6: Represent the Expenditure on different items Uses of Pie Graph

- It is effective for illustrating proportions of a total in a visually striking manner.
- A pie diagram is used when a population is stratified, and each stratum needs to be displayed as a percentage.

Advantages of Pie Diagram

- The visual format is simple and easy to grasp.
- Data can be shown as parts of a whole.
- It acts as an effective communication tool for audiences that may not have prior knowledge of the topic.
- It enables quick visual comparisons, facilitating immediate analysis and understanding.

Disadvantages of Pie Graph

- Its effectiveness decreases when there are too many segments.
- When too many data points are present, even with labels and numbers, the segments can become cluttered and hard to read.
- As it only represents a single data set, it is not suitable for comparing multiple sets.
- This limitation can complicate the analysis and assimilation of information for the reader.

2.7. CHECK YOUR PROGRESS

Question 1: The given table represents the patient's body temperature recorded every hour in a hospital. Draw the line graph for the given information:

Time	9 am	10 am	11 am	12 noon	1 pm	2 pm	3pm
Temperature	34	35	38	37	34	35.5	36.5

Question 2: A word-nonsense syllables associated test was administered a on a student of class X to demonstrate the effect of practice on learning. The data so obtained may be student from the following table:

Trial No.	1	2	3	4	5	6	7
Score	4	5	8	8	10	13	12

Question 3: Construct a suitable bar diagram for following data:

College	**MAM**	**GWCPrade**	**Commerce**	**Total**
E	1200	800	600	2600
F	700	500	600	1800

Question 4: Represent the following data of Faculty-wise distribution of students, by a multiple bar diagram

College	**Arts**	**Science**	**Commerce**
A	1200	600	500
B	1000	800	650
C	1400	700	850
D	750	900	300

Question 5: construct the following data by a pie diagram

Item:	Food	Clothing	Recreation	Education	Rent	Miscellaneous
Expenditure:	87	24	11	13	25	20

2.8. GRAPHS FOR CONTINUOUS FREQUENCY DISTRIBUTION

The most commonly used graphs for charting a continuous frequency distribution are

- Histogram
- Frequency Polygon
- Frequency Curve
- Ogive or cumulative frequency curve

2.8.1 HISTOGRAM

Histogram for a continuous frequency distribution is constructed using adjacent rectangles where area of a rectangle represents the frequency of a class interval, i.e.

There are two ways of drawing rectangles.

- Height of the rectangle is proportional to the frequency and width is taken as unity. In case of unequal class intervals, frequencies are needed to be adjusted according to the following formula:

  where, . And, take the scale width of rectangle as unit for the smallest class interval.

- Another way is to draw height of the rectangles proportional to relative frequency and the width of the rectangle is simply proportional to the class magnitude.

Types of Histograms

There are two types of histograms

- Histogram with equal class interval
- Histogram with unequal class interval

Type I. Histogram with equal class interval

Consider the distribution of marks of students in **Table 5** . This distribution is of equal class interval and the class width in 10. So, the scale of the width of rectangles can be taken as 10 marks = 1 unit and take the height of the rectangles in proportion to the frequencies.

Table 5: Distribution of Marks of Students

Marks	No. of Students
10-20	18
20-30	20
30-40	22
40-50	21
50-60	19
60-70	10
70-80	10
80-90	20
90-100	10
Total	150

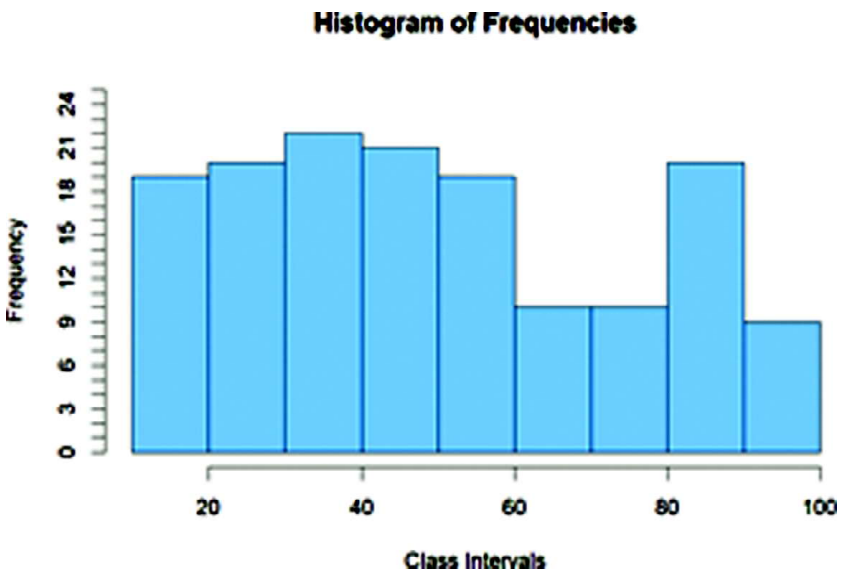


Figure 7: Distribution of Marks of Students

The histogram gives us an approximate idea about the nature of the distribution. By drawing the rectangles to represent the frequencies of the class interval assumes that the frequencies are equally distributed in the interval. However, this may not be really true. Frequency polygon and frequency curve can give better idea about the distribution.

Type II: Histogram with unequal class interval

Steps for Converting an exclusive and inclusive series

Step:1 Find the difference between the lower limit of second class interval and upper limit of the first class interval. (Say h)

Step:2 Subtract h/2 from the lower limit of each and every class and add h/2 to upper limit of every class

Step: 3 Now write down the frequency distribution with continuous class intervals

Step:4 Draw the histogram of the new distribution.

Example 4: Draw a Histogram of following data:

Class Interval	12-16	17-21	22-26	27-31	32-36	37-41	42-46
Frequency	2	6	7	5	3	7	9

Solution: Here, given histogram is not continuous. If we represent the given data by a graph, we shall not get a histogram. In a histogram the bars or rectangles are continuous without gaps. Therefore, we shall have to make class continuous by taking actual class limits.

Class Interval	11.5-16.5	16.5-21.5	21.5-26.5	26.5-31.5	31.5-36.5	36.5-41.5	41.5-46.5
Mid- Points	14	19	24	29	34	39	44
Frequency	2	6	7	5	3	7	9

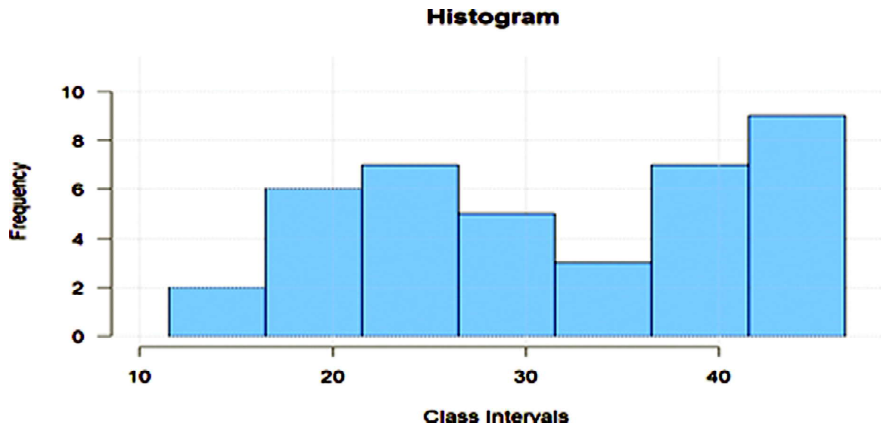


Figure 8: Represent the Histogram of Given Data Advantages of Histogram

- It is simple to create and easy to understand.
- It facilitates a quick and clear comprehension of the distribution.
- It offers greater precision compared to a frequency polygon.

Drawbacks of Histogram

- Multiple distributions cannot be plotted on the same axes with a histogram.
- Comparing more than one frequency distribution on the same axes is not achievable.
- It cannot be represented in a smooth form.

Importance of Histogram

- **Understanding Score Distribution**: It reveals how scores are spread within the group, indicating whether they cluster at the lower or higher end or are uniformly distributed across the scale.
- **Visual Representation**: This displays the data in a graphical format.

2.8.2 FREQUENCY POLYGON AND THE FREQUENCY CURVE

Frequency polygon is an alternative way of representing frequency distribution. It can be derived from the histogram by joining the mid-points of the tops of the consecutive rectangles by straight line. When these points are joined by smooth curve, it gives frequency curve. They are useful in getting approximate idea of the shape of the distribution.

Working Rules:

- **Calculate Midpoints**: Determine the midpoint for each class interval by averaging the lower and upper boundaries.
- **List Frequencies**: Write down the frequency for each class interval.
- **Plot Points**: On a graph, plot points using the midpoints of the class intervals on the x-axis and the corresponding frequencies on the y-axis.
- **Connect Points**: Join the plotted points with straight lines to form the frequency polygon.
- **Label the Graph**: Add titles, axis labels, and a legend if necessary for clarity.
- **Analyze the Graph**: Examine the shape of the frequency polygon to understand the distribution of the data.

The frequency polygon and frequency curve for the distribution in **Example 4** are presented in **Figure 9 to Figure 11**, respectively.

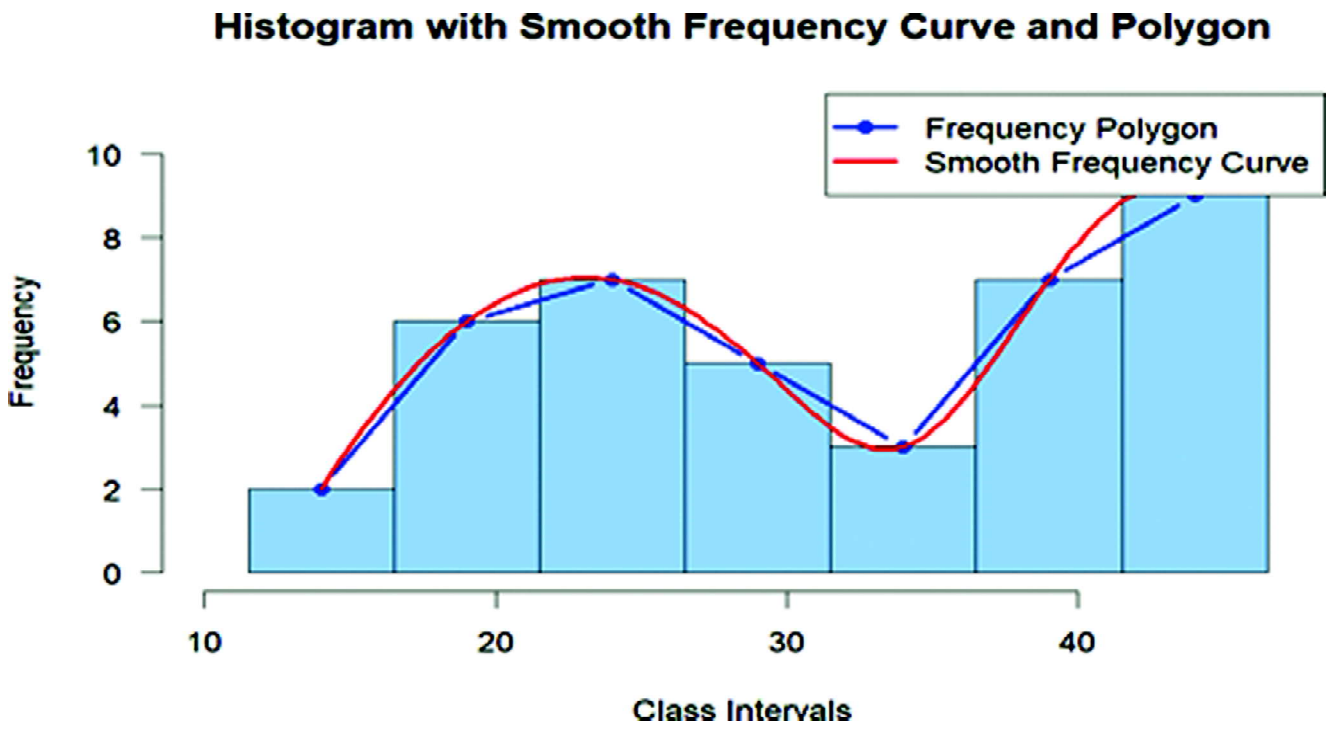


Figure 9: Frequency Polygon and Frequency Curve for Given Data

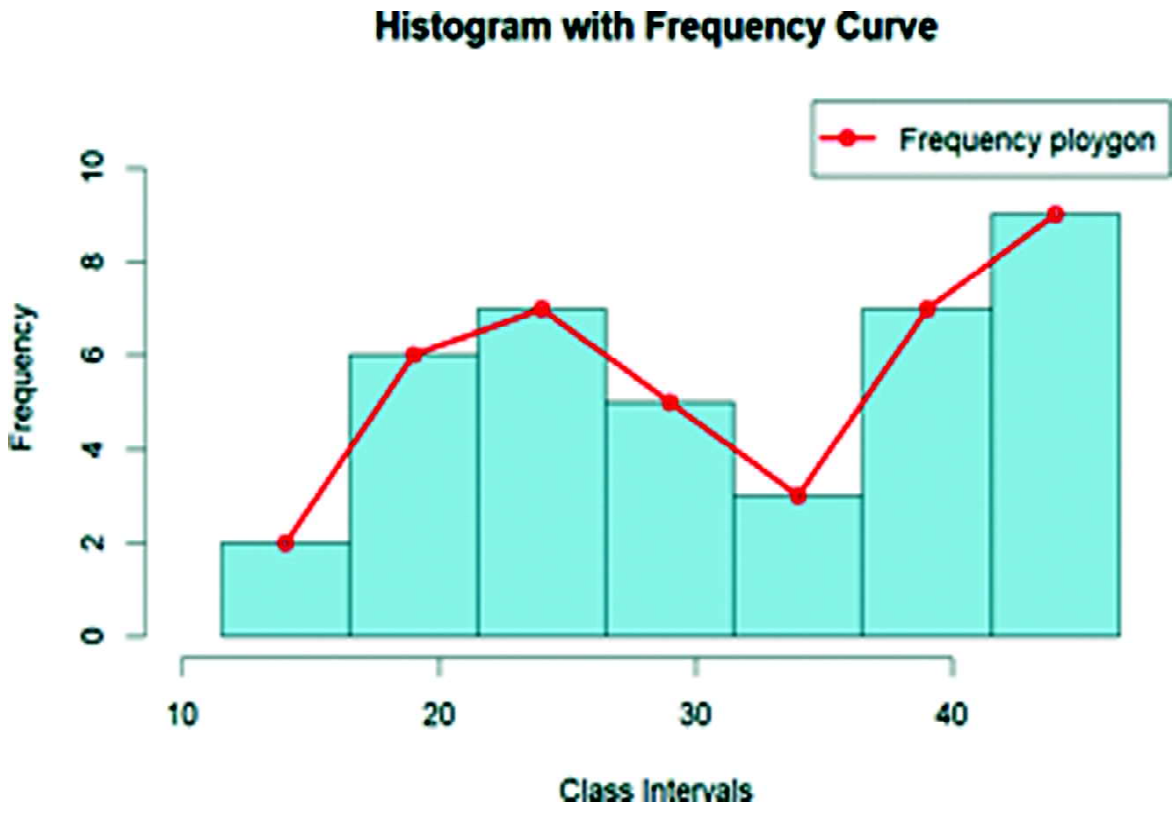


Figure 10: Represent the Frequency Polygon for Given Data

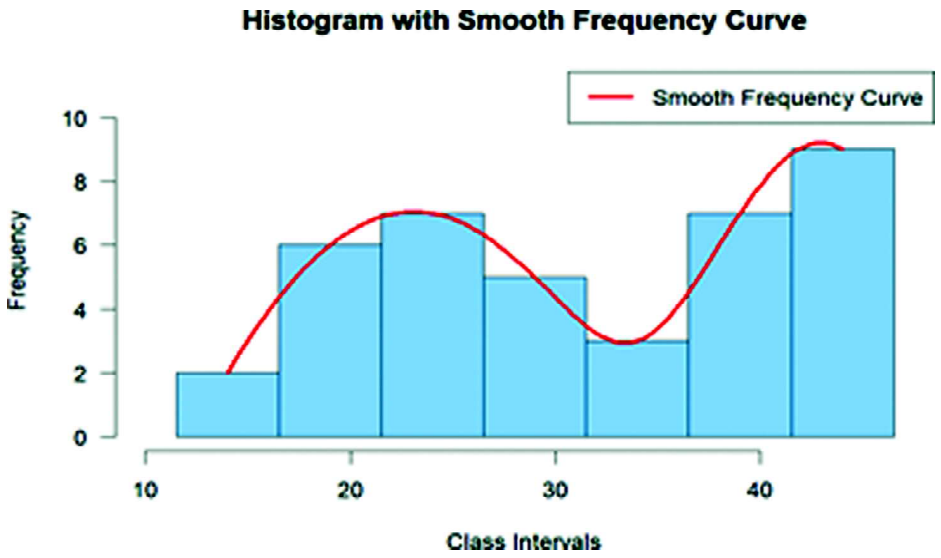


Figure 11: Frequency Curve for Distribution of the Given Data

Advantages:

- **User-Friendly**: It is straightforward to create and easy to comprehend.
- **Dual Distribution Plotting**: Two distributions can be plotted simultaneously on the same axes.
- **Comparative Analysis**: Frequency polygons allow for effective comparison between two distributions.
- **Smooth Representation**: The graph can be smoothed for clearer visualization.

Limitations:

- **Lacks Precision**: The representation is less precise.
- **Inaccurate Area Representation**: It does not accurately reflect the area under the frequency for each interval.

Applications of Frequency Polygon

- **Usage for Distribution Comparison**: The frequency polygon is employed to compare two or more distributions.
- **Visual Representation of Data**: It displays the data graphically.
- **Insight into Score Distribution**: It shows how scores are distributed in one or more groups, highlighting whether they are concentrated at the lower or higher ends or evenly distributed across the scale.

2.8.3 OGIVE OR CUMULATIVE FREQUENCY CURVES

The graphical representation of cumulative frequency distribution for continuous frequency distribution is a ogive. The cumulative frequencies are plotted against the corresponding class boundaries and the successive

points are joined by using smooth curve. The obtained curve is known as ogive or cumulative frequency curve.

There are two types of Ogives:

- Less than Ogive
- More than Ogive

For the less-than ogive, the cumulative frequencies are plotted against the upper limits of each class interval. The curve drawn from the data cumulated downward is known as less than ogive.

For the more-than ogive, the cumulative frequencies are plotted against the lower limits of each class interval. The curve drawn from the data cumulated upward as more than ogive. An ogive is used to find median, quartiles, deciles and percentiles etc.

Working Rules: Less than Ogive

Step I: Represent the upper limits (lower limit) of classes along X-axis.

Step II: Represent the cumulative frequency of respective frequency along y-axis.

Step III: Plot the points corresponding to upper (lower) class limits and cumulative frequencies less than the respective upper limits**.**

Step IV: Join the points plotted in step III by a free hand.

Step V: It is assumed that the class preceding the first class (succeeding the last class) in the classification exist and its frequency is zero. plot the point corresponding to this hypothetical point and join it to point of the first class. i.e. this point is (Lower limit of first class). The Curve so obtained is Less than Ogive or Cumulative frequency curve less than upper class limit.

Example 5: Construct a less-than ogive for the given frequency distribution.

I.Q	60-70	70-80	80-90	90-100	100-110	110-120	120-130
No. of Students	2	5	12	31	39	10	4

Solution:

I.Q	**Frequency**	**Less than**
60-70	2	2
70-80	5	7
80-90	12	19
90-100	31	50
100-110	39	89
110-120	10	99
120-130	4	103

Working Rules: More than Ogive

Step I: Represent the lower limit of classes along X-axis.

Step II: Represent the cumulative frequency of respective frequency along y-axis.

Step III: Plot the points corresponding to lower class limits and cumulative frequencies less than the respective upper limits**.**

Step IV: Join the points plotted in step III by a free hand.

Step V: It is assumed that the class succeeding the last class in the classification exist and its frequency is zero. Plot the point corresponding to this hypothetical point and join it to point of the first class. i.e. this point is (Lower limit of first class). The Curve so obtained is less than Ogive or Cumulative frequency curve less than upper class limit.

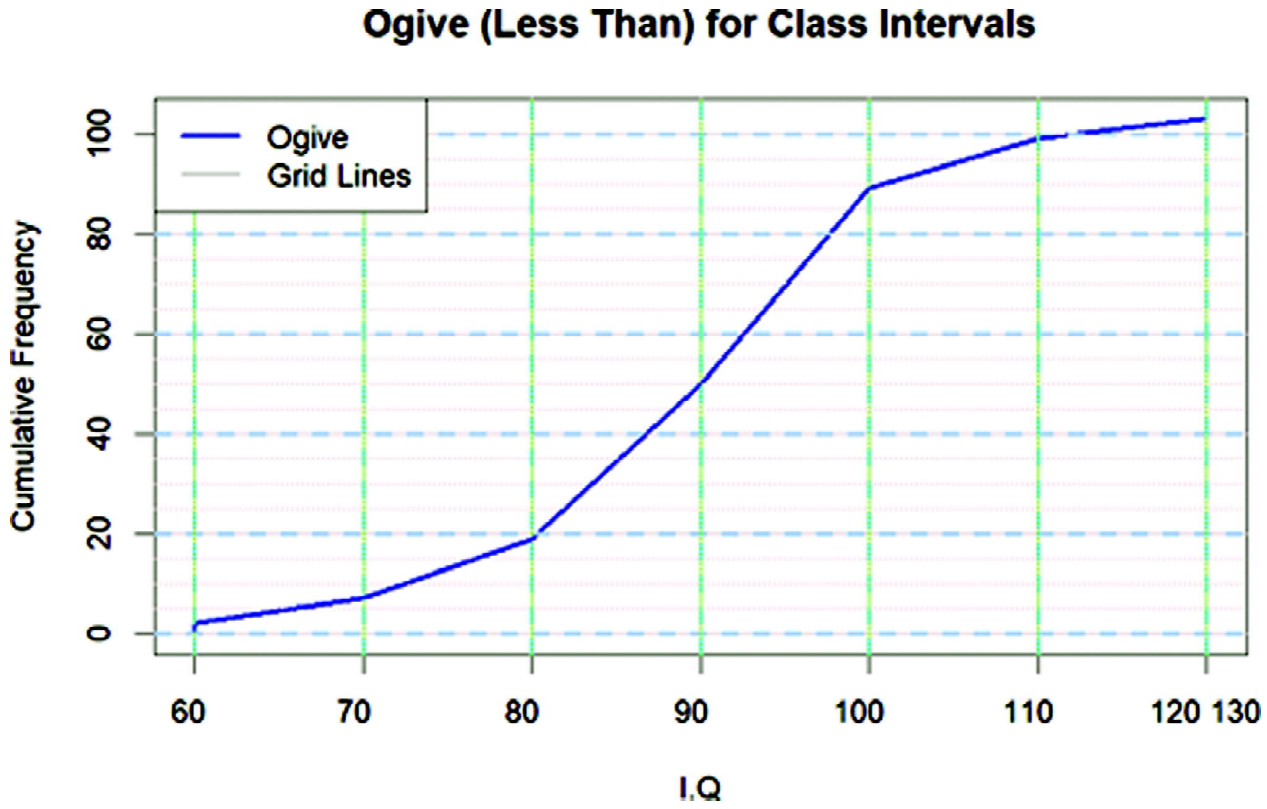


Figure 12: Represent the Graphical Representation of I.Q

Example 6: Plot more than ogive for given frequency distribution.

Wages	20	40	60	80	100
No. of Worker	41	92	156	194	201

Solution:

Weekly Wages	No. of Worker	More than
20	41	684
40	92	643
60	156	156
80	194	395
100	201	201

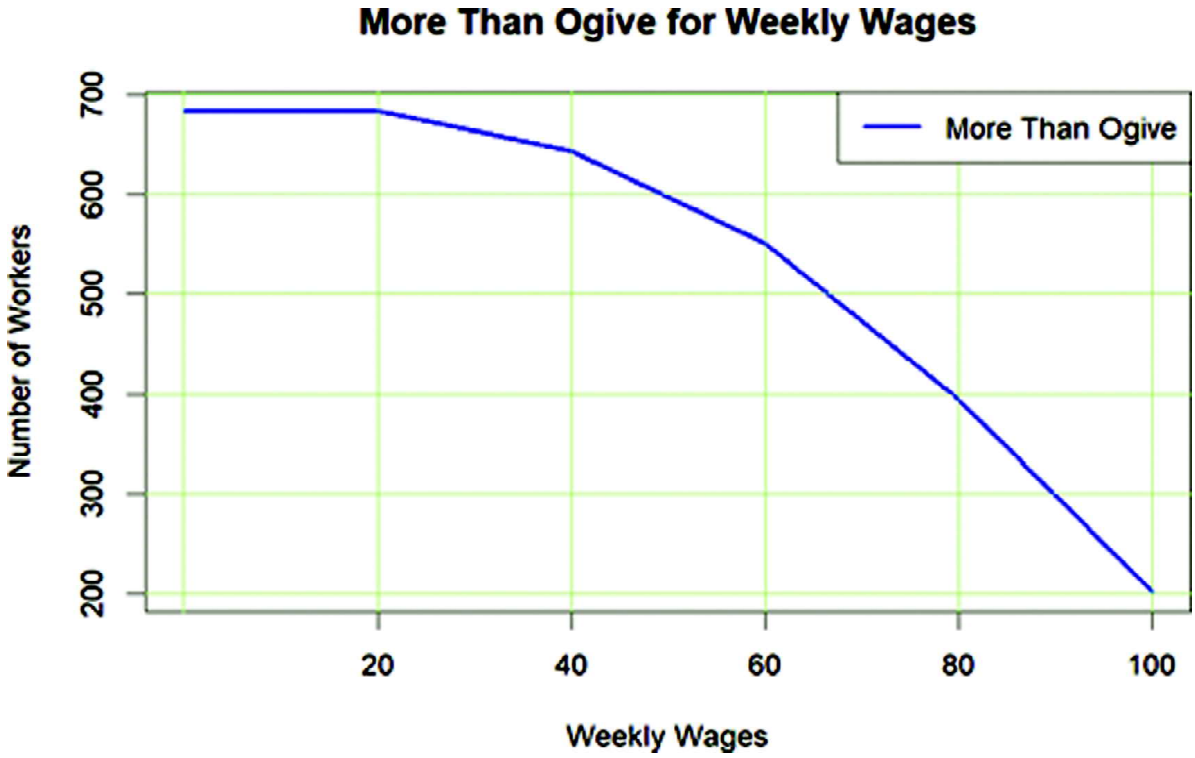


Figure 13: Represent the More than Ogive for Weekly Wages

Example 7: Plot both the more-than ogive and less-than ogive for the following data:

Cost of Product	4-6	6-8	8-10	10-12	12-14	14-16
Frequency	13	111	182	105	19	7

Solution:

Cost of Product	Frequency	Less than	More than
4-6	13	13	437
6-8	111	124	424
8-10	182	306	313
10-12	105	411	131
12-14	19	430	26
14-16	7	437	7

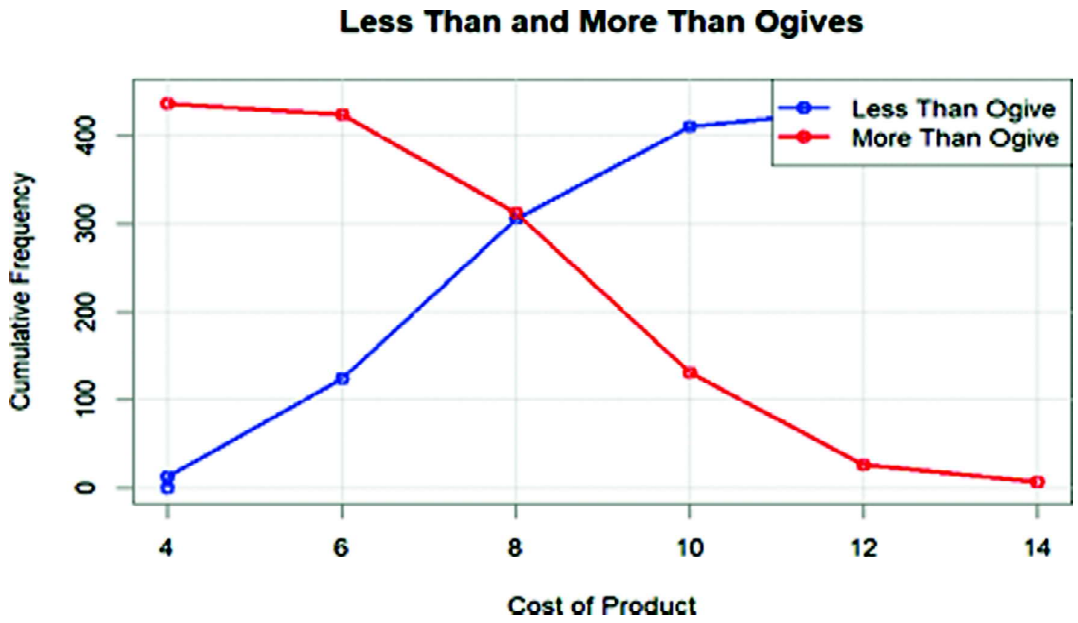


Figure 14: Represent the More than and Less than Ogive for Cost Production Advantages:

- **Clear Visual**: Ogives show a visual picture of how many values are below a certain point, making it easier to understand data.
- **Helps Analyze Data**: You can easily find important numbers like medians (the middle value) and percentiles (like the 25th percentile) using an ogive.
- **Comparison Tool**: You can use ogives to compare different sets of data to see how they stack up against each other.
- **Identify Trends**: They help you spot trends, such as whether data is evenly spread out or skewed to one side.
- **Summarizes Data**: Ogives summarize large amounts of data, so you don't have to look at every single number.

Disadvantages

- **Less Detail**: While ogives summarize data, they might hide specific details, like exact frequencies of each group.
- **Can Be Confusing**: If you're not familiar with them, ogives can be hard to understand.
- **Sensitive to Intervals**: The way you group the data (class intervals) can change the shape of the ogive, and bad choices can mislead you.
- **Not for All Data**: Ogives work best with continuous data (like heights) and might not be as useful for discrete data (like the number of people).

2.8.4 STEM-LEAF PLOT

One disadvantage of using a histogram to summarize data is that the original data aren't preserved in the graph. A stem-and-leaf plot, on the other hand, summarizes the data and preserves the data at the same time. A Stem and Leaf Plot is a special table where each data value is split into

- a "stem" (the first digit or digits) and
- a "leaf" (usually the last digit).

Example 8: Represent the following data using stem-leaf plot.

- 15, 16, 21, 23, 23, 26, 26, 30, 32, 41

Stem	Leaf
1	5 6
2	1 3 3 6 6
3	0 2
4	1

- Stem 2 and leaf 3 represent 23.

The "stem" is used to group the scores and each "leaf" shows the individual scores within each group.

Example 9: Obtain the original data set from the following stem-leaf plot.

Stem	Leaf
1	3 5 8
2	5 6 6 7 8 9
3	2 3 4 4
4	2 3

The observations are 13, 15, 18, 25, 26, 26, 27, 28, 29, 32, 33, 34, 34, 42, 43.

Example 10: Suppose data on long jump by 10 friends are as follows:

2.3, 2.5, 2.5, 2.7, 2.8 3.2, 3.6, 3.6, 4.5, 5.0

Solution:

Stem	Leaf
2	3 5 5 7 8
3	2 6 6
4	5
5	0
Stem "2" Leaf "3" means 2.3	

Example 11: Test scores of 50 students:

93	77	67	72	52	83	66	84	59	63
75	97	84	73	81	42	61	51	91	87
34	54	71	47	79	70	65	57	90	83
58	69	82	76	71	60	38	81	74	69
68	76	85	58	45	73	75	42	93	65

Solution:

Stem	**Leaf**	**Frequency**
3	4 8	2
4	2 2 5 7	4
5	1 2 4 7 8 8 9	7
6	0 1 3 5 5 6 7 8 9 9	10
7	0 1 1 2 3 3 4 5 5 6 6 7 9	13
8	1 1 2 3 3 4 4 5 7	9
9	0 1 3 3 7	5

2.8.5 BOXPLOT

A boxplot (or box-and-whisker plot) is a standardized way of displaying the distribution of data based on a five-number summary: minimum, first quartile (Q1), median (Q2), third quartile (Q3), and maximum.

Structure of a Boxplot:

- **Box:** The box itself represents the interquartile range (IQR), which contains the middle 50% of the data. It spans from Q1 to Q3.
- **Median Line:** A line inside the box shows the median (Q2), indicating the center of the data.
- **Whiskers:** The lines extending from the box (the whiskers) typically extend to the smallest and largest values that are not considered outliers. They represent data points within 1.5 times the IQR from Q1 and Q3.
- **Outliers:** Points that fall outside the whiskers are marked separately (often with dots or stars) and indicate data points that are significantly different from the rest.

Importance of Boxplot:

Boxplots provide a quick visual summary of a dataset, allowing viewers to grasp its central tendency and

spread at a glance.

Identification of Outliers: They effectively highlight outliers, helping analysts to identify anomalies in data that may require further investigation.

Comparison across Groups: When multiple boxplots are displayed side by side, they facilitate comparisons between different groups or categories, making it easy to observe differences in distributions.

Visualizing Distribution: Boxplots help in understanding the skewness of the data. For example, if the median line is closer to Q1 or Q3, it indicates skewness in the data.

Robustness: Boxplots are less sensitive to extreme values compared to other graphical methods like histograms, making them reliable for summarizing data distributions.

Boxplots are an invaluable tool in data analysis, promoting clarity and understanding. Their ability to convey complex information in a straightforward manner makes them essential for researchers, analysts, and decision-makers across various disciplines. By effectively visualizing data, boxplots enhance the interpretability of statistical results and support data-driven decision-making.

Example 12: How do sales figures compare across different regions? Sales figures for regions A, B, and C:

Region A: 200, 220, 250, 300, 320

Region B: 150, 180, 200, 250, 280

Region C: 400, 420, 450, 500, 550

Solution: A boxplot for each region can be created side by side. You can compare the medians and IQRs to see which region has higher sales and the variability within each region.

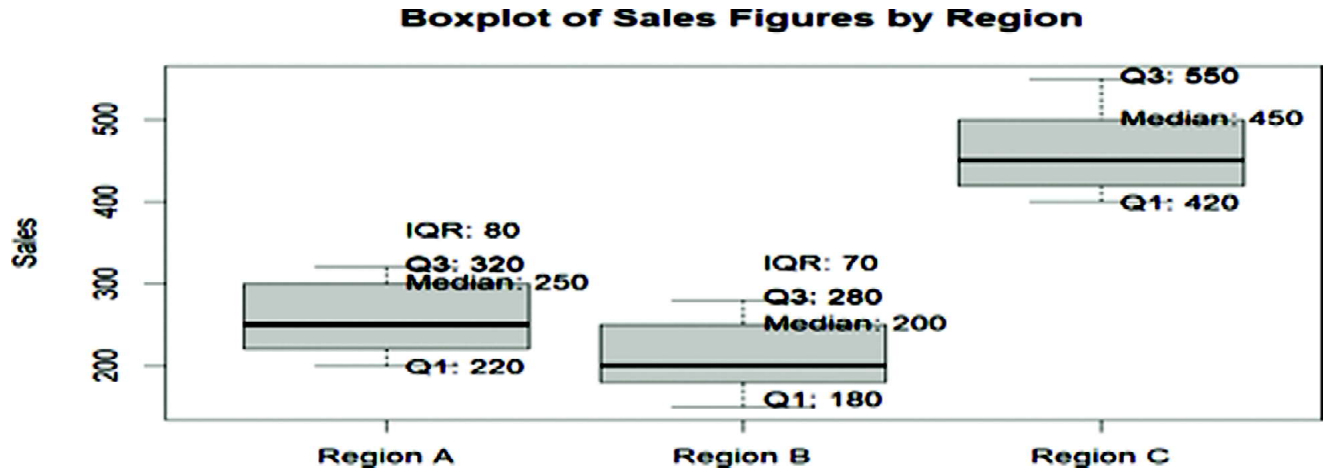


Figure 15: Boxplot compare sales of three different region

Example 13: Are there any outliers in the dataset of weights?

Weights (in kg): 55, 58, 60, 62, 70, 85, 90, 120.

Solution: The boxplot will identify any outliers based on the IQR method. In this case, the weight of 120 kg may appear as an outlier.

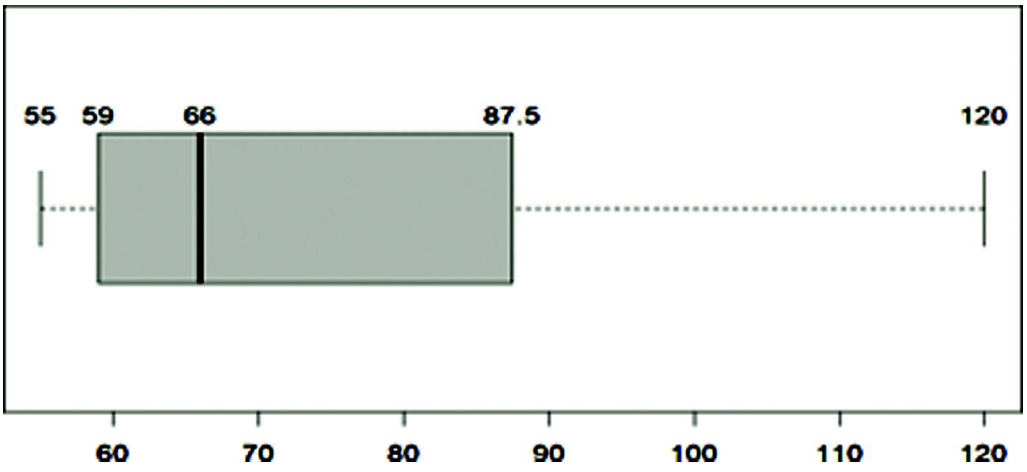


Figure 16: Boxplot for represent outlier

Example 14: What is the range of test scores in a class?

Test scores: 45, 67, 72, 78, 82, 85, 90, 92, 95.

Solution: The boxplot will show the minimum, Q1, median, Q3, and maximum. The range can be calculated as the difference between the maximum and minimum scores, which can be visually confirmed using the whiskers of the boxplot.

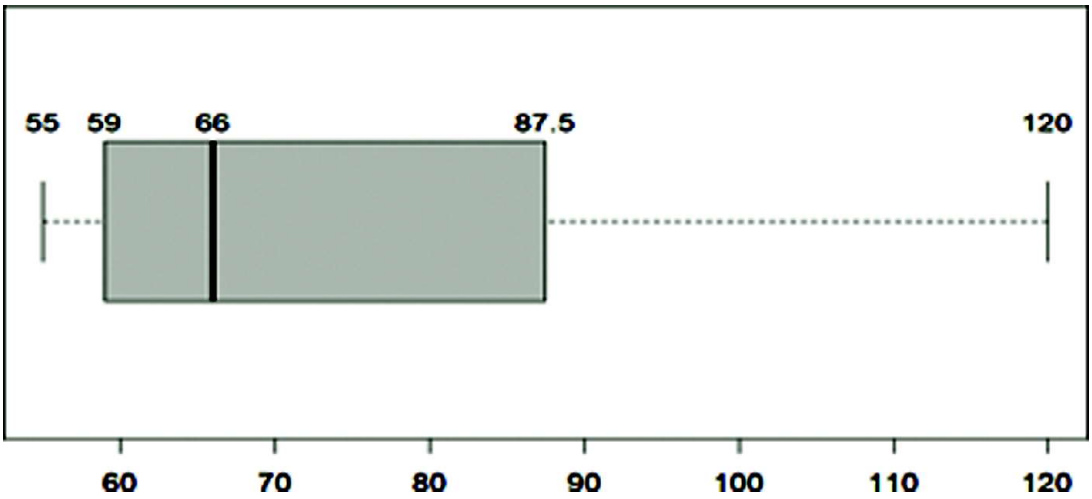


Figure 17: Boxplot represent the test scores.

2.9. CHECK YOUR PROGRESS

Question 1: Define a histogram. Explain how it differs from a bar chart.

Question 2: Given the following data on the number of books read by a group of students in a month:

Number of Books	Frequency
0-1	5
2-3	8
4-5	12
6-7	4
8-9	1

a. Construct a histogram for the data.

b. Describe the shape of the histogram and what it indicates about students' reading habits.

Question 3: Given the following data on the number of hours studied by students:

Hours Studied	Frequency
0-1	4
2-3	6
4-5	10
6-7	5
8-9	2

a. Create a frequency polygon for the data.

b. Explain the shape of the frequency polygon and what it indicates about students' study habits.

2.10. LET US SUM UP

Graphical representation of data uses visual tools like bar charts, histograms, pie charts, line graphs, scatter plots, and box plots to present information in an easily understandable way. These visuals are crucial for identifying patterns, trends, and relationships, helping to simplify complex data. For example, bar charts are effective for comparing categories, while line graphs track changes over time. Scatter plots reveal correlations, and box plots highlight data distribution and outliers. The importance of graphical representation lies in its ability to make data more accessible, allowing for clearer analysis and communication of findings,

especially in educational contexts, such as displaying student performance or survey results.

2.11. KEY POINTS/GLOSSARY

1. Frequency Distribution: A summary of how often different values occur in a dataset. Helps visualize data distribution and identify patterns.

2. Graphical Representation: Makes complex data easier to understand and interpret.

- Bar Chart: Compares different categories using bars.
- Histogram: Displays frequency distributions for continuous data.
- Pie Chart: Represents proportions of a whole.
- Boxplot: Displays data distribution through quartiles, highlighting median and outliers.

3. Ogive: A cumulative frequency graph that represents the accumulation of frequencies. Useful for determining medians and percentiles**.**

4. Pictograph: A visual representation using pictures or symbols to represent data.

2.12. SELF-ASSESSMENT QUESTIONS

Question1. Why is graphical representation of data important in understanding and analyzing information?

Question 2. What does a bar chart show?

a) Relationships between two variables
b) Categories and their values
c) Changes over time
d) Data spread

Question 3. What type of data is best represented by a pie chart?

a) Data over time
b) Parts of a whole
c) Numerical values
d) Comparisons of different categories

Question 4. What does a scatter plot show?

a) Data points connected by lines
b) The relationship between two variables
c) Data in percentage form
d) Data arranged in a circle

Question 5. What does a box plot help to identify?

a) Data trends
b) Data distribution and outliers
c) Parts of a whole
d) Data comparisons

2.13. LESSON END EXERCISE

Question 1: How does the use of bar graphs, pie charts, and line graphs enhance the understanding of educational outcomes and achievement gaps?

Question 2: Describe what a frequency curve is and how it is related to a histogram. Explain the process of converting a histogram into a frequency curve.

Question 3: Plot both the more-than ogive and less-than ogive for the following data:

Age	14-16	16-18	18-20	20-22	22-24	24-26
Frequency	130	111	182	105	190	70

Question 4: How can graphical representations of student performance data help educators identify trends and make informed decisions?

Question 5: Construct the more- than ogive for given frequency distribution.

Monthly Earning	20-40	40-60	60-80	80-100	100-120
No. of Worker	4	9	15	9	20

Question 6: Construct the less -than ogive for the following frequency distribution.

I.Q	60	70	80	90	100	110	120
No. of Students	20	59	80	38	39	60	40

Question 7: Construct a pie diagram to represent the given data:

Male	Female	Boys	Girls	Total
2000	1800	4200	2000	10000

Question 8: A person spends his time on different activities daily (in hours):

Activities	Exercise	Office Work	Travelling	Watching TV	Listening Music	Sleeping	Miscellaneous
No. of spent	1	8	2	2	1	7	3

Draw a pie chart for this information**.**

2.14. SUGGESTED READINGS

- Singh, Simon (2000). The Code Book: The Science of Secrecy from Ancient Egypt to Quantum Cryptography (1st Anchor Books ed. ed.). New York: Anchor Books. ISBN 0- 385-49532-3.

- Gupta, S. C., & Kapoor, V. K. (2020). Fundamentals of mathematical statistics. Sultan Chand & Sons.
- Tufte, E. R. (2001). The Visual Display of Quantitative Information (2nd ed.) (p.178). Cheshire, CT: Graphics Press.
- Goon, A.M., Gupta, M.K. and Dasgupta, B. (1968). Fundamental of Statistics, Vol I, World Press, Kolkata.
- Goon, A.M., Gupta, M.K. and Dasgupta, B. (1968). Fundamental of Statistics, Vol II, World Press, Kolkata.
- Gupta, S.C. and Kapoor, V.K. (2020). Fundamentals of Mathematical Statistics, 12th Ed., Sultan Chand and Sons.
- Moore, D.S. (2009). The Basic Practice of Statistics. 5th Ed., W H Freeman.
- Neter, J (1987). Applied Statistics. 3rd Ed. Allyn and Bacon. • Rowntree, D. (2018). Statistics without Tears: An Introduction for Non-Mathematicians. Penguin Press
- Starnes, D.S (2022). The Practice of Statistics, 8th Ed., W H Freeman. • Ummer, E.K. (2022). Basic Statistics for Economics, Business and Finance, Atlantic Publication.

LESSON : 3

MEASURES OF CENTRAL TENDENCY: TOOLS FOR SUMMARIZING DATA EFFECTIVELY

Structure

3.1 INTRODUCTION

Central tendency is a key concept in statistics that helps to identify the central or most typical value in a data set, providing a summary measure that represents the general trend of the data. The three main measures of central tendency are the mean, the median, and the mode. The mean is the arithmetic average, calculated by summing all values and dividing by the number of values, but it can be heavily influenced by outliers or extreme values. The median is the middle value when the data is ordered and is often more reliable than the mean when the data is skewed or contains outliers. The mode is the value that occurs most frequently, and a set can have one, more than one, or no mode at all. These measures are essential for summarizing data, making comparisons, and drawing conclusions.

In education, central tendency plays a crucial role by helping educators, researchers, and policymakers understand and summarize student performance, trends, and behaviours. By using measures like the mean, median, and mode, educators can track student progress, evaluate the effectiveness of teaching strategies, and make data-driven decisions. For example, the mean helps determine the average score of a class, while the median provides insight into the typical student's performance, especially in cases where data is skewed by outliers. The mode can identify the most common responses or performance levels. Understanding central tendency enables educators to offer targeted support, personalize learning, and create fair and effective educational environments.

3.2 LEARNING OBJECTIVE

After reading this lesson, student will be able to:

- know the meaning and need of measures of central tendency and measures of dispersion
- Learners will identify and differentiate between the mean, median, and mode.
- Learners will evaluate the strengths and weaknesses of the mean, median, and mode in different contexts, especially in the presence of outliers or skewed data.
- Learners will explore how understanding central tendency can help ensure fair assessment practices and create effective learning environments for all students.

3.3 CENTRAL TENDENCY/ LOCATION/AVERAGE

- A **central tendency/location/average** is a single number that represents the whole data, which is known as an average of the data.
- A central tendency is a score that indicates where the approximate centre of the distribution of the data to be located. It also explains about the shape and nature of the distribution.
- According to **Croxton and Cowden** "An average is a single value within the range of the data that is used to represent all the values in the series. Since an average is somewhere within the range of data, it is sometimes called a measure of central value".

Aims and Roles of Averages

1. **Summarizing Large Data Sets**: Large amounts of data can be difficult to digest. Averages provide a single figure that captures the essence of the information, making it easier to understand and remember
2. **Facilitating Comparison:** Averages are useful for quantitative analysis, reducing extensive observational data to a single value or calculation, which simplifies comparisons.
3. **Clarifying Relationships:** Averages are important for revealing relationships between different data groups or variables, offering valuable insights.
4. **Assisting Decision-Making:** Averages act as reference points for decision-makers. Many analyses and planning decisions are based on the overall values of these variables.

Characteristics of an Effective Average / Measure of Central Tendency

1. **Clear Definition:** An effective average should be easily understood and have a single interpretation.
2. **Consistency:** Its value should remain the same regardless of the method or formula used for calculation.
3. **Inclusiveness:** The average should consider all items in the dataset, so that removing any single item will change its value.
4. **Resistance to Outliers**: The average should not be significantly affected by extreme values; one or two outliers shouldn't distort its overall value.
5. **Comprehensive Dependence**: It should take into account every item in the dataset.
6. **Sampling Stability:** The average should exhibit stability when calculated from different samples of the same population, producing similar results.
7. **Analytical Utility:** It should allow for further quantitative and statistical analysis, enabling the use of measures such as dispersion and correlation.

3.4. MEASURES OF CENTRAL TENDENCY

The following are five commonly used measures of central tendency or average:

I. Arithmetic average or arithmetic mean or simple mean

II. Median

III. Mode

IV. Geometric mean

V. Harmonic mean

The arithmetic mean, harmonic mean and geometric mean are typically referred to as mathematical averages, while the mode and median are known as positional averages..

3.4.1 ARITHMETIC MEAN

Arithmetic mean is the most commonly used measure of central tendency. These are methods to find the arithmetic mean:

(i) Direct Method

(ii) Indirect Method

(i) **Direct Method:**

Arithmetic mean for ungrouped data:

Let $X_1, _2, X_3, \ldots X_n$, are the n observations, then their arithmetic mean is given as:

$$X = \frac{X_1 + X_2 + X_3 + \ldots X_n}{n} \text{ or } = \frac{\Sigma_{i=1}^{n} X_1}{n}$$

In case of frequency distribution:

$$X = 1\ X_2,\ X_3,\ X_4,\ \ldots.\ \ldots.\ X_n$$

$$f = f_1, f_2, f_3, f_4\ \ldots\ \ldots f_n$$

Arithmetic mean is given by:

$$\mathrm{X} = \Sigma_{i=1}^{n} 1 \frac{fix}{}, \text{ where } \mathrm{N} = \Sigma_{i=1}^{n} f$$

Note: *In case of grouped frequency distribution, the value of X is taken as mid value of the respective class interval.*

ii) **Indirect Method**

The deviation method is used to calculate the mean of ungrouped data when the frequencies and values of the variables are significantly large, making it challenging to compute the arithmetic mean directly.

Assumed Mean Method: In this context, the provisional mean AAA is defined as the value of X (the mid-point of the class interval) that corresponds to the highest frequency or which comes near the middle value of the frequency distribution. This number is called the Assumed mean. Then the arithmetic mean is given by the formula.

i) Ungrouped data

$$X = A + \sum_{i=1}^{n} \frac{d_i}{N}$$

Suppose A be any assumed mean (or any assumed number), d the deviation from the assumed number

Working Rule for Ungrouped Data:

Step I: Identify the variables of the discrete series as X (or x)

Step II: Choose any item from the series, preferably the middle one, and label it as A. This value is known as the assumed mean.

Step III: Determine the difference X"A and label it as d, representing the deviation of any variate from A.

Step IV: Calculate the sum " 5ØQÜ5ØVÜ.

Step V: Use the following formula to find the arithmetic mean

$$X = A + \sum_{i=1}^{n} \frac{d_i}{N}$$

where $d_i = X_i - A$

5ØVÜ=1

Working Rules for Grouped Data:

Step I: In the case of discrete series, denote the variable by x or X and the corresponding frequency by f.

Step II: Choose any item from the x series, ideally the middle one, and label it as A. This value is known as the assumed mean.

Step III: Calculate the difference x"A and denote it as d.

Step IV: Multiply each f by its corresponding d and record the products in a column labelled *fd.*

Step V: Compute the sum " 5ØSÜ5ØQÜ

Step VI: Use the following formula to determine the arithmetic mean

$$X = A + X = A + \sum_{i=1}^{n} \frac{d_i}{N}$$

where $d_i = X_i - A$

Step Deviation Method:

$$X = A + \sum_{i=1}^{n} \frac{d_i}{N} * n$$

where $d_i = \frac{x_i A}{h}$ and h is the **"width of the class interval.**

Example 1. Calculate the arithmetic mean of the following distribution

Variants (X):	6	7	8	9	10	11	12
Frequency (f):	20	43	57	61	72	45	39

Solution:

X	**f**	**fX**
6	20	120
7	43	301
8	57	456
9	61	549
10	72	720
11	45	495
12	39	468
	N=337	∑ f X=3109

$$\bar{x} = \frac{3109}{337} = 9.225$$

Example 2. Find the arithmetic mean of the following by direct and assumed mean methods both:

Class :	20-30	30-40	40-50	50-60	60-70	70-80
Frequency :	8	26	30	20	16	15

Solution:

By direct method

$$X = \frac{\Sigma fX}{N} = \frac{3109}{337} = 9.22$$

Class	**Mid Value (X)**	**Frequency (f)**	**f.X**	**d= X-A**	**f.d**
20-30	25	8	200	-20	-160
30-40	35	26	910	-10	-260
40-50	45	30	1350	0	0
50-60	55	20	1100	10	200
60-70	65	16	1040	20	320
70-80	75	15	1125	30	450
Total		N = 115	∑ f X = 5725		∑f d = 550

By direct method

$$X = \frac{\Sigma fX}{N} = \frac{5725}{115} = 49.78$$

By assumed mean method,

Let assumed mean A= 45.

$$X = A + \frac{\Sigma fd}{N}$$

$$= 45 + (550/115) = 49.78$$

Example 3: Calculate the mean using the Step Deviation Method:

Age(X)	10-20	20-30	30-40	40-50	50-60
Frequency	5	8	12	10	5

Solution: First, we calculate the midpoint (xxx) for each class interval:

$$X = \frac{Upper\ limit + Lower\ Limit}{2}$$

Age	**Frequency**	**Mid-Point**	fx	di	**Fd**
10-20	5	15	75	-2	10
20-30	8	25	200	-1	8
30-40	12	35	420	0	0
40-50	10	45	450	1	10
50-60	5	55	275	2	10
Total	40		1420		2

$$\text{Mean} = \Sigma_{i=1}^{n} \frac{fx}{N} - \frac{1420}{40} = 35.5$$

For the Step Deviation Method, we simplify calculations by selecting an assumed mean (AM). Let's assume A=35A = 35A=35 (which is close to our actual mean).

Next, we calculate *d* values, which represent the deviation from the assumed mean, normalized by the class width. The class width here is 10.

$$d = \frac{x - A}{h}$$

where h = 10.

Mean by Step Deviation:

$$X = A + \sum_{i=1}^{n} \frac{f_i d_i}{N} * h$$

$$X = 35 + \sum_{i=1}^{n} \frac{2}{40} * 10$$

$$= 35 + 0.5 = 35.5$$

Properties of Mean

Property 1: The algebraic sum of the deviations of all the variates from their arithmetic mean equals zero.

Property 2: The sum of the squares of the deviations of a set of values is minimum when taken about mean.

Property 3: If each observation is multiplied by p($\neq 0$), then mean of the new observation is p . $\overline{x}$

Property 4: If each observation is divided by p (‘‘0), then mean of the new observation is $\frac{x}{p}$

Merits and Demerits of Mean;

Merits of Mean:

1. **Simplicity:** From the view point of calculation and usage, arithmetic mean is the simplest of all the measures of central tendency
2. **Certainty**: Arithmetic mean is a certain value; it has no scope for estimated values.
3. **Based on all items**: Arithmetic mean is based on all items in a series. It is, therefore representative value of the different items
4. **Algebraic treatment:** Arithmetic mean is capable of further algebraic treatment. It is therefore, extensively used in statistical analysis.
5. **Stability:** Arithmetic mean is a stable measure of central tendency. This is because changes in the sample of a series have minimum effect on the arithmetic average.
6. **Basis of comparison**: Being stable and certain, arithmetic mean can be easily used for comparison.
7. **Accuracy test:** Arithmetic mean can be tested for its accuracy as a representative value of the series.

Demerits of Mean:

1. **Affected by Outliers**: If there are very high or very low numbers in the data, they can change

the mean a lot, making it not very reliable.

2. **Not Always Accurate**: If the data is skewed (not balanced), the mean might not show the true middle value. Sometimes, the median (the middle number) is better.
3. **Ignores Data Spread**: The mean doesn't show how spread out the numbers are, which is important to understand the data.
4. **Treats All Values the Same**: The mean treats every number equally, which might not be fair in some situations where certain numbers matter more.
5. **Only for Numbers**: You can only calculate the mean for numerical data, so it doesn't work for categories like colors or names.
6. **Can Mislead with Small Samples**: In small groups of data, the mean can be easily influenced by just one or two numbers, making it misleading.

Some Common Uses of the Mean

1. **Averaging Scores**: Schools often use the mean to calculate average grades for students, giving a quick view of their overall performance.
2. **Understanding Income**: Economists use the mean to find the average income in a region, helping to analyze economic conditions.
3. **Comparing Data**: The mean helps compare different groups, like finding the average height of boys versus girls in a class.
4. **Weather Data**: Meteorologists calculate the mean temperature over a month to understand weather trends.
5. **Sports Statistics**: In sports, the mean can show the average points scored by a player over a season, helping to evaluate performance.
6. **Business Analysis**: Companies might use the mean to find the average sales over time, helping them make decisions about products.
7. **Health Studies**: Researchers often use the mean to find the average weight or blood pressure of a group, which can help in health assessments.

3.4.2. MEDIAN

The median is defined as the measure of the central value, when the given terms (i.e., values of the variate) are arranged in the ascending or descending order of magnitudes. In other words, the median is value of the variate for which total of the frequencies above this value is equal to the total of the frequencies below this value.

"*The median is the value of the variable which divides the group into two equal parts one part comprising all values greater, and the other all values less than the median".* To understand this concept let us take an example:

The marks received, by seven students in a paper of English are 16, 22, 28, 35, 38, 49, 52, 57, 59 the maximum marks being 60, then the median marks is 38 since it is the value of the 6^{th} term, which is situated such that the marks of 1^{st}, 2^{nd} ,$3^{rd,}$ 4^{th} and 5^{th} students are less than this value and those of 7^{th}, 8^{th} ,9^{th} and 10^{th} students are greater than this value.

Computation of Median

a) Median in individual series.

Let n be the number of values of a variate (i.e. total of all frequencies). First of all we write the values of the variate (i.e., the terms) in ascending or descending order of magnitudes.

Here two cases arise:

Case 1. If n is odd then value of (n+1)/2th term gives the median.

Case2. If n is even then there are two central terms i.e. $\left(\frac{n}{2}\right)$ and $\left(\frac{fn/2^{th}}{2}\right)$ these two values in these positions gives the median.

(b) Median in continuous series (or grouped series).

In this case, the median (Md) is computed by the following formula

$$Md = L_1 + \frac{\frac{n}{2} - cf}{f} \times i$$

Md = median

L_1 = lower limit of median class

cf = total of all frequencies before median class f = frequency of median class

i = class width of median class.

Example 3: According to the census of 1991, following are the population figure, in thousands, of 10 cities:

1400, 1250, 2600, 1670, 1800, 700, 650, 570, 488, 2500, 2100, 1700, 2200.

Find the median.

Solution. Arranging the terms in ascending order.

488, 570, 650, 700, 1250, 1400, 1670, 1800, 2100, 2200, 2400, 2600.

Here n =12, therefore the median is the mean of the measure of the 6th and 7th terms.

Here 6th term is 1400 and 7th term is 1670.

Median M_d = (1400+1670)/2

= 1535 thousands.

Examples 4. Find the median for the following distribution:

Wages in Rs. :	0-10	10-20	20-30	30-40	40-50	60-70	70-80
No. of workers :	22	38	46	35	20	40	45

Solution. We shall calculate the cumulative frequencies.

Wages in Rs.	No. **of Workers(f)**	**CumulativeFrequencies (c.f.)**
0-10	22	22
10-20	38	60
20-30	46	106
30-40	35	141
40-50	20	161
50-60	40	201
60-70	45	246

Here N = 246. Therefore, median is the measure of $\left(N+\frac{1}{2}\right)^{th}$

term i.e 123.5st term. Clearly

123.5st term is situated in the class 20-30. Thus 20-30 is the median class. Consequently,

$$M_d = L_1 + \frac{\frac{n}{2} - cf}{1} \times i$$

$$= 20 + \frac{\frac{246}{2} - 60}{46} \times 10$$

$$= 20 + 13.69 = 33.69$$

Example 5: Find the Median of the following data, if the marks scored by the students in a class test out of 50 are,

Marks	0-10	10-20	20-30	30-40	40-50
Number of Students	6	8	10	8	8

Solution:

Marks(X)	**Number of Students(f)**	**Cumulative Frequency(c.f)**
0-10	6	6
10-20	8	14
20-30	10	24
30-40	8	32
40-50	8	40

Here N = 40. Therefore, median is the measure of $\left(N+\frac{1}{2}\right)^{th}$ term i.e 20.5^{st} term. Clearly 20.5^{st} term is situated in the class 20-30. Thus 20-30 is the median class.

Consequently.

$$M_d = L_1 + \frac{\frac{n}{2} - cf}{f} \times i$$

$$= 20 + \frac{\frac{40}{2} - 14}{10} \times 10$$

$$= 20 + 6 = 26$$

Merits of Median:

1. **Not Affected by Outliers**: The median isn't influenced by very high or very low numbers, so it gives a better middle value when there are extreme values in the data.
2. **Better for Skewed Data**: In cases where data is not evenly spread out, the median often represents the central point more accurately than the mean.
3. **Easy to Understand**: The median is simple to calculate and understand, making it accessible for most people.
4. **Works with Ordinal Data**: The median can be used with ordered data (like rankings) where calculating a mean doesn't make sense.

Demerits of Median:

1. **Ignores All Data Points:** The median only looks at the middle value, ignoring all other numbers,

which can be less informative about the overall data set.

2. **Less Useful in Small Samples:** In very small groups, the median can be less reliable because it may not represent the overall trend well.
3. **Not as Commonly Used:** People often think of the mean first, so using the median might confuse some when comparing averages.
4. **Requires Sorting:** To find the median, you need to arrange the data in order, which can take extra time for large data sets.

Uses of Median:

1. It is useful in those cases where numerical measurement is not possible.
2. It is also useful in those cases where mathematical calculations cannot be made in order to obtain the mean
3. It is generally used in studying phenomena like skill, honesty, intelligence etc.

Example 13: The following data are pertaining to the number of members in a family. Find median size of the family.

Number of members (x)	1	2	3	4	5	6	7	8	9	10
Frequency (F)	1	3	5	6	10	13	9	5	3	2

3.4.3 MODE

The term “mode” originates from the French word “la mode,” which translates to “in fashion.” Dr. A. L. Bowley defines the mode as the value of a graded quantity in a statistical group where the number of occurrences is greatest; it is also known as the point of highest density or the predominant value. In simpler terms, the mode is the value that occurs most frequently in the distribution, marking the point of maximum frequency or density. A data set with two modes is called bimodal. A data set with three modes is called trimodal.

- Examples: Single Mode
  Data Set = 3, 6, 9, 2, 6, 5, 8
  Mode = 6
- Examples: Bimodal
  Data Set = 3, 6, 2, 3, 6, 2, 4, 8
  Modes = 2 and 6
- Examples: Trimodal

Data Set = 2, 6, 2, 8, 6, 4, 8

Modes = 2, 6, and 8

Mode for individual series of observations:

Mode for Grouped data

In the case of grouped frequency distribution, calculation of mode just by looking into the frequency is not possible. To determine the mode of data in such cases we calculate the modal class. Mode lies inside the modal class. The mode of data is given by the formula:

$$\text{Mode} = L_1 + f_1 - \frac{f_0}{2f_1 - f_0 - f_1} \times i$$

L_1 is the Lower limit of the model class

f_1: frequency of model class

f_0: frequency of the class just preceding to the modal class

f_2: frequency of the class just following of the modal class

i: length of class interval

Example 6. Compute the mode of the following distribution:

Class	:	0-7	7-14	14-21	21-28	28-35	35-42	42-49
Frequency	:	19	25	36	72	51	43	28

Solution: Here maximum frequency 72 lies in the class-interval 21-28. Therefore 21-28 is the modal class.

$L_1 = 21, f_1 = 72,\ f_0 = 36, f_2 = 51,\ i = 7$

$$\text{Mode} = L_1 + \frac{f_1 - f_0}{2f_1 - f_0 - f_1} \times i$$

$$= 21 + \frac{72 - 36}{2 \times 72 - 36 - 51} \times 7$$

$$= 21 + 252 / 57$$

$$= 21 + 4.421052$$

$$= 25.421052$$

Example 7: Calculate mode for the following:

Sales (Interval)	Frequency
10-20	4
20-30	6
30-40	10
40-50	8
50-60	2

Solution: The highest frequency is 10 and corresponding class interval is 40-50, which is the modal class.

$L_1 = 30, f_1 = 10, f_0 = 60, f_2 = 8, i = 10$

$$\text{Mode} = L_1 + \frac{f_1 - f_0}{2f_1 - f_0 - f_1} \times i$$

$$= 30 + \frac{10-6}{2\times10-6-8}$$

$$= 30 + 3.33$$

$$= 33.33$$

Example 8 : If the mean and median of a moderately asymmetrical series are 25 and 28, respectively, what is the most probable mode?

Solution: To find the mode, we can use the empirical formula

Mode = 3x Median – 2x Mean

= 3 x28 – 2x25

= 84 – 50

= 34

Advantages and Disadvantages of the Mode:

Advantages:

- The mode is easy to understand and calculate.
- The mode is not affected by extreme values.
- The mode is easy to identify in a data set and in a discrete frequency distribution.
- The mode is useful for qualitative data.
- The mode can be computed in an open-ended frequency table.
- The mode can be located graphically.

Disadvantages:

- The mode is not defined when there are no repeats in a data set.
- The mode is not based on all values.
- The mode is unstable when the data consist of a small number of values.
- Sometimes data have one mode, more than one mode, or no mode at all.

3.4.4 CHECK YOUR PROGRESS

1. The.....is the measure of central tendency that represents the most frequent value in a data set.
2. The.....is the sum of all values in a data set divided by the number of values
3. The.....is the middle value when a data set is ordered from least to greatest.
4. In a symmetrical distribution, the.....,....., and _____ are all equal.
5. The.....is particularly useful when dealing with skewed data, as it is not affected by outliers.
6. The.....is the best measure of central tendency when the data set contains outliers, because it is not influenced by extreme values.
   (Answer: median)

3.4.5 GEOMETRIC MEAN

If X_1, X_2.....X_n are n values of the variate X, none of which is zero. Then their geometric mean G is defined by

$$G = (X_1, X_2, ...X_n)$$

If f_1, f_2 .....f_n are the frequencies of X_1, X_2 .....X_n respectively, then geometric mean G is given by

$$G = (X_1 f_1 . X_2 f_2 ... X_n f^{\frac{1}{n}}) \ldots (1)$$

1 2 5Ø[Ü

where $N = f_1 + f_2 + ... + f_n$

Taking logarithm of (1), we get

$$\log (G) = \frac{1}{N} \Sigma_{i=1}^{n} f_i (X)$$

Our objective s to find the geometric mean so takes antilog of (2). Hence the geometric mean will be

$$G = \text{antilog} \left(\frac{1}{N} \sum_{i=1}^{n} f_i \log(X_i) \right)$$

Example 9: Calculate the geometric mean (G.M.) of the following series of monthly income of a batch of families 170, 200, 450, 1200, 1000.

X	Log X
170	2.23
200	2.30
450	2.65
1200	3.07
1000	3.00
Total	13.25

G.M. = Antilog $\sum_n^{\log xi}$ = Antilog 13.35 = Antilog 2.65

Example 10. Compute the geometric mean of the following distribution:

Marks :	0-10	10-20	20-30	30-40
No. of Students :	5	8	3	4

Class	Mid value X	No. of students	Log10 X	(f log X)
0-10	5	5	0.6990	3.4950
10-20	15	8	1.1761	9.4088
20-30	25	3	1.3979	4.1937
30-40	35	4	1.5441	6.1764
		N = Σf = 20		Σ (f log X) =23.2739

Merits of Geometric Mean

1. It is based on all the items.
2. It is capable of further algebraic treatment.
3. It gives less weight to large items and more to small items.

Demerits of Geometric Mean

1. It is difficult to compute.
2. It is not easy to understand.
3. If observations have negative values in the series, it can't be computed.

3.4.6 HARMONIC MEAN

The harmonic mean can be expressed as the reciprocal of the arithmetic mean of the reciprocals of the given set of observations.

The harmonic mean H of the positive real numbers $X_1, X_2, \ldots, X_n$ is defined to be

$$HM - \frac{n}{\frac{1}{x_1}+\frac{1}{x_2}+\ldots+\frac{1}{X_n}}$$

For the frequency distribution $X_i | f_i$ the harmonic mean will be given by the formula:

$$HM = \frac{N}{\frac{f}{X_1^1}+\frac{f}{X_2^2}+\ldots+\frac{1}{X_n^n}}$$

Example 11. Calculate the harmonic mean of the marks received in the examination listed below.

Marks :	15	22	28	34	42	44	48
No. of Students:	13	17	8	15	24	12	10

Solution:

Marks X	**No. of students (f)**	$\frac{1}{X}$	$t\frac{1}{X}$
15	13	0.066666667	0.866666667
22	17	0.045454545	0.772727273
28	8	0.035714286	0.285714286
34	15	0.029411765	0.441176471
42	24	0.023809524	0.571428571
44	12	0.022727273	0.272727273
48	10	0.020833333	0.208333333
	$N = \Sigma f = 99$		$\Sigma f\left(\frac{1}{X}\right)$ =3.418773873

$$\text{HM} = \frac{N}{\Sigma f\left(\frac{1}{X}\right)} = \frac{99}{} = 28.957$$

Example12: The marks obtained by several students in a class are provided below. Please calculate the harmonic mean.

Marks	20	35	40	45	50
Number of Students	10	15	10	15	20

Solution:

Marks (X)	No. of students (f)	$\frac{1}{x}$	$t.\frac{1}{x}$
20	10	0.05	0.5
35	15	0.028	0.42
40	10	0.025	0.25
45	15	0.022	0.33
40	20	0.025	0.50
	$N = \Sigma f = 70$		$\Sigma f\left(\frac{1}{X}\right) = 2$

$$HM\ \frac{N}{\Sigma f\left(\frac{1}{X}\right)} = \underline{70} = 35$$

Merits of Harmonic Mean

1. **Comprehensive:** It considers all observations in a series, meaning no data point can be overlooked in its calculation.
2. **Algebraic Utility:** It supports further algebraic manipulation and analysis.
3. **Effective for Equal Goals:** It yields better results when the same endpoints are used across different means.
4. **Emphasis on Smaller Values:** It prioritizes smaller values within the series.
5. **Usability with Negatives:** It can be computed even if the series includes negative values.
6. **Normalizing Skewed Distributions:** It helps transform a skewed distribution into a more normal shape.
7. **Smoother Curve:** It produces a curve that is straighter compared to those generated by the arithmetic or geometric means.

Demerits of Harmonic Mean

1. **Complexity:** It may be challenging for individuals with a basic understanding to comprehend.
2. **Cumbersome Calculation:** Calculating it can be tedious, as it involves finding the reciprocals of the numbers.
3. **Limited Accuracy for Equal Means:** It does not provide better or more accurate results when the same means are used for different outcomes.
4. **Restricted Algebraic Manipulation:** Its algebraic treatment is not as extensive as that of the arithmetic mean.
5. **Sensitivity to Extreme Values:** It is significantly affected by the values of extreme items.
6. **Inapplicability with Zero:** It cannot be calculated if any item in the series is zero.

3.4.7 CHECK YOUR PROGRESS

Question 1: Calculate the arithmetic mean of m arks obtained by 9 students in statistics are given below.

50 35 70 50 63 35 60 35 58

Question 2: Calculate the arithmetic mean and median of the following distribution

Variate :	12	14	15	16	17	18	19
Frequency:	340	343	407	379	462	435	359

Question 3: Calculate the mode for given data.

Variate :	0-20	20-40	40-60	60-80	80-100
Frequency:	31	44	39	58	12

Question 4: Discuss which measure of central tendency (mean, median, or mode) provides the best representation of the store's average daily sales and why.

Question 5: A group of students participated in a mathematics competition, and their scores (out of 100) are as follows: 45,55,60,65,70,75,80,85,90,95

i) Use the geometric mean formula to calculate the geometric mean of the scores.

ii) Use the harmonic mean formula to calculate the harmonic mean of the scores.

Question 6: Explain the differences between mean, median, and mode. In what situations would each measure be most appropriate to use?

Question7: Calculate the average income (Use geometric mean)

No. of families	2	5	7	9	10	12	15	16	17
income per head	5000	1000	2000	6000	2500	3000	4000	4500	2000

3.5. LET US SUM UP

Measures of central tendency, including the mean, median, and mode, are statistical tools used to describe the central or typical value in a data set. The mean represents the arithmetic average, providing a comprehensive measure when the data is evenly distributed. The median, the middle value when data is arranged in ascending order, is especially useful when dealing with skewed data or outliers, as it is less affected by extreme values. The mode identifies the most frequent value, highlighting the most common observation in a dataset. Together, these measures offer a clear understanding of the data's general trend, aiding in data interpretation and decision-making.

3.6. KEY POINTS/GLOSSARY

Mean: The average of a set of values, calculated by summing all values and dividing by the number of values.

Media: The middle value of a dataset when it is ordered from least to greatest

Calculation: i) If N is odd, the median is the middle number.

ii) If N is even, the median is the average of the two middle numbers.

Mode: The value that occurs most frequently in a dataset. A dataset can have one mode (unimodal), more than one mode (bimodal or multimodal), or no mode at all. Useful for categorical data to identify the most common category.

Central Tendency: A statistical measure that identifies a single value as representative of an entire dataset. Measures: Includes mean, median, and mode.

3.7. SELF-ASSESSMENT QUESTIONS

Question 1: How would you use the mean to calculate the average monthly expenses of your household?

Question 2: In a classroom, if you were to determine the typical score on a test, would you use the mean, median, or mode? Why?

Question 3: How can the median be more helpful than the mean when analyzing a set of salaries in a company where a few individuals earn exceptionally high wages?

Question 4: If you are analyzing the number of hours people sleep each night in a group of friends, how might the mode help you understand the most common sleeping habits?

Question 5: How would you interpret the mean of a data set in a situation where the data is heavily skewed, such as average household income in a country with extreme wealth disparities?

3.8. LESSON END EXERCISE

Question 1: Given the following set of test scores: 72, 85, 91, 78, and 88, calculate the mean, median, and mode.

Question 2: A company has recorded the number of units sold each month for the past five months: 150, 200, 250, 175, and 180. What is the median number of units sold?

Question 3: Which measure of central tendency (mean, median, or mode) would you recommend for analyzing the average income of a group where there are extreme outliers, and why?

Question 4: In a survey about the number of hours people exercise each week, the results are as follows: 3, 5, 7, 10, 10, and 12 hours. What is the mode and what does it tell you about the survey respondents' exercise habits?

3.9. SUGGESTED READINGS

1. Gupta, S. C., & Kapoor, V. K. (2020). Fundamentals of mathematical statistics. Sultan Chand & Sons.
2. Tufte, E. R. (2001). The Visual Display of Quantitative Information (2nd ed.) (p.178). Cheshire, CT: Graphics Press.
3. Goon, A.M., Gupta, M.K. and Dasgupta, B. (1968). Fundamental of Statistics, Vol I, World Press, Kolkata.

4. Goon, A.M., Gupta, M.K. and Dasgupta, B. (1968). Fundamental of Statistics, Vol II, World Press, Kolkata.

5. Gupta, S.C. and Kapoor, V.K. (2020). Fundamentals of Mathematical Statistics, 12th Ed., Sultan Chand and Sons.

6. Moore, D.S. (2009). The Basic Practice of Statistics. 5th Ed., W H Freeman.

LESSON : 4

EXPLORING MEASURES OF DISPERSION: UNDERSTANDING DATA VARIABILITY

Structure

4.1 INTRODUCTION

Partition and Dispersion are important statistical concepts that help describe the distribution and spread of data. Partition refers to dividing a data set into smaller, more manageable intervals or segments, such as quartiles, percentiles, and deciles. Quartiles divide the data into four equal parts, with the first quartile (Q1) representing the 25^{th} percentile, the second quartile (Q2) as the median (50^{th} percentile), and the third quartile (Q3) representing the 75^{th} percentile. These measures help identify the range within which most data points lie, offering insight into the concentration and spread of values. On the other hand, dispersion measures how spread out or varied the data is. The range is the simplest measure of dispersion, representing the difference between the highest and lowest values in the data set. More advanced measures like variance and standard deviation quantify how much the data points differ from the mean. The variance calculates the average of squared deviations from the mean, while the standard deviation, which is the square root of variance, is expressed in the same unit as the data and is easier to interpret. High variance or standard deviation indicates a wide spread of data, while low values suggest that the data points are closely grouped around the central value. Together, partition and dispersion provide a more comprehensive understanding of data by revealing how data points are distributed and how much variability exists within the data set.

4.2. LEARNING OBJECTIVES

After reding this lesson, student will be able to:

- distinguish between absolute and the relative measures of dispersion
- apply the various measures of dispersion and
- Learners will be able to define dispersion and recognize its role in describing the spread or variability of data
- Students will be able to calculate the range, variance, and standard deviation for a given data set.
- calculate and compare the different measures of central tendency and measures of dispersion

4.3 PARTITION VALUES

The values which divide the series into a certain number of equal parts, when the data is arranged in ascending or descending order. Some commonly used partition values are quartiles which divide the dataset into 4 equal parts, deciles in 10 equal parts and percentiles divide the dataset in 100 equal parts.

4.3.1. QUARTILES

Quartiles are the typical values of the variable, which divides the whole data set into four equal parts. There will be 3 quartiles, named as first, second and third quartile and are denoted as Q_1, Q_2, and Q_3,

respectively.

25%	25%	25%	25%
Q_1	Q_2	Q_3	

There are 25% of the observations below Q_1 and 75% of the observations are above the first quartile. Q_3 is the value which has 75% observations below and 25% observations are greater than the value of Q_3. There will be 50% below and above the value of Q_2, i.e., it is nothing but median. Q_0 and Q_4 will be minimum and maximum values in the dataset.

The quartiles can be calculated as follows:

From Simple Series of observations

Arrange the observations in ascending order and calculate $\frac{iN}{4}$. The value at $\frac{iN}{4}$ position will the i^{th} quartile.

i^{th} Quartile $(Q_i) = \left(\frac{iN}{4}\right)^{th}$ value, $i = 1,2,3$

From Simple frequency distribution

Obtain cumulative frequencies (less than type)

Qi = The value corresponding to the c.f. just greater than $\frac{iN}{4}$, $i = 1,2,3$

From Grouped frequency distribution

When a grouped frequency distribution is given, first ensure that class intervals are of exclusive type. If not, convert them into exclusive type class intervals. Calculate cumulative frequencies (less than type). First find the i^{th} quartile Class. The class interval having c.f. just greater than $\frac{iN}{4}$ is the i^{th} quartile class. Apply the following formula.

$$Q_i = l_i + \frac{\left(\frac{iN}{4} - cf_i\right)}{f_i} \times h_i,\ i = 1,2,3$$

l_i= lower boundary of quartile class

f_i= frequency of quartile class

h_i= height of quartile class

cf_i= cumulative frequency of the class preceding i^{th} quartile class

4.3.2. DECILES

Deciles are another partition value which divide the whole dataset into 10 equal parts. 9 points will be required for such partition, hence, there are 9 values of decile. The i^{th} decile is denoted as D_i which has $(i \times 10)\%$ of the observations below $D_i, i = 1,2, \dots, 9$.

10%	10%	10%	10%	10%	10%	10%	10%	10%
D_1	D_2	D_3	D_4	D_5	D_6	D_7	D_8	D_9

The deciles can be calculated as follows:

From Simple Series of observations

Arrange the observations in ascending order and calculate $\frac{iN}{10}$. The value at $\frac{iN}{10}th$ position will the i^{th} decile.

i^{th} Decile $(D_i) = \left(\frac{iN}{10}\right)^{th}$ value, $i = 1,2, \dots 9$

From Simple frequency distribution

Obtain cumulative frequencies (less than type)

$D_i =$ The value corresponding to the c.f. just greater than $\frac{iN}{10}$, $i = 1,2, \dots, 9$

From Grouped frequency distribution

When a grouped frequency distribution is given, first ensure that class intervals are of exclusive type. If not, convert them into exclusive type class intervals. Calculate cumulative frequencies (less than type).

Find the i^{th} decile Class. The class interval having c.f. just greater than $\frac{iN}{10}$ will be the i^{th} decile class. Then apply the following formula:

$$D_i = l_i + \frac{\left(\frac{iN}{10} - cf_i\right)}{f_i} \times h_i, \quad i = 1,2, \dots 9$$

l_i= lower boundary of i^{th} decile class

f_i= frequency of i^{th} decile class

h_i= height of i^{th} decile class

cf_i= cumulative frequency of the class preceding i^{th} decile class

4.3.3. PERCENTILES

Percentiles divide the dataset into 100 equal parts. The i^{th} percentile can be calculated as follows:

From Simple Series of observations

Arrange the observations in ascending order

i^{th} Percentile $(P_i) = \left(\frac{iN}{100}\right)^{th}$ value, $i = 1,2, \ldots 99$

From Simple frequency distribution

Obtain cumulative frequencies (less than type)

$P_i =$ The value corresponding to the c.f. just greater than $\frac{iN}{100}$, $i = 1,2, \ldots, 99$

From Grouped frequency distribution

When a grouped frequency distribution is given, first ensure that class intervals are of exclusive type. If not, convert them into exclusive type class intervals. Calculate cumulative frequencies (less than type).

Find the i^{th} percentile class. The class interval having c.f. just greater than $\frac{iN}{100}$ will be the percentile class. Then apply the following formula:

$$P_i = l_i + \frac{\left(\frac{iN}{100} - cf_i\right)}{f_i} \times h_i, \quad i = 1,2, \ldots 99$$

l_i= lower boundary of percentile class

f_i= frequency of percentile class

h_i= height of percentile class

cf_i= cumulative frequency of the class preceding i^{th} percentile class

Remark: Some texts also suggest using $\frac{i(N+1)}{k}$ instead of $\frac{iN}{k}$, where k is the number of partitions. k takes value 4, 10 or 100 for quartiles, deciles and percentiles, respectively.

Graphical Method for finding partition values:

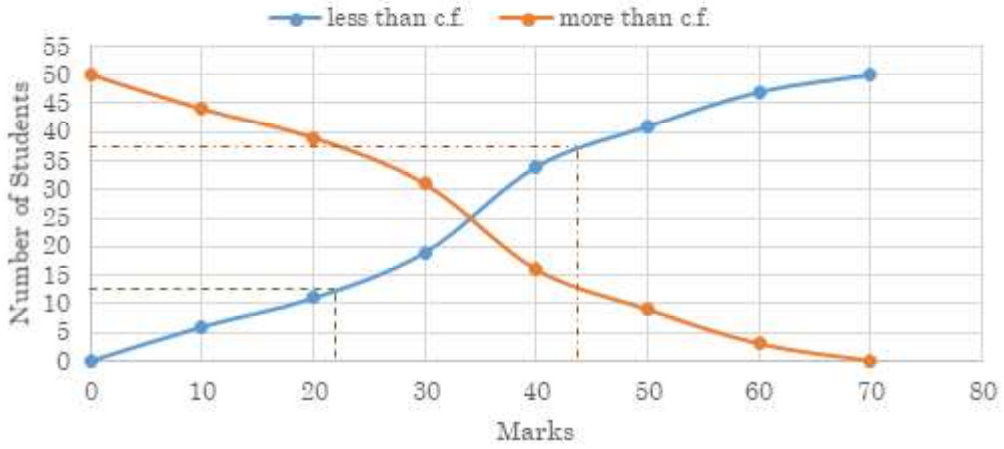
Partition values can be located graphically using ogives (less than or more than cumulative frequency curves).

Since first quartile divides the observations such that 25% ($\frac{N}{4}$ th) of the observations are less than that value, hence find the abscissa of $\frac{N}{4}$ for the less than cumulative frequency curve will give the first quartile (Q_1). Some examples of other partition values are given below:

Partition value	Abscissa of
1st Quartile	N/4 for less than cumulative frequency curve, or 3N/4 for more than cumulative frequency curve
3rd Quartile	3N/4 for less than cumulative frequency curve, or N/4 for more than cumulative frequency curve
4th Decile	4N/10 for less than cumulative frequency curve, or 6N/10 for more than cumulative frequency curve
k^{th} Percentile	$\frac{kN}{100}$ for less than cumulative frequency curve

Example 13: Find quartiles, 4th decile and 95th percentile of the marks of students given in the following table.

Marks	0-10	20-30	30-40	40-50	50-60	60-70
No. of Students	5	8	15	7	6	3



First quartile:

$\frac{N}{4} = \frac{50}{4} = 12.5$, the corresponding abscissa on less than curve is 22, $Q_1 = 22$

That means, 25% of the students obtained less than 22 marks in the examination.

Second quartile:

$\frac{N}{2} = \frac{50}{2} = 25$, the corresponding abscissa on less than curve is 34, $Q_2 = 34$.

That means, 50% of the students scores less of equal to 34 marks.

***Third quartile*:**

$\frac{3N}{4} = 3 \times \frac{50}{4} = 37.5$, the corresponding abscissa on less than curve is 44, $Q_3 = 44$.

This means that 25% of the students secured more or equal to 44 marks.

4th Decile:

$\frac{4N}{10} = \frac{4\times50}{10} = 20$, the corresponding abscissa on less than curve is 31, $D_4 = 31$.

This value means that there are 40% students who secured less or equal to 31 marks in the examination.

95th percentile:

$\frac{95N}{100} = \frac{95\times50}{100} = 47.5$, the corresponding abscissa on less than curve is 60.5, $P_{95} = 60.5$.

This implies that only 5% of the students obtained 60.5 marks or higher in the examination. In other words, we can say that this is the minimum marks of the top 5% students in that examination.

4.3.4. CHECK YOUR PROGRESS

$\frac{95N}{100} = \frac{95\times50}{100} = 47.5$, the corresponding abscissa on less than curve is 60.5,

1. The _____ divides a data set into four equal parts, each representing 25% of the data.
2. The first quartile (Q1) is also known as the _____, which separates the lowest 25% of the data from the rest.
3. The _____ is the value that separates the lowest 10% of the data from the rest of the data.
4. The median of the data set corresponds to the _____, which is the second quartile (Q2).
5. In a data set, the _____ is the value that separates the top 10% from the rest of the data.
6. The _____ is the average of the first quartile and the third quartile, and it represents the range of the middle 50% of the data.

4.4. DISPERSION

Dispersion refers to the degree to which data points in a dataset are spread out or vary from the central value, typically the mean. It provides insights into the consistency and variability of the data, highlighting how much the individual values differ from each other.

Understanding dispersion is crucial because it affects the interpretation of data. For example, a dataset with low dispersion indicates that the values are closely clustered around the mean, suggesting uniformity. Conversely, high dispersion signifies greater variability, which can imply a wider range of outcomes or greater uncertainty.

In practical terms, dispersion helps in making informed decisions, comparing different datasets, and assessing the reliability of statistical analyses. It plays a vital role in various fields, including finance, quality control, and research, guiding analysts in understanding the behavior of data.

4.5. MEASURES OF DISPERSION

The measure of central tendency gives a single value to describe the whole data set. However, it may not be sufficient to describe the nature of the dataset. For example, consider the following data in which gives the scores of three students in a series of tests.

Student 1:	440	460	480	500	520	540	560
Student 2:	350	400	450	500	550	600	650
Student 3:	260	340	420	500	580	660	740

From these observations, it can be seen that the average marks (A.M.) of all three students is 500. So, can we conclude that the performance of all three students is same in the series of tests? The answer will be no. When we look at the observations carefully, we see that student 1 scored more closely to 500 marks in all the tests as compared to student 2 and 3. Whereas, student 3 performed very poor in some tests while very good in some tests. Or in other words, we can say that the scores for student 1 are closely knitted, i.e., less scatteredness, for student 2, the scatteredness of the scores is little more as compared to student 1 and the scattered in scores of student 3 is maximum among the three students. It is quite clear that averages (Or measures of central tendency) only provide insight into one aspect of the distribution of observations: the central location of the data. Further, to describe the distribution of observations on a variable, it is necessary to have some numeric quantity which gives a good measure of scatteredness (also called dispersion, variability).

The Yule's criteria for an ideal measure of central tendency can be considered for measures of dispersion as well. The criteria are as follows:

1. It should be rigidly defined.
2. It should be readily comprehensible and easy to calculate.
3. It should be based upon all the observations.
4. It should be suitable for further mathematical treatment.
5. It should be affected as little as possible by fluctuations of sampling.
6. It should not be affected much by extreme values

In descriptive analysis, there could be a need to have either absolute measures of dispersion or relative measures of dispersion. An absolute measure is measured in terms of the original units of the observations, whereas, a relative measure does not have any unit. The absolute measures of dispersion are not suitable for comparative study of characteristics of two or more series.

4.5.1 ABSOLUTE MEASURES OF DISPERSION

There are four measures of dispersion

i. Range

ii. Quartile Deviation

iii. Mean Deviation

iv. Standard Deviation

4.5.1.i. RANGE

It is a distance measure. The range of a series of observations can be simply calculated by taking the difference between the maximum and minimum values in the dataset.

Range = Max. Value –Min. Value

It is clear that the range tells that how far the lowest and highest values in the dataset are.

The range is a crude measure of dispersion.

Advantages

- It is easy to calculate and easy to understand.

Disadvantages

- It does not account all observations in the dataset
- It is highly affected by extreme values.
- It is affected by sampling fluctuations.

Example 14: Find the range and coefficient of range of the weights of 10 students from the following data:

50 60 40 30 50 80 100 20 10

Solution: Arrange the data in ascending order, we get

10 20 30 40 50 50 60 80 100

Largest value (L) = 100 smallest value(S) = 10

Range = L-S = 100-10

Coefficient of Range= $\frac{L-S}{L+S}$ $\frac{100-10}{100+10} = \frac{90}{110} = 0.9$

Example 15: Find the range and the coefficient of rage of marks of 50 students.

Marks	0-100	10-20	20-30	30-40	40-50
No. of Students	13	15	5	10	7

Marks	No. of Students
0-10	13
10-20	15
20-30	5
30-40	10
40-50	7

Largest L = 50 and Smallest S= 0

Range =L-S = 0

Coefficient of Range = $\frac{L-S}{L+S}\frac{50-0}{50+0} = \frac{50}{50} = 1$

4.5.1. ii. QUARTILE DEVIATION

Quartile deviation (Q.D.) is another distance measure of dispersion. It is calculated as follows:

$$Q.D. = \frac{(Q_3 - Q_1)}{2}$$

Where Q_1 and Q_3 are first and third quartiles. The difference between first and third quartile, $Q_3 - Q_1$ is known as interquartile range; and Q.D. is also known as semi-interquartile range. Quartile deviation is useful especially when the distribution features open-ended class intervals (such as less than or more than types). In such case, other measures cannot be calculated.

Coefficient of Quartile Deviation given by

$$= \frac{(Q_3 - Q_1)}{(Q_3 + Q_1)}$$

Advantages

- It is easy to calculate and easy to understand.

Disadvantages

- It uses only the middle 50% of the observations.
- As it only uses the middle 50% of the observations, it is not affected by extreme values.

However, to overcome the limitation of quartile deviation, that it only uses middle 50% of the observations

some other measures, mean deviation and standard deviation, are given. These measures give the average distance of all the observations from an average value.

Example 16: Given the following data set 5,8,12,15,20,22,25,30

Calculate the first quartile (Q_1), third quartile (Q_3), and then determine the quartile deviation.

Solution: Firstly, calculate Q_1 and Q_3, for that, arrange the data in order of magnitude

5, 8,12,15, 20, 22, 25, 30

N=8 (even number of observations)

$Q_2 = \frac{1}{2}[15 + 20] = 17.5$

$Q_1 = \frac{8+12}{2} = 10$

$Q3 = \frac{22+25}{2} = 23.5$

Quartile Deviation: (Q.D.) =

Example 17: The numbers of pieces of junk mail received by 10 families during the past month.

41, 33, 28, 21, 29, 19, 14, 31, 39, 36

Find the range and Q.D.

Solution:

Minimum value in the data set = 14

Maximum value in the data set = 41

Range = Max. Value- Min. Value = 41–14 =33

Quartile Deviation (Q.D.) = $\frac{(Q_3 - Q_1)}{2}$

Calculation of Quartile Deviation

First calculate the first and third quartile. For that, arrange the data in order of magnitude

Position	1	2	3	4	5	6	7	8	9	10
Values	14	19	21	28	29	31	33	36	39	41

$\frac{N}{4} = \frac{10}{4} = 2.5$, and $\frac{3N}{4} = \frac{3\times10}{4} = 7.5$

will be the value at 2.5th position. We can take A.M. of the values 19 and 21 at 2nd and 3rd positions, respectively.

$$Q_1 = \frac{19 + 21}{2} = 20$$

Similarly, $Q_3 = \frac{33 + 36}{2} = 34.5$

The inter-quartile range $Q_3 - Q_1 = 34.5 - 20 = 14.5$ and the

$$Q.D. = \frac{34.5-20}{2} = 7.25.$$

Example 18: Given the following data set of test scores:

55,60,65,70,75,80,85,90,95,

Calculate the first quartile (Q1) and the third quartile (Q3) and determine the quantile deviation.

Solution: Calculate Q_1 and Q_3

Step 1: Arrange the data in ascending order:

The data is already arranged:

55,60,65,70,75,80,85,90, 95

Step 2: Find Q1:

The position of Q1 is given by $\frac{N+1}{2}$, where N is the number of data points.
Here, n = 9, So

$$Q_1 = \frac{9 + 1}{4} = 2.5$$

Q_1 is the average of the 2nd and 3rd values

$$Q_1 = \frac{60 + 65}{2} = 62.5$$

Step 3: Find Q3:

The position of Q3 is given by $\frac{3N+1}{4}$,

$$Q_3 = \frac{3(9 + 1)}{4} = 7.5$$

Q_3 is the average of the 7th and 8th values

$$Q_3 = \frac{85 + 90}{2} = 87.5$$

Quartile Deviation: (Q.D.) = $\frac{(Q_3 - Q_1)}{2}$

Quartile Deviation: (Q.D.) = $\frac{(87.5+62.5)}{2} = 12.5$

The quartile deviation of 12.5 indicates that, on average, the test scores in the middle 50% of the data are spread out by 12.5 points around the median. This reflects the variability in test performance among the students. A larger quartile deviation would suggest more variability in scores, while a smaller one would indicate more consistency in test performances.

4.5.1.iii. MEAN DEVIATION

It finds the absolute distance of all the observations from an average value (mean, median or mode) and then the average of those absolute distances gives the mean deviation about the respective average. The formula is given as follows:

Let $x_1, x_2, \ldots, x_N$ be *N* observations on a variable

$$\text{M.D. about A} = \frac{1}{N}\sum_{i=1}^{N}|x_i - A|$$

When $(x_i, f_i),\ i = 1,2,\ldots,n$ be a frequency distribution,

$$\text{M.D. about A} = \frac{1}{n}\sum_{i=1}^{n} f_i|x_i - A|$$

Where is an average (Mean, Median or Mode) and is the number of distinct values/classes. When the frequency distribution is grouped, the mid-value of the i^{th} class interval will be x_i.

Advantages

- It is easy to calculate and easy to understand.
- It is based upon all observations.

Disadvantages

- Taking the absolute of the differences creates artificiality.
- It is not appropriate for additional mathematical analysis. For e.g., given mean deviation of two groups of observations, we cannot determine the mean deviation of the combined observations.
- It is affected by extreme values and sampling fluctuations.

Example 19: Given the following data set of values:

15,22,28,35,40,45,50,55,60,65,70,75

Calculate the mean of the data set.

Compute the mean deviation about the mean.

Solution: Calculate the Mean:

Step 1: Sum of the data points:

15+22+28+35+40+45+50+55+60+65+70+75=550

Step 2: Number of data points (n):

n=12

Step 3: Mean Deviation (μ)

$$\text{Mean} = \frac{550}{12} = 45.83$$

Step 4: $\sum_{i=1}^{N} |x_i - mean| =$

30.83+23.83+17.83+10.83+5.83+0.83+4.17+9.17+14.17+19.17+24.17+29.17=191.17

Step 5: Mean Deviation (μ) $= \frac{191.17}{12} = 15.93$

The mean deviation of approximately 15.93 indicates that, on average, the values in the data set deviate by about 15.93 units from the mean. This reflects the variability in the data around the average value.

Example 20: Consider the following frequency distribution of students' scores in a test:

Class Interval	0-10	10-20	20-30	30-40	40-50	50-60	60-70
Frequency	8	12	10	8	3	2	7

i) Determine the mean score.

ii) Calculate the mean deviation from the mean.

Solution:

Class Interval	**Frequency**	**Mid-Point**	**d=X-A**	**f.d**	**X-29**	**$\lvert X-\bar{X}\rvert$**	**$f\lvert X-\bar{X}\rvert$**
0-10	8	5	-30	-240	-24	24	192
10-20	12	15	-20	-240	-14	14	168
20-30	10	25	-10	-100	-4	4	40
30-40	8	35	0	0	6	6	48
40-50	3	45	10	30	16	16	48
50-60	2	55	20	40	26	26	52
60-70	7	65	30	210	36	36	252
	N= 50			$\sum fd = -300$			$\sum f\lvert X-\bar{X}\rvert$=800

Let A=35 be the assumed mean

Mean $\bar{X} = A + \frac{\sum fd}{N} = 35 + \frac{-300}{50} = 29$

Mean Deviation about Mean = $\frac{1}{N}\sum_{i=1}^{n} f_i|X-\bar{X}| = \frac{800}{50} = 16$

Coefficient of Mean Deviation: The relative measure of dispersion associated with the mean deviation is called the coefficient of mean deviation. It is given by following formula:

Coefficient of Mean Deviation (C.V)=$\frac{MD}{\bar{x}}$, where $\bar{x}$ is the arithmetic mean.

$\frac{MD}{M_d}$ and $\frac{MD}{M_o}$, mean deviation about median and mode respectively.

Example 21: Calculate the mean deviation from the mean for the following data:

Class Interval	0-4	4-8	8-12	12-16	16-20
Mid-Values	4	6	8	5	2

Solution:

Class Interval	**Mid Value**	**Frequency**	**fx**	**d=$\lvert X-\bar{X}\rvert$**	**fd**
0-4	2	4	8	7.2	28.8
4-8	6	6	36	3.2	19.2
8-12	10	8	80	0.8	6.4
12-16	14	5	70	4.8	24.0
16-20	18	2	36	8.8	17.6
		N=25	$\sum fx = 230$		$\sum fd = 96.0$

Now, Arithmetic Mean : $\frac{230}{25} = 9.2$

Mean Deviation (MD): $\frac{96}{25} = 3.84$

4.5.1.iv. STANDARD DEVIATION

It is the most commonly used measure of dispersion. It represents the positive square root of the average (A.M.) of the squared distances of all observations from their arithmetic mean.

Let $x_1, x_2, \ldots, x_N$ be N observations on a variable

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (x_i - \underline{x})^2}$$

When (x_i, f_i), $i = 1,2, \ldots, n$ be a frequency distribution

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

Where, represents the number of distinct values or classes. In a grouped frequency distribution, the mid-value of the i^{th} class interval is denoted as x_i.

Advantages

- It is based upon all observations.

Disadvantages

- Taking square root makes it a little difficult to calculate.
- It is suitable for further mathematical treatment.
- It is affected by extreme values and sampling fluctuations.
- It cannot be calculated for open-end classes.

The square of standard deviation is known as variance and denoted by σ^2.

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

Properties of Variance:

1. Variance is unaffected by changes in the origin but is influenced by changes in scale.
2. A zero variance indicates that the observations are constant.

Example 22: The following table gives the number of errors made in a baseball game.

No. of Errors	No. of Games
0	11
1	14
2	9
3	7
4	3
5	1
Total	45

Calculate Q.D., mean deviation and standard deviation for number of errors made.

Solution

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
No. of Errors (X)	**No. of Games (f)**	$f_i\, x_i$	**c.f.**	$\lvert x_i - \bar{x}\rvert$	$f_i \lvert x_i - \bar{x}\rvert$	$f_i \lvert x_i - median\rvert$	$f_i \lvert x_i - mode\rvert$	$f_i(x_i - \bar{x})2$
0	11	0	11	1.56	17.16	11	11	26.7696
1	14	14	25	0.56	7.84	0	0	4.3904
2	9	18	34	0.44	3.96	9	9	1.7424
3	7	21	41	1.44	10.08	14	14	14.5152
4	3	12	44	2.44	7.32	9	9	17.8608
5	1	5	45	3.44	3.44	4	4	11.8336
Total	**45**	**70**			**49.8**	**47**	**47**	**77.112**

$Mean\ (\bar{x}) = \frac{1}{N}\sum_{i=1}^{n} \quad f_i x_i = \frac{70}{45} = 1.56$

The maximum frequency in this distribution is 14 and the corresponding value is 1, hence Mode = 1

First Quartile: N/4 = 45/4 = 11.25

The cumulative frequency that exceeds 11.25 is 25, and the corresponding value of *X* is 1, therefore $Q_1 = 1$

Second Quartile/Median: $\frac{N}{2} = \frac{45}{2} = 22.5$

The cumulative frequency that exceeds 22.5 is 25 and the corresponding value of *X* is 1, therefore $Q_2 = 1$

Third Quartile: 3N/4 = 3*45/4=33.75

The cumulative frequency that exceeds 33.75 is 34 and the corresponding value of *X* is 2, therefore $Q_3 = 2$

Quartile Deviation (Q.D.) = $\frac{Q_3 - Q_1}{2} = \frac{2-1}{2} = 0.5$

M.D. about mean $= \frac{1}{N}\sum_{i=1}^{n} \quad f_i|x_i - \bar{x}| = \frac{49.8}{45} = 1.11$

M.D. about median $= \frac{1}{N}\sum_{i=1}^{n} \quad f_i|x_i - median| = \frac{47}{45} = 1.04$

M.D. about mode $= \frac{1}{N}\sum_{i=1}^{n} \quad f_i|x_i - mode| = \frac{47}{45} = 1.04$

Variance $(\sigma^2) = \frac{1}{N}\sum_{i=1}^{n} \quad f_i(x_i - \bar{x})^2 = \frac{77.112}{45} = 1.71$

S.D. $(\sigma) = \sqrt{1.71} = 1.31$

Example 23: Find quartile deviation, mean deviation about mean, median and mode and standard deviation of the marks of students given in the following table.

Marks	No. of Students
0-10	6
10-20	5
20-30	8
30-40	15
40-50	7
50-60	6
60-70	3
Total	50

Calculate Standard Deviation.

Solution:

Marks	**No. of Students**	**Mid-Value (x)**	$\boldsymbol{d = \frac{x - 35}{10}}$	$\boldsymbol{fd}$	$\boldsymbol{f_i d_i^2}$
0-10	6	5	-3	-18	54
10-20	5	15	-2	-10	20
20-30	8	25	-1	-8	8
30-40	15	35	0	0	0
40-50	7	45	1	7	7
50-60	6	55	2	12	24
60-70	3	65	3	9	27
Total	**50**	**245**		**-8**	**140**

$\bar{d} = -0.16$

$\bar{x} = 35 + 10 \times (-0.16) = 33.4$

Variance $(\sigma_d^2) = \frac{1}{N}\sum_{i=1}^{n} \quad f_i d_i^2 - \bar{d}^2 = \frac{140}{50} - (-0.16)^2 = 2.7744$

$\sigma_X^2 = h^2\sigma_d^2 = 10^2 \times 2.7744 = 277.44$

Standard Deviation $(\sigma) = 16.66$

Example 24: Consider the following data set representing the test scores of 15 students:

78,85,92,88,76,95,89,84,91,80,87,90,82,86, 79

i) Calculate the mean of the data set.

i) Calculate the variance.

ii) Calculate the standard deviation.

Solution:

The mean is calculated using the formula:

i) $\bar{x} = \frac{\sum x_i}{n} = \frac{78+85+92+88+76+95+89+84+91+80+87+90+82+86+79}{15}$

$= \frac{1305}{15} = 87$

- The variance is calculated using the formula:

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

firstly, calculate $\sum(x_i - \bar{x})^2 = 81 + 4 + 25 + 1 + 121 + 64 + 4 + 9 + 16 + 49 + 0 + 9 + 25 + 1 + 64 = 403$

Variance $(\sigma^2) = \frac{403}{15} = 26.87$

- The standard deviation is the square root of the variance:

Standard Deviation $(\sigma) = \sqrt{\sigma^2} = \sqrt{26.81} = 5.19$

4.5.2 RELATIVE MEASURES OF DISPERSION

Absolute measures of dispersion can be inadequate for comparing variability between two groups in the following situations:

a) When the units of measurement differ between the two groups.

b) When there is a significant difference between the average values of the two groups.

For example, if the variances for height (in centimeters) and weight (in kilograms) are found to be 20 cm² and 15 kg², respectively, it is clear that these values cannot be directly compared, as centimeters and kilograms are incompatible units.

	Daily expenses (in Rs.)							Mean	S.D.
Family A	450	600	380	450	650	470	240	405.71	138.58
Family B	4500	5200	4000	4800	4000	4500	3700	4535.71	1031.22

When we examine the standard deviation of the daily expenses for families A and B, we find that family B exhibits significantly higher variability compared to family A. However, upon reviewing the mean expenses of both groups, we observe a substantial difference in their average spending. An analysis of the following figures, which display the daily expenses with a red marker indicating the mean expense, reveals that family A actually shows slightly more variability in their daily expenses than family B.

In this example, even though the when the units of measurement are consistent, absolute measures are not the most suitable choice for comparing variability. Relative measures of dispersion provide insights

into variability in relation to the average value. Based on different absolute measures of dispersion, we can derive several relative measures of dispersion.

1. Coefficient of dispersion based on Range $= \frac{Max.Value - Min.Value}{Max.Value + Min.Valu}$
2. Coefficient of dispersion based on Quartile Deviation $= \frac{(Q_3 - Q_1)/2}{(Q_3 + Q_1)/2} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$
3. Coefficient of dispersion based on Mean Deviation $= \frac{Mean\ Deviation\ about\ an\ Average}{Average}$
4. Coefficient of dispersion based on Standard Deviation $= \frac{\sigma}{\bar{x}}$

Standard deviation is the most widely accepted measure of dispersion, and when applicable, the coefficient of dispersion based on standard deviation is generally favoured over other options. For easier interpretation, this measure has been adapted and is commonly referred to as the coefficient of variation (C.V.).

	Daily expenses (in Rs.)							Mean	S.D.	C.V
Family A	450	600	380	450	650	470	240	405.71	138.58	34.15
Family B	4500	5200	4000	4800	4000	4500	3700	4535.71	1031.22	22.7

$C.V._A = 34.15$ tells that there is 34.15% variability in the daily expenses with respect to the average expenditure of the family A.

$C.V._B = 22.7$ tells that there is 22.7% variability in the daily expenses relative to the average expenditure of the family B.

Family B has a smaller coefficient of variation, indicating that, although they spend significantly more on average than Family A, their daily expenses show less variability (or greater consistency) compared to Family A.

4.6. CHECK YOUR PROGRESS

Question 1: Explain the concepts of range, quartile deviation, standard deviation, and mean deviation. How does each measure provide insights into the spread of a data set?

Question 2: In what situations might the range be a useful measure of dispersion? Provide an example to illustrate your point.

Question 3: Below are the test scores of a group of students:

Student	1	2	3	4	5	6	7
Score	4	7	9	5	9	2	6

i) Determine the range of the scores.

ii) Find the and Q_1 and Q_3, and then calculate the quartile deviation.

iii) Compute the standard deviation of the scores, showing all calculations.

iv) Find the mean deviation of the scores.

Question 4: The following table shows the exam scores of students along with their frequencies:

Score	50	60	70	80	90
Frequency	7	8	4	8	6

i) Compute the standard deviation of the scores, showing all calculations.

ii) Find the mean deviation of the scores.

Question 5: The following table shows the height (in cm) of plants in a garden, grouped into ranges:

Height Range	10-20	20-30	30-40	40-50	50-60	60-70	70-80
Frequency	14	17	19	15	19	12	16

Compute the standard deviation and Variance.

Question 6: The following table represents the monthly expenditure (in $) of families in a neighbourhood:

Expenditure Range	1000-2000	2000-3000	3000-4000	4000-5000	5000-6000	6000-7000	7000-8000
Frequency	17	19	15	15	19	20	16

Compute the standard deviation and Variance.

4.7. LET US SUM UP

Measures of dispersion, such as range, variance, and standard deviation, provide insights into the spread or variability of data within a dataset. The range represents the difference between the highest and lowest values, giving a quick sense of the extent of variation. Variance measures the average squared deviation from the mean, offering a deeper understanding of how individual data points differ from the mean. Standard deviation, the square root of variance, expresses the spread in the same units as the original data, making it easier to interpret. Together, these measures help assess the consistency of data and are essential for understanding the variability and reliability of statistical analysis.

4.8. KEY POINTS/GLOSSARY

1. **Variability:** The extent to which data points in a dataset differ from each other**.**
   - **i) Range:** The difference between the maximum and minimum values**.**
   - **ii) Variance:** The average of the squared differences from the mean.
   - **iii) Standard Deviation:** The square root of the variance.
2. **Partition Values:** Values that divide a dataset into equal parts or intervals. Help in understanding the distribution of data and identifying relative standings within a dataset.
3. **Quartiles:** Specific types of partition values that divide a dataset into four equal parts.

   First Quartile (Q1): The value below which 25% of the data falls.

   Second Quartile (Q2): The median, dividing the dataset into two halves (50%).

   Third Quartile (Q3): The value below which 75% of the data falls.

4.9. SELF-ASSESSMENT QUESTIONS

Question 1: If you are comparing the salaries of employees at a company, how would the standard deviation help you understand salary variation across different departments?

Question 2: When assessing the consistency of your weekly grocery spending, why might you calculate the range or variance rather than just looking at the total amount spent?

Question 3: If you track the daily temperatures in your city over a month, how would the range of temperatures help you understand the overall weather variation for the month?

Question 4: You're comparing the daily number of visitors to a website over the last week. How would you use the standard deviation to understand whether the visitor numbers are consistent or vary widely from day to day?

Question 5: In a study about daily commute times for employees, how could a high variance in the commute times indicate potential problems with transportation or infrastructure in the area?

4.10. LESSON END EXERCISE

Question 1: Given the following set of ages: 22, 25, 30, 35, 40, calculate the range and explain what this measure tells you about the spread of ages in the group.

Question 2: In a set of data for monthly rainfall (in cm): 10, 12, 8, 15, 20, 25, calculate the variance and standard deviation. What do these values reveal about the variability in rainfall over the month?

Question 3: Consider the following data of exam scores: 70, 75, 80, 85, 90. Calculate the range, variance, and standard deviation. How would you interpret these measures in terms of the consistency of student performance?

Question 4: A local company records the weekly sales figures (in thousands): 100, 120, 110, 140, 160. Calculate the range and standard deviation, and describe what these figures tell you about sales performance variability.

Question 5: In a set of five test scores: 65, 75, 85, 90, 95, calculate the variance and standard deviation. How would a lower or higher standard deviation change your understanding of these scores in terms of performance consistency?

4.11. SUGGESTED READINGS

1. Gupta, S. C., & Kapoor, V. K. (2020). Fundamentals of mathematical statistics, Sultan Chand & Sons.
1. Tufte, E. R. (2001). The Visual Display of Quantitative Information (2nd ed.) (p.178). Cheshire, CT: Graphics Press.
2. Goon, A.M., Gupta, M.K. and Dasgupta, B. (1968). Fundamental of Statistics, Vol I, World Press, Kolkata.
3. Goon, A.M., Gupta, M.K. and Dasgupta, B. (1968). Fundamental of Statistics, Vol II, World Press, Kolkata.
4. Gupta, S.C. and Kapoor, V.K. (2020). Fundamentals of Mathematical Statistics, 12th Ed., Sultan Chand and Sons.
5. Moore, D.S. (2009). The Basic Practice of Statistics. 5th Ed., W H Freeman.

LESSON : 5

NORMAL DISTRIBUTION: KEY CONCEPTS, APPLICATIONS, AND THE FOUNDATION OF STATISTICAL INFERENCE

Structure

5.1 INTRODUCTION

So far you have learnt in descriptive statistics, how to organize a distribution of scores and how to describe its shape, central value and variation. You have used histogram and frequency polygon to illustrate the shape of a frequency distribution, measures of central tendency to describe the central value and measures of variability to indicate its variation. All these descriptions have gone a long way in providing information about a set of scores, but you also need procedures to describe individual scores or cutting point scores to categorize the entire group of individuals on the basis of their ability or the nature of test paper, which a psychometrician or teacher has used to assess the outcomes of the individual on a certain ability test. For example, suppose a teacher has administered a test designed to appraise the level of achievement and a student has got some score on the test. What did that score mean? The obtained score has some meaning only with respect to other scores either the teacher may be interested to know how many students lie within the certain range of scores? Or how many students are above and below certain referenced score? Or how many students may be assigning A, B, C, D etc. grades according to their ability? To have an answer to such problems, the curve of Bell shape, which is known as Normal curve, and the related distribution of scores, through which the bell-shaped curve is obtained, generally known as Normal Distribution, is much helpful. Thus, the present lesson presents the concept, characteristics and use of Normal Distributions and Normal Curve, by suitable illustrations and explanations.

5.2 LEARNING OBJECTIVES

After reading this lesson, you will be able to:

- Explain the concept of normal distribution and normal probability curve;
- Draw the normal probability curve on the basis of given normal distribution;
- Explain the theoretical basis of the normal probability curve;
- Elucidate the Characteristics of the normal probability curve and normal distribution;
- Analyze the normal curve obtained on the basis of large number of observations;
- Describe the importance of normal distribution curve in mental and educational measurements;
- Explain the applications of normal curve in mental measurement and educational evaluation;
- Read the table of area under normal probability curve;
- Compare the non-normal with normal Distribution and express the causes of divergence from normalcy; and
- Explain the significance of skewness and kurtosis in the mental measurement and educational evaluation.

5.3 NORMAL DISTRIBUTION/ NORMAL PROBABILITY CURVE

Carefully look at the following hypothetical frequency distribution, which a teacher has obtained after examining 150 .students of class IX on a Mathematics achievement test.

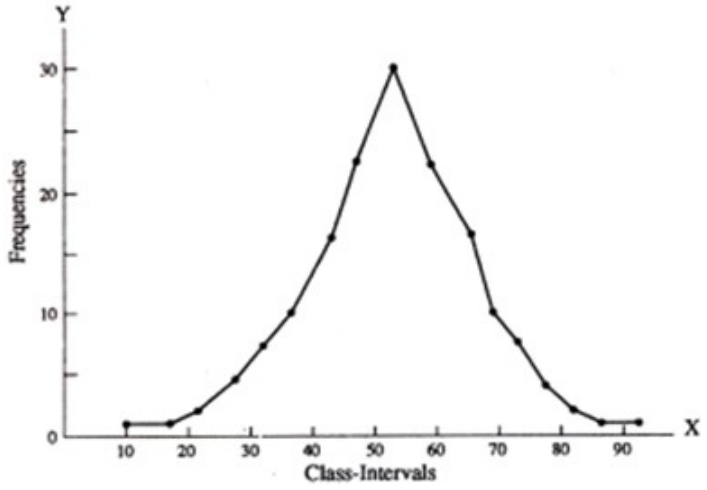
Table 5.1: Frequency distribution of the mathematics achievement test scores

Class Intervals	**Tallies**	**Frequency**
85 – 89	I	1
80 – 84	II	2
75 – 79	IIII	4
70 – 74	~~IIII~~ II	7
65 – 69	~~IIII~~ ~~IIII~~	10
60 – 64	~~IIII~~ ~~IIII~~ ~~IIII~~ I	16
55 – 59	~~IIII~~ ~~IIII~~ ~~IIII~~ ~~IIII~~	20
50 – 54	~~IIII~~ ~~IIII~~ ~~IIII~~ ~~IIII~~ ~~IIII~~ ~~IIII~~	30
45 – 49	~~IIII~~ ~~IIII~~ ~~IIII~~ ~~IIII~~	20
40 – 44	~~IIII~~ ~~IIII~~ ~~IIII~~ I	16
35 – 39	~~IIII~~ ~~IIII~~	10
30 – 34	~~IIII~~ II	7
25 – 29	IIII	4
20 – 24	II	2
15 – 19	I	1
	Total	**150**

Are you able to find some special trend in the frequencies shown in the column 3 of the above table? Probably yes! The concentration of maximum frequencies (f = 30) lies near a central value of distribution and frequencies gradually tapper off symmetrically on both the sides of this value.

Now, suppose if you draw a frequency polygon with the help of above distribution, you will have a curve as shown in the figure 5.1

Figure 5.1: Frequency Polygon of the data given in Table 5.1



The shape of the curve in Figure 5.1 is just like a 'Bell' and is symmetrical on both the sides. If you compute the values of Mean, Median and Mode, you will find that these three are approximately the same (M = 52; Md = 52 and Mo = 52).

This Bell-shaped curve technically known as Normal Probability Curve or simply Normal Curve and the corresponding frequency distribution of scores, having just the same values of all three measures of central tendency (Mean, Median and Mode) is known as Normal Distribution.

Many variables in the physical (e.g. height, weight, temperature etc.) biological (e.g. age, longevity, blood sugar level and behavioural (e.g. Intelligence; Achievement; Adjustment; Anxiety; Socio-Economic-Status etc.) sciences are normally distributed in the nature. This normal curve has a great significance in mental measurement. Hence to measure such behavioural aspects, the Normal Probability Curve in simple terms Normal Curve worked as reference curve and the unit of measurement is described as σ (Sigma).

5.4 THEORETICAL BASE OF THE NORMAL PROBABILITY CURVE

The normal probability curve is based upon the law of Probability (the various games of chance) discovered by French Mathematician Abraham Demoiver (1667-1754). In the eighteenth century, he developed its mathematical equation and graphical representation also.

The law of probability and the normal curve that illustrates it are based upon the law of chance or the probable occurrence of certain events. When anybody of observations conforms to this mathematical form, it can be represented by a bell-shaped curve with definite characteristics.

5.5 CHARACTERISTICS OR PROPERTIES OF NORMAL PROBABILITY CURVE (NPC)

The characteristics of the normal probability curve are:

1. **The Normal Curve is Symmetrical:** The normal probability curve is symmetrical around its vertical axis called ordinate. The symmetry about the ordinate at the central point of the curve implies that the size, shape and slope of the curve on one side of the curve is identical to that of the other. In other words, the left and right halves to the middle central point are mirror images, as shown in the figure 5.2.

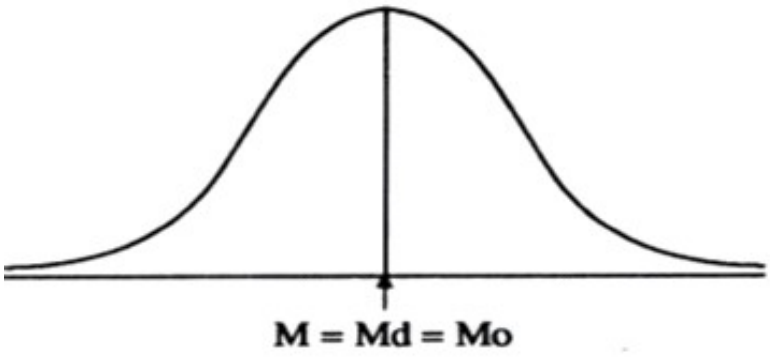


Fig. 5.2

2. **The Normal Curve is Unimodal:** Since there is only one maximum point in the curve, thus the normal probability curve is unimodal, i.e. it has only one mode.

3. **The Maximum Ordinate occurs at the Center:** The maximum height of the ordinate always occurs at the central point of the curve, that is the mid-point. In the unit normal curve, it is equal to 0.3989.

4. **The Normal Curve is Asymptotic to the X Axis:** The normal probability curve approaches the horizontal axis asymptotically; i.e. the curve continues to decrease in height on both ends away from the middle point (the maximum ordinate point); but it never touches the horizontal axis. Therefore, its ends extend from minus infinity (- ∞) to plus infinity (+ ∞).

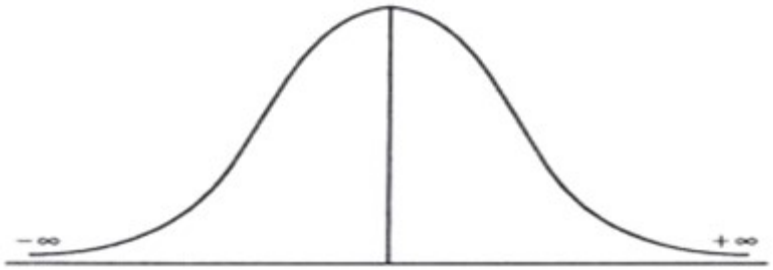


Figure 5.3

5. **The Height of the Curve declines Symmetrically:** In the normal probability curve the height declines symmetrically in either direction from the maximum point.

6. **The Points of Influx occur at point ±1 Standard Deviation (± 1σ):** The normal curve changes its direction from convex to concave at a point recognized as point of influx. If you draw the perpendiculars from these two points of influx of the curve to the horizontal X axis; touch at a distance one standard deviation unit from above and below the mean (the central point).

7. **The Total Area under Normal Curve may be also considered 100 Percent Probability:** The total area under the normal curve may be considered to approach 100 percent probability; interpreted in terms of standard deviations. The specified area under each unit of standard deviation is shown in the figure 5.4.

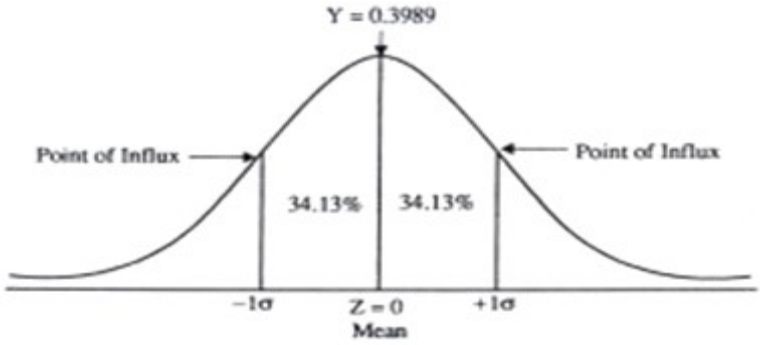


Figure 5.4

8. **The Total Area under Normal Curve may be also considered 100 Percent Probability:** The total area under the normal curve may be considered to approach 100 percent probability; interpreted in terms of standard deviations. The specified area under each unit of standard deviation is shown in the figure 5.5.

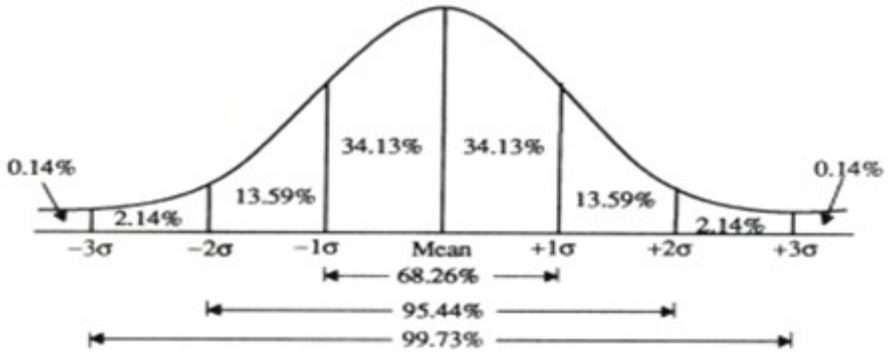


Figure 5.5: The Percentage of the Cases Failing Between Successive Standard Deviation in Normal Distribution

9. **The Normal Curve is Bilateral:** The 50% area of the curve lies to the left side of the maximum central ordinate and 50% of the area lies to the right side. Hence the curve is bilateral.

10. **The Normal Curve is a mathematical model in behavioural Sciences Specially in Mental Measurement:** This curve is used as a measurement scale. The measurement unit of this scale is ± 1σ (the unit standard deviation).

5.6 INTERPRETATION OF NORMAL CURVE/ NORMAL DISTRIBUTION

What do normal curve/ normal distribution indicate? Normal curve has great significance in the mental measurement and educational evaluation. It gives important information about the trait being measured.

If the frequency polygon of observations or measurements of certain trait is a normal curve, it is a indication that

1. The measured trait is normally distributed in the universe.
2. Most of the cases i.e. individuals are average in the measured trait and their percentage in the total population is about 68.26%.
3. Approximately 15.87% (50-34.13%) of cases are high in the trait measured.
4. Similarly, 15.87% of cases are low in the trait measured.
5. The test which is used to measure the trait is good.
6. The test which is used to measure the trait has good discrimination power as it differentiates between poor, average and high ability group individuals.
7. The items of the test used are fairly distributed in terms of difficulty level.

5.7 IMPORTANCE OF NORMAL DISTRIBUTION

The Normal distribution is by far the most used distribution in inferential statistics because of the following reasons:

1. Number of evidences are accumulated to show that normal distribution provides a good fit or describe the frequencies of occurrence of many variable facts in biological statistics, e.g. sex ratio in births, in a country over a number of years. The anthropometrical data, e.g. height, weight, etc. The social and economic data e.g. rate of births, marriages and deaths. In psychological measurements e.g. Intelligence, perception span, reaction time, adjustment, anxiety etc. In errors of observation in physics, chemistry, astronomy and other physical sciences.
2. The normal distribution is of great value in educational evaluation and educational research, when you make use of mental measurement. It may be noted that normal distribution is not an actual distribution of scores on any test of ability or academic achievement, but is, instead, a mathematical model. The distributions of test scores approach the theoretical normal distribution as a limit, but the fit is rarely ideal and perfect.

5.8 APPLICATION OF NORMAL DISTRIBUTION

There are number of applications of normal curve in the field of psychology as well as educational measurement and evaluation.

These are:

i) To determine the percentage of cases (in a normal distribution) within given limits or scores.

ii) To determine the percentage of cases that are above or below a given score or reference point.

iii) To determine the limits of scores which include a given percentage of cases to determine the percentile rank of an individual or a student in his own group.

iv) To find out the percentile value of an individual on the basis of his percentile rank.

v) Dividing a group into sub-groups according to certain ability and assigning the grades.

vi) To compare the two distributions in terms of overlapping.

vii) To determine the relative difficulty of test items.

5.9 TABLE OF AREAS UNDER THE NORMAL PROBABILITY CURVE

How do you use all the above applications of normal curve in mental as well as in educational measurement and evaluation? It is essential first to know about the Table of areas under the normal curve.

The table 5.2 gives the fractional parts of the total area under the normal curve found between the mean and ordinates erected at various σ (sigma) distances from the mean.

The normal probability curve table is generally limited to the areas under unit normal curve with N = 1, σ = 1. In case, when the values of N and ó are different from these, the measurements or scores should be converted into sigma scores (also referred to as standard scores or z scores). The process is as follows:

$$z = \frac{X - M}{\sigma} \quad or \; z = \frac{x}{\sigma}$$

Where, z = Standard Score X = Raw Score
M = Mean of X Scores σ = Standard Deviation of X Scores

The table of areas of normal probability curve are then referred to find out the proportion of area between the mean and the z value.

Though the total area under the N.P.C. is 1, but for convenience, the total area under the curve is taken to be 10,000 because of the greater ease with which fractional parts of the total area, may be then calculated.

The first column of the table, x/σ gives distance in tenths of ó measured off on the base line for the normal curve from the mean as origin. In the row, the x/ó distance are given to the second place of the decimal.

To find the number of cases in the normal distribution between the mean, and the ordinate erected at a distance of 1σ unit from the mean, you go down the x/σ column until 1.0 is reached and in the next column under .00 you take the entry opposite 1.0, namely 3413. This figure means that 3413 cases in 10,000; or 34.13 percent of the entire area of the curve lies between the mean and 1σ. Similarly, if you have to find the percentage of the distribution between the mean and 1 .56σ, say, you go down the x/σ column to 1.5, then across horizontally to the column headed by .06, and note the entry 44.06. This is the percentage of the total area that lies between the mean and 1.56σ.

Table 5.2: Fractional parts of the total area (taken as 10,000) under the normal probability curve, corresponding to distance on the baseline between the mean and successive points laid off from the mean in units of standard deviation.

x/σ	**.00**	**.01**	**.02**	**.03**	**.04**	**.05**	**.06**	**.07**	**.08**	**.09**
0.0	0000	0040	0080	0120	0160	0199	0239	0279	0319	0359
0.1	0398	0438	0478	0517	0557	0596	0636	0675	0714	0753
0.2	0793	0832	0871	0910	0948	0987	1026	1064	1103	1141
0.3	1179	1217	1255	1293	1331	1368	1406	1443	1480	1517
0.4	1554	1591	1628	1664	1700	1736	1772	1808	1844	1879
0.5	1915	1950	1985	2019	2054	2088	2123	2157	2190	2224

0.6	2257	2291	2324	2457	2389	2422	2454	2486	2517	2549
0.7	2580	2611	2642	2673	2704	2734	2764	2794	2823	2852
0.8	2881	2910	2939	2967	2995	3023	3051	3078	3106	3133
0.9	3159	3186	3212	3238	3264	3290	3315	3340	3365	3389
1.0	3413	3438	3461	3485	3508	3531	3554	3577	3599	3621
1.1	3643	3665	3686	3708	3729	3749	3770	3790	3810	3830
1.2	3849	3869	3889	3907	3925	3944	3962	3980	3997	4015
1.3	4032	4049	4066	4082	4099	4115	4131	4147	4162	4177
1.4	4192	4207	4222	4236	4251	4265	4279	4292	4306	4319
1.5	4332	4345	4357	4370	4383	4394	4406	4418	4429	4441
1.6	4452	4463	4474	4484	4495	4505	4515	4525	4535	4545
1.7	4554	4564	4573	4582	4591	4599	4608	4616	4625	4633
1.8	4641	4649	4656	4664	4671	4678	4686	4693	4699	4706
1.9	4713	4719	4726	4732	4738	4744	4750	4756	4761	4767
2.0	4772	4778	4783	4788	4793	4798	4803	4808	4812	4817
2.1	4821	4826	4830	4834	4838	4842	4846	4850	4854	4857
2.2	4861	4864	4868	4871	4875	4878	4881	4884	4887	4890
2.3	4893	4896	4898	4901	4904	4906	4909	4911	4913	4916
2.4	4918	4920	4922	4925	4927	4929	4931	4932	4934	4936
2.5	4938	4940	4941	4943	4945	4946	4948	4949	4951	4952
2.6	4953	4955	4956	4957	4959	4960	4961	4962	4963	4964
2.7	4965	4966	4967	4968	4969	4970	4971	4972	4973	4974
2.8	4974	4975	4976	4977	4977	4978	4979	4979	4980	4981
2.9	4981	4982	4982	4988	4984	4984	4985	4985	4986	4986
3.0	4986.5	4986.9	4987.4	4987.8	4988.2	4988.6	4988.9	4989.3	4989.7	4990.0
3.1	4990.3	4990.6	4991.0	4991.3	4991.6	4991.8	4992.1	4992.4	4992.6	4992.9
3.2	4993.129									
3.3	4995.166									
3.4	4996.631									
3.5	4997.674									
3.6	4998.409									
3.7	4998.922									
3.8	4999.277									
3.9	4999.519									
4.0	4999.683									
4.5	4999.966									
5.0	4999.997133									

You have so far considered only σ distances measured in the positive direction from the mean. For this you have taken into account only the right half of the normal curve. Since the curve is symmetrical about the mean, the entries in table 5.2 apply to distances measured in the negative direction (to the left) as well as to those measured in the positive direction. If you have to find the percentage of the distribution between mean and -1.28σ, for instance, you take entry 3997 in the column .08, opposite 1.2 in the x/σ column. This entry means that 39.97 percent of the cases in the normal distribution fall between the mean and -1.28σ.

For practical purposes you take the curve to end at points -3 σ and +3 σ distant from the mean as the normal curve does not actually meet the base line. Table of area under normal probability curve shows that 4986.5 cases lie between mean and ordinate at +3 σ. Thus, 99.73 percent of the entire distribution, would lie within the limits -3 σ and +3 σ. The rest 0.27 percent of the distribution beyond ±3 σ is considered too small or negligible except where N is very large.

5.9.1 CHECK YOUR PROGRESS

1 What formula is to use to convert raw score X into standard score i.e. z score.

2 What is the reference point on the normal probability curve.

3 Mean value of the z scores is_____

4 The value of standard deviation of z scores is_____

5 The total area under the N.P.C. is always_____

6 The negative value of z scores shows that_____

5.10 POINTS TO BE KEPT IN MIND WHILE CONSULTING TABLE OF AREA UNDER NORMAL PROBABILITY CURVE

The following points are to be kept in mind to avoid errors, while consulting the N.P.C. Table.

1. Every given score or observation must be converted into standard measure i.e. Z score, by using the following formula:

$$z = \frac{X - M}{\sigma}$$

2. The mean of the curve is always the reference point, and all the values of areas are given in terms of distances from mean which is zero.
3. The area in terms of proportion can be converted into percentage, and
4. While consulting the table, absolute values of z should be taken. However, a negative value of z shows that the scores and the area lie below the mean and this fact should be kept in mind while doing further calculation on the area. A positive value of z shows that the score lies above the mean i.e. right side.

5.11 PRACTICAL PROBLEMS RELATED TO APPLICATION OF THE NORMAL PROBABILITY CURVE

Under the section you have studied the Application of normal Distribution/ Normal Curve in mental and

educational measurements. Now how the practical problems related to this application are solved you go through the following examples carefully and thoroughly**.**

a) To determine the percentage of cases/number of scores in a normal distribution within given limits of scores.

The Normal Probability Curve helps us to determine:

i. What percent of cases fall between two scores of a distribution.

ii. What percent of scores lie above a particular score of a distribution.

iii. What percent of scores lie below a particular score of a distribution.

Sometimes researchers are often interested to know the number of cases or individuals that lie in between two points or two limits. For example, a teacher may be interested to know that how many students of his class got marks in between 60% and 70% in the annual examination, or he may be interested in how many students of his class got marks above 80%. In those cases, one can use the concept of NPC.

Example 5.1: An adjustment test was administered on a sample of 500 students of class VIII. The mean of the adjustment scores of the total sample obtained was 40 and standard deviation obtained was 8, what percentage of cases lie between the score 36 and 48, if the distribution of adjustment scores is normal in the universe.

Solution: In the problem it is given that N = 500, M = 40 and σ = 8.

You have to find out the total percentage of the students who obtained score in between 36 and 48 on the adjustment test. To find the required percentage of cases, first you have to find out the z scores for the raw scores (X) 36 and 48, by using the formula:

$$z = \frac{X - M}{\sigma}$$

∴z score for raw score 36 is

$$z_1 = \frac{36 - 40}{8} = -0.5$$

Similarly, z score for raw score 48 is

$$z_2 = \frac{48 - 40}{8} = +1$$

According to table of area under Normal Probability curve (N.P.C.) i.e. Table No. 3.2, the total area of the curve lies in between M to +1 σ is 34.13 and in between M to -0.5 σ is 19.15.

∴ The total area of the curve in between -0.5 σ to +1 σ is 19.15 + 34.13 = 53.28.

Thus, the total percentage of students who got scores in between 36 and 48 on the adjustment test is 53.28.

Example 5.2: A reading ability test was administered on the sample of 200 cases studying in IX class. The mean and standard deviation of the reading ability test score was obtained 60 and 10 respectively. Find how many cases lie in between the scores 40 and 70. Assume that reading ability scores are normally distributed.

Solution: Given, N = 200, M = 60, σ = 10, X_1 = 40 and X_2 = 70.

To find out the total no. of cases in between the two scores 40 and 70. You first have to find out the total percentage of cases lie in between Mean and 40 & mean and 70.

For the purpose, first the given raw scores (40 & 70) should be converted into z scores by using the formula:

$$z = \frac{X - M}{\sigma}$$

$$\therefore z_1 = \frac{40 - 60}{10} = -2$$

Similarly,

$$z_2 = \frac{70 - 60}{10} = +1$$

According to Table 5.2, the area of the curve in between M and -2σ is 47.72% and in between M and +1σ is 34.13%.

$\therefore$ The total area of the curve in between -2σ to +1σ is = 47.72 + 34.13 = 81.85%.

Therefore, the total no. of cases in between the two scores 40 and 70 are = $\frac{81.85 \times 200}{100}$ = 163.7 or 164. Thus total no. of cases who got scores in between 40 and 70 are = 164.

b) To determine the percentage of cases, lie above or below a given score or reference point.

You can use NPC table, in order to determine the percentage of cases that lie above or below a given score or reference point.

Example 5.3: An intelligence test was administered on a group of 500 cases of class V. The mean I.Q. of the students was found 100 and the S.D. of the I.Q. scores was 16. Find how many students of class V having the I.Q. below 80 and above 120.

Solution: Given, M = 100, σ = 16, X_1 = 80 and X_2 = 120.

To find out:

i. he total no. of cases below 80

ii. The total no. of cases above 120

To find the required no. of cases first you have to find z scores of the raw scores X_1 = 80 and X_2 = 120 by using the formula

$$z = \frac{X - M}{\sigma}$$

Thus,

$$z_1 = \frac{80 - 100}{16} = -\frac{20}{16} = -1.25$$

Similarly,

$$z_2 = \frac{120 - 100}{16} = \frac{20}{16} = +1.25$$

According to NPC table (Table 5.2) the total percentage of area of the curve lie in between Mean to 1.25 σ is = 39.44.

According to the properties of N.P.C. the 50% area lies below to the mean i.e. in left side and 50% area lie above to the mean i.e. in right side.

Thus, the total area of NPC curve below M = (100) is = 50 – 39.44 = 10.56.

Similarly, the total area of NPC curve above M = (100) is = 50 – 39.44 = 10.56.

Therefore, total cases below to the I.Q. 80 = $\frac{10.56 \times 500}{100}$ = 52.8 = 53 Approx.

Similarly, total cases above to the I.Q. 120 = $\frac{10.56 \times 500}{100}$ = 52.8 = 53 Approx.

Thus, in the group of 500 students of V class there are total 53 students having I.Q. below 80. Similarly, there are 53 students who have I.Q. above 120.

c) To determine the limits of scores which includes a given percentage of cases

Sometime a researcher is interested to know the limits of the scores in which a specified group of individuals lies, in that case he/ she can use NPC.

Example 5.4: An achievement test of mathematics was administered on a group of 75 students of class VIII. The value of mean and standard deviation was found 50 and 10 respectively. Find limits of the scores in which middle 60% students lies.

Solution: Given that, N = 75, M = 50, σ = 10.

You need to find out the value of the limits of middle 60% cases i.e. X_1 and X_2. As per given condition (middle 60% cases), 30%-30% cases lie left and right to the mean value of the group. According to the formula :

$$z = \frac{X - M}{\sigma}$$

If the value of M, σ and z is known, the value of X can be found out. In the given problem the value of M and σ are given. You can find out the value of z with the help of the NPC Table (table 5.2) as the area of the curve situated right and left to the mean (30%-30% respectively) is also given.

According to the table 5.2 the value of z_1 and z_2 of the 30% area is ± 0.84 σ.

Therefore, by using formula

$$z_1 = \frac{X_1 - M}{\sigma}$$

$$-0.84 = \frac{X_1 - 50}{10}$$

Or,

$X_1 = 50 - 0.84 \times 10 = 41.60 = 42$ approx.

Similarly,

$$z_2 = \frac{X_2 - M}{\sigma}$$

$$+0.84 = \frac{X_2 - 50}{10}$$

Or,

$X_1 = 50 + 0.84 \times 10 = 58.4 = 58$ approx.

Thus, $X_1 = 42$ and $X_2 = 58$.

Therefore, the middle 60% cases of the entire group (75, students) got marks on achievement test of mathematics in between 42 to 58.

d) To determine the percentage rank of the individual in his group.

The percentile rank is defined as the percentage of cases lie below to a certain score (X) or a point. Sometime an investigator is interested to know the position of an individual or a person in his own group on the bases of the trait is measured, in those cases NPC is very useful.

Example 5.5: In a group of 60 students of class X, Sumit got 75% marks in board examination. If the mean of whole class marks is 50 and S.D. is 10. Find the percentile rank of the Sumit in the class.

Solution: In this case, you have to find out the total percentage of cases (i.e. the area of N.P.C.) lie below to the point X = 75.

To find the total required area (shaded part) of the curve, it is essential first to know the area of the curve lie in between the points 50 and 75.

This area can be determined very easily, by taking up the help of N.P.C. Table (table 5.2), if you know the value of z of score 75.

According to the formula

$$z = \frac{X - M}{\sigma}$$

Thus,

$$z = \frac{75 - 50}{10} = \frac{25}{10} = 2.5$$

According to NPC table, the area of the curve lies M and +2.50 σ is 49.387.

In the present problem you have determined 49.38% area lies right to the mean and 50% area lies to the left of the Mean. (According to the characteristics of NPC, see section 3.1.4 characteristic no. 9).

Thus, according to the definition of percentile the total area of the curve lies below to the point X = 75 is = 50 + 49.38% = 99.38% or 99% Approx.

Therefore, the percentile rank of the Sumit in the class is 99th.

In other words, Sumit is the topper student in the class, remaining 99% students lie below to him.

e) To find out the percentile value of an individual's percentile rank.

Some time you are interested to know the percentage of score a person or an individual have got on the test paper having a specific percentile rank in the group.

Example 5.6: An intelligence test was administered on a large group of students of class VIII. The mean and standard deviation of the scores was obtained 65 and 15 respectively. On the basis intelligence test if the Ramesh's percentile rank in the class is 80, find what is the score of the Ramesh, he got on the test?

Solution: Given, M = 65, σ = 15, and PR = 80.

In this case you need to find out the value of P_{80}.

Now, as per definition of percentile rank, the 30% area of the curve lie from mean to the point P_{80} and 50% are lie to the left side of the mean.

The z value of the 30% area of the curve lies in between M and P80 is = +0.85 σ (table 5.2).

You know that

$$z = \frac{X - M}{\sigma}$$

or

$$+0.85 = \frac{X - 65}{15}$$

or

$$X = 65 + 15 \times 0.85$$
$$= 65 + 12.75 = 77.75 \text{ or } 78 \text{ Approx.}$$

Thus, Ramesh's intelligence score on the test is 78.

f) Dividing a group of individuals into sub-group according to the level of ability or a certain trait. If the trait or ability is normally distributed in the universe.

In some situations, you are making qualitative evaluation of the person or an individual on the basis of trait or ability, and assign them grades like A, B, C, D, E etc. or 1st grade, 2nd grade, 3rd grade etc. or High, Average or Low. For example, a company evaluate their salesman as A grade, B grade and C grade salesman. A teacher provides A, B, C etc. grades to his students on the basis of their performance in the examination. A psychologist may classify a group of persons on the basis of their adjustment as highly adjusted, Average and poorly adjusted. In such conditions, always there is a question that how many persons or individuals, you have to provide A, B, C, D and E etc. grades to the individuals and categorize them in different groups. In such situation NPC can prove very helpful.

Example 5.7: A company wants to classify the group of salesmen into four categories as Excellent, Good, Average and Poor on the basis of the sale of a product of the company, to provide incentive to them. If the number of salesmen in the company is 100, their average sale of the product per week is 10,00,000 Rs. and standard deviation is Rs. 500/-. Find the number of salesmen to place as Excellent, Good, Average and Poor.

Solution: As per property of the N.P.C. you know that total area of the curve is 6σ over a range of -3σ to $+3\sigma$.

According to the problem, the total area of the curve is divided into four categories.

Therefore, area of each category is $6\sigma/4 = \pm 1.5\sigma$. It means the distance of each category from the mean on the curve is 1.5σ respectively.

ii. Total % of salesman in "Good" category

According to N.P.C. Table, the Total area of the curve lies in between M an $+1.5\sigma$ is = 43.32%

∴ The total % of salesman in "Good" category is 43.32%

iii. Total % of salesman in "Average" category

Total area of the curve lies in between Mean and -1.5σ is also = 43.32%

∴ The total % of salesman in Average category is 43.32%

iv. Total % of salesman in "Excellent" category

The total area of the curve from M to $+3\sigma$ and above is = 50%

∴ The total % of salesman in the category Excellent is = 50 – 43.32 = 6.68%

v. Total % of salesman in "Poor" category

The total area of the curve from M to -3σ and below is = 50%

∴ The total % of the salesman in the poor category is = 50 – 43.32 = 6.68%

Thus,

i. The number of salesmen should place in "Excellent" category

$$= \frac{6.68 \times 100}{100} = 6.68 = 7\ approx.$$

ii. The number of salesmen should place in "Good" category

$$= \frac{43.32 \times 100}{100} = 43.32 = 43\ approx.$$

iii. The number of salesmen should place in "Average" category

$$= \frac{43.32 \times 100}{100} = 43.32 = 43\ approx.$$

iv. The number of salesmen should place in "poor" category

$$= \frac{6.68 \times 100}{100} = 6.68 = 7\ approx.$$

□ Total=100

Example 5.8: A group of 1000 applicants who wishes to take admission in a psychology course. The selection committee decided to classify the entire group into five sub-categories A, B, C, D and E according to their academic ability of last qualifying examination. If the range of ability being equal in each sub category, calculate the number of applicants that can be placed in groups A, B, C, D and E.

Solution: Given, N = 1000

In this case you have to find out the 1000 cases to be categorized into five categories A, B, C, D, and E.

You know that the base line of a normal distribution curve is considered extend from -3 σ to +3 σ that is range of 6 σ.

Dividing this range by 5 (the five subgroups) to obtain σ distance of each category, i.e. the z value of the cutting point of each category.

$$z = \frac{6\sigma}{5} = \pm 1.20$$

(It is to be noted here that the entire group of 1000 cases are divided into five categories. The number of subgroups is odd number. In such condition the middle group or middle category (c) will lie equally to the center i.e. M of the distribution of scores. In other words, the number of cases of "c" category or middle category remain half to the left area of the curve from the point of mean and half of the right area of the curve from the mean.

∴ the limits of "c" category is $= \frac{1.2\sigma}{2} = \pm 0.60\ \sigma$

i.e. the "c" category will remain on NPC curve in between the two limits -0.6 σ to +0.6 σ.

Now, the limits of B category, Lower limit = +0.6 σ and Upper limit = 0.60 σ + 1.20 σ or = +1.80 σ

The limits of A category, Lower limit = + 1.8 σ and Upper limit = + 3 σ and above.

Similarly, the limits of D category, Upper limit = - 0.6 σ and Lower limit = (- 0.60 σ) + (-1.20 σ) or = -1.80 σ.

The limits of E category, Upper limit = -1.8 σ and Lower limit = -3 σ and below.

i. The total % area of the NPC for A category:

According to NPC Table (1.6.1) the total % of area in between Mean to +1.80 σ is = 46.41.

∴ The total % of area of the NPC for A category is = 50 – 46.41 = 3.59.

ii. The total % Area of the NPC for B category:

According to NPC Table (5.2) the total % of Area in between Mean and + 0.60 σ is = 22.57.

∴ The total % area of NPC for B category is = 46.41 – 22.57 = 23.84.

iii. The total % area of the NPC for C category:

According to NPC table the total % area of NPC in between M and + 0.06 σ is = 22.57.

Similarly, the total % area of NPC in between M and – 0.06 σ is also = 22.57.

∴ The total % area of NPC for C category is = 22.57 + 22.57 = 45.14.

iv. In similar way the total % area of NPC for D category is = 23.84.

v. The total % area of NPC for E category is = 3.59

Thus, the total number of applicants (N = 1000) in

i. A category is = $\frac{3.59 \times 1000}{100}$ = 35.9 = 36

ii. B category is = $\frac{23.84 \times 1000}{100}$ = 238.4 = 238

iii. C category is = $\frac{45.14 \times 1000}{100}$ = 451.4 = 452

iv. D category is = $\frac{23.84 \times 1000}{100}$ = 238.4 = 238

v. E category is = $\frac{3.59 \times 1000}{100}$ = 35.9 = 36

❒ Total= 1000

g) To compare the two distributions in terms of overlapping.

If scores of two groups on a particular variable are normally distributed. What you know about the group is the mean and standard deviation of both the groups. And you want to know how much the first group over-laps the second group or vice-versa at that time you can determine this by using the table area under NPC.

Example 5.9: A numerical ability test was administered on 300 graduate boys and 200 graduate girls. The boys Mean score is 26 with S.D. (σ) of 4. The girls' mean. Mean score is 28 with a σ = 8. Find the total number of boys who exceed the mean of the girls and total number of girls who got score below to the mean of boys.

Solution: Given,

For Boys, N = 300, M = 26 and σ = 6
For Girls, N = 200, M = 28 and σ = 8

You need to find

i. Number of boys who exceed the mean of girls.

ii. Number of girls who scored below to the mean of boys.

As per given conditions, first you have to find the number of cases above the point 28.

(The mean of the numerical ability scores of girls) by considering M = 26 and σ = 6.

Second, you to find no. of cases below to the point 26 (The mean score of the boys), by considering M = 28 and σ = 8.

1. The z score of X (28) is = $\frac{28-26}{6} = \frac{2}{6}$

or = + 0.33

According to NPC Table (3.2) the total % of area of the NPC from M = 26 to + 0.33 σ is 12.93

∴ The total % of cases above to the point 28 is = 50 – 12.93 = 37.07

Thus, the total number of boys above to the point 28 (mean of the girls) is = $\frac{37.07 - 300}{100}$ = 111.21 = 111

2. The z score of X (26) is = $\frac{26-28}{8} = -\frac{2}{8} = -0.25$

According to the NPC table the total % of area of the curve in between M = 28 and -0.25 σ is 9.87.

∴ Total % of cases below to the point 26 is = 50 – 9.87 = 40.13

Thus, the total number of girls below to the point 26 (mean of the boys) is = $\frac{40.13 - 200}{100}$ = 80.26 = 80

Therefore,

i. The total number of boys who exceed the mean of the girls in numerical ability is = 111

ii. The total number of girls who are below to the mean of the boys is = 88

h) To determine the relative difficulty of a test items or problem

When it is known that what percentage of students successfully solved a problem, you can determine the difficulty level of the item or problem by using table area under NPC.

Example 5.10: In a mathematics achievement test meant for 10th standard class, question 1, 2 and 3 are solved by the students 60%, 30% and 10%, respectively. find the relative difficulty level of each Question. Assume that solving capacity of the students is normally distributed in the universe.

Solution: Given, the percentage of the students who are solving the test items (Qs) of a question paper correctly.

Here, you have to find the relative difficulty level of each item of the test paper given.

First of all, you shall mark the relative position of test items on the basis of percentage of students solving the items successfully on the NPC scale.

Q.No.3 of the test paper is correctly solved by the 10% students only. It means 90% students unable to attend the Q.No. 3. On the NPC scale, these 10% cases lie extreme to the right side of the mean.

Similarly, 30% students who are solving Q.No. 2 correctly also lying to the right side of the curve. While the 60% students who are solving Q.No. 1 correctly are lying left side of the NPC curve. Now, you have to find out the z value of the cut point of each item (Q.No.) on the NPC base line.

i. The z value of the cut point of Q.No. 3 The total percentage of cases lie in between mean and cut point of Q.No. 3 is = (50% - 10%) in right half of NPC

   ∴ The z value of the right 40% of area of the NPC is = 11.28 σ

ii. The z value of the cut point of Q.No.2 The total percentage of cases lie between the mean and cut point of Q.No. 2 is = 20% (50% - 30%) in right half of NPC

   ∴ The z value of the right 20% area of the NPC is = + 0.52 σ

iii. The z value of the cut point of Q.No. 3 The total percentage of cases lie between the mean and cut point of Q.No. 3 is = (60% - 50%) in left half of NPC

   ∴ The z value of the left of 10% of area = - 0.25 σ Therefore corresponding z value of each item (Q) passed by the students is

Item (Q.No.)	Passed By	z value	Z difference
3	10%	+ 1.28 σ	-
2	30%	+ 0.52 σ	0.76 σ
1	60%	- 0.25 σ	0.77 σ

You may now compare the three questions of the mathematics achievement test, Q.No. 1 has a difficulty value of 0.76 σ higher than the Q.No. 2. Similarly, the Q.No. 2 has a difficulty value of 0.77 σ higher than the Q.No. 3. Thus, the Q.No. 1, 2 and 3 of the mathematics achievement tests are the good items having equal level of difficulty and are quite discriminative.

5.12 DIVERGENCE IN NORMALITY (THE NON-NORMAL DISTRIBUTION)

In a frequency polygon or histogram of test scores, usually the first thing that strikes one is the symmetry or lack of it in the shape of the curve. In the normal curve model, the mean, the median and the mode all coincide and there is perfect balance between the right and left halves of the curve. Generally, two types of divergence occur in the normal curve.

1. Skewness
2. Kurtosis

1. **Skewness:** A distribution is said to be "skewed" when the mean and median fall at different points in the distribution and the balance i.e. the point of center of gravity is shifted to one side or the other to left or right. In a normal distribution the mean equals, the median exactly and the skewness is of course zero ($S_K = 0$). There are two types of skewness which appear in the normal curve.

   a. **Negative Skewness:** Distribution said to be skewed negatively or to the left when scores are massed at the high end of the scale, i.e. the right side of the curve are spread out more gradually toward the low end i.e. the left side of the curve. In negatively skewed distribution the value of median will be higher than that of the value of the mean.

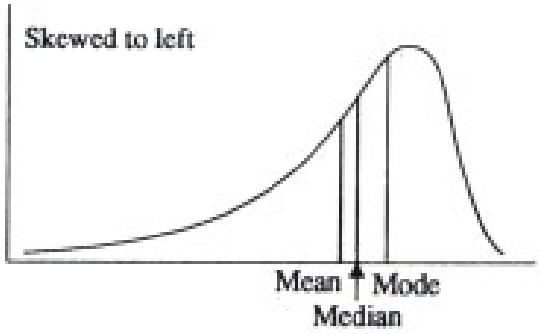


Figure 5.6: Negative Skewness

   b. **Positive Skewness:** Distributions are skewed positively or to the right, when scores are massed at the low; i.e. the left end of the scale, and are spread out gradually toward the high or right end as shown in the figure 5.7.

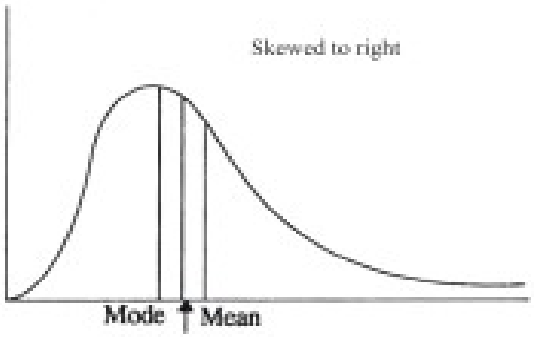


Figure 5.7: positive Skewness

2. **Kurtosis:** The term kurtosis refers to (the divergence) in the height of the curve, especially in the peakness. There are two types of divergence in the peakness of the curve.

   a. **Leptokurtosis:** Suppose you have a normal curve which is made up of a steel wire. If you push both the ends of the wire curve together. What would happen in the shape of the curve? Probably your answer may be that by pressing both the ends of the wire curve, the curve become more peeked i.e. its top become narrower than the normal curve and scatterdness in the scores or area of the curve shrink towards the center.

Thus, in a Leptokurtic distribution, the frequency distribution curve is more peaked than to the normal distribution curve. Leptokurtic distributions have more kurtosis than the normal distribution. These distributions have heavy tails that are longer and contain more extreme values. In short, there is a greater tendency for outliers because the tails are thicker. They have values of greater than 3 or positive excess values (> 0). Examples of leptokurtic distribution are t-distribution, laplace distribution etc.

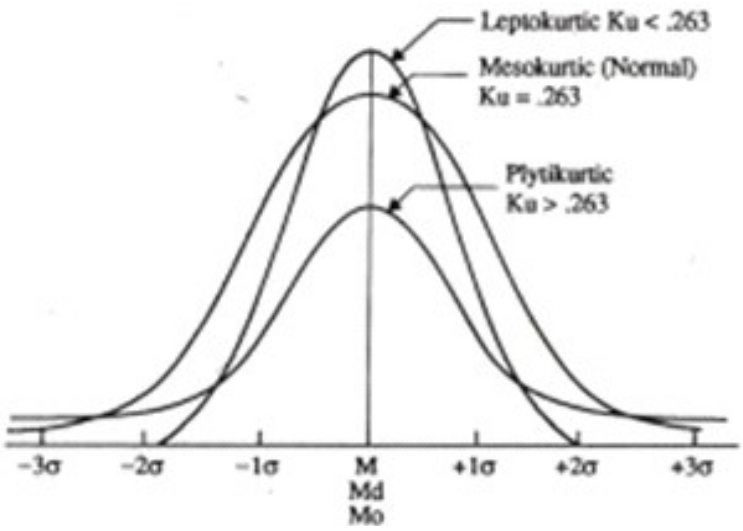


Figure 5.8: Kurtosis in the Normal Curve

b. Platykurtosis: Now suppose we put a heavy pressure on the top of the wire made normal curve. What would be the change in the, shape of the curve? Probably you may say that the top of the curve become flatter than to the normal.

Thus, a distribution of flatter Peak than to the normal is known Platykurtic distribution.

Platykurtic distributions have less kurtosis than the normal distribution. They have lighter tails that are shorter and contain fewer outliers because the tails are thinner. These distributions have values of less than 3 or negative excess values (< 0). Examples of Platykurtic distributions are uniform distribution, beta distribution etc.

c. Mesokurtosis: In Mesokurtosis, you have the same curve as normal curve.

Thus, a distribution which has same curve as normal curve is known as Mesokurtic distributions. Mesokurtic distributions have kurtosis values close to that of a normal distribution. These distributions have a value of approximately 3 or an excess value near zero. When sample data have kurtosis values that are notably different from the normal distribution, it indicates that the population might not follow a normal distribution.

5.13 FACTORS CAUSING DIVERGENCE IN THE NORMAL DISTRIBUTION/ NORMAL CURVE

The reasons on why distribution exhibit skewness and kurtosis are numerous and often complex, but a careful analysis of the data will often permit the common causes of asymmetry. Some of common causes are –

1. **Selection of the Sample:** Selection of the subjects (individuals) produce skewness and kurtosis in the distribution. If the sample size is small or sample is biased one, skewness is possible in

the distribution of scores obtained on the basis of selected sample or group of individuals.

If the scores made by small and homogeneous group are likely to yield narrow and leptokurtic distribution. Scores from small and highly heterogeneous groups yield platykurtic distribution.

2. **Unsuitable or Poorly Made Tests:** If the measuring tool or test is inappropriate, or poorly made, the asymmetry is possible in the distribution of scores. If a test is too easy, scores will pile up at the high end of the scale, whereas the test is too hard, scores will pile up at the low end of the scale.

3. **The Trait being Measured is Non-Normal:** Skewness or Kurtosis or both will appear when there is a real lack of normality in the trait being measured, e.g. interest, attitude, suggestibility, deaths in an old age or early childhood due to certain degenerative diseases etc.

4. **Errors in the Construction and Administration of Tests:** The unstandardized with poor item-analysis test may cause asymmetry in the distribution of the scores. Similarly, while administrating the test, the unclear instructions – Error in timings, Errors in the scoring, practice and motivation to complete the test all these factors may cause skewness in the distribution.

5.14 MEASURING DIVERGENCE IN THE NORMAL DISTRIBUTION / NORMAL CURVE

The divergence in normal distribution/normal curve has a significant role in construction of the ability and to test the representativeness of a sample taken from a large population. Further the divergence in the distribution of scores or measurements obtained of a certain population reflects some important information about the trait of population measured. Thus, there is a need to measure the two divergences i.e. skewness and kurtosis of the distribution of the scores.

a) Measuring Skewness

There are two methods to study the skewness in a distribution.

i. Observation Method

ii. Statistical Method

1. **Observation Method:** There is a simple method of detecting the directions of skewness by the inspection of frequency polygon prepared on the basis of the scores obtained regarding a trait of the population or a sample drawn from a population.

   Looking at the tails of the frequency polygon of the distribution obtained, if longer tail of the curve is towards the higher value or upper side or right side to the center or mean, the skewness is positive. If the longer tail is towards the lower values or lower side or left to the mean, the skewness is negative.

2. **Statistical Method:** To know the skewness in the distribution we may also use the statistical method. For the purpose we use measures of central tendency, specifically mean and median values and use the following formula

$$S_k = \frac{3(Mean - Median)}{\sigma}$$

Another measure of skewness based on percentile values is as under

$$S_k = \frac{P_{90} - P_{10}}{2} - P_{50}$$

Here, it is to be kept in mind that the above two measures are not mathematically equivalent. A normal curve has the value of $S_K = 0$. Deviations from normality can be negative and positively direction leading to negatively skewed and positively skewed distributions, respectively.

b) Measuring Kurtosis

For judging whether a distribution lacks normal symmetry or peakness; it may be detected by inspection of the frequency polygon obtained. If a peak of curve is thin and sides are narrow to the center, the distribution is leptokurtic and if the peak of the frequency distribution is too flat and sides of the curve are deviating from the center towards ± 4σ or ± 5σ than the distribution is platikurtic.

Kurtosis can be measured by following formula using percentile values.

$$K_U = \frac{Q}{P_{90} - P_{10}}$$

Where,
Q = quartile deviation, P_{10} = 10th percentile and P_{90} = 90th percentile

A normal distribution has $K_U = 0.263$. If the value of K_U is less than 0.263 ($K_U < 0.263$), the distribution is leptokurtic and if K_U is greater than 0.263 ($K_U > 0.263$), the distribution is platykurtic.

5.15 CHECK YOUR PROGRESS

7 The positive value of z scores shows that

8 In a distribution what percentage of frequencies are lie in between

a) -1 σ to + 1 σ

b) -2 σ to + 2 σ

c) -3 σ to + 3 σ

9 A numerical ability test was administered on 500 graduate boys and 700 graduate girls. The boys Mean score is 26 with S.D. (σ) of 4. The girls' mean. Mean score is 28 with a S.D = 8. Find

- **i.** Number of boys between the two means 26 and 28
- **ii.** Number of girls between the two means 26 and 28
- **iii.** Number of boys below to the mean of girls
- **iv.** Number of girls above to the mean of boys
- **v.** Number of boys above to the Md of girls which is 28.20
- **vi.** Number of girls exceed to the Md of the boys which is 26.20

10 The three test items 1, 2 and 3 of an ability tests are solved by 10%, 20% and 30% respectively. What are the relative difficulty values of these items?

11 The observation given in the example 4, i.e. M = 50 and S.D. (σ) = 10

- **i.** Find the limits of the scores middle 30% cases
- **ii.** Find the limits of the scores middle 75% cases
- **iii.** Find the limits of the scores middle 50% cases

12 In a test of 200 items, each correct item has 1 mark. If M = 100, σ = 10

- **i.** Find the position of Rohit in the group who secured 85 marks on the test.
- **ii.** Find the percentile rank of Sunita she got 130 marks on the test.

5.16 LET'S US SUM UP

The normal distribution is a very important concept in statistics because most of the variables used in analytical research are assumed to be normally distributed. In statistical researches, each variable has a specific mean and standard deviation, there is a family of normal distribution rather than just a single distribution. However, if you know the mean and standard deviation for any normal distribution you can transform it into the standard normal distribution. The standard normal distribution is the normal distribution in standard score (z) form with mean equal to 0 and standard deviation equal to 1.

Normal curve is much helpful in psychological and educational measurement and educational evaluation. It provides relative positioning of the individual in a group. It can also be used as a scale of measurement in behavioural sciences. The normal distribution is a significant tool in the hands of teacher and researchers. Through which he can decide the nature of the distribution of the scores obtained on the basis of measured variable. Also, he can decide about his own scoring process which is very lenient or hard; he can Judge

the difficulty level of the test items in the question paper and finally he may know about his class, whether it is homogeneous to the ability measured or it is heterogeneous one.

5.17 KEY POINTS

- **Skewness:** Skewness is a measure of asymmetry of the probability distribution of a random variable about its mean.
- **Kurtosis:** Kurtosis is a measure of "tailedness" of the probability distribution of a random variable. In other words, it is a measure whether a data is heavy tailed or light tailed in relation to normal distribution.
- **Normal Probability Curve:** A normal curve is a bell-shaped curve, bilaterally symmetrical and continuous frequency distribution curve.
- **Normal Probability Distribution:** A continuous probability distribution for a variable is called as normal probability distribution or simply normal distribution. It is also known as Gaussian/ Gauss or Laplace – Gauss distribution.
- **Standard score:** Standard score or z-score is a transformed score which shows the number of standard deviation units by which the value of observation (the raw score) is above or below the mean.

5.18 SELF-ASSESSMENTS

1 Define a Normal Probability Curve.

2 Write the properties of Normal Distribution.

3 Mention the conditions under which the frequency distribution can be approximated to the normal distribution

4 Define the following:

i. Skewness

ii. Kurtosis

5 In case of normal distribution what should be the value of skewness.

6 In case of normal distribution what should be the value of Kurtosis.

7 How you can instantly study the skewness in a distribution.

8 What is the formula to measure skewness in a distribution?

9 What indicates the kurtosis of a distribution?

10 What formula is used to calculate the value of kurtosis in a distribution?

11 How you decide that a distribution is leptokurtic or platykurtic?

5.19 LESSON END EXERCISE

1 Given a distribution of scores with a mean of 24 and S.D. of 8. Assuming normality what percentage of the cases will fall between 16 and 32.

2 Given a distribution of scores with a mean of 40 and S.D. of 8. Assuming normality what percentage of cases will lie above and below the score 36.

3 In a distribution of scores of a doss Pinky's percentile rank in statistics is 65. The mean of the distribution is 55 with a standard deviation of 10. Find but the raw score of Pinky in Statistics.

4 An achievement test of mathematics was administered on a group of 75 students of class V. The value of mean and standard deviation was found 50 and 10 respectively. Find the limits of the scores middle 1) 30% cases, 2) 75% cases and 3) 50% cases.

5 Given a distribution of scores with a mean of 20 and S.D. of 5. If you assume normality what limits will include the middle 75% of cases.

6 In a test of 200 items, each correct item has 1 mark. If M = 100, σ = 10. Find the position of Rohit in the group who secured 85 marks on the test. Also, find the percentile rank of Sunita she got 130 marks on the test.

7 If M = 100, σ = 10 Find the values of P_{75}, P_{10}, P_{50} and P_{80}.

8 A company wants to classify the group of salesmen into Six categories as excellent, very good, good average, poor and very poor on the basis of the sale of a product of the company, to provide incentive to them. If the number of salesmen in the company is 200, their average sale of the product per week is 10,00,000 Rs. and standard deviation is Rs. 800/-. Find the number of salesmen in each category as per their sales ability.

9 An achievement test was administered to the 600 8th grade students. The teacher wants to assign these students in to 4 grades namely A, B, C and D according to their performance in the test. Assuming the normality of the distribution of scores calculate the number of students can be placed in each group.

10 A group of 3000 applicants who wishes to take admission in a psychology course. The selection committee decided to classify the entire group into three sub-categories A, B and C according to their academic ability of last qualifying examination. Find how many applicants will be the categorized in group A, B and C.

5.20 SUGGESTED READINGS

- Gupta, S.C. and Kapoor, V.K. (2017): "*Fundamental of Mathematical Statistics".* S Chand Publication.
- Aggarwal, Y.P. (1986): *"Statistical Methods-Concepts, Applications and Computation"*. New Delhi: Sterling Publishers Pvt. Ltd.
- Veeraraghavan, V and Shetgovekar, S. (2016): "*Textbook of Parametric and Nonparametric Statistics"*. Delhi: Sage.
- Mood, A.M., Graybill, F. and Boes, D. (2017): " *Introduction to theory of Statistics*". Mcgraw hill.

LESSON : 6

ANALYZING CORRELATION: ASSESSING THE STRENGTH AND DIRECTION OF RELATIONSHIPS IN EDUCATION

Structure

6.1 INTRODUCTION

In previous lesson, you have learnt about the Normal probability curve which provides relative positioning of the individual in a group on the basis of their ability or any other measurement. Now, you will learn measures of relationship and how to examine the relationship between two variables. The measures of relationship study the relationship between two or more variables in a given data series. When you study the relationship between two variables in a population, it is known as bivariate population. When you study more than two variables in a population, it is known as multivariate population. The relationship among variables can be of two types – correlation and cause and effect. Based on these relationships, there are two types of analysis namely Correlation Analysis and Regression Analysis. But in this lesson, you will study only correlation analysis. As the summer heat rises, hill stations, are crowded with more and more visitors. Ice-cream sales become brisker. Thus, the temperature is related to number of visitors and sale of ice-creams. Similarly, as the supply of tomatoes increases in your local mandi, its price drops. When the local harvest starts reaching the market, the price of tomatoes drops from a princely Rs 40 per kg to Rs 4 per kg or even less. Thus, supply is related to price. **Correlation analysis** is a means for examining such relationships systematically. It deals with questions such as:

- Is there any relationship between two variables?
- If the value of one variable changes, does the value of the other also change?
- Do both the variables move in the same direction?
- How strong is the relationship?

6.2 LEARNING OBJECTIVES

After reading this unit, you would be able to

- Describe the concept of correlation;
- Explore the types of correlation;
- Describe the scatter diagram;
- Calculate the product moment correlation coefficient, spearman's rank correlation and partial correlation coefficient;
- Describe the properties of correlation coefficient; and
- Describe uses and limitation of correlation

6.3 CORRELATION ANALYSIS

Suppose that you have the height and weight of each one of a group of 12-year-old girls. You can, of

course, compute the mean and standard deviation of height; and you can do the same for weight. But a more important question may well be: Is there a correlation, or relationship, between height and weight? Are the tall girls apt to weigh more or less than the short girls? Do the heavy girls tend to be above or below the average in height? When two such sets of measurements are associated so that the measurements in one set are related to those in the other set, you can say that the two sets of measurements are correlated. Correlation is a statistical measure that indicates the extent to which two or more variables fluctuate together. A positive correlation indicates the extent to which those variables increase or decrease in parallel; a negative correlation indicates the extent to which one variable increases as the other decreases. When the fluctuation of one variable reliably predicts a similar fluctuation in another variable, there's often a tendency to think that the change in one causes the change in the other. However, correlation does not imply causation. There may be an unknown factor that influences both variables similarly.

Correlation studies and measures the direction and intensity of relationship among variables. Correlation measures covariation, not causation. Correlation should never be interpreted as implying cause and effect relation. The presence of correlation between two variables X and Y simply means that when the value of one variable is found to change in one direction, the value of the other variable is found to change either in the same direction or in the opposite direction, but in a definite way. For simplicity, let's assume here that the correlation, if it exists, is linear, i.e. the relative movement of the two variables can be represented by drawing a straight line on graph paper.

Correlation shows whether and how strongly pairs of variables are related. Although this correlation is fairly obvious, your data may contain unsuspected correlations. You may also suspect there are correlations, but don't know which are the strongest. An intelligent correlation analysis can lead to a greater understanding of your data.

6.4 TYPES OF CORRELATION

Correlation is commonly classified into two categories-

1. Positive Correlation
2. Negative Correlation

1. **Positive Correlation:** The correlation is said to be positive when the variables move together in the same direction. In simple words, correlation is positive or direct when the values of variables increase/decreases together. For example, When the income rises, consumption also rises and when income falls, consumption also falls; Sale of ice-cream and temperature move in the same direction as well. Positive correlation can be either strong or weak. Following figure 6.1a and figure 6.1b shows the graphical representation of positive correlation.

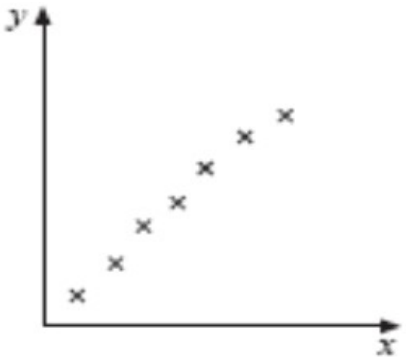


Figure 6.1a: Strong positive correlation

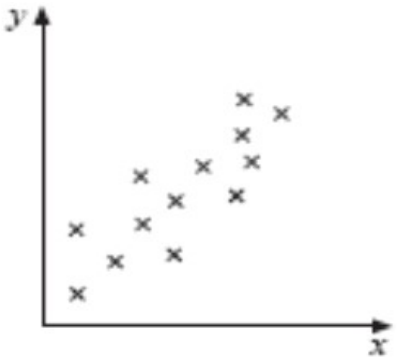


Figure 6.1b: Weak positive correlation

Figure 6.1a shows **strong positive correlation** between x and y. The points lie close to a straight line with y increasing as x increases.

Figure 6.1b shows **weak positive correlation** between x and y. The trend shown is that y increases as x increases but the points are not close to a straight line.

2. **Negative Correlation:** The correlation is negative when the variables move in opposite directions, which means that correlation is Negative when one value decreases as the other increases or vice versa. For example, when the price of apples falls its demand increases and when the prices rise its demand decreases; when you spend more time in studying, chances of your failing decline and when you spend less hours in your studies, chances of scoring low marks/grades increase. Like positive correlation, negative correlation can also be strong and weak. Following figures (6.2a and 6.2b) gives the graphical representation of strong and weak negative correlation.

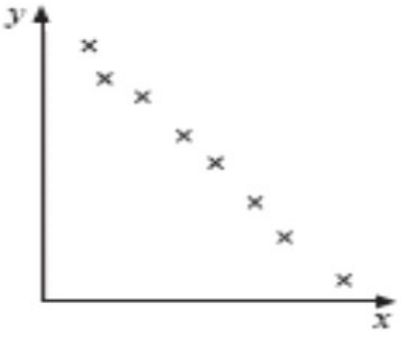


Figure 6.2a: Strong negative correlation

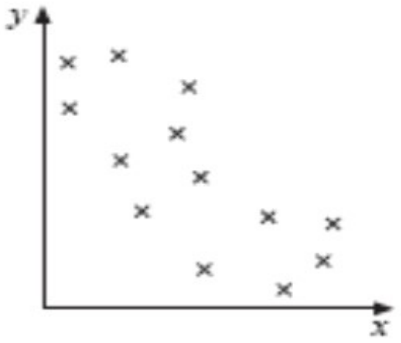


Figure 6.2b: Weak negative correlation

Figure 6.2a shows **strong negative correlation** between x and y. The points lie close to a straight line, with y decreasing as x increases.

Figure 6.2b shows **weak negative correlation** between x and y. The trend shown is that y decreases as x increases but the points do not lie close to a straight line.

Note: Majorly there are only two types of correlation viz positive correlation and negative correlation. But in real life there exists some conditions where you will come across a term called **No correlation.** Two variables are said to have no correlation if there are not related to each other, which means that they do not have any kind of association between them. Figure 6.3 shows the graphical representation of no correlation between x and y; the points are distributed randomly on the graph.

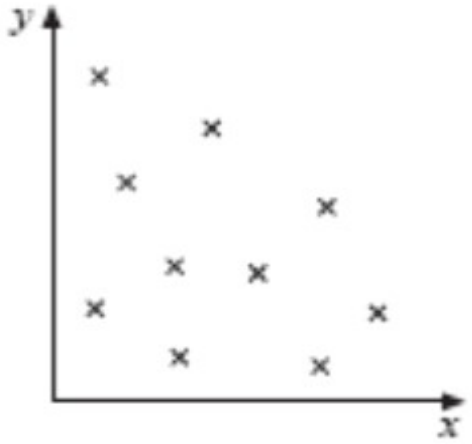


Figure 6.3: No Correlation

6.5 SCATTER DIAGRAM

A scatter diagram is a useful technique for visually examining the form of relationship, without calculating any numerical value. In this technique, the values of the two variables are plotted as points on a graph paper. From a scatter diagram, one can get a fairly good idea of the nature of relationship. In a scatter diagram the degree of closeness of the scatter points and their overall direction enable us to examine the relationship. If all the points lie on a line, the correlation is perfect and is said to be unity. If the scatter points are widely dispersed around the line, the correlation is low. The correlation is said to be linear if the scatter points lie near a line or on a line.

Scatter diagrams spanning over figure 6.1 to figure 6.3 give you an idea of the relationship between two variables. Figure 6.1b, shows a scatter around an upward rising line indicating the movement of the variables in the same direction. When x rises y will also rise. This is positive correlation. In figure 6.2b, the points are found to be scattered around a downward sloping line. This time the variables move in opposite directions. When x rises y falls and vice versa. This is negative correlation. In figure 6.3, there is no upward rising or downward sloping line around which the points are scattered. This is an example of no correlation. In figure 6.1a and figure 6.2a the points are no longer scattered around an upward rising or downward falling line. The points themselves are on the lines. This is referred to as perfect (strong) positive correlation and perfect (strong) negative correlation, respectively.

A careful observation of the scatter diagram gives an idea of the nature and intensity of the relationship.

6.6 KARL PEARSON'S PRODUCT MOMENT CORRELATION

This is also known as Pearson's coefficient of correlation and simple correlation coefficient. It gives a precise numerical value of the degree of linear relationship between two variables X and Y. The linear relationship may be given by

$$Y = a + bX$$

This type of relation may be described by a straight line. The intercept that the line makes on the Y-axis is given by a and the slope of the line is given by b. It gives the change in the value of Y for very small change in the value of X. On the other hand, if the relation cannot be represented by a straight line as in

$$Y = X^2$$

the value of the coefficient will be zero. It clearly shows that zero correlation need not mean absence of any type of relation between the two variables. Let X_1, X_2, ..., X_N be N values of X and Y_1, Y_2, ..., Y_N be the corresponding values of Y. In the subsequent presentations the subscripts indicating the unit are dropped for the sake of simplicity. The arithmetic means of X and Y are defined as

$$\bar{X} = \frac{\sum X}{N} \; and \; \bar{Y} = \frac{\sum Y}{N}$$

and their variances are as follows

$$\sigma_X^2 = \frac{\sum (X - \bar{X})^2}{N} = \frac{\sum X^2}{N} - \bar{X}^2$$

And

$$\sigma_Y^2 = \frac{\sum (Y - \bar{Y})^2}{N} = \frac{\sum Y^2}{N} - \bar{Y}^2$$

The standard deviations of X and Y respectively are the positive square roots of their variances. Covariance of X and Y is defined as

$$Cov(X, Y) = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{N} = \frac{\sum xy}{N}$$

Where, $x = X - \bar{X} \; and \; y = Y - \bar{Y}$ are the deviations of the i^{th} value of X and Y from their mean values respectively.

The sign of covariance between X and Y determines the sign of the correlation coefficient. The standard deviations are always positive. If the covariance is zero, the correlation coefficient is always zero. The product moment correlation or the Karl Pearson's measure of correlation is given by

$$r = \frac{\sum xy}{N \sigma_X \sigma_Y}$$

or

$$r = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum (X - \bar{X})^2} \sqrt{\sum (Y - \bar{Y})^2}}$$

or

$$r = \frac{N \sum XY - \sum X \sum Y}{\sqrt{N \sum X^2 - (\sum X)^2} \sqrt{N \sum Y^2 - (\sum Y)^2}}$$

6.7 ASSUMPTIONS OF COEFFICIENT OF CORRELATION

Assumptions of a Pearson correlation have been intensely debated. It is therefore not surprising, but nonetheless confusing, that different statistical resources present different assumptions. In reality, the coefficient can be calculated as a measure of a linear relationship without any assumptions. However, proper inference on the strength of the association in the population from which the data were sampled (what one is usually interested in) does require that some assumptions be met:

1. As is actually true for any statistical inference, the data are derived from a random, or at least representative, sample. If the data are not representative of the population of interest, you cannot draw meaningful conclusions about that population.
2. Both variables are continuous, jointly normally distributed, random variables. They follow a bivariate normal distribution in the population from which they were sampled.
3. If there is a relationship between jointly normally distributed data, it is always linear. Therefore, if the data points in a scatter plot seem to lie close to some curve, the assumption of a bivariate normal distribution is violated.
4. There are no relevant outliers. Extreme outliers may have undue influence on the Pearson correlation coefficient. While it is generally not legitimate to simply exclude outliers, running the correlation analysis with and without the outlier(s) and comparing the coefficients is a possibility to assess the actual influence of the outlier on the analysis. For data with relevant outliers, Spearman correlation is preferred as it tends to be relatively robust against outliers.
5. Each pair of x–y values is measured independently from each other pair. Multiple observations from the same subjects would violate this assumption. The way to deal with such data depends on whether you are interested in correlations within subjects or between subjects.

6.8 PROPERTIES OF COEFFICIENT OF CORRELATION

Following are the properties of correlation coefficient 'r':

1. Correlation Coefficient 'r' has no unit. It is a pure number. It means units of measurement are not part of r. r between height in feet and weight in kilograms, for instance, could be say 0.7.
2. A negative value of r indicates an inverse relation. A change in one variable is associated with change in the other variable in the opposite direction. When price of a commodity rises, its demand falls. When the rate of interest rises the demand for funds also falls. It is because now funds have become costlier.
3. If r is positive the two variables move in the same direction. When the price of coffee, a substitute

of tea, rises the demand for tea also rises. Improvement in irrigation facilities is associated with higher yield. When temperature rises the sale of ice-creams becomes brisk.

4. The value of the correlation coefficient lies between minus one and plus one, $-1 \leq r \leq 1$. If, in any exercise, the value of r is outside this range it indicates error in calculation.

5. If r = 0, the two variables are uncorrelated. There is no linear relation between them. However other types of relation may be there.

6. If r = 1 or r = –1 the correlation is perfect. The relation between them is exact.

7. A high value of r indicates strong linear relationship. Its value is said to be high when it is close to +1 or –1.

8. A low value of r indicates a weak linear relation. Its value is said to be low when it is close to zero.

9. The magnitude of r is unaffected by the change of origin and change of scale. Given two variables X and Y let us define two new variables.

$$U = \frac{X - A}{B}; V = \frac{Y - C}{D}$$

Where, A and C are assumed means of X and Y respectively. B and D are common factors and of same sign. Then,

$$r_{XY} = r_{UV}$$

This. property is used to calculate correlation coefficient in a highly simplified manner, as in the step deviation method.

Example 6.1: Find the correlation coefficient between advertisement expenditure and profit for the following data:

Advertisement expenditure	30	44	45	43	34	44
Profit	56	55	60	64	62	63

Solution: To find out the correlation coefficient between advertisement expenditure and profit, you have Karl Pearson's formula given as:

$$r = r_{XY} = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum(X - \bar{X})^2}\sqrt{\sum(Y - \bar{Y})^2}}$$

X	**Y**	$X-\bar{X}$	$(X-\bar{X})^2$	$Y-\bar{Y}$	$(Y-\bar{Y})^2$	$(X-\bar{X})(Y-\bar{Y})$
30	56	−10	100	−4	16	40
44	55	4	16	−5	25	−20
45	60	5	25	0	0	0
43	64	3	9	4	16	12
34	62	−6	36	2	4	−12
44	63	4	16	3	9	12
$\sum X$ **=240**	$\sum Y$ **=360**		$\sum(X-\bar{X})^2$ **=202**		$\sum(Y-\bar{Y})^2$ **=70**	$\sum(X-\bar{X})(Y-\bar{Y})$ **=32**

Here, N=6

Now,

$$\bar{X}=\frac{\sum X}{N}=\frac{240}{6}=40$$

$$\bar{Y}=\frac{\sum Y}{N}=\frac{360}{6}=60$$

substituting the values from above table in the formula, you have

$$r=r_{XY}=\frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sqrt{\sum(X-\bar{X})^2}\sqrt{\sum(Y-\bar{Y})^2}}=\frac{32}{\sqrt{(202)(70)}}$$

$$r=\frac{32}{\sqrt{14140}}=\frac{32}{118.91}=0.27$$

Hence, the correlation coefficient between expenditure on advertisement and profit is 0.27. This indicates that the correlation between expenditure on advertisement and profit is positive and you can say that as expenditure on advertisement increases (or decreases) profit increases (or decreases). Since, it lies between 0.25 and 0.5 it can be considered as weak positive correlation coefficient.

Example 6.2: Calculate Karl Pearson's coefficient of correlation (using raw score method) between price and demand for the following data.

Price	17	18	19	20	22	24	26	28	30
Demand	40	38	35	30	28	25	22	21	20

Solution: In this example, correlation coefficient is calculated using following formulae:

$$r = \frac{N\sum XY - \sum X \sum Y}{\sqrt{N\sum X^2 - (\sum X)^2}\sqrt{N\sum Y^2 - (\sum Y)^2}}$$

X	**Y**	**X^2**	**Y^2**	**XY**
17	40	289	1600	680
18	38	324	1444	684
19	35	361	1225	665
20	30	400	900	600
22	28	484	784	616
24	25	576	625	600
26	22	676	484	572
28	21	784	441	588
30	20	900	400	600
$\sum X$ = 204	**$\sum Y$ = 259**	**$\sum X^2$ = 4794**	**$\sum Y^2$ = 7903**	**$\sum XY$ = 5605**

Here, N=9

On substituting the values from above table in the formulae, you have

$$r = \frac{N\sum XY - \sum X \sum Y}{\sqrt{N\sum X^2 - (\sum X)^2}\sqrt{N\sum Y^2 - (\sum Y)^2}}$$

$$r = \frac{9(5605) - (204)(259)}{\sqrt{9(4794) - (204)^2}\sqrt{9(7903) - (259)^2}}$$

$$r = \frac{50445 - 523836}{\sqrt{43146 - 41616}\sqrt{71127 - 67081}} = \frac{-2391}{\sqrt{(1530)(4046)}}$$

$$r = \frac{-2391}{2488.0474} = -0.96$$

6.8.1 CHECK YOUR PROGRESS

1. What does correlation analysis examine?
2. Name the three types of correlation.
3. What does a scatter diagram show?
4. What is the formula for Karl Pearson's correlation coefficient?
5. What range does the coefficient of correlation r lie in?

6. What assumption is made about the relationship between variables in correlation analysis?
7. What does a positive correlation mean?
8. What is the assumption of homogeneity of variance in correlation analysis?
9. Name one property of the coefficient of correlation.

6.9 COEFFICIENT OF DETERMINATION

The **coefficient of determination** (R^2) is the proportion of the variance in the dependent variable that is predicted from the independent variable. It indicates the level of variation in the given data set. It can be interpreted as the proportion of variance in 1 variable that is accounted for by the other. If the correlation coefficient between height and weight of toddlers is 0.42, then corresponding coefficient of determination (R^2) will be 0.18, suggesting that about 18% of the variability in the height of toddlers can be "explained" by the relationship with their weight. As more than 80% of the variability is yet unexplained, there must be 1 or more other relevant factors that are related to variability in height and weight of toddlers.

- The coefficient of determination is the square of the correlation (r); thus, it ranges from 0 to 1.
- The coefficient of determination is equal to the square of the correlation between the x and y variables i.e., $R^2 = (r)^2$
- The value of R^2 lies between [0,1] i.e., $0 \leq R^2 \leq 1$.
- If R^2 is equal to 0, then the dependent variable cannot be predicted from the independent variable.
- If R^2 is equal to 1, then the dependent variable can be predicted from the independent variable without any error.
- If R^2 is between 0 and 1, then it indicates the extent that the dependent variable can be predictable. If R^2 of 0.10 means, it is 10 percent of the variance in the y variable is predicted from the x variable. If 0.20 means, 20 percent of the variance in the y variable is predicted from the x variable, and so on.

Properties of Coefficient of Determination (R^2):

1. It helps to get the ratio of how a variable which can be predicted from the other one, varies.
2. If you want to check how clear it is to make predictions from the data given, you can determine the same by this measurement.
3. It helps to find Explained variation / Total Variation
4. It also lets us know the strength of the association(linear) between the variables.

5. If the value of r^2 gets close to 1, The values of y become close to the regression line and similarly if it goes close to 0, the values get away from the regression line.
6. It helps in determining the strength of association between different variables.

6.10 SPEARMAN'S RANK CORRELATION

Spearman's rank correlation is a non-parametric test that is used to measure the degree of association between two variables. It was developed by the British psychologist C.E. Spearman. It is used when the variables cannot be measured meaningfully as in the case of price, income, weight etc. Ranking may be more meaningful when the measurements of variables are suspect. Consider the situation where you are required to calculate the correlation between height and weight of students in a remote village. Neither measuring rods nor weighing scales are available. The students can be easily ranked in terms of height and weight without using measuring rods and weighing scales.

There are also situations when you are required to quantify qualities such as fairness, honesty etc. Ranking may be a better alternative to quantification of qualities. Moreover, sometimes the correlation coefficient between two variables with extreme values may be quite different from the coefficient without the extreme values. Under these circumstances rank correlation provides a better alternative to simple correlation. Rank correlation coefficient and simple correlation coefficient have the same interpretation. Its formula has been derived from simple correlation coefficient where individual values have been replaced by ranks. These ranks are used for the calculation of correlation. This coefficient provides a measure of linear association between ranks assigned to these units, not their values. It is the Product Moment Correlation between the ranks. It does not assume any assumptions about the distribution of the data and is the appropriate correlation analysis when the variables are measured on a scale that is at least ordinal. The following formula is used to calculate the Spearman's rank correlation coefficient:

$$\rho = r_s = 1 - \frac{6 \sum {d_i}^2}{n(n^2 - 1)}$$

Where,
$\rho = r_s$= Spearman rank correlation coefficient
d_i = the difference between the ranks of corresponding values X_i and Y_i (i.e., $d_i = X_i - Y_i$)
n = number of values in each data set

When the ranks are repeated the formula is

$$\rho = r_s = 1 - \frac{6\left[\sum {d_i}^2 + \frac{m_1({m_1}^2 - 1)}{12} + \frac{m_2({m_2}^2 - 1)}{12} + \cdots\right]}{n(n^2 - 1)}$$

Where, m_1, m_2, ..., are the number of repetitions of ranks and $\frac{m_1({m_1}^2-1)}{12}$, $\frac{m_2({m_2}^2-1)}{12}$, ..., are their corresponding

correction factors. This correction is needed for every repeated value of both variables. If three values are repeated, there will be a correction for each value. Every time m_1 indicates the number of times a value is repeated.

All the properties of the simple correlation coefficient are applicable here. Like the Pearson Coefficient of correlation, it lies between 1 and –1 (ρ = +1 indicates a perfect association of ranks, ρ = zero indicates no association between ranks and ρ = -1 indicates a perfect negative association of ranks. The closer ρ to zero, the weaker the association between the ranks.) However, generally it is not as accurate as the ordinary method. This is due the fact that all the information concerning the data is not utilized. The first differences of the values of items in the series, arranged in order of magnitude, are almost never constant. Usually, the data cluster around the central values with smaller differences in the middle of the array. If the first differences were constant, then r and r_k would give identical results. The first difference is the difference of consecutive values. Rank correlation is preferred to Pearson coefficient when extreme values are present. In general, r_k is less than or equal to r.

The calculation of rank correlation will be illustrated under three situations.

1. The ranks are given.
2. The ranks are not given. They have to be worked out from the data.
3. Ranks are repeated.

Case 1: When ranks are given

Example 6.3: Five persons are assessed by three judges in a beauty contest. You have to find out which pair of judges has the nearest approach to common perception of beauty**.**

	Competitors				
Judge	**1**	**2**	**3**	**4**	**5**
A	1	2	3	4	5
B	2	4	1	5	3
C	1	3	5	2	4

Solution: There are 3 pairs of judges necessitating calculation of rank correlation thrice. Also. Here n=5

The rank correlation between A and B is calculated as follows:

A	**B**	**d**	**d^2**
1	2	-1	1
2	4	-2	4
3	1	2	4
4	5	-1	1
5	3	2	4
Total			14

Since,

$$\rho = 1 - \frac{6\sum {d_i}^2}{n(n^2-1)}$$

$$\rho = 1 - \frac{6(14)}{5(5^2-1)} = 1 - \frac{84}{120} = 1 - 0.7$$

$$\rho = 0.3$$

The rank correlation between A and C is calculated as follows:

A	**C**	**d**	**d^2**
1	1	0	0
2	3	-1	1
3	5	-2	4
4	2	2	4
5	4	1	1
Total			10

Substituting these values in formula of the rank correlation, you have

$$\rho = 1 - \frac{6(10)}{5(5^2-1)} = 1 - \frac{60}{120} = 1 - 0.5$$

$$\rho = 0.5$$

Finally, the rank correlation between B and C is calculated as follows:

B	**C**	**d**	**d^2**
2	1	1	1
4	3	1	1
1	5	-4	16
5	2	3	9
3	4	-1	1
Total			28

Now, substituting these values in the formulae of rank correlation coefficient, you have

$$\rho = 1 - \frac{6(28)}{5(5^2-1)} = 1 - \frac{168}{120} = 1 - 1.4$$

$$\rho = -0.4$$

Thus, the perceptions of judges A and C are the closest. Judges B and C have very different tastes.

Case 2: When the ranks are not given

Example 6.4: You are given the percentage of marks, secured by 5 students in Economics and Statistics. Calculate the rank correlation between marks of students in two subjects.

Student	Marks in Statistics (X)	Marks in Economics (Y)
A	85	60
B	60	48
C	55	49
D	65	50
E	75	55

Solution: Since, in this example you are not given with the ranks of the student. So, you first need to rank the students according to their marks in each subject and then find the difference in the ranks.

Student	Marks in Statistics (X)	Marks in Economics (Y)	Ranking in Statistics (R_X)	Ranking in Economics (R_Y)	d	d^2
A	85	60	1	1	0	0
B	60	48	4	5	-1	1
C	55	49	5	4	1	1
D	65	50	3	3	0	0
E	75	55	2	2	0	0
Total						2

Here, n=5

Now, substitute the calculated values in the following formulae:

$$\rho = 1 - \frac{6\sum {d_i}^2}{n(n^2-1)}$$

$$\rho = 1 - \frac{6(2)}{5(5^2-1)} = 1 - \frac{12}{120} = 1 - 0.1$$

$$\rho = 0.9$$

Hence, rank correlation between the marks in Statistics and Economics is 0.9.

Case 3: When the ranks are repeated

Example 6.5: Find the Spearman's rank correlation between X and Y

X	25	45	35	40	15	19	35	42
Y	55	60	30	35	40	42	36	48

Solution: In order to compute rank correlation, you first need to rank the variables and the find the corresponding differences. Here, n=8

X	Y	Ranks of X	Ranks of Y	d	d^2
25	55	6	2	4	16
25	60	1	1	0	0
35	30	4.5	8	-3.5	12.25
40	35	3	7	-4	16
15	40	8	5	3	9
19	42	7	4	3	9
35	36	4.5	6	-1.5	2.25
42	48	2	3	-1	1
Total					65.5

Since, X has the value 35 both at the 4th and 5th rank. Hence, both are given the average rank i.e., $\frac{4+5}{2}$= 4.5th rank.

The necessary correction thus is

$$\frac{m(m^2-1)}{12} = \frac{2(2^2-1)}{12} = \frac{2(4-1)}{12} = \frac{6}{12} = \frac{1}{2}$$

Using Formulae

$$\rho = 1 - \frac{6\left[\sum {d_i}^2 + \frac{m(m^2-1)}{12}\right]}{n(n^2-1)}$$

$$= 1 - \frac{6[65.5+0.5]}{8(8^2-1)} = 1 - \frac{396}{504} = 1 - 0.786$$

$$\rho = 0.214$$

Thus, there is positive rank correlation between X and Y. Both X and Y move in the same direction. However, the relationship cannot be described as strong.

6.11 PARTIAL CORRELATION COEFFICIENT

Two variables, A and B, are closely related. The correlation between them is partialled out, or controlled for the influence of one or more variables, then it is called as partial correlation. So, when it is assumed that some other variable is influencing the correlation between A and B, then the influence of this variable(s) is partialled out for both A and B. Hence it can be considered as a correlation between two sets of residuals. As an example, consider, a simple case of correlation between A and B is partialled out for C. This can be represented as $r_{AB.C}$ which is read as correlation between A and B partialled out for C. The correlation between A and B can be partialled out for more variables as well. The range partial correlation coefficient is same as correlation coefficient i.e. -1 to 1.

For example, the researcher is interested in computing the correlation between anxiety and academic achievement controlled from intelligence. Then, correlation between academic achievement (A) and anxiety (B) will

be controlled for Intelligence (C). This can be represented as: $r_{\text{Academic Achievement(A) Anxiety (B). Intelligence (C)}}$. To calculate the partial correlation (r_p) you will need a data on all three variables. The computational formula is as follows:

$$r_P = r_{AB.C} = \frac{r_{AB} - r_{AC}\, r_{BC}}{\sqrt{(1 - r_{AC}^2)(1 - r_{BC}^2)}}$$

Example 6.6: If $r_{12} = 0.60$, $r_{13} = 0.50$ and $r_{23} = 0.45$, then calculate $r_{12.3}$, $r_{13.2}$ and $r_{23.1}$

Solution: You have,

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

$$r_{12.3} = \frac{0.60 - (0.50)(0.45)}{\sqrt{(1 - 0.50^2)(1 - 0.45^2)}} = \frac{0.60 - 0.23}{\sqrt{(1 - 0.25)(1 - 0.20)}}$$

$$r_{12.3} = \frac{0.37}{\sqrt{(0.75)(0.80)}} = \frac{0.37}{\sqrt{0.60}} = \frac{0.37}{0.77}$$

$$r_{12.3} = 0.48$$

Similarly,

$$r_{13.2} = \frac{r_{13} - r_{12}\, r_{32}}{\sqrt{(1 - r_{12}^2)(1 - r_{32}^2)}}$$

$$r_{13.2} = \frac{0.50 - (0.60)(0.45)}{\sqrt{(1 - 0.60^2)(1 - 0.45^2)}} = \frac{0.50 - 0.27}{\sqrt{(1 - 0.36)(1 - 0.20)}}$$

$$r_{13.2} = \frac{0.23}{\sqrt{(0.64)(0.80)}} = \frac{0.23}{\sqrt{0.512}} = \frac{0.23}{0.72}$$

$$r_{13.2} = 0.32$$

Also,

$$r_{23.1} = \frac{r_{23} - r_{21}\, r_{31}}{\sqrt{(1 - r_{21}^2)(1 - r_{31}^2)}}$$

$$r_{23.1} = \frac{0.45 - (0.60)(0.50)}{\sqrt{(1 - 0.60^2)(1 - 0.50^2)}} = \frac{0.45 - 0.30}{\sqrt{(1 - 0.36)(1 - 0.25)}}$$

$$r_{23.1} = \frac{0.15}{\sqrt{(0.64)(0.75)}} = \frac{0.15}{\sqrt{0.48}} = \frac{0.15}{0.69}$$

$$r_{23.1} = 0.22$$

Example 6.7: From the following data, obtain $r_{12.3}$, $r_{13.2}$ and $r_{23.1}$
Solution: To obtain the partial correlation coefficients $r_{12.3}$, $r_{13.2}$ and $r_{23.1}$

You need r_{12}, r_{13} and r_{23}, which can be obtained from following table

S. No.	X_1	X_2	X_3	$(X_1)^2$	$(X_2)^2$	$(X_3)^2$	X_1X_2	X_1X_3	X_2X_3
1	20	12	13	400	144	169	240	260	156
2	15	13	15	225	169	225	195	225	195
3	25	16	12	625	256	144	400	300	192
4	26	15	16	676	225	256	390	416	240
5	28	23	14	784	529	196	644	392	322
6	40	15	28	1600	225	784	600	1120	420
7	38	28	14	1444	784	196	1064	532	392
Total	192	122	112	5754	2332	1970	3533	3245	1917

Here, N=7

$$r_{12} = \frac{N\sum X_1X_2 - \sum X_1 \sum X_2}{\sqrt{N\sum {X_1}^2 - (\sum X_1)^2}\sqrt{N\sum {X_2}^2 - (\sum X_2)^2}}$$

$$r_{12} = \frac{7(3533) - (199)(122)}{\sqrt{(7(5754) - (199)^2)(7(2332) - (122)^2)}}$$

$$r_{12} = \frac{1307}{\sqrt{(3414)(1440)}} = \frac{1307}{2217.24} = 0.59$$

Similarly, you can compute

$$r_{13} = \frac{N\sum X_1X_3 - \sum X_1 \sum X_3}{\sqrt{N\sum {X_1}^2 - (\sum X_1)^2}\sqrt{N\sum {X_3}^2 - (\sum X_3)^2}}$$

$$r_{13} = \frac{7(3245) - (199)(112)}{\sqrt{(7(5754) - (199)^2)(7(1970) - (112)^2)}} = 0.59$$

And

$$r_{23} = \frac{N\sum X_2X_3 - \sum X_2 \sum X_3}{\sqrt{N\sum {X_2}^2 - (\sum X_2)^2}\sqrt{N\sum {X_3}^2 - (\sum X_3)^2}}$$

$$r_{23} = \frac{7(1917) - (122)(112)}{\sqrt{(7(2332) - (122)^2)(7(1970) - (112)^2)}} = -0.18$$

Now, you calculate $r_{12.3}$

You have, $r_{12} = 0.59$, $r_{13} = 0.59$ and $r_{23} = -0.18$, then

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

$$r_{12.3} = \frac{0.59 - (0.59)(-0.18)}{\sqrt{(1 - 0.59^2)(1 - 0.18^2)}} = \frac{0.696}{\sqrt{(0.65)(0.97)}}$$

$$r_{12.3} = \frac{0.696}{0.79} = 0.88$$

Similarly,

$$r_{13.2} = \frac{r_{13} - r_{12}\, r_{32}}{\sqrt{(1 - r_{12}^2)(1 - r_{32}^2)}} = \frac{0.59 - (0.59)(-0.18)}{\sqrt{(1 - 0.59^2)(1 - 0.18^2)}}$$

$$r_{13.2} = 0.88$$

And

$$r_{23.1} = \frac{r_{23} - r_{21}\, r_{31}}{\sqrt{(1 - r_{21}^2)(1 - r_{31}^2)}} = \frac{-0.18 - (0.59)(0.59)}{\sqrt{(1 - 0.59^2)(1 - 0.59^2)}}$$

$$r_{23.1} = -0.81$$

6.12 USES OF CORRELATION

Correlation can be used for varied purposes that have been discussed as follows:

1. **Validity and reliability:** Validity and reliability are important aspects of psychological testing and correlation can be used to obtain validity and reliability of a psychological test. Validity is whether a test is measuring what it is supposed to measure and reliability provides information about consistency of a test.

2. **Verification of theory**: Correlation can also be used to verify or test certain theories by denoting whether relationship exists between the variables. For example, if a theory states that there is a relationship between parenting style and resilience, the same can be tested by computing correlation for the two variables.

3. **Putting variables in groups:** Variables that show positive correlation with each other can be grouped together and variables that show negative correlation can be grouped separately based on the coefficient of correlation obtained.

4. **Computation of further statistical analysis:** Based on the results obtained after computing correlation, various statistical techniques can be used like regression. Further, correlation is also used for multivariate statistical analysis, especially for techniques like Multivariate Analysis of Variance (MANOVA), Multivariate Analysis of Covariance (MANCOVA), Discriminant Analysis, Factor analysis and so on.

5. **Based on correlation, one can decide whether or not to determine prediction:** By computing correlation, it is not possible to predict one variable based on another variable, but based on the information that two or more variables are significantly related to each other, further statistical techniques can be used to make predictions. For example, if you obtain a positive correlation between family environment and adjustment of children, then further statistical techniques can be employed to find if adjustment of children can be predicted based on family environment.

6.13 LIMITATIONS OF CORRELATION

Some of the limitations of correlation have been discussed as follows:

1. As was stated earlier, correlation will not provide any information about cause-and-effect relationship or causation.

2. The coefficient of correlation, mainly, Pearson's product moment correlation and Spearman's rank order correlation are suitable, when there is a linear relationship between the variables.

3. With regard to distributions that are discontinuous, the coefficient of correlation obtained may be overestimated or higher.

4. Sample variations can have an effect on correlation (as is also true with other statistical techniques).

5. In case of pooled sample, the correlation will be determined by relative position of the scores in X and Y dimensions or variables.

6.14 CHECK YOUR PROGRESS

1. Pearson's formula for correlation coefficient r_{xy}_____
2. Correlation Coefficient is not affected by changes in_____and_____
3. The quantity R^2 is known as_____
4. The range of Pearson's correlation coefficient is_____
5. If $r = 0$, it depicts_____association.
6. The range of partial correlation coefficient is_____
7. The estimation of correlation between yield of wheat and chemical fertilizers eliminating the effect

of pesticides and manures is an example of_____correlation.

8. When the coefficient of correlation r between the variates is 0, the rank correlation is also 0 and two variates are independent. (True/False)
9. The two variates X and Y are such that (X+Y) is always equal to 100. Therefore, the correlation is perfect and positive. (True/False)
10. The correlation coefficient can be calculated only, if both variates are in same unit. (True/False)

6.15 LET'S US SUM UP

In this lesson, you have learnt about the Pearson's correlation coefficient. Pearson's correlation coefficient is useful to calculate correlation between two relatively continuous variables. Calculation of Pearson's correlation is possible with two methods: Deviation score method and raw scores method. The coefficient obtained can be interpreted on the basis of the strength and the direction. Range, unreliability of the measurement, outliers, and curvilinearity are the factors that need to be considered while interpreting the correlation coefficient. Using correlation coefficient as a descriptive-statistics does not require assumptions. However, the use of sample correlation to estimate population parameter (ρ) requires assumptions. In this lesson you also learned about the coefficient of determination. You also learned about spearman's rank correlation coefficient which is used to see the association between two qualitative characteristics and about partial correlation coefficient which is used when we are interested in controlling for one or more variable. You can judicially be able to use this coefficient for understanding the correlation between two variables.

6.16 KEY POINTS

- **Correlation:** Degree of association between two variables.
- **Correlation Coefficient**: A number lying between -1 (Perfect negative correlation) and + i (perfect positive correlation) to quantify the association between two variables.
- **Covariance:** This is the joint variation between the variables X and Y.
- **Scatter Diagram:** An ungrouped plot of two variables, on the X and Y axes.
- **Partial Correlation:** It measures the degree of association between two random variables, with the effect of a set of controlling random variables removed.

6.17 SELF-ASSESSMENTS

1 Why is r preferred to covariance as a measure of association?

2 Can r lie outside the –1 and 1 range depending on the type of data?

3 Does correlation imply causation?

4 When is rank correlation more precise than simple correlation coefficient?

5 Does zero correlation mean independence?

6 Can simple correlation coefficient measure any type of relationship?

7 Collect the price of five vegetables from your local market every day for a week. Calculate their correlation coefficients. Interpret the result.

8 Measure the height of your classmates. Ask them the height of their bench mate. Calculate the correlation coefficient of these two variables. Interpret the result.

9 List some variables where accurate measurement is difficult.

10 Interpret the values of r as 1, –1 and 0.

11 Why does rank correlation coefficient differ from Pearson correlation coefficient?

6.18 LESSON END EXERCISE

1. Calculate the simple correlation coefficient between wing length & tail length of the following 12 birds of a particular species.

Wing length(cm) x	10.4	10.8	11.1	10.2	10.3	10.2	10.7	10.5	10.8	11.2	10.6	11.4
Tail length (cm) y	7.4	7.6	7.9	7.2	7.4	7.1	7.4	7.2	7.8	7.7	7.8	8.3

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

X	65	66	57	67	68	69	70	72
Y	67	56	65	68	72	72	69	71

3. The following data relates to the yield in grams(y) and the matured pods (x) of 10 groundnut plants. Work out the correlation coefficient and comment on their relationship.

X	14	34	20	16	11	11	20	17	22	17
Y	16	40	21	18	14	13	20	35	17	27

4. Find the persons coefficient of correlation between price and demand from the following data.

Price	11	13	15	17	18	19	20
Demand	30	29	24	24	21	18	15

5. The scores for nine students in physics and math are as follows:

Physics	35	23	47	17	10	43	9	6	28
Mathematics	30	33	45	23	8	49	12	4	31

Compute the student's ranks in the two subjects and compute the Spearman rank correlation.

6. Calculate a Spearman rank correlation coefficient of the following data:

Statistics	20	23	8	29	14	12	11	20	17	18
Science	20	25	11	24	23	16	12	21	22	26

7. From the following data, obtain $r_{12.3}$, $r_{13.2}$ and $r_{23.1}$

X_1	40	44	42	45	40	45	40	40	42	41
X_2	18	20	26	24	20	25	23	19	18	16
X_3	52	51	50	48	47	52	50	51	49	50

8. From the following data, obtain $r_{12.3}$, $r_{13.2}$ and $r_{23.1}$

X_1	12	7	9	15	14	18	18
X_2	10	7	16	15	8	12	10
X_3	7	9	4	8	10	12	8

9. If r12 = 0.87, r13 = 0.82 and r23 = 0.62, compute partial correlation $r_{12.3}$, $r_{13.2}$ and $r_{23.2}$

10. In trivariate distribution, r_{12} = 0.8, r_{23} = 0.6 and r_{13} = 0.6. Compute $r_{12.3}$

1.19 SUGGESTED READINGS

- Gupta, S.C. and Kapoor, V.K. (2017): "*Fundamental of Mathematical Statistics".* S Chand Publication.
- Sharma, A.K. (2005): "*Correlation and Regression*". Discovery Publishing House.
- Draper, N. and H. Smith, (1966). "*Applied Regression Analysis*". John Wiley: New York.
- Wilcox, R. R. (1996). "*Statistics for Social Sciences*". San Diego: Academic Press.

LESSON : 7

TESTING HYPOTHESES: METHODS, TECHNIQUES, AND REAL-WORLD APPLICATIONS

Structure

7.1. INTRODUCTION

Many a time, we strongly believe some results to be true. But after taking a sample, we notice that one sample data does not fully support the result. The difference is due to i) the original belief being wrong, or ii) the sample being slightly one sided. In this unit and the next, we shall study a class of problems where the decision made by a decision maker depends primarily on the strength of the evidence thrown up by a random sample drawn from a population. We can elaborate this by an example where the purchase manager of a machine tool making company has to decide whether to buy casting from a new supplier or not. The new supplier claims that his casting has higher hardness than those of the competitors. If the claim is true then it would be in the interest of the company to switch from the existing suppliers to the new suppliers because of the higher hardness, all other conditions being similar. However, if the claim is not true, the purchase manager should continue to buy from the existing suppliers. He needs a tool which allows him to test such a claim.

Testing of hypothesis provides such a tool to the decision maker. If the purchase manager were to use this tool, He would ask the new supplier to deliver a small number of castings. The sample of castings will be evaluated and based on the strength of the evidence produced by the sample, The purchase manager will accept or reject the claim of the new supplier and accordingly make his decision. The claim made by the new supplier is the hypothesis that needs to be tested and a statistical procedure which allows us to perform such a test is called *testing of hypothesis*.

What is the hypothesis?

Setting up and testing hypothesis is an essential part of statistical hypothesis. A hypothesis, or more specifically a statistical hypothesis, is some statement about a population parameter or about a population distribution. If the population is large, there is no way of analyzing the population or of testing the hypothesis directly. Instead, the hypothesis is tested on the basis of the outcome of random sample. Our hypothesis for the example situated in 7.1 could be that the mean hardness of castings provided by the new supplier is less than or equal to 20, where 20 is the mean hardness of castings supplied by the existing suppliers. The hypotheses are often statements about population parameters like expected value and variance. For example, H_0 might be the statement that the expected value of the height of the 10 years old boys in the Indian population, is not different from that of 10 years old girls.

A Two-action Decision Problems

That season problem faced by the purchase manager in 7.1 above has two alternative courses of action- either to buy from the new supplier or not to buy from the new supplier. The alternative chosen depends on whether the claim made by the new supplier is accepted or rejected. Now, the claim made by the new supplier can be formulated as a statistical hypothesis as has been done in 7.1 above. Therefore, the decision

made or the alternative chosen depends primarily on whether the hypothesis is accepted or rejected.

7.2. LEARNING OBJECTIVES

After reading this lesson, students should be able to:

- understand the meaning of statistical hypothesis and the important basic concepts.
- appreciate the importance of the significance level of significance and the P-value of the test.
- learn the steps involved in conducting a test of hypothesis
- examples on real-life for understanding the concepts.

7.3 SOME BASIC CONCEPTS

For the purpose of decision making, one should be familiar with some basic terminologies**:**

Population: Before discussing about hypothesis testing procedures, we have to first understand about the term '*population*'. In a statistical analysis, if we are interested in the evaluation of the general magnitude and also the study of the variation relating to individuals belonging to a group with respect to one or more characteristics. This group of individuals under consideration is called population or universe. Thus, in statistics, population is an aggregate of objects whether animate or inanimate, under study. The population may be finite or infinite.

Finite Population is that population in which number of observations are finite or countable. For example, books in a library, number of students in a class and so on. And, *Infinite Population* is that population which is uncountable. For Example, number of stars in the sky, algae on earth etc.

It is obvious that complete enumeration of the population for any statistical investigation is rather impracticable. For example, if we want to study the average income of the people in India, then in that case, we will have to enumerate all the individuals who are earning in the country, which is impracticable or very difficult task.

Complete enumeration is also not possible when the population is infinite. If in case the inspection is destructive e.g., inspection of crackers, explosive materials, etc., 100% inspection, though possible, is not at all desirable. Also, even if the inspection is not destructive or finite, 100% inspection is not taken because of causes like time factor, financial and administrative implications etc., and we take the help of sampling.

Population size: The number of units in the population is called its population size. It is being represented by N.

Sample. A part of the population is called a *sample.* In other words, a finite subset of statistical individuals

in a population is called a sample. The number of individuals in a sample is called the sample size.

To determining the population characteristics, instead of enumerating the entire population, a sample individual only are observed. Then the characteristics of sample are used to estimate the population. For example, on observing the sample of a rice, we arrive at a decision whether to purchase or to reject that whole rice. In such approximations there is a chance of error involved, and is known as sampling error. Sampling error is inherent and unavoidable in any sampling scheme. But if we consider time and cost, sampling results are in considerable gains which not only help in interpretation of results but also in the subsequent handling of the data.

We use sampling in our day-to day life. For example, in a grocery shop we assess the quality of wheat, sugar, rice or any other commodity by taking a handful of it from the bag and then we decide whether to purchase it or not. A housewife tests the cooked products normally by taking a part of it and come to know if they are properly cooked or not.

It may be noted here that a sample should be a true representative of the population.

Sample Size: The total number of units in the sample is termed as sample size and is denoted by n. When the sample size is less than 30 ($n<30$) we call sample as small sample otherwise (i.e. for $n \geq 30$) the sample in hand is known as large sample.

Parameter: To avoid the verbal confusion about the statistical constants about the population, like mean μ and variance σ^2, population correlation co-efficient (r) etc, we use the term parameter. It is a function of population values

Statistic: To avoid the verbal confusion about the statistical constants about the sample, like mean $\bar{x}$ and variance s^2, we use the term statistic. It is a function of sample values only.

Sampling Error: Sampling error has its origin in sampling and arise due to the fact that only part of population is taken into consideration. If the estimated value of statistics is equal to parameter then there is no sampling error and if the estimated value of statistics is not equal to parameter then there is a sampling error.

Level of Significance: The size of type I error (α) is also known as level of significance abbreviated as l.o.s. Generally, we take $\alpha = 0.5$ or 0.01. Here $\alpha = 0.05$ indicates that there is possibility of taking wrong decision in 5% cases.

Degree of freedom: The number of observations in a sample less the number of constraints imposed upon them is called degrees of freedom. It is abbreviated as *d.f.* and denoted by Greek letter ν (nu). For instance, if a sample is of size n and one constraint is imposed on these n observations then degrees of freedom will be $\nu = n - 1$ for test statistic.

Simple and Composite Hypotheses

In general sense, if a hypothesis specifies only one value or exact value of the population parameter then it is known as simple hypothesis. And if a hypothesis specifies not just one value but a range of values that the population parameter may assume is called a composite hypothesis. In other words, a hypothesis which completely specifies parameter(s) of a theoretical population (probability distribution) is called a *simple hypothesis* and if does not specify all the parameters is called *composite hypothesis*.

For example, (i)a customer of motorcycle wants to test whether the claim of motorcycle of certain brand gives the average mileage 60 km/liter is true or false. As in this examples, Customer of motorcycle may write the claim or postulate the hypothesis "the motorcycle of certain brand gives the average mileage 60 km/liter". Here, we are concerning the average mileage of the motorcycle so let μ represents the average mileage then our hypothesis becomes $\mu = 60$ km / liter.

(ii) a doctor wants to test whether new medicine is really more effective for controlling blood pressure than old medicine. In this example, doctor may write the claim or postulate the hypothesis "the new medicine is really more effective for controlling blood pressure than old medicine." Here, we are concerning the average effect of the medicines so let μ_1 and μ_2 represent the average effect of new and old medicines respectively on controlling blood pressure then our hypothesis becomes $\mu_1 > \mu_2$.

As in the above examples, the hypothesis postulated in (i) $\mu = 60$ km/liter is simple hypothesis because it gives a single value of parameter ($\mu = 60$), whereas the hypothesis postulated in (ii) $\mu_1 > \mu_2$ is composite hypothesis because it does not specify the exact average value.

Null and Alternative Hypotheses

As stated earlier, a hypothesis is a statement about a population parameter or about a population distribution. A hypothesis is a statement supposed to be true till it is proved false. In any testing of hypothesis problem, we are faced with a pair of hypotheses such that one and only one of them is always true. One of this pair is called the Null Hypothesis and the other one is called the alternative hypothesis. The Null Hypothesis is represented as H_0. Null hypothesis is the hypothesis of no difference. *According to Prof R.A. Fisher, Null hypothesis is the hypothesis which is tested for possible rejection under the assumption that it is true.*

For example, in case of a single statistic, H_0 will be that the sample statistic does not differ significantly from the hypothetical parameter value and in the case of two statistics, H_0 will be that the sample statistics do not differ significantly.

Any hypothesis which is complementary to the Null Hypothesis is called an alternative hypothesis usually denoted by H_1. For example, if we want to taste the null hypothesis that the population has a specified mean μ_0, (say), i.e., H_0: $\mu = \mu_0$then the alternative hypothesis could be:

i) $H_1: \mu \neq \mu_0$ (i.e., $\mu > \mu_0$ or $\mu < \mu_0$) ii) $H_1: \mu < \mu_0$ iii) $H_1: \mu > \mu_0$

The alternative hypothesis in one is known as two tailed alternative and the alternative in ii) and iii) are known as right tailed and left tailed alternatives respectively. The setting of alternative hypothesis is very important since it enables us to decide whether we you have to use a single tailed (left or right) or two tailed.

For example, if it is assumed that the mean of the weights of the population of college is 60 kg and if the population mean is represented by μ, we set up our hypothesis, as follows:

Null hypothesis will be: The mean of the population is 55 kg, i.e. $H_0: \mu = 55\ kg$ and the alternative hypothesis will be: $H_1: \mu < 55\ kg$ or $H_1: \mu > 55\ kg$

It is clear that both H_0 and H_1 cannot be true at the same time and also that one of them will always be true. At the end of our testing procedure, if we come to the conclusion that H_0 should be rejected, this also amounts to saying that H_1 should be accepted and vice versa.

The final conclusion once the test has been carried out is always given in terms of the null hypothesis. We either 'reject H_0 in favour of H_1' or 'do not reject H_0'; we never conclude 'reject H_1' or even 'accept H_1'. If we conclude 'do not reject H_0', this does not necessarily mean that the null hypothesis is true. It only suggests that there is not sufficient evidence against H_0 in favour of H_1; rejecting the null hypothesis then suggests that the alternative hypothesis may be true.

Hypothesis testing refers to the process of using statistical analysis to determine if the observed differences between two or more samples are due to random chance (as stated in the null hypothesis) or to true differences in the samples (as stated in the alternate hypothesis). A null hypothesis (H_0) is a stated assumption that there is no difference in parameters (mean, variance) for two or more populations. The alternate hypothesis (H_1) is a statement that the observed difference or relationship between two populations is real and not the result of chance or an error in sampling. Hypothesis testing is the process of using a variety of statistical tools to analyze data and, ultimately, to fail to reject or reject the null hypothesis. From a practical point of view, finding statistical evidence that the null hypothesis is false allows you to reject the null hypothesis and accept the alternate hypothesis.

7.3.1. CRITICAL REGION

A critical region, also known as the rejection region, is a set of values for the test statistic for which the null hypothesis is rejected, i.e. if the observed test statistic is in the critical region then we reject the null hypothesis and accept the alternative hypothesis. In other words, the critical region is the region of values that corresponds to the rejection of the null hypothesis at some chosen probability level. If ω is the critical region and if $t=t\ (x_1,x_2,\ldots,x_n)$ is the value of the statistic based on the random sample of size n, then

$$P(t \in \omega | H_0) = \alpha, \qquad P(t \in \bar{\omega} | H_1) = \beta$$

where, $\bar{\omega}$ is the critical region and is complementary to $\bar{\omega}$, is called *acceptance region.*

Also, $\omega \cup \bar{\omega} = S$ and $\omega \cap \bar{\omega} = \phi$

The critical values are tabulated and thus obtained from the appropriate table. If the absolute value of the statistic is larger than the tabulated value, then the data points belong to the critical region.

7.3.2. CONFIDENCE LIMITS AND LEVEL OF SIGNIFICANCE

The limits (or range) within which the hypothesis should lie with specified probabilities are called Confidence limits or fiduciary limits. Fixing the limits totally depend upon the accuracy desired. Generally, the limits are fixed such that the probability that the difference will exceed the limit is 0.05 or 0.01. These levels are called level of significance and are expressed as 5% or 1% level of significance.

In using the hypothesis-testing procedure to determine if the null hypothesis should be rejected, the person conducting the hypothesis test specifies the maximum allowable probability of making a Type-I error, called the *level of significance* for the test. In other words, we can say that the Probability 'α' that a random value of the statistic *t* belongs to critical region is called *level of significance*. It is the probability of Type-I error is called level of significance. Common choices for the level of significance are $\alpha = 0.05$ and $\alpha = 0.01$. Although most applications of hypothesis testing control the probability of making a Type I error, they do not always control the probability of making a Type-II error.

7.3.3. P-VALUE

A concept known as the p-value provides a convenient basis for drawing conclusions in hypothesis-testing applications. The p-value is a measure of how likely the sample results are, assuming the null hypothesis is true; the smaller the p-value, the lesser likely are the sample results reliable. If the p-value is less than α, the null hypothesis can be rejected; otherwise, the null hypothesis cannot be rejected. The p-value is often called the observed level of significance for the test. A P-value of 0.05 or lower generally considered statistically significant. P-value can serve as an alternative to (or in addition to) pre-selected confidence levels for hypothesis testing.

The p-value also depends on the type of the test. If test is one-tailed then the p-value is defined as:

For right-tailed test: p-value = *P[Test Statistic (T) ≥ observed value of the test statistic]*

For left-tailed test: p-value = *P[Test Statistic (T) ≤ observed value of the test statistic]*

Procedure of taking the decision about the null hypothesis on the basis of p-value: To take the decision about the null hypothesis based on p-value, the p-value is compared with level of significance (á) and if p-value is equal or less than a then we reject the null hypothesis and if the p-value is greater than a we do not reject the null hypothesis.

7.4. CHECK YOUR PROGRESS 1

1. What do you mean by null hypothesis and alternative hypothesis?

   ……

   ……

2. What is meant by simple and composite hypothesis?

   ……

   ……

3. How do you define critical region and acceptance region?

   ……

   ……

4. What is the procedure for taking decision about the null hypothesis on the basis of P value?

   ……

   ……

7.5. ONE-TAILED AND TWO-TAILED TESTS

In any test, the critical region is represented by a portion of the area under the probability curve after sampling distribution of the test statistic.

A test of a statistical hypothesis where the alternative hypothesis is one-tailed (right-tailed or left-tailed) is called 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.

In the right-tailed test ($H_1: \mu > \mu_0$), the critical reason lies entirely in the right tail of the sampling distribution of $\bar{x}$, while for the left-tailed test $H_1: \mu < \mu_0$, the critical region is entirely in the left tail of the distribution.

A test of statistical hypothesis where the alternative hypothesis is two-tailed 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 two tailed test and in such a case the critical reason is given by the portion of the area lying in both side of the probability curve of the test statistic.

In a particular problem, whether one-tailed or two-tailed test is to be applied depends entirely on the nature of the alternative hypothesis. If the alternative hypothesis is two-tailed, we apply two-tailed test and if alternative hypothesis is one-tailed, we apply one-tailed test.

For example, suppose there are two population Brands of bulbs, one manufactured standard process (with mean life μ_1) and the other manufactured by some new technique (with mean life μ_2). If we want to test if the bulbs differ significantly, then our null hypothesis is $H_0: \mu_1 = \mu_2$and the alternative will be $H_1: \mu_1 \neq \mu_2$, thus giving us two-tailed test. If you want to test if the bulbs produced by the new process have higher average life than those produced by standard process, then we have $H_0: \mu_1 = \mu_2$ and $H_1: \mu_1 < \mu_2$, thus giving us left-tailed test. Similarly, for testing if the product of new process is inferior to that of standard process, then we have $H_0: \mu_1 = \mu_2$ and $H_1: \mu_1 > \mu_2$, thus giving us right-tailed test. Thus, the decision about applying a two-tailed test or a single tailed test will depend on the problem under study.

7.6. CRITICAL VALUES OR SIGNIFICANT VALUES

The critical value at a certain significance level can be thought of as a cut-off point. It is also called significant value. The value of test statistic which separates the critical region and the acceptance region is called the critical value or significant value. If a test statistic on one side of the critical value result in accepting the null hypothesis, a test statistic on the other side will result in rejecting the null hypothesis. It depends upon:

i) The level of significance used, and

ii) The alternative hypothesis whether it is two-tailed or single-tailed.

Paris has been pointed out earlier, for large samples, the standardized variable corresponding to the statistic *t*, viz.,

$$Z = \frac{t - E(t)}{S.E.(t)} \sim N(0,1) \quad \ldots (*)$$

Asymptotically as $n \to \infty$. The value of Z given (*) by under the null hypothesis is known as test statistic. The critical value of the test statistic at level of significance α for two tailed tests is given by z_α, where, z_α is determined by the equation:

$$P(|Z| > z_\alpha) = \alpha \quad \ldots (a)$$

i.e., z_α is the value so that the total area of the critical region on both tails is α. Since normal probability curve is a symmetrical curve, from (a), we get

$P(Z > z_\alpha) + P(Z < -z_\alpha) = \alpha \Rightarrow P(Z > z_\alpha) + P(Z > z_\alpha) = \alpha$(By symmetry)
$2P(Z > z_\alpha) = \alpha \Rightarrow P(Z > z_\alpha) = \alpha/2$

In other words, 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 $\alpha/2$.

In case of single-tail alternatives, the critical value z_α is determined so that total area to the right of it (for right-tailed test) is α and for left-tailed test total area to the left of $(-z_\alpha)$ is α, i.e.,

For right-tailed test: $P(Z > z_\alpha) = \alpha$ … (b)
For left-tailed test: $P(Z > -z_\alpha) = \alpha$ … (c)

Thus, the significant or critical value of Z for a single-tailed test (left or right) at level of significance 'a' is same as the critical value of Z for the two-tailed test at level of significance '2α'.

We give below, the critical value of Z at commonly used level of significance for both two-tailed and single-tailed tests. These values have been obtained from equations (a), (b), (c), On using the Normal Probability Tables.

Critical Value (z_α)	Level of Significance (α)		
	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	$Z_\alpha = -2.33$	$Z_\alpha = -1.645$	$Z_\alpha = -1.28$

Note: If *n* is small, then the sampling distribution of the test statistic Z will not be normal and in that case, we can't use the above significant values which have been obtained from normal probability curves. In this case, viz., *n* small (usually less than 30), we use the significant values based on the exact distribution of the statistic Z [defined in (**)], which turns out of be *t, F* and chi-square. These significant values have been tabulated for different values of *n* and α.

7.7. ERRORS IN TESTING OF HYPOTHESIS

Since our conclusions are based on the evidence produced by the sample and since variations from one sample to another can never be eliminated until the sample is as large as the population itself, it is possible that the conclusion drawn is incorrect which leads to an error. In testing any hypothesis, we get only two results either we accept or we reject it. We did not know whether it is true or false. Hence four possibilities may arise:

i) The hypothesis is true but test rejects it (Type-I Error).

ii) The hypothesis is false but test accepts it (Type-II Error).

iii) The hypothesis is true and test accepts it (correct decision).

iv) The hypothesis is false and test rejects it (correct decision).

Table 7.1 gives a summary of possible results of any hypothesis test

True state of nature		Decision	
		Reject H_0	*Do not reject H_0*
	H_0	Type-I Error	Correct Decision
	H_1	Correct Decision	Type-II Error

A Type-I error corresponds to rejecting H_0 when H_0 is actually true, and a Type-II error corresponds to accepting H_0 when H_0 is false. The probability of making a Type-I error is denoted by α, and the probability of making a Type-II error is denoted by β.

Type-I Error

In a hypothesis test, a Type-I error occurs when the null hypothesis is rejected when it is in fact true; that is, H_0 is wrongly rejected. For example, in a clinical trial of a new drug, the null hypothesis might be that the new drug is no better, on average, than the current drug; that is H_0: there is no difference between the two drugs on average. A Type-I error would occur if we concluded that the two drugs produced different effects when in fact there was no difference between them.

A Type-I error is often considered to be more serious, and therefore more important to avoid, than a Type-II error. The hypothesis test procedure is therefore adjusted so that there is a guaranteed 'low' probability of rejecting the null hypothesis wrongly; this probability is never 0. This probability of a Type-I error can be precisely computed as,

$$P(\text{Type-I error}) = \text{significance level} = \alpha$$

$\alpha = P[X \in \omega | H_0]$ where $X = (X_1, X_2, \ldots, X_n)$ is a random sample and ω is the rejection region
and, $1-\alpha = 1-P[\text{Reject } H_0 / H_0 \text{ is true}]$

$= P[\text{Do not reject } H_0 / H_0 \text{ is true}]$

$= P[\text{Correct decision}]$

The $(1-\alpha)$ is the probability of correct decision and it correlates to the concept of $100(1-\alpha)\%$ confidence interval used in estimation.

If we do not reject the null hypothesis, it may still be false (a Type-II error) as the sample may not be big enough to identify the falseness of the null hypothesis (especially if the truth is very close to hypothesis). For any given set of data, Type-I and Type-II errors are inversely related; the smaller the risk of one, the higher the risk of the other. A Type-I error can also be referred to as an error of the first kind or *Rejection Error*.

Type-II Error

In a hypothesis test, a Type-II error occurs when the null hypothesis, H_0, is not rejected when it is in

fact false. For example, in a clinical trial of a new drug, the null hypothesis might be that the new drug is no better, on average, than the current drug; that is H_0: there is no difference between the two drugs on average. A Type-II error would occur if it was concluded that the two drugs produced the same effect, that is, there is no difference between the two drugs on average, when in fact they produced different ones. A Type-II error is frequently due to sample sizes being too small.

The probability of a Type-II error is symbolized by β and written as:

P (Type-II error) = β (but is generally unknown)

A Type-II error can also be referred to as an error of the second kind or *Acceptance Error*.

$\beta = P[Do\ not\ reject\ H_0\ when\ H_0\ is\ false]$

$= P[Do\ not\ reject\ H_0\ when\ H_1\ is\ true]$

$= P[Do\ not\ reject\ H_0\ /\ H_1\ is\ true]$

$= P[X \in \bar{\omega} | H_1]$ where, $\bar{\omega}$ is the non-rejection region.

and,

$1-\beta = 1-P[Do\ not\ reject\ H_0\ /\ H_1\ is\ true]$

$= P[Reject\ H_0\ /\ H_1\ is\ true]$

$= P[Correct\ decision]$

The (1-β) is the probability of correct decision and also known as "power of the test". Since it indicates the ability or power of the test to recognize correctly that the null hypothesis is false, therefore, we wish a test that yields a large power. We say that a statistical test is ideal if it minimizes the probability of both types of errors and maximizes the probability of correct decision. But for fix sample size, α and β are so interrelated that the decrement in one result into the increment in other. So, minimization of both probabilities of type-I and type-II errors simultaneously for fixed sample size is not possible without increasing sample size. Also, both types of errors will be at zero level (i.e. no error in decision) if size of the sample is equal to the population size. But it involves huge cost if population size is large. And, it is not possible in all situations such as testing of blood. Depending on the problem in hand, we have to choose the type of error which has to minimize. For this, we have to look at a situation, suppose there is a decision-making problem and there is a rule that if we make type-I error, we lose10 rupees and if we make type-II error we lose 1000 rupees. In this case, we try to eliminate the type-II error, since it is more expensive. In another situation, suppose the Delhi police arrests a person whom they suspect is a murderer. Now, policemen have to test hypothesis:

H_0: Arrested person is not criminal.

H_1: Arrested person is a criminal.

The type-I error is α = P [Reject H_0 when it is true]

That is, suspected person who is actually an innocent will be sent to jail when H_0 rejects, although H_0 being a true.

Consider another example, suppose a manufacturer produces some type of articles of good quality. A purchaser by chance selects a sample randomly. It so happens that the sample contains many defective articles and it leads the purchaser to reject the whole product. Now, the manufacturer suffers a loss even though he has produced a good article of quality. Therefore, this Type-I error is called *"Producers risk"*. On the other hand, if the entire lot is accepted on the basis of a sample and the lot is not really good, the consumers are put in loss. Therefore, this Type-II error is called the *"Consumers risk"*.

Generally, strong control on α is necessary. It should be kept as low as possible. In test procedure, we prefix it at very low level like α = 0.05 (5%) or 0.01 (1%).

7.8. PROCEDURE FOR FORMULATING HYPOTHESES AND STATING CONCLUSIONS

i) State the null hypothesis, H_0 and it will always contain an equality sign. In testing of hypothesis, we make our predictions about the population parameters.

ii) State the alternative hypothesis H_1 which is opposite of Null hypothesis and contains either 'not equals to', 'less than' or 'greater than' signs depending upon the problem at hand.

iii) Level of significance or alpha level for the hypothesis test. This is represented by α which is the probability used to define the very unlikely sample outcomes, if the null hypothesis is true.

iv) Choose the appropriate test statistic.

v) Set the criteria for a decision.

vi) If the sample evidence supports the alternative hypothesis, the null hypothesis will be rejected and the probability of having made an incorrect decision (when in fact $H_0$is true) is α, a quantity that can be manipulated to be as small as the researcher wishes.

vii) If the sample does not provide sufficient evidence to support the alternative hypothesis, then conclude that the null hypothesis cannot be rejected on the basis of your sample. In this situation, you may wish to collect more information about the phenomenon under study.

7.9. CHECK YOUR PROGRESS 2

1. What is the procedure for taking decision about the null hypothesis on the basis of P value?

   ……

   ……

2. What do you understand by one tailed and two tailed tests?

   ……

   ……

3. Define type I error & type II error? Which error is most powerful?

   ……

   ……

4. What is the procedure for formulating hypotheses and stating conclusions?

   ……

   ……

5. What is significant value?

   ……

   ……

7.10. LET US SUM UP

Setting up and testing hypothesis is an essential part of statistical hypothesis. A statistical hypothesis, is some statement about a population parameter. If the population is large, there is no way of analyzing the population or of testing the hypothesis directly. Instead, the hypothesis is tested on the basis of the outcome of random sample.

Testing of hypothesis provides such a tool to the decision maker. Many a time, we strongly believe some results to be true. But after taking a sample, we notice that one sample data does not fully support the result. The difference is due to i) the original belief being wrong, or ii) the sample being slightly one sided. The decision made or the alternative chosen depends primarily on whether the hypothesis is accepted or rejected.

7.11. KEY POINTS

- **Parameter**: The statistical constants about the population, like mean μ and variance σ^2, population

correlation co-efficient (r) etc.

- **Statistic:** The statistical constants about the sample, like mean $\bar{x}$ and variance s^2.
- **Sampling error:** A statistic gives an estimate to the parameter. The difference between the value of a statistic and the value of the corresponding parameter is called sampling error.
- **Hypothesis:** The statistical statement regarding the form of a parameter of the distribution is called hypothesis.
- **Null hypothesis:** It is hypothesis of no difference or no effect denoted by Ho.
- **Alternative Hypothesis**: An alternative statement to null hypothesis that we believe is true is called alternative hypothesis. It is denoted by H_1.

7.12. SELF-ASSESSMENT QUESTIÕNS

1. What do you understand by sampling distribution?
2. What is meant by a statistical hypothesis? What are two types of errors of decision that occur in testing of hypothesis?
3. What are the steps for testing of hypothesis?
4. What is critical region? What do you understand by critical value?
5. What do you understand by small sample and large sample?
6. What is the utility of P-value?
7. How do you describe Type I error and Type II error?
8. Write down the procedure to formulate hypothesis in decision making?
9. Give the rule for acceptance and rejection of hypothesis based on P value.

7.13. LESSON END EXERCISE

1. A company has replaced its original technology of producing electric bulbs by CFL technology. The company manager wants to compare average life of bulbs manufactured by original technology and new technology CFL. Write null and alternative hypothesis. Also, sat about the one tailed and two tailed tests.
2. If we have null and alternative hypotheses as

$$H_0: \theta = \theta_0 \text{ and } H_0: \theta \neq \theta_0$$

Then corresponding test will be

i) Left-tailed test ii) Right-tailed test iii) Two-tailed test

Write about the correct option.

3. The test whether one tailed or two tailed depends on

i) Null hypothesis ii) Alternative hypothesis iii) Neither null nor Alternative hypothesis iv) both null and alternative hypothesis

7.14. SUGGESTED READINGS

- Goon, A.M., Gupta, M.K. and Dasgupta, B. (1968). Fundamentals of Mathematical Statistics, Vol II, World Press, Kolkata.
- Gupta, S.C., & Kapoor, V.K. (2020). Fundamental of Mathematical Statistics. Sultan Chand and Sons.
- Gupta, S.C., & Kapoor, V.K. (2020). Fundamental of Applied Statistics. Sultan Chand and Sons.
- Gupta, S.P. (2021). Statistical Methods. 46th ED., Sultan Chand and Sons.

LESSON : 8

ADVANCED STATISTICAL METHODS: T-TESTS, ANOVA, CHI-SQUARE, AND F-TESTS FOR HYPOTHESIS TESTING

Structure

8.1. INTRODUCTION

Sometimes complete enumeration of the population for any statistical investigation is impracticable. For instance, if we want to study the average income of the people in India, then in that case, we will have to enumerate all the individuals who are earning in the country, which is very difficult task.

Complete enumeration is also not possible when the population is infinite. If in case the inspection is destructive e.g., inspection of crackers, explosive materials, etc., 100% inspection, though possible, is not at all desirable. Also, even if the inspection is not destructive or finite, 100% inspection is not taken because of causes like time factor, financial and administrative implications etc., and we take the help of sampling.

To determining the population characteristics, instead of enumerating the entire population, a sample individual only are observed. Then, the characteristics of sample are used to estimate the population. For example, on observing the sample of a rice, we arrive at a decision whether to purchase or to reject that whole rice. In such cases where sample is used there is a chance of error involved, and is known as sampling error. Sampling error is inherent and unavoidable in any sampling scheme. But if we consider time and cost, sampling results are in considerable gains which not only help in interpretation of results but also in the subsequent handling of the data.

We use sampling in our day-to day life for example, in a grocery shop we check the quality of wheat, sugar, rice or any other commodity by taking a handful of it from the bulk and then we decide whether to purchase it or not. A housewife tests the cooked products normally by taking a part of it and come to know if they are properly cooked or not.

8.2. LEARNING OBJECTIVES

Upon successful reading this lesson, students should be able to:

- Understand the statistical tests for small sample as well as for large sample.
- learn the assumptions for statistical tests involved in conducting statistical tests.
- examples on real-life for understanding the concepts.

8.3. TYPES OF SAMPLING

There are different types of sampling techniques used in Statistics; some are given below.

1. Non-Probability sampling

In non-probability sampling, the sample units are selected by non-random method. In this method, all population members do not have equal chance of being selected in the sample. Some of the common types of non-probability sampling are given below:

(i) Subjective sampling

(ii) Purposive sampling

(iii) Judgement sampling

(iv) Quota Sampling

(v) Snowball sampling

2. Probability sampling

Probability sampling uses random selection of sample units. It is the sampling technique in which the samples taken from a large population are selected on the basis of probability theory. Some of the probability sampling techniques are:

(i) Simple Random sampling

(ii) Stratified sampling,

(iii) Systematic sampling.

(iv) Cluster sampling

8.4. STANDARD ERROR

Standard error is defined as the standard deviation of the sampling distribution of a statistic. It is abbreviated as *S.E.* For large samples, the standard error of some of the well-known statistics are given below:

S.No.	Statistic	Standard Error
1	Sample mean, $\bar{x}$	$\sigma/\sqrt{n}$
2	Observed sample proportion, 'p'	$\sqrt{PQ/n}$
3	Sample S.D., *'s'*	$\sqrt{\sigma^2/2n}$
4	Sample variance, 's^2'	$\sigma^2\sqrt{2/n}$
5	Sample correlation coefficient, 'r'	$(1-\rho^2)/\sqrt{n}$, ρ being the population correlation coefficient.
6	Difference of two means, $(\bar{x}_1 - \bar{x}_2)$	$\sqrt{\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}}$
7	Difference of two *s.d.'s,* $(s_1 - s_2)$	$\sqrt{\frac{{\sigma_1}^2}{2n_1} + \frac{{\sigma_2}^2}{2n_2}}$
8	Difference of two sample proportions, $(p_1 - p_2)$	$\sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}$
9	Sample median	$1.25331\ \sigma/\sqrt{n}$
10	Sample quartile	$1.36263\ \sigma/\sqrt{n}$

Significance of Standard Error:

In large sample theory, for testing the hypothesis, *S.E.* plays a fundamental role. It enables us to find the probable limits within which the parameter may be expected to lie.

Suppose, t is any statistic, then for testing the large samples, the test statistic is given by:

$$Z = \frac{t - E(t)}{\sqrt{V(t)}} \sim N(0,1)$$

Therefore, for large samples,

$$Z = \frac{t - E(t)}{S.E.(t)} \sim N(0,1)$$

The reciprocal of the *S.E.* is taken as the precision of the statistic i.e.,

$$S.E.(p) = \sqrt{PQ/n} \text{ and } S.E.(\bar{x}) = \sigma/\sqrt{n}$$

Therefore, from above expressions, the *S.E.* of sample proportion and sample mean vary inversely as the square root of the sample size 'n'.

The probable limits for population standard deviation are given by $s \pm 3\sqrt{\sigma^2/2n}$ and population mean are $\bar{x} \pm 3\,\sigma/\sqrt{n}$.

Thus, we can say that, *S.E.* is inversely proportional to sample size. *S.E.* can be reduced by increasing the sample size but it may result in increase in time, cost and labour etc.

8.5. STATISTICAL TESTS

Some important statistical tests are discussed below:

8.5.1. CHI-SQUARE TEST

Chi-Square variate is defined as the square of standard normal variate (*pronounced as Kai-square*) with one degrees of freedom*(d.f.)*

Thus, if $X \sim N(\mu, \sigma^2)$, then $Z = \frac{X-\mu}{\sigma} \sim N(0,1)$

therefore, $Z^2 = \left(\frac{X-\mu}{\sigma}\right)^2$ is a Chi-Square variate with 1 *d.f.*

In general, if X_i; $i=1,2,\ldots,n$ are n independent normal variates with mean μ_i and variance ${\sigma_i}^2$, then

$\chi^2 = \sum_{i=1}^{n} \left(\frac{X_i-\mu_i}{\sigma_i}\right)^2$, is a Chi-Square variate with n *d.f.*

Chi-Square test is a statistical test used to compare observed results with expected results. It is a statistical procedure for determining the differences between observed and expected frequencies pertaining to any

particular phenomenon. It is used to determine whether your data is significantly different from what you expected. It is also known as Pearson's Chi-Square. This test involves categorical data. Categorical variables can be nominal or ordinal. Chi-Square test is of three types:

- The Chi-Square goodness of fit test, is used to test whether the frequency distribution of a categorical variable is different from your expectations.
- The Chi-Square test of independence of attributes, used to test whether the two categorical variables are related to each other.
- The Chi-Square test for homogeneity, is used to test whether two are more samples are drawn from the same population or from different populations.

8.5.2. USES OF CHI-SQUARE TEST

Chi-Square is a non-parametric test which is being extensively used for the following reasons:

1. This test is a distribution-free method, which does not rely on assumption that the data are drawn from a given parametric family of probability distributions.
2. This is easier to compute and simple enough to understand as compared to parametric test.
3. This test can be used in the situations where parametric tests are not appropriate or measurements prohibit the use of parametric tests.

8.5.3. APPLICATIONS OF CHI-SQUARE DISTRIBUTION

Chi-Square has large number of applications in Statistics. Let us discuss some of them:

i)To test the independence of attributes.

ii) To test if the hypothetical value of the populations variances is $\sigma^2=\sigma_0^2$.

iii) To test the 'goodness of fit'.

iv) To combine various probabilities obtained from independent experiments to give a single test of significance.

v) To test the homogeneity of independent estimates of the population variance.

vi) To test the homogeneity of independent estimates of the population correlation coefficient.

8.5.3.1. Inferences about a Population Variance

Let us consider a random sample x_i, *(i=1,2,…,n)* of size *n*. Suppose we want to test if a random sample has been drawn from a normal population with a specified variance $\sigma^2=\sigma_0^2$(say).

The null hypothesis will be: $H_0 : \sigma^2=\sigma_0^2$, the statistics:

$$\sum_{i=1}^{n}\left[\frac{(x_i-\bar{x})^2}{{\sigma_0}^2}\right]=\frac{1}{{\sigma_0}^2}\left[\sum_{i=1}^{n}{x_i}^2-\frac{\sum_{i=1}^{n}{x_i}^2}{n}\right]=\frac{n{s}^2}{{\sigma_0}^2} \qquad \ldots(**)$$

The statistic in eq (**) follows a Chi-Square distribution with *(n-1) d.f.*

The next step is to compare the calculated value with the tabulated value of chi-square for *(n-1)* degrees of freedom for any certain level of significance (we usually use 5%). If the calculated value is less than the tabulated value, we may retain the null hypothesis i.e. we may accept the null hypothesis and we conclude that value is not significant. If calculated value is greater than the tabulated value, we may reject the null hypothesis and we can conclude that the value is significant.

Remarks:*1) This test can be applied only if the population under study from which the sample has been drawn is normal.*

2) We can use Fisher's approximation if the sample size drawn is large i.e. greater than 30:

$$\sqrt{2\chi^2} \sim N\left(\sqrt{2n-1}, 1\right), \textit{ i.e., } Z = \sqrt{2\chi^2} - \sqrt{2n-1} \sim N(0,1)$$

Example 1: The following data shows the 11 measurements of the same subject:

2.5, 2.4, 2.3, 2.5, 2.7, 2.3 2.6, 2.5, 2.6, 2.5, 2.7

It is believed that the precision of an instrument is no more than 0.14. Carry out the test at 1% level of significance.

Solution:

Let us set up the Null Hypothesis, H_0: $\sigma^2 = 0.14$ against the alternative hypothesis, H_1: $\sigma^2 > 0.14$.

Here, $\sum X = 27.6, \bar{X} = 2.51, \sum (X - \bar{X})^2 = 0.1891$

Under the Null Hypothesis, H_0: $\sigma^2 = 0.14$, the test statistic is given by:

$$\chi^2 = \frac{ns^2}{\sigma^2} = \sum \frac{(X-\bar{X})^2}{\sigma^2} = \frac{0.1891}{0.14} = 1.35 \sim {\chi_{10}}^2$$

Computation of sample variance

X	$X-\bar{X}$	$(X-\bar{X})^2$
2.5	-0.01	0.0001
2.4	-0.11	0.0121
2.3	-0.21	0.0441
2.5	-0.01	0.0001
2.7	+0.19	0.0361
2.3	-0.21	0.0441
2.6	+0.09	0.0081
2.5	-0.01	0.0001
2.6	+0.09	0.0081
2.5	-0.01	0.0001
2.7	+0.19	0.0361

Since the calculated value of χ^2 is less than the tabulated value of χ^2 which is 23.2 for 10 *d.f.* at 1%

level of significance, it is a not significant. Hence, null hypothesis may be accepted. Therefore, we may conclude that the data are consistent with the assumption that the precision of the instrument is 0.14.

Example 2: Suppose a random sample of size 50 is drawn from a normal population with standard deviation s =10. Test the hypothesis that the data support the assumption of σ = 15.

Solution: Let us consider the hypothesis,

Null hypothesis, H_0: σ = 15 against the alternative hypothesis H_1: σ = 15

We are given n=50, s = 10. Now, $\chi^2 = \frac{ns^2}{\sigma^2} = \frac{50\times100}{225}$ =22.22

Since n is large, using (***), the test statistics is: $Z = \sqrt{2\chi^2} - \sqrt{2n-1} \sim N(0,1)$

$\therefore Z = \sqrt{44.44} - \sqrt{99}$ = 6.66-9.95 = -3.29 $\Longrightarrow |Z| = 3.29$

Since $|Z| > 3$, it is significant at all the levels of significance. Therefore, H_0 is rejected and we may conclude that $\sigma \neq 15$.

8.5.3.2. A Test for Independence of Attributes

This test is used to test the association of two or more attributes. Let us suppose we have N observations classified according to two attributes A and B. The attribute A is divided into 'r' classes i.e., $A_1, A_2, \ldots, A_r$ and attribute B is divided into 's' classes i.e., $B_1, B_2, \ldots, B_s$. If the attributes are divided into more than two classes then it is called manifold classification. By applying the test of independence for attributes on the given observations, we try to find out whether the attributes have some association or they are independent. The association between the attributes may be positive, negative or absence of association. For example, we can find out whether there is any association between regularity in class and the grade of the passing students. In order to test whether the attributes are associated or not we take the null hypothesis that there is no association in the attributes under study i.e., the two attributes are independent. The various cell frequencies can be expressed in the following table known as $(r \times s)$ *manifold contingency table* where, (A_i) is the number of persons possessing the attribute A_i, *(i=1, 2, ..., r);* (B_j) is the number of persons possessing the attribute B_j, *(j=1, 2, ..., s);* $(A_i\ B_j)$ is the number of persons possessing both the attributes A_i and B_j, *(i=1, 2, ..., r; j=1, 2, ..., s).* Note that, $\sum_{i=1}^{r}(A_i) = \sum_{j=1}^{s}(B_j) = N$, where, N is the total frequency.

A / B	A_1	A_2	…	A_i	…	A_r	Total
B_1	$(A_1\ B_1)$	$(A_2\ B_1)$	…	$(A_i\ B_1)$	…	$(A_r\ B_1)$	(B_1)
B_2	$(A_1\ B_2)$	$(A_2\ B_2)$	…	$(A_i\ B_2)$	…	$(A_r\ B_2)$	(B_2)
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
B_j	$(A_1\ B_j)$	$(A_2\ B_j)$	…	$(A_i\ B_j)$	…	$(A_r\ B_j)$	(B_j)
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
B_s	$(A_1\ B_s)$	$(A_2\ B_s)$	…	$(A_i\ B_s)$	…	$(A_r\ B_s)$	(B_s)
Total	(A_1)	(A_2)	…	(A_i)	…	(A_r)	N

The exact test for independence of attribute is very complicated but a fair degree of approximation, for large samples, is given by:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s}\left[\frac{\{(A_iB_j)-(A_iB_j)_0\}^2}{(A_iB_j)_0}\right] = \sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(f_{ij}-e_{ij})^2}{e_{ij}} \quad \ldots (***)$$

where, f_{ij} is the observed frequency and e_{ij} is the expected frequency and it follows χ^2- variate with *(r-1)(s-1) d.f.*

Remark: $\varphi^2 = \frac{\chi^2}{N}$ *is known as mean-square contingency. Since, the limits for f^2 and ?^2vary in different cases, they cannot be used for establishing the closeness of the association between the qualitative variables under study. Therefore, another measure suggested by Prof Karl Pearson, known as coefficient of "coefficient of mean square contingency" and is denoted by C, which is given by:* $\sqrt{\frac{\chi^2}{\chi^2+N}} = \sqrt{\frac{\varphi^2}{1+\varphi^2}}$. *Noted that, C is always less than unity. The maximum value of C depends on the number of classes into which A and B are divided i.e., r and s. In a r×s contingency table, the maximum value of C=* $\sqrt{\left(r-\frac{1}{r}\right)}$. *Since, the maximum value of C differs for different classifications, viz., r×r (r=2, 3, 4, 5, ...), strictly speaking the value of C obtained from different types of classifications are not comparable.*

Example 3: Suppose a random sample of students of *ABC* University was selected and asked their opinion about 'autonomous colleges'. The equal number of each sex is included within each class-group. The results are given below.

Class	Numbers		Total
	Favouring Autonomous colleges	Opposed to Autonomous colleges	
B.A./B.Sc./B.Com. Part I	120	80	200
B.A./B.Sc./B.Com. Part II	130	70	200
B.A./B.Sc./B.Com. Part III	70	30	100
M.A./M.Sc./M.Com.	80	20	100
Total	400	200	600

Test the hypothesis that opinions are independent of the class groupings at 5% level of significance.

Solution: Let us set up the null hypothesis that the opinions about autonomous colleges are independent of the class-groupings.

Under the null hypothesis of independence:

$$E(120) = \frac{400\times200}{600} = 133.33 \; ; \; E(130) = \frac{400\times200}{600} = 133.33 \; ; \; E(70) = \frac{400\times100}{600} = 66.67$$

The test statistic is given by:

$$\chi^2 = \sum_{i=1}^{n} \left[\frac{(f_i - e_i)^2}{e_i} \right] \sim \chi^2_{(r-1)(s-1)}$$

Calculation for χ^2

f_i	e_i	$f_i - e_i$	$(f_i - e_i)^2$	$\frac{(f_i - e_i)^2}{e_i}$
120	133.33	-13.33	177.6889	1.3327
130	133.33	-3.33	11.0889	0.0832
70	66.67	3.33	11.0889	0.1663
80	66.67	13.33	177.6889	2.6652
80	66.67	13.33	177.6889	2.6652
70	66.67	3.33	11.0889	0.1663
30	33.33	-3.33	11.0889	0.3327
20	33.33	-13.33	177.6889	5.3312
Total =400	**400**			**12.7428**

Tabulated value of χ^2 for (4-1)× (2-1) d.f. at 5% level of significance is 7.815.

Conclusion: The calculated value of χ^2 is 12.7428 and tabulated value at 5% for 3 *d.f.* is much less than the calculated value of χ^2. Therefore, the null hypothesis is rejected and we may conclude that that the opinions about autonomous colleges are not independent of the class-groupings.

8.5.3.2. A Test for homogeneity

The Chi-Square test of homogeneity is an extension of the Chi-Square test of independence. Such test indicates whether two or more independent samples are drawn from the same population or from different populations. Instead of one sample as we use in the independence problem, we shall now have two or more samples. Suppose a test is given to students in two different higher security schools. The sample size in both the case is the same. The question we have to ask is there any difference between the two higher secondary schools? In order to find the answer, we have to set up the null hypothesis that the two samples came from the same population. The word 'homogeneous' is used frequently in statistics to indicate 'the same' or 'equal'. Accordingly, we can say that we want to test in our example whether the two samples are homogeneous the test is called a test of homogeneity.

Example 1: The following information was obtained on the demand for a particular spare part in a factory which was found to vary from day-to-day:

Days	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
No. of parts demanded	1125	1110	1124	1126	1120	1115

Test the hypothesis that the number of parts demanded does not depends on day of week at 5% level of significance.

Solution: Set up the null hypothesis that, H_0: the number of spare parts demanded does not depends on day of the week, against the alternative hypothesis, H_1: the number of spare parts demanded depends on day of the week.

The test statistic is given by:

$$\chi^2 = \sum_{i=1}^{n} \left[\frac{(\boldsymbol{f_i - e_i})^2}{\boldsymbol{e_i}} \right] \sim \chi^2_{(n-1)}$$

Under the null hypothesis, the expected frequencies of the number of spare parts demanded on each day of week is given by:

$$\frac{1125+1110+1124+1126+1120+1115}{6} = 1120.$$

Calculation for $\boldsymbol{\chi^2}$

Days	$\boldsymbol{f_i}$	$\boldsymbol{e_i}$	$(\boldsymbol{f_i - e_i})^2$	$\frac{(\boldsymbol{f_i - e_i})^2}{\boldsymbol{e_i}}$
Monday	1125	1120	25	0.022
Tuesday	1110	1120	100	0.089
Wednesday	1124	1120	16	0.014
Thursday	1126	1120	36	0.032
Friday	1120	1120	0	0
Saturday	1115	1120	25	0.022
Total	**6720**	**6720**		**0.179**

Tabulated value of χ^2 at 5% level of significance for (6-1) = 5 *d.f.* is 11.07.

Since calculated value is less than the tabulated value, it is not significant. So, the null hypothesis is rejected and we may conclude that the number of spare parts demanded does not depend on day of the week.

5.5.3.3. A Test of goodness of fit

This test is the most important test of Chi-Square. This test is used to test whether an observed frequency distribution differ from an estimated frequency distribution. This test was proposed by Prof. Karl Pearson in 1900 and is known as *"Chi-Square test for goodness of fit"*. It enables us to find if the deviation of the experiment from theory is just by chance or it is really due to the inadequacy of the theory to fit the observed data. This is an approximate test for large values of *n*.

Suppose *fi* (i=1, 2, …, n) are the observed frequencies and *ei* (i=1, 2, …, n) are the corresponding expected frequencies, then the test statistic is given by:

$$\chi^2 = \sum_{i=1}^{n} \left[\frac{(\boldsymbol{f_i} - \boldsymbol{e_i})^2}{\boldsymbol{e_i}} \right] \sim \chi^2_{(n-1)}$$

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

Decision rule: If calculated value of χ^2 is less than the tabulated value of χ^2 at a level of significance for (n-1) *d.f.*, we accept the null hypothesis H_0 and we say that test is not significant otherwise we reject the null hypothesis and then in that case we say that value is significant.

Conditions for validity of chi-square test for 'goodness of fit':

For the validity of chi-square test for 'goodness of fit' between theory and experiment, the following conditions must be satisfied:

1. Sample should be random i.e. the sample observations should be independent.
2. Constraints on the cell frequencies should be linear, for example, $\sum_{i=1}^{n} f_i = \sum_{i=1}^{n} e_i$.
3. The total frequency, *N* should be reasonably large, say, greater than 50.
4. No theoretical cell frequency should be less than 5. If any of the cell frequency is less than 5 then for the application of chi-square it must be pooled either with the preceding frequency or with the succeeding frequency so that the pooled frequency is more than 5 and finally adjust for the degrees of freedom lost in pooling.

Remark: *It may be noted that chi-square test depends only on the set of observed and expected frequencies and on degrees of freedom. It does not make any assumptions regarding the parent population from which the observations are taken. Since chi-square does not involve any population parameters it is termed as statistic and the test is known as the non-parametric test or distribution-free test.*

Example 1: The following information was obtained on the demand for a particular spare part in a factory which was found to vary from day-to-day:

Days	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
No. of parts demanded	1125	1110	1124	1126	1120	1115

Test the hypothesis that the number of parts demanded does not depends on day of week at 5% level of significance.

Solution: Set up the null hypothesis that, H_0: the number of spare parts demanded does not depends on

day of the week, against the alternative hypothesis, H_1: the number of spare parts demanded depends on day of the week.

The test statistic is given by:

$$\chi^2 = \sum_{i=1}^{n} \left[\frac{(f_i - e_i)^2}{e_i} \right] \sim \chi^2_{(n-1)}$$

Under the null hypothesis, the expected frequencies of the number of spare parts demanded on each day of week is given by:

$$\frac{1125+1110+1124+1126+1120+1115}{6} = 1120.$$

Calculation for χ^2

Days	f_i	e_i	$(f_i - e_i)^2$	$\frac{(f_i - e_i)^2}{e_i}$
Monday	1125	1120	25	0.022
Tuesday	1110	1120	100	0.089
Wednesday	1124	1120	16	0.014
Thursday	1126	1120	36	0.032
Friday	1120	1120	0	0
Saturday	1115	1120	25	0.022
Total	**6720**	**6720**		**0.179**

Tabulated value of χ^2 at 5% level of significance for (6-1) = 5 *d.f.* is 11.07.

Since calculated value is less than the tabulated value, it is not significant. So, the null hypothesis is rejected and we may conclude that the number of spare parts demanded does not depend on day of the week.

Example 2: Two researchers adopted different sampling techniques while investigating the same group of students to find the number of students falling in different intelligence levels. The results are as follows:

Researcher	No. of students in each level				Total
	Below Average	Average	Above Average	Genius	
X	86	60	44	10	200
Y	40	33	25	2	100
Total	126	93	69	12	300

Would you say that the sampling techniques adopted by the two researchers are significantly different?

Solution: Let us set up the null hypothesis that the data obtained are independent of the sampling techniques adopted by the two researchers i.e. H_0: there is no significant difference between the sampling techniques

used by the two researchers for collecting the required data against the alternative hypothesis, H_1: there is a significant difference between the sampling techniques used by the two researchers for collecting the required data.

In this case, we have a 4×2 contingency table and therefore *d.f.* is (4-1) (2-1) = 3×1 = 3.

Calculation of Expected frequencies:

Under the null hypothesis of independence, we have,

$E(86) = \frac{200\times126}{300}$=84; $E(60) = \frac{200\times93}{300}$=62 and so on.

Calculation of χ^2

Researchers	Types of Students	f_i	e_i	$(f_i - e_i)^2$	$\frac{(f_i - e_i)^2}{e_i}$
X	Below Average	86	84	2	0.048
	Average	60	62	-2	0.064
	Above Average	44	46	-2	0.087
	Genius	10	8	2	0.500
Y	Below Average	40	42	-2	0.095
	Average	33	31	2	0.129
	Above Average Genius	25} 2}	23} 4}	0	0
Total		**6720**	**6720**		**0.923**

$$\chi^2 = \sum_{i=1}^{n}\left[\frac{(f_i - e_i)^2}{e_i}\right] \sim \chi^2_{(r-1)(s-1)-1}$$

$$= 0.923 \sim \chi^2_{(4-1)(2-1)-1}$$

$$= 0.923 \sim \chi^2_2 \qquad [\because 1\ d.f. \text{ is lost in pooling}]$$

Tabulated value of χ^2 at 5% level of significance for 2 *d.f.* is 5.991.

Conclusion: Since, calculated value of χ^2 is 0.923 and tabulated value is more than tabulated value at 5% level of significance for 2 *d.f.* We accept the null hypothesis and we may conclude that there is no significant difference between the sampling techniques used by the two researchers for collecting the required data.

8.6. YATE'S CORRECTION

In a 2×2 contingency table, the number of d.f. is (2-1)(2-1) = 1. If anyone of the theoretical cell frequencies is less than 5 then use of pooling method for Chi-Square test results in Chi Square with 0 degrees of

freedom. Since 1 degree of freedom is lost in pooling which is meaningless, in this case, F. Yates provide the correction method which is called 'Yates Correction for continuity'. Since χ^2 is a continuous distribution and if any of the cell frequencyis less than 5 then it loses its continuity. This consists in adding 0.5 to the cell frequency which is less than 5 and then adjusting for the remaining cell frequencies accordingly. The chi-square test for 'goodness of fit' is then applied without pooling method.

For a 2×2 contingency table $\begin{array}{|c|c|}\hline a & b \\ \hline c & d \\ \hline\end{array}$ we have $\chi^2 = \frac{N(ad-bc)^2}{(a+c)(b+d)(a+b)(c+d)}$

According to Yate's correction, we add (or subtract) 0.5 from *a* and *d* and subtract (add) 0.5 to *b* and *c* so that the marginal totals are not disturbed at all. Thus, corrected value of χ^2 is given by:

$$\chi^2 = \frac{N[(a\mp 0.5)(d\mp 0.5)-(b\pm 0.5)(c\pm 0.5)]^2}{(a+c)(b+d)(a+b)(c+d)} \qquad (8.5.1.)$$

On solving (8.5.1.), we get,

$$\chi^2 = \frac{N\left[|ad-bc|-\frac{N}{2}\right]^2}{(a+c)(b+d)(a+b)(c+d)}$$

Note: 1. It is recommended by authors and it seems quite logical in the light of the above discussion that Yate's correction be applied to every 2×2 table, even if no theoretical cell frequency is less than 5.

2. If *N* is large, the use of Yate's correction will make little deviation in the value of χ^2. If N is small, the application of Yate's correction may overstate the probability.

8.7. CHECK YOUR PROGRESS 1

1. Define Standard error. What is its significance?

   …….....

   …….....

2. What do you mean by sampling?

   …….....

   …….....

3. What are different types of sampling?

   …….....

   …….....

4. Define Chi Square test. What are its uses.

……

……

5. Write down the applications of Chi Square test.

……

……

6. What are the conditions for validity of Chi Square test for goodness of fit?

……

……

7. What do you understand by goodness of fit?

……

……

8. Under what condition Yate's correction is used?

……

……

9. Which test is used to test the significance of independence of attributes?

……

……

10. What do you mean by t-test? What are the assumptions for t-test?

……

……

11. What are the applications of student's t-test?

……

……

12. What are the assumptions for t-test for difference of two means?

……

……

13 Write a note on paired t-test?

……

……

8.8. STUDENT'S t- TEST

Suppose *xi* (*i*=1, 2, …, n) be a random sample of size n drawn from a normal population with mean μ and variance σ^2. Then the student's t is defined by the statistic:

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \sim t_{n-1}$$

where, $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the sample mean and $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ is an unbiased estimate of the population variances σ^2.

8.8.1. ASSUMPTIONS OF T-TEST

The following are the assumptions which are made to the student-*t* test:

1. The sample observations are independent i.e. sample is random.
2. Parent population from which the sample has been drawn is normal.
3. Population standard deviation σ is unknown.

8.8.2. APPLICATIONS OF STUDENT'S T- TEST

In statistics, Student's t- test has a wide number of applications. Some of which are enumerated below:

1. To test if the sample mean differs significantly from the hypothetical value population mean, μ_0 (say).
2. To test the significance of the difference between two sample means.
3. To test the significance of an observed sample correlation coefficient and sample regression coefficient.
4. To test the significance of observed partial correlation coefficient.

Let us discuss the applications in detail, one by one.

8.8.2.1. t-Test for single mean

Suppose we want to test: i) if the sample mean differs significantly from the hypothetical value μ_0 of population mean or ii) if a random sample of size *n i.e.*x_i *(i=1, 2, ..., n)* has been drawn from a normal

population with a specified mean, say, μ_0.

Under the null hypothesis, H_0:

i) there is no significant difference between sample mean $\bar{x}$ and population mean μ_0 or

ii) the sample has been drawn from the population with mean μ_0.

The test statistic is given by:

$$t = \frac{\bar{x} - \mu}{\frac{S}{\sqrt{n}}} \sim t_{n-1}$$

where, $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the sample mean and sample variance, $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$.

Conclusion: Compare the calculated value of t with tabulated value of t at a level of significance for *(n-1)* degrees of freedom. If calculated value is less than the tabulated value, we accept the null hypothesis otherwise we reject it.

Remark: 95% confidence interval is given by: $\bar{x} \pm \frac{t_{0.05}S}{\sqrt{n}}$ and 99% confidence interval can be calculated by $\bar{x} \pm \frac{t_{0.01}S}{\sqrt{n}}$.

Example 1. A machinist is making spare parts with diameters of axle as 0.700 inch. A random sample of 10 parts indicates a mean-diameter of 0.742 inch with a standard deviation of 0.040 inch. Test whether the work is meeting the specifications at 5% level of significance.

Solution. We are given that: population mean, $\mu = 0.700$ inch, sample mean $\bar{x}$ is 0.742 inch, sample standard deviation, S is 0.040 inch and n=10.

Set up the null hypothesis that the product is meeting the specifications i.e., Ho: $\mu = 0.7000$ inch, against the alternative hypothesis that the product is not meeting the specifications i.e., H_1: $\mu \neq 0.7000$ inch.

The test statistic is given by:

$$t = \frac{\bar{x} - \mu}{\frac{S}{\sqrt{n}}} \sim t_{n-1}$$

$$= \frac{0.742 - 0.700}{\frac{0.040}{\sqrt{10}}} \sim t_9$$

$$= 3.15$$

Tabulated value of t for 5% level of significance for 9 *d.f.* is 2.26.

Conclusion: Here, calculated value of t is 3.15 and tabulated value of *t* at 5% level of significance for 9 *d.f.* is 2.26. Since, calculated value of *t* is more than the tabulated value, the value is significant and hence we reject the null hypothesis. Therefore, we may conclude that the product is not meeting the specifications.

Example 2. The average weekly sales of detergent powder in departmental stores were 146.3 per store. An advertising campaign was organized and after the campaign the mean weekly sales in 22 stores for a typical week increased to 153.7 and showed a standard deviation of 17.2. Test whether the advertising campaign successful or not at 5% level of significance?

Solution. Given that: n = 22, sample mean, $\bar{x}$ is 153.7, s=17.2

Set up the null hypothesis that the advertising campaign is not successful, i.e, Ho: μ = 146.3 against the alternative hypothesis that advertising campaign is successful i.e., Ho: μ > 146.3 (right-tailed).

The test statistic is given by:

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \sim t_{n-1}$$

$$= \frac{153.7 - 146.3}{\frac{17.2}{\sqrt{21}}} \sim t_{21}$$

$$= 9.03$$

Tabulated value of *t* for 21 *d.f.* at 5% level of significance for singled test is 1.72.

Conclusion. Since calculated value of *t* is much greater than the tabulated value of *t*, it is highly significant and hence null hypothesis is rejected. Therefore, we may conclude that advertising campaign is successful.

8.8.2.2. t-Test for Difference of Two Means

Suppose two independent random samples x_i (i=1, 2, …, n_1) and y_j(j=1, 2, …, n_2) of sizes n_1 and n_2 have been drawn from two normal populations with means $\mu_X$and $\mu_Y$respectively.

Under the null hypothesis (H_0) that the samples have been drawn from the normal populations with means μ_X and μ_Y and under that assumption that the population variances are equal i.e. $\sigma_X{}^2 = \sigma_Y{}^2 = \sigma^2$(say), the test statistic is given by:

$$t = \frac{(\bar{x} - \bar{y}) - (\mu_X - \mu_Y)}{s\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t_{n_1+n_2-2}$$

where, $\bar{x} = \frac{1}{n_1}\sum_{i=1}^{n_1} x_i$is the sample mean of random sample *X* and $\bar{y} = \frac{1}{n_2}\sum_{j=1}^{n_2} y_j$ is the sample mean of

random sample Y and sample variance, $S^2 = \frac{1}{n_1+n_2-2}\left[\sum_{i=1}^{n_1}(x_i-\bar{x})^2 + \sum_{j=1}^{n_2}(y_j-\bar{y})^2\right]$ is an unbiased estimate of the common population variance σ^2.

Under the null hypothesis i.e. H_0: $\mu_X = \mu_Y$, the above expression of *t*-statistic can be written as:

$$t = \frac{(\bar{x}-\bar{y})}{s\sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \sim t_{n_1+n_2-2}$$

Remark: You can calculate the 95% confidence interval using the formula: $(\bar{x}-\bar{y}) \pm \frac{t_{0.05}S}{\sqrt{n}}$ and 99% confidence interval using the formula $(\bar{x}-\bar{y}) \pm \frac{t_{0.01}S}{\sqrt{n}}$.

8.8.2.3. Assumptions for the t-Test for Difference of Two Means

There are three fundamental assumptions of t-Test for Difference of Two Means as under:

1. The two samples are random and independent of each other.
2. Parent populations from which the samples have been drawn are normally distributed.
3. The population variances are equal and unknown i.e. $\sigma_1^2 = \sigma_2^2 = \sigma^2$ (say).

Example1 : The following table gives data on gain in weight (in kgs) of goats fed on two diets *A* and *B*:

Diet A	25	32	30	34	24	14	32	24	30	31	35	25			
Diet B	44	34	22	10	47	31	40	30	32	35	18	21	35	29	22

Test whether the two diets differ significantly or not as regards their effect on increase in weight.

Solution: Let us set up the null hypothesis, H_0 that there is no significant difference between the mean increase in weight due to the Diet A and Diet B i.e. H_0: $\mu_X = \mu_Y$.

Against the alternative hypothesis, H1 that there is a significant difference between the mean increase in weight due to the Diet *A* and Diet *B* i.e. H_1: $\mu_X \neq \mu_Y$(two-tailed).

Here, $n_1 = 12, n_2 = 15$; $\bar{x} = \frac{1}{n_1}\sum_{i=1}^{n_1} x_i = \frac{336}{12} = 28$; $\bar{y} = \frac{1}{n_2}\sum_{j=1}^{n_2} y_j = \frac{450}{15} = 30$

$S^2 = \frac{1}{n_1+n_2-2}\left[\sum_{i=1}^{n_1}(x_i-\bar{x})^2 + \sum_{j=1}^{n_2}(y_j-\bar{y})^2\right] = \frac{1}{12+15-2} \times [380+1410] = 71.6$

Under the null hypothesis (H_0):

$$|t| = \left|\frac{(\bar{x}-\bar{y})}{s\sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}\right| \sim t_{n_1+n_2-2}$$

$$= \left| \frac{(28-30)}{8.46 \times \sqrt{\left(\frac{1}{12}+\frac{1}{15}\right)}} \right| \sim t_{12+15-2}$$

$$= 0.609 \sim t_{25}$$

Diet A			Diet B		
x	$x-\bar{x}$	$(x-\bar{x})^2$	y	$y-\bar{y}$	$(y-y)^2$
25	-3	9	44	14	196
32	4	16	34	4	16
30	2	4	22	-8	64
34	6	36	10	-20	400
24	-4	16	47	17	289
14	-14	196	31	1	1
32	4	16	40	10	100
24	-4	16	30	0	0
30	2	4	32	2	4
31	3	9	35	5	25
35	7	49	18	-12	144
25	-3	9	21	-9	81
			35	5	25
			29	-1	1
			22	-8	64
Σx =336	$\sum(x-\bar{x})$ = 0	$\sum(x-\bar{x})^2$ = 380	Σy =450	$\Sigma(y-\bar{y})$ =0	$\sum(x-\bar{x})^2$ = 1410

Tabulated value of t at 5% level of significance for 25 degrees of freedom is 2.06.

Decision: Here, calculated value of $|t|$ is 0.609 and tabulated value of t at 5% level of significance for 25 degrees of freedom is 2.06. Since calculated value is less than the tabulated value, null hypothesis may be accepted and we may conclude that there is no significant difference between the mean increase in weight due to the Diet *A* and Diet *B*.

8.8.2.4. Paired t-Test for Difference of two Means

Paired t-Test is used when the two samples are not independent i.e., they are related to each other. It is a statistical method to test that compares the means of two variables for a single group. This test is also called *'before and after test'*.

Suppose we consider the case when i) :

the sample sizes are equal i.e., n_1=n_2=n (say).

the sample sizes are equal i.e., n_1=n_2=n (say) and ii) the two samples are not independent but the sample observations are paired together, that is, the pair of observations (x_i, y_i), *(i= 1, 2, ..., n)* corresponds to

the same sample unit.

The problem is to test whether the sample means differ significantly or not i.e. to test whether the mean difference between pairs of measurements is zero or not.

Therefore, the test statistic is given by:

$$t = \frac{\bar{d}}{\frac{s}{\sqrt{n}}} \sim t_{n-1}$$

where, $\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i$; $d_i = x_i - y_i (i = 1, 2, \dots, n)$ are the increments which are due to fluctuations of sampling, $S^2 = \frac{1}{n-1} \times \sum_{i=1}^{n_1} (d_i - \bar{d})^2$.

Remark: You can calculate the 95% confidence interval using the formula: $\bar{d} \pm \frac{t_{0.05} * S}{\sqrt{n}}$ and 99% confidence interval using the formula $\bar{d} \pm \frac{t_{0.01} * S}{\sqrt{n}}$.

Example 1: The following results of increase in weight due to calcium supplement were observed in animals:

Animal No.		1	2	3	4	5	6	7	8	Total
Weight in lb	Before supplement	53	52	51	49	47	52	50	53	407
	After supplement	55	53	52	52	50	54	54	53	423

Test the hypothesis whether is significant effect of calcium supplement on increase in weight?

Solution: Suppose the sample of 'before supplement' is denoted by *X* and the sample of 'after supplement' is denoted by *Y*.

Let us consider the null hypothesis, H_0: there is no significant effect of calcium supplement on increase in weight i.e., H_0: $\mu_X = \mu_Y$ against the alternativehypothesis, H_1: there is a significant effect of calcium supplement on increase in weight i.e., H_0: $\mu_X = \mu_Y$ against the alternative hypothes is H_1: $\mu_X > \mu_Y$.

The test statistic is given by:

$$t = \frac{\bar{d}}{\frac{s}{\sqrt{n}}} \sim t_{n-1}$$

where, $\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i$; $d_i = x_i - y_i (i = 1, 2, \dots, n)$;

$S^2 = \frac{1}{n-1} \times \sum_{i=1}^{n_1} (d_i - \bar{d})^2 = \frac{1}{n-1} \times \left[\sum_{i=1}^{n} {d_i}^2 - \frac{\left(\sum_{i=1}^{n} d\right)^2}{n}\right]$

S.No.	1	2	3	4	5	6	7	8	Total
X	53	52	51	49	47	52	50	53	407
Y	55	53	52	52	50	54	54	53	423
d=X-Y	-2	-1	-1	-3	-3	-2	-4	0	-16
d^2	4	1	1	9	9	4	16	0	44

Here, $\bar{d}=\frac{1}{n}\sum_{i=1}^{n} d_i=\frac{-16}{8}=-2$; $S^2=\frac{1}{n-1}\times\sum_{i=1}^{n_1}(d_i-\bar{d})^2=\frac{1}{8-1}\times\left(44-\frac{256}{8}\right)=1.714$

$$\therefore |t|=\left|\frac{\bar{d}}{\frac{S}{\sqrt{n}}}\right|\sim t_{n-1}$$

$$=\left|\frac{-2}{\frac{1.309}{\sqrt{8}}}\right|\sim t_7$$

$$=4.32$$

Tabulated value of t at 5% level of significance for 7 *d.f.* is 1.90

Decision: Since the calculated value of $|t|$ is more than the tabulated value of *t* at 5% level of significance for 7 *d.f.*, we reject our null hypothesis and we may conclude that there is a significant effect of calcium supplement on increase in weight i.e., H_0: $\mu_X=\mu_Y$ against the alternative hypothesis H_1: $\mu_X > \mu_Y$.

8.8.2.5. t-Test for testing the significance of an observed sample correlation coefficient

This test was given by Prof. R.A. Fisher. Suppose '*r*' be the observed sample correlation coefficient in a sample of '*n*' pairs of observations from a bivariate normal population, then under the null hypothesis: population correlation coefficient is zero i.e., H_0: $\rho=0$.

The test statistic is given by:

$$t=\frac{r}{\sqrt{1-r^2}}\sqrt{n-2}$$

and it follows student's *t* – distribution with *(n-2) d.f.*

Decision: If the calculated value of *t* comes out to be greater than tabulated value at α level of significance, it is called significant and we reject or null hypothesis otherwise accept it.

Example 1: The correlation coefficient of 27 pair of observations from a normal population is 0.6. Is this value of correlation significant?

Solution: We set up the null hypothesis that the population correlation coefficient is zero i.e., H_0: $\rho=0$ against the alternative hypothesis H_0: $\rho\neq 0$.

The test statistic is given by:

$$t = \frac{r}{\sqrt{1-r^2}}\sqrt{n-2} \quad \sim t_{n-2}$$

$$= \frac{0.6}{\sqrt{1-0.6^2}} \times \quad \sim t_{25}$$

$$= 3.75$$

Decision: Calculated value of *t* is 3.75 and the tabulated value of *t* at 5% level of significance for 25 degrees of freedom is 2.06. Since the calculated value is greater than the tabulated value, we reject our null hypothesis. Hence, we may conclude that $\rho \neq 0$.

8.9. F-TEST

F-distribution is defined as the ratio of two independent variates *X* and *Y* divided by their respective degrees of freedom ϑ_1 and ϑ_2 and it follows Snedecor's *F* distribution with $(\vartheta_1, \vartheta_2)$ degrees of freedom, i.e.,

$$F = \frac{\frac{X}{\vartheta_1}}{\frac{Y}{\vartheta_2}} \quad \sim F(\vartheta_1, \vartheta_2)$$

8.9.1. ASSUMPTIONS OF F-TEST

F-Test has several assumptions that are made to the F-Test are:

i) *Independence:* The observations within each group are independent of each other.

ii) *Normality:* The data within each group follows a normal distribution.

iii) *Random samples:* The samples must be random.

iv) *Homogeneity of Variances:* The variance in each group being compared are approximately equal.

8.9.2. APPLICATIONS OF F-TEST

F-Test has the following applications in statistical theory:

i) F-Test for equality of two population variances

ii) F-Test for equality of several means

iii) F-Test for testing the significance of an observed multiple correlation coefficient.

iv) F-Test for testing the significance of an observed sample correlation coefficient ratio.

v) F-Test for testing the linearity of regression

Let us discuss the applications of F-Test in detail:

8.9.2.1. F-Test for equality of two population variances

Let us consider two random samples x_i(i=1, 2, …, n_1) and y_j (j=1, 2, …, n_2) of sizes n_1 and n_2 have been drawn from two normal populations with same variance σ^2 (say).

Suppose we want to test i) whether the two independent samples have been drawn from the normal populations with the same variance σ^2.

ii) whether the two independent estimates of population variances are homogenous or not.

Therefore, under the null hypothesis that i) the population variances are equal i.e., ${\sigma_X}^2 = {\sigma_Y}^2 = \sigma^2$ (say), or ii) whether the two independent estimates of population variances are homogenous or not, the F-statistic is given by:

$$F = \frac{\text{Larger sample variance}}{\text{Smaller sample variance}} = \frac{{S_X}^2}{{S_Y}^2} \quad \sim F(n_1, n_2)$$

where, ${S_X}^2 = \frac{1}{n_1 - 1}\left[\sum_{i=1}^{n_1}(x_i - \bar{x})^2\right]$; ${S_Y}^2 = \frac{1}{n_2 - 1}\left[\sum_{j=1}^{n_2}(y_j - \bar{y})^2\right]$ are the unbiased estimates of the population variances ${\sigma_X}^2$ and ${\sigma_Y}^2$ respectively.

Remarks: If $F(n_1, n_2)$ represents an F-variate with n_1 and n_2 degrees of freedom, then $F(n_2, n_1)$ is distributed as 1/ $F(n_1, n_2)$ variate.

Example 1: The following data gives the estimates of two independent samples of sizes 8 and 7 respectively:

Sample I	**9**	**11**	**13**	**11**	**15**	**9**	**12**
Sample II	**10**	**12**	**10**	**14**	**9**	**8**	**10**

Test whether the estimates of population variances differ significantly at 5% level of significance.

Solution: Let us consider sample I as random variable X and sample II as random variable Y. Set up null hypothesis that there is no significant difference between the two population variances i.e. H_0:${\sigma_1}^2 = {\sigma_2}^2$ against the alternative hypothesis, H_1:${\sigma_1}^2 \neq {\sigma_2}^2$.

Here, $\bar{x} = \frac{1}{n_1}\sum_{i=1}^{n_1} x_i = \frac{94}{8} = 11.75$; $\bar{y} = \frac{1}{n_2}\sum_{j=1}^{n_2} y_j = \frac{73}{7} = 10.42$

X	**9**	**11**	**13**	**11**	**15**	**9**	**12**	**14**	$\sum X = 94$
Y	**10**	**12**	**10**	**14**	**9**	**8**	**10**		$\sum Y = 73$
$(X-\bar{X})$	-2.75	-0.75	1.25	-0.75	3.25	-2.75	0.25	2.25	
$(X-\bar{X})^2$	7.5625	0.5625	1.5625	0.5625	10.5625	7.5625	0.0625	5.0625	$\sum(X-\bar{X})^2$=28.4375
$(Y-\bar{Y})$	-0.42	1.58	-0.42	3.58	-1.42	-2.42	1.58		
$(Y-\bar{Y})^2$	0.1764	2.4964	0.1764	12.8164	2.0164	5.8564	2.4964		$\sum(Y-\bar{Y})^2$=26.0348

$S_X{}^2 = \frac{1}{n_1 - 1}\left[\sum_{i=1}^{n_1}(x_i - \bar{x})^2\right] = \frac{28.4375}{7} = 4.0625$;

$S_Y{}^2 = \frac{1}{n_2 - 1}\left[\sum_{j=1}^{n_2}(y_j - \bar{y})^2\right] = \frac{26.0348}{6} = 4.3391$

Therefore, the test statistic is given by:

$$F = \frac{\text{Larger sample variance}}{\text{Smaller sample variance}} = \frac{4.3391}{4.0625} = 1.07 \quad \sim F(7,8)$$

Tabulated value of *F* at 5% level of significance for (7,8) degrees of freedom is 3.73.

Decision: Since Calculated value is less than the tabulated value at 5% level of significance for (7, 8) degrees of freedom, we may accept or null hypothesis. Hence, we may conclude that estimates of two population variances do not differ significantly.

Example 2: The following data gave the result of two random samples:

Sample	Size	Sample Mean	Sum of square of deviation from mean
1	10	15	90
2	12	14	108

Test whether the samples come from the same normal population at 5% level of significance.

Solution: (Try Yourself)

Hint: We know that normal population has two parameters viz., mean μ and variance σ^2. In order to test that the two samples come from the same normal population, we have to test i) the equality of population variance i.e.,$\mu_1 = \mu_2$ and ii) the equality of population variance i.e., $\sigma_1{}^2 = \sigma_2{}^2$.

Set up the null hypothesis, H_0: $\mu_1 = \mu_2$ and $\sigma_1{}^2 = \sigma_2{}^2$, i.e., the two samples have been drawn from the same normal population against the alternative hypothesis that they are not.

Equality of means can be tested by using *t*-Test of difference of two means F-Test is used for testing equality of two population variances. Since, *t*-Test is based on the assumption of equality of two population variance. Therefore, to check the assumption, first we will test equality of means by using *t*-Test of difference of two means and then we will apply F-Test for equality of two population variances.

If the assumption of equality of two variance satisfies then we will proceed for *t*-Test otherwise we will conclude the result with acceptance of alternative hypothesis.

8.9.2.2. F-Test for testing the significance of an Observed Multiple Correlation Coefficient

Let us suppose *R* is the observed multiple correlation coefficient of a variate with *k* other variates in a

random sample of size n from a (*k+1*) variate population. Prof. R.A. Fisher proved that under the null hypothesis (H_0) that the multiple correlation coefficient in the population is zero. Then, the test statistic is given by

$$F = \frac{R^2}{1-R^2} \times \frac{n-k-1}{k} \sim F(k, n-k-1)$$

Decision: If the calculated value is less than the tabulated value at a level of significance for $(h-1, N-h)$ degrees of freedom, we may accept our null hypothesis otherwise we reject it.

8.9.2.3. F-Test for testing the significance of an Observed Sample Correlation Ratio

Under this testing, the null hypothesis (H_0): Population correlation ratio is zero i.e., $\eta = 0$.

The test statistic is given by:

$$F = \frac{\eta^2}{1-\eta^2} \times \frac{N-h}{h-1} \sim F(h-1, N-h)$$

Decision: If the calculated value is less than the tabulated value at a level of significance for (*h*-1,*N-h)* degrees of freedom, we may accept our null hypothesis otherwise we reject it.

8.9.2.4. F-Test for testing the Linearity of Regression

Let us consider a sample of size *N* arranged in *h* arrays drawn from a bi-variate normal population.

The test statistic for testing the hypothesis of linearity of regression is given by:

$$F = \frac{\eta^2 - r^2}{1-\eta^2} \times \frac{N-h}{h-2} \sim F(h-2, N-h)$$

Decision: If the calculated value is less than the tabulated value at a level of significance for (*h*-2,*N-h)* degrees of freedom, we may accept our null hypothesis otherwise we reject it.

8.9.2.5. F-Test for testing the equality of several means

t-Test is used to test the significance of difference between two means only. When there are more than two samples under consideration then *t*-Test cannot be applied. In that case F-Test is used to test the significance of equality of more than two population means.

This test is carried out by the technique of *Analysis of Variance* which we have discussed in the next chapter. This test plays a fundamental role in *Design of Experiments* in Agricultural Statistics.

8.9.2.6. Relation between *t* and F Distributions

If a statistic *t* follows Student's *t* distribution with *n d.f.* then t^2 follows F distribution with (*1, n*) *d.f.*

Symbolically, *if* $t \sim t_{(n)}$

then $t^2 \sim F_{(1,n)}$

8.9.2.7. Relation between F and χ^2 Distributions

If a statistic *t* follows χ^2 distribution with n_1 *d.f.* then in $F(n_1, n_2)$ distribution if we let $n_2 \to \infty, \chi^2 = n_1 F$ follows χ^2 distribution with n_1*d.f.*

8.10 TESTS OF SIGNIFICANCE FOR LARGE SAMPLES

For large number of trials (or for large values of *n*), almost all the distributions, e.g., Poisson, binomial, negative binomial distribution etc., are very closely approximated by normal distribution. Thus, we apply the normal tests for large samples which are based on the fundamental property called *area property* of the normal probability curve.

If $X \sim N(\mu, \sigma^2)$then $Z = \frac{X - E(X)}{\sqrt{V(X)}} = \frac{X-\mu}{\sigma}$ is called standard normal variate and it follows *N(0, 1).*

Therefore, from the table of normal probability, we have,

$$P(-3 \leq Z \leq 3) = 0.9973 \quad \Rightarrow P(|Z| > 3) = 1 - P(|Z| \leq 3) = 0.0027$$

This means that in all probability we should expect a standard normal variate to lie between ±3.

Also, $P(-1.96 \leq Z \leq 1.96) = 0.95 \quad \Rightarrow P(|Z| \leq 1.96) = 0.95$

$$\Rightarrow P(|Z| > 1.96) = 1 - 0.95 = 0.05$$

and $P(|Z| < 2.58) = 0.99 \Rightarrow P(|Z| > 2.58) = 0.01$

At 5% level of significance, the significant value of Z is 1.96 and at 1% level of significance, the significant value of Z is 2.58.

8.10.1. PROCEDURE FOR TESTING OF HYPOTHESIS

Below are the various steps which one must follow in a systematic manner for testing of a statistical hypothesis for large samples:

1. *Null Hypothesis:* Set up the null hypothesis H_0 for the parameter.
2. *Alternative hypothesis:* Set up an alternative hypothesis H_1 for the parameter. Here, we will enable about the Here, we will enableto decide whether we have to use a single-tailed (left or right) or two-tailed test.
3. *Level of significance:* We decide about the level of significance i.e. α valueon which the results

are to tested or compared.

4. *Test Statistic:* Computation of test statistic, under H_0, is given by:

$$Z = \frac{t - E(t)}{S.E.(t)} \sim N(0,1)$$

5. Conclusion/Decision: In the last step, we compare our calculated value with tabulated value at given level of significance for large samples. If calculated value is less than the tabulated value i.e. if $|Z| < z\alpha$ then we say it is not significant and, in this case, null hypothesis is rejected. But, If calculated value is more than the tabulated value i.e. if $|Z| > z\alpha$ then we say it is significant and, in this case, null hypothesis is accepted.

Remark: For practical purposes, $n > 30$ is considered as a large sample.

8.11 SAMPLING OF ATTRIBUTES

By attribute we mean characteristics or quality. In sampling of attributes, we consider the sampling from population which is divided into two mutually exclusive and exhaustive classes in which one class possessing the particular, say, *A* and the other class not possessing the attribute. The presence of attribute may be termed as success and absence of the attribute is termed as failure. In this case, we get the sample of size *n* with outcomes in the form of success and failure and hence identified as *n* independent Bernoulli trials with constant of probability *P* of success for each trial. Then, the probability of *x* successes in *n* trials is given by the binomial probability distribution:

$p(x) = \binom{n}{x} P^x Q^{n-x}$; $x = 0, 1, 2, \ldots, n$.trails with constant probability *P* of success for each trial. Then, the probabili

where, $Q - 1 - P$.

8.11.1 TEST OF SIGNIFICANCE FOR SINGLE PROPORTION

Let the number of successes in n trials be denoted by *X* and the probability of success is denoted by *P* for each trial, then $E(X) = nP$ and $V(X) = nPQ$ where, $Q = 1 - P$ is called the probability of failure.

For large sample size *n*, $Z = \frac{X - E(X)}{\sqrt{V(X)}} = \frac{X - nP}{nPQ} \sim N(0,1)$

Therefore, for large value of n we apply the normal test.

Remark: 1) Observed proportion of success$=\frac{X}{n} = p, (say)$

where, n is the sample size and X is the number of persons possessing the given attribute. ∴

$E(p) = E\left(\frac{X}{n}\right) = \frac{1}{n}E(X) = \frac{1}{n}\, nP = P$. Hence, sample proportion *'p'* is an unbiased estimate of the population proportion P.

Also, $V(p) = V\left(\frac{X}{n}\right) = \frac{1}{n}V(X) = \frac{1}{n^2}\, nPQ = \frac{PQ}{n}$

and, $S.E.(p) = \sqrt{\frac{PQ}{n}}$

2) 95% confidence interval for P can be calculated by: $P \pm 1.96 \times \sqrt{\frac{PQ}{n}}$

and, 99% confidence interval can be calculated by: $P \pm 2.58 \times \sqrt{\frac{PQ}{n}}$

Example 1: An unbiased die is thrown 8,000 times and a throw of 2 or 3 is observed 3,250 times. Show that the throw of die cannot be regarded as unbiased one. Also, find the 95% confidence limits between which the probability of a throw of 2 or 3 lies.

Solution: Let X denote the number of successes which is equal to 3,250. Given that, n is 8,000.

Therefore, probability of getting a throw of 2 or 3 is $\frac{1}{6}+\frac{1}{6}=\frac{1}{3}$.

Set up the null hypothesis (H_0): Die is an unbiased i.e., $P=\frac{1}{3}$ against the alternative hypothesis (H_1): Die cannot be regarded as unbiased one i.e., $P \neq \frac{1}{3}$.

Under the null hypothesis, the test statistic is given by:

$$Z = \frac{X - nP}{\sqrt{nPQ}} \sim N(0,1)$$

$$= \frac{3250 - 8000 \times \frac{1}{3}}{\sqrt{8000 \times \frac{1}{3} \times \frac{2}{3}}} = \frac{583.33}{1777.77} = 0.3281 \sim N(0,1)$$

Since, $|Z| > 3$, null hypothesis is rejected and hence we may conclude that the die cannot be regarded as unbiased one i.e., $P \neq \frac{1}{3}$.

Therefore, 95% confidence interval is given by: $\hat{P} \pm 3 \times \sqrt{\frac{\hat{P}\hat{Q}}{n}}$,

where, $\hat{P} = \frac{3250}{8000} = 0.41$ and $\hat{Q} = 1 - \hat{P} = 1 - 0.41 = 0.59$

$$\therefore \hat{P} \pm 3 \times \sqrt{\frac{\hat{P}\hat{Q}}{n}} = 0.41 \pm 3 \times \sqrt{\frac{0.41 \times 0.59}{8000}} = (0.3935, 0.4265).$$

Example 2: From the sample of 500 people in Calcutta, 260 are wheat eaters and 240 are ice eaters. Test the hypothesis whether both wheat and rice are equally popular in the state at 1% level of significance?

Solution: Given that: n = 500; X is the number of wheat eaters which are equal to 260 and number of rice eaters are 240.Sample proportion of wheat eaters,$p = \frac{X}{n} = \frac{260}{500} = 0.52$ and q=1-p=1-0.52=0.48.

Set up the null hypothesis that both wheat and rice eaters are equally popular in the state i.e., P=0.5against the alternative that they are not i.e., $P \neq 0.5$.

Under H0, the test statistic is given by:

$$Z = \frac{p - P}{\sqrt{\frac{PQ}{n}}} \sim N(0,1)$$

$$= \frac{0.52 - 0.5}{\sqrt{\frac{0.5 \times 0.5}{500}}} = 0.8849$$

Conclusion: Since the calculated value of Z <2.58 (i.e. at 1% level of significance), we may accept or null hypothesis. Hence, we may conclude that both wheat and rice eaters are equally popular in the state.

Note:1) 99% confidence interval can be calculated by: $Z \pm 2.58 \times \sqrt{\frac{PQ}{n}}$.

2) 95% confidence interval can be calculated by: $Z \pm 1.96 \times \sqrt{\frac{PQ}{n}}$.

8.11.2. TEST OF SIGNIFICANCE OR DIFFERENCE OF PROPORTIONS

This test is used to compare the prevalence of attribute in two distinct populations with respect to some attribute, A (say) among their members. Suppose, X_1 and X_2 be the no. of individuals from random samples of sizes n_1 and n_2 possessing the given attribute a from two populations respectively. Then, sample proportions are defined by $p_1 = \frac{X_1}{n_1}$ and $p_2 = \frac{X_2}{n_2}$.

Suppose P_1 and P_2 are the two population proportions, then, $E(p_1) = P_1$ and $E(p_2) = P_2$.

Also, $V(p_{1)} = \frac{P_1 Q_1}{n_1}$ and $V(p_{2)} = \frac{P_2 Q_2}{n_2}$.

For large samples, the sample proportions p_1 and p_2 are independently and asymptotically normally distributed. Therefore, $(p_1 - p_2)$ are also normally distributed. Then the standard normal variate corresponding to the difference $(p_1 - p_2)$ is given by:

$$Z = \frac{(p_1 - p_2) - \mathrm{E}(p_1 - p_2)}{\sqrt{V(p_1 - p_2)}} \sim N(0,1)$$

Set up the null hypothesis that there is no significant difference between the two sample proportions i.e., H_0: $P_1 = P_2$ against the alternative hypothesis that there is a significant difference between the two sample proportions i.e., H_1: $P_1 \neq P_2$.

Under the null hypothesis, H_0: $P_1 = P_2$, $\mathrm{E}(p_1 - p_2) = \mathrm{E}(p_1) - \mathrm{E}(p_2) = P_1 - P_2 = 0$

and, $\mathrm{V}(p_1 - p_2) = \mathrm{V}(p_1) + \mathrm{V}(p_2)$. Since the two samples are independent, the covariance term $\mathrm{Cov}(p_1, p_2)$ vanish.

$$V(p_1 - p_2) = \frac{P_1 Q_1}{n_1} + \frac{P_2 Q_2}{n_2} = PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \qquad \left[\because \text{Under } \mathrm{H}_0\text{: } \mathrm{P}_1 = \mathrm{P}_2 = \mathrm{P} \text{ (say) and } \mathrm{Q}_1 = \mathrm{Q}_2 = \mathrm{Q} \text{(say)}\right]$$

Therefore, the test statistic is given by:

$$Z = \frac{(p_1 - p_2)}{\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0,1)$$

In general, when the proportion of *A*'s in the populations from which the two samples have been drawn are not known then under the null hypothesis, H_0: $P_1 = P_2 = P$ an unbiased estimate of the population proportion base on the sample information is given by:

$$\hat{P} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{X_1 + X_2}{n_1 + n_2}$$

Note: $E(\hat{P}) = \frac{1}{n_1 + n_2} E(n_1 p_1 + n_2 p_2) = \frac{1}{n_1 + n_2} E[(n_1 p_1) + E(n_2 p_2)]$

$$= \frac{1}{n_1 + n_2}[n_1 P_1 + n_2 P_2]$$

Since, under H_0: $P_1 = P_2 = P$, $E(\hat{P}) = P$

Example 1: A project of flyover is proposed to build near a residential area. A random sample of 300 men and 450 women were selected to get responses on the proposal of flyover. Out of 300 men, 200 men were in favour of the proposal and out of 450 women, 325 were in favour of proposal. Test the hypothesis that whether the proportion of men and women in favour of proposal are same or not at 5%

level of significance.

Solution: Given that, $n_1 = 300, n_2 = 450$

Let X_1 = the no.of men in favour of proposal= 200 and X_2 = the no. of women in favour of proposal. Therefore, p_1 = Proportion of the men in favour of proposal = X_1/n_1 = 200/300= 0.67

p2 = Proportion of the women in favour of proposal = $\frac{X_1}{n_1} = \frac{200}{300} = 0.67$

Set up the null hypothesis that there is no significance difference between the opinion of men and women in favour of proposal i.e., $H_0: P_1 = P_2 = P$ (say).

Against the alternative hypothesis that there is a significance difference between the opinion of men and women in favour of proposal i.e., $H_1: P_1 \neq P_2$.

The test Statistic is given by:

$$Z = \frac{(p_1 - p_2)}{\sqrt{\hat{P}\hat{Q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0,1)$$

where, $\hat{P} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{X_1 + X_2}{n_1 + n_2} = \frac{200+325}{300+450} = 0.7$

$\Longrightarrow$ $\hat{Q} = 1 - \hat{P} = 1 - 0.7 = 0.3$

$\therefore$ $Z = \frac{(0.67-0.72)}{\sqrt{0.7 \times 0.3\left(\frac{1}{300} + \frac{1}{450}\right)}} = -1.47$

$\Longrightarrow$ $|Z| = 1.47$

Decision: Since calculated value of |Z| is less than 1.96 (5% level of significance), we may accept our null hypothesis. Hence, we may conclude that there is no significance difference between the opinion of men and women in favour of proposal.

Note: 1) The 95% confidence limits for $(p_1 - p_2)$ is given by:

$$(p_1 - p_2) \pm 1.96 \times \sqrt{\left(\frac{P_1 Q_1}{n_1} + \frac{P_2 Q_2}{n_2}\right)}$$

2) The 99% confidence limits for $(p_1 - p_2)$ is given by:

$$(p_1 - p_2) \pm 2.58 \times \sqrt{\left(\frac{P_1 Q_1}{n_1} + \frac{P_2 Q_2}{n_2}\right)}$$

Example 2: Out of a random sample of400 students of the University teaching departments, 300 failed in the examination. Out of 500 students of an affiliated college, 300 failed in the examination. Test whether the proportion of failures in the university teaching departments is more than the proportions of failures in the University teaching departments and affiliated colleges taken together? Use 5% level of significance.

Solution: Given that: $n_1 = 400$ and $n_2 = 500$. Let X_1 = no. of students of the university teaching departments who have failed = 300 and X_2 = no. of students of the affiliated colleges who have failed =300.

Let p_1=proportion of failed students from university teaching depertments

$$= \frac{X_1}{n_1} = \frac{300}{400} = 0.75$$

and, p_2= proportion of failed students from affiliated colleges

$$= \frac{X_2}{n_2} = \frac{300}{500} = 0.60.$$

Set up the null hypothesis, H_0 that there is no significance difference between p_1 and $p = \hat{P}$, where, $\hat{P}$ is the proportion of failed students in the university teaching departments and affiliated colleges taken together (in other words called pooled estimate).

The test statistic is given by:

$$Z = \frac{(p - p_1)}{S.E.\text{of}\,(p - p_1)} \sim N(0,1)$$

where, $S.E.\text{of}\,(p - p_1) = \sqrt{\left(\frac{pq}{n_{1+n_2}} \times \frac{n_2}{n_1}\right)}$ and $p = \hat{P} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{400 \times 0.75 + 500 \times 0.60}{400 + 500}$ =0.67

$\therefore \quad S.E.\text{of}\,(p - p_1) = \sqrt{\left(\frac{pq}{n_{1+n_2}} \times \frac{n_2}{n_1}\right)} = \sqrt{\left(\frac{0.67 \times 0.33}{400 + 500} \times \frac{500}{400}\right)} = 0.018.$

$\Longrightarrow \quad Z = \frac{(p - p_1)}{S.E.\text{of}\,(p - p_1)} = \frac{0.67 - 0.75}{0.018} = -4.08$

$\therefore \quad |Z| = 4.08$

Conclusion: Since calculated value of |Z|>3, it is significant. Hence, we may reject our null hypothesis and we may conclude that there is a significance difference between p_1 and $\hat{P}$.

8.12. SAMPLING OF VARIABLES

By sampling of variables, we mean the variables that can measured in numbers, for example, weight, height, age, income, expenditure etc. In sampling of variables, each member of the population under study provides the value of the variable and the total of these values forms the frequency distribution for the population

under study. A random sample of size *'n'* is then taken from the population using any of the sampling techniques, which is same as choosing a sample of size n from a given variable from the distribution.

8.12.1. UNBIASED ESTIMATE OF POPULATION MEAN AND VARIANCE

Consider a random sample $x_1, x_2, \dots, x_n$ of size n from a population of size *N* (here, *N* is assumed to be large) with mean μ and variance σ^2.

Then, the sample mean $\bar{x}$ and sample variance S^2 are given by:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \dots (1)$$

and, $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ … (2)

Taking expectation on both sides of (1), we get,

$$E(\bar{x}) = E\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(x_i)$$

Since $x_i, (i = i, 2, \dots, n)$ is a random sample from Population $X_i, (i = 1, 2, \dots, N)$, therefore, it can take any of the values $X_1, X_2, \dots, X_N$ each with equal probability 1/N.

Therefore, $E(x_i) = \frac{1}{N}X_1 + \frac{1}{N}X_2 + \cdots + \frac{1}{N}X_N = \frac{1}{N}(X_1 + X_2 + \cdots + X_N) = \mu$ … (3)

$\Rightarrow E(\bar{x}) = \frac{1}{n}\sum_{i=1}^{n}\mu = \frac{1}{n}\, n\mu = \mu \Rightarrow E(\bar{x}) = \mu$ … (4)

Hence, the sample mean is an unbiased estimator of population mean.

Next, on taking expectation on both sides on (2), we have,

$E(s^2) = E\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right] = E\left(\frac{1}{n}\sum_{i=1}^{n} {x_i}^2 - \bar{x}^2\right) = \frac{1}{n}\sum_{i=1}^{n} E({x_i}^2) - E(\bar{x}^2)$ … (5)

We have, $V(x_i) = E\big(x_i - E(x_i)\big)^2 = E(x_i - \mu)^2$

$= \frac{1}{N}\{(X_1 - \mu)^2 + (X_2 - \mu)^2 + \dots + (X_N - \mu)^2\} = \sigma^2$ … (6)

And, $V(x) = E(x^2) - \{E(x)\}^2 \Rightarrow E(x^2) = V(x) + \{E(x)\}^2$ … (7)

Or, in particular, $E({x_i}^2) = V(x_i) + \{E(x_i)\}^2 = \sigma^2 + \mu^2$ … (8)

From (7), we can get, $E(\bar{x}^2) = V(\bar{x}) + \{E(\bar{x})\}^2$ … (9)

Also, $V(\bar{x}) = \frac{\sigma^2}{n}$, on substituting in (9), we get, $E(\bar{x}^2) = \frac{\sigma^2}{n} + \mu^2$, where, σ^2 is the population variance.

Now, substituting $E(x_i{}^2)$ and $E(\bar{x}^2)$ in (5), we have,

$$E(s^2) = \frac{1}{n}\sum_{i=1}^{n}(\sigma^2 + \mu^2) - \left(\frac{\sigma^2}{n} + \mu^2\right) = \frac{n-1}{n}\sigma^2 \quad \ldots (10)$$

Therefore, sample variance is not an unbiased estimator of population variance.

Rearranging (10), we get, $\frac{n-1}{n}E(s^2) = \sigma^2$. Therefore, $E\left(\frac{n\,s^2}{n-1}\right) = \sigma^2$.

$E\left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2\right] = \sigma^2$ i.e., $E(S^2) = \sigma^2$, where, $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ is an unbiased estimator of population variance σ^2.

Remark: Standard Error of sample mean: Let n is the size of the random sample. The variance of the sample mean is $\frac{\sigma^2}{n}$, where, s is the population standard deviation. The standard error of mean of the random sample from population with variance σ^2 is $\frac{\sigma}{\sqrt{n}}$.

8.12.2. TEST OF SIGNIFICANCE OF SINGLE MEAN

Suppose a random sample $x_i, (i = i, 2, \ldots, n)$ of size n is drawn from a normal population of size *N* with mean μ and variances σ^2. Then, we have to prove that the sample mean $\bar{x}$ is normally distributed with mean μ and variance σ^2/n. For large samples, the standard normal variate corresponding to $\bar{x}$ is:

$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \quad \sim N(0,1)$$

Here, the null hypothesis is that: there is no significant difference between sample mean x ¯and population mean (μ).

Note:1) The 95% Confidence interval is given by: $\bar{x} \pm 1.96 \times \frac{\sigma}{\sqrt{n}}$

2) The 99% Confidence interval is given by: $\bar{x} \pm 2.58 \times \frac{\sigma}{\sqrt{n}}$

Example 1 : A sample of 900 members is selected with standard deviation 2.61cms. and mean 3.4 cms. Test whether the sample is selected from a large population of standard deviation 2.61 cms. and mean 3.25 cms.? Use 5% level of significance.

Solution: Given that: $n = 900, \bar{x} = 3.4\ cms$; $s = 2.61\ cms$; $\mu = 3.25\ cms$; $\sigma = 2.61\ cms$

Under the null hypothesis, H_0: $\mu = 3.25\ cms$

Against, H_1: $\mu \neq 3.25\ cms$ (two-tailed).

The test statistic is given by:

$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$$

$$= \frac{3.40 - 3.25}{\frac{2.61}{\sqrt{900}}} =$$

Conclusion: Since, calculated value of Z is less than 3 i.e. |Z|<1.96, we may accept our null hypothesis and we may conclude that μ=3.25 *cms*.

8.12.3. TEST OF SIGNIFICANCE FOR DIFFERENCE OF MEANS

Suppose two independent random samples x_i(i=1, 2, …, n_1) and y_j(j=1, 2, …, n_2) of sizes n_1 and n_2 have been drawn from two normal populations with means μ_1, μ_2 and variances ${\sigma_1}^2$, ${\sigma_2}^2$ respectively.

For large samples, $\bar{x}_1 \sim N\left(\mu_1, \frac{{\sigma_1}^2}{n_1}\right)$ and $\bar{x}_2 \sim N\left(\mu_2, \frac{{\sigma_2}^2}{n_2}\right)$. Also, $(\bar{x}_1 - \bar{x}_2)$ is also a normal variate because $\bar{x}_1$ and $\bar{x}_2$ are the independent normal variates.

Under the null hypothesis (H_0) that there is no significant difference between the two sample means and under that assumption that the population variances are equal i.e. ${\sigma_X}^2 = {\sigma_Y}^2 = \sigma^2$ (say), the test statistic is given by:

$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sigma\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0,1)$$

Under H_0: $\mu_1 = \mu_2$, then

$$Z = \frac{(\bar{x}_1 - \bar{x}_2)}{\sigma\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim$$

where, $\bar{x} = \frac{1}{n_1}\sum_{i=1}^{n_1} x_i$ is the sample mean of random sample X and $\bar{y} = \frac{1}{n_2}\sum_{j=1}^{n_2} y_j$ is the sample mean of random sample Y.

For, ${\sigma_1}^2 \neq {\sigma_2}^2$, the test statistic is given by:

$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}\right)}} \sim N(0,1)$$

Under H_0: $\mu_1 = \mu_2$,

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

In case the population variances σ_i^2 are unknown, then the sample estimates s_i^2 can be used. Then, the test statistic will be: $Z = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\left(\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}\right)}}$

Remarks: 1) The 95% Confidence interval is given by: $(\bar{x}_1 - \bar{x}_2) \pm 1.96 \times \sqrt{\left(\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}\right)}$

2) The 99% Confidence interval is given by: $(\bar{x}_1 - \bar{x}_2) \pm 2.58 \times \sqrt{\left(\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}\right)}$

Example 1: Two independent random samples of sizes 400 and 100 were drawn from two universities and from the data of their weights (in kgs.) means and standard deviations are calculated.

	Sample Sizes	Mean	S.D.
University A	400	55	10
University B	100	57	15

Test the hypothesis that there is no significant difference between the means at 5% level of significance.

Solution: Given that: n_1= 400 and n_2= 100; $\bar{x}_1$= 55 and $\bar{x}_2$= 57; s_1= 10 and s_1= 15.

Set up the null hypothesis that: there is no significant difference between two sample means.

Against, the alternative hypothesis that: there is a significant difference between two sample means.

Since, the population variances are unknown, then the test statistic is given by:

$$Z = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\left(\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}\right)}} \sim N(0,1)$$

$$= \frac{(55 - 57)}{\sqrt{\left(\frac{10^2}{400} + \frac{15^2}{100}\right)}}$$

Since, calculated value of $|Z|$<1.96, it is not significant. We may accept our null hypothesis and therefore, we may conclude that there is a no significant difference between two sample means.

Example 2: In Firm A, the average hourly wage of a sample of 150 workers was Rs. 2.56 and S.D. is Rs. 1.08. In Firm *B*, the average hourly wage of a sample of 200 workers was Rs. 2.87 and S.D. is Rs. 1.28. Test the assumption that the hourly wage paid *B* are more than the hourly wage paid by Firm *A*.

Solution: Given that: n_1= 150 and n_2= 200; $\bar{x}_1$= 2.56 and $\bar{x}_2$= 2.87; s_1= 1.08 and s_2 = 1.28

Set up the null hypothesis that: there is no significant difference between the average hourly wage paid

by Firm *A* and Firm *B*, i.e., H0: $\mu_1 = \mu_2$.

Against, the alternative hypothesis that: the average hourly wage paid by Firm *B* is more than the average hourly wage paid by Firm *A*, i.e., H_1: $\mu_1 < \mu_2$.

The test statistic is given by:

$$Z = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\left(\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}\right)}} = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\left(\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}\right)}} \sim N(0,1)$$

$$= \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\left(\frac{1.08^2}{150} + \frac{1.28^2}{200}\right)}} = -2.46$$

Decision: For one-tailed test, the significant value of Z at 5% level of significance is 1.645.

Since the calculated value of |Z| is greater than the significant value, we reject or null hypothesis. Hence, we may conclude that the average hourly wage paid by Firm *B* is more than the average hourly wage paid by Firm *A*.

8.12.4. TEST OF SIGNIFICANCE FOR THE DIFFERENCE OF STANDARD DEVIATIONS

Suppose two independent random samples x_i(i=1, 2, …, n_1) and y_j(j=1, 2, …, n_2) of sizes n_1 and $n_2$with variances s_1 and s_2 respectively have been drawn from two normal populations. Then, under the null hypothesis that there is no significant difference between two population standard deviations, i.e., H_0: $\sigma_1 = \sigma_2$. Against the alternative hypothesis, H_1: $\sigma_1 \neq \sigma_2$.

The test statistic is given by:

$$Z = \frac{s_1 - s_2}{S.E.(s_1 - s_2)} \sim N(0,1), \text{ for large samples.}$$

$$= \frac{s_1 - s_2}{\sqrt{\left(\frac{{\sigma_1}^2}{2n_1} + \frac{{\sigma_2}^2}{2n_2}\right)}} \sim N(0,1)$$

If ${\sigma_1}^2$ and ${\sigma_2}^2$ are unknown then for large samples we can use their sample estimates of variances ?s_1?^2 and ${s_1}^2$ and ${s_2}^2$.. Then, the test statistics will be:

$$Z = \frac{s_1 - s_2}{\sqrt{\left(\frac{{s_1}^2}{2n_1} + \frac{{s_2}^2}{2n_2}\right)}} \sim N(0,1), \text{ for large samples.}$$

Remarks: 1) The 95% Confidence interval is given by: $(s_1 - s_2) \pm 1.96 \times \sqrt{\left(\frac{{\sigma_1}^2}{2n_1} + \frac{{\sigma_2}^2}{2n_2}\right)}$

2) The 99% Confidence interval is given by: $(s_1 - s_2) \pm 1.96 \times \sqrt{\left(\frac{{\sigma_1}^2}{2n_1} + \frac{{\sigma_2}^2}{2n_2}\right)}$

Example 1: A large organization has two Factories *A* and *B*. Factory *A* produces 100 electric bulbs and Factory *B* produces 200 electric bulbs. The manager of the organization suspected that the efficiency of two factories is not same. He carried out a test to check the variability of the life of the bulbs produced by each factory. The results are given below:

	Factory A	Factory B
No. of bulbs in sample	100	200
Average Life	1,100 hrs.	900 hrs.
Standard Deviation	240 hrs.	220 hrs.

Test the hypothesis that the difference between the variability of the life of bulbs produced in each factory is significant. Use 1% level of significance.

Solution: Given that: n_1= 100 and n_2= 200; $\bar{x}_1$= 1,100 hrs. and $\bar{x}_2$= 900 hrs.; s_1 = 240 hrs. and $s_2 = 220$ hrs.

Set up the null hypothesis that there is no significance difference between the variability of the life of bulbs produced in each factory, i.e., H_0: $\sigma_1 = \sigma_2$.

Against, the alternative hypothesis that there is a significance difference between the variability of the life of bulbs produced in each factory, i.e., H_0: $\sigma_1 \neq \sigma_2$.

The test statistic under null hypothesis is given by:

$$Z = \frac{240 - 220}{\sqrt{\left(\frac{240^2}{2 \times 100} + \frac{220^2}{2 \times 200}\right)}} \sim N(0,1)$$

$$= \frac{240 - 220}{\sqrt{\left(\frac{240^2}{2 \times 100} + \frac{220^2}{2 \times 200}\right)}} = 0.9889$$

Conclusion: Since calculated value of *Z* is less than 2.58 (at 1% level of significance), we may accept our null hypothesis and we may conclude that there is no significance difference between the variability of the life of bulbs produced in Factory *A* and Factory *B*.

8.13. ANALYSIS OF VARIANCE

Variance is defined as a measurement of dispersion that gives the quantitative value for the amount of variation from the population mean, a particular sample mean shows or the amount of variation shown by the data points in a group from their group mean. Variance can also be termed as the ratio of the squared

sum of differences of values from their mean to the total number of values. It is the average squared sum of difference from means of a sample.

In some decision-making situations, sample data may be divided into various groups i.e. sample may be supposed to have consisted of *k* sub-samples. There are interest lies in examining whether the total sample can be considered as homogeneous or there is some indication that sub-samples have been drawn from different populations. So, in these situations, we have to compare the mean values of various components groups, with respect to one or more criteria.

The total variation present in a data set of data may be partitioned into a number of non-overlapping components as per the nature of the classification. The systematic procedure to achieve this is called *analysis of variance*. The *analysis of variance* is a powerful statistical tool for tests of significance. The test of significance based on t-distribution is an adequate procedure only for testing the significance of the difference between 2 sample means. In a situation when we have three or more samples to consider at a time an alternate procedure is needed for testing the hypothesis that all the samples are drawn from the same population i.e., they have the same means. If we comparing two means ANOVA will produce the same results as t-test for independent (dependent) samples.

Analysis of variance involve the use of variance for testing the significance of the difference between two or more samples under study. ANOVA was developed by statistician and evolutionary biologist Prof. Ronald A. Fisher in 1920's to deal with the problem in the agronomical data. Variations is inherent in nature. In any set of numerical data, the total variation is due to number of causes which may be classified as (i) Assignable causes, and (ii) chance causes. Variation due to assignable causes can be detected and measured but the variations due to the chance cause cannot be detected and is beyond the control of human hands and hence cannot be traced separately. According to Prof. R. A. Fisher, "Analysis of Variance (ANOVA) is the "Separation of variance describable to one group of causes from the variance ascribable to other group". By the technique of ANOVA, the total variation in the sample data can be expressed as the sum of its non-negative components where each of these components is a measure of the variation due to some specific independent source or factor or cause.

Unlike t/z tests, the test carried out is called the F-test that can be done even when they are more than two groups of samples drawn from the population. It is more useful method than t/z test because this shows the possibility of the interaction effect amongst the two types of independent variables chosen which cannot be calculated using t/z test. The name is derived from the fact that in order to test for statistical significance between means, we are actually analyzing (comparing) variances. A response variable related to one or more explanatory variables, usually is categorical.

It should be clearly understood that ANOVA does not compare several variances. It only compares several means simultaneously.

8.13.1 ASSUMPTIONS FOR ANOVA

Analysis of Variance test is based on the test statistics F (called Variance Ratio). So, for the validity of F-test following assumptions are made:

i) Parent population from which the sample has been drawn is normal.

ii) The sample observations are independent.

iii) Various treatments and environmental effects are additive in nature.

iv) Homoscedastity indicates that variance of several groups (treatments) are homogenous.

ANOVA can be classified as one-way ANOVA or two-way ANOVA, depending upon the one-way classification of data and two-way classification of data respectively.

8.13.2. ADVANTAGES OF ANOVA

1. It is a powerful technique to compare several populations means simultaneously and thus results in saving of time and money as compared with other techniques available for comparing the several means.
2. It consists in classifying as well as cross classifying statistical data and results in testing if the means of a specified classification differ with each other or not.
3. It enables us to determine the importance of particular classification.
4. It is also used in testing the linearity of fitted regression line or to test the significance of the correlation ratio.

8.13.3. ONE-WAY CLASSIFICATION

Let us consider a random sample Y of N observations y_{ij}, $(i=1, 2, ..., k\ ;\ j=1, 2, ..., n_i)$. Suppose these N observations of random sample Y are grouped on some basis into k classes of sizes $n_1, n_2, ..., n_k$ respectively. Here, $N = \sum_{i=1}^{k} n_i$.

In ANOVA, the total variation in the observations y_{ij} can be divided into two components as follows:

1. The variations due to the different bases of classification, known as treatments i.e., the variations *between the classes.*
2. The inherent variability of the random variables within the class observations i.e., the variations *within the classes.*

One-Way ANOVA Table

Class	Sample observations				Total	Mean
1	y_{11}	y_{12}	$\cdots$	y_{1n_1}	$T_{1.}$	$\bar{y}_{1.}$
2	y_{21}	y_{22}	$\cdots$	y_{2n_2}	$T_{2.}$	$\bar{y}_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$	$\vdots$
i	y_{i1}	y_{i2}	$\cdots$	y_{in_i}	$T_{i.}$	$\bar{y}_{i.}$
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$	$\vdots$
k	y_{k1}	y_{k2}	$\cdots$	y_{kn_k}	$T_{k.}$	$\bar{y}_{k.}$

The variations between the classes are due to assignable causes and can be detected and eliminated or controlled by human endeavour. The variations *within the classes* are due to the chance causes which cannot be detected and if detected are beyond human control. Therefore, ANOVA technique is used to examine the variations *between the classes* and test if there is any significant difference between the class means because the different classes have inherent variability within the separate classes.

The mathematical model for one-way ANOVA is given by:

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}; \quad i = 1, 2, \dots, k; j = 1,2, \dots, n_i \qquad \dots \text{(A)}$$

where,

- y_{ij} is the yield from the i^{th} treatment in j^{th} row,
- μ is the general mean effect (called fixed effect). It means when there are no chance causes or no any treatment differences then the yield of each treatment and is given by,

$$\mu = \sum_{i=1}^{k} \frac{n_i \mu_i}{N}$$

- α_i is the i^{th} the ith treatment and ε_{ij} is the error due to chance and is defined by:

$$\alpha_i = \mu_i - \mu, (i = 1, 2, \dots, k)$$

- ε_{ij} is the error due to the chance.

8.13.4. FIXED EFFECT MODEL AND RANDOM EFFECT MODEL

Fixed Effect Model. Suppose out of a large number of classes (treatments), only the k-classes (treatments) in the model (A) have been particularly chosen by the experimenter. Here, in this case *a*'s are fixed or unknown constant and then the model (A) is known as the fixed effect model. In this model we would like to estimate *a*'s and test some hypothesis about *a*'s. In the fixed effect model, only to the k-treatments (factor levels) considered in the experiment and the conclusions about the test of hypothesis regarding

the parameters *a*'s will apply only on the selected treatments. These conclusions, therefore, cannot be applied to those treatments (factor levels) which were not considered in the experiment.

Random Effect Model. Suppose we have a large number of classes or levels of factor under consideration and we wish to test if all these class effects are equal or not. Here, in this case, null hypothesis is of the homogeneity of class treatment effects. It may not be possible to include all the factor-levels in the experiment due to consideration of money, time or administrative convenience. In such a case, we consider only a random sample of classes in the experiment and after examining the sample data, we draw conclusions based on sample information which would be valid for all the classes or factor levels and decide whether included in the experiment of not. In such a situation the parameters *a*'s in the model (A) will not be considered as fixed constants but will be random samples and hence, the model is known as random effect model.

Note that, in the random effect model if the null hypothesis is rejected, then we cannot apply the t-test to test the significance of the difference between two class (treatment) effects because all the treatments are not included in the experiment.

8.13.5. ANOVA FOR FIXED EFFECT MODEL

If out of a large number of treatments, only the k-classes (treatments) in the model (A) are of interest, then the Fixed Effect Model is given by:

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}; \quad i = 1, 2, \dots, k; j = 1,2, \dots, n_i \qquad \dots \text{(A)}$$

where, α_i's are the Fixed constants.

Assumptions of Fixed Effect Model:

- All the sample observations are independent and follows $N(\mu_i, {\sigma_e}^2)$.
- All the effects are additive in nature.
- ε_{ij}, the error term, are *i.i.d.*$N(0, {\sigma_0}^2)$, which means $E(\varepsilon_{ij}) = 0$ and $V(\varepsilon_{ij}) = 0\ \forall i$ and .

Statistical Analysis of model (A):

We want to test the equality of various population means.

Set up the null hypothesis, H_0: $\mu_1 = \mu_2 = \cdots = \mu_k = \mu$,

or, H_0: $\alpha_1 = \alpha_2 = \cdots = \alpha_k$

Against the Alternative hypothesis: At least two treatment means $\mu_1, \mu_2, \dots, \mu_k$ differ significantly.

Let us write, $\bar{y}_{i.}$= mean of the *i*th class = $\sum_{j=1}^{n_i} \frac{y_{ij}}{n_i}$; *i=1, 2, ..., k.*

and, Grand total (*G*)=$\sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij}$

therefore, $\bar{y}_{..}$= overall mean = $\frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij} = \frac{1}{N} \sum_{i=1}^{k} n_i \bar{y}_{i.}$

Correction Factor *(C.F.)* = G^2/N.

Raw Sum of Squares (*R.S.S.*) = $\sum_i \sum_j y_{ij}{}^2$

Total Sum of squares *(T.S.S)* = *R.S.S.*− *C.F* = $\sum_i \frac{T_{i.}{}^2}{n_i} - C.F.$

Error Sum of Squares (S.S.E) = *T.S.S.*− *S.S.T.*

Next, we prepare an ANOVA table which is given below

ANOVA Table

Sources of Variations	Degrees of Freedom	Sum of Squares	Mean Sum of Squares	F-Ratio
(1)	(2)	(3)	(4) = (3)÷(2)	
Treatment	*k-1*	S.S.T.	M.S.S.T.	F=M.S.S.T./M.S.S.E. $\sim F(k-1, N-k)$

At last, we present our conclusion. If calculated value of *F* comes out to be more than the tabulated value of *F* at a level of significance for $(k-1, N-k)$ degrees of freedom, we reject our null hypothesis otherwise we accept it.

Note: *If null hypothesis is accepted then we can conclude our result by giving approval to null hypothesis. But if null hypothesis is rejected then we have to proceed further to check which pair of treatment means differ significantly. For this we have to find the* ***'Least Significant Difference'*** *or* ***'Critical Difference'*** *between two pair of means.*

Example 1: A production company has purchased three new machines A_1, A_2 and A_3 of different makes. The manufacturer wants to determine the efficiency of the machines for manufacturing the units. From each machine, a five hourly production figure are observed at random. The results are given in the adjoining table. Use ANOVA technique and determine whether the three machines A_1, A_2 and A_3 are significantly different in their average speeds. Use 5% level of significance.

	Machine		
	A_1	A_2	A_3
Observations	25	31	24
	30	39	30
	36	38	28
	38	42	25
	31	35	28

Solution: Given that $n_1 = n_2 = n_3 = 5; k = 3; N = n_1 + n_2 + n_3 = 15$

Here, the machine A_1, A_2 and A_3 are the factors of variations. We set up the null hypothesis that all the machines are equally efficient i.e., $H_0: \mu_1 = \mu_2 = \mu_3$.

Against the alternative hypothesis, H_1: at least two of the means are not equal.

Calculation table is given by:

Machine	Sample observations (y_{ij})					Total	
A_1	25	30	36	38	31	$T_1. = 160$	$T_1.^2 = 25,600$
A_2	31	39	38	42	35	$T_2. = 185$	$T_2.^2 = 34,225$
A_3	24	30	28	25	28	$T_3. = 135$	$T_3.^2 = 18,225$
Total						$G = \sum_i \sum_j y_{ij} = 480$	$\sum_I T_{i.}^2 = 78,050$

Raw Sum of Square (R.S.S.) $= \sum_i \sum_j y_{ij}^2 = 25^2 + 30^2 + \cdots + 28^2 + 25^2 + 28^2 = 15,810$.

Grand Total, $G = \sum_i \sum_j y_{ij} = 160+185+135 = 480$

Correction Factor, $C.F. = \frac{G^2}{n} = \frac{(480)^2}{15} = \frac{230400}{15} = 15360$

Total S.S.(T.S.S.) = R.S.S. – C.F.=15, 810 – 15, 360 = 450.

Treatment Sum of Squares (S.S.T.) $= \sum_i \frac{T_i.^2}{n_i} - \text{C.F.} = \left(\frac{T_1.^2}{n_1} + \frac{T_2.^2}{n_2} + \frac{T_3.^2}{n_3}\right) - \text{C.F}$

$$= \frac{78,050}{5} - 15,360 = 250$$

Error Sum of Squares (S.S.E.) = T.S.S. − S.S.T. = 450 −250 = 200

ANOVA Table

Sources of Variations	Degrees of Freedom	Sum of Squares	Mean Sum of Squares	F-Ratio
(1)	(2)	(3)	(4) = (3) ÷(2)	
Treatment (Machines)	3-1=2	250	250/2=125	F = 125/16.67 = 7.4985 ~ $F(2, 12)$

The tabulated value of F for (2, 12) *d.f.* at 5% level of significance is 3.89.

Conclusion: Since, calculated value of F is more than the tabulated value of F at 5% level of significance for (2, 12) degrees of freedom, we reject our null hypothesis. Hence, we may conclude that the machines differ significantly.

Since, in the present problem, null hypothesis is rejected, so we will find the Least Significant Difference between the pair of treatment means.

The standard error of the difference between any two treatment means $\bar{y}_{i.}$ and $\bar{y}_{j.}$ is given by:

$$S.E.\left(\bar{y}_{i.}-\bar{y}_{j.}\right)=\sqrt{M.S.S.E.\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}=\sqrt{{s_e}^2\left(\frac{1}{5}+\frac{1}{5}\right)}=\sqrt{\frac{2\times 16.67}{5}}=2.58$$

Therefore, Least Significant Difference (*L.S.D.*) $=t_{N-k}\left(\frac{\alpha}{2}\right)\times S.E.\left(\bar{y}_{i.}-\bar{y}_{j.}\right)$

$$=t_{12}(0.025)\times 2.58=2.18\times 2.58=5.62$$

Table for Comparison of treatment means

Treatment (Machine)	Mean hourly output	Difference	Comparison with L.S.D.	Result
A_1	$\bar{y}_{1.}$=160/5=32	$\lvert\bar{y}_{1.}-\bar{y}_{2.}\rvert=5$	< 5.62	Not Significant
A_2	$\bar{y}_{2.}$=185/5=37	$\lvert\bar{y}_{1.}-\bar{y}_{3.}\rvert=5$	< 5.62	Not Significant
A_3	$\bar{y}_{3.}$=135/5=27	$\lvert\bar{y}_{2.}-\bar{y}_{3.}\rvert=10$	> 5.62	Significant

From the table, we can see that the pair $\lvert\bar{y}_{2.}-\bar{y}_{3.}\rvert$ is significant. Therefore, we can conclude that the Machines A_2 and A_3 differ significantly.

8.14. CHECK YOUR PROGRESS 2

1. Define F-test. What are its applications?

   ……

   ……

2. Write down the relation between t & F distributions.

   ……

   ……

3. What is the relationship between F & Chi Square distribution?

   ……

   ……

4. How do you test the significance for large samples? Write down its procedure.

……

……

5. Write down the procedure to test the significance for single proportion.

……

……

6. Write down the difference between sampling of attributes and sampling of variables.

……

……

7. Define analysis of variance. What are its advantages?

……

……

8. What do you understand by one-way classification and two-way classification?

……

……

9. What are the assumptions of ANOVA? Discuss one-way ANOVA.

……

……

10. What do you mean by fixed effect model and random effect model?

……

……

8.15. LET US SUM UP

t-Test is a test for small samples and is usually used to test the significance of the hypothetical value of population mean and also test the significance of difference between means of two samples.

F-Test is used to test the significance of the difference of variances of two samples. It is also known as variance ratio test.

t-Test, F-test and Z-Test are most commonly used parametric tests.

Chi-Square test is a non-parametric test and is used to test the association between two samples. This test is important in the sense that it is also used to test goodness of fit and also test the independence of attributes.

When a statistic is used to estimate the parameter, the number of degrees of freedom depends upon the restrictions placed. Therefore, the number of degrees of freedom will vary from one statistic to another. Degrees of freedom are basically the sample size less the number of restrictions imposed.

Analysis of variance is used to test the significance of the difference between the means of more than two samples. ANOVA deals with variances rather to deal with means and their standard error of the difference exist between the means.

8.16. KEY POINTS

- **Sampling:** Sampling is a technique of selecting a sample from the population.
- **Non-Probability sampling:** In non-probability sampling, the sample units are selected by non-random method.
- **Probability sampling:** It is the sampling technique in which the samples taken from a large population are selected on the basis of probability theory.
- **Standard error:** It is defined as the standard deviation of the sampling distribution of a statistic.
- **ANOVA:** The analysis of variance is a powerful statistical tool for tests of significance. Analysis of variance involve the use of variance for testing the significance of the difference between two or more samples under study.

8.17. SELF-ASSESSMENT QUESTIONS

1. How do you compare calculated value of a statistic with the critical value?
2. What do you understand by degrees of freedom? In F-Test, which variance is placed on the numerator?
3. How many degrees of freedom are associated with the variation in the data for a comparison of four means for independent samples each containing 10 cases?
4. What is the mathematical relationship between t and F test?
5. What is the procedure to test the significance of large samples?

8.18. LESSON END QUESTIONS

Example 1: Suppose a manufacturing company wants to test the mean life of its 4 brands of tyres. The company selected the random sample of all brands. The results are shown in an adjoining table.

Brand 1	*Brand 2*	*Brand 3*	*Brand 4*
20	19	21	15
23	15	19	17
18	17	20	16
17	20	17	18
	16	16	

Test the hypothesis that the mean life of each brand type is same. Use 1% level of significance.

Example 2: Two hundred bolts were selected at random from the output of each of the five machines. The number of defective bolts found were 5, 9, 13, 7 and 6. Is there a significant difference among the machines? Test at 5% level of significance.

Example 3: A random sample of 400 oranges was taken from a large load and 40 were found to be bad. Find the standard error of the proportion of bad oranges and also find out the confidence interval for bad oranges of load lies. Use 5% level of significance.

Example 4: A cigarette manufacturing company claims that the brand A of the cigarette outsells the brand B by 9%. 40 out of the sample of 200 smokers prefer brand A and 20 out of another random sample of 120 smokers prefer brand B. Use 5% level of significance to test whether the claim is correct or not.

Example 5: Random samples of students were drawn from two different colleges and from their weights (in kgs.), standard deviations and means were calculated. Test the significance of the difference between the two means.

	Mean	S.D.	Sample size
College A	6	2	10
College B	9	3	15

Use 5% level of significance to test the assumption.

Example 6: A random sample of heights of 6 soldiers are (in inches): 63, 62, 65, 68, 71, 72 and another random sample of heights of 10 sailors are 62, 61, 66, 71, 72, 68, 69, 73, 67, 70. Test whether if there is any significant difference between the average heights of soldiers and sailors. Use 5% level of significance.

Example 7: A *LIC* agent has claimed that he has insured policy holders whose average age is less than 30 years old. The following data of age distribution shows the random sample of 100 policy holders who had insured through him:

Age on last birthday	16-20	21-25	26-30	31-35	36-40
No. of persons	10	24	22	30	16

Test whether his claim is true or not? Use 5% level of significance.

8.19 SUGGESTED READINGS

- Goon, A.M., Gupta, M.K. and Dasgupta, B. (1968). Fundamentals of Mathematical Statistics, Vol II, World Press, Kolkata.
- Gupta, S.C., & Kapoor, V.K. (2020). Fundamental of Mathematical Statistics. Sultan Chand and Sons.
- Gupta, S.C., & Kapoor, V.K. (2020). Fundamental of Applied Statistics. Sultan Chand and Sons.
- Gupta, S.P. (2021). Statistical Methods. 46th ED., Sultan Chand and Son

• • •

t-test table

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence Level										

Appendix - B

Chi-square distribution Table

df	χ^2 .995	χ^2 990	χ^2 .975	χ^2 .950	χ^2 .900	χ^2 .100	χ^2 .050	χ^2 .025	χ^2 .010	χ^2 .005
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154	79.490
60	35.534	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	140.169

Appendix – C

F-distribution (Upper tail probability = 0.05)

df_1=1	2	3	4	5	6	7	8	9	10	12
df_2=1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40
3	10.13	9.552	9.277	9.117	9.014	8.941	8.887	8.845	8.812	8.786
4	7.709	6.944	6.591	6.388	6.256	6.163	6.094	6.041	5.999	5.964
5	6.608	5.786	5.409	5.192	5.050	4.950	4.876	4.818	4.772	4.735
6	5.987	5.143	4.757	4.534	4.387	4.284	4.207	4.147	4.099	4.060
7	5.591	4.737	4.347	4.120	3.972	3.866	3.787	3.726	3.677	3.637
8	5.318	4.459	4.066	3.838	3.688	3.581	3.500	3.438	3.388	3.347
9	5.117	4.256	3.863	3.633	3.482	3.374	3.293	3.230	3.179	3.137
10	4.965	4.103	3.708	3.478	3.326	3.217	3.135	3.072	3.020	2.978
11	4.844	3.982	3.587	3.357	3.204	3.095	3.012	2.948	2.896	2.854
12	4.747	3.885	3.490	3.259	3.106	2.996	2.913	2.849	2.796	2.753
13	4.667	3.806	3.411	3.179	3.025	2.915	2.832	2.767	2.714	2.671
14	4.600	3.739	3.344	3.112	2.958	2.848	2.764	2.699	2.646	2.602
15	4.543	3.682	3.287	3.056	2.901	2.790	2.707	2.641	2.588	2.544
16	4.494	3.634	3.239	3.007	2.852	2.741	2.657	2.591	2.538	2.494
17	4.451	3.592	3.197	2.965	2.810	2.699	2.614	2.548	2.494	2.450
18	4.414	3.555	3.160	2.928	2.773	2.661	2.577	2.510	2.456	2.412
19	4.381	3.522	3.127	2.895	2.740	2.628	2.544	2.477	2.423	2.378
20	4.351	3.493	3.098	2.866	2.711	2.599	2.514	2.447	2.393	2.348
21	4.325	3.467	3.072	2.840	2.685	2.573	2.488	2.420	2.366	2.321
22	4.301	3.443	3.049	2.817	2.661	2.549	2.464	2.397	2.342	2.297
23	4.279	3.422	3.028	2.796	2.640	2.528	2.442	2.375	2.320	2.275
24	4.260	3.403	3.009	2.776	2.621	2.508	2.423	2.355	2.300	2.255
25	4.242	3.385	2.991	2.759	2.603	2.490	2.405	2.337	2.282	2.236
26	4.225	3.369	2.975	2.743	2.587	2.474	2.388	2.321	2.265	2.220
27	4.210	3.354	2.960	2.728	2.572	2.459	2.373	2.305	2.250	2.204
28	4.196	3.340	2.947	2.714	2.558	2.445	2.359	2.291	2.236	2.190
29	4.183	3.328	2.934	2.701	2.545	2.432	2.346	2.278	2.223	2.177
30	4.171	3.316	2.922	2.690	2.534	2.421	2.334	2.266	2.211	2.165
40	4.085	3.232	2.839	2.606	2.449	2.336	2.249	2.180	2.124	2.077
60	4.001	3.150	2.758	2.525	2.368	2.254	2.167	2.097	2.040	1.993
120	3.920	3.072	2.680	2.447	2.290	2.175	2.087	2.016	1.959	1.910
Inf	3.842	2.996	2.605	2.372	2.214	2.099	2.010	1.938	1.880	1.831