EDU-3016

GAUHATI UNIVERSITY

Centre for Distance and Online Education

M.A. Third Semester

(Under CBCS)

EDUCATION

Paper: EDU 3016
EDUCATIONAL STATISTICS

Contents:

Block- I : **Normal Distribution**

Block- II : **Regression and Correlation**

Block- III : **The Significance of the Other Statistics and the Difference between Means**

Block- IV : **Analysis of Variances**

Block- V : **Chi Square Test**

SLM Development Team:

Head, Department of Education, G.U.

Programme Coordinator, MA in Education, GUCDOE

Dr. Manoj Kr. Nayak, Assistant Professor, GUCDOE

Course Coordination:

Dr. Debahari Talukdar	Director, GUCDOE
Dr. Purabi Baishya	Programme Coordinator, GUCDOE,
Dr. Manoj Kr. Nayak	Assistant Professor, GUCDOE
Dipankar Saikia	Editor SLM, GUCDOE

Contributors:

Taslima Nasrin (Block- I: Unit-1 & 2)
Assistant Professor, GUCDOE

Dr. Manoj Kr. Nayak (Block- I: Unit-3 & 4)
Assistant Professor, Dept. of Education, GUCDOE

Dr. Nabanita Sarmah (Block- II)
Assistant Professor, Dept. of Education
Biswanath College

Dr. Prasanta Khanikar (Block- III & IV)
HoD, Dept. of Education, Nawgaon College

Dr. Marami Goswami (Block- V)
Professor & HoD, Dept. of Education
Dr. ABS College of Education (NERIM)

Content Editing:

Dr. Sima Kalita (Block- I, II & V)
Assistant Professor, Dept. of Education
Gauhati University

Dr. Ajaya Kumar Mohanty (Block- III & IV)
Assistant Professor, Dept. of Education
Ravanshaw University, Cuttack, Odisha

Cover Page Design & Type Setting:

Bhaskar Jyoti Goswami	GUCDOE
Nishanta Das	GUCDO

ISBN: 978-81-989447-0-2

June, 2025

CONTENTS:

Block- I: Normal Distribution **(Page: 04-73)**

Unit 1 : Normal Probability Curve

Unit 2 : Properties of Normal Probability Curve

Unit 3 : Measures of Asymmetry (Skewness and Kurtosis)

Unit 4 : Applications of Normal Probability Curve

Block- II: Regression and Correlation **(Page: 74-110)**

Unit 1 : Regression Equation, Regression and Prediction

Unit 2 : Product Moment Correlation and Scatter Diagram

Block- III: The Significance of the Other Statistics and the Difference between Means **(Page: 111-198)**

Unit 1 : Statistical Inferences

Unit 2 : Standard Error of Mean of Large and Small Sample

Unit 3 : Significance of Mean and Other Statistics

Unit 4 : Paramatric Test

Unit 5 : The Hypotheses

Unit 6 : Significance of the Difference Between Means

Block- IV: Analysis of Variances **(Page: 199-260)**

Unit 1 : Meaning, Natur and Uses of Analysis of Variance

Unit 2 : One Way and Two Way Analysis of Variance

Unit 3 : Difference Between One Way and Two Way Analysis of Variance

Unit 4 : Steps for Calculating The Analysis of Variance

Block- V: Chi Square Test **(Page: 261-373)**

Unit 1 : Non-Parametric Test Introduction to Chi-Square Test

Unit 2 : Assumptions of Chi-Square Tests

Unit 3 : Uses and Significance of Chi-Square Test

Unit 4 : Testing Null Hypothesis of Independence in 2×2 Contingency Table

BLOCK- I: NORMAL DISTRIBUTION

Unit 1 : Normal Probability Curve

Unit 2 : Properties of Normal Probability Curve

Unit 3 : Measures of Asymmetry (Skewness and Kurtosis)

Unit 4: Applications of Normal Probability Curve

UNIT-1
NORMAL PROBABILITY CURVE

Unit Structure:

1.1 Introduction
1.2 Objectives
1.3 Meaning of Normality
1.4 Meaning of Probability
1.5 Normal Probability Curve
1.6 Characteristics/Nature of Normal Probability Curve
1.7 Importance of Normal Probability Curve
1.8 Summing Up
1.9 Questions and Exercise
1.10 Answer to Check Your Progress
1.11 References and Suggested Readings

1.1 Introduction

As we know, Statistics helps people make informed decisions using various types of data. To understand and describe data, we use histograms, frequency polygons, measures of central tendency and variability. Similarly, we also need procedures to describe individual scores to categorize the entire group of individuals on the basis of their abilities which a teacher has used to assess the outcomes of the individual on a certain ability test. The related distribution of scores through which the bell shaped curve is obtained is generally known as Normal Distribution, comes in. These concepts are crucial. In this unit, we will be focussing on meaning, characteristics/nature, and importance of normal probability curve.

1.2 Objectives

After going through this unit, you will be able to-

- understand the meaning of normality and probability
- understand the meaning of Normal Probability Curve
- Know the characteristics/nature of Normal Probability Curve.
- Discuss the importance of normal probability curve.

1.3 Meaning of Normality

The term normal refers to the ‘average’ or mean value, commonly used in statistical analysis, particularly in fields like education, psychology and sociology. In these areas, individuals who meet a specific standard or benchmark are considered ‘normal’, whereas those who exceed or fall short of this standard are referred to as away from normality. For example, when looking at student’s intelligence, only a few have a high IQ (above 110), and a few have a low IQ (below 90). Most of the students have an average IQ between 80 and 110. When plotting the frequency distribution of student’s scores, it is evident that most scores are concentrated around the mean. This characteristic of data being centred around the mean is referred to as normality.

1.4 Meaning of Probability

Probability is the chance, possibility or likelihood of an event to occur. Probability laws play a significant role in our daily lives and various fields. For instance, predictions related to weather or any other thing such as, “it will probably be hot tomorrow”, “Probably, I may not be able to come tomorrow” are grounded in probability laws that there are chances for the events to occur tomorrow but there is no certainty. In research and statistics, the probability theory is of great importance in making predictions. In this way, the probability of an event helps researchers create hypothesis to continue with their research and experiments and it's a starting point and a reference for drawing conclusions based on these probabilities. Probability is essential in various fields including science, finance, social sciences artificial intelligence and mathematics. By understanding probability, researchers can make informed decisions and conclusions.

Garrett's definition of probability is as follows:

> ''Probability of a given event is defined as the expected frequency of occurence of the event among events of a like sort". This means that the expected frequency of occurence may be based upon a knowledge of the conditions determining the occurence of the phenomenon as in determining or coin-tossing or upon empirical data as in mental and social measurements.

Stop to Consider

The theory of probability originated in the 17th century. Blaise Pascal and Pierre de Format are considered as the originator of probability theory. Later, in 1954, Antoine Gumband showed interest in Probability. Since then, many statisticians have built upon and refined his ideas.

1.5 Normal Probability Curve

In most graphs, scores tend to cluster around the centre and decrease towards the extremes. This pattern is common in quantitative data, where scores concentrate around the middle and taper off towards the high and low ends. If a line is drawn through the central peak, the two resulting parts will typically be symmetrical in shape and equal in area. This phenomenon is often observed in measurements of mental and social traits, which tend to be distributed symmetrical around the mean. This type of distribution is known as a normal probability curve. Data from random events, such as coin tosses or dice rolls, often forms a frequency curve that closely resembles the normal curve when plotted on a graph. This similarity is due to the probabilistic nature of these events, leading to the normal curve being commonly referred to as the normal probability curve.

Let us understand Normal Probability curve with an example- When we measure height or weight in a large group, we will find that there will be more persons with average height and few persons who are taller or shorter compared to the average. When we plot this data on a graph, it forms a normal curve.

The normal probability curve was independently developed by Laplace and Carl Friedrich Gauss (1777-1855), who were studying experimental errors in physics and astronomy. They discovered that these errors followed a normal distribution. As a result, the curve is also known as the normal curve of error, where 'error' refers to deviations from the true value. In recognition of Gauss's contribution, the curve is also referred to as the Gaussian Curve.

The distribution of data doesn't always result in a normal or approximately normal curve. In cases, where individual scores differ significantly deviate from the average, the resulting curve will also deviate from the normal curve shape. These deviations from normality can occur in two main ways:

1) Skewness
2) Kurtosis

Skewness :

The distribution in which Mean and Median are not aligned or fall on different points known as Skewed Distribution and this tendency of distributions is known as Skewness.

Skewness occurs when a curve lacks symmetry. Unlike a perfectly symmetrical normal curve, where the mean, median and mode are equal, skewed distributions have different values for these measures and are asymmetrical, leaning to the left or right.

There are two types of skewed distributions. A distribution is negatively skewed when most scores are above the average, while a distribution is positively skewed when most scores fall below the average.

Kurtosis

Kurtosis refers to the shape of a distribution. When there are very few individuals whose scores are near to the average score for their group, the curve representing such a distribution becomes 'flattened' in the middle. On the other hand, when there are too many cases in the central area, the distribution curve appears too 'peaked' in comparison to normal. These two characteristics, 'flatness' or 'Peakedness' describes kurtosis.

In the coming units, we will be discussing skewness and kurtosis more in a detailed manner.

1.6 Characteristics/Nature of Normal Probability Curve

To understand the Normal Probability Curve, it is very necessary to understand its characteristics. So, some of the main characteristics are given below-

1) **Symmetrical-** The Normal Probability Curve is symmetrical around the central point. The size, shape and slope of the

curve on one side of the curve is identical to that of the other. This means that the two sides of the curve are equal indicating that it is perfectly symmetrical.

2) **Bell-shaped**- The curve is bell-shaped with most data points concentrated around the mean. The curve is highest at the mean and it decreases gradually as we move away from the mean.

3) **Unimodal-** The Normal Probability Curve is unimodal. It has a single peak or mode, indicating that there is only one data point with the highest frequency.

4) **Mean, median and mode are the same**- The measures of central tendency mean, median and mode are all equal in a Normal Probability Curve. This means that 50% of the curve is above the mean and the other 50% is below it.

5) **Asymtotic to the X-axis** – The normal curve does not meet the base line and there are infinite distances to the right and left of the mean. As the curve is extended further and it comes nearer to X-axis, but it never touches the same.

6) **Bilateral-** The normal curve is bilateral because 50% of the curve lies to the left and 50% of the curve lies to the right sides of the curve. Hence, the curve is bilateral.

7) **Standard deviation**- The standard deviation is a measure of the spread or dispersion of the data. About 68% of the data points fall within 1 standard deviation of the mean. About 95% of the data points fall within 2 standard deviation of the mean. About 99.7% of the data points fall within 3 standard deviation of the mean.

8) **Height of the curve declines symmetrically**- The normal probability curve is symmetric about the mean (μ), meaning

that the curve is mirror-image on both sides of the mean. This symmetry results in height of the curve declining symmetrically on both sides of the mean.

9) **As a measurement scale** – The normal probability curve is used as a measurement scale. The measurement unit of this scale is ± 1 (the unit standard deviation).

10) **Free from skewness**- The normal probability curve is free from skewness, i.e, it's coefficient of skewness amounts to zero.

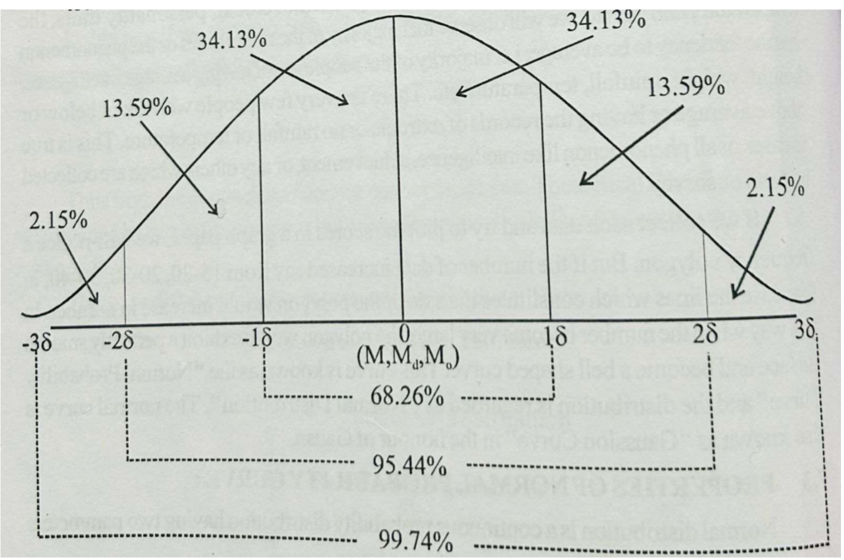


Fig.1.1

Normal Probability Curve

Look into Fig1.1, It shows that the Mean, Median and Mode, they fall in the centre point of the curve. The curve extends on both sides -3σ distance on the left to + 3σ distance on the right. If we measure from mean to 1, the score is 34.13%. Similarly, from 1 to 2, the score is 13.59%. Again, from 2 to 3, the score is 2.15%. The normal curve doesnot meet the baseline and there are infinite distances to the right and left of the mean.

Check Your Progress

1. Write any two characteristics of Normal Probability Curve?
2. NPC touches the baseline (True/False)
3. NPC is free from skewness (True/false)

1.7 Importance of Normal Probability Curve

The Normal Probability Curve (NPC), which is also known as the Gaussian Distribution or Bell Curve, is a fundamental concept used in statistics and data analysis and it has numerous applications in various fields including social sciences, medicine, psychology etc. Its importance are discussed below-

1) The normal distribution plays a crucial role in educational assessment and research, particularly when measuring mental abilities. While it serves as a valuable mathematical model in behavioural sciences, its essential to note that it's an idealized representation rather than an exact reflection of real-world test scores. In reality, test score distributions often approximate the theoritical normal distribution, but rarely achieve a perfect fit.

2) Normal Probability Curve is a versatile statistical tool that helps describe the frequencies of occurrence of various types of data including 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, adjustment, anxiety etc. Overall, the Normal Probability Curve is a

powerful tool for understanding and analysing continous data in a wide range of fields.

3) The normal probability curve helps in computing percentiles and percentile ranks in a given normal distribution.

4) The normal distribution is the most widely used probability distribution. This is applicable for continous random variable. If a random variable assumes any value within range is known as continuous random variable.

5) The normal curve, also known as the curve of error, helps to understand standard errors of measurement. For instance, when calculating means from multiple samples of the same population, these means will be normally distributed around the mean of the entire population. The distance of a particular sample mean from the average can help determine its standard error of measurement.

6) Assuming normal distribution, qualitative data like ratings, grades, and categorical ranks can be transformed and combined to calculate an average rating for each individual.

7) Normal curve represents a model distribution that can be used to assess if a distribution is normal or not and if not, in what way it diverges from the normal.

8) The normal distribution plays a crucial role in various inferential statistical tests, including z-tests, t-tests, and F-tests.

9) The normal distribution is a continuous distribution and plays significant role in statistical theory and inference.

Stop to Consider

- In statistics, the term 'Probability' refers to the expected frequency (chance) of occurrence of an event among all possible events.
- The literal meaning of the term normal is average.

The Normal Probability Curve is considered as an ideal degree of peak or kurtosis.

- A continous probability distribution for a variable is called as normal probability distribution or simply normal distribution.
- Z-score is used for standardising the raw data.
- The 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.
- A random variable means a real valued functions defined over a sample space. For every value of random variable, there is associated probability. If a random variable takes only integer values, it is known as discrete random variable. If a random variable assumes any value within range, it is known as continuous random variable.
- Normal distribution is for continuous random variable and for discrete random variables, there are binomial distribution, Poisson Distribution, Multinomial Distribution.
- Poisson distribution is again classified as- Normal distribution (Z), Students distribution (t), Chi-square distribution (X), Fishers distribution (F).

1.8 Summing Up:

Coming to the last part of this unit, it can be said that this unit has tried to the familiarize you with the meaning, characteristics/nature, and importance of normal probability curve. Thus, we can summarize the unit as-

- A Normal Probability Curve, also known as a Bell Curve, is a graphical representation of a normal distribution.
- The curve extends on both sides -3 (σ) distance on the left to +3(σ)distance on the right.
- A Normal Probability Curve is based upon the Law of Probability, discovered by French mathematician Abraham Demoivre. He developed its mathematical equation and graphical representation in the 18^{th} century. This law is based on chance and probability and when data follows this mathematical equation, it forms a distinctive bell-shaped curve.
- A Normal Probability curve is a symmetric, bell-shaped curve that shows how data points are distributed around the mean (average) value. In other words, it is not skewed. The value skewness of this curve is zero.
- Since the curve never touches the baseline, the mean serves as the starting point for working with the normal curve.
- The data distribution doesn't always form a normal curve. In cases, where individual scores differ significantly from the average, the resulting curve will also deviate from the normal curve shape. These deviations from normality can occur in two main ways- Skewness and Kurtosis.
- Regarding the importance of normal probability curve, the curve has wide applications in the field of measurement concerning education, psychology and sociology. The curve is used as a model, also used for computing

percentiles and percentile ranks in a given normal conditions, also used for transforming and combining qualitative data and also used as ability grouping and so on.

1.9 Questions and Exercises:

1. What do you understand by the term normal distribution and normal curve.
2. Write the characteristics/nature of Normal probability curve.
3. What do you understand by the term divergence from normality?
4. Why normal curve doesnot meet the baseline?

1.10 Answer to Check your Progress- 1

Two characteristics of NPC are-

a) It is symmetrical and bell-shaped.
b) It is unimodal i.e. it has only one mode.

Answer to CYP 2- False

Answer to CYP3- True

Answer to SQ1:

A normal distribution is a way that data is spread out. It means that most data points are close to the average and fewer data points are far away from the average.

A normal curve, also called a bell-curve, is a graph that shows a normal distribution. It shaped like a bell, with the majority of data points clustered around the average, and tapering off gradually towards the extremes.

Answer to SQ2:

The characteristics of normal probability curve are:

1) It is bell shaped.
2) It is symmetrical.
3) Mean, median and mode are same.
4) It is unimodal.
5) It is asymptotic.

Answer to SQ3:

Divergence from normality occurs when the scores of individuals in the group seriously deviate from the average, the curve representing these distributions also deviate from the shape of a normal curve.

Answer to SQ4:

Normal curve doesn’t touch the baseline because the normal curve has infinite tails, meaning it extends indefinitely in both directions. This makes it impossible for the curve to touch the baseline.

1.11 References and Suggested Readings

- Garrett, Henry E. (1984), Statistics in Psychology and Education. Vakils, Feffer and Simons Pvt. Ltd., Bombay,
- Bhandarkar, K.M. (2007), Statistics in Education. Neelkamal Publications PVT. LTD., Hyderabad.
- Saha, Kaberi. (2023), Statistical Analysis in Social Sciences, Mani Manik Prakash, Guwahati.
- Mangal, S.k. (2021), Statistics in Psychology and Education. PHI Learning Private limited, Delhi.
- Goswami, Sadhana. (2013), Statistics in Education. Shanti Prakashan, Guwahati.
- Koul, Lokesh, (2021,), Methodology of Educational Research, Vikash Publishing House Private Limited

*****

UNIT-2
PROPERTIES OF NORMAL PROBABILITY CURVE

Unit Structure:

2.1 Introduction

2.2 Objectives

2.3 Concept of Probability

2.4 Probability Distribution (Theoretical Distribution)

2.5 Normal Distribution

2.6 Properties of Normal Probability Curve

2.7 Uses of Normal Probability Curve

2.8 Summing up

2.9 Questions and Exercises

2.10 References and Suggested Readings

2.1 Introduction

Dear Learners, in the previous unit, you have understood the meaning, characteristic/nature and importance of normal probability curve. As you came to know that the normal distribution has great significance in statistical work and also serves as a good approximation of many discrete distribution when dealing with large samples. Normal probability curve is very helpful to find out the individual position of a particular students in a group or the percentage of students who scored above say 80% in a particular group etc and it is also of great help when we were required to put grades like A, B,C,D,E and so on instead of marks. In this unit, we will be focussing on meaning, properties and uses of normal probability curve.

2.2 Objectives

After going through this unit, you will be able to-

- Understand the concept of probability
- explain about Probability Distribution (Theoretical Distribution)
- Know about the Normal Distribution.
- discuss the properties of normal probability curve
- Gain insight about the uses of normal probability curve

2.3 Concept of Probability

Probability or Chance is a very familiar concept we use daily, but in statistics, it has a specific meaning. The theory of probability emerged in the 17^{th} century. Later in 1954, Antoine Gornband had taken interest in Probability. Garrett's definition of probability is as follows:

Probability of a given event is defined as the expected frequency of occurance of the event among events of a like sort". This means that the expected frequency of occurance may be based upon acknowledge of the conditions determining the occurance of the phenomenon as in determining or coin-tossing or upon empirical data as in mental and social measurements.

For example – When tossing a one-rupee coin 10 times, probability suggests that heads and tails will appear an equal number of times, approximately 5 times each. Since the coin has only two sides i.e head and tail, the probability of appearance of each sides appearing is 50%

2.4 Probability Distribution (Theoretical Distribution)

Frequency distribution can be determined using two approaches: actual frequency counts or mathematical estimation. Sometimes, we mathematically deduce the expected frequency distribution of certain population on the basis of some theoretical considerations. This is called a Theoretical Distribution or Probability Distribution. For instance, we expect heads and tails to happen 50% of the time. Again, if we flip the coin 100 times, we still expect heads and tails to happen around 50 times each. But in actual practice, we might get heads 42 times (this is the actual or observed frequency). So there's a difference between what we expect to happen (theoretical) and what actually happens (observed).

So, from the above discussions we can conclude that, Theoretical Distribution are not obtained by actual observations but are mathematically deduced on certain theoretical assumptions. There are different types of theoretical assumptions. These are discrete and continuous.

STOP TO CONSIDER

1. If a random variable takes only integer values, it is known as discrete random variable. Binomial, Poisson distribution and multinomial distribution are for discrete random variables.

2. Poisson distribution is again classified as- Normal distribution(Z) , students distribution (t) , chi-square distribution (X), fishers distributions(F)

3. If a random variable assumes any value within range, it is known as continous random variable. Normal distribution is for continous random variable.

2.5 Normal Distribution:

The Normal Distribution, initially introduced by Abraham De Moivre in 1733, which was later developed by Laplace and Gauss in 1809, and its now considered the most significant theoretical distributions. The term ‘normal’ refers to the average. In nature, most phenomenon, such as intelligence, height, weight, personality traits etc tend to cluster around the average. Majority of the people are having average intelligence, height, weight etc but there are very few people who are of below or above average or having the records of extreme or no rainfall or temperature. Collecting data on these phenomena and plotting it on a graph reveals a frequency polygon. As more data is added, the polygon’s lines increases, say from 15-20, 20-30,30-40, in that case the lines which constitutes the side of the polygon would increase in number. As the number become very large, the polygon’s sides become smoother, eventually forming a bell-shaped curve known as the Normal Probability Curve or Gaussian Curve, named after Gauss and the distribution is regarded as “Normal Distribution”.

The Normal Probability Curve, also known as a Gaussian curve or bell curve, is a probability distribution that describes how data points are spread out around a central value, called the mean. It is a continous, smooth, and symmetric curve that illustrates the probability density of the data points. The normal probability curve is of great value in educational research when we make use of mental measurement. The normal distribution is not an actual distribution of scores on any test of ability or academic achievement, but instead, a mathematical model. The distribution of test scores tends to approach a normal distribution, but the fit is rarely perfect.

2.6 Properties of Normal Probability Curve

The Normal Probability Curve (NPC) also known as the Gaussian distribution or bell curve, is a fundamental concept in statistics. The NPC is characterized by its symmetrical, bell-shaped curve which is defined by its mean and standard deviation. Understanding the properties of NPC is crucial in statistics, as it forms the basis for many statistical tests, confidence intervals, and hypothesis testing.

The following are the properties of Normal Probability Curve (NPC):

1) The normal distribution has key characteristics that are easy to spot in graphs. The mean, median and mode are exactly the same. The distribution is symmetric about the mean-half the values fall below the mean and half above the mean.
2) The total area of the normal curve is within z ± 3 standard deviation below and above the mean.
3) At the points, where the curve changes from curving upward to curving downward are called inflection points.
4) The z-scores or the standard scores in NPC towards the right from the mean are positive and towards the left from the mean are negative.
5) The normal distribution can be described by two values- the mean and the standard deviation.
6) The mean is the location parameter. The mean determines where the peak of the curve is centred. Increasing the mean moves the curve right, while decreasing it moves the curve left.
7) The standard deviation is the scale parameter. The standard deviation stretches or squeezes the curve. A small standard

deviation results in narrow curve. A large standard deviation leads to a wider curve.

8) The curve is perfectly symmetrical around its central axis, meaning both sides of the curve are mirror images of each other in terms of size, shape and slope.

9) The Mean, Median and Mode of the normal distribution are same and fall at the same place.

10) The Normal Curve is Unimodal. It has only one peak (mode) making it a unimodal distribution.

11) The Normal Probability Curve approaches the X-axis asymptotically, meaning it gradually decreases in height towards the extremes, but never actually intersects or touches the horizontal axis in the baseline.

12) About 68% of the curve falls within the limit of plus or minus 1σunit from the mean; about 95% of the curve area falls within the limit of plus or minus 2σunit from the mean and about 99.7% of the curve area falls within the limit of plus or minus 3σunit from the mean.

13) When a normal probability curve is distorted, it becomes asymmetrical, causing the mean, median and mode to no longer align. This distortion can occur in two main ways- skewness and kurtosis.

14) The fractional areas in between any two given z-scores are identical in both halves of the normal curve, for example, the fractional area between the z-scores of +1 σ identical to the z-scores of -1 σ. Further, the height of the ordinates at a particular z-score in both the halves of the normal curve is same, for example, the height of an ordinate at +1 σ is equal to the height of an ordinate at -1 σ.

15) Any normal variate with mean μ and standard deviation(σ) can be converted into corresponding SNV as

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

Check Your Progress

1. What is the shape of normal probability curve?
   a) Symmetric and bell-shaped
   b) Asymmetric and skewed
   c) Rectangular and uniform
   d) Triangular and unimodal
2. In normal distribution :
   a) Mean=median=mode
   b) Mean≠median≠mode
   c) Mean=median≠mode
   d) Mean≠median=mode
3. Which of the following is an importance of normal distribution ?
   a) It Simplifies complex data
   b) It is essential for hypothesis testing
   c) It enables statistical inference
   d) All of the above

Stop to Consider

- The normal probability curve is symmetric and bell-shaped.
- 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.
- Numerous statistical problems requires the assumptions of a normal population to be solvable.

- Normal distribution also helps to identify outliers and anomalies in data.

- The Normal Probability Curve (NPC) is used to convert raw scores into z-scores, also known as standard scores. A z-score indicates how many standard deviations an observations is away from the mean. If z-score is equal to 0, it is on the mean.

- A positive z-score indicates that the raw score is higher than the mean average. For example- If a student scored 80 on a test with a mean of 70 and a standard deviation of 5, their z-score would be:

$$z = \frac{80 - 70}{5} = 2$$

This positive z-score (z = 2) indicates that the student scored 2 standard deviation above the mean which is a very good score.

- A negative z-score reveals the raw score is below the mean average. For example- If a student scored 60 on a test with a mean of 70 and a standard deviation of 5, their z-score would be:

$$z = \frac{60 - 70}{5} = -2$$

This negative z-score (z = -2) indicates that the student scored 2 standard deviation below the mean, which is a below-average score.

2.7 Uses of Normal Probability Curve

The Normal Probability Curve has many uses. They are described below:

1) Converting Raw scores into Comparable Standard Normalized scores-

To compare scores from different tests, we need to standardize them with the help of normal curve that enables meaningful comparisons across tests with varying scales providing a fair and accurate assessment of an individual's performance. For converting a given raw score into a z score, we subtract the mean of the scores of distribution from the respective raw scores and divide it by the standard deviation of the distribution.

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

2) **Comparing scores**-The NPC helps compare scores from different groups or tests. For example- The NPC converts test scores into percentiles and percentile ranks to see how a student ranks compared to others, making it an essential tool in statistical analysis and data interpretation.

3) **Hypothesis Testing**- The normal curve provides researchers a theoretical foundation for hypothesis testing, enables researchers to make informed decisions about their data and statistical analysis. NPC facilitates the decision to reject or fail to reject the null hypothesis based on the p-value or critical region.

4) **Used for transforming and combining qualitative data**- The NPC helps to convert ratings (e.g A,B,C) and grades (e.g A,B,C) into numbers (e.g. 1-5 or 1-10) and also helps to assign numbers to categorical ranks (e.g. High, Medium, Low). Therefore, by using NPC we can transform and combine qualitative data in a meaningful way, making it easier to analyse and compare.

5) **Also used for ability grouping –** With the help of normal curve, a group of individuals may be grouped into certain categories such as A,B,C,D,E (Very good, good, average, poor, very poor) in terms of some trait (assumed to be normally distributed).

6) **Helps to study anthropometrical data-** NPC is also used to study anthropometrical data, that involves measuring human body dimensions such as height, weight, and body proportions. By applying NPC to anthropometrical data, researchers and

practitioners can gain valuable insights into human body dimensions, growth patterns and health outcomes, ultimately informing policies, designs, and interventions that promote human health and well-being.

7) **Understanding and Applying the concept of Standard Errors of Measurement-**The normal curve which is also known as curve of error helps in understanding the concept of standard errors of measurement. For example- If we compute mean for the distributions of various samples taken from a single population, then these means will be found to be distributed normally around the mean of the entire population. The sigma distance of a particular sample mean may help us determine the standard error of measurement for the mean of that sample.

8) **Understanding data-** The normal curve is also used in showing how most values are close to the average, with fewer very high or very low values. Example- In a class, most students score around 70-80 on a test, with fewer scoring very low or very high.

9) **Used in determining the relative difficulty of test items-**The normal curve is a simple yet effective tool of scaling test items for difficulty, enabling educators to determine the relative difficulty of test questions, problems and other test items.

10) **Generalize about population**- The NPC is used to generalize about population from which the samples are drawn by calculating the standard error of mean and other statistics.

Check Your Progress

1. Which of the following is a use of the Normal Probability Curve?
   a) To analyse categorical data
   b) To analyse numerical data

c) To determine the correlation between two variables.

d) To determine the probability of an event.

2. Which of the statements is false about the Normal Probability Curve?

a) It is symmetrical about the mean.

b) It is asymmetrical

c) The mean, median and mode are equal.

d) The area under the curve is 1

3. What is the purpose of standardizing a normal distribution?

a) To change the mean.

b) To change the standard deviation

c) To compare the different datasets

d) To make the distribution skewed

Stop to Consider

- The NPC has numerous real-world applications in fields such as medicines, finance and social sciences.
- Through NPC, it becomes easier to study certain socio-economic data like rates of birth, marriage etc under the prevailing circumstances.
- Through NPC, it is also easier to present intelligence as measured by standard tests, educational test scores, speed of association etc.

2.8 Summing Up

Coming to the last part of this unit, it can be said that this unit has tried to familiarize you with the concept of probability, normal distribution, meaning of normal probability curve, properties and uses of normal probability curve.

1) The Normal Probability Curve (NPC) is a graphical representation of a normal distribution, also known as a Gaussian distribution or bell curve. It is a continuous probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean.

2) The Normal Probability Curve (NPC) is widely used in statistics, data analysis and many fields such as medicines, social sciences, and finances to model and analyse continuous data.

3) The Normal Probability Curve (NPC) is standardized by converting raw scores to z-scores, which have a mean of 0 and a standard deviation of 1.

4) The formula to convert a raw score to a z-score is:

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

   Where:

   z = z-score

   X = raw score

   M = mean

   σ = standard deviation

5) Given Mean = 49.5 and SD = 14.3 for a distribution, change the score of 80 into z or sigma score

Solution

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

Where,

X = 80, M = 49.5, SD = 14.3

z = 80-49.5 /14.3

= 30.5 /14.3

= 2.13

2.9 Questions and Exercises

1) Given Mean = 24, and SD = 8, change the raw score of 16 into z or sigma score.
2) Given Mean = 40 and SD = 8, change the raw score of 36 into z or sigma score.
3) If Z = 2.5, then T? Convert Z to T.
4) The NPC has an infinite range (True/False)
5) Z-score is the most popular standard score (True/False)

Answer to Check Your Progress 1-

Symmetric and Bell-shaped.

Answer to CYP 2- Mean = Median = Mode

Answer to CYP 3-All of the above

Answer to CYP 4- To analyse numerical data.

Answer to CYP 5 - It is asymmetrical.

Answer to CYP 6 – To compare the different datasets.

Answer to SQ1-

Solution,

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

Where,

X= 16, Mean = 24, SD = 8

z = 16 – 24/ 8

= - 8/8

= -1

Answer to SQ2-

Solution,

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

Where,

X= 36, Mean = 40, SD = 8

z =36 – 40/ 8

= - 4/8

= -0.5

Answer to SQ3 –

Solution,

T = 10z + 50

z = 2.5

T = ?

T = 10 x 2.5 + 50

= 25 + 50

= 75

Answer to SQ4 – True

Answer to SQ5 - True

2.10 References and Suggested Readings

- Garrett, Henry E. (1984), Statistics in Psychology and Education. Vakils, Feffer and Simons Pvt. Ltd., Bombay,
- Bhandarkar, K.M. (2007), Statistics in Education. Neelkamal Publications PVT. LTD., Hyderabad.
- Saha, Kaberi. (2023), Statistical Analysis in Social Sciences, Mani Manik Prakash, Guwahati.
- Mangal, S.k. (2021), Statistics in Psychology and Education. PHI Learning Private limited, Delhi.
- Goswami, Sadhana. (2013), Statistics in Education. Shanti Prakashan, Guwahati.
- Koul, Lokesh, (2021,), Methodology of Educational Research, Vikash Publishing House Private Limited

*****

UNIT-3

MEASURES OF ASYMMETRY

(Skewness and Kurtosis)

Unit Structure:

3.1 Introduction

3.2 Objectives

3.3 Meaning and Definitions of skewness

3.4 Positively Skewed Curve

3.5 Characteristics / Properties of Positively Skewed Curve (PSC)

3.6 Negatively Skewed Curve

3.7 Characteristics of Negatively Skewed Curve

3.8 Index of Skewness

3.9 Kurtosis

3.10 Types of kurtosis

3.11 Summing Up

3.12 Questions and Exercises

3.13 References and Suggested Readings

3.1 Introduction:

As you know that normal probability curve is a bell shaped curve. In the previous unit you have learnt about the concept and characteristics of normal probability curve. Also you know the importance and uses of the curve. In this unit you will be able to know that, if a curve is not normal then what dimension it will appear. Therefore, in this unit you will learn the measures of asymmetry or you can say divergence of normality. The most important of all the frequency curves is the normal curve which is a symmetrical curve. It can be regarded as the limit of a bionomial curve when the number of trials becomes very large. Demoivre discovered this curve while working on problems in games of chance. Laplace and Gauss discovered independently that errors of observation in the physical sciences followed the normal law. Distributions that arise from data taken from the fields of biology, sociology etc, are found to be approximately normal.

Under normal conditions if a graph is drawn from the scores of a group of students then its shape appears like a Bell and this bell-shaped graph is known as the Normal Probability Curve. It is also known as the equivalent curve or balanced curve or Gaussian curve, named after Gauss, whose contribution is unparalleled in its development. The credit for development of theory and laws of probability goes to the gamblers in the seventeenth century. The gamblers often consulted the great scientists Viz Galileo Galile and Pascal and they used to ask these great scientists regarding numbers for gambling. In turn, the scientists used to give numbers on the basis of the laws of probability, which proved to be fruitful for the gamblers.

The credit for developing the theory of probability in complete form goes to Gauss, a great mathematician of the 19th century. The law/theory of probability was thereafter used in several important endeavors of human life.

3.2 Objectives:

After going through this unit you will be able to:

- Understand the Meaning and Definitions of skewness.
- Know the Characteristics / Properties of Positively Skewed Curve (PSC)
- Analyze the index of Skewness
- Understand the meaning of kurtosis and explain various types of kurtosis
- Discuss about Kurtosis

3.3 Meaning and Definition of Skewness:

The distribution in which Mean and Median fall on different points is known as Skewed Distribution and this tendency of distribution is known as Skewness.

In a normal distribution Mean and Median fall on the same point; hence in this case skewness of distribution is zero. As the difference between Mean and Median increases the Normal Probability decreases in the same proportion and skewness increases (of course, in the same proportion)

Skewness of distribution can be of two types, which gives birth to two types of curves—

- Positively skewed curve
- Negatively skewed curve

3.4 Positively Skewed Curve:

Definition

The distribution in which most of the frequencies are concentrated in the class intervals bearing lower values is known as positively skewed distribution and the curve obtained from this distribution is called the Positively Skewed Curve.

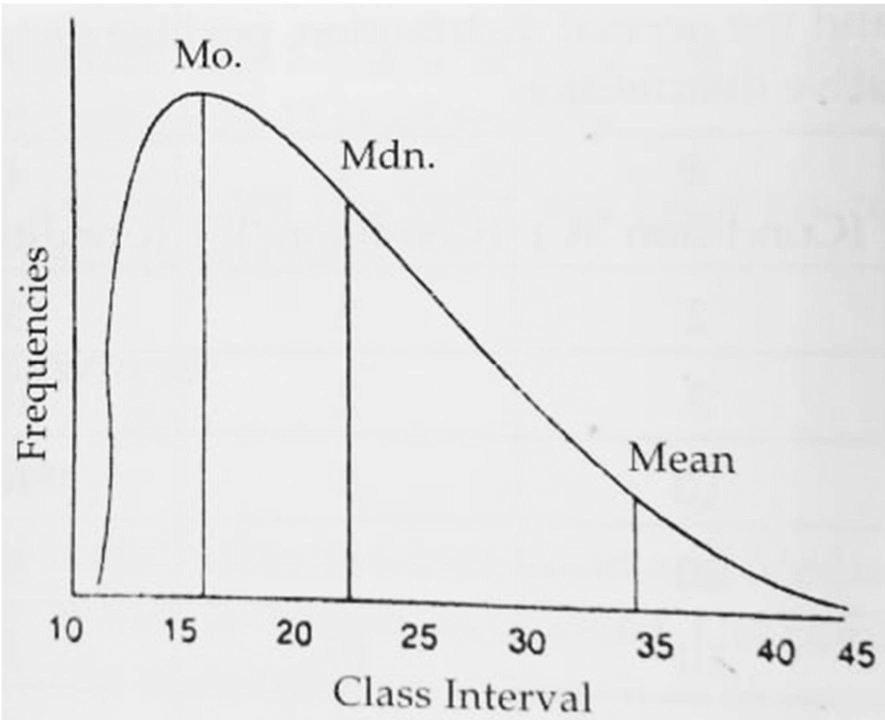
Let us study the following example in order to understand the normal distribution, positive distribution and negative distribution.

C.I	**f (condition 'A')**	**f (condition 'B')**	**f (condition 'C')**
40-44	2	2	20
35-39	3	2	10
30-34	10	3	10
25-29	20	3	3
20-24	10	10	3
15-19	3	10	2
10-14	2	20	2

………………………… Example (44)

Condition 'A' is an ideal state of normal distribution, because here the values of Mean, Median and Mode are exactly identical (M = Mdn = Mo = 27). If weights of 500 grams each are placed in two pans of a balance (which is technically accurate) then both the pans will be in a balanced state. Same is the case with this distribution. Zero point of this distribution is 20 and there are 15 frequencies on each side of the zero point.

If the weight of 900 grams is placed in the right pan and that of 100 gms is placed in the left pan of a balance, then the right pan shall go down and the left pan will rise upwards. Similar is the case with condition 'B'. In this distribution most of the frequencies are concentrated in the class intervals of lower values and therefore skewness has been created in the distribution. Graph drawn for this distribution will be a positively skewed curve, which is shown below.



Positively Skewed Curve

Stop to Consider

The most important of all the frequency curves is the normal curve which is a symmetrical curve. It can be regarded as the limit of a bionomial curve when the number of trials becomes very large. Demoivre discovered this curve while working on problems in games of chance. Laplace and Gauss discovered independently that errors of observation in the physical sciences followed the normal law.

The credit for developing the theory of probability in complete form goes to Gauss, a great mathematician of the 19th century. The law/theory of probability was thereafter used in several important endeavors of human life.

Check Your Progress

Q. What do you mean by skewed distribution?

Q. what are the types of skewness distribution.

Q. Who developed the theory of probability ?

3.5 Characteristics / Properties of Positively Skewed Curve (PSC)

1. The distribution in which most of the frequencies are concentrated in the CI of lower values is known as positive distribution and graph drawn on the basis of this distribution is called positively skewed curve.
2. The graph drawn for the scores of classroom, in which there are low achievers or failures, shows positive skewness.

3. If the subject is difficult and most of the students could not understand the same then their scores shall be low and the graph drawn for the scores shall be positively skewed.
4. If the teaching is ineffective or the students did not prepare well for the examination, then the graph drawn from the obtained marks shall be positively skewed.
5. If the valuer did assign fewer marks to the examinees, then the graph drawn from such scores shall be positively skewed.
6. If the question paper consists of questions having higher difficultly level, then the graph drawn from the scores obtained, shall be positively skewed.
7. In a positively skewed curve the value of Mean is highest, the value of Median is lower than the Mean, and the value of Mode is lowest, which can be denoted symbolically as

M > Mdn > Mo

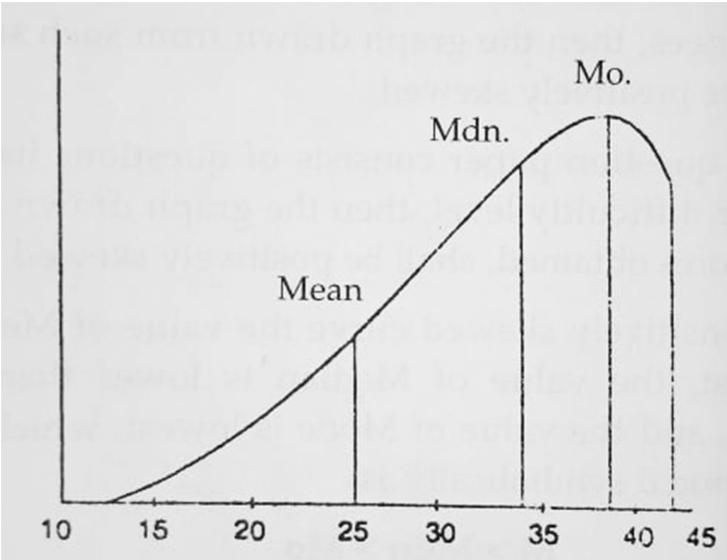
8. In a positively skewed curve, if we start from the zero point then the measures of central tendencies fall in the following order-Mode, Median and Mean.

3.6 Negatively Skewed Curve:

1. The distribution in which most of the frequencies or scores are concentrated in the CI of higher values is known as negative distribution and the graph drawn on the basis of this distribution is called negatively skewed curve.

If a weight of 900 gms. is placed in the left pan and that of 100 gms. is placed in the right pan, then the left pan will go down and right pan will go upwards, creating imbalance. Same is the state with condition 'C' (Pl. refer example 44)

In this distribution most of the frequencies are concentrated in the Cl of higher values and has given birth to skewness. The graph drawn from such distribution shall be a negatively skewed curve shown as follows –



Negatively Skewed Curve

3.7 Characteristics of Negatively Skewed Curve:

Characteristics of Negatively Skewed Curve (inferences that can be drawn)

1. Graph obtained for the distribution in which most of the frequencies are concentrated in the CI of higher values is called negatively skewed curve.
2. If there are intelligent and studious students in a class then the graph drawn for their scores is a negatively skewed curve.
3. If a certain subject is easy and understanding of students is better than the students shall score higher marks in that particular subject and the graph drawn for such scores shall be a negatively skewed curve.

4. If teaching is very effective then the graph drawn for the scores of students shall be negatively skewed.
5. If the examiner is very liberal in giving marks, then the graph drawn from the scores of students shall be a negatively skewed curve.
6. In a negatively skewed curve the value of Mean is less than Median and the value of Median is less than Mode.

This condition is exactly opposite to that of positively skewed curve. In negatively skewed curve the position and values of measures of Central Tendency can be shown as under.

$$\mathbf{M < Mdn < Mo}$$

7. In a negatively skewed curve if we start from the zero point then the measures of Central Tendency will fall in the following order Mean, Median, Mode.

Stop to Consider

In a positively skewed curve the value of Mean is highest, the value of Median is lower than the Mean, and the value of Mode is lowest, which can be denoted symbolically as

$$\mathbf{M > Mdn > Mo}$$

In a positively skewed curve, if we start from the zero point then the measures of central tendencies fall in the following order-Mode, Median and Mean.

Check Your Progress

Q. Write any two Characteristics of Positively Skewed Curve?

Q. Draw a negatively skewned curve.

Q. Write any two Characteristics of negatively Skewed Curve.

3.8 Index of Skewness:

We have studied the definitions of skewness and the properties of positively and negatively skewed curves. Actually, skewness of a distribution can be denoted by skewness index.

The formula for computation of skewness index is as under –

$$SK_1 = \frac{3\ (Mean - Median)}{\sigma}$$

or

$$SK_1 = \frac{3\ (M - Mdn)}{\sigma}$$

In the above formula -
SK_1 = Skewness Index
M = Mean
Mdn = Median
σ = Standard Deviation
3 is a constant in the formula.

Let us study the following example for understanding the application of this formula.

Mean of a distribution is 20 and its Median is 15. If the Standard Deviation is 10, what will be the skewness index of the distribution?

..Example (45)

SK_1 = ? M = 20
Mdn = 15 σ = 10
By keeping these values in formula (42) we get –

$$SK_1 = \frac{3\ (20 - 15)}{10}$$

$$= \frac{3\ (5)}{10}$$

$$= \frac{15}{10}$$

= 1.5

... SK_1 of the distribution shall be 1.5.

Stop to Consider

The formula for computation of skewness index is as under –

$$SK_1 = \frac{3\ (Mean - Median)}{\sigma}$$

Check Your Progress

Q. Write the abbreviation of the following.

SK_1=…………..

M=……………

Mdn=………

σ =……………

3.9 Kurtosis:

When there are very few individuals whose scores are near to the average score for their group, the curve representing such a distribution becomes 'flattened' in the middle. On the other hand, when there are too many cases in the central area, the distribution curve becomes too 'peaked in comparison with the normal curve.

Both these characteristics of being flat or peaked are used to describe the term kurtosis.

The curves with kurtosis are of three types.

3.10 Types of Kurtosis:

1. Leptokurtic Curve
2. Platykurtic Curve
3. Mesokurtic Curve

1. Leptokurtic Curve

If maximum frequencies in a distribution are concentrated around the Mean, then the number of frequencies falling between - 1 σ to +1 σ is more than the frequencies falling within the range in case of Normal Probability Curve and the curve obtained for this distribution is called Leptokurtic curve.

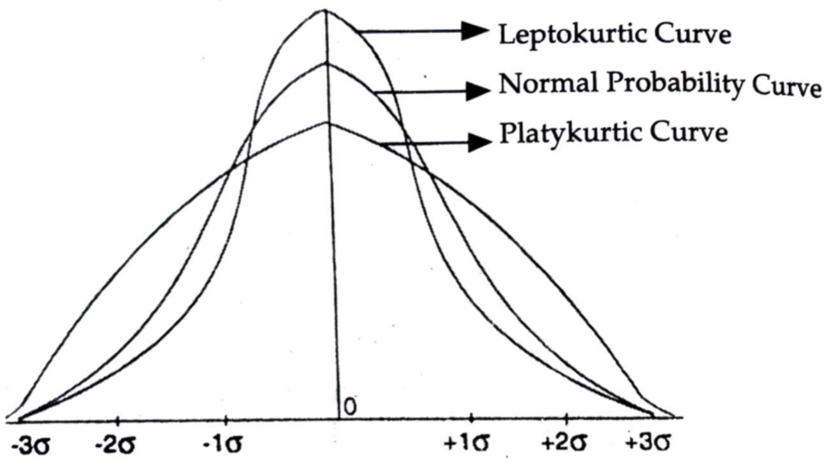
2. Platykurtic Curve

In exactly opposite condition (to that of above) the scores are not concentrated around the Mean. Therefore, the number of frequencies falling between - 1 σ to +1 σ is lower in comparison with the Normal Probability Curve and the curve drawn for this distribution is known as Platykurtic curve.

In the following figure all the three types of curves i.e. Normal Probability Curve, Leptokurtic Curve and Platykurtic Curve are shown.

3. Mesokurtic Curve

A frequency distribution is said to be mesokurtic when it almost resembles the normal (neither too flat nor too peaked).



Formula for calculation of Kurtosis:

$$Ku = \frac{Q}{P_{90} - P_{10}} \qquad \text{............... Formula (43)}$$

where,

Ku = Kurtosis

Q = Quartile Deviation

P90 = 90th percentile

P10 = 10th percentile

For calculating Ku for Normal Probability Curve by using the above formula the following factors are to be taken into account -

1. The value of Q should be taken as 0.6745 of the value of SD (i.e. 67.45% of the value of SD; the value of Q is around 2/3 of SD).

2. The sigma distance of P from the Mean should be 90 taken as +1.28 and the sigma distance of P_1, from the Mean should be taken as - 1.28.

By keeping these values in formula (43) we get-

$$Ku = \frac{0.6745}{[1.28 - (-1.28)]}$$

$$= \frac{0.6745}{2.56}$$

= 0.2634

Therefore, the value of kurtosis for the Normal Probability Curve is 0.2634. If the value of kurtosis is more than this value then the curve is leptokurtic and if the said value is less than 0.2634 then it is platykurtic (fig.9).

Self-Asking Questions

Q. Explain the meaning of kurtosis.

…………………………………………………………………………
…………………………………………………………………………
…………………………………………………………………………
…………………………………………………………………………
…………………………………………………………………………

Q. Explain the types of curves with kurtosis by the help of diagrams.

…………………………………………………………………………
…………………………………………………………………………
…………………………………………………………………………
…………………………………………………………………………

Q. Mention the formula for kurtosis.

…………………………………………………………………………
…………………………………………………………………………
…………………………………………………………………………
…………………………………………………………………………

3.11 Summing Up:

In summary, we can draw the following inferences from the leptokurtic curve.

1. Most of the students in the group are average (mediocre).
2. Intelligent and dull students are less in number.
3. The examiner has allotted average marks to most of the students.
4. The teacher may have used only one method of teaching for the whole class by neglecting individual differences.
5. Questions with more or less similar difficulty level may have been included in the question paper by the examiner.
6. Discrimination Index (DI) of the questions selected is less.
7. The examiner may have allotted more or less similar marks to the examinees.

The following inferences can be drawn on the basis of platykurtic curve.

1. Questions included in question paper may not have been framed by following general trends.
2. The teaching is ineffective.
3. Valuation of papers may have been improper.
4. Discrimination index of questions asked in the exam may be of higher order.
5. There may have been variation in the difficulty level of questions.
6. The group of students is heterogeneous.

3.12 Questions and Exercises:

1. Explain the law of probability.
2. Draw a well-labeled diagram of the Normal Probability Curve and state the percentage of scores falling in its different areas.
3. Write seven characteristics of the Normal Probability Curve.
4. Explain the meaning of skewness.
5. Draw a positively skewed curve and state its characteristics.
6. Draw a negatively skewed curve and mention its important characteristics.
7. State the formula for calculation of skewness index.
8. Explain the meaning of kurtosis. Explain the types of curves with kurtosis by the help of diagrams.
9. Mention the formula for kurtosis.

3.13 References and Suggested Readings:

- Garrett, Henry E. (1984), Statistics in Psychology and Education. Vakils, Feffer and Simons Pvt. Ltd., Bombay,
- Bhandarkar, K.M. (2007), Statistics in Education. Neelkamal Publications PVT. LTD., Hyderabad.
- Saha, Kaberi. (2023), Statistical Analysis in Social Sciences, Mani Manik Prakash, Guwahati.
- Mangal, S.k. (2021), Statistics in Psychology and Education. PHI Learning Private limited, Delhi.
- Goswami, Sadhana. (2013), Statistics in Education. Shanti Prakashan, Guwahati.
- Koul, Lokesh, (2021,), Methodology of Educational Research, Vikash Publishing House Private Limited

*****

UNIT- 4
APPLICATIONS OF NORMAL PROBABILITY CURVE

Unit structure:

4.1 Introduction
4.2 Objectives
4.3 Applications of Normal Probability Curve
4.4 Illustration of the Applications of the Normal Curve
4.5 Examples of Application of the Normal Curve
4.6 Summing Up
4.7 Questions and Exercises
4.8 Answer to Check Your Progress
4.9 References and Suggested Readings

4.1 Introduction

Dear learners, in the previous units you have understood the basic concept of normal probability curve. You have also understood the characteristics and importance of normal curve. Similarly, you have gone through the concept of skewness and kurtosis. The present unit is too interesting for you. In this unit you can see how normal probability curve can be applied in various situations. So let us start the unit with some practical work.

4.2 Objectives

After going through this unit you will be able to-

- Understand various applications of normal probability curve
- Illustrate the applications of the normal curve
- Analyse examples of application of the normal curve

4.3 Applications of the Normal Curve

Normal curve has wide significance and applications in the field of measurement concerning education, psychology and sociology. Some of its main applications are discussed in the following sections:

- **Normal curve can be use as a Model:**

Normal curve represents a model distribution. It can be used as a model to

- compare various distributions with it i.e. to say, whether the distribution is normal or not and if not, in what way it diverges from the normal;
- compare two or more distributions in terms of overlapping; and
- evaluate students' performance from their scores.

- **Computing Percentiles and Percentile Ranks:**

Normal probability curve may be conveniently used for computing percentiles and percentile ranks in a given normal distribution.

- **Understanding and Applying the Concept of Standard Errors of Measurement:**

The normal curve as we have pointed out earlier is also known as the normal curve of error or simply the curve of error on the grounds that it helps in understanding the concept of standard errors of measurement. For example, if we compute mean for the distributions of various samples taken from a single population, then, these means will, be found to be distributed normally around the mean of the entire population. The sigma distance of a particular sample mean may help us determine the standard error of measurement for the mean of that sample.

- **Ability Grouping:**

A group of individuals may be conveniently grouped into certain categories as A, B, C, D, E (Very good, good, average, poor, very poor) in terms of some trait (assumed to be normally distributed), with the help of a normal curve.

- **Transforming and Combining Qualitative Data:**

Under the assumption of normality of the distributed variable, the sets of qualitative data such as ratings, letter grades and categorical ranks on a scale may be conveniently transformed and combined to provide an average rating for each individual.

- **Converting Raw Scores into Comparable Standard Normalized Scores:**

Sometimes, we have records of an individual's performance on two or more different kinds of assessment tests and we wish to compare his score on one test with the score on the other. Unless the scales of these two tests are the same, we cannot make a direct comparison. With the help of a normal curve, we can convert the raw scores belonging to different tests into standard normalized scores like sigma (or z scores) and T-scores. For converting a given raw score into a z score, we subtract the mean of the scores of the distribution from the respective raw scores and divide it by the standard deviation of the distribution $\left(i.e, z = \frac{X-M}{\sigma}\right)$, In this way, a standard z score clearly indicates how many standard deviation units a raw score is above or below the mean and thus provides a standard scale for the purpose of valuable comparison. Since z values may carry negative signs and decimals, they are converted into T values by multiplying by some constant and added to a second constant, i.e. using the formula T score = 10z + 50.

- **Determining the Relative Difficulty of Test Items:**

Normal curve provides the simplest rational method of scaling test items for difficulty and therefore, may be conveniently employed for determining the relative difficulty of test questions, problems and other test items.

Stop to Consider

Normal curve has wide significance and applications in the field of measurement concerning education, psychology and sociology. It can be applied as:

- a model
- computing percentiles and percentile ranks
- understanding and applying the concept of standard errors of measurement
- ability of grouping
- transforming and combining qualitative data
- converting raw scores into Standard Scores
- determining the relative difficulty of test items

4.4 Illustration of the Applications of the Normal Curve

Now, let us study some of the applications of the normal curve discussed so far, with the help of a few examples. Solutions to these problems require:

- ✓ a knowledge of the conversion of raw scores into z scores and vice versa, and
- ✓ the use of the normal curve (NPC table) showing the fractional parts of the total area of the curve in relation to sigma distances.

- **Converting Raw Scores into z Scores and Vice Versa:**

The relationship between z scores and raw scores can be expressed by the formula

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

Where,

X = a given raw score

M = Mean of the distribution of X scores

σ = SD of the distribution of X scores

Example 4.1: Given Mean = 49.5 and SD = 14.3 for a distribution, change the score of 80 into-z or sigma score.

Solution:

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

where X = 80 M = 49.5 SD = 14.3

Putting these values in the formula, we get

$$z = \frac{80 - 49.5}{14.3} = \frac{30.5}{14.3} = 2.13$$

Example 4.2: Given M = 48 σ = 8 for a distribution, convert a z score (σ score) of the value 0.625 into a raw score.

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

Putting known values in the given formula, we get

$$0.625 = \frac{X - 48}{8}$$

Or

X = 0.625 x 8 + 48=5.0+48= 53

- **Making Use of the Table of Normal Curve:**

The normal probability curve (given in the Appendix) provides the fractional parts of the total area (taken as 10,000) under the curve in relation to the respective sigma distances from the mean. This table may therefore be used to find the fractional part of the total area when z scores or sigma scores are given and also to find the sigma or z scores, when the fractional parts of the total area are given. Let us illustrate it with the help of examples.

Example 4.3: From the table of the normal distribution, read the areas from mean to 2.73σ.

Solution: We have to look for the figure of the total area given in the table corresponding to $2.73\,\sigma$ score. For this we have to first locate 2.7 σ distance in the first column headed by

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

(σ scores) and then move horizontally in the row against 2.7 until we reach the place below the sigma distance .03 (lying in column 4). The figure 4968 gives the fractional parts of the total area (taken as 10000) corresponding to the $2.73\,\sigma$ distance (lying on the right side) from the mean of the curve. Consequently, 4968/10000 or 49.68 percent of the cases may be said to lie between the mean and $2.73\,\sigma$.

Example 4.4: From the table of the normal distribution read the value of sigma score from the mean for the corresponding fractional area 3729.

Solution: The figure of the area 3729 located in the table lies in front of the row of the σ distance 1.1 and below the column headed by the σ distance .04. Consequently, the corresponding sigma distance from the area 3729 may be computed as 1.14.

Self-Asking Question

Q.1. What is the formula of Z?

Q.2 Write the abbreviation of the following—

X =

M =

σ =

4.5 Examples of Application of the Normal Curve

Case I: Comparing scores on two different tests:

Example 4.5: A student obtains 80 marks in Maths and 50 in English. If the mean and SD for the scores in Maths are 70 and 20 and for the scores in English are 30 and 10 find out in which subject, Maths or English, he did better?

Solution: Here, from the given data, direct comparison of his status in Maths and in English cannot be made because the marks achieved do not belong to the same scale of measurement. For putting them into a common scale, let us convert these two raw scores into z scores. Here,

Raw scores in maths (X_1) = 80, $M_1 = 70$ and $\sigma_1 = 20$,

and

Raw score in English = (X_2) =50, $M_2 = 30$ and $\sigma_2 = 10$

Therefore,

z score in Maths = $\frac{X_1 - M_1}{\sigma_1} = \frac{80-70}{20} = 0.5$

and

z score in English = $\frac{X_2 - M_2}{\sigma_2} = \frac{50-30}{10} = 2$

We can thus conclude that he did better in English than in Maths.

Case II: To determine percentage of the individuals whose scores between two given scores:

Example 4.6: In a sample of 1000 cases, the mean of test scores is 14.5 and standard deviation is 2.5. Assuming normality of distribution many individuals scored between 12 and 16?

Solution: Both the raw scores 12 and 16 have to be converted scores (sigma scores)

Z score equivalent to raw score $12 = \frac{X-M}{\sigma}$

$$= \frac{12 \quad .5}{2.5}$$

$$= \frac{-2.5}{2.5} = -1\sigma$$

z score equivalent to raw score $16 = \frac{16-14.5}{2.5} = \frac{1.5}{2.5}$

$$= 0.6\sigma$$

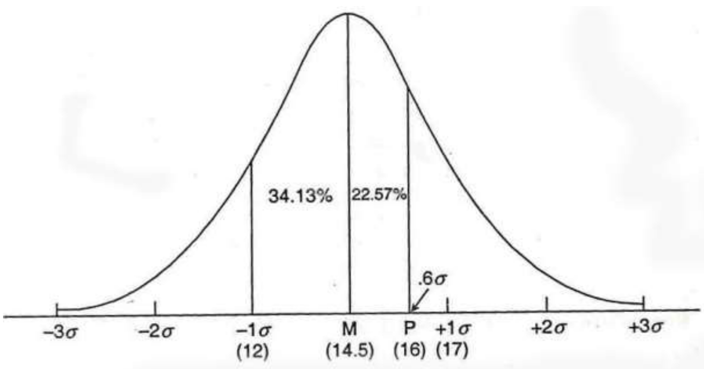


Figure Showing cases lying between scores 12 and 16.

From the normal curve table, we see that 2257 (out of 10000), i.e. 22.57% cases lie between mean and 0.6 σ. Similarly, between -1 σ and mean, 3413, i.e. 34.13% cases lie. In this way, it may be easily concluded that 22.57 + 34.13 = 56.7% or 567 individuals out of 1000 score between 12 and 16.

Case III: To determine percentage of the individuals scoring above a given score point:

Example 4.7: In a sample of 500 cases, the mean of the distribution is 40 and standard deviation 4. Assuming normality of distribution, find how many individuals score above 47 score point.

Solution: z score equivalent to the raw score 47 is given by

$\frac{X-M}{\sigma} = \frac{47-40}{4} = \frac{7}{4} = 1.75\ \sigma$

From the normal curve table, it may be found that 4599 out of 10,000 or 45.99 percent cases lie between mean and 1.756. It is also known that a total of 50 percent cases lie on both sides of the mean.

Therefore, it may be easily concluded that in all 50-45.99 = 4.01 or 4% individuals or 20 individuals out of 500 score above the score point 47.

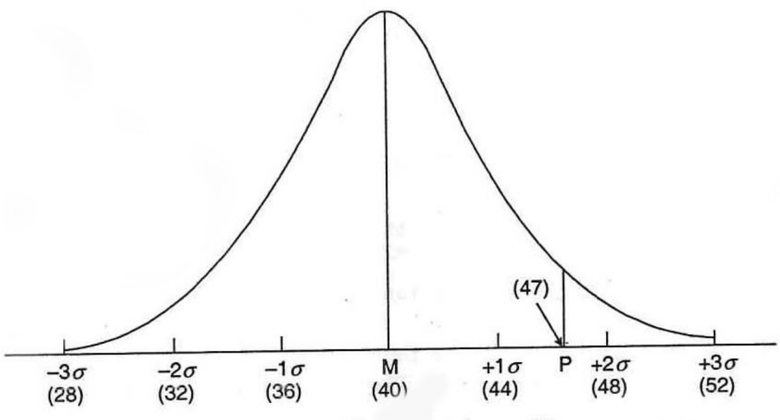


Figure showing the cases above 47 score point.

Check Your Progress

Q.1. What do you mean by the sigma distance? State the formula for its calculation.
Q.2 Mean of a normal distribution is 56 and its Standard Deviation is 6. Calculate the percentage of scores falling between 44 and 62.
Q.3 Mean of a normal distribution is 40 and its Standard Deviation is 5. Calculate the percentage of students scoring marks above the score, 30.

Case IV: To determine percentile rank, i.e. percentages of cases lying below a given score point:

Example 4.8: Given a normal distribution N = 1000, Mean = 80, SD 16; find (i) the percentile rank of the individual scoring 90 and (ii) the total number of individuals whose scores lie below the score point 40.

Solution: (i) The percentile rank is essentially a rank or the position of an individual (on a scale of 100) decided on the basis of the individual's score. In other words, here we have to determine the percentage of cases lying below the score point 90. For this, let us first transform the raw score into the standard z score with the help of the formula

$$Z = \frac{X-M}{\sigma} = \frac{90-8}{16} = \frac{10}{16}$$

$$= 0.625 \text{ sigma score} = 0.625\ \sigma$$

Now, we have to determine the total percentage of cases lying below 0.625 σ (standard scores).

The sigma score 0.625 is halfway between 0.62 and 0.63. Therefore, we have to interpolate the area lying between M and 0.625 from the given normal curve table as,

$2324 + \frac{2357-2324}{2} = 2324 + \frac{33}{2} = 2340.5$ or 2341

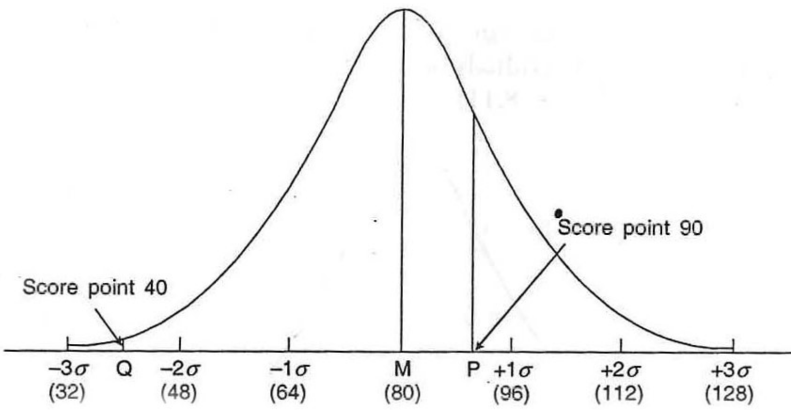


Figure shows percentile rank of 90 and cases lying below 40.

We can say that 23.41% cases lie between M and 0.625 σ distance. But 50% of cases lie up to the Mean. Therefore, it may be concluded that there are of the individuals whose scores lie. 50 + 23.41 = 73.41% below the score point 90 (shown as point P in the figure) or we may say that percentile rank of the individual scoring 90 is 73.

Solution: (ii) The z score equivalent to the raw score 40 is

$$\frac{X-M}{\sigma}=\frac{40-80}{16}=\frac{-40}{16}= 2.5\sigma$$

From the normal curve table, we may find that 4938 out of 10,000 or 49.38 percent cases lie between M and - 2.5 σ. Therefore, it may be easily concluded that, in all, 50-49.38= 0.62 percent of the cases lie below the given score point 40 (shown by Q in the figure), or

$$\frac{0.62}{100} \times 1000 = 6.2$$

i.e. 6 individuals achieve below the score point 40.

Case V: To determine the limits of the scores between which a certain percentage of the cases lie:

Example 4.9: If a distribution is normal with M = 100 and SD = 20 find out the two points between which the middle 60 percent of the cases lie.

Solution: It may be seen from the figure that the middle 60% of the cases are distributed in such a way that 30% (or 3000 out of 10,000) of the cases lie to the left and 30% to the right of the mean or 30% above the mean and 30% below the mean.

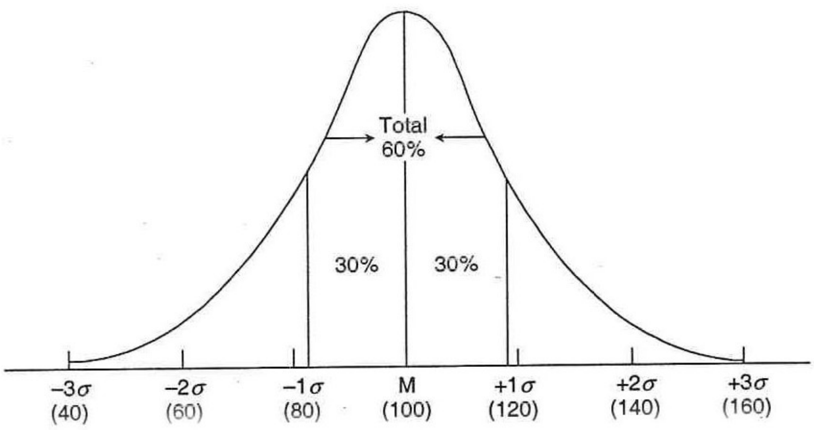


Figure shows the area covered by middle 60 percent.

From the normal curve (Table given in the Appendix), we have to find out the corresponding distance for the 3000 fractional parts of the total area under the normal curve. There is a figure of 2995 for area in the table (very close to 3000) for which we may read sigma value as 0.840. It means that 30% of the cases lie on the right side of the curve between M and 0.840 and similarly 30% of the cases lie on the left side of the figure between M. and -0.840. The middle 60% of the cases, therefore, fall between the mean and standard score +0.840. We have to convert the standard z scores to raw scores with the help of the formulae

$$Z = \frac{X_1 - M}{\sigma_1},\ Z = \frac{X_2 - M}{\sigma_2}$$

or

$$0.84 = \frac{X_1 - 100}{20}, \qquad \text{and} - 0.84 = \frac{X_2 - 100}{20}$$

$$X_1 = 16.8 + 100 \ and \ X_2 = 100 - 16.8$$

$$X_1 = 116.8 \quad and \quad X_2 = 83.2$$

After rounding the figures, we have the scores 117 and 83 that include the middle 60 percent of the cases.

Case VI: To find out the limits in terms of scores which include the highest given percentages of the cases:

Example 4.10: Given a normal distribution with a mean of 120 and SD of 25, what limits will include the highest 10% of the distribution (see the below Figure).

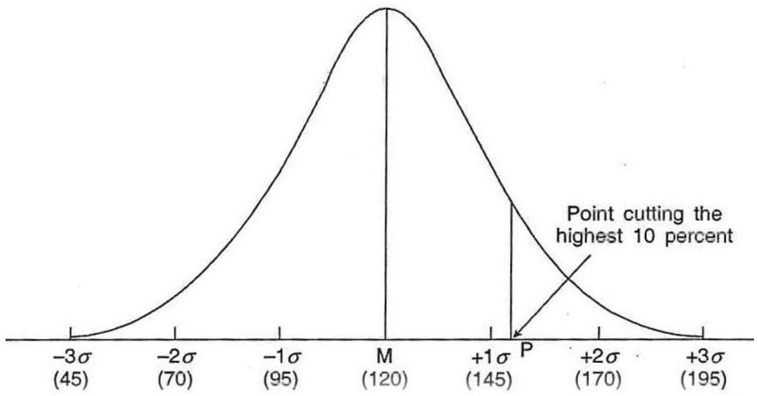


Figure shows the area covered by highest 10%.

Since 50% of the cases of a normally distributed group lie in the right half of the distribution, the highest 10% of the cases will have 40% of the cases between its lower limits and the mean of the distribution. From the table of the normal curve, we know that 3997 cases in 10,000 or 40% of the distribution are between the mean and 1.28 σ. Therefore, the lower limit of the highest 10% of the cases is M + 1.28 σ or 120 + 1.28 x 25 (as here in the present example, M = 120 and σ = 25). Thus, the lower limit of the highest 10% of the cases is 120 +32 =152 and the upper limit of the highest 10% of the cases will be the highest score in this distribution.

Case VII: To determine the percentile points or the limits in terms of scores which include the lowest given percentages of the cases:

Example 4.11: Given a normal distribution, N = 1000 M = 80 and σ = 16 determine the percentile $P_{(30)}$

Solution: In determining percentile $P_{(90)}$ we have to look for a score point on the scale of measurement below which 30 percent of the cases lie.

It may be clearly observed from the figure given below that such a score point on the scale of measurement will have 20% of the cases lying on the left side of the mean. From the table of the normal curve, let us try to find out the corresponding distances from the mean for the 20% or 2000 out of 1000 cases. By interpolation, it may be taken as 0.525 σ and its sign will be negative since it lies on the left side of the distribution. Therefore, the required score here will be:

$M - 0.525\sigma$ or $80 - 0.525 \times 16 = 80 - 8.4 = 71.6$ or 72

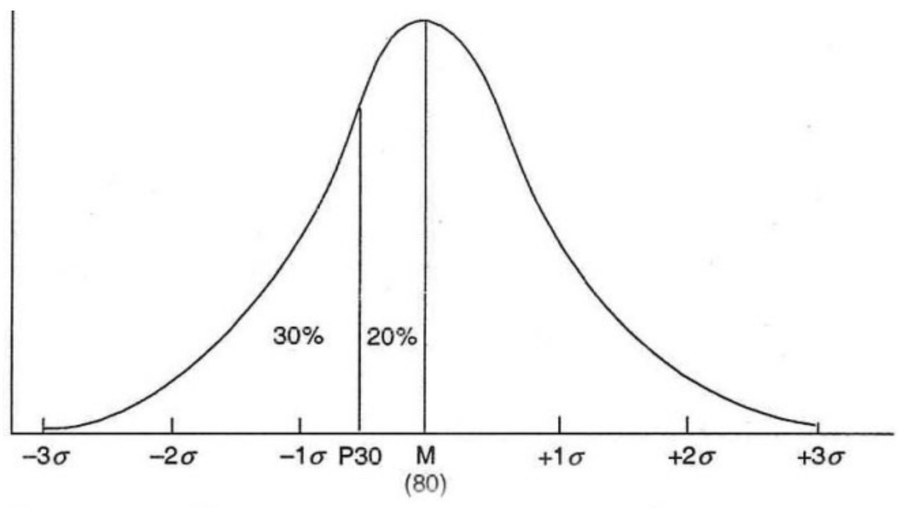


Figure shows the area covered by the lowest 30 percent.

Here, we can say about the limits in terms of scores which include the lowest 30% of the cases. The upper limit of these cases may now be given by the score point 71.6 and the lower limit will be the lowest score of the distribution.

Check Your Progress

Q.4. 6000 students appeared for B.Ed. examination conducted by Gauhati University. After analysis of results, the Mean was found to be 50 and SD was 5. Results of the examination are in accordance with Normal Probability Curve.

On the basis of the above information answer the following questions -

4.1 Calculate the number of students obtaining marks 60 and above.

4.2 How many students will fall within the score range 40 and 65?

4.3 How many students shall score the marks below 45?

4.4 What will be percentile rank of a student who scored 60 marks?

4.5 Calculate how many percent of student will fall within the score range 35 and 55 also calculate their exact number.

Case VIII: To determine the relative difficulty value of the test items:

Example 4.12: Four problems A, B, C, and D have been solved by 50%, 60%, 70%, and 80% respectively of a large group. Compare the difference in difficulty between A and B with the difference between C and D.

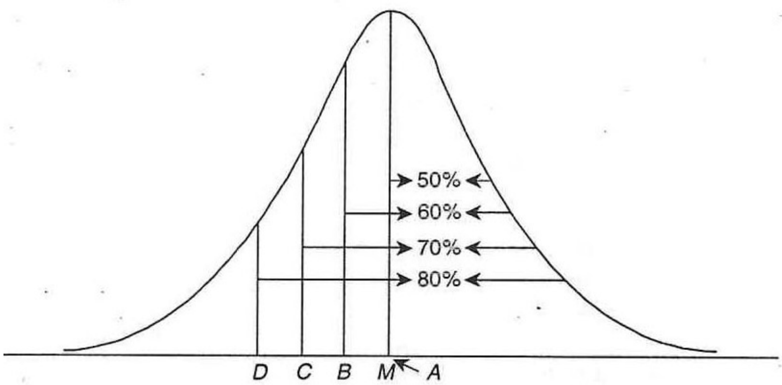


Figure shows the problems solved by different percentages of the cases.

In the case of a large group, the assumption about normal distribution of the ability of a group in terms of the achievement on a test holds good. The percentages of the students who are able to solve a particular problem are counted from the extreme right. Therefore, while starting on the base line of the curve from the extreme right, up to the point M (mean of the group), we may cover the 50% of the cases who can solve a particular problem. For the rest 60, 70 or 80 percent (more than 50%) we have to proceed on the left side of the baseline of the curve. Above figure represents well, the percentage of cases who are able to solve a particular problem.

For problem A, we see that it has been solved by 50% of the group. It is also implied that 50% of the group has not been able to solve it. Therefore, we may say that it was an average problem having zero difficulty value.

In the case of problem B, we see that it has been solved by 60% of the group. It has been a simple problem in comparison to A as 10% more individuals in the group are able to solve it. For determining the difficulty value of this problem, we have to find the σ distance from the mean of these 10% of the individuals. From the normal curve table we see that 10% (1000 out of 10,000) cases fall at a sigma distance of 0.253 from the mean. Here we have to interpolate as the given table contains 0.25 σ = 0987 and 0.26 σ = 1026; therefore, 0.253 σ = 1000 (approx.). Therefore, the difficulty value of problem B will be taken as – 0.253 σ.

Similarly, we may determine the difficulty value of problem C passed by 70% of the group (20% more individuals of the group than the average). From the table, we know that 20% (2000 out of 10,000 cases fall at a sigma distance of 0.525 σ from the mean. Therefore, the difficulty value of problem C will be taken as -0.525 σ and the difficulty value of problem D passed by 80% (30% more individuals than the average) will be -0.840σ.

Let us represent all the determined difficulty values as under:

Problem	Solved by (%)	Difficulty value	Relation difficulty value (σ differences)
A	50	- 0	
B	60	- 0.253 σ	- 0. 253 σ
C	70	-0.525 σ	
B	80	- 0.840 σ	0.315 σ

It may be seen that problem B is simpler than problem A by having 0.253 σ less difficulty value and similarly problem D is simpler than the problem C by having 0.315 σ less difficulty value.

Example 4.13: Three questions are solved by 20%, 30% and 40% respectively of a large unrelated group. If we assume the ability measured by the test questions to be distributed normally, find out the relative difficulty value of these questions.

Solution:

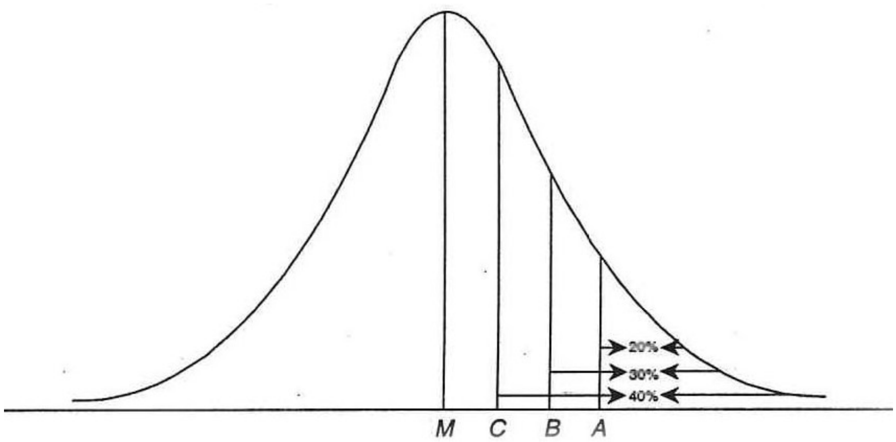


Figure shows problems passed by different percentages of the students.

Questions A, B and C are solved by 20%, 30% and 40% of the group. Therefore, from the point of view of difficulty, question A is the most difficult one in comparison to questions B and C. Question A has been solved by only 20% of the group. Our first task is to find for question A, a point on the base line such that 20% of the entire group (who has solved this question) lies below this point. The highest 20% in a normal curve as shown in the above figure must have 30% of the cases between its lower limit and the mean. From NPC table, we find that 30% or 3000 out of 10,000 cases must fall between M and 0.844 σ. For finding out the σ value for 3000, we have to interpolate between the values 2985 and 3023 respectively for 0.84 σ and 0.85 σ. In this way, the difficulty value for question A is 0.844 σ.

Similarly, the difficulty value for question B, solved by 30% of the group may be determined by locating a point on the baseline above and below which 30% and 70% cases lie. The highest 30% as shown in the figure must have 20% of the cases between its lower limit and mean. From NPC table, we find that 20% (2000 out of 10,000) cases must fall between M and 0.525 σ. Therefore, the difficulty value of question will be 0.525 σ.

Question C has been solved by 40% and failed by 60% of the e group. The point on the base line separating the individuals who have passed and failed may be located such that it has 10% (1000 out 10,000) cases lying between its lower limit and the mean. From NPC table, we find the corresponding σ value for 10% or 1000 cases which is 0.253 σ. Hence the difficulty value of question C is 0.253 σ.

We may summarize the results as follows:

Question	Solved by %	Difficulty value
A	20	0.844 σ
B	30	0.525 σ
C	40	0.253 σ

Case IX: To divide a given group into categories according to an ability or trait assumed to be distributed normally:

Example 4.14: There is a group of 200 students that has to be classified into five categories, A, B, C, D and E, according to ability, the range of ability being equal in each category. If the trait counted under ability is normally distributed; tell how many students should be placed in each category A, B, C, D and E.

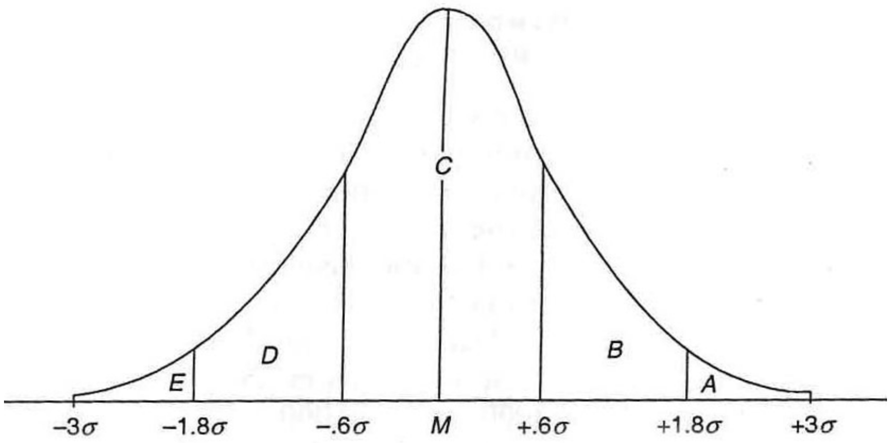


Figure shows division of the area into five equal categories.

As the trait under measurement is normally distributed, the whole group divided into five equal categories may be represented diagrammatically as shown in the above figure. It shows that the base line of the curve, considered to extend from -3 σ to +3 σ, i.e. over a range of 60, may be divided into five equal parts. It gives 1.2 σ as the portion of the base line to be allotted to each category. This allotment may be made in the manner as shown in Figure 8.18 with the various categories demarcated by erecting perpendiculars. Here, group A covers the upper 1.2 σ segment (falling between 1.80 and 30), group B the next 1.20 group C lies 0.6 σ to the right and 0.6 σ to the left of the mean groups D and E covers the same relative positions on the left side of the mean as covered by the groups B and A on the right side of the mean.

After the area of the curve occupied by the respective categories is demarcated the next problem is to find out from the normal curve table the percentage of cases lying within each of these areas.

Let us begin the task with area A. It extends from 1.80 to 30. To know the percentage of cases falling within this area, we read from the table the cases lying between mean and 30 (4986 or 49.86%) and then between M and 1.80 (4641 or 4641%). The difference (49.86 46.41 = 3.45%) will yield the required percentage of the whole group belonging to category A. Therefore, group A may be said to comprise 3.45% or 3.5% of the whole group.

Similarly, group B will cover the cases lying between 0.60 and 1.80. We find from the table that the percentage of the whole group lying between M and 1.80 is 46.41 and between M and 0.60 is 22.57. Therefore, group B may be said to comprise 46.4122.57 of the entire group. 23.84%

Group C extends from -0.60 to 0.60 on both sides of the mean. The normal curve table tells us that 22.57% cases lie between M and 0.60 and a similar percentage of cases, i.e. 22.57, lies between M and -0.60. Therefore, group C may be said to comprise 22.57 x 2 = 45.14 or 45% of the entire group.

Groups D and E, as can be seen from Figure 8.18, are identical to groups B and A and, therefore, may be found to consist of the same percentage of cases as covered in groups B and A, respectively, i.e. 23.8 and 0.5 percent of the whole group.

Now, we may summarize the percentage and number of students in each category in the following manner:

	A	B	C	D	E
Percentage of the whole group in each category	3.5	23.8	45.0	23.8	3.5
Number of students in each category out of the total 200 students	7.0	47.6	90.0	47.6	7.0
Number of students in whole number	7	48	90	48	7

4.6 Summing Up

The literal meaning of the term normal is average. Most of the things and attributes in nature are distributed in a normal way. There are quite a few persons who deviate noticeably from average and a very few who markedly differ from average. If the data in terms of the results of tests, surveys and experiments performed on a randomly selected sample or population are plotted on a graph paper, we are likely to get a typical curve often resembling a vertical cross-section of a bell. This bell-shaped curve is named as normal curve. It is a perfectly symmetrical curve along the vertical middle line. Where the scores of individuals in the group seriously deviate from the average, the curves representing these distributions also deviate from the shape of a normal curve. This deviation from the normality tends to vary either in terms of skewness or in terms of kurtosis.

Skewness refers to the lack of symmetry. In a distribution showing skewness, there is no symmetry between the right and left halves of the curve. There are two types of skewed distributions-negative and positive-indicating the inclination of the curve more towards left or right. It is computed by the formulae

$$S_k = \frac{3\,(M - Mdn)}{\sigma}$$

$$S_k = \frac{P_{90} + P_{10}}{2} - P_{50}$$

Kurtosis refers to the flatness or peakedness of a frequency distribution as compared with the normal. If it is more flat than the normal, it is named as platy-kurtic and if more peaked, then it is called as lepto-kurtic. Kurtosis is computed by the formula

$$K_u = \frac{Q}{P_{90} + P_{10}}$$

A normal curve shows interesting properties and typical characteristics like the following:

1. Mean, median and mode are the same.

2. The value of the measure of skewness is zero.

3. The value of the measure of kurtosis is 0.263.

4. It approaches but never touches the base line at the extremes.

5. The mean of the distribution is used as the starting point. The curve has its maximum height at this point. The distance travelled along the base line from this point is measured in the unit of the Standard deviation of the distribution (σ).

6. The curve extends from - 3σ to 3σ covering a total area of 10,000 (taken arbitrarily). 68.26 percent of this area falls within the limits $\pm 1\sigma$, 95.44 percent within the $\pm 2\sigma$, and 99.74 percent within $\pm 3\sigma$, thus leaving 26 cases in 10,000 beyond the range $\pm 3\sigma$,

7. Limits of the distances $\pm$ 1.96σ include 95 percent and the limits $\pm 2.58\sigma$include 99 percent of the total cases.

Normal curve has a wide significance and applicability and may be used

1. as a model for comparing various distributions and distributing school marks and categorical ratings;
2. to compute percentiles and percentile ranks;
3. to understand and apply the concept of standard errors of measurement;
4. for the purpose of ability grouping;
5. for transforming and combining qualitative data;
6. for converting raw scores into comparable standard norma-lized scores; and
7. for determining the relative difficulty of test items.

4.7 Questions and Exercises

1. What is a normal curve? Why is it named as normal probability curve, normal curve of error or Gaussian curve?
2. What do you understand by the term divergence from normality? Point out the main types of such divergent curves and throw light on the concepts of skewness and kurtosis.
3. Define and explain the terms skewness and kurtosis along with their main types.
4. Discuss the chief characteristics and properties of a normal curve.
5. What do you understand by the terms normal distribution and normal curve? Bring out the main applications of the concept of normal curve in the fields of education, psychology and sociology.

6. Given the following data regarding two distributions,

	Mathematics	Physics
M	60	33
SD	8	9
Achievement scores of a student	70	67

find out whether the student did better in Maths or Physics.

7. Given a normal distribution with a mean of 50 and standard deviation of 15?

(a) What percent of the cases will lie between the scores of 47 and 60?

(b) What percent of the cases will lie between 40 and 47, and

(c) What percent of the group is expected to have scores greater than 68?

134 Statistics in Psychology and Education

8. In a normal distribution with a mean of 54 and SD of 5, calculate the following:

(a) $Q_1$and Q_3(P_{25} and P_{75})

(b) Two scores between which lies the middle 50% cases.

(c) Number of persons out of 600 who score below 57.

9. Given N = 100 M = 28.52 SD = 4.66 assuming normality of the given distribution find

(a) What percent of cases lie between 23-25 and

(b) What limits include the middle 60%?

10. On the assumption that IQ's are normally distributed in the population with mean of 100 and standard deviation of 15, what percentage age of the cases fall

(a) above 135 IQ

(b) above 120 IQ

(c) below 90 IQ

(d) between 75 and 125 IQ

11. Assuming a normal distribution of scores, a test has a mean score of 100 and a standard deviation of 15. Compute the following:

(a) Score that cuts off the top 10%,

(b) Score that cuts off the lower 40%,

(c) Percentage of cases above 90,

(d) Score that occupies the 68th percentile rank, and

(e) Score limits of the middle 68%.

12. Four tests are passed by 15%, 50%, 60% and 75% respectively of a large unrelated group. Assuming normality, find the relative difficulty value of each problem.

13. Given three test items. 1, 2 and 3 passed by 50%, 40% and 30% respectively of a large group. On the assumption of normality of distribution, what percentage of this group must pass test item 4 in order for it to be as much more difficult than 3 as 2 is more difficult than 1?

14. There is a group of 1000 individuals to be divided into 10 sub-groups, i.e. A, B, C, D, E, F, G, H, I and J respectively according to a trait supposed to be distributed normally. What number of individuals should be placed in each of these sub-groups?

15. Scores on a particular psychological test are normally distributed with a mean of 50 and SD of 10. The decision is made to use a letter grade system, as follows: A 10%, B 20%, C 40%, D 20% and E 10%. Find the score intervals for the five letter grades.

4.8 Answers to Check Your Progress

Q.2. 81.85%

Q.3. 97.92%

Q.4.1 137 students

Q.4.2 97.58%

Q.4.3 952 students

Q.4.4 Percentile Rank 97.72

Q.4.5 83.99% students, 5039 students.

4.9 References and Suggested Readings

- Garrett, Henry E. (1984), Statistics in Psychology and Education. Vakils, Feffer and Simons Pvt. Ltd., Bombay,
- Bhandarkar, K.M. (2007), Statistics in Education. Neelkamal Publications PVT. LTD., Hyderabad.
- Saha, Kaberi. (2023), Statistical Analysis in Social Sciences, Mani Manik Prakash, Guwahati.
- Mangal, S.K. (2021), Statistics in Psychology and Education. PHI Learning Private limited, Delhi.
- Goswami, Sadhana. (2013), Statistics in Education. Shanti Prakashan, Guwahati.
- Koul, Lokesh, (2021,), Methodology of Educational Research, Vikash Publishing House Private Limited

*****

Block- II: Regression and Correlation

Unit 1 : Regression Equation, Regression and Prediction

Unit 2 : Product Moment Correlation and Scatter Diagram

UNIT-1

REGRESSION EQUATION, REGRESSION AND PREDICTION

Unit Structure:

1.1 Introduction

1.2 Objectives

1.3 Regression

1.3.1 Significance of Regression

1.3.2 Regression Equation

1.3.3 Prediction from the Regression line

1.4 Difference between correlation and Regression

1.5 Uses of Regression Analysis

1.6 Summing Up

1.7 Questions and Exercises

1.8 Answer to Check Your Progress

1.9 References and Suggested Readings

1.1 Introduction:

Regression analysis is a branch of statistical Theory that is widely used in almost all disciplines. It is an attempts to determine strength and character of the relationship between a dependent variable and one or more independent variable.

The term 'regression' was first used by Francis Galton with reference to the inheritance of stature. When analysts and researchers use the term regression by itself they are typically referring to linear regression focusing on developing a linear model to explain the relationship between predictor variables and a numeric outcome variable. In its formal statistical sense, regression also includes non-linear models that enables a functional relationship between predictors and outcome variables. According to Pedhazur, regression analysis has two uses in scientific literature: prediction including classification and explanation.

1.2 Objectives:

After going through this unit you will be able to:

- Explain about the basics of Regression.
- Explain about what is Regression equation.
- Describe about Regression Equation in score form.
- Describe about the procedure of using Regression equation.
- Explain about the meaning of prediction from regression line.
- Explain about the advantages and disadvantages of Regression and Prediction.

1.3 Regression

Regression is the determination of the measure of a statistical relationship between two or more variables in terms of original units of data. Dictionary meaning of the word regression is 'stepping back' or 'going back'. To be precise, regression analysis helps individuals to determine how changes in one variable are associated with changes in another. It is like finding a mathematical formula that best fits the data and allows to make predictions etc. understand the impact of different factors on outcome. Thus, Regression analysis is done for estimating or predicting the unknown value of one variable from the known value of the other variable. This is very useful statistics tool which is used in natural and social sciences.

1.3.1 Significance of Regression:

The co-efficient of correlation tells us to the way in which two variables are related to each other but it cannot prove to be a good estimate for predicting the change in one variable. For example, we cannot predict the IQ scores of a student with the help of academic achievement scores unless this correlation is perfect correlations are hardly found to be perfect in most of the data related to education and psychology. So, for reliable prediction we generally use the concept of regression lines and regression equations.

Stop to consider

- Francis Galton in later half of nineteenth century first use the word 'Regression'.
- Regression analysis is a statistical tool that helps identify and quantity the relationship between variables.

1.3.2 Regression Equation:

In statistics, a regression equation to come up with an equation like model. This equation like model helps to represent the pattern and patterns present in data.

A regression equation can be defined as a statistical noodle, used to determine the specific relationship between the prediction variable and the outcome variable.

The regression equation of y on × deviating town from the means of y and × is as follows:

$$y = r\frac{\delta y}{\delta x} \times x$$

The regression equation of x on y, deviations taken from the means of x and y is as follows:

$$y = r\frac{\delta x}{\delta y} \times y$$

Stop to consider:

- Regressions equation of y variable on x variable is represented by →

$$y - My = r\frac{\delta y}{\delta x}(X - M_x)$$

- Regression equation of x variable of y variable is represented by→

$$x - M_x = r\frac{\delta y}{\delta x}(Y - M_y)$$

Check Your Progress

i. What is regression equation?

ii. Mention two important significance of regression.

1.3.3 Prediction from the Regression line:

Regression equation and prediction is very useful in psychology and education and also useful in educational guidance and prognosis. In statistics, prediction is the process of using a model to estimate the outcome of an event or situation that is currently unknown eg- if parents were very tall the children tended to be tall but shorter than their parents. If parents were very short the children tended to be short but taller than their parents were. Let us now understand the use of regression with the help of some examples:

Case I: Computation of regression equations when all the desired statistics are given.

Example: 2.1

Given the following data

Marks in History (X)	Marks in English (Y)
M_x = 75.00	M_y = 70.00
δ_x = 6.00	δ_y = 8.00

r = 0.72

Determine the regression equations and predict.

1. The marks is English of a student whose marks in History are 65 and
2. The marks in History of a student whose marks in English are 50.

Solution:

The equation for the prediction of Y is :

$$Y - M_y = r\,\frac{\delta_y}{\delta_x}\,(X - M_x)$$

Substituting all the known values

$$Y - 70 = 0.72 \times \frac{8}{6}\,(X-75)$$

$$= 0.96X - 72$$

$$Y = 0.96X - 2$$

When

X – 65, Y = 0.96×65-2 = 62.4-2=60.4

The equation for the prediction of X is

$$X - M_x = r\,\frac{\delta_y}{\delta_x}\,(Y - M_y)$$

Substituting the values in this equation, we get

$$X = 75 = 0.72 \times \frac{6}{8} \ (Y-70)$$

$$X = 75 + 0.54\ (Y - 70)$$

$$X = 75 + 0.54Y - 37.8$$

$$X = 0.54Y + 37.2$$

When $Y = 50$, we have

$$X = 0.54 \times 50 + 37.2 = 27.00 + 37.2 = 64.2$$

Example 2.2:

The following are the data given for test Maths and English

Maths (X)	English (Y)
Mean = 60	Mean = 50
SD = 10	SD = 12

r = 0.70

(a) Determine both the regression equation.

(b) Predict the probable score in English (y) of a student whose score in History is 70

(c) Predict the probable score in Maths (x) of a student whose score in English is 65.

Solution:

Here

$M_x = 60$ $\qquad$ $M_y = 50$

$\delta_x = 10$ $\qquad$ $\delta_y = 12$

And $r = 0.70$

(a) $\therefore$ The two regression equation are:

$Y - M_y = r\frac{\delta_y}{\delta_x}(X - M_x)$

$Y - 50 = 0.70\frac{12}{10}\ (X - 60)$

$Y - 50 = 0.70 \times 1.2\ (X - 60)$

$Y - 50 = 0.84 \times (X - 60)$

$Y = 0.74x - 50.4 + 50$

$= 0.84x - 0.4$

And $X - M_x = r\frac{\delta_y}{\delta_x}\ (Y - M_y)$

$X - 60 = 0.70\ \frac{12}{10}\ \ (Y - 50)$

$X - 60 = 0.70 \times 0.83\ (Y - 50)$

$X - 60 = 0.58\ (Y - 50)$

$X - 60 = 0.58Y - 29$

$X = 0.39y - 31.98 + 70$

$\therefore$ The regression equation

$Y = 0.8x + 0.4$

$X = 0.58y + 31$

(a) The probable score on English (Y)

$Y - M_y = \ r\frac{\delta_y}{\delta_x}(X - M_x)$ [Here score is History = 70]

$Y = r\frac{\delta_y}{\delta_x}(X - M_x) + M_y$

$Y = 0.70\ \frac{12}{10}(70 - 60) + 50$

$Y = 0.70 \times 1.2\ (10) + 50$

$Y = 0.70 \times 12 + 50$

$Y = 0.84 + 50 = 58.4$

$\therefore$ The probable score in English is 58, when the score in history is 70.

(b) The probable score is Maths (X)

$X - M_x = r\frac{\delta_y}{\delta_x}(Y - M_y)$

by substituting the values in the formula We get

$$X = r\frac{\delta_y}{\delta_x}(Y - M_y) + M_x \text{ (here Y = 65)}$$

$$X - 0.70\,\frac{12}{10}(65 - 50) + 60$$

$$X = 0.70 \times 0.83\ (15) + 60$$

$$X = 0.58\ (15) + 60$$

$$X = 8.7 + 60 = 68.7$$

The probable score in Maths of a student is 68.7 whose score in English is 65.

Case II: Computation of regression equations directly from raw data.

Given the following data, find the two regression equations.

X — 2, 3, 6, 4, 5, 4

Y — 1, 3, 4, 2, 5, 3

Solution : The regression equations are

$$Y - M_y = r\frac{\delta_y}{\delta_x}(X - M_x)$$

$$X - M_x = r\frac{\delta_x}{\delta_y}(Y - M_y)$$

In these equations are :

$$Y - M_y = r\frac{\delta_y}{\delta_x}(X - M_x)$$

$$X - M_x = r\frac{\delta_x}{\delta_y}(Y - M_y)$$

In these equations, $\sqrt[r]{\delta_y \mid \delta_x}$ and $\sqrt[r]{\delta_x \mid \delta_y}$ (called regression co-efficient) are calculated by using following formula :

$$r\frac{\delta_y}{\delta_x} = \frac{N\Sigma XY - \Sigma X.\Sigma Y}{N\Sigma X^2 - (\Sigma X)^2}$$

$$r\frac{\delta_x}{\delta_y}=\frac{N\Sigma XY-\Sigma X.\Sigma Y}{N\Sigma Y^2-(\Sigma Y)^2}$$

Individuals	X	Y	XY	X^2	Y^2
A	2	1	2	4	1
B	3	3	9	9	9
C	6	4	24	36	16
D	4	2	8	16	4
E	5	5	25	25	25
F	4	3	12	16	9
N = 6	$\Sigma X = 24$	$\Sigma Y = 18$	$\Sigma XY = 80$	$\Sigma X^2 = 106$	$\Sigma Y^2 = 64$

Here,

$$M_x = \frac{\Sigma X}{N} - \frac{24}{6} = 4$$

$$M_y = \frac{\Sigma Y}{N} - \frac{18}{6} = 3$$

$$r\frac{\delta_y}{\delta_x} = \frac{6\times 80 - 24\times 18}{6\times 106 - 24\times 24} = \frac{480 - 432}{636 - 276}$$

$$= \frac{48}{60} = \frac{4}{5} = 0.8$$

$$r\frac{\delta_x}{\delta_y} = \frac{6\times 80 - 24\times 18}{6\times 64 - 18\times 18} = \frac{480 - 432}{384 - 324}$$

$$= \frac{48}{60} = \frac{4}{5} = 0.8$$

Now the equation for the prediction

of y is Y – 3 = 0.8 (X–4)

or Y = 3 + 0.8X – 32

or $\quad Y = 0.8X - 0.2$

The equation for the prediction of

$X \text{ is} = 0.8\,(Y - 3)$

$X = 4 + 0.8Y - 2.4$

$X = 0.8Y + 1.6$

Stop to Consider

In statistics prediction is the process of using a model to estimate the outcome of an event or situation that is currently unknown.

Check Your Progress — 2

(i) What is prediction from regression line?

(ii) If the two regression lines are as under

$Y = a + bx$

$Y = c + dy$

What is the correlation coefficient between various X and Y?

a) $\sqrt{bc}$ b) $\sqrt{ca}$ c) $\sqrt{ad}$ d) $\sqrt{bd}$

Self asking questions

1. Given the following data for two tests :

History (X)	Civics (Y)
Mean = 25	Mean = 30
SD = 1.7	SD = 1.6

Co efficient of correlation $r_{xy} = 0.95$

(a) Determine both the regression equations

(b) Predict the probable score in Civics of a student whose score in History is 40.

(c) Predict the probable score in History of a student whose score in Civics is 50.

1.4 Difference between correlation and Regression:

Both the correlation and regression analysis helps us in studying the relationship between two variables yet they differ in their approach and objectives. Let see the differences:

Correlation	**Regression**
1. Correlation between two series is not necessarily a cause and effect relationship. There may be no cause and affect relationship between the variables under study and yet they may be correlated.	1. Regression analysis can be used to make prediction about the dependent variable based on the independent variant.
2. It is used for testing and verifying the relation between two variables and gives limited information.	2. Besides verification it is used for the prediction of one value in relationship to the other given value.
3. The coefficient of correlation is a relative measure. The range of relationship lies between – 1 and +1	3. Regression coefficient is an absolute figure. If we know the value of the independent variable, we can find the value of the dependent variable.
4. In correlation both the variables x and y are random.	4. In regression x is random variable and y is a fined variable. Sometimes both the variables may be random variables.
5. In the coefficient of correlation is position then two variables are positively correlated and vice-versa.	5. The regression coefficient explains that the decrease in one variable is associated with the increase in the other variables.

1.5 Uses of Regression Analysis:

1. In statistical analysis of demand curves supply curves, production function, cost function, consumption function etc., regression analysis is widely used.
2. Regression analysis predicts the values of dependent variables from the values of independent variables.
3. Regression analysis helps in establishing a functional relationship between two or more variables.
4. We can calculate coefficient of correlations (r) and coefficient of determination (r2) with the help of regression coefficient.

Stop to Consider

- In correlation analysis the degree and direction of relationship between the variables are studied.

But in regression analysis the nature of relationship is studied.

- The main uses of regression are fore casting, time modeling and to find out cause–effect relationship between variables.

Check Your Progress– 3

1. Mention one significant difference between regression and correlation.
2. Mention two important uses of regression analysis.

Self-Asking Questions

1. What is the Arithmetical representation of the coefficient of correlation insisting bet –1 and +1?
2. Define regression.

1.6 Summing Up :

Regression analysis is done for estimating or predicting the unknown value statistical tool which is used in natural and social sciences. Regression equations are algebraic expression of the regression lives. Since there are two regression lines, so there are two regression equations. The regression equations of x and y is used to describe the variations in the values of x for given changes in y and regression equation of y or x is used to describe the variation in the values of y for given changes in x.

Regression and predictions are widely in education and behavioural sciences. As regression questions and prediction estimates the values of the dependent variables from the values of the independent variables. Therefore, this equations can be used in education and vocational guidance. But we cannot predict things to happen exactly in the sense way as they happen in practice. This variation between the predicted values or scores and the observed values or scores is termed as the error in prediction.

1.7 Questions and Exercises :

1. A group of fine students obtained the following scores on two achievement tests x and y.

Students	A	B	C	D	E
Scores in X test	12	13	14	11	10
Scores in y test	14	20	22	12	12

i Determine both the regression equations

ii. If a student scores 12 in test x, predict his probable score in test y.

iii. If a student scores 9 in test y, Predict his probable score is test x.

2 A group of five students obtained the following scores on two achievement tests X and Y:

Students score in	A	B	C	D	E
X test	10	11	12	9	9
Y test	12	18	20	10	10

a) Determine both the regression equations.
b) If a student scores 15 in test X, predict his probable score in test Y.
c) If a student scores 5 in test Y, predict his probable score in test X.

1.8 Answer to Check Your Progress:

Check your progress Answer : 1

1. A regression equations can be defined as a statistical equation which can be used to determine the specific relationship bet^n the prediction variable and outcome variable.
2. The two important significance of regression are—

i) It shows whether changes observed in the dependent variable are associated with changes in one or more of the independent variable.

ii) Regression helps economists and financial analysts with challenges ranging from asset valuation to making predictions.

Check your progress Answer : 2

1. A regression line allows you to prediction the value of a dependent variable based on the value of an independent variable.

2. Ans = $\sqrt{bd}$

Check your progress Answer : 3

1. The correlation is perfectly negative when the value of the co-efficient of correlation is –1, and when the correlation is perfectly positive which value of coefficient of correlation is +1.

2. In simple regression is based on r, there are two regression equations :

$$Y - M_y = r\frac{\delta_y}{\delta_x}(X - M_x)$$

$$X - M_x = r\frac{\delta_x}{\delta_y}(Y - M_y)$$

1.9 References and Suggested Readings:

Pedhazur E.I., Multiple Regression in behavioral Research 3rd. ed. Forth Worth Tx: Harcount Brace College Publishers : 1997 (google scholar)

Garrett, H.E., Statistics in Psychology and Education, 6th Indian ed., Vakils, Fefferv and Simon, Bombay, 1971.

Tate, M.W., Statistics in Educaiton, Mc Graw-Hill, New York, 1948.

Mangal, S.K., "Statistics in psychology and education, wnd Edition, PAI Learning Private Limited, Delhi-110092 (2019)

Saha, Dr. Kaberi, Statistics in Education and Psychology. 1st Edition, Asian Books Private Limted, New Delhi, 2012.

*****

UNIT-2

PRODUCT MOMENT CORRELATION AND SCATTER DIAGRAM.

Unit Structure:

2.1 Introduction
2.2 Objectives:
2.3 Meaning of Correlation:
 2.3.1 Types of Correlation
2.4 Computation of coefficient of correlation:
 2.4.1 Product Moment Method:
 2.4.2 Computation of r from ungrouped data
2.5 Scattered diagram
2.6 Questions and Exercises
2.7 Answer to Check Your Progress
2.8 References and Suggested Readings

2.1 Introduction:

In statistical analysis related to the study of relationship between two variables is known as B9-variate analysis. In education and psychology, there are times when one needs to known whether there exists any relationship among the different attributes or abilities of the individual or they are independent of each other.

Eg—

- Does scholastic achievement depend upon the general intelligence of a child?
- Is it true that the height of the children increases with the increase of their ages?
- Is it true that dull children tend to be more neurotic than bright children?

2.2 Objectives:

After going through this unit you will be able to —

- Explain about product moment correlation and its various uses.
- Explain about scattered diagram and its uses in correlation.

2.3 Meaning of Correlation:

For expressing the degree of relationship quantitatively between two sets of measures or variables, we usually take the help of an indent that is known as coefficient of correlation. The relationship between two or more series of variables or sets of data is known as correlation. If the change in one variable appears to be accompanied by a change in the other variables, the two variables are said to be correlated and this inter-dependence is called correlation or co variation.

It is a kind of ratio which expresses the extent to which changes in one variable are accompanied by changes in the other variables. It involves no units and varies from –1 (indicating perfect negative correlation) to +1 (indicating perfect positive correlation). In case the coefficient of correlation in zero, it indicates zero correlation between two sets of Measures.

In the words of A.M. Turtle, "correlation is an analysis the co-variations between two or more variables"

According to Ya-Wen-Chau, "Correlation Analysis attempts to determine the degree of relationship between variables."

2.3.1 Types of Correlation

Correlation is classified into various types : Such as positive and negative linear and non-linear, bi-serial, partial or multiple correlations. Some of the most significant types of correlation have been discussed as follows—

*** Positive Correlation :**

If the two variables tend to on one together in the same direction i.e. an increase in the value of the one variable is accompanied by an increase in the value of the other variable. On the otherhand if a decrease in the value of one variable is accompanied by a decrease in the value of other, then the correlation is called positive correlation eg– A student scores first in the one test and also scores firs in one test and also scores first in another test. Similarly a student who ranked first in one test also ranked third in another test. In such cases relationship is regarded as positive.

*** Negative Correlation:**

Negative correlation occurs when an increase in one variable decreases the value of another. It imply when the relationship bet^{n} two variables are negative that is when a high degree of one trait may be associated with the low degree of another traint in that case the correlation bet^{n} the two traits may be considered as negatives.

*** Zero Correlation:**

When there is no linear depending then no correlation occurs.

Eg- body weight and intelligence shoes size and monthly salary etc. The zero correlation is the mid point of the range –1 +1.

*** Linear Correlation :**

When the relationship bet^{n} two sets of scores or variables can be represented graphically by a straight line then it is known as linear correlation.

Eg-

x	2	4	6	8	10	12
y	3	6	9	12	15	18

Here the ratio of change bet^n the two variables (x and y) is the same.

*** Non linear correlation:**

If the amount of change in one variable does not bear a constant ratio of the amount of change in the other then the relation is called curvilinear or non linear correlation. In multiple correlation more than two variables simultaneously.

eg– quantity of money and price level is simple and the relationship of price, demand and supply of a commodity are example of multiple correlation.

Stop to Consider

- Linear correlation represents the simplest kind of correlation. The relationship between two variables are expressed by a ratio which is known as co efficient of correlation.
- It signifies no of units from +1 to –1 (perfect positive of perfect negative correlation).

Check your progress – 1

1. What is co-efficient of correlation?
2. Mention two important types of correlation?

Self-Asking Question

What do you mean by linear correlation? Enumerate the importance of computing correlation in the field of education.

……………………………………………………………………

……………………………………………………………………

……………………………………………………………………

……………………………………………………………………

…………………………………………………………………….

2.4 Computation of coefficient of correlation:

There are two different methods of computing co-efficient of linear correlation :

1. Rank difference Method
2. Product Moment Method

1. Rank difference Method:

Rank difference method of computing co-efficient of correlation was developed by the famous psychologists and statistician Edward Spearman in 1904. This method is based on ranks of the individuals in the characteristics or variables A and B. It is designated by the G reok alphabet g (Rho). eg– there are some situations in order of merit on two variables when these two sets of rank have agreement bet^{n} them. We measures the degrees of relationship by rank difference correlation.

The formula for calculating co-efficient of correlation through rank difference method is as follows:

$$\text{P (Rho)} = 1 - \frac{6\Sigma d^2}{N(N^2 - 1)}$$

where

P (Rho) = The value of rank difference co-efficient of correlation.

D = difference between rank 1 (R_1) and Rank 2 (R_2)

N = Number of scores or individuals or observations

Stop to consider

- Spearman's Rank difference method is most simple method of computing coefficient of correlation.
- The formula of Rank difference method is $\text{P (Rho)} = 1 - \frac{6 \times \Sigma d^2}{N(N^2 - 1)}$

Example 2.1

Following series in the data of 10 students in two trialy of list with gap of two weeks:

Students	A	B	C	D	E	F	G	H	I	J
Trial I (x)	10	15	11	14	16	20	10	8	7	9
Trial II (y)	16	16	24	18	22	24	14	10	12	14

Compute the co-efficient of correlation between the scores of two trials by rank difference Method.

Solutions:

Table – 1 Computation of Spearman's Rank difference Co-efficient of Correlation.

Students	Trial I (x)	Trial II (y)	R_1	R_2	d (R_1-R_2)	d2
A	10	16	6.5	5.5	1	1
B	15	16	3	5.5	2.5	6.25
C	11	24	5	1.5	3.5	12.25
D	14	18	4	4	0	0
E	16	22	2	3	1	1
F	20	24	1	1.5	0.5	0.25
G	10	14	6.5	7.5	1	1
H	8	10	9	10	1	1
I	7	12	10	9	1	1
J	9	14	8	7.5	0.5	0.25
N+10						$\Sigma d^2 = 24$

$$\text{P (Rho)} = 1 - \frac{6 \times \Sigma d^2}{N(N^2 - 1)}$$

$$= 1 - \frac{6 \times 24}{10(10^2 - 1)}$$

$$= 1 - \frac{144}{10(100 - 1)}$$

$$= 1 - \frac{144}{990}$$

$$= 1 - 0.145$$

$$P = 0.855$$

The Rank difference Co-efficient of correlation (P.(Rho) is 0.86

*** Merits & Limitations of Rho (P) :**

- Spearman's Rank order Co-efficient of correlation can be easily computed.
- It is quite easy to interpret P.

Some of the limitations of rank difference methods.

- When the interval data is converted into rank order data, the information about the size of the score difference is lost.

Check Your Progress : 2

1. Rank difference Method was discovered by _________.
2. Write one limitation of Rank difference Method.

Self-Asking Question :

1. Find rank correlation co-efficient from the following data and interpret the results :

Individuals	A	B	C	D	E	F	G	H
Marks in Hindi	30	40	50	20	10	45	22	18
Marks in English	55	75	60	12	11	38	25	15

2. Find the correlation between the following two sets of raw scores and Interpret the results.

Individuals	A	B	C	D	E	F	G	H	I	J
Test X	13	12	10	8	7	6	6	4	3	1
Test Y	7	11	3	7	2	12	6	2	9	6

2.4.1 Product Moment Method:

Karl Pearson, a great biometrician and statistician, suggested a mathematical method for measuring the magnitude of linear relationship between two variables. It is popularly known as Pearson's co-efficient of correlation and it is denoted by the symbol of r. If the two variables under study are x and y, the following formula suggested by Karl Pearson can be used for measuring the degree of relationship

$$rxy = \frac{\Sigma xy}{N\delta \times \delta y}$$

Where

rxy = correlation between x and y (two sets of scores)

x = Deviation of any x – score from the mean in the list x.

y = Deviation of the corresponding y score from the mean in testy y.

Σ xy = Same of all the products of Deviation (Each x deviation multiplied by its corresponding y deviation)

δx = Standard deviation of the distribution of the scores in test.

N = Total no of scores or frequencies.

In this formula, the basic quantity to determine the degree of correlation between two sets of variables x and y is Σ xy/N is known as the product moment and the corresponding correlation is called the product moment correlation.

Stop to Consider

- Karl Pearson is a great biometrician and statistician.
- Karl Pearson's method of measuring degree of relationship in also known as Pearson's Product Moment Method.
- Symbolic form of product Moment method is $r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 . \Sigma y^2}}$

Check Your Progress : 3

1. Product Moment Method is also known as ________.
2. Who discovered the Product Moment Method of coefficient of correlation?

Self-Asking Questions

What is Pearson's Product Moment Method of computing correlation? Why has it been named so?

……………………………………………………………………………
……………………………………………………………………………
……………………………………………………………………………

2.4.2 Computation of r from ungrouped data:

For computation of r from ungrouped data the formula—

$$r = \frac{\Sigma xy}{N\delta \times \delta y}$$

Eg—

Individuals	Scores in test X	Score in test Y	x	y	xy	x^2	y^2
A	15	60	–10	10	–100	100	100
B	25	70	0	20	0	0	400
C	20	40	–5	–10	50	25	100
D	30	50	5	0	0	25	0
E	35	30	10	–20	–200	100	400
						$\Sigma x^2=250$	$\Sigma y^2=1000$

Mean of series x, Mx = 25

Mean of series y, My = 50

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

$$= \frac{-250}{\sqrt{250{,}000}}$$

$$= -\frac{1}{2}$$

$$= -0.5$$

∴ Product moment correlation coefficient = – 0.5

Example 2.2 :

Calculation of Pearsons Product moment ‘r’ from raw scores:

Subjects	Scores (x)	Scores (y)	xy	x^2	y^2
A	5	12	60	25	144
B	3	15	45	9	225
C	2	11	22	4	121
D	8	10	80	64	100
E	6	18	108	36	324
N=5	$\Sigma x = 24$	$\Sigma y = 66$	$\Sigma xy = 315$	$\Sigma x^2 = 138$	$\Sigma y^2 = 914$

In this situation the value of

N = 5

$\Sigma x = 24$

$\Sigma y = 66$

$\Sigma xy = 315$

$\Sigma x^2 = 138$

$\Sigma y^2 = 1914$

Using the formula we find that

$$rxy = \frac{N.\Sigma xy - \Sigma x.\Sigma y}{\sqrt{[N.\Sigma x^2 - (\Sigma x)^2][N.\Sigma y^2 - (\Sigma y)^2]}}$$

$$= \frac{(5)(315) - (24)(66)}{\sqrt{[(5)(138) - (24)^2][(5)(914) - (66)^2]}}$$

$$= \frac{1575 - 1584}{\sqrt{[(690 - 576)][(4570 - 4356)]}}$$

$$= \frac{-9}{\sqrt{114(214)]}}$$

$$= \frac{-9}{\sqrt{24396}}$$

$$= \frac{-9}{156.19}$$

$$= -0.06$$

The computed value of Pearson's Product Moment 'r' is –0.06.

Example 2.3 :

Compute the correlation bet^n the two sets of scores given below by using the method, when deviations are taken from zero—

	A	B	C	D	E	F	G	H	I	J	K	L
X	74	71	71	67	65	61	62	60	59	56	54	50
Y	70	36	34	28	30	32	30	26	28	34	25	22

Calculation of r when deviation are Original Scores S (AMS = 0)

	Test I X	Test II Y	x^2	y^2	xy
A	74	40	5476	1600	2960
B	71	36	5041	1296	2556
C	71	34	5041	1156	2414
D	67	28	4489	784	1876
E	65	30	4225	900	1950
F	61	32	3721	1024	1952
G	62	30	3844	900	1860
H	60	26	3600	676	1560
I	59	28	3481	784	1652
J	56	34	3136	1156	1904
K	54	25	2916	624	1350
L	50	22	2500	484	1100
$\Sigma = 72$	$\Sigma x = 750$	$\Sigma y = 365$	$\Sigma x^2 = 47470$	$\Sigma y^2 = 11385$	$\Sigma xy = 23134$

Mx = 750 ÷ 12 = 62.50

My = 365 ÷ 12 = 30.42

Here

$$r = \frac{N.\Sigma xy - \Sigma x.\Sigma y}{\sqrt{[N.\Sigma x^2 - \Sigma x^2][N.\Sigma y^2 - \Sigma y^2]}}$$

∴ Substituting the required value in the above formula we get—

$$= \frac{12 \times 23134 - 750 \times 365}{\sqrt{[12 \times 47470 - (750)^2 \times 12 \times 11385 - (365)^2]}}$$

$$= \frac{277608 - 273750}{\sqrt{(56940 - 562500) \times (136620 - 133225)}}$$

$$= \frac{3858}{\sqrt{24240300}}$$

$$= \frac{3858}{4923.44}$$

$$= 0.78$$

$\therefore$ The coefficient of correlation is 78 which is highly positive.

Advantages and Limitations of Pearson's Co-efficient of Correlation :

Advantages :

- The product moment deals with the size of the scores as well as with their position in the series.

Disadvantages :

- Pearson's product moment co-efficient of correlation always assumes linear relationship regardless of the fact, whether that assumption is true or not.

Stop to Consider

- Computation of r from upgrouped data the formula is —

$$r = \frac{\Sigma xy}{N.\delta x.\delta y}$$

Check Your Progress – 4

1. Mention one Advantage of product moment Method.
2. Mention the method formula of calculating Co-efficient of correlation by product moment method.

Self Asking Questions:

1. Calculate coefficient of correlation by product moment method from the following:

X	21	22	23	24	25	26	27
Y	16	15	17	18	19	20	21

…………………………………………………………………………

…………………………………………………………………………

…………………………………………………………………………

…………………………………………………………………………

2.5 Scattered Diagram :

Construction of Scattered Diagram:

Examples:2.4

Ten students have obtained the following scores on tests in History and Hindi. Express these scores through a scattered diagram.

Individuals	A	B	C	E	E	F	G	H	I	J
Scores in History (X)	13	12	10	8	7	6	6	4	3	1
Scores in Hindi (Y)	7	11	3	7	2	12	6	2	9	6

Procedure:

To construct a scatter diagram from the given data, we usually proceed in the following manner:

1. First of all, the given x and y scores in the two subjects are arranged into two frequency distributions by a suitable choice of

size and the number of class intervals. In the present case, we can select these classes in the two distributions as under →

Scores in History	
X	f
13-14	1
11-12	1
9-10	1
7-8	2
5-6	2
3-4	2
1-2	1
	N=10

Scores in Hindi	
Y	f
11-12	2
9-10	1
7-8	2
5-6	2
3-4	1
1-2	2
	N=10

2. A table is constructed with columns and rows. In the first row of the table, the Class Intervals for the variable x are to be written and along the left margin, the class Intervals for the variable y are to be written. In the present case, we begin the listening of class intervals belonging to x on the top from 1-2 and end with 13-14 and y on the left margin from class intervals 1-2 to 11-12. Here, we have a correlation matrix of size equal to the number of class intervals in the x distribution multiplied by the number of class intervals in the y distribution. We also add one column and one row to the matrix for the purpose of presenting the summation of the frequencies in the respective rows and columns.

3. Now we come again to the individual scores in test x and y we make one tally mark for each individual's x and y scores. For example, for the individuals C who score 10 in test x and 3 in test y,

we place a tally mark in the cell at the intersection of the column for interval 9-10 in x and the row for interval 3-4 in y. Similarly, the scores 1 and 6 of individuals J, a tally mark is put in the cell at the intersection of the column for interval 1-2 and row for interval 5-6. All other individuals may be similarly tailed in their cells.

4. After completion of the task of tallying, these tally works are translated into cells frequencies. These frequencies are written in each of the cells.

5. The cell frequencies in each of the rows will be summed up and sums will be written in the last column under the heading fy. Similarly the cell frequency in each of the columns may also be summed up and the sums are recorded in the bottom row under the heading fx.

6. The frequencies recorded in the last column under the leading fy are the summed up. This sum represents the total number of frequencies (N) for the distribution y. Similarly, by seeming up the frequencies recorded in the bottom row under the head fx we may got the total of all the frequencies (N) under the distribution (x).

Scattered Diagram for data given in Example x score in History.

Y scores on Hindi

Class Intervals	1-2	3-4	5-6	7-8	9-10	11-12	13-14	fy
11-12			/1			/1		2
9-10		/1						1
7-8				/1			/1	2
5-6	/1		/1					2
3-4					/1			1
1-2		/1		/1				2
fx	1	2	2	2	1	1	1	10

Stop to Consider:

- A scattered diagram helping to identify potential correlation bet^{n} two variables.
- It displays visual representation of data, making it easier to spot patterns, trends and outliers.
- It predict the behavior of one variable on the other.

Change Your Progress: 5

i. What is scattered diagram?

ii. When to use scattered diagram

Self Asking Questions:

Find the Pearson's product moment r from the following Scatter diagram (correlation table) and interpret the results scores on Test X.

Scores on Test y

Class Intervals	0-4	5-9	10-14	15-19	20-24	25-29	30-34	Total
15-17			1			1	2	4
12-14		1	1	3	1	2	1	9
9-11		1	4	5	5	1		16
6-8	2	3	5	4	5	2		21
3-5	2	5	1		1	1		10
0-2	2	1		1				4
Total	6	11	12	13	12	7	3	64

2.6 Questions and Exercises

1. Compute co-efficient of correlation by Pearson's Product Moment Method from the following data.

Subject	Scores (X)	Scores (Y)
A	5	12
B	3	15
C	2	11
D	8	10
E	6	18

1. Compute the co efficient of correlation by Pearson's Product moment method from the following data:

Test X	60	54	45	36	39	59	48	47	54	66
Test Y	72	70	36	80	73	78	83	84	50	56

2.7 Answer to Check Your Progress

Check your progress Answers : 1

1. Co-efficient of correlation is a kind of ratio which express the extent to which changes in one variable are accompanied by charge in the other variable.
2. The two important types of correlation are positive and negative correlation.

Check your progress Answer : 2

1. Rank difference Method was discovered by Spearman.

2. One important limitation of Rank difference Method was _____ It only consider the relative order of data points disregarding the actual numerical difference bet^{n} values.

Check your progress Answer : 3

1. Product Moment Method is also known as Pearson's Product Moment correlation coefficient.
2. Karl Pearson discovered the Product Moment Method of coefficient of correlation.

Check your progress Answer : 4

1. One advantage of Product moment method is ___ It is widely used method to measure the strength and direction of a linear relationship bet^{n} two variables.
2.

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

$$\left(\frac{\Sigma \mathrm{xy}}{\sqrt{\Sigma \mathrm{x}^2 . \Sigma \mathrm{y}^2}} \right)$$

2.8 References and Suggested Readings:

Aggarwal, Y.P. (1990) Statistical Methods, Concepts, Applications and Computations, New Delhi : Sterling Publishers Pvt. Ltd.

Garrel, H.E. & Wood Worth, R.S (1981). Statistics in Psychology and Education (Tenth Indian Reprint) Bombay: Vakils, Feffer and Simons Ltd.

Tata, M.W, Statistics in Education, Mc. Graw Hill, New York. 1948.

Mangal S.K. Statistics in Psychology and Education, 2nd Edition, PHI Learning Private Limited, Delhi – 110092 (2019)

Saha, Dr. Kaberi, Statistics in Education and Psychology 1st Edition, Asian Books Private Limited, New Delhi, 2012

*****

Block- III: The Significance of the Other Statistics and the Difference between Means

Unit 1 : Statistical Inferences

Unit 2 : Standard Error of Mean of Large and Small Sample

Unit 3 : Significance of Mean and Other Statistics

Unit 4 : Paramatric Test

Unit 5 : The Hypotheses

Unit 6 : Significance of the Difference Between Means

UNIT-1
STATISTICAL INFERENCES

Unit Structure:

1.1 Introduction
1.2 Objectives
1.3 Meaning of statistical inferences
1.4 Need for statistical inferences
1.5 Importance of statistical inferences
1.6 Statistical Decision Theory
1.7 Level of Significance
1.8 Degrees of Freedom
1.9 Confidence interval
1.10 Concept of t distribution
1.11 Summing Up
1.12 Questions and Exercises
1.13 References and Suggested Readings

1.1 Introduction:

Statistical Interference is the process of using data analysis to deduce properties of an underlying distribution of probability. Inferential statistical analysis infers properties of a population for example by testing hypothesis and deriving estimates. It is assumed that the observed data set is sampled from a large population.

Inferential statistics can be contrasted with description statistics is solely concerned with properties of the observed data and it does not test on the assumption that the data come from a larger population. In machine learning, the term inference is sometimes used instead to mean "make a prediction, by evaluating an already trained model". In this context deducting properties of the model is referred to as training or learning and using a model for prediction is referred to as inference.

Any Statistical inference requires some assumptions. A statistical model is a set of assumptions concerning the generation of the observed data and similar data. Descriptions of statistical models usually emphasize the role of population quantities of interest, about which we wish to inference. "Descriptive statistics are typically used as a preliminary step before more formal inferences are draw",

The four pillars of statistical inference are:

i) Significance
ii) Generalization
iii) Estimation and
iv) Causation

1.2 Objectives:

After going through this unit you will be able to-

- know the meaning of statistical inferences
- Understand the need and importance of statistical inferences
- Explain what is Statistical Decision Theory?
- Discuss the Level of Significance and the Concept of t distribution
- Define Degrees of Freedom and Confidence interval

1.3 Meaning of Statistical Inference

The primary objective of statistical inference is to enable us to generalize from a sample to some larger population of which the sample is a part. As it is not possible to measure all of the members of a given population and hence we must usually be content with samples drawn from this population. Furthermore, owing to differences in the composition of our samples, means and σ computed from such groups will tend to be sometimes larger and sometimes smaller than their population values. Ordinarily, we have only the single sample; and our problem becomes one of determining how well we can infer or estimate the M_{pop}, for example, from the one sample mean. Means and other measures computed from samples are called statistics, and are subject to what are called "fluctuations of sampling". Measures descriptive of a population, on the other hand, are called parameters and are to be thought of as fixed reference values.

We do not know, of course, the parameters of a given population. But we can under specified conditions -forecast the parameters from our sample statistics with known degrees of accuracy. The degree to which a sample mean represents its parameter is an index of the significance or trustworthiness of the computed sample mean. When we have calculated a statistic, therefore we may ask ourselves this question: "How good an estimate is this statistic of the parameter based upon the entire population from which my sample was drawn?" The aspects of this, is being in this unit to provide methods which will enable us to answer this question for the mean, the median, and for other statistics.

As you know that, statistical inference is a fundamental part of statistics that allows us to draw conclusions about a population based on data collected from a sample. Let us breakdown of its need and importance:

1.4 Need for Statistical Inference

1. **Impracticality of Full Population Data:**
    - It's often impossible or too costly to collect data from an entire population (e.g., surveying every voter in a country).

- Statistical inference allows us to make educated guesses about the population using a manageable sample.

2. **Uncertainty and Variability:**

- Real-world data is variable and uncertain.
- Statistical inference provides tools (like confidence intervals and hypothesis tests) to make sense of this uncertainty.

3. **Decision-Making Under Uncertainty:**

- Businesses, governments, and researchers need to make decisions with incomplete data.
- Statistical inference helps guide these decisions with quantitative evidence.

4.Testing Hypotheses:

- Allows us to formally test assumptions (hypotheses) about a population.
- **Example:** A teacher wants to know if a new teaching method improves test scores. By comparing test scores from students who used the new method versus the old, statistical inference can confirm or reject the effectiveness.

5. **Estimating Parameters:**

- Enables estimation of population parameters (like the mean or proportion) using sample statistics.
- **Example:** A market researcher uses a sample survey to estimate the average amount customers are willing to pay for a new product.

6. **Quality Control and Improvement:**

- Vital for monitoring and improving product quality in manufacturing.

- **Example:** A factory inspects a sample of items from a production batch. Inference tells them whether the entire batch meets quality standards.

Stop to Consider

Statistical inference is the process of drawing conclusions about a population based on data from a sample. It's a corner stone of data analysis and decision-making in various fields like science, business, healthcare, and economics.

1.5 Importance of Statistical Inference

1. **Generalization from Sample to Population:**
    - Enables researchers to draw conclusions beyond the observed data.
    - For example, medical studies use inference to generalize treatment effects from a group of patients to the general population.
2. **Testing Hypotheses:**
    - Allows us to test assumptions or theories (e.g., "Does a new drug work better than the existing one?").
    - Hypothesis testing helps validate or reject scientific claims.
3. **Confidence in Results:**
    - Inference provides measures like confidence intervals and p-values that express the reliability of results.

- This helps distinguish between real effects and random chance.

4. **Guidance in Policy and Strategy:**
    - In fields like economics, public health, marketing, and education, statistical inference supports evidence-based policy and strategy development.
5. **Supports Predictive Modeling:**
    - It forms the basis of many predictive analytics techniques, crucial in data science, AI, and machine learning.

Stop to Consider

Statistical inference is essential because it transforms raw sample data into meaningful insights about the larger population, enabling informed decision-making across virtually all scientific, industrial, and governmental fields.

Check Your Progress

Q.1 what is the need of statistical inferences?

Q.2 state the importance of statistical inferences.

1.6 Statistical Decision Theory

We often make decisions about population on the basis of our sample data information. Such decisions are called statistical decisions. For example, we may wish to decide on the basis of our sample data whether a particular teaching method is better than the other, whether a new serum is really effective in curing a particular disease or a particular procedure is better is better than another.

Null Hypotheses

In answering attitude towards reading, we formulate a hypothesis that there is no difference between boys' and girls' attitude towards reading Or Boys' and girls' attitude towards reading do not differ.

(i.e.. any observed differences are due merely to fluctuations in sampling from the same population). Such hypotheses are often called **"null hypotheses"** and are denoted by Ho.

Alternative Hypotheses

Any hypothesis that differs from a given hypothesis is called an alternative hypothesis. For example, in answering attitude towards reading, we formulate a hypothesis that "there is a difference between boys and girls attitude towards reading" Or "Boys' and girls' attitude towards reading differ". A hypothesis alternative to the null hypothesis is denoted by H1, or H A.

1.7 Level of Significance

In a testing a given hypothesis, the maximum probability with which we would be willing to risk a Type I error is called the **level of significance**, or a **significance level**, of the test. This probability, often denoted by a, is generally specified before any samples are drawn so that the results obtained will not influence our choice.

In practice, a significance level of 0.05 or 0.01 is costmary, although other vales are used. If, for example, the 0.05 (or 5%) significance level is chosen in designing a decision rule, then there are about 5 chances in 100 that we would reject the hypothesis when it should be accepted, that is, we are about 95% **confident** that we have made the right decision. In such case we say that the hypothesis has been rejected at the 0.05 significance level which means that the hypothesis has a 0.05 probability of being wrong.

1.8 Concept of t distribution

We usually (however, quite arbitrarily) take a sample of 30 cases or more as a large sample and a sample of less than 30 as small. We have seen that the means of large samples randomly drawn from the same population produce a normal distribution (Figure 9.1). However, when samples are less than 30 in size, the distribution does not take the shape of a normal curve. It is symmetrical but leptokurtic. In other words, it becomes more peaked in the middle and has relatively more area in its tails. Such a distribution is known as the / distribution. It was developed originally in 1908 by W.S. Gosset, who wrote under the pen name "Student". distribution is sometimes called "Student" distribution in honour of its discoverer.

It should be clearly understood that I distribution is not a single curve (as in the case of a perfect normal curve) but a whole family of curves, the shape of each being a function of sample size. As the size of the sample increases and includes about 30 cases, the / distribution takes the shape of a normal curve.

The nature of the distribution is also linked to the number of degrees of freedom. A different / distribution exists for each number of degrees of freedom. As the number of degrees of freedom increases, the / distribution takes the shape of a normal curve. This fact may be well illustrated through Table C of the distribution given in the Appendix. In this table, we find the values of f in relation to the given degrees of freedom and level of confidence. We observe that, as the number of degrees of freedom increases and approaches infinity, approaches the values 1.96 and 2.58, respectively, at 5% and 1% levels of confidence (it is seen that these are the values of sigma scores at the 5% and 1% levels of confidence in the case of a normal distribution).

1.9 Degrees of Freedom (df)

The concept of degrees of freedom is explained like this - if we have 5 scores namely 5, 6, 7, 8, and 9. The mean is 7. The deviations of these scores from the mean are: -2, -1, 0, 1, 2. The sum of these deviations is zero. Of the five deviations, only 4 (N-1) can be selected "freely" (i.e. are independent) as the condition that the sum equal zero immediately restricts the value of (fixes) the 5th deviate.

The SD is, of course, based upon the squares of the deviations taken around the mean. There are N df for computing the mean, but only (N-1) are available for the SD as one df is lost in calculating the mean. Therefore, for N. take (N-1) degrees of freedom. The degrees of freedom are not always (N-1) however, but will vary with the problem and the restrictions imposed. In estimating the dependability of an r, for example (which depends upon the deviations from two means), the df are (N-2).

Self-Asking Question

Q.1 what do you mean by null hypothesis?

…………………………………………………………………………

…………………………………………………………………………

…………………………………………………………………………

…………………………………………………………………………

Q.2 what is level of significance?

…………………………………………………………………………

…………………………………………………………………………

……………………………………………………………………

……………………………………………………………………

Q.3 define the concept of t distribution.

……………………………………………………………………

……………………………………………………………………

……………………………………………………………………

……………………………………………………………………

1.10 Confidence Intervals

What is a Confidence Interval?

A confidence interval is a range around the measurement that conveys how the precise is the measurement. We all most all come across in some situations about the measurement in question is a proportion or a rate. Confidence intervals are often seen in the newspapers about weather reporting, and exit polls reporting at the time of elections, etc.

What does a Confidence Interval Tell Us?

The confidence interval tells us more than just the possible range around the estimate. It also tells us about how stable the estimate is. A stable estimate is one that would be close to the same value if the survey were repeated. An unstable estimate is one that would vary from one sample to another. Wider confidence intervals in relation to the estimate itself indicate instability.

How are Confidence Intervals Calculated?

Confidence intervals are calculated based on the standard error of a measurement. Once the standard error is calculated, the confidence interval is determined by multiplying the standard error by a constant that reflects the level of significance desired, based on the normal distribution. The constant for 95 percent confidence intervals is 1.96.

Finding a Confidence Interval for a Small Sample

If we have a small set of data (under 30 items), we want to use the t distribution instead of the normal distribution to construct the confidence interval.

Example: Construct a 98% Confidence Interval for the following data. 45.55.67, 45,68,79,98,87, 84, 82.

The formula for constructing a Confidence Interval with the t-distribution

$$\bar{X} \pm t\frac{s}{\sqrt{n}}$$

Step-1: Find the mean, μ and standard deviation, for the data.

$\mu = 71 \quad \sigma = 18.172$

Step-2: find df: df = 9

Step-3: Subtract the confidence level 98% from 1, then divide by two, we get alpha level.

(1 -.98) / 2 =.01

Step-4: Look at the t-distribution table. For **df** = 9 and σ = 0.01 the table value is **2.821**.

Degrees of freedom in the left column of the t distribution table.

Step-5: Divide SD by the square root of your sample size. (step1)

18.172 /√(10)) = **5.75**

Step-6: Multiply step 4 by step 5.

2.821 ×5.75 = **16.22**

Step-7: For the lower end of the range, subtract step 6 from the mean (Step 1)

71-16.22 = **54.78**

Step-8: For the upper end of the range, add step 6 to the mean (Step 1).

71+16.22 = **87.22**

This is how we find a confidence interval using the t-distribution.

1.11Summing up

Let us sum up the unit. In this unit we have studied that statistical inference is the process of drawing conclusions about a population based on data from a sample. It's a corner stone of data analysis and decision-making in various fields like science, business, healthcare, and economics. It is essential because it transforms raw sample data into meaningful insights about the larger population, enabling informed decision-making across virtually all scientific, industrial, and governmental fields. We have also studied the level of significance, concept of t distribution and degrees of freedom too.

1.12 QUESTIONS AND EXERCISES

I. Short Answer Questions

1. Explain statistical inference
2. What is standard Error?
3. What is the formula for probable Error of Mean?
4. What is level of significance?
5. Explain degrees of freedom (df) with example.
6. What is confidence intervals?
7. Calculate PEr, if r = 67 and N = 10
8. Calculate PEM when SD is 5.4 and N is 40.

9. "Null Hypothesis is always better" explain.

10. Compute the 't' value for the following data.

M1 = 70 M2= 60 $\sigma 1$= 12 $\sigma 2$= 10 N1= 20 N2 = 25

II. Essay Type Questions

1. Explain statistical inference and its uses.
2. "Type-1 and Type-II Errors are inter related" discuss.
3. Briefly discuss about the statistical decision theory.
4. If N = 20 Mean = 42.25 and SD = 13.81 . Can we depend on this mean when the sample taken is assumed to be normal?

1.13 References and Suggested Readings

Garrett, Henry E. (1984), Statistics in Psychology and Education. Vakils, Feffer and Simons Pvt. Ltd., Bombay,

Bhandarkar, K.M. (2007), Statistics in Education. Neelkamal Publications PVT. LTD., Hyderabad.

Saha, Kaberi. (2023), Statistical Analysis in Social Sciences, Mani Manik Prakash, Guwahati.

Mangal, S.K. (2021), Statistics in Psychology and Education. PHI Learning Private limited, Delhi.

Packiam, S (2023), Statistics Applied in Educational Research, Neelkamal Publications PVT. LTD., Hyderabad.

Goswami, Sadhana. (2013), Statistics in Education. Shanti Prakashan, Guwahati.

Koul, Lokesh, (2021,), Methodology of Educational Research, Vikash Publishing House Private Limited.

*****

UNIT-2
STANDARD ERROR OF MEAN OF LARGE AND SMALL SAMPLE

Unit Structure:

2.1 Introduction

2.2 Objectives

2.3 Measures of Errors

    2.3.1 Standard Error

    2.3.2 Type I and type II Error

    2.3.3 Probable Error

2.4 Significance of the sample mean and other statistics

2.5 Standard error in computing the significant value of the mean

2.6 Summing Up

2.7 Questions and Exercises

2.8 References and Suggested Readings

2.1 Introduction

The **Standard Error of the Mean (SEM)** is a statistical measure that quantifies how much the sample mean of a dataset is expected to vary from the true population mean. In simple terms, it tells us how precise our sample mean is as an estimate of the population mean.

When we collect data and calculate a sample mean, we're using that sample to make inferences about a larger population. However, due to random sampling variability, different samples will produce slightly different means. The SEM gives us an estimate of this variability. It is calculated using the formula:

$$SE_M = \sigma_M = \frac{\sigma}{\sqrt{N}}$$

where:

- σ is the sample standard deviation,
- N is the sample size.

The SEM decreases as the sample size increases, meaning that larger samples provide more reliable estimates of the population mean. SEM is often used in constructing confidence intervals and in hypothesis testing, making it a crucial concept in inferential statistics.

2.2 Objectives

After going through this unit you will be able to-

- Understand the concept of standard error
- Know various Measures of Errors
- Describe the significance of the sample mean and other statistics
- Analyse standard error in computing the significant value of the mean

2.3 Measures of Errors

2.3.1 Standard Error

The terms "standard error" and "standard deviation" are often confused. The contrast between these two terms one need to understand and appreciate them.

The standard deviation **(SD)** is a *measure of variability*. When we calculate the sample mean we are usually interested not in the mean of this particular sample, but interested in the mean of the whole population. If we take-up various samples' means, means will vary from sample to sample; the way this variations occurs is described by the "sampling distribution fluctuation" of these means. We can estimate how much the sample means will vary from the standard deviation of this sampling distribution, which we call it as the standard error **(SE)** estimate of the mean. This measure is the precision of the sample mean.

The standard error of the mean is the **standard deviation of those sample means over all possible samples drawn from the population.**

Secondly, the standard error of the mean can refer to an estimate of that standard deviation, computed from the sample of data being analyzed at the time.

Mathematically, the standard error of the mean formula is given by:

$$SE_M = \sigma_M = \frac{\sigma}{\sqrt{N}}$$

$SE_M = \sigma_M$ = standard error of the mean

σ = the standard deviation of the original distribution

N = the sample size

It can be seen from the formula that the standard error of the mean decreases as N increases. This is expected because if the mean at each step is calculated using many data sets, then there by a small deviation in one value will cause less effect on the final mean reaching the exact mean or true mean of the population. This occurs when the standard error of mean is zero.

The standard deviation is considered as one of the best measures of dispersion, which gauges the dispersion of values from the central value. On the other hand, the standard error is mainly used to check the reliability and accuracy of the estimate and so, the smaller the error, the greater is its reliability and accuracy. The standard error is also used to calculate P values in many circumstances.

2.3.2 Type I and Type II Error

If we reject a hypothesis when it should be accepted, we say that a Type I error has been made. If, on the other hand, we accept a hypothesis when it should be rejected, we say that a Type II error has been made. In either case, a wrong decision or error in judgment has occurred.

The difference between Type 1 and Type II errors is that in the first one we reject Null Hypothesis even if it's true, and in the second case we accept Null Hypothesis even if it's false.

Type I Error

Example: Your Hypothesis: Men are better drivers than women. Null Hypothesis. Men are not better drivers than women. Type 1 Error happens if we reject Null Hypothesis, but in reality we should have accepted it (because men are not better drivers than women). Similarly, if we accept Null Hypothesis, but in reality we should have rejected it, then Type II error is made.

Type II Error

Example: Your hypothesis states that people riding a motorcycle are more hostile than those driving a car. Null Hypothesis would be: People riding a motorcycle are as hostile as those driving a car. Type II Error is made if you conclude that both groups have the same level of hostility, but you should have rejected Null Hypothesis since there is a difference in hostility between these two groups. In other words, if you accept Null Hypothesis and state that both motorcycle riders and car drivers are equally hostile, then you made a Type II Error.

In order for decision rules to be good (or test of hypotheses) to be good, they must be designed so as to minimize errors of decision. This is not a simple matter, because for any given sample size, an attempt to decrease one type of error is generally accompanied by an increase in the other type of error. In practice, one type of error may be more serious than the other, and so a compromise should be reached in favour of limiting the more serious error. The only way to reduce both types of error is increase the sample size, which may or may not be possible.

Stop to Consider

Mathematically, the standard error of the mean formula is given by:

$$SE_M = \sigma_M = \frac{\sigma}{\sqrt{N}}$$

$SE_M = \sigma_M$ = standard error of the mean

σ = the standard deviation of the original distribution

N = the sample size

2.3.3 Probable Error

Probable Error is nothing but a range within one error on either side of the mean will include 50% of the data values. This is 0.67456

Probable Error defines the half range of an interval about a central point for the distribution such that half of the values from the distribution will lies within the interval and half outside.

Probable Error is basically the correlation co-efficient hat is fully responsible for the value of the co-efficient and its accuracy. In other words, the Probable Error (P.E) is the value which is added or subtracted from the co-efficient of correlation (r) to get the upper limit and lower limit respectively, within which the value of correlation expectedly lies.

$$PE_r = 0.6745\frac{1 - r^2}{\sqrt{N}}$$

Where, r =co-efficient of correlation

N = No. of observation.

$$PE_M = 0.6745 SE_M$$

Where, $SE_M = \frac{\sigma}{\sqrt{N-1}}$ $\qquad$ $PE_\sigma = 0.6745 \frac{\sigma}{\sqrt{N-1}}$

Check Your Progress

Q.1 what do you mean by standard error ?

Q.2 what is the formula for calculating standard error?

2.4 Significance of the sample mean and other statistics

The representative values like mean, median, standard deviation, etc. calculated from the samples are called statistics and those directly computed from the population are named as parameters. The statistics computed from the samples may be used to draw inferences and estimates about the parameters. In an ideal situation, we expect any sample statistic to give a true estimate of the population statistic. The degree to which a sample mean (or other statistics) represents its parameter, is an indication of the significance of the computed sample mean.

Therefore, in a situation in which we approach the element of a representative sample rather than the element of the entire population and compute the sample mean and other statistics for the estimation of related parameters, we have to make sure of the significance or trustworthiness of the computed sample statistics. In other words, we have to say how far we can rely on the computed mean or some other statistic, to predict or estimate the value of the related parameter (the true mean).

2.5 Standard Error in Computing the Significant Value of the Mean

Assume that we have knowledge of the true mean (mean of the population). Also suppose that we have taken 100 representative samples from the population and computed their respective sample means. If we analyze the distribution of these sample means, we will come to know that a majority of these sample means are clustered around the population mean and in the case of a large sample (number of cases more than 30), the distribution will be found normal as shown in Figure 2.1. The mean of the sample will be a fairly good or a somewhat true estimate of the population mean. Some of these sample means will deviate from the population mean either on the positive or the negative side, while most of them will show negligibly small deviations.

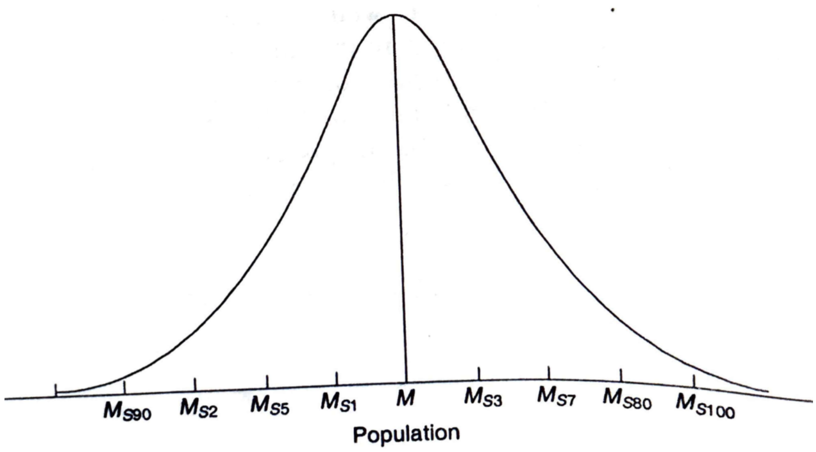


Figure 2.1 Normal distribution of the means of the samples.

The measure of dispersion or deviation of these sample mean scores, may also be calculated in the form of standard deviation. In an ideal case, a sample mean must represent the population mean (or true mean) but on application, a sample mean is likely to s likely to differ from its parameter or true value. This difference is known as the possible error that may occur in estimating the true mean from a given sample mean. The curve representing the distribution of sample means possessing lesser or greater error of estimation for the population mean is called the curve of the error and the standard deviation of this distribution of sample means is known as the standard error of the mean.

In a large sample, the standard error of the mean may be computed with the help of the following formula:

Standard Error of the Mean,

$$SE_M \quad or\ \sigma_M = \frac{\sigma}{\sqrt{N}}$$

where N stands for total No. of cases in the sample and σ for standard deviation of the distribution of the sample means.

In the true sense, must represent the standard deviation of the population. But as we seldom know this value, the SD of the given sample, the significance of whose mean we have to estimate, may be used in this formula.

2.6 Summing Up

Let us conclude with these two terms. The terms "standard error" and "standard deviation" are often confused. The contrast between these two terms one need to understand and appreciate them.

The standard deviation **(SD)** is a *measure of variability*. When we calculate the sample mean we are usually interested not in the mean of this particular sample, but interested in the mean of the whole population. If we take-up various samples' means, means will vary from sample to sample; the way this variations occurs is described by the "sampling distribution fluctuation" of these means. We can estimate how much the sample means will vary from the standard deviation of this sampling distribution, which we call it as the standard error **(SE)** estimate of the mean. This measure is the precision of the sample mean.

The standard error of the mean is the **standard deviation of those sample means over all possible samples drawn from the population.**

Secondly, the standard error of the mean can refer to an estimate of that standard deviation, computed from the sample of data being analyzed at the time.

Mathematically, the standard error of the mean formula is given by:

$$SE_M \quad or\ \sigma_M = \frac{\sigma}{\sqrt{N}}$$

$SE_M = \sigma_M$ = standard error of the mean

σ = the standard deviation of the original distribution

N = the sample size

2.7 Questions and Exercises

1. What do you understand by the terms population and sampling? Define the terms statistics and parameter. How do we use statistics for estimating the parameters?
2. Explain the concept of standard error for determining the significance of the mean and other statistics.
3. What do we mean by the term significance of sample mean? Explain in detail.
4. Explain the concept of confidence interval and confidence limits as used in determining the significance of mean and other statistics.

2.8 References and Suggested Readings

Garrett, Henry E. (1984), Statistics in Psychology and Education. Vakils, Feffer and Simons Pvt. Ltd., Bombay,

Bhandarkar, K.M. (2007), Statistics in Education. Neelkamal Publications PVT. LTD., Hyderabad.

Saha, Kaberi. (2023), Statistical Analysis in Social Sciences, Mani Manik Prakash, Guwahati.

Mangal, S.K. (2021), Statistics in Psychology and Education. PHI Learning Private limited, Delhi.

Packiam, S (2023), Statistics Applied in Educational Research, Neelkamal Publications PVT. LTD., Hyderabad.

Goswami, Sadhana. (2013), Statistics in Education. Shanti Prakashan, Guwahati.

Koul, Lokesh, (2021,), Methodology of Educational Research, Vikash Publishing House Private Limited.

*****

UNIT-3

SIGNIFICANCE OF MEAN AND OTHER STATISTICS

Unit Structure:

3.1 Introduction
3.2 Objectives
3.3 Significance of Mean for Large Samples
3.4 Significance of Mean for Small Samples
3.5 Process of Determining the Significance of Small Sample Means
3.6 Significance of some other statistics
3.7 Summing Up
3.8 Questions and Exercises
3.9 References and Suggested Readings

3.1 Introduction:

In education and psychology, we have to compute mean and several other representative values for studying the characteristics of a certain population. But it is neither feasible nor practicable to approach each and every element of the population. For example, if we wish to know about the average height of the Indian male of 25 years of age, then, we have to catch hold of all the young men of 25 years of age (necessarily identified as citizens of India), measure their heights and compute the average. This average will yield a true value of the desired population mean. But to do so is not a simple task. It is quite impracticable as well as inessential to approach every Indian male of 25 years of age. The convenient as well as practical solution lies in estimating the population mean from the sample means. Here, we may take a few representative samples of appropriate size from the randomly selected districts and states in India. Suppose that we have taken 100 such samples. Then, the means of these samples may yield the desired average height of the whole population of the Indian male of 25 years of age. In this case sampling is needed. However, in the present unit we will see the significance of mean and other statistics.

3.2 Objectives:

After going through this unit you will be able to-

- Understand the significance of mean for large samples
- Understand the significance of mean for small samples
- Know the process of determining the significance of small sample means
- Discuss the significance of some other statistics

3.3 Computation of Significance of Mean for Large Samples

Example 3.1: In a psychological test, a sample of 500 college students of Guwahati city is found to possess the mean score of 95.00 and SD of 25. Test the significance of this mean. In other words how far can this mean be trusted to estimate the man of the entire population of the college students studying in the Guwahati city..

Solution. We have seen that means of large samples randomly drawn from the same population produce a normal distribution. In the present case a sample of 500 students is quite a large sample. Therefore, the standard error of the mean may be computed by the formula

$$SE_M \ or \ \sigma_M = \frac{\sigma}{\sqrt{N}} = \frac{25}{\sqrt{500}} = \frac{25}{\sqrt{5 \times 10 \times 10}} = \frac{25}{10\sqrt{5}} = 1.12$$

Significance of the given mean at 5% and 1% levels of confidence may be given through the following confidence limits.

At 5% level of confidence

$M \pm 1.96\,\sigma \quad = 95 \pm 1.96 \times 1.12$

$= 95 \pm 2.2$

$= 92.8$ to 97.2

It means that there are only 5 chances out of 100 that the population mean or true mean will lie beyond the limit 92.8-97.2.

At 1% level of confidence

$M \pm 2.58\sigma_M \quad = 95 \pm 2.58 \times 1.12$

$= 95 \pm 2.89$

$= 92.11$ to 97.89

It means that there is only 1 chance out of 100 that the true mean will lie beyond the limit 92.11-97.89.

3.4 Significance of Mean for Small Samples

Concept of t distribution. We usually (however, quite arbitrarily) take a sample of 30 cases or more as a large sample and a sample of less than 30 as small. We have seen that the means of large samples randomly drawn from the same population produce a normal distribution (Figure 2.1). However, when samples are less than 30 in size, the distribution does not take the shape of a normal curve. It is symmetrical but leptokurtic. In other words, it becomes more peaked in the middle and has relatively more area in its tails. Such a distribution is known as the / distribution. It was developed originally in 1908 by W.S. Gosset, who wrote under the pen name

"Student". I distribution is sometimes called "Student" distribution in honour of its discoverer.

It should be clearly understood that / distribution is not a single curve (as in the case of a perfect normal curve) but a whole family of curves, the shape of each being a function of sample size. As the size of the sample increases and includes about 30 cases, the distribution takes the shape of a normal curve.

The nature of the distribution is also linked to the number of degrees of freedom. A different / distribution exists for each number of degrees of freedom. As the number of degrees of freedom increases, the distribution takes the shape of a normal curve. This fact may be well illustrated through Table C of the / distribution given in the Appendix. In this table, we find the values of 1 in relation to the given degrees of freedom and level of confidence. We observe that, as the number of degrees of freedom increases and approaches infinity, t approaches the values 1.96 and 2.58, respectively, at 5% and 1% levels of confidence (it is seen that these are the values of sigma scores at the 5% and 1% levels of confidence in the case of a normal distribution).

Concept of degrees of freedom. The number of independent variables is usually called the number of degrees of freedom. This term has been borrowed from Coordinate Geometry and Mechanics where

the position of a point or a body is specified by a number of independent variables, i.e. the coordinates. Each coordinate corresponds to one degree of freedom of movement. Constraints on the body reduce the number of degrees of freedom.

The concept of degrees of freedom is widely used in dealing with small sample statistics. The symbol df is frequently used to represent the degrees of freedom. The 'freedom' in the term degrees of freedom signifies 'freedom to vary'. Therefore, the degrees of freedom associated with a given sample are determined by the number of observations (the number of values of the variables) that are free to vary. The degrees of freedom are reduced through the constraints or restrictions imposed upon the observations or variables. As a general rule one degree of freedom is lost for each constraint or restriction imposed. The number of degrees of freedom in any case, is given by the formula

No. of degrees of freedom =No. of observations or variables -No. of constraints or restrictions

Let us illustrate it with the help of some examples.

If we have five scores as 12, 10, 7, 6 and 5, the mean is 8 and the deviations of these scores from their mean are 4, 2, 1, 2, and -3 respectively. The sum of these deviations is zero. In consequence, if any four deviations are known, the remaining deviation may be automatically determined. In this way, out of the five deviations, only four (i.e. N 1) are free to vary as the condition that "the sum equal to zero" suddenly imposes restriction upon the independence of the 5th deviant. Thus, we may observe that in a sample for the given set of observations or scores (N), one degree of freedom is lost when we employ the mean of the scores for computing the variance of standard deviation (by calculating deviation of the scores from the mean). Originally there were 5 (N = 5) degrees of freedom in computing the mean because all the observations or scores were quite independent. But as we made use of the mean for computing variance and standard deviation, we lost one degree of freedom.

Let us now take the case of the sampling distribution of means and decide about the degrees of freedom. For the purpose of illustration, let us assume that there are only three samples in our sampling distribution of means. The means of these samples are M1, M_2 and M3.The mean of these means, say M, will be our population mean. We know that the sum of the deviations around the mean of any distribution is zero. Therefore, in this case $(M-M_1) + (M-M_2)+(M-M_3)$ will be equal to zero. We see that up to the point of computation of M, we had M_1, M_2 and M_3 as independent observations, but after computing M. we lose one degree of freedom as the value of M_1, M_2 and M_3 is automatically fixed in the light of the value of M. In this way, we may find that when we employ the sample mean as an estimate of the parameter (population mean), one degree of freedom is lost and we are left with N - 1 degrees of freedom for estimating the population variance and standard deviation.

However, the degrees of freedom are not always N - 1 in all cases.It varies with the nature of the problem and the restriction imposed. For example, in the case of correlation between two variables where we need to compute deviations from two means, the number of restrictions imposed goes up to two and consequently, the number of degrees of freedom becomes N - 2 Similarly (as we would see in Chapters 11 and 12 of this text), in the case of contingency tables of Chi square test and analysis of variance, we may have some other different formulae for the computation of number of degrees of freedom. However, in all cases, we will always have a common feature conveying that the number of observations or values in a given data minus the number of restrictions imposed upon this data, constitute the number of degrees of freedom for that data.

Check Your Progress

Q.1 what is the concept of t distribution?

Q.2. what is the concept of degrees of freedom?

3.5 Process of Determining the Significance of Small Sample Means

It may be summarized under the following steps:

Determining the standard error of the mean. In small samples (N less than 30), the standard error of the mean is computed by the formula:

$$SE_M \ or \ S_M = \frac{S}{\sqrt{N}}$$

where N is the size of the sample and s is the standard deviation.

Here the value of s is calculated using the formula

$$s = \sqrt{\frac{\Sigma x^2}{N-1}}$$

(where x = deviation scores from the mean) instead of the formula

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

used for computing SD in large samples.

Use of t distribution in place of normal distribution. As we discussed earlier, the distribution of sample means in the case of small samples takes the shape of a t distribution instead of a normal distribution. Therefore we try to make use of the / distribution.

Determining the degrees of freedom. As we know, the shape of a r distribution depends upon the number of degrees of freedom. The numbers of degrees of freedom thus have to be determined by using the formula

$$df = N - 1$$

where N stands for the number of cases in the samples.

STOP TO CONSIDER

In small samples (N less than 30), the standard error of the mean is computed by the formula:

$$SE_M \ or \ S_M = \frac{S}{\sqrt{N}}$$

Using the t distribution table. Like the normal curve table, / distribution table is also available. Table C given in the appendix represents such a table. From this table, we can read the value of t for the given degrees of freedom at the specific level of probability (5% or 1%) decided at the beginning of the experiment or collection of data.

Determining the confidence interval. After locating the values of t at the 5% and 1% levels of confidence with known degrees of freedom, we may proceed to determine the range or confidence interval for the population mean, indicated as follows:

1. From the first value of t, say t_1 read from the table, we can say that 95% of the sample means lie between the limits $M \pm t_1 \times$ Standard error of the mean and the remaining 5% fall beyond these limits.

2. From the second value of t, say 12, read from the table, we can say that 99% of the sample means lie between the limits $M \pm t_2 \times$ Standard error of the mean and that only 1% fall these limits.

Let us illustrate the above process through an example.

Example 3.2: In a particular test there were 16 independent observations of a certain magnitude with a mean of 100 and SD of 24. Find out (at both 0.05 and 0.01 levels of confidence) the limits of the confidence interval for the population (or true) mean.

Solution. The sample consists of 16 observations and hence may be regarded as a small sample. Therefore,

$$SE_M \text{ or } s_M = \frac{s}{\sqrt{N}} = \frac{24}{\sqrt{16}} = \frac{24}{4} = 6$$

Number of degrees of freedom = N - 1 = 16 - 1 = 15 From Table C. we read the value of t for 15 degrees of freedom at the points 0.05 and 0.01 (5% and 1% levels of confidence). These values are 2.13 and 2.95. Therefore, the limits of confidence interval at the 0.05 level is

$$M \pm 2.13 \times SE_M = 100 \pm 2.13 \times 6$$
$$= 100 \pm 12.78$$

or from 87.22 to 112.78

The limits of confidence interval at .01 level are

$$M \pm 2.95 \times SE_M = 100 \pm 2.95 \times 6$$
$$= 100 \pm 17.70$$

or from 82.30 to 117.70

3.6 Significance of Some Other Statistics

Like mean, the significance of some other statistics like median, standard deviation, quartile deviation, percentages, correlation coefficient, etc. may also be determined by computing the standard errors of these estimates. For large samples (N > 30) the respective formulae are given below:

Standard Error of a Median

$\sigma_{Md} = \frac{1.253\sigma}{\sqrt{N}}$ (in terms of σ)

$\sigma_{Md} = \frac{1.858Q}{\sqrt{N}}$ (in terms of Q)

where σand Q represent the standard deviation and quartile deviation respectively of the given sample.

Standard Error of a Quartile Deviation

$SE_Q \ or \ \sigma_Q = \frac{0.786\sigma}{\sqrt{N}}$ (in terms of σ)

$\sigma_Q = \frac{1.17Q}{\sqrt{N}}$ (in terms of Q)

Standard Error of a Standard Deviation

$$SE_\sigma \ or \ \sigma_\sigma = \frac{\sigma}{\sqrt{2N}}$$

Standard Error of the Coefficient of Correlation

$$SE_r \ or \ \sigma_r = \frac{1 - r^2}{\sqrt{N}}$$

$$SE_\rho \ or \ \sigma_\rho = \frac{1.05(1 - \rho)^2}{\sqrt{N}}$$

After determining the standard error of the required statistics with the help of these formulae, we determine their significance at 5% or 1% level of confidence by computing the limits of the respective confidence intervals. Let us illustrate it through the following example.

Example 3.3: The performance on an intelligence test of 225 students of grade X is as follows:

Median =90.8 and SD = 3.5

Determine the confidence limits at the 0.05 and 0.01 levels for estimation of the population median.

Solution. Standard error of the median

$\sigma_{Md} = \frac{1.253\sigma}{\sqrt{N}}$

Here, σ= 3.5 N = 225

Substituting the respective values in the formula, we obtain

$$\sigma_{Md} = \frac{1.253 \times 3.5}{\sqrt{225}} = \frac{4.3855}{15} = 0.292$$

The confidence interval at 5% level of confidence is

$$\sigma_d \pm 1.96\sigma_{Md} = 90.8 \pm 1.96 \times 0.292 = 90.8 \pm 0.572$$

or

from 85.08 to 91.372

The confidence interval at 1% level of confidence is

$$M_d \pm 2.58\sigma_{Md} = 90.8 \pm 2.58 \times 0.292 = 90.8 \pm 0.753$$

or

from 87.27 to 91.553

Coefficient of Correlation r

Case I:When the sample is large (preferably N = 100 or more) and value of close to $\pm$.50. In all such cases the significance of r may be tested through the computation of its standard error by the following formula

Standard error of r,

$$\sigma_r = \frac{1-r^2}{\sqrt{N}}$$

Let us take an example as an illustration.

Example 3.4: Here, r = 0.52 and N = 173 Therefore,

$$\sigma_r = \frac{1-(.52)^2}{\sqrt{173}} = \frac{1-0.2704}{13.15} = \frac{0.7296}{13.15} = 0.055$$

By assuming the sampling distribution of r to be normal we have the following confidence intervals at 5% and 1% level for delimiting our population r.

Interval at 5% level of confidence is

$r \pm 1.96 \times \sigma_r = 0.52 \pm 1.96 \times 0.055 = 0.52 \pm 0.108$

or from 0.412 to 0.628

The interval at 1% level of confidence is

$r \pm 2.58 \times \sigma_r = 0.52 \pm 2.58 \times 0.055 = 0.52 \pm 0.142$

or from 0.378 to 0.662

In this way, we may conclude that the estimated r is at least as large as 0.378 and no larger than 0.662.

Case II:When we have a large sample (N is 30 or greater) and any value of r (whether too high or too low). In all the above cases, we have converted this value into function for testing the significance of a given value r. (This conversion may be done with the help of Table D given in the Appendix.) After this, using the following formula, we find the standard error of z

$$\sigma_z = \frac{1}{\sqrt{N-3}}$$

Subsequently, confidence interval at the 5% and 1% levels of confidence may be found by using formulae z $\pm$1.96 $\times\sigma_z$ and z$\pm$ 2.58$\times\sigma_z$. These values are again converted into with the help of the conversion Table D.

The whole process can be better understood through the following illustration.

Example 3.5: Given r = 0.78 and N = 84, find the confidence interval at 0.05 and 0.01 of confidence for estimating true r.

Solution. Thez value for the given (as read from Table D given in the Appendix) is 1.05.

The standard error of z, ie.

$$\sigma_z = \frac{1}{\sqrt{N-3}} = \frac{1}{\sqrt{84-3}} = \frac{1}{\sqrt{81}} = \frac{1}{9} = 0.11$$

The confidence interval at the 5% level for the true z is

$z \pm 1.96 \times \sigma_z = 1.05 \pm 1.96 \times 0.11 = 1.05 \pm 0.2156$

$= 1.05 \pm 0.216$ (up to three decimals)

or from 0.834 to 1.266, i.e.

from 0.83 to 1.27 (rounded to two decimals)

These values of z are again converted into values with the help ofTable D. Thus, the confidence interval at 5% level for the true r = from0.68 to 0.85.

The confidence interval at 1% level for the true z is

$z \pm 2.58 \times \sigma_z = 1.05 \pm 2.58 \times 0.11 = 1.05 \pm 0.2838 = 1.05 \pm 0.28$ (rounded to two decimals)

or from 0.77 to 1.33

Converting these z values back into values from Table D. We get confidence interval at the 1% level for the true r, as 0.645 to 0.87.

Case III:When we make use of the null hypothesis to test r. This method of testing the significance of r developed by R.A. Fisher by making use of the distribution and testing the significance of an obtained r against the null / hypothesis, shows that the population r is. in fact, zero.

Fisher in his work demonstrated that when the population ris zero, a parameter / can be estimated by the formula,

$$t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}}$$

The values of t for different samples are distributed as in a f distribution However, in actual practice, there is no need to compute the value of (t ratio) from a given r. It can be done with the help of

Table E given in the Appendix. From this table, we can read directly, given the degrees of freedom, the values of that would be significant at 0.05 and 0.01 levels of confidence and then take the appropriatedecision about the significance of the given r This method may be employed in the case of large as well as smallsamples with any obtained value of r. The procedure can be under stood through the following illustration.

Example3.6: Given r = 0.52 and N = 72 test the significance of the given r.

Solution. Let us set up a null hypothesis saying that the population r is in fact zero and see whether it is accepted or rejected in view of the given data.

Here, N = 72. Therefore, the number of degrees of freedom will be N - 2 = 72 - 2 = 70. We read the values of r at 0.05 and 0.01 levels of confidence from Table E (given in the Appendix) for df = 70. These are 0.232 and 0.302 respectively. Therefore, the given value of r, ie. 0.52 (being much higher than 0.232 or 0.302) is highly significant at both 5% and 1% levels. The hypothesis that population is in fact, zero, is rejected and we may take the computed value of r as quite trustworthy and significant.

3.7 Summing Up

In studying the characteristics of a certain population, it is neither feasible nor practical to approach each and every element of the population. Therefore, we usually resort to sampling. A sample of the appropriate size is drawn from the population and the desired statistics, mean, median, SD, and the like are calculated. These sample statistics are then used to draw an estimate about the parameters, ie. the mean, median, SD, correlation coefficient and so on of the whole population. The degree to which a sample statistics represents its parameter is an index of the significance of trustworthiness of the computed sample statistic-mean, median and so on.

Testing the significance or trustworthiness of the computed sample mean and of other statistics requires the computation of standard error of the sample mean and other statistics. Suppose we take a number of samples from the same population and compute means for these samples. The distribution of these sample means in case of large samples, i.e. N >30 will be normal and may be shown by a normal curve. Standard deviation of this distribution will be known as the standard error of the mean.

In the case of a large sample, the standard error of mean may be computed by the formula

$$\sigma_M = \frac{\sigma}{\sqrt{N}}$$

of median by the formula

$$\sigma_{Md} = \frac{1.253\sigma}{\sqrt{N}}$$

of SD by $\sigma_\sigma = \frac{\sigma}{\sqrt{2N}}$

of quartile deviation by $\sigma_Q = \frac{0.786\sigma}{\sqrt{N}}$

and of r by $\sigma_r = \frac{1-r^2}{\sqrt{N}}$

In all these formulae, N represents the total number of cases in the sample and σ the SD of the distribution of the given sample.

In the case of small samples (N<30), standard error of mean maybe computed by the formula

$$SE_M \ or \ s_M = \frac{s}{\sqrt{N}}$$

where s is the SD of the sample which may be computed by using the formula

$$s = \sqrt{\frac{\Sigma x^2}{N-1}}$$

where x represents deviation scores from the mean.

After computing the standard error of the mean (or of other statistics), testing of the significance of the mean (or of other statistic) requires the computation of the limits of confidence intervals fixed on the basis of levels of confidence or significance (degree of confidence or trustworthiness required in a particular situation). 0.05 and 0.01 arethe two levels of confidence or significance that are usually employed.

The values in terms of the scores of the limits M $\pm 1.96\sigma_M$ and M $\pm 2.58\sigma_M$ are called *confidence limits* and the interval they contain is *confidence interval* at the 0.05 and 0.01 levels of confidence respectively.

The limits of the confidence intervals given by the values M $\pm 1.96\sigma_M$ and M $\pm 2.58\sigma_M$tell us how far a sample mean should miss population it may be taken as a trustworthy estimate of the population mean. The same thing may also be said about the trustworthiness or significance of other statistics, median, SD, etc. In this way, we can point out the significance of the mean and of other statistics, first by computing SE and then determining the limits of confidence intervals at 0.05 or 0.01 levels of significance.

In the case of small samples, the distribution of sample means takes the shape of a t distribution, and not normal as in the case of large samples. Here we have to read the values of t from a table of tdistribution, for given degrees of freedom (computed by the formula df = N - 1 in the case of mean, median, SD, etc.) at the specific level confidence 0.05 or 0.01. To determine the limits of confidence interval, we use the formula M $\pm$ t × Standard error of the mean.

Testing the significance of coefficient of correlation when the sample is quite large (N = 100 or more) and value of r close to ±0.05 requires the use of the formula

$$\sigma_r = \frac{1 - r^2}{\sqrt{N}}$$

for computing SE of r. If we have a large sample (N = 30or > 30) and any value of r, too high or too low, use of z function is appropriate. Conversion of r into z or vice versa is made possible through a conversion table, and standard error of z is computed by the formula

$$\sigma_z = \frac{1}{\sqrt{N - 3}}$$

Another suitable method for testing the significance of r is provided by Fisher by making use of the distribution and testing the significance of an obtained against the null hypothesis that the population is, in fact, zero.

3.8 Questions and Exercises

1. How would you proceed to determine the significance of a given sample mean in a large sample? Illustrate with the help of a hypothetical example.

2. How does the procedure for determining the significance of the mean of a small sample differ from the mean of a large sample?

Illustrate with the help of a hypothetical example the process of determining the significance of a small sample mean.

3. Discuss in brief the process of determining the significance of the statistics like median, standard deviation and quartile deviation.

4. Discuss the different methods used to determine the significance of a correlation coefficient between two variables of a given sample.

5. From the following data, compute the standard of mean and establish the confidence intervals for the location of true mean at 05 and 01 levels.

	Sample A	Sample B	Sample C	Sample D	Sample E	Sample F
Mean	40	45	60	50	30	80
SD	4	6	2	8	3	12
N	125	400	16	900	25	626

10. Are the following values of r significantly different from zero?

(a) r = 0.3 for N = 25

(b) r = 0.60 for N = 15

6. Given r =0.48 and N = 25

(a) find the standard error of r and determine the limits of confidence intervals at 0.05 and 0.01 levels.

(b) convert the given r into a z function; σ_z and determine the limits of the confidence intervals 0.05 and 0.01 levels.

(c) by using the concept of null hypothesis, find whether the given r is significant at 0.05 or 0.01 levels.

7. On an attitude scale, the performance of a group of 36 student teachers was recorded as below:

M_d = 30.4, SD = 2.4

How well does this median represent the median of the population from which this sample was drawn?

3.9 References and Suggested Readings

Garrett, Henry E. (1984), Statistics in Psychology and Education. Vakils, Feffer and Simons Pvt. Ltd., Bombay,

Bhandarkar, K.M. (2007), Statistics in Education. Neelkamal Publications PVT. LTD., Hyderabad.

Saha, Kaberi. (2023), Statistical Analysis in Social Sciences, Mani Manik Prakash, Guwahati.

Mangal, S.K. (2021), Statistics in Psychology and Education. PHI Learning Private limited, Delhi.

Packiam, S (2023), Statistics Applied in Educational Research, Neelkamal Publications PVT. LTD., Hyderabad.

Goswami, Sadhana. (2013), Statistics in Education. Shanti Prakashan, Guwahati.

Koul, Lokesh, (2021,), Methodology of Educational Research, Vikash Publishing House Private Limited.

*****

UNIT-4
PARAMATRIC TEST

Unit structure:

4.1 Introduction

4.2 Objectives

4.3 Meaning of Parametric Test

4.3.1Key Features of Parametric Tests

4.3.2Common Parametric Tests

4.3.3Examples of Parametric Tests

4.4 Nature of Parametric Test

4.4.1Key Aspects of the Nature of Parametric Tests

4.4.2 Examples Demonstrating Nature

4.5 Uses of parametric test

4.6 Differences between parametric and non parametric test

4.7 Summing up

4.8 Questions and Exercise

4.9 References and Suggested Readings

4.1 Introduction:

In the realm of statistical analysis, data often tells a story — one of patterns, relationships, and underlying truths. To uncover these insights with precision and confidence, statisticians turn to a powerful set of tools known as **parametric tests**. These tests are rooted in classical statistical theory and are distinguished by their reliance on underlying assumptions about the population from which the data is drawn.

At their core, parametric tests assume that the data follows a certain distribution — most commonly, the **normal distribution** — and that the population parameters, such as the mean and standard deviation, are both known and meaningful. This structured framework allows parametric tests to be highly efficient, especially when these assumptions are met, offering robust conclusions about differences, associations, or effects within and between groups.

Whether it's comparing the means of two samples through a *t-test*, analyzing variance via *ANOVA*, or exploring relationships with *regression analysis*, parametric tests form the backbone of inferential statistics. Their elegance lies in their mathematical rigor and the depth of inference they provide when applied appropriately.

In essence, parametric tests do not merely examine numbers — they interpret them through the lens of probability theory, transforming raw data into evidence-based decisions that can drive research, policy, and innovation forward.

4.2Objectives:

After going through this unit you will be able to-

- Understand the meaning of parametric test
- Know the nature of parametric test
- Describe various uses of parametric test
- Differentiate parametric test and non parametric test

4.3 Meaning of Parametric Test:

A parametric test is a type of statistical test that makes certain assumptions about the parameters of the population distribution from which the sample is drawn. Most importantly, parametric tests assume that the data follow a specific distribution, usually a normal distribution.

A parametric test is a statistical test that is based on assumptions about the parameters (such as the mean and standard deviation) of the population distribution from which the sample data are drawn.

The word "parametric" comes from "parameter," which refers to a value that describes a characteristic of a population (e.g., the population mean μ\muμ, standard deviation σ\sigmaσ).

So, parametric tests use sample data to make inferences about these population parameters, assuming that the data follow a known distribution—usually a normal distribution.

4.3.1Key Features of Parametric Tests:

- **Assume normality**: Data is assumed to come from a population that follows a normal distribution.
- **Require interval or ratio data**: These tests are typically used for data measured on an interval or ratio scale.
- **More powerful**: When the assumptions are met, parametric tests tend to be more powerful than non-parametric tests, meaning they are more likely to detect a true effect.

4.3.2 Common Parametric Tests:

- **t-test**: Compares the means of two groups.
- **ANOVA (Analysis of Variance)**: Compares means among three or more groups.
- **Pearson correlation**: Measures the linear relationship between two variables.
- **Linear regression**: Models the relationship between dependent and independent variables.

STOP TO CONSIDER
1. **Assumptions about distribution**: • Data is normally distributed (bell curve). • Homogeneity of variance (equal variance across groups). • Interval or ratio scale measurement. 2. **Use of population parameters**: • These tests estimate and compare values like **means, variances**, etc. 3. **Greater statistical power**: • When the assumptions are true, parametric tests are more sensitive and accurate than non-parametric tests.

4.3.3 Examples of Parametric Tests:

Test Name	Purpose
t-test	Compare the means of two groups
Z-test	Compare sample and population means
ANOVA	Compare means across multiple groups
Pearson correlation	Measure linear relationship between variables
Linear regression	Predict outcome using one or more predictors

Check Your Progress
Q.1 Write the key features of Parametric Tests. Q.2 Give some examples of parametric test. Q.3 what is the purpose of parametric test?

4.4 Nature of Parametric Test

The nature of parametric tests refers to their fundamental characteristics, behavior, and the conditions under which they operate effectively. Understanding this helps in determining when and why these tests are appropriate.

4.4.1 Key Aspects of the Nature of Parametric Tests:

- **Assumption-Based**

  Parametric tests are assumption-driven. They assume:

  - The population follows a normal distribution.
  - The data has homogeneity of variances (equal variance across groups).
  - Observations are independent of each other.
  - Data is measured on an interval or ratio scale (quantitative).

- **Focus on Population Parameters**

  These tests aim to estimate or test hypotheses about population parameters like:

  - Mean (μ\muμ)
  - Standard deviation (σ\sigmaσ)
  - Proportions (p)

- **High Statistical Power**

  - When assumptions are met, parametric tests have greater statistical power, meaning they are more likely to detect a true effect or difference if one exists.
  - Used for Precise Measurement
  - They are best suited for precise, numerical data rather than categories or ranks.

- **Sensitive to Outliers and Violations**
  - These tests are sensitive to violations of assumptions (e.g., non-normality or unequal variances), which can lead to incorrect conclusions.

4.4.2 Examples Demonstrating Nature:

- A **t-test** assumes normality of data and is used to compare two means.
- An **ANOVA** assumes equal variance and compares more than two means.
- **Linear regression** models the relationship between variables using assumptions of normality and constant variance of residuals.

Stop to Consider

Key Aspects of the Nature of Parametric Tests:

- Assumption-Based
- Focus on Population Parameters
- High Statistical Power
- Sensitive to Outliers and Violations

Check Your Progress

Q.4 Write basic nature of parametric test.

4.5 Uses of Parametric Test:

Parametric tests are widely used in statistical analysis to make inferences about a population based on sample data. They are particularly useful when certain assumptions about the data (like normal distribution and equal variances) are met.

Main Uses of Parametric Tests:

1. Comparing Means

- To determine whether there is a significant difference between group means.
- Example:
    - t-test: Compares means of two groups.
    - ANOVA: Compares means across three or more groups.

2. Testing Hypotheses

- Used to test null and alternative hypotheses about population parameters.
- Example: Testing if the mean income in one city is different from another.

3. Estimating Population Parameters

- Help estimate values like the population mean, variance, or proportion from sample data.
- Example: Using a Z-test to estimate a population proportion.

4. Measuring Relationships

- Used to find the strength and direction of relationships between variables.
- Example:
- Pearson correlation: Measures the linear relationship between two variables.

5. Predictive Analysis

- Used in regression analysis to model and predict values.
- Example:

- Linear regression: Predicts the value of a dependent variable based on one or more independent variables.

6. Quality Control and Experimental Design

- Widely used in scientific experiments, medical research, economics, and engineering to analyze experimental results and control processes.

Stop to Consider

Examples of Common Parametric Tests and Their Uses:

Test Name	Use Case Example
t-test	Comparing average blood pressure in two groups
Z-test	Testing if a sample mean differs from a known value
ANOVA	Comparing test scores across multiple classrooms
Pearson correlation	Checking the relationship between height and weight
Linear regression	Predicting sales based on advertising spend

Self-Asking Questions

Q. 5 Write true or false:

a) Parametric test can be used when we compare two mean scores.

b) For testing hypothesis parametric test cannot be used.

c) Z test is not a parametric test.

Ans: a) …………… b) ……………… c)…………….

4.6 Difference between Parametric and Non-Parametric Tests

Feature	Parametric Test	Non-Parametric Test
1. Assumptions	Assumes underlying distribution (usually normal distribution)	No strict assumptions about the data distribution
2. Data Type	Used for quantitative data (interval or ratio scale)	Can be used for qualitative or ordinal data
3. Use of Parameters	Involves population parameters (e.g., mean, standard deviation)	Does not involve population parameters directly
4. Statistical Power	Generally has more statistical power if assumptions are met	Less powerful, but more flexible when assumptions are violated
5. Examples	t-test, Z-test, ANOVA, Pearson correlation, Linear regression	Mann-Whitney U test, Kruskal-Wallis test, Wilcoxon signed-rank test, Spearman correlation
6. Sensitivity to Outliers	Sensitive to extreme values and outliers	Less affected by outliers
7. Sample Size	Performs best with larger sample sizes	Can be used with small sample sizes
8. Output Interpretation	Provides results like mean differences, confidence intervals	Focuses on ranks or medians, not means

4.7 Summing Up:

A parametric test is a statistical method used to make inferences about population parameters, under the assumption that the data follow a known distribution (typically normal). These tests are powerful but require that certain assumptions are met..

The nature of parametric tests is analytical, assumption-dependent, and powerful. They are designed for quantitative data and focus on drawing inferences about population parameters, assuming the data follows certain statistical conditions.

Parametric tests are used for comparing means, testing hypotheses, measuring relationships, predicting outcomes, and analyzing experimental data, provided the underlying assumptions (like normality) are met.

A parametric test is a statistical technique used to analyze data that meets certain assumptions about the population parameters and the distribution of the data—primarily that the data follows a normal distribution. These tests are designed for quantitative (interval or ratio) data and are more powerful and precise than non-parametric tests when their assumptions are satisfied. However, it is

- **Based on assumptions**: Requires normal distribution, equal variances, and independent observations.
- **Estimates population parameters**: Such as mean, standard deviation, and proportion.
- **Used for**:
  - ✓ Comparing means (e.g., **t-test**, **ANOVA**)
  - ✓ Measuring relationships (e.g., **Pearson correlation**)
  - ✓ Predicting outcomes (e.g., **Linear regression**)
  - ✓ Hypothesis testing (e.g., **Z-test**)

- **More accurate and powerful**: If assumptions are met, results are highly reliable.
- **Not suitable**: For non-normal or ordinal data, or when sample size is small.

Parametric tests are essential tools in statistics for making accurate inferences about populations, provided that the data meets their underlying assumptions. They are widely used in research, business, healthcare, and social sciences due to their efficiency and effectiveness in analyzing numerical data.

- Parametric tests are best for normally distributed, continuous data and are more accurate when their assumptions are met.
- Non-parametric tests are ideal for non-normal, ordinal, or small sample data, and don't rely on strict assumptions.
- Choose parametric tests when assumptions are met, and non-parametric tests when they are not.

4.8 Questions and Exercises:

Q.1 What do you mean by parametric test. State the nature of parametric test with suitable examples.

Q.2 What is the use of parametric test in educational statistics.

Q.3 State the purpose of parametric test in educational statistics.

Q.4 Differentiate parametric and not parametric test with suitable examples.

4.9 References and Suggested Readings:

Garrett, Henry E. (1984), Statistics in Psychology and Education. Vakils, Feffer and Simons Pvt. Ltd., Bombay, Bhandarkar, K.M. (2007), Statistics in Education. Neelkamal Publications PVT. LTD., Hyderabad.

Saha, Kaberi. (2023), Statistical Analysis in Social Sciences, Mani Manik Prakash, Guwahati.

Mangal, S.K. (2021), Statistics in Psychology and Education. PHI Learning Private limited, Delhi.

Packiam, S (2023), Statistics Applied in Educational Research, Neelkamal Publications PVT. LTD., Hyderabad.

Goswami, Sadhana. (2013), Statistics in Education. Shanti Prakashan, Guwahati.

Koul, Lokesh, (2021,), Methodology of Educational Research, Vikash Publishing House Private Limited.

*****

UNIT-5
THE HYPOTHESES

Unit Structure:

5.1 Introduction
5.2 Objectives
5.3 Meaning of Hypothesis
5.4 Types of Hypothesis
5.5 Level of Significance
5.6 One tailed and Two Tailed Test of Significance
5.7 Type I and Type II Error in Testing Hypothesis
5.8 Summing Up
5.9 Questions and Exercises
5.10 References and Suggested Readings

5.1 Introduction:

Researchers select variables to study from a variety of sources Common sense and curiosity have led to a number of surprising research findings. Observation of the world around us is a valuable source of inspiration for empirical research. Theories guide research by organizing facts and generating knowledge. By examining past research, we can identify questions that need to be answered. Based on the idea you want to investigate, a hypothesis is formulated that makes a statement about how variables are related to one another.

Thus, Hypothesis is a powerful tool in the scientific inquiry for finding out truth. Hypothesis makes use of Inductive philosophy which emphasizes on observation and then the logic of Deductive philosophy which emphasizes reasoning. It enables us to relate theory to observation and observation to theory.

Hypothesis is a proposition which is to be tested and is the very subject of the investigations. It is an idea or a suggestion that is based on known facts and is used as a basis for reasoning or further investigation. Ex: I.Q and achievement are positively related.

Basic to good scientific research is a theory which serves as a point of departure for the successful investigation of a problem. As more and more facts relevant to a theory are gathered, tentative generalizations can be made from them. These generalizations are usually referred to as a set of postulates. Deducing from a set of postulates one formulate a hypothesis and tested.

5.2 Objectives:

After going through this unit you will be able to-

- Understand the meaning of hypothesis
- Know various types of hypothesis
- Discuss level of significance
- Differentiate one tailed and two tailed test of significance
- Explain type I and type II error in testing hypothesis

5.3 Definition of Hypothesis:

Dictionary meaning of Hypothesis is a supposition or proposed explanation made on the basis of limited evidence as a starting point for further investigation. In science, a hypothesis is an idea or explanation that test through study and experimentation. Outside science a theory or guess can also be called hypothesis. A hypothesis is something more than a wild guess but less than a well-establishment theory.

Hypothesis may be defined as a possible statement of a potential relationship between two variables. It is also defined as "A Hypothesis is a statement of possible relationship between dependent and independent variable". It is a specific testable prediction about expected study.

According to George A Lend berg(1942), Hypothesis is a tentative generalization, the validity of which remains to be tested in its most elementary stage, the hypothesis may be very much guess, imaginative proportion which becomes the basis for action or investigation.

Hypothesis is generally considered the most important instrument in research. It main function is to suggest new functions and ideas. In social sciences where direct knowledge of population parameters is rare hypothesis testing is offer used for deciding whether couple data supports our purpose.

A Hypothesis often follows a basic format of "If (this happens) then (this will happens)" one way to structure the hypothesis is to describe what will happen to "dependent variable" if it makes change to the "independent variable" the basic format might be "If (these changes are made to a certain independent variable) then observe (a change in a specific dependent variable)".

A few examples

"Students who eat breakfast will perform better on Maths exam than students who do not eat breakfast".

"Students who experience test anxiety prior to an English exam will get higher scores than students who do not experience test anxiety

5.4 Characteristics of Good Hypothesis:

A good hypothesis posse the following certain attributes:

i) Power of Prediction: One of the valuable attribute of a goodhypothesis is to predict for future. It not only clears the present problematic situation but also predict for the future that what would be happened in the coming time.

ii) Close to Observable Things: A hypothesis must have close contactwith observable things. It does not believe on air castles but it is based on observations those things and objects which cannot observe, for that hypothesis cannot be formulated.

iii) Simplicity: A hypothesis should be so dabble to every lay man. PV. Young says "A hypothesis would be simple, if a researcher has more in sight towards the problem." A hypothesis should be as sharp as razor's blade". So, it should be simple without complexity.

iv) Clarity: A hypothesis must be conceptually clear. It should be clear from ambiguous in formations. The terminology used in it must be clear and acceptable to everyone.

v) Testability: A good hypothesis should be tested empirically. It should be stated and formulated after verification and deep observation. Thus testability is the primary feature of hypothesis.

vi) Relevance to Problem: A hypothesis is guidance for the identification and solution of the problem. So, it must be accordance to the problem.

vii) Specific Ability: It should be formulated for a particular and specific in nature. It should not include generalization. If generalization exists, then a hypothesis cannot reach to the correct conclusions.

viii) Relevant to Statistical Techniques: Hypothesis must be relevant to the techniques which are available for testing and analysis. A researcher must know about the workable techniques before finalizing hypothesis.

ix) Fruitful for New Discoveries: It should be able to provide new suggestions and way of knowledge it must create new discoveries of knowledge J.S. Mill says that "Hypothesis is the best source of new knowledge. It creates new ways of discoveries."

x) Consistency and Harmony: Internal harmony and consistency is a major characteristic of good hypothesis. It should be out of contradictions and conflicts. There must be a close relationship between variables which one dependent on other.

Stop to Consider

Hypothesis is a supposition or proposed explanation made on the basis of limited evidence as a starting point for further investigation. In science, a hypothesis is an idea or explanation that test through study and experimentation. Outside science a theory or guess can also be called hypothesis. A hypothesis is something more than a wild guess but less than a well-establishment theory.

5.5 Formulation of Hypothesis:

A hypothesis is a tentative assumption drawn from knowledge and theory which is used as a guide in the investigation of other facts and theory which are not yet known.

The formulation of hypothesis is the one of the most difficult and most crucial steps in the entire scientific process.

It is the central core of the study that directs the selection of the data, the experimental design, the statistical analysis and the conclusions drawn from the study.

Hypothesis may be defined as a possible statement of a potential relationship between two variables. It is also defined as "A Hypothesis is a statement of possible relationship between dependent and independent variable". It is a specific testable prediction about expected study.

In order to form a good hypothesis one should follow these steps.

- Collect as many as observations about a topic or problem.
- Evaluate the observations and look for possible causes of the problem.
- Create a list of possible explanations that it might want to explore.
- After development of some possible hypothesis think of ways to conform or disprove each hypothesis through experimentation.

Check Your Progress
Q.1 What do you mean by hypothesis?
Q.2 Write any two characteristics of a good hypothesis.

5.6 Functions of Hypothesis in the Research Process:

Hypothesis plays an important role any research the following are the some important functions.

- It is a temporary guess, concerning some truth, which enables a researcher to start looking.
- It helps to establish the specifics of what to look for and may offer possible solutions of the problem.
- Each hypothesis may lead to another hypothesis.
- Preliminary hypothesis may become final hypothesis.

- Each hypothesis provides the researcher with a definite statement which may be objectively tested and accepted or rejected, and allows for interpretation of the findings that is related to the original purposes of the study.

5.7 The Hypothesis in the Scientific Method:

In the scientific method, whether it involves research in Psychology and Education. A hypothesis represents what the researches think will happen in an experiment. The scientific method involves the following steps.

1) Forming a question.

ii) Performing background research

iii) Creating a hypothesis.

iv) Designing an experiment.

v) Collecting data.

vi) Analyzing the results.

vii) Drawing Conclusions.

viii) Communicating the results.

5.8 Importance of Hypothesis:

The form of hypotheses to be tested can be very controversial. The full form is prepared mostly by experienced research workers because this form of statement more readily defines the mathematical model to be utilised in the statistical test of the hypotheses.

Hypotheses are important because they are helpful in

1) pinpointing of problems.

ii) determining the relevancy of facts,

iii) indicating research design.

iv) presenting explanation,

v) providing framework of conclusions, and

vi) stimulating further research.

Self Asking Questions

Q.1 What is the functions of hypothesis in the research process?

………………………………………………………………………
………………………………………………………………………
………………………………………………………………………
………………………………………………………………………

Q.2 State the importance of hypothesis in educational research.

………………………………………………………………………
………………………………………………………………………
………………………………………………………………………
………………………………………………………………………

5.9 Types & Forms of Hypothesis

5.9.1 Types of Hypothesis

There are two types of hypotheses:

1. Universal Hypotheses: Universal hypotheses assert that therelationship in question holds for all the variables that are specified, for all times, and at all places.

Ex: For all rats, if they are rewarded for tuning left, then they will turn left in a T maze.

2. Existential Hypotheses: Existential hypotheses assert that the relationship stated in the hypothesis holds for at least one particular case (Existential implies that one exists)

Ex: There is at least one rat, if he is rewarded for turning left, then he will turn left in a T maze.

5.9.2 Forms of Hypothesis:

There are three forms of hypotheses:

1. Declarative Form Hypotheses: Researcher makes a positive statement of what he expects the outcome of the study.

Ex: "There will be a significant difference in the instructional standards of Private schools as compared with Govt. schools."

2. Null Form Hypotheses

Ex: "There will be no significant difference in instructional standards of Private schools as compared with Govt. schools."

3. Question Form Hypotheses

Ex.: "Is there a significant difference in instructional standards of Private schools and Govt. schools?"

These hypotheses can be tested statistically: therefore, they are also termed as statistical hypotheses.

5.10 Statistical Hypothesis:

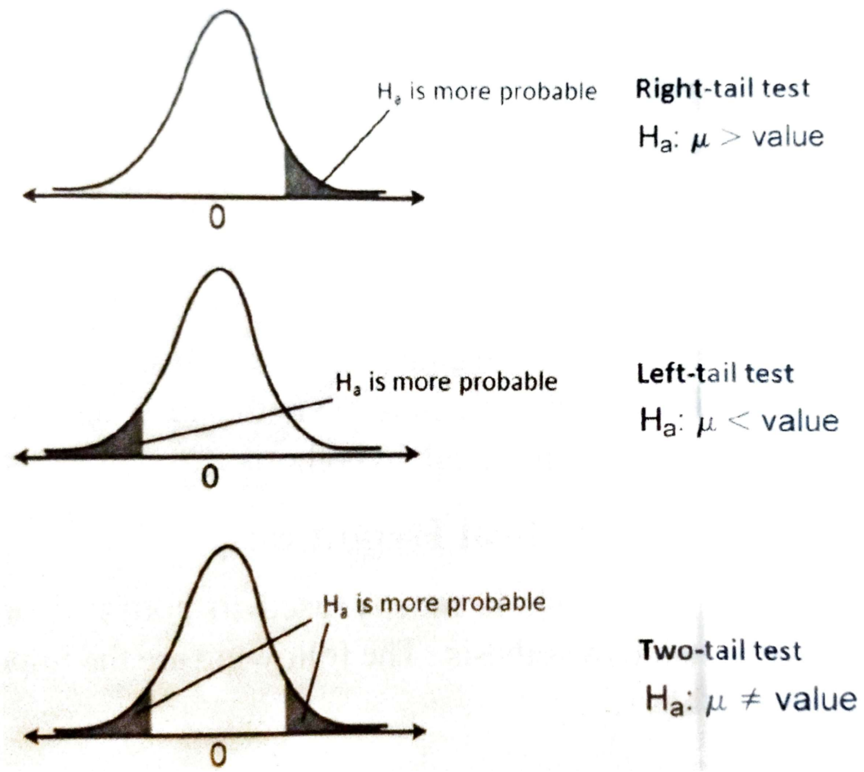


Fig.5.1: Statistical Hypothesis Testing

A Statistical hypothesis, sometimes called confirmatory data analysis is a hypothesis that is testable on the basis of observing a process that is modeled via a set of random variables. A statistical hypothesis is a method of statistical inference.

5.10.1 Definitions of Statistical Hypothesis:

A statistical hypothesis is a hypothesis concerning the parameters or form of the probability distribution for a designated population or more generally probabilistic mechanism which is supposed to generate the observations.

A statistical hypothesis is a formal claim about a state of nature structured within the frame work of statistical model.

Example

One could claim that the median time to failure from (accelerated) electro migration of the chip population.

5.10.2 Types of Statistical Hypothesis:

There are two types of statistical hypothesis (1) Null hypothesis and (ii) Alternative hypothesis.

i) Null Hypothesis (H_o): A Null hypothesis is a type of hypothesis used in statistics that proposes that there is no difference between certain characteristics of a population (or data-gathering process).

In inferential statistics, the null hypothesis is a general statement or default position that there is no relationship between two measured phenomena or no association among groups.

ii) Alternative Hypothesis (H_A): The alternative hypothesis is a position that states something is happening a new theory is preferred instead of a old one. (Null hypothesis-H). It is usually consistent with the research hypothesis. Because it is constructed the literature review, previous studies etc.. However, the research hypothesis is sometimes consistent with the null hypothesis.

5.10.3 Importance of Statistical Hypothesis:

The statistical hypothesis is mostly used to conform or reject the assumption which is called hypothesis. The following are the important areas to use statistical hypothesis.

1) Statistical hypothesis tests are important for qualifying answers to questions about samples of data

The interpretation of a statistical hypothesis test requires a correct understanding of P-Values and critical values.

m) Regardless of the significance level, the finding of hypothesis tests may still contain errors.

5.11 Testing Hypothesis:

Hypothesis testing is an act in statistics where by an analysts tests an assumption regarding a population parameter. The methodology employed by the analyst depends upon nature of the data used and the reason for the analysts. It is used to assess the plausibility of a hypothesis by using sample data. Such data may come from a data generating process. The word 'population' will be used for both of these cases in the following descriptions.

Hypothesis testing is used to assess the plausibility of a hypothesis by suing sample data.

The test provides evidence concerning the plausibility of the hypothesis, given the data.

Statistical analysts test a hypothesis by measuring and examining a random sample of the population being analyzed.

5.12 Methods of Measuring Significance:

Tests of Hypothesis and Significance Rules

If we suppose that a particular hypothesis is true but find that the results observed in a random sample differ markedly from the results expected under the hypothesis (i.e., expected on the basis of pure chance, using sampling theory), then we would say that the observed differences are significant and would thus be inclined to reject the hypothesis (or at least not accept it on the basis of the evidence obtained).

Procedures that enable us to determine whether observed samples differ significantly from the results expected, and thus help us decide whether to accept or reject hypotheses, are called tests of hypotheses, tests of significance, rules of decisions, or simply decision rules.

Steps is Hypothesis Testing

There are five (5) major steps in testing hypothesis:

Step-1: Specifying the Null Hypothesis (Ho).

Step-2: Specifying the alternate Hypothesis (HA).

Step-3: Set the significance level.

Step-4: Calculate the Test Statistic and corresponding value.

Step-5: Drawing conclusion.

Method of sample hypothesis testing is as follows:

The Mean population of IQ is 100 with a S.D of 15. Test the hypothesis at 0.05 level.

Step-1: State the Null Hypothesis. The accepted fact is that the population Mean is 100. So, Ho, M = 100

Step-2: State the Alternative Hypothesis. The accepted mean of the population is 100.

So, H_A, M = 100

Step-3: The significance level is 0.05.

Step-4: Select Z-test and calculate value.

Step-5: Draw the conclusion whether to reject or accept with the value of Z.

5.13 Summing Up:

Lets us conclude the unit here. Hypothesis is a tentative generalization, the validity of which remains to be tested in its most elementary stage, the hypothesis may be very much guess, imaginative proportion which becomes the basis for action or investigation. Hypothesis is generally considered the most important instrument in research. It main function is to suggest new functions and ideas. In social sciences where direct knowledge of population parameters is rare hypothesis testing is offer used for deciding whether couple data supports our purpose.

The formulation of hypothesis is the one of the most difficult and most crucial steps in the entire scientific process. It is the central core of the study that directs the selection of the data, the experimental design, the statistical analysis and the conclusions drawn from the study.

Hypothesis plays an important role any research the following are the some important functions.

- It is a temporary guess, concerning some truth, which enables a researcher to start looking.
- It helps to establish the specifics of what to look for and may offer possible solutions of the problem.
- Each hypothesis may lead to another hypothesis.
- Preliminary hypothesis may become final hypothesis.
- Each hypothesis provides the researcher with a definite statement which may be objectively tested and accepted or rejected, and allows for interpretation of the findings that is related to the original purposes of the study.

However, there are three forms of hypotheses

1. Declarative Form Hypotheses

2. 2. Null Form Hypotheses

3. Question Form Hypotheses

Similarly, hypothesis testing is an act in statistics where by an analysts tests an assumption regarding a population parameter. The methodology employed by the analyst depends upon nature of the data used and the reason for the analysts. It is used to assess the plausibility of a hypothesis by using sample data. Such data may come from a data generating process.

5.14 Questions and Exercises:

I. Short Answer Questions.

1. What is Hypothesis?
2. Explain the steps in formulation of hypothesis.
3. What are the functions of Hypothesis?
4. Explain the Statistical Hypothesis?
5. Explain Null Hypothesis.
6. Give two examples for statistical hypothesis?
7. What is Alternative Hypothesis?
8. Explain the 3 forms of hypothesis.
9. What are the functions of Hypothesis?
10. "Hypothesis plays a vital role in research." Explain.

II. Essay Type Questions.

1. What is Hypothesis? Explain its importance in research?
2. Explain the characteristics of Good Hypothesis.
3. Explain the steps of Scientific method.
4. What are the types of hypothesis? Explain briefly.
5. Discuss the importance of statistical Hypothesis

5.15 References and Suggested Readings:

Garrett, Henry E. (1984), Statistics in Psychology and Education. Vakils, Feffer and Simons Pvt. Ltd., Bombay,

Bhandarkar, K.M. (2007), Statistics in Education. Neelkamal Publications PVT. LTD., Hyderabad.

Saha, Kaberi. (2023), Statistical Analysis in Social Sciences, Mani Manik Prakash, Guwahati.

Mangal, S.K. (2021), Statistics in Psychology and Education. PHI Learning Private limited, Delhi.

Packiam, S (2023), Statistics Applied in Educational Research, Neelkamal Publications PVT. LTD., Hyderabad.

Goswami, Sadhana. (2013), Statistics in Education. Shanti Prakashan, Guwahati.

Koul, Lokesh, (2021,), Methodology of Educational Research, Vikash Publishing House Private Limited.

*****

UNIT-6
SIGNIFICANCE OF THE DIFFERENCE BETWEEN MEANS

Unit structure:

6.1 Introduction
6.2 Objectives
6.3 Computation of Critical Ratio
6.4 Interpretation of Values
6.5 Verification of Null Hypothesis
6.6 Method of Testing the Hypothesis in Educational Research
6.7 Meaning of Level of Significance
6.8 Summing Up
6.9 Questions and Exercises
6.10 References and Suggested Readings

6.1 Introduction:

In order to assess the significance of difference between/among means of various groups, Critical Ratio (C.R.), t-Test and Analysis of Variance (F-test) are used. In educational research, the appropriate statistical technique is selected on the basis of sample and needs/objectives of the research.

When there are two groups of sample, which are large in size, then the technique of C.R. is useful. If there are two groups, which are small in size, (e.g. N = 30 or less) then it is better to use t-test. Here the most important point to be kept in mind is that the method for calculation of the value of C.R. or t is the same and their values are also same for a given example. But while verifying the significance of the value of CR it is not essential to take into account the value of degrees of freedom (df) while for verification of the significance of t-value it is absolutely essential to take into account the value of df. In other words if the size of sample is large then the index of significance is denoted by CR and if the size of sample is small then the same is denoted by t.

If a researcher is required to find out the significance of difference of means of more than two groups then the technique of analysis of variance (F. ratio) is extremely useful. F-test can also be employed to find out the significance of difference between the Means of the two groups. But since the method of calculation of F is complicated, it may be used in case of need only (i.e. when there are more than two groups.)

6.2 Objectives:

After going through this unit you will be able to-

- Compute critical ratio or t value
- Interpret t Values
- Verify null hypothesis
- Know the methods of testing the hypothesis in educational research
- Understand the meaning of Level of Significance

6.3Computation of Critical Ratio

Formula for computation of CR or t-value –

$t = \frac{D}{\sigma D}$………………..Formula (44)

D $=M_1 \sim M_2$…………Formula (45)

where,

D = difference between the means of two groups.

M_1 = Mean of the first group.

M_2 = Mean of the second group.

$$\sigma D = \sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}} \text{ ... ...Formula (46)}$$

where,

σD = Standard error

σ_1 = SD of the first group

N_1 = No. of scores in the first group.

σ_2 = SD of the second group

N_2 = No. of scores in the second group.

By combining formulae 45 and 46 we get –

$$t = \frac{M_1 \sim M_2}{\sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}}} \text{ ....Formula (47)}$$

When the groups are small then t is computed with the help of the following method -

There are 10 students each in group A and B. A test of 50 marks was administered on both the groups. Scores of the students are given below. By applying t-test find out whether groups A and B differ significantly in their achievement.

Group A score	x_1	x_1^2	Group B Score	x_2	x_2^2
18	4	16	24	7	49
34	12	144	38	7	49
22	0	0	35	4	16
15	7	49	19	12	144

17	5	25	41	10	100
32	10	100	33	2	4
33	11	121	22	9	81
23	1	1	38	7	49
15	7	49	42	11	121
11	11	121	18	13	169
$\Sigma X = 220$ $M_1 = 22$			$\Sigma X = 310$ $M_2 = 31$		

For calculation of t-value let us calculate the value N_1 and N_2

$N_1 = 10 - 1$

$= 9$

$N_2 = 10 - 1$

$= 9$

In order to calculate the value of SD of group B we can apply the following formula.

$$SD = \sqrt{\frac{\Sigma x_1^2 + \Sigma x_2^2}{N_1 + N_2}} \quad \ldots\ldots\ldots\ldots\ldots\ldots Formula\ (48)$$

where,

SD = Standard Deviation (σ)

Σx_1^2 = Total of squares of deviations of scores from the mean of first group

Σx_2^2 = Total of squares of deviations of scores from the mean in second group.

N_1 = No. of scores in the first group.

N_2= No. of scores in the second group.

$$SD = \sqrt{\frac{626 + 782}{18}}$$

$$= \sqrt{\frac{1408}{18}}$$

$$= \sqrt{78.2222}$$

$$= 8.844$$

$$SD = \sqrt{\frac{626 + 782}{18}}$$

$$SE_D = 8.844\sqrt{\frac{10 + 10}{100}}$$

$$= 8.844\sqrt{\frac{20}{100}}$$

$$= 8.844\ \sqrt{.2}$$

$$= 8.844 \times .4472$$

$$= 3.955 \qquad (\therefore D = 31 - 22 = 9)$$

$$t = 2.75 \qquad (\therefore t = \frac{D}{SE_D})$$

6.4 Interpretation of Values:

Research hypothesis can be stated as under -

There is no significant difference between the achievement of students in group A and B.

Table

Sl.No	Group	Mean	Standard deviation	No of Cases	t-value
1	A	$M_1 = 22$	Combined for both the groups 8.844	$N_1 = 10$	2.275
2	B	$M_2 = 31$		$N_2 = 10$	

df = (N - 2) = 18

Table value for df 18 = 2.10 (0.05 level)

To be significant at 0.05 level the value of t (for df 18) must be 2.10 or more (see table in the appendix). The value of t obtained is 2.275, which is more than the table Value; so the null hypothesis is rejected (the hypothesis framed proves to be false).

In other words the students in group A and group B differ significantly in their achievement.

6.5 Verification of Null Hypothesis:

In educational research the null hypotheses are verified with the help of 't' values according to the following procedure -

Example

In section 'A' of Standard IX there are 70 students. For mathematics, the Mean of the class is 60 and the Standard Deviation is 5. In section 'B' there are 65 students. For mathematics the Mean of the class is 55 and the SD is 4. On the basis of this information, find out the significance of difference between the Means of section A and B at 0.01 and 0.05 level of significance.

……………Example (47)

The summary of given information can be stated as under-

Section	Mean	Standard Deviation	No. of Students
A	60 (M_1)	5 (σ_1)	70 (N_1)
B	55(M_2)	4 (σ_2)	65 (N_2)

According to formula-

$$t = \frac{60 - 55}{\sqrt{\frac{5^2}{70} + \frac{4^2}{65}}}$$

$$t = \frac{5}{\sqrt{\frac{25}{70} + \frac{16}{65}}}$$

$$t = \frac{5}{\sqrt{0.357 + 0.246}}$$

$$t = \frac{5}{\sqrt{0.603}}$$

$$t = \frac{5}{0.7766}$$

$$t = 6.438$$

It is essential to find out the value of df in order to verify the significance of the obtained value of CR or t. It must be kept in mind that any distribution consists of columns and rows. One frequency for the column and one for that of row are fixed. Hence for columns we get (C-1) and for rows we get (r - 1) By combining these we get the following formula for calculation of the degrees of freedom

df = (n - 2) …………….Formula (47a)

For the present example we can calculate the value of df as under-

df = (n - 2)

n= $(N_1 + N_2)$

= (70+65)

= 135

df = (135 - 2)

= 133

According to the table of 't' values (pl. refer to table 'T') the value of t to be significant for df 133 at 0.01 level must be 2.62 or more.

The obtained value of t(t = 6.438) is more than the table value and it is significant at 0.01 level of significance. If the value of t is significant at 0.01 level,then it is obviously significant at 0.05 level also (But the reverse is not true).

In view of educational research zero or null hypothesis in case of this example can be framed as under.

There is no significant difference in achievement in mathematics between the students of section A and section B.

Or

The students in Section A and B do not differ significantly in their achievement in mathematics.

For this hypothesis we have already seen that the value of t is 6.438, which is significant at 0.01 level. On the basis of this evidence we can draw the following conclusion -

The value of t obtained is significant at 0.01 level, hence the null hypothesis is rejected; so it can safely be concluded that the students of section A and B differ significantly in their achievement in mathematics.

The most important thing to be kept in mind while testing the hypothesis is that if the obtained value of t is significant at 0.01 level or 0.05 level then the null hypothesis is rejected and if it is not significant then the null hypothesis is accepted.

In educational research, sometimes the positive or negative hypothesis is framed for statistical testing but mostly the null or zero hypotheses is framed, owing to its advantages. In case the hypothesis is positive, it is to be accepted if obtained value of t is significant and the same is to be rejected if t value is insignificant.

In order to clarify the method of testing the hypothesis let us study yet another example -

An aptitude test was administered on two groups of students. The first group consisted of 60 boys and the second group consisted of 50 girls. After analysis of the obtained scores the values of various measures and quantities were found as under.

Group	No. of students	Mean	Standard Deviation
Boys	60	80	5.4
Girls	50	81	6.3

On the basis of this information find out the significance of difference between the Means of these two groups .....Example (48).

In view of the educational research the null hypothesis for example (48) may be stated as under -

There is no significant difference in the aptitude of the boys and girls.

In order to verify the statistical significance, we have to find out the value of t by using the formula -

$$t = \frac{M_1 \sim M_2}{\sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}}}$$

By keeping the values of various quantities in the formula (47) above –

$M_1 = 80$ $\quad$ $M_2 = 81$

$\sigma_1 = 5.4$ $\quad$ $\sigma_2 = 6.3$

$N_1 = 60$ $\quad$ $N_2 = 50$

$$t = \frac{80 - 81}{\sqrt{\frac{(5.4)^2}{60} + \frac{(6.3)^2}{50}}}$$

$$t = \frac{1}{\sqrt{\frac{29.16}{60} + \frac{39.69}{50}}}$$

$$t = \frac{1}{\sqrt{0.486 + 0.7938}}$$

$$t = \frac{1}{\sqrt{1.2798}}$$

$$t = \frac{1}{1.1313}$$

$$t = 0.8839$$

$$df = (n - 2)$$

$$= (110 - 2)$$

$$= 108$$

Interpretation

To be significant at 0.05 level the value of t for df 108 must be 1.98 or more (Pl. see the table of t values).

For given example, the value of t obtained is 0.8839, which is not significant at 0.05 level. Hence, the null hypothesis is accepted. In other words, there is no significant difference between the aptitude of the boys and girls.

The value of t which is insignificant at 0.05 level is insignificant at 0.01 level also, since at this level thevalue required for significance is always higher than that of the value required at 0.05 level.

6.6 Method of Testing the Hypothesis in Educational Research:

1. First of all the null hypothesis related with educational research should be framed and stated in an appropriate manner.

2. Then the concerned information should be given in tabular form (e.g. groups, no. of students in each group, Mean, SD etc.)

3. The obtained value of t should be mentioned.

4. Degrees of freedom (df) should be calculated by using the formula df = (n-2).

5. On the basis of the df the significance of t value should be verified first at 0.01 level and if it is not significant at this level, then at 0.05 level.

While testing the significance the table of t values should be used carefully. The significance of the t values is always verified on the basis of df. In our sample if df = 108 and if the table shows the value of df 125 after df 100 then df 100 should be taken as the base as it is more close to 108 as compared to 125.

The technique of interpolation may be applied in order to get the exact table value corresponding to the given df.

If the value of t is not significant at 0.05 level of significance then it also follows that it is not significant at 0.01 level also and in fact there is no need to write down it as such. By the same logic if the value of t is significant at 0.01 level then there is no need to writedown that it is significant at 0.05 level also. If suppose we consider 0.05 level as B.Sc. & 0.01 level as M.Sc. degree then a person who is not B.Sc. cannot possess M.Sc. degree. On the other hand a person who is M.Sc. will essentially possess B.Sc. degree.

6.7 Meaning of Level of Significance:

Suppose for df 100 the significant value of t is 1.98; then it means that if a Normal Probability Curve is drawn for the 100 scores then 95 scores out of 100 will fall within the area - 1.98 0 to 1.98 σ; hence the level of significance shall be 5% (100-95) which is known as 0.05 level of significance.

Similarly, for df 100 the value of significance of t is 2.63 (Pl. refer table t); then 99% of scores (or frequencies) will fall within the area - 2.63 to + 2.636 of the Normal Probability Curve. Hence the level of significance will be (100-99) 1%. This is called as 0.01 level of significance.

Self-Asking Questions

Q.1 what is the formula for computation of critical ratio?

…………………………………………………………………………

…………………………………………………………………………

…………………………………………………………………………

Q.2 Discuss the methods of testing the hypothesis in educational Research.

…………………………………………………………………………

…………………………………………………………………………

…………………………………………………………………………

…………………………………………………………………………

Q.3 What do you mean by level of significance ?

…………………………………………………………………………
…………………………………………………………………………
…………………………………………………………………………
…………………………………………………………………………

6.8 Summing Up:

In carrying out various studies and experiments in the fields of education, psychology and sociology, we often need to test the significance of the difference between two sample means (independent or correlated) drawn from same or different populations.

The process of determining the significance of difference between two given sample means varies with respect to the largeness or smallness of the sample as well as the relatedness or unrelatedness (independence) of the samples. A sample having 30 or more cases is usually treated as a large sample while a sample containing less than 30 cases is considered a small sample. The samples are called uncorrelated or independent when they are drawn at random from totally different and unrelated groups or when uncorrelated tests are administered to the same sample. However, on a broader outlook, it may be summarized as follows:

1. Establishment of a null hypothesis.

2. Choosing a suitable level of significance, 5% or 1%.

3. Determining the standard error of the difference between means of two samples.

4. Determining standard score values in terms of z (for large samples) and in terms of 1 (for small samples).

5. Determining the critical value of z (for large samples) from the normal curve table and critical value of t (for small samples) from the table for the computed value of degrees of freedom.

6. If the computed value of z ort in the given problem reaches the critical table value of z or 1, then it is to be taken as significant, and consequently the null hypothesis stands rejected. If this falls short of the critical value, the null hypothesis is not rejected.

7. The critical values of z or are different in the case of two-tailed and one-tailed tests.

When we are interested only in knowing the magnitude of the difference between means, a two-tailed test is employed but in case the direction is also needed, then a one-tailed test is used. The critical values to be taken are summarized in the following:

Level of significance	*Two-tailed test*		*One-tailed test*	
	z-values	*t-values*	*z-values*	*t-values*
5% level	1.96	Read under the column P = .05	1.64	Read under the column P = .10
1% level	2.58	Read under the column P = 0.1	2.33	Read under the column P = .02

8. If the null hypothesis is rejected, we say that the difference found in the sample means is trustworthy and real but it is not rejected, we have to conclude that the difference between the means is not real; it may occur by chance or due to sampling fluctuations.

The process of computing standard error of the difference between the means shows in essence the difference on account of the size or nature of the samples. Different formula used in the various situations are as follows:

Case I: Large but independent (uncorrelated) samples

$$SE_D = \sqrt{\sigma_{M_1}^2 + \sigma_{M_2}^2} = \sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}}$$

Case II: Small but independent (uncorrelated) samples

$$SE_D = \sigma\sqrt{\frac{1}{N_1} + \frac{2}{N_2}}$$

where σor pooled SD is given by the formula

$$\sigma = \sqrt{\frac{\Sigma x_1^2 + \Sigma x_2^2}{(N_1 - 1) + (N_2 - 1)}}$$

Here, x_1 and x_2 are given by X_1-M_1 and X_2-M_2, respectively.

Case III: Correlated samples-large and small

1. For the two groups matched in pairs as well as in case if a single group is tested twice (before starting the experiment and after), here we use the formula

$$\sigma_D = \sqrt{\sigma_{M_1}^2 + \sigma_{M_2}^2 - 2r\sigma_{M_1}\sigma_{M_2}}$$

2. For the two groups matched in terms of the group as a whole (i.e. Mean and SD), the formula is

$$\sigma_D = \sqrt{(\sigma_{M_1}^2 + \sigma_{M_2}^2)(1 - r^2)}$$

A special method may be employed for small samples giving original raw scores, to determine / values directly without computing the standard error of the difference between means, as

$$t = \frac{\Sigma D}{\sqrt{\frac{N\Sigma D^2 - (\Sigma D)^2}{N-1}}}$$

where D stands for the difference in scores for the two samples.

6.9 Questions and Exercises:

1. What do you mean by the Critical Ratio (C.R.)? Do the terms Critical Ratio and t-value have the same meaning?

2. What do you mean by the Standard Error? How it is calculated?

3. Explain the meaning of the term level of significance. Bring out the difference between 0.05 and 0.01 level of significance.

4. After administering an achievement test on a sample taken for experimental research the following data were obtained. Calculate the valueof t and test the significance of the obtained value of t at 0.05 level or 0.01 level of significance:

Group	No. of Students	Mean	SD
Pre-test	60	55	4.6
Post-test	60	61	5.2

5. Environment education knowledge test was administered on two groups. The first group consisted of 150 boys and the second group consisted of 140 girls. Information obtained is given in the following table. Find out whether there exists any significant difference between the knowledge level of the boys and girls?

Group	No. of Students	Mean	SD
Boys	150	62.2	7.8
Girls	140	61.5	6.6

Answers

4. t = 6.694 Significant at 0.01 level.
Null hypothesis is to be rejected.
5. t = 0.8268 Not significant at 0.05 level.
Null hypothesis is to be accepted.

6.10 References and Suggested Readings

Garrett, Henry E. (1984), Statistics in Psychology and Education. Vakils, Feffer and Simons Pvt. Ltd., Bombay,

Bhandarkar, K.M. (2007), Statistics in Education. Neelkamal Publications PVT. LTD., Hyderabad.

Saha, Kaberi. (2023), Statistical Analysis in Social Sciences, Mani Manik Prakash, Guwahati.

Mangal, S.K. (2021), Statistics in Psychology and Education. PHI Learning Private limited, Delhi.

Packiam, S (2023), Statistics Applied in Educational Research, Neelkamal Publications PVT. LTD., Hyderabad.

Goswami, Sadhana. (2013), Statistics in Education. Shanti Prakashan, Guwahati.

Koul, Lokesh, (2021,), Methodology of Educational Research, Vikash Publishing House Private Limited

*****

Block- IV: Analysis of Variances

Unit 1: Meaning, Natur and Uses of Analysis of Variance

Unit 2: One Way and Two Way Analysis of Variance

Unit 3: Difference Between One Way and Two Way Analysis of Variance

Unit 4: Steps for Calculating the Analysis of Variance

UNIT-1

MEANING, NATUR AND USES OF ANALYSIS OF VARIANCE

Unit structure:

4.1Introduction

4.2 Objectives

4.3 Meaning of Analysis of Variance

4.4 Nature of Analysis of Variance

4.5 Uses of Analysis of Variance

4.6 Assumptions of ANOVA

4.7 Summing Up

4.8 Questions and Exercises

4.9 References and Suggested Readings

4.10 Answer to check your progress

4.1Introduction

Dear learner, we have studied how critical ratio or t-test is applied to find out the significance of difference between the Means of any two groups. t-test is easy and there are quite a few advantages of it, however, the greatest drawback of t-test is that if the number of groups to be compared is more than two, then the t-test becomes futile. A teacher or researcher many a times needs to compare the performance of more than two groups. For example, if a teacher wants to compare the performance of the students divided into four sections, of class IX and to whom he/she is teaching mathematics, and then in such a case t-test is of no use.

Similarly, if a teacher/researcher has used more than two methods of instruction and wishes to compare the achievement of the groups under experiment in order to assess the effectiveness of the methods then also the t-test proves to be useless. In order to overcome all such problems the famous statistician **R.A. Fisher** has developed the technique of Analysis of Variance. His monumental work received wide recognition and the technique was named as F-ratio after him. By the help of F-ratio we can compare the performance/achievement of any given number of groups. F-ratio is widely used in educational and other types of researches in order to verify the significance of difference among two or more than two groups. F-ratio is the most reliable and useful technique for such comparisons.

If the groups in a given sample are not correlated, then Analysis of Variance (ANOVA) is used. If the groups to be compared are correlated then the Analysis of Co-variance (ANCOVA) is used. If there are two dependent variables then Two-Way ANOVA is used and if there are more than two dependent variables then Multi-Variate Analysis of Variance -MANOVA is used.

4.2 Objectives

After going through this unit you will be able to-

- Understand the meaning of analysis of variance
- Know the nature of analysis of variance
- Discuss the uses of analysis of variance
- Analyse various assumptions of ANOVA

4.3 Meaning of Analysis of Variance

A composite procedure for testing simultaneously the difference between several sample means is known as the *analysis of variance*. It helps us to know whether any of the differences between the means of the given samples are significant. If the answer is 'yes', we examine pairs (with the help of the t test) to see just where the significant differences lie. If the answer is 'no', we do not proceed further.

In such a test, as the name implies, we usually deal with the analysis of the variances. Variance is simply the arithmetic average of the squared deviation from their means. In other words, it is the square of the standard deviation (variance = σ^2). Variance has a quality which makes it especially useful. It has an additive property, which the standard deviation with its square root does not possess. Variance on this account can be added up and broken down into components. Hence, the term 'analysis of variance' deals with the task of analyzing of breaking up the total variance of a large sample or a population consisting of a number of equal groups or sub-samples into two components (two kinds of variances), given as follows:

1. *"Within-groups" variance*. This is the average variance of the members of each group around their respective group means, i.e. the mean value of the scores in a sample (as members of each group may vary among themselves).

2. *"Between-groups" variance.* This represents the variance of group means around the total or grand mean of all groups, i.e. the best estimate of the population mean (as the group means may vary considerably from each other).

Let us make this clearer with the help of the example cited in the beginning of this unit.

In this study, there were 40 boys in all, belonging to four different localities. If we add the IQ scores of these 40 boys and divide the sum by 40, we get the value of the grand or general mean, i.e. the best estimate of the population mean. There are 4 groups (samples) of 10 boys each. The mean of the IQ scores of 10 boys in each group is called the group mean and in this way, there will be 4 group means, which will vary considerably from each other.

Now the question arises, as to how far does an IQ score of a particular boy belonging to a particular sample or group deviate from the grand mean for 40 boys. We may observe that the deviation has two parts, i,e. deviation of the score from the mean of that particular group and the deviation of the mean of the group from the grand mean.

For deriving more useful results, we can use variance as a measure of dispersion (deviation from the mean) in place of some useful measures like standard deviation. Consequently, the variance of an individual's score from the grand mean may be broken into two parts (as pointed out earlier), viz. within-groups variance and between-groups variance. Commenting on the above aspects regarding the meaning and nature of analysis of variance and the subsequent procedure adopted in the analysis of variance technique, Lindquist (1970) writes:

The basic proposition is that from any set of groups of n cases each, we may, on the hypothesis that all groups are random samples from the same population, derive two independent estimates of the population variance, one of which is based on the variance of group means, the other on the average variance of within groups. The test of this hypothesis then consists of determining whether or not the *ratio (F) between these estimates lies below the value in the table for F that corresponds to the selected level of significance.*

In this way, the technique of analysis of variance as a single composite test of significance, for the difference between several group means demands the derivation of two independent estimates of the population variance, one based on variance of group means (between groups variance) and the other on the average variance within the groups (within groups variance.) Ultimately, the comparison of the size of between groups variance and within-groups variance called F-ratio denoted by

$$\frac{\text{between} - \text{groups variance}}{\text{within} - \text{groups variance}}$$

is used as a critical ratio for determining the significance of the difference between group means at a given level of significance.

However, in statistics variance is represented by using the square of deviation and it is denoted by o' (sigma square). F-ratio is a more advanced accurate and better measure than that of Standard Deviation because here the variance can be analyzed and conclusion regarding variables can be drawn thereof.

We can get information regarding the following matters through analysis of variance.

- Variance caused on account of difference between/among means of two or more groups.
- Variance caused due to difference between /among two or more distributions.
- Variance caused because of error in sample and measure.

4.4 Nature of Analysis of Variance

The following are some of the nature of analysis of variance.

- **Parametric test:** ANOVA is a parametric test, meaning it assumes that the data follows a normal distribution.
- **Comparing means:** ANOVA compares the means of many groups to determine if there is a significant difference between them.
- **F-statistic:** ANOVA uses F-statistic to determine the significance of difference between the means.
- **Variability:** ANOVA measures the variability between and within groups.

Stop to Consider

The famous statistician **R.A. Fisher** has developed the technique of Analysis of Variance. His monumental work received wide recognition and the technique was named as F-ratio after him. By the help of F-ratio we can compare the performance/achievement of any given number of groups. F-ratio is widely used in educational and other types of researches in order to verify the significance of difference among two or more than two groups. F-ratio is the most reliable and useful technique for such comparisons.

4.5 Uses of Analysis of Variance

The uses of analysis of variance are as follows:

- **Comparing treatment effects:** ANOVA is used to compare the effects of different treatments or interventions on a response variable.

- **Analysing experimental data:** ANOVA is commonly used to analyse data from experiments, such as comparing the means of different groups.
- **Determining significant differences:** ANOVA is used to determine whether there are significant differences between the means of three or more groups.
- **Identifying interactions:** ANOVA can be used to identify interactions between two or more factors.
- **Validating assumptions:** ANOVA can be used to validate assumptions about the data, such as normality and equal variances.
- **Research studies:** ANOVA is widely used in research studies in comparing means of different groups.
- **Quality control:** ANOVA is used in quality control to compare means of different batches or processes.
- **Social Science:** ANOVA is used in social science to compare means of different groups or populations.

Stop to Consider

In a given sample, if the groups are not correlated, then Analysis of Variance (ANOVA) is used. If the groups to be compared are correlated then the Analysis of Co-variance (ANCOVA) is used. If there are two dependent variables then Two-Way ANOVA is used and if there are more than two dependent variables then Multi-Variate Analysis of Variance -MANOVA is used.

Check Your Progress

Q.1 Who has developed the technique of Analysis of Variance.

Q. 2. Write any two nature of Analysis of Variance.

4.6 Assumptions of ANOVA

The following are the fundamental assumptions for the use of analysis of variance technique:

1. The dependent variable which is measured should be normally distributed in the population.
2. The individuals being observed should be distributed randomly in the groups.
3. Within-groups variances must be approximately equal.
4. The contributions to variance in the total sample must be additive.

Check Your Progress
Q.3. Write any five uses of Analysis of Variance.
Q.4. Write any two assumptions of Analysis of Variance.

4.7 Summing Up

Let us sum up the unit. So far as we have gone through this unit we have understood that the famous statistician **R.A. Fisher** has developed the technique of Analysis of Variance. His monumental work received wide recognition and the technique was named as F-ratio after him. By the help of F-ratio we can compare the performance/achievement of any given number of groups. F-ratio is widely used in educational and other types of researches in order to verify the significance of difference among two or more than two groups. F-ratio is the most reliable and useful technique for such comparisons.

ANOVA is a parametric test. It compares the means of many groups to determine if there is a significant difference between them. It also uses F-statistic to determine the significance of difference between the means. It measures the variability between and within groups. ANOVA can be used while we compare treatment effects, analyse experimental data, determine significant differences, Identify interactions, validate assumptions, research studies, quality control and also while conducting social science research.

Also in another condition i,e. if the groups are not correlated, then also Analysis of Variance (ANOVA) can be used. If the groups to be compared are correlated then the Analysis of Co-variance (ANCOVA) is used. If there are two dependent variables then Two-Way ANOVA is used and if there are more than two dependent variables then Multi-Variate Analysis of Variance -MANOVA is used.

4.8 Questions and Exercises

Q.1. What is analysis of variance? State its nature.

Q.2. Discuss various uses of analysis of variance.

Q.3. What are the fundamental assumptions of ANOVA ?

4.9 Answer to check your progress

Ans.1 :R.A. Fisher has developed the technique of Analysis of Variance.

Ans.2 : The nature of ANOVA are as follows:

(a) ANOVA compares the means of many groups to determine if there is a significant difference between them.

(b) ANOVA measures the variability between and within groups.

Ans.3 : The uses of analysis of variance are as follows:

- ✓ Comparing treatment effects,
- ✓ Analysing experimental data
- ✓ Determining significant differences
- ✓ Identifying interactions and
- ✓ Validating assumptions

Ans.4 : The two fundamental assumptions for the use of analysis of variance techniques are as follows:

- The dependent variable which is measured should be normally distributed in the population.
- Within-groups variances must be approximately equal.

4.10 References and Suggested Readings

Garrett, Henry E. (1984), Statistics in Psychology and Education. Vakils, Feffer and Simons Pvt. Ltd., Bombay,

Bhandarkar, K.M. (2007), Statistics in Education. Neelkamal Publications PVT. LTD., Hyderabad.

Saha, Kaberi. (2023), Statistical Analysis in Social Sciences, Mani Manik Prakash, Guwahati.

Mangal, S.K. (2021), Statistics in Psychology and Education. PHI Learning Private limited, Delhi.

Packiam, S (2023), Statistics Applied in Educational Research, Neelkamal Publications PVT. LTD., Hyderabad.

Goswami, Sadhana. (2013), Statistics in Education. Shanti Prakashan, Guwahati.

Koul, Lokesh, (2021,), Methodology of Educational Research, Vikash Publishing House Private Limited

*****

UNIT-2
ONE WAY AND TWO WAY ANALYSIS OF VARIANCE

Unit structure:

2.1 Introduction
2.2 Objectives
2.3 One way Analysis of Variance
2.4 Two way Analysis of Variance
2.5 Difference between T-test and F-test
2.6 Factor Analysis
2.7 Summing Up
2.8 Questions and Exercises
2.9 References and Suggested Readings

2.1 Introduction:

In the previous unit you have understood the concept and meaning of analysis of variance. It generally helps us to know whether any of the differences between the means of the given samples are significant. If the answer is 'yes', we examine pairs (with the help of the t test) to see just where the significant differences lie. If the answer is 'no', we do not proceed further. Is not it? So, in such a test, we usually deal with the analysis of the variances. Variance is simply the arithmetic average of the squared deviation from their means. In other words, it is the square of the standard deviation (variance = σ^2). Variance has a quality which makes it especially useful. It has an additive property, which the standard deviation with its square root does not possess. Variance on this account can be added up and broken down into components. Hence, the term 'analysis of variance' deals with the task of analyzing of breaking up the total variance of a large sample or a population consisting of a number of equal groups or sub-samples into two components or two types of variances. These are:

1. "Within-groups" variance.
2. "Between-groups" variance.

So, in this unit we will discuss one way and two way analysis of variance.

2.2 Objectives:

After going through this unit you will be able to-

- Clear the concept of analysis of Variance
- Understand what is one way and two way Analysis of Variance
- Differentiate between T-test and F-test
- Discuss about factor analysis

2.3 One way Analysis of Variance

One-way ANOVA is used when there is only one independent variable with several groups or levels or categories, and the normally distributed response or dependent variables are measured, and the means of each group of response or outcome variables are compared.

a) Computation of Variance (F-Ratio)

Variance is computed in the following manner

Total Variance = Variance caused due to difference among the means of two or more groups (SSM)+ Variance caused out of difference in distributions of two or more groups (SSW) + Variance of error (e).

Formula: σ^2 = SSM + SSW + e....... Formula (1)

Let us study the following example in order to understand the technique of ANOVA -

A researcher divided 30 students of class VIII into five groups A. B. C. D and E and taught Physics to these groups by using the following methods.

Group	**Method**
A	Lecture Method
B	Laboratory Method
C	Discussion Method
D	Project Method
E	Self-learning Method

After completion of the teaching process an achievement test of 50 marks was administered on the five groups. Scores of the students are as follows. On the basis of the given information find out whether the groups differ significantly in their achievement? Which is the most effective method for teaching of Physics?

	A	**B**	**C**	**D**	**E**
	16	17	19	34	33
	22	23	32	33	37
	18	19	20	35	35
	19	22	27	32	34
	26	36	35	30	32
	37	33	23	34	33
Total of Scores	138	150	156	198	204
Means	23	25	26	33	34
Total of all scores	**846**				

Combined Mean = 28.29($\frac{846}{30}$)

.......Example (1)

Let us calculate F-ratio in order to find out the significance of difference among Means of the groups.

b) Method of Calculation

Method for calculation of F-ratio is as under

1. First of all the Means of the groups are calculated and then the combined Mean is calculated; for the present example it is 28.2 ($\frac{846}{30}$)

2. Since the SD is calculated from the raw scores and the size of sample may also vary, correction term has also to be calculated: the value of correction term is computed by using the formula –

$C = \frac{\Sigma X^2}{N}$ ……………………. Formula (2)

Where,

C = correction term

ΣX^2 = Square of total of all scores

N = No. of scores

$C = \frac{(846)^2}{30}$

$= \frac{715716}{30}$

= 23857.2

3. In the third step all scores are squared and the total of all squares is obtained.

(Total sum of squares)

Total $(SS_t = 16^2 + 22^2 + 18^2 \ldots\ldots + 32^2 + 33^2) - C$

=(256 + 484 +324.......+1024+1089) − C

=25288 −23857.2

= 1430.8

4. In the fourth step, the sum of squares within conditions (teaching methods) A, B, C, D and E is calculated (SS_M) in the following manner.

$SS_M = (138)^2 + (150)^2 + (156)^2 + (198)^2 + (204)^2, - C$

$= \frac{19044+22500+24336+39204+41616}{6} - C$

$= \frac{146700}{6} - C$

= 24450 – C

= 24450 –23857.2

= 592.8

Sum of squares within conditions

SSw =Total SSt – SSm among means

= SSt – SSm

=1430.8-592.8

= 838

C) Summary of ANOVA

Source of variation	*df*	*Sum of Scores (SS)*	*Mean Squares Variance*	*SD*
Among the means of conditions (Methods)	4	592.8	148.2	5.79

Within conditions	25	838	33.52	$\sqrt{33.52}$
Total	**29**			

$$F = \frac{148.2}{33.52}$$

=4.42

F-ratio for the given example is 4.42. in order to verify the significance of this value we shall refer table F.

d) Interpretation of Results

For interpretation of the obtained value we are required to calculate degrees of freedom (df). In the summary of ANOVA figure in the column 'among the means of conditions' is 4 and it is to be treated as denominator (This value will be always less by 1 than the number of groups. Since the number of groups is 5, this value will be 5-1 i.e. 4.) In the same table, the figure in the column 'within conditions' is 25 and it is to be treated as the denominator. (There are five groups. One value is fixed in each group and 5 values are free 5x5 = 25). Thus value of df for this example is 4/25. In table F, value 4 is to be searched in the vertical columns and the value 25 in the horizontal row. The value, which appears at the crossing point of the said column and row, is the table value significant either at .01 level or .05 level of significance. Interpolation of values is to be done, if required.

For the present example df = 4/25. In table F the table value for df 4/25 at .05 level is 2.78 and the same at .01 level is 4.18. The value of F obtained 4.42 is significant at 0.01 level. Therefore, it can safely be concluded that the five groups differ significantly in their achievement.

Mean achievement of group E is the highest among all groups. The students who learned physics by self-learning method scored better than those of their counterparts; so self-learning is the most effective method for acquiring knowledge in physics.

Stop to Consider

One-way ANOVA is used when there is only one independent variable with several groups or levels or categories, and the normally distributed response or dependent variables are measured, and the means of each group of response or outcome variables are compared.

Self-Asking Questions

Q.1 write the formula of the following:

σ^2=……………………….

C=………………………..

Q.2 what are the methods for calculation of F-ratio?

……………………………………………………………………

……………………………………………………………………

……………………………………………………………………

……………………………………………………………………

Check Your Progress

Instruction: write the following abbreviations.

C =

ΣX^2 =

N =

2.4 Two way Analysis of Variance

When there are two independent variables each with multiple levels and one dependent variable in question the ANOVA becomes two-way. The two-way ANOVA shows the effect of each independent variable on the single response or outcome variables and determines whether there is any interaction effect between the independent variables. Two-way ANOVA has been popularized by Ronald Fisher, 1925, and Frank Yates, 1934. Years later in 2005, Andrew Gelman proposed a different multilevel model approach of ANOVA.

a) Logic of Two-way ANOVA

We now have two factors, and we will test the main effects and interaction.

- **Main effect (Factor A: Rows):** The independent effects of Length

of Exposure on Perceived Ability

- **Main effect (Factor B: Columns):** The independent effects of Type of Exposure on Perceived Ability
- **Interaction:** The joint effects of Length and Type of Exposure on Perceived Ability

Hypotheses

Main-Effects

H_o: All groups have equal means

H_1: At least 1 group is different from the other

Interaction

H_o: There is no interaction between factors A and B. All the mean differences between treatment conditions are explained by the main effects.

H_1: There is an interaction between factors A and B. The mean differences between treatment conditions are not what would be predicted from the overall main effects of the two factors.

Logic Variance Parsing

One-way ANOVA has only one factor to test: SSt= SSw + SSb

Two-way ANOVA has multiple factors to test: SSt= SSw+(SS FactorA +SS FactorB +SS A×B)

Note that in the two-way ANOVA: SSB=(SS FactorA +SS FactorB +SS A×B)

F-Ratio Logic

Omnibus F (same as one-way)

F=MSB –MSW

This will still tell us if one cell is different from another, but we also have to calculate F-tests on Factors A, B, and the A x B

$F = MS_{\text{Factor A}} / MS_w$

$F = MS_{\text{Factor B}} / MS_W$

$F = MS_{A \times B} / MS_W$

b) Method of Computation

Problem-1: The yields per acre of four different plant crops grown on lots treated with three different types of fertilizers are detailed below. Using Mean deviation method, determine at the 0.01 level significance whether there is a difference in yield per acre (i) due to the fertilizers and (ii) due to crops.

	Crop I	Crop II	Crop III	Crop IV
Fertilizer A	4.5	6.4	7.2	6.7
Fertilizer B	8.8	7.8	9.6	7.0
Fertilizer C	5.9	6.8	5.7	5.2

Solution: Tabulate the data in a tabular form so as to compute the row totals, the row means, the column totals, column means, and the grand total and the grand mean.

	Crop I	Crop II	Crop III	Crop Iv	Row Total	Row Mean
Fertilizer A	4.5	6.4	7.2	6.7	24.8	6.2
Fertilizer B	8.8	7.8	9.6	7.0	33.2	8.3
Fertilizer C	5.9	6.8	5.7	5.2	23.6	5.9
Column Total	19.2	21.0	22.5	18.9	Grand Total=81.6	Grand Mean=6.8
Column Mean	6.4	7.0	7.5	6.3		

Step-1: Find the variation of the row means from the grand mean VR

$VR = [(6.2 - 6.8)^2 + (8.3 - 6.8)^2 + (5.9 - 6.8)^2]=13.68$

Step-2: Find the variation of column means from the grand mean V_C

$V_C = [(6.4 - 6.8)^2 + (7.0 - 6.8)^2 + (7.5 - 6, 8)^2 + (6.3 - 6.8)^2]=2.82$

Step-3: Find total variation V.

$V = (4.5-6.8)^2 + (6.4-6.8)^2 + (7.2 - 6.8)^2 + ... + (5.7-6.8)^2 + (5.2-6.8)^2$

$=23.08$

Step-4: Find the random variation V_E

$V_E = V - V_R - V_C$

$= 23.08 - 13.68 - 2.82$

$= 6.58$

Step-5: Now prepare Analysis of Variance table

Analysis of Variance

Variation	*df*	*Mean Square*	*F*
VR =13.68	2	6.84	
$V_C = 2.82$	0.94	0.86	
$V_E = 6.58$	1.097		6.24
V =23.08	11		

Now you see the F-table. At the 0.05 significance level with 2 and 6 degrees of freedom, F 0.95 = 5.14 . Our calculated value 6.24 > 5.14, we can reject the hypothesis that the row means are equal and so we can conclude that at the 0.05 level, there is a significant difference in yield due to fertilizers.

Since the F value corresponding to the differences in column means is less than 1, we can conclude that there is no significant difference in yield due to the crops.

2.5 Difference between T-test and F-test

So far as we have worked out on one way and two way ANOVA, now let us see the difference between T-test and F –test. The difference between t-test and f-test can be drawn clearly on the following grounds:

- A univariate hypothesis test that is applied when the standard deviation is not known and the sample size is small is t-test. On the other hand, a statistical test, which determines the equality of the variances of the two normal datasets, is known as f-test.
- The t-test is based on T-statistic follows Student t-distribution, under the null hypothesis. Conversely, the basis of the f-test is F-statistic follows Snedecor f-distribution, under the null hypothesis.
- The t-test is used to compare the means of two populations. In contrast, f-test is used to compare two population variances.

2.6 Factor Analysis

Factor analysis is a statistical approach that can be used to analyze large number of interrelated variables and to categorize these variables using their common aspects.

The approach involves finding a way of representing correlated variables together to form a new smaller set of derived variables with minimum loss of information. So, it is a type of a data reduction tool and it removes redundancy or duplication from a set of correlated variables.

Also, factors are formed that are relatively independent of one another. But since it require the data to be correlated, so all assumptions that apply to correlation are relevant here.

Main Types

There are two main types of factor analysis. The two main types are:

- **Principal component analysis -** this method provides a unique solution so that the original data can be reconstructed from the results. Thus, this method not only provides a solution but also works the other way round, i.e., provides data from the solution. The solution generated includes as many factors as there are variables.
- **Common factor analysis -** this technique uses an estimate of common difference or variance among the original variables to generate the solution. Due to this, the number of factors will always be less than the number of original factors. So, factor analysis actually refers to common factor analysis.

Main Uses

The main uses of factor analysis can be summarized as given below. It helps us in:

- **Identification of underlying factors-** the aspects common to many variables can be identified and the variables can be clustered into homogeneous sets. Thus, new sets of variables can be created. This allows us to gain insight to categories.
- **Screening of variables -** it helps us to identify groupings so that we can select one variable to represent many.

Example

Let us consider an example to understand the use of factor analysis.

Suppose we want to know whether certain aspects such as "task skills" and "communication skills" attribute to the quality of "leadership" or not. We prepare a questionnaire with 20 items. 10 of them pertaining to task elements and 10 to communication elements.

Before using the questionnaire on the sample we use it on a small group of people, who are like those in the survey. When we analyze the data we try to see if there are really two factors and if those factors represent the aspects of task and communication skills.

In this way, factors can be found to represent variables with similar aspects.

2.7 Summing Up

Let us sum up the unit. In this unit you have learnt that the term 'analysis of variance' deals with the task of analyzing of breaking up the total variance of a large sample or a population consisting of a number of equal groups or sub-samples into two components or two types of variances. These are:

1. "Within-groups" variance.
2. "Between-groups" variance.

Also, in this context we can say that one-way ANOVA is used when there is only one independent variable with several groups or levels or categories, and the normally distributed response or dependent variables are measured, and the means of each group of response or outcome variables are compared. However, Variance is computed in the following manner:

Total Variance = Variance caused due to difference among the means of two or more groups (SSM)+ Variance caused out of difference in distributions of two or more groups (SSW) + Variance of error (e).

Formula: $\sigma^2 = SSM + SSW + e$....... Formula (1)

Similarly, when there are two independent variables each with multiple levels and one dependent variable in question the ANOVA becomes two-way. The two-way ANOVA shows the effect of each independent variable on the single response or outcome variables and determines whether there is any interaction effect between the independent variables. Two-way ANOVA has been popularized by Ronald Fisher, 1925, and Frank Yates, 1934. Years later in 2005, Andrew Gelman proposed a different multilevel model approach of ANOVA.

Moreover, in this unit you have clearly differentiate between T-test and F-test and also able to discuss factor analysis too.

2.8 Questions and Exercises

Short Answer Questions

Q.1 What is Analysis of Variance?

Q.2 Explain the difference between t-test and F-test.

Q.3 What is Factor Analysis ?

Long Answer Questions

Q.1 Why two-way ANOVA is superior to one-way ANOVA?

Q.2 In a learning experiment 10 subjects are assigned at random to each of the 6 groups. Each group performed the same task but under slightly different experimental conditions. Find F ratio. (Ans:0.93)

G.No	**Scores of Subjects**										**Total**
	1	2	3	4	5	6	7	8	9	10	
1	41	40	39	41	39	41	36	35	35	37	384

2	40	36	40	34	34	39	36	34	41	37	371
3	36	33	29	30	45	39	33	32	34	34	345
4	14	38	51	41	36	36	36	32	38	36	358
5	41	35	52	41	34	10	44	26	54	30	367
6	55	36	41	36	48	36	42	42	34	40	410
	Total										2235

2.9 References and Suggested Readings

Garrett, Henry E. (1984), Statistics in Psychology and Education. Vakils, Feffer and Simons Pvt. Ltd., Bombay,

Bhandarkar, K.M. (2007), Statistics in Education.Neelkamal Publications PVT. LTD., Hyderabad.

Saha, Kaberi. (2023), Statistical Analysis in Social Sciences, Mani Manik Prakash, Guwahati.

Mangal, S.K. (2021), Statistics in Psychology and Education. PHI Learning Private limited, Delhi.

Packiam, S (2023), Statistics Applied in Educational Research, Neelkamal Publications PVT. LTD., Hyderabad.

Goswami, Sadhana. (2013), Statistics in Education. Shanti Prakashan, Guwahati.

Koul, Lokesh, (2021,), Methodology of Educational Research, Vikash Publishing House Private Limited.

*****

UNIT-3

DIFFERENCE BETWEEN ONE WAY AND TWO WAY ANALYSIS OF VARIANCE

Unit structure:

3.1 Introduction
3.2 Objectives
3.3 Analysis of Variance
3.4 Differences between one-way and two-way ANOVA
3.5 Differences on the basis of features
3.6 Uses of one-way and two-way ANOVA
3.7 Summing Up
3.8 Questions and Exercises
3.9 References and Suggested Readings

3.1 Introduction:

We have studied in the previous unit that the famous statistician **R.A. Fisher** has developed the technique of Analysis of Variance. His monumental work received wide recognition and the technique was named as F-ratio after him. By the help of F-ratio we can compare the performance/achievement of any given number of groups. F-ratio is widely used in educational and other types of researches in order to verify the significance of difference among two or more than two groups. F-ratio is the most reliable and useful technique for such comparisons.

If the groups in a given sample are not correlated, then Analysis of Variance (ANOVA) is used. If the groups to be compared are correlated then the Analysis of Co-variance (ANCOVA) is used. If there are two dependent variables then Two-Way ANOVA is used and if there are more than two dependent variables then Multi-Variate Analysis of Variance -MANOVA is used.

3.2Objectives:

After going through this unit you will be able to-

- Differentiate between one-way and two-way ANOVA
- Differentiate on the basis of features
- Know the uses of one-way and two-way ANOVA

3.3 Analysis of Variance:

As you know that in statistics variance is represented by using the square of deviation and it is denoted by o' (sigma square). F-ratio is a more advanced accurate and better measure than that of Standard Deviation because here the variance can be analyzed and conclusion regarding variables can be drawn thereof.

We can get information regarding the following matters through analysis of variance.

- Variance caused on account of difference between/among means of two or more groups.
- Variance caused due to difference between /among two or more distributions.
- Variance caused because of error in sample and measure.

There are two types of ANOVA. One way ANOVA and two way ANOVA. But there are some differences between these two on the basis of various parameters. Let us see the key differences between one-way and two-way ANOVA.

3.4 Differences Between One-Way and Two-Way ANOVA:

The key differences between one-way and two-way ANOVA are summarized clearly below.

1. A one-way ANOVA is primarily designed to enable the equality testing between three or more means. A two-way ANOVA is designed to assess the interrelationship of two independent variables on a dependent variable.

2. A one-way ANOVA only involves one factor or independent variable, whereas there are two independent variables in a two-way ANOVA.

3. In a one-way ANOVA, the one factor or independent variable analyzed has three or more categorical groups. A two-way ANOVA instead compares multiple groups of two factors.

4. One-way ANOVA need to satisfy only two principles of design of experiments, i.e. replication and randomization. As opposed to two-way ANOVA, which meets all three principles of design of experiments which are replication, randomization and local control.

3.5 Differences on the Basis of Features

The **main difference** between **one-way ANOVA** and **two-way ANOVA** lies in the **number of independent variables (factors)** and the **type of interactions** being tested.

Here's a breakdown of the differences:

Feature	One-Way ANOVA	Two-Way ANOVA
Number of Independent Variables	1 (single factor)	2 (two factors)
Purpose	Tests the effect of one factor on a dependent variable	Tests the effect of two factors and their interaction on a dependent variable
Interaction Effect	Cannot test for interaction	Can test for interaction between the two factors
Example	Comparing test scores of students by **teaching method** (e.g., traditional, online, hybrid)	Comparing test scores by **teaching method** and **student gender**, including if gender and teaching method interact
Design Complexity	Simpler	More complex
Model Output	One F-statistic (for the factor)	Three F-statistics (factor A, factor B, and A×B interaction)
Interpretation	Focuses on whether group means differ due to one factor	Allows understanding of whether each factor and their combination affect the outcome

Stop to Consider

The **main difference** between **one-way ANOVA** and **two-way ANOVA** lies in the **number of independent variables (factors)** and the **type of interactions** being tested.

Self-Asking Questions

Q.1 What is the purpose of one-way and two-way ANOVA ?

……
……
……
……

Q.2 What is the Interaction effect of one-way and two-way ANOVA?

……
……
……
……

3.6 Uses of one-way and two-way ANOVA:

- Use **one-way ANOVA** when you're analyzing the effect of **one categorical independent variable** on a continuous dependent variable.
- Use **two-way ANOVA** when you have **two categorical independent variables** and want to know:
- The **main effects** of each variable
- Whether there's an **interaction** effect between them

3.7 Summing Up:

Let us sum up the unit in brief. After reading this unit we have once again understood the concept and meaning of analysis of variance. No doubt there are certain differences between one way and two way ANOVA which has mentioned in this unit. Moreover, uses of one way and two way ANOVA is also important in educational statistics. The purpose of one way ANOVA is to tests the effect of one factor on a dependent variable whereas the purpose of two way ANOVA is to tests the effect of two factors and their interaction on a dependent variable. So, on the basis of the concept and theory we shall move forward to the next unit. We shall do some exercise by following various steps in calculating one way and two way ANOVA.

3.8 Questions and Exercises:

Q.1 What do you mean by Analysis of variance ?

Q.2 Differentiate between one-way and two- way ANOVA.

Q.3 Discuss the uses of analysis of variance with examples.

3.9 References and Suggested Readings

Garrett, Henry E. (1984), Statistics in Psychology and Education. Vakils, Feffer and Simons Pvt. Ltd., Bombay,

Bhandarkar, K.M. (2007), Statistics in Education. Neelkamal Publications PVT. LTD., Hyderabad.

Saha, Kaberi. (2023), Statistical Analysis in Social Sciences, Mani Manik Prakash, Guwahati.

Mangal, S.K. (2021), Statistics in Psychology and Education. PHI Learning Private limited, Delhi.

Packiam, S (2023), Statistics Applied in Educational Research, Neelkamal Publications PVT. LTD., Hyderabad.

Goswami, Sadhana. (2013), Statistics in Education. Shanti Prakashan, Guwahati.

Koul, Lokesh, (2021,), Methodology of Educational Research, Vikash Publishing House Private Limited

*****

UNIT-4

STEPS FOR CALCULATING THE ANALYSIS OF VARIANCE

Unit structure:

4.1Introduction

4.2 Objectives

4.3 Steps for Calculating the Analysis of Variance

4.4 Two Way Analysis of Variance

4.5 Summing Up

4.6 Questions and Exercises

4.7 References and Suggested Readings

4.1Introduction:

As we have discussed in the previous units that the deviation of an individual's score belonging to a sample or a group of population from the grand mean can be divided into two parts: (i) deviation of the individual's score from his group mean; and (ii) deviation of the group mean from the grand mean. Consequently, the total variance of the scores of all individuals included in the study may be partitioned into within-group variance and between-groups variance. The formula used for the computation of variance (σ^2) is $\sum x^2 / N$ i.e. sum of the squared deviation from the mean value divided by total frequencies. Hence, by taking N as the common denominator, the total sum of the squared deviation of scores around the grand or general mean for all groups combined can be made equal to the sum of the two partitioned, between-groups and within-groups sum of squares. Mathematically, Total sum of squares (around the general mean)

= between-groups sum of squares + within-groups sum of squares

or

$$S_t^2 = S_b^2 + S_w^2$$

Hence the procedure for the analysis of variance involves the following main tasks:

(i) Computation of total sum of squares S_t^2

(ii) Computation of between-groups sum of squares S_b^2

(iii) Computation of within-groups sum of squares S_w^2

(iv) Computation of F-ratio

(v) Use of t test (if the need for further testing arises).

All these tasks may be carried out in a series of systematic steps. Let us try to understand these steps by adopting the following terminology:

X = Raw score of any individual included in the study (any score entry in the given table)

$\sum X$ = Grand or general mean

$\frac{\sum X}{N}$= Grand Sum

$X_{1,} X_2, \ldots$..denote means of the first group, second group,

$n_{1,} n_{2,} n_{3,}$ = No. of individuals in first, second and third groups

$\frac{\sum X_1}{N}, \frac{\sum X_2}{N},$ denote scores within first group, second group,

N = Total No. of scores or frequencies

= $n_1 + n_2 + n_3 + \ldots$..

4.2 Objectives:

After going through this unit you will be able to-

- Understand the concept of analysis of variance
- Describe the steps for calculating the analysis of variance
- Explain the concept of two-way ANOVA

4.3 Steps for Calculating the Analysis of Variance:

Let us now outline the steps.

Step 1.*Arrangement of the given table and computation of some initial values.* In this step, the following values needed in computation are calculated from the experimental data arranged in proper tabular form:

(1) Sum of squares,$\sum X_1$, $\sum X_2$…..and the grand sum, $\sum X$

(ii) Group means, ,$\frac{\sum X_1}{N}$, $\frac{\sum X_2}{N}$ and the grand mean $\frac{\sum X}{N}$

(iii) Correction term C computed by the formula

$$C = \frac{(\sum X)^2}{N} = \frac{Square\ of\ the\ grand\ sum}{Total\ No\ of\ cases}$$

Step 2. *Arrangement of the given table into squared-form table and calculation of some other values.* The given table is transformed into a squared-form table by squaring the values of cach score given in the original table and then the following values are computed:

(i) $\sum X_1^2, \sum X_2^2 \sum X_3^2$,…..

(ii) $\sum X^2$

Step 3.*Calculation of total sum of squares*. The total sum of squares around the general mean is calculated with the help of the following formula:

$$S_t^2 = \sum X^2 - \text{Correction value (C)}$$

$$= \sum X^2 - \frac{(\sum X)^2}{N}$$

Step 4.*Calculation of between-group sum of squares.* The value of the between-groups sum of squares may be computed with the help of the following formula:

$$S_b^2 = \frac{(\sum X)^2}{n_1} + \frac{(\sum X)^2}{n_2} + \frac{(\sum X)^2}{n_3} + \frac{(\sum X)^2}{n_1} - C$$

$= \frac{(sum\ of\ scores\ in\ group\ I)^2}{No.of\ scores\ in\ group\ I}$

$+ \frac{(sum\ of\ scores\ in\ group\ II)^2}{No.of\ scores\ in\ group\ II}$

+….- C (correction value)

Step 5.*Calculation of within-groups sum of squares.* Between-groups and within-groups sum of squares constitute the total sum of squares. Therefore, after steps 3 and 4, the value of within-groups sum of squares (sum of squares of deviation within each group about their respective group means) may be calculated by subtracting the value of between-groups sum of squares from the total sum of squares. Its formula, therefore, goes thus:

Within groups sum of squares,

$S_w^2 = S_t^2 - S_b^2$

Step 6.*Calculation of the number of degrees of freedom.* All these sums of squares calculated in steps 3-5 possess different degrees of freedom given by

Total sum of squares, (S_t^2)= N - 1

Between-groups sum of squares, (S_b^2) = K - 1

Within-groups sum of squares, (S_w^2) = (N – 1) – (K – 1) = N - K

where N represents the total number of observations, scores or frequencies and K, the number of groups in the research study.

Step 7.*Calculation of F-ratio.* The value of F-ratio furnishes a comprehensive or overall test of significance of the difference between means. For its computation, we have to arrange the data and computation work in the following manner:

Source of various	Sum of Squares	df	Mean square various
Between –groups	S_b^2 $(Computed\ in\ step$ 4	K -1	$\frac{S_b^2}{K-1}$
Within-groups	S_w^2 (Computed I step 5)	N – K	$\frac{S_w^2}{N-K}$

F = $\frac{Mean\ sqaure\ variance\ between-groups}{Mean\ square\ variance\ within-groups}$

Step 8.*Interpretation of F-ratio.* F-ratios are interpreted by the use of the critical value of F-ratios given in Table R of the Appendix. This table has the number of degrees of freedom for the greater mean square variance across the top and the number of degrees of freedom in the smaller mean square variance on the left-hand side. If our computed value of F is equal to or greater than the critical tabled value of F at a given level of significance 0.05 or 0.01, it is assumed to be significant and consequently, we reject the null hypothesis of no difference among these means at that level of significance. However, a significant F does not tell us which of the group means differ significantly; it merely tells us that at least one mean is relatively different from some other. Consequently, there arises a need for further testing to determine which of the differences between means are significant.

In case our computed value of F is less than the critical tabled value of F at a given level of significance, it is taken as non-significant and consequently, the null hypothesis cannot be rejected. Then there means will be significant. In a summarized form the above analysis may be represented as follows :

F→ Significant → Null hypothesis rejected → Need for further testing

F→ Non-significant → Null hypothesis not rejected → No need for further testing

Generally, as and when we get the value of F as less than 1, we straightaway interpret it as non-significant resulting in the non-rejection of the null hypothesis.

Step 9. *Testing differences between means with the t test.* When F is found significant, the need for further testing arises. We take pairs of the group means one by one for testing the significance of differences. The t test provides an adequate procedure for testing the significance when we have means of only two samples or groups at a time for consideration. Therefore, we make use of the t test to test the differences between pairs of means.

As we have seen in previous block, the usual formula for computation of t values is

$$t = \frac{D}{\sigma_D} = \frac{Difference\ between\ two\ means}{Standard\ error\ of\ the\ difference\ between\ the\ means}$$

and σ_D is computed by the formula

$$\sigma_D = \sigma\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$$

Where

σ = Pooled SD of the samples drawn from the same population

n_1, n_2 = Total No. of cases in samples I and II, respectively

In the analysis of variance technique, within-groups means square variance provides us the value ofσ^2, the square root of which can give us the required pooled SD of the samples or groups included in our study.

The degrees of freedom for within-groups sum of squares are given by the formula N –K. with these degrees of freedom we can read the t values from Table C given in the Appendix at the 0.05 and 0.01

levels of significance. If the computed value of t at 0.05 or 0.01 levels, we can reject the null hypothesis at that level of significance.

Proceeding similarly, we take other pairs for testing the difference between means and arrive at conclusions.

Stop to consider

(1) Sum of squares, $\sum X_1$, $\sum X_2$.....and the grand sum, $\sum X$

(ii) Group means, , $\frac{\sum X_1}{N}$, $\frac{\sum X_2}{N}$ and the grand mean $\frac{\sum X}{N}$

(iii) Correction term C computed by the formula

$$C = \frac{(\sum X)^2}{N} = \frac{Square\ of\ the\ grand\ sum}{Total\ No\ of\ cases}$$

Check Your Progress

$S_t^2 =$

$S_b^2 =$

$S_w^2 =$

F =

t =

Let us illustrate now the whole process of using the analysis of variance technique with the help of an example.

Example 4.1: The aim of an experimental study was to determine the effect of three different techniques of training on the learning of a particular skill. Three groups, each consisting of seven students of class IX, assigned randomly, were given training through these different techniques. The scores obtained on a performance test were recorded as follows:

Group I	Group II	Group III
3	4	5
5	5	5
3	3	5
1	4	1
7	9	7
3	5	3
6	5	7

Test the difference between groups by adopting the analysis of variance technique.

Solution.

Step 1. *Original table computation*

Table 4.1 Organization of Data Given in Example 4.1

Rating of the coaching experts

Group I ((X_1)	Group II (X_2)	Group III (X_3)	Total
3	4	5	12
5	5	5	15
3	3	5	11
1	4	1	6
7	9	7	23
3	5	3	11
6	5	7	18
$\sum X_1 = 28$	$\sum X_2 = 35$	$\sum X_3 = 33$	$\sum X = 96$ Grand Total

Here,

$n_1 = n_2 = n_3 = 7$, N = $n_1 + n_2 + n_3 = 21$

Group means = $\frac{\sum X_1}{n_1} = \frac{28}{7} = 4$

$\frac{\sum X_1}{n_2} = \frac{35}{7} = 5$

$\frac{\sum X_3}{n_3} = \frac{33}{7} = 4.71$

Correction term C = $\frac{(\sum X)^2}{N} = \frac{96 \times 96}{21} = \frac{9216}{21} = 438.85$

Step 2. *Squared – table computation*

Table 4.2 Squared Values of the Original Data

X_1^2	X_2^2	X_3^2	Total
9	16	25	-50
25	25	25	75
9	9	25	43
1	16	1	18
49	81	49	179
9	25	9	43
36	25	49	110
$\sum X_1^2 = 138$	$\sum X_2^2 = 197$	$\sum X_3^2 = 183$	$\sum X^2 = 518$

Step 3 *Total sum of squares*

$S_t^2 = \sum X^2 - \frac{(\sum X)^2}{N} = \sum X^2 - C = 518 - 438.85 = 79.15$

Step 4 *Between-groups sum of squares (S_b^2):*

$S_b^2 = \frac{(\sum X_1)^2}{n_1} + \frac{(\sum X_2)^2}{n_2} + \frac{(\sum X_3)^2}{n_3} - C$

$= \frac{28 \times 28}{7} + \frac{35 \times 35}{7} + \frac{33 \times 33}{7} - 438.85$

$= \frac{784 + 1225 + 1089}{7} - 438.85$

$= 442.57 - 438.85 = 3.72$

Step 5. *Within-groups sum of squares* (S_W^2). *This is obtained as*

$S_W^2 = S_t^2 - S_b^2 = 79.15 - 3.72 = 75.43$

Step 6. *Number of degrees of freedom*

For total sum of squares,

(S_t^2) = N -1 = 21- 1 = 20

For between-groups sum of squares,

(S_b^2) = k – 1 = 3 – 1 = 2

For within-groups sum of squares,

S_w^2 = N – K = 21 – 3 = 18

Step 7. *Calculation of F-ratio*

Table 4.3 Computation of Values of Mean Square Variance

Source of variation	Sum of squares	df	Mean square variance
Between- groups	$S_b^2 = 3.72$	2	3.72/2 = 1.86
Within-groups	$S_w^2 = 75.43$	18	75.43/18 = 4.19

$F = \frac{Mean\ square\ variance\ between-groups}{Mean\ square\ variance\ within-groups} = \frac{1.86}{4.19} = 0.444$

Step 8.*Interpretation of F-ratio.* The F-ratio-table (Table given in the Appendix) is referred to for 2 degrees of freedom for smaller mean square variance on the left-hand side, and for 18 degrees of freedom for greater mean square variance across the top. The critical values of F obtained by interpolation are as follows:

Critical value of Critical value of at 0.05 level of significance at 0.01 level of significance F = 19.43 F = 99.44

Our computed value of F (444) is not significant at both the levels of significance arid hence, the null hypothesis cannot be rejected and we may confidently say that the differences between means are not significant and therefore, there is no need for further testing with the help of t test.

Example 4.2: In a study, the effectiveness of the methods of memorization was to be determined. For this purpose, three groups of 10 students, each randomly selected from class VII of a school were taken and each group was made to adopt a particular method of memorization. In the end, the performance was tested. The number of presented below :

Group I	: 12,	10,	11,	11,	8,	10,	7,	9,	10,	6
Group II	: 14,	8,	19,	15,	10,	11,	13,	12,	9,	12
Group III	: 8,	11,	13,	9,	7,	5,	6,	8,	7,	10

Apply the analysis of variance technique for testing the significance of the difference between group means.

Solution.

Step 1. Arrangement of the data into a proper table and initial computation (Table 4.4).

Table 4.4 Organization of Data Given in Example 4.2

Group I (X_1)	Group II (X_2)	Group III (X_3)	Total
12	14	8	34
10	8	11	29
11	19	13	43
11	15	9	35
8	10	7	25
10	11	5	26
7	13	6	26
9	12	8	29
10	9	7	26
6	12	10	28
$\sum X_1 = 94$	$\sum X_2 = 123$	$\sum X_3 = 84$	$\sum X = 301$

Here,

$n_1 = n_2 = n_3 = 10$

$N = n_1 + n_2 + n_3 = 30$

Group means = $\frac{\sum X_1}{n_1} = 9.4, \frac{\sum X_2}{n_2} = 12.3, \frac{\sum X_3}{n_3} = 8.4$

Correction term C = $\frac{(\sum X)^2}{N} = \frac{301 \times 301}{30} = \frac{90.601}{30} = 3020$

Step 2.Squared table computation. This can be obtained as in table 4.5

Table 4.5 Squared Values of the Original Data

X_1^2	X_2^2	X_3^2	Total
144	196	64	404
100	64	121	285

121	361	169	651
121	225	81	427
64	100	49	213
100	121	25	246
49	169	36	254
81	144	64	289
100	81	49	230
36	144	100	280
$\sum X_1^2 = 916$	$\sum X_2^2 = 1605$	$\sum X_3^2 = 758$	$\sum X^2 = 3279$

Step 3. Total sum of squares (S_t^2), This is given by

$S_t^2 = \sum X^2 \frac{(\sum X)^2}{N} = X^2 - C$

= 3279 – 3020 = 259

Step 4. Between-groups sum of squares(S_b^2) This is obtained as

$S_b^2 = \frac{(\sum X_1)^2}{n_1} + \frac{(\sum X_2)^2}{n_2} + \frac{(\sum X_3)^2}{n_3} - C$

$= \frac{94 \times 94}{10} + \frac{123 \times 123}{10} + \frac{84 \times 84}{10} - 3020$

= 883.6 + 1512.9 + 705.6 – 3020

= 3102.1 – 3020 = 82.1

Step 5. Within-groups sum of squares S_w^2 This is given by

$S_w^2 = S_t^2 - S_b^2 = 259 - 82.1 = 176.9$

Step 6. Number of degrees of freedom

Df for S_t^2 = N – 1 = 30-1 = 29

Df for S_b^2 = K – 1 = 3 – 1 = 2

Df for $S_w^2 = N - K = 30 - 3 = 27$

……………………………………

Total 4.6 Computation of the Values of Mean Square Variance

Source of variance	Sum of squares	Df	Mean square variance
Between of variance	$S_b^2 = 82.1$	2	$\frac{82.1}{2} = 41.05$
Within groups	$S_w^2 = 176.9$	27	$\frac{176.9}{27} = 6.55$

$$F = \frac{Mean\ square\ variance\ between\text{–}groups}{Mean\ square\ variance\ within\text{–}groups} = \frac{41.05}{6.55} = 6.27$$

Step 8. Interpretation of F-ratio. The F-ratio table (Table R in the Appendix) is now referred for 2 degrees of freedom for greater mean quare variance and 27 degrees of freedom for smaller mean square 'ariance.

Critical values of F = 3.35 F = 5.49 at 0.05 level of significance

Critical values of F = 5.49 at 0.01 level of significance

The computed value of F. i.e. 6.27, is much higher than both the critical values of F at 0.05 and 0.01 levels of significance. Hence, it should be taken as quite significant. Consequently, we have to reject he null hypothesis. Thus, a significant difference definitely exists between the group means. Let us further test to find out where these differences exist.

Step 9. Application of i test. The group means in our study are as follows:

$M_1 = 9.4$, $M_2 = 12.3$, $M_3 = 8.4$

Let us first test the difference between $M_1$and M_2i.e.

12.3 – 9.4 = 2.9

$$t = \frac{D}{\sigma_D} = \frac{2.9}{\sigma_D}$$

$$\sigma_D = \sigma \left(\frac{1}{n^1} + \frac{1}{n^2}\right)$$

here , the square root of the value of within-groups mean square triance, is

$\sqrt{6.55} = 2.559$

$n_1 = n_2 = 10$

$\sigma_D = 2.559 \left(\frac{1}{10} + \frac{1}{10}\right)$

$= 2.559 \times .2 = 0.5118$

= 0.52 (approx..)

Therefore,

$t = \frac{2.9}{0.52} = 5.57$

and df for within groups sum of squares is

N-K = 27

From the t-distribution table (Table in the Appendix):

Critical values of t = 2.05 at 0.05 level of significance

Critical values of t = 2.77 at 0.01 level of significance

The computed value of t, i.e. 5.57, is much higher than the critical values of t at both these levels. Hence it is to be regarded as quite significant and thereby null hypothesis stands rejected, the result being that the difference between $M_1$and M_2 is quite significant and real.

Now, take M_2 and M_3 Here,

D = M_2 - M_3 = 12.3 - 8.4 = 3.9

$t = \frac{D}{\sigma_D} = \frac{3.9}{0.52} = 7.5$

This computed t value is quite significant at both levels of significance, the result being given by that the difference between M_2 and M_3 is also quite significant and real.

Finally, we take$M_1$and M_3

D = $M_1 - M_3$ = 9.4 - 8.4 = 1

t = $\frac{D}{\sigma_D} = \frac{1}{0.52} = 1.92$

It can be seen that the computed value of t, i.e. 1.92, is much lesser than the critical values 2.05 and 2.77 at 0.05 and 0.01 levels of significance. Hence it is not significant and, consequently, the null hypothesis cannot be rejected. Thus, it can be said that the difference between $M_1$and M_3 is not significant and real. It may occur by chance or due sampling fluctuation.

In conclusion, it can be said that of the total three, only two of the difference, M_1- M_2 and M_2 - M_3 are significance and trustworthy

4.4 Two-Way Analysis of Variance

So far, we have dealt with one-way analysis of variance involving one experimental variable. However, experiments may be conducted in the fields of education and psychology for the simultaneous study of two experimental variables. Such experiments involve two-way classification based on two experimental variables. Let us make some distinction between the need for carrying out one-way and two way analyses of variance through some illustrations.

Suppose that we want to study the effect of four different methods of teaching. Here, the method of teaching is the exper-imental variable (independent variable which is to be applied at four levels). We take four groups of students randomly selected from a class. These four groups are taught by the same teacher in the same school but by different methods. At the end of the session, all the groups are tested through an achievement test by the teacher. The mean scores of these four groups are computed. If we are interested in knowing the significance of the differences between the means of these groups, the best technique is the analysis of variance. Since only one experimental variable (effect of the method of teaching) is to be studied, we have to carry out one-way analysis of variance.

Let us assume, that there is one more experimental or independent variable in the study, in addition to the method of teaching. Let it be the school system at three levels which means that three school systems are chosen for the experiment. These systems can be: government school, government-aided school and public school. Now the experiment will involve the study of 4 * 3 groups. We have 4 groups each in the three types of schools (all randomly selected). The achievement scores of these groups can then be compared by the method of analysis of variance by establishing a null hypothesis that neither the school system nor the methods have anything to do with the achievement of pupils. In this way, we have to simultaneously study the impact of two experimental variables, each having two or more levels, characteristics or classifications and hence we have to carry out the two-way analysis of variance.

In the two-way classification situation, an estimate of population variance, i.e. total variance is supposed to be broken into: (i) variance due to methods alone, (ii) variance due to school alone, and (iii) the residual variance in the groups called *interaction variance* ($M \times S$; M= methods, S= schools) which may exist on account of the following factors:

1. Chance

2. Uncontrolled variables like lack of uniformity in teachers

3. Relative merits of the methods (which may differ from one school to another).

In other words, there may be interaction between methods and schools which means that although no method may be regarded as good or bad in general, yet it may be more suitable (or unsuitable) for a particular school system. Hence, the presence of interaction variance is a unique feature with all the experimental studies involving two or more experimental variables. In such problems, requiring two or more ways of analysis of variance, we will have to take care of the residual or interaction variance in estimating their population variance. If the null hypothesis is true, variance in terms of the method should not be significantly different from the interaction variance. Similarly, the variance due to schools may also be compared with the interaction variance. For all such purposes of comparison, the F-ratio test is used, as explained later in this chapter.

The design of the above two-way classification experiment can be further elaborated by having more experimental or independent variables, each considered at several levels. As the number of variables increases, the order of interaction also increases. For example, if school (S), teacher (T), and methods (M) are taken as three independent variables, there will be 3 first-order interactions: S×T T×M and S×M and the second-order interaction We S×T×M will have 4 × 3 × 4 (if there are 4 levels of teacher i.e. 48 randomized groups in this study and the problem of testing the difference between means will require three way analysis of variance technique.

According to the needs and requirements of the experiment, a research worker has to select a particular experimental design involving two or more independent or experimental variables. A discussion of these designs in terms of their construction, administration and analysis is beyond the scope of the hook. Hence, the readers are advised to consult some of the books given in the reference section like Edwards (1961), Hicks (1964), Lindquist (1968) and Guilford (1973). However, let us briefly study the method of carrying out two-way analysis of variance for testing the difference between group means. As an illustration, let us take the case of an experimental design where a single group is being assessed more than once.

Example-4.3 In a research study there were two experimental or independent variables: a seven-member group of player and three aches who were asked to rate the players in terms of a particular trait 1 a ten-point scale. The data were recorded as under:

Rating by three coaches	Players						
	1	2	3	4	5	6	7
A	3	5	3	1	7	3	6
B	4	5	3	4	9	5	5
C	5	5	5	1	7	3	7

Table 4.7 Squared Values of Original Data

Players	Rating of coaches			Total of rows	Square of the total of rows
	By A (X_1)	By B (X_2)	By C (X_3)		
1	3	4	5	12	144
2	5	5	5	15	225
3	3	3	5	11	121
4	1	4	1	6	36

5	7	9	7	23	529
6	3	5	3	11	121
7	6	5	7	18	324
Total of columns	$\sum X_1$= 28	$\sum X_2$= 35	$\sum X_3$= 33	$\sum X$= 96 (Grand Total)	
Square of the totals of the columns	$(\sum X_1)^2$= 784	$(\sum X_2)^2$= 1225	$(\sum X_3)^2$= 1089	$(\sum X)^2$= 9216	

Here

N = $n_1 + n_2 + n_3$ = 7 + 7 + 7 = 21

C = $\frac{(\sum X)^2}{N} = \frac{9216}{21}$ = 438.85

Step 2. Arrangement of the table in square form and computation of some essential values (Table 4.8)

Table 4.8 Squared Values of Original Data

X_1^2	X_2^2	X_3^2	Total
9	16	25	50
25	25	25	75
9	9	25	43
1	16	1	18
49	81	49	179
9	25	9	43
36	25	49	110
Total $\sum X_1^2$ = 138	$\sum X_2^2$ = 197	$\sum X_3^2$= 183	$\sum X^2$ = 518

Step 3.Calculation of the total sum of squares (S_t^2) (around grand mean).

$S_t^2 = \sum X^2 - \frac{(\sum X)^2}{N} = \sum X^2 - C$

= 518 – 438.85 = 79.15

Step 4. Calculation of the sum of squares for rows(S_t^2)(between the means of players to be rated by three coaches).

$S_t^2 = \frac{144+225+121+36+529+121+}{C}$ – C

= $\frac{1500}{3}$ – 438.85 = 61.15

Step 5.Calculation of the sum of squares for columns (S_t^2) between the means of coaches rating 7 players).

$S_t^2 = \frac{784 \times 1225 \times 108}{7}$ – C

= 442.57 – 438 – 85 = 3.72

Step 6. Calculation of the interaction or residual sum of squares (S_t^2) The interaction sum of squares is given by

$(S_t^2) = S_t^2 - (S_t^2 + S_t^2)$

= 79.15- (61.15 + 3.72) = 14.28

Step 7.Computation of F-ratios.

Table 4.9 illustrates the computation of mean square variance value.

Source of variation	df	Sum of squares	Mean square variance
Row (players)	$(r$-1) = 6	61.5	$\frac{61.15}{6}$ = 10.19
Columns (coaches)	$(r$-1) = 2	3.72	$\frac{3.72}{2}$ = 1.86
Interaction or residual	$(r$-1)$(c$-1) =12	14.28	$\frac{14.28}{12}$ = 1.19

F (for rows) = $\frac{Mean\ square\ variance\ between\ rows\ (players)}{Mean\ square\ variance\ in\ terms\ of\ interaction}$

$= \frac{10.19}{1.19} = 8.56$

F (for columns) $= \frac{Mean\ square\ variance\ between\ columns\ (coaches)}{Mean\ square\ variance\ in\ terms\ of\ interaction}$

$= \frac{1.86}{1.19} = 1.56$

Step 8. Interpretation of F ratios. Table in the appendix gives the critical values F.

Table 4.10 Critical Values of F and Significance of Computed F

Kind of F	df for greater mean square variance	df for smaller mean square variance	Critical values of F		Judgement about the significance of computed f	Conclusion
			at 0.05 level	at 0.01 level		
For rows	2	12	3.88	6.93	Significant at both levels	Null hypothesis rejected
For Columns	6	12	3.00	4.82	Not significant at both levels	Null hypothesis not rejected

Here, F (for rows) is highly significant and hence null hypothesis rejected. It indicates that the coaches did discriminate among the players. The second F (for columns) is insignificant and hence, the null hypothesis cannot be rejected. It indicates that the coaches did not fer significantly among themselves in their ratings of the players. In her words, their ratings may be taken as trustworthy and reliable.

4.5 Summing Up:

In performing experiments and carrying out studies in the fields of education and psychology, we often come across situations where we are required to test the significance of the differences among the means of several samples instead of only two sample means, capable of testing with the help of z or t tests. In such situations, a technique known as analysis of variance proves quite useful. It provides a composite test to find out, simultaneously, the difference between several sample means and thus helps us tell, whether any of the differences between means of the given samples are significant. If the answer is 'yes' we examine pairs of sample means (with the help of the usual test) to see just where the significant differences lie. If the answer is 'no', we do not require further testing.

The technique of analysis of variance demands the analysis or breaking apart of the total squared deviation scores or variance of the scores of all individuals (sigma ^ 2) ^ * included in the study around the grand mean (mean of the total scores) into two parts-(i) within-groups variance, i.e. the squared deviation scores of the individuals from their group means and, (ii) between-groups variance, i.e. the squared deviation scores of the group mean from the grand mean. In actual computation work:

1. The total variance$S_t^2$is first calculated and then the $S_b^2$the difference between the two automatically gives S_w^2The formulac run as follows:

$$S_t^2 = \sum X^2 - \frac{(\sum X)^2}{N}$$

$$S_b^2 = \frac{(\sum X_1)^2}{n_1} + \frac{(\sum X_2)^2}{n_2} + \frac{(\sum X_3)^2}{n_3} + \ldots + \frac{(\sum X_n)^2}{n_n} - C$$

$$S_w^2 = S_t^2 - S_b^2$$

2. The degrees of freedom for St and S are K-1 and NK, where K represents the number of groups and N, total observations.

3. Then, the F-ratios desired for testing the significance is calculated by the formula

$$F = \frac{Mean\ square\ variance\ between{-}group}{Mean\ squa\quad variance\ within{-}groups} = \frac{S_b^2\ /\ K{-}1}{S_w^2/N{-}K}$$

4. The computed F-ratios are then interpreted by the use of critical F values read from the table with the degrees of freedom at the given level of significance. In case our computed F is found non-significant, then there is no need for further testing but if it is found significant then the use of t test is made for testing further the difference between sample means taken two at a time.

In the experiments involving two experimental or independent variables instead of one, the significance of the difference between several means is tested with the help of two-way analysis of variance instead of one-way analysis of variance. In the two-way analysis total variance is supposed to be broken into three parts instead of two as done in the one-way analysis. These are: (i) variance due to one variable, (ii) variance due to other variable, and (iii) interaction or residual variance on account of the supposed interaction between variables. Here, instead of one, two F-ratios are computed with the help of the formulae:

$$F\ (\text{for one variable}) = \frac{Mean\ square\ variance\ between\ one\ varibale}{Mean\ square\ variance\ in\ term\ of\ interaction}$$

$$F\ (\text{for the other variable}) = \frac{Mean\ square\ variance\ between\ othe\quad variable}{Mean\ square\ variance\ in\ terms\ of\ interactio}$$

and then the interpretation is made by comparing these F-ratios with the critical F values read from the table for computed degrees of freedom at a given level of significance.

For its successful application, the analysis of variance needs the fulfilment of assumptions like the normal distribution of the dependent variable, random distribution in the groups of individuals under study, homogeneity of the within-groups variances and the analysis of total variances into component variances.

4.6 Questions and Exercises

1. What do you understand by the technique, analysis of variance, used for the analysis of statistical data?

2. In which circumstances is the technique of analysis of variance preferred to the usual z or tests for testing the significance of the difference between in sans? Discuss with the help of examples.

3. Discuss the various steps involved in the application of analysis of variance for testing the difference between groups. Illustrate with the help of a hypothetical example.

4. What is analysis of variance? Point out the underlying assumptions in its application.

5. The following are error scores on a psychomotor test for four groups of equal subjects tested under four experimental conditions:

Group I	Group II	Group III	Group IV
4	9	2	7
5	10	2	7
1	9	6	4
0	6	5	2
2	6	2	7

Apply the analysis of variance to test the null hypothesis.

6. The following data represent the scores of six students (selected randomly) in each of the five sections of class VIII of a school on a vocabulary test;

Section A	Section B	Section C	Section D	Section E
32	19	17	12	12
17	26	26	15	15
28	30	30	10	36
24	17	35	20	17
21	34	20	18	20
38	15	15	30	25

Test the null hypothesis for vocabulary ability of the different sections of class VIII.

In a research study, a Marital Adjustment Inventory was administered on four random samples of the individual. Test the hypothesis that they are from the same population.

10. The following data represent the marks of eight students by three teachers in terms of their performances on a particular 22 skill:

Sample I	Sample II	Sample III	Sample IV
16	24	16	25
7	6	15	19
19	15	18	16
24	25	19	17
31	32	6	42
	24	13	45
	29	18	

8. The following data were collected from the students of three groups in terms of correctly recalled nonsense syllables under three methods of memorization:

Method I	Method II	Method III
18	20	7
29	17	15
28	21	9
26	19	13
24	29	5
31	26	11
30	25	
27		

Compute the means for separate groups and apply the analysis of variance technique to test the significance of the difference between groups.

9. In a learning experiment, three subjects made the following number of correct responses in a series of 7 trials.

	Trials						
Subjects	1	2	3	4	5	6	7
A	5	9	3	7	9	3	7
B	6	8	4	5	2	4	3
C	5	7	3	5	9	3	7

Apply the technique of two-way analysis to test the difference between means.

10. The following data represent the marks of eight students by three teachers in terms of their performances on a particular skill:

	Marks given by the teachers		
Students	A	B	C
1	8	4	2
2	6	3	6
3	8	4	8
4	10	8	8
5	12	6	8
6	14	10	9
7	16	8	10
8	20	11	8

Apply the technique of analysis of variance to answer the following questions:

(a) Did the teachers discriminate among the students in their markings?

(b) Did the teachers differ among themselves in their markings?

4.7 References and Suggested Readings

(c) Garrett, Henry E. (1984), Statistics in Psychology and Education. Vakils, Feffer and Simons Pvt. Ltd., Bombay,

(d) Bhandarkar, K.M. (2007), Statistics in Education. Neelkamal Publications PVT. LTD., Hyderabad.

(e) Saha, Kaberi. (2023), Statistical Analysis in Social Sciences, Mani Manik Prakash, Guwahati.

(f) Mangal, S.K. (2021), Statistics in Psychology and Education. PHI Learning Private limited, Delhi.

(g) Packiam, S (2023), Statistics Applied in Educational Research, Neelkamal Publications PVT. LTD., Hyderabad.

(h) Goswami, Sadhana. (2013), Statistics in Education. Shanti Prakashan, Guwahati.

(i) Koul, Lokesh, (2021,), Methodology of Educational Research, Vikash Publishing House Private Limited.

*****

Block- V: Chi Square Test

Unit 1 : Non-Parametric Test Introduction to Chi-Square Test

Unit 2 : Assumptions of Chi-Square Tests

Unit 3 : Uses and Significance of Chi-Square Test

Unit 4 : Testing Null Hypothesis of Independence in 2×2 Contingency Table

UNIT-1

NON-PARAMETRIC TEST

INTRODUCTION TO CHI-SQUARE TEST

Unit Structure:

1.1 Introduction
1.2 Objectives
1.3 Introduction: Non-Parametric Test
    1.3.1 Meaning of Non-Parametric Test
    1.3.2 Key Features of Non-Parametric Tests
    1.3.3 Nature of Non-Parametric Test
    1.3.4 Common Types of Non-Parametric Tests
    1.3.5 Differences between Parametric and Non-parametric:
    1.3.6 Advantages of Non-Parametric Tests
    1.3.7 Limitations of Non-Parametric Tests
1.4 Chi-Square Test: Introduction:
    1.4.1 Meaning of Chi-Square Test
    1.4.2 Categories and Frequency Tables
    1.4.3 Observed and Expected Frequencies
    1.4.4 Steps for Solving Problems Using the Chi-Square (χ^2) Test
    1.4.5 Nature of Chi-Square Test
    1.4.6 Conditions for the Application of the Chi-Square (χ^2) Test
    1.4.7 Advantages of the Chi-Square Test
    1.4.8 Limitations of the Chi-Square Test
1.5 Summing Up
1.6 Questions and Exercises
1.7 Answers to Check Your Progress
1.8 References and Suggested Readings

1.1: Introduction:

Statistical analysis is essential in research across various disciplines, offering tools to derive meaningful insights from data. Broadly, statistical tests are categorized into two types: parametric and non-parametric tests. Parametric tests depend on specific assumptions regarding population parameters, whereas non-parametric tests are distribution-free, making them suitable for situations where these assumptions are not met. Non-parametric tests are especially valuable when analyzing ordinal or nominal data, working with small sample sizes, or dealing with datasets that do not conform to a normal distribution. The Chi-Square Test is a fundamental tool for examining categorical data, helping researchers determine whether a significant association exists between two categorical variables or if an observed frequency distribution deviates from an expected one. This unit focuses on non-parametric tests, with a particular emphasis on the Chi-Square Test, one of the most widely applied statistical methods. Through this unit, learners will gain both a conceptual understanding and practical knowledge of non-parametric methods in statistical analysis.

1.2: Objectives

By the end of this unit, learners will understand:

- Define non-parametric tests and distinguish them from parametric tests.
- Explain the significance of non-parametric tests in statistical analysis.
- Identify the key features and assumptions of non-parametric tests.
- Discuss situations where non-parametric tests are preferred over parametric tests.

- Define the Chi-square test and its purpose in statistical analysis.
- Identify different types of Chi-square tests (e.g., test for goodness of fit and test for independence).

1.3: Introduction: Non-Parametric Test

Statistical analysis is an essential tool in research, helping scholars and practitioners make informed decisions based on data. Learners, you have already come to know that statistical methods are broadly classified into parametric and non-parametric tests.

- Have you ever encountered a situation where your data does not fit the usual statistical assumptions?
- What if your data is not normally distributed or is measured in categories rather than numbers?

In such cases, non-parametric tests become an excellent tool for analysis. Thus, while parametric tests assume specific conditions about population parameters and distributions, non-parametric tests provide an alternative when these assumptions are not met.

1.3.1: Meaning of Non-Parametric Test

A non-parametric test is a type of statistical test that does not require assumptions about the underlying population distribution required for parametric tests. These are statistical methods used to describe characteristics of a population, test hypotheses, examine relationships between variables, or analyze differences across populations, time periods, or related constructs. Unlike parametric tests, which require data to be normally distributed and measured on an interval or ratio scale, non-parametric tests can handle ordinal or nominal data, making them highly flexible and applicable in diverse research scenarios.

Imagine you are conducting a survey where participants rank their preferences (e.g., favorite teaching methods). Since, these rankings are **ordinal data** (not numerical measurements), traditional parametric tests may not work. This is where **non-parametric tests** come in.

These methods do not require assumptions about the distribution of population data and do not necessitate interval-level measurement (Hebel, 2002). These tests are often referred to as distribution-free tests because they do not rely on specific parameters (such as mean and standard deviation) to conduct analysis. The term distribution-free tests refer to statistical tests that do not assume a specific probability distribution (such as normal distribution) for the population from which the sample is drawn. In parametric tests like the **t-test** or **ANOVA**, the data must typically follow a normal distribution for the results to be valid. However, non-parametric tests can be used even when the distribution is unknown, skewed, or irregular**.**

Why Are They Called Distribution-Free?

1. **No Assumption of Normality:** Parametric tests require data to be normally distributed, but distribution-free tests work with any distribution.
2. **Applicable to Skewed and Non-Normal Data:** If data are highly skewed or have outliers, parametric tests may give misleading results, while non-parametric tests remain reliable.
3. **Useful for Small Samples:** When the sample size is too small to determine the distribution accurately, distribution-free tests provide a practical alternative.

Works with Various Data Types: Unlike parametric tests that require numerical data (interval or ratio scale), non-parametric tests can handle ordinal **(ranked) and** nominal (categorical) data.

1.3.2: Key Features of Non-Parametric Tests

According to Robson (1994), non-parametric tests are appropriate when analyzing nominal or ordinal variables and when the assumptions required for parametric tests are not met. A key characteristic of non-parametric statistical tests is that they do not impose specific conditions on population parameters. Additionally, they do not require the same level of measurement precision as parametric tests. Most non-parametric tests are designed for ordinal-scale data, while some can also be applied to nominal-scale data. The key features of these tests are -

1. **No Assumption of Normality:** One of the most significant advantages of non-parametric tests is that they do not require the data to follow a normal distribution. In parametric tests like the t-test or ANOVA, normality is a crucial assumption, but non-parametric tests can be applied to data that are skewed, irregular, or unknown in distribution. This makes them highly versatile for analyzing real-world data that may not always meet strict statistical assumptions.

2. **Applicable to Small Samples:** Non-parametric tests are particularly useful when working with small sample sizes where parametric assumptions (such as normality and equal variance) may not hold. Since parametric tests often require a larger dataset to achieve reliable results, non-parametric methods serve as a practical alternative when the available sample is limited.

3. **Useful for Categorical and Ranked Data:** Unlike parametric tests, which require interval or ratio data (numerical values with meaningful differences), non-parametric tests can handle ordinal (ranked) and nominal (categorical) data. This makes them ideal for analyzing

survey responses, Likert scale ratings, preference rankings, and other non-numeric data types.

4. **Robust and Flexible:** Non-parametric tests are less sensitive to outliers and extreme values, which can significantly affect the results of parametric tests. Because these tests rely on ranks or categorical groupings rather than precise numerical values, they remain reliable even in the presence of anomalies in the dataset. This robustness makes them suitable for real-world applications where data inconsistencies are common.
5. **Simpler Computational Process:** The calculations involved in non-parametric tests are generally simpler compared to parametric tests. Many non-parametric methods rely on ranking or counting rather than complex mathematical formulas involving mean, variance, or standard deviation. This simplicity makes them easier to apply, interpret, and implement, especially for researchers who may not have an extensive statistical background.

To conclude, Non-parametric tests provide a flexible, distribution-free, and robust approach to statistical analysis, making them particularly useful for small sample sizes, ranked or categorical data, and datasets with outliers. Their simplicity and applicability in diverse research fields make them an essential tool in statistical studies.

1.3.3: Nature of Non-Parametric Test:

The nature of non-parametric tests can be understood in terms of their statistical approach, applicability, and interpretation.

1. Statistical Approach in Non-Parametric Tests

Non-parametric tests use a different approach compared to parametric tests when analyzing data. Instead of relying on exact numerical values (like mean and standard deviation) these tests use ranking, frequency counting, or categorical analysis instead of relying on exact numerical values like mean and standard deviation. This makes them more flexible and applicable to different types of data, especially when parametric assumptions are not met. These tests focus on the order of values and group comparisons rather than specific measurements. This approach makes them more flexible and widely applicable, especially when dealing with ordinal or categorical data.

A. How Do Non-Parametric Tests Work?

- **Ranking Instead of Exact Values:** Instead of directly using numerical values, non-parametric tests rank the data from lowest to highest. Example: If we have five students' scores (45, 78, 66, 89, 52), instead of using the actual scores, we assign them ranks: 1st, 2nd, 3rd, 4th, and 5th place based on their performance.
- **Frequency Counting:** These tests count how often a certain category or value appears in the data. Example: In a survey where respondents choose "Yes" or "No," a non-parametric test will compare the number of people who selected each option rather than calculating averages or variances. Similarly, if we survey 100 people about their favorite color, we count how many chose red, blue, green, etc**.** and compare the frequencies.
- **Categorical Analysis:** Non-parametric tests analyze data that falls **into** groups or categories**,** such as gender (male, female, other) or education level (primary, secondary, higher). These tests analyze group differences or associations without

assuming that the data follows a normal distribution. Example: A Chi-square test can determine if there is a significant association between gender and preference for a particular product. Similarly, if we want to compare customer satisfaction across three brands (Brand A, Brand B, Brand C), a non-parametric test will assess how responses (e.g., "Satisfied," "Neutral," "Dissatisfied") vary among the brands without assuming that satisfaction scores are normally distributed.

B. What Do Non-Parametric Tests Evaluate?

- **Relationships**: They examine how two variables are connected (e.g., Spearman's rank correlation measures whether higher income is associated with higher job satisfaction). Example: **Spearman's Rank Correlation** is used to measure the relationship between two variables, such as **student attendance and exam performance**.

- **Differences**: They compare groups without assuming normal distribution (e.g., Mann-Whitney U test compares two groups like male and female students' test scores). Example: The **Mann-Whitney U Test** can compare the effectiveness of two teaching methods by analyzing student ranks rather than exact scores.

- **Associations**: They determine if two categorical variables are related (e.g., Chi-square test checks if gender and voting preference are linked). Example: The **Chi-Square Test** helps analyze whether customer preferences are linked to their age group.

The non-parametric statistical approach is simple, flexible, and reliable, especially when working with small samples, ranked data, or categorical information. It helps researchers find meaningful patterns, relationships, and group differences without relying on strict mathematical assumptions. This makes non-parametric tests widely applicable in education, healthcare, business, and social sciences**.**

2. Applicability of Non-Parametric Tests

Non-parametric tests are particularly useful in situations where traditional parametric tests may not be suitable. They are applied in the following conditions:

- **When Data is Ordinal or Nominal:** Non-parametric tests are ideal when dealing with data that is ranked (ordinal) or categorized (nominal) rather than numerical. ***Example:*** A survey measuring customer satisfaction with responses like "Very Satisfied," "Satisfied," "Neutral," "Dissatisfied and "Very Dissatisfied" can be analyzed using non-parametric methods since the data represents rankings rather than exact numerical values. Similarly, gender (male/female/other) or education level (primary, secondary, higher education) are nominal categories that can be analyzed using non-parametric techniques.

- **When the Sample Size is Small:** Parametric tests require a large sample size to produce reliable results, especially when estimating population parameters like the mean and standard deviation. Non-parametric tests, however, can provide meaningful insights even when the sample size is small and does not meet the conditions for normality. ***Example:*** If a researcher is studying the effectiveness of a new teaching method in a small classroom of 10 students, a non-parametric test would be a more appropriate choice than a parametric test, which might require a larger group.

- **When Data is skewed or Contains Outliers:** In real-world scenarios, data is not always symmetrically distributed. It can be skewed (having more values on one side) or contain outliers (extreme values that distort averages). Parametric tests assume normal distribution and can be misleading when data does not follow this assumption. Non-parametric tests, on the other hand, are not influenced by skewed distributions or extreme values, making them a more reliable choice. ***Example:*** Suppose a dataset records income levels in a population. A few billionaires in the sample might inflate the average income, making it an inaccurate representation of the majority. A non-parametric test can analyze median incomes instead of the mean, providing a more realistic comparison.

- **When Population Parameters Are Unknown or Undefined:** Parametric tests rely on knowing or estimating key population parameters such as mean, variance, and standard deviation. If these parameters are unknown, unreliable, or difficult to estimate, parametric methods may not be applicable. Non-parametric tests do not require these assumptions and can still analyze relationships, differences, or trends effectively. ***Example:*** If a researcher is analyzing ancient historical data where population statistics are unknown or missing, non-parametric tests would allow meaningful comparisons without requiring exact population parameters.

3. Interpretation of Non-Parametric Tests

Since non-parametric tests rely on rankings, categorical groupings, or frequency counts instead of exact numerical values, their results are interpreted differently compared to parametric tests.

- **Focus on Statistical Significance Rather Than Precise Estimates:** In parametric tests, results are often expressed in terms of means, standard deviations, and effect sizes. In non-parametric tests, results are interpreted based on statistical significance rather than exact numerical differences. The tests determine whether a relationship or difference exists, rather than how large the difference is in absolute numbers. ***Example:*** If we use the Mann-Whitney U test to compare the effectiveness of two teaching methods, the result will indicate whether one method is significantly better than the other, but it won't give a precise numerical difference in student performance scores.

- **Trend-Based Conclusions Instead of Exact Predictions:** Since non-parametric tests use rankings or category comparisons, they ighlight trends and patterns rather than providing precise predictive models. ***Example:*** A Spearman's rank correlation test may show that as income increases, happiness levels tend to rise, but it won't quantify the exact increase in happiness for a given income level.
- **Less Emphasis on Distributional Properties:** Because these tests do not assume a specific distribution, results can be generalized across different types of datasets. The emphasis is on whether a significant association or difference exists rather than on fitting the data into a predefined mathematical model.

To sum up Non-parametric tests are powerful tools for analyzing data when standard parametric assumptions are not met. They are especially useful for ordinal and nominal data, small sample sizes, skewed distributions, and cases where population parameters are unknown. When interpreting their results, the focus is on statistical significance, trends, and categorical comparisons, rather than precise numerical estimates. This makes non-parametric methods widely applicable in social sciences, medical research, market analysis, and education.

Let's think about some real-life situations:

- ☑ **Your data is in ranks or categories** – For example, student satisfaction levels (e.g., high, moderate, low).
- ☑ Your **sample size is small** – When you don't have enough data to assume a normal distribution.
- ☑ **Your data is skewed or has outliers** – If extreme values affect the results, non-parametric tests are a safer option.

1.3. 4. Common Types of Non-Parametric Tests

There are several widely used non-parametric tests, each serving different purposes. Some of the most common ones include:

Test Name	Purpose	Used For
Chi-square Test	Tests independence between categorical variables	Nominal data
Mann-Whitney U Test	Compares two independent groups	Ordinal or non-normal continuous data
Wilcoxon Signed-Rank Test	Compares two related samples	Ordinal paired data
Kruskal-Wallis Test	Compares more than two independent groups	Ordinal or non-normal continuous data
Friedman Test	Compares repeated measures on the same subjects	Ordinal or ranked data

Each of these tests has specific applications and is used when parametric tests like t-tests and ANOVA are not appropriate.

1.3.5: Differences between Parametric and Non-parametric:

Differences between parametric and non-parametric tests are based on the assumptions they make about the data. Below is a comparison of the two:

1. **Definition:**

- ***Parametric Test:*** A statistical test that assumes the underlying data follows a known probability distribution, usually a normal distribution.
- ***Non-Parametric Test:*** A statistical test that does not require assumptions about the distribution of the data.

2. **Assumptions**

- ***Parametric Test:***
  - — Data should follow a normal distribution.
  - — The sample size should be large.
  - — Homogeneity of variance (equal variance) is assumed.
  - — Data should be measured on an interval or ratio scale.
- ***Non-Parametric Test:***
  - — No assumption about the normality of data.
  - — Can be used with small sample sizes.
  - — Does not require homogeneity of variance.
  - — Can be applied to ordinal, nominal, or skewed data.

3. **Examples**

- **Parametric Tests**:
  - — t-test (for comparing means of two groups)
  - — ANOVA (for comparing means of multiple groups)

— Pearson's correlation **(**for measuring relationships between variables)

— Regression Analysis

- **Non-Parametric Tests**:

— **Mann-Whitney U test** (alternative to t-test for two independent groups)

— **Kruskal-Wallis test** (alternative to ANOVA for multiple groups)

— **Spearman's rank correlation** (alternative to Pearson's correlation)

— **Chi-square test** (for categorical data)

4. **When to Use**

- Use parametric tests when the data meets all assumptions (normality, homogeneity, sufficient sample size).
- Use non-parametric tests when the data is non-normal, has outliers, or is measured on ordinal or nominal scales.

In essence, **b**oth parametric and non-parametric tests have their own significance in statistical analysis. Parametric tests are preferred for their efficiency and statistical power when assumptions are met, whereas non-parametric tests provide flexibility and robustness when dealing with non-normal or small datasets.

Summary: Comparison of Parametric Tests and Non parametric Test

BASIS FOR COMPARISON	PARAMETRIC TEST	NONPARAMETRIC TEST
Meaning	A statistical test, in which specific assumptions are made about the population parameter, is known as parametric test.	A statistical test used in the case of non-metric independent variables, is called non-parametric test.
Basis of test statistic	Distribution	Arbitrary
Assumption	Data must be normally distributed	No normality assumption required
Measurement level	Interval or ratio(Numerical)	Nominal or ordinal(Categories, Ranks)
Measure of central tendency	Mean	Median
Dependence on Mean & Standard Deviation	Yes	No
Information about population	Completely known	Unavailable
Applicability	Variables	Variables and Attributes
Correlation test	Pearson	Spearman
Example	t-test, ANOVA	Mann-Whitney U test, Chi-square test

1.3.6: Advantages of Non-Parametric Tests

Non-parametric tests provide several benefits, making them useful in statistical analysis, particularly when data does not meet the assumptions required for parametric tests.

1. **Applicable to Various Data Types** – These tests can be used with different types of data, including ordinal and nominal scales, making them more versatile than parametric tests, which require interval or ratio data.
2. **Minimal Assumptions** – Unlike parametric tests, non-parametric methods do not require the data to follow a normal distribution or have equal variances, making them suitable for diverse datasets.
3. **Suitable for Small Samples** – Non-parametric tests are effective even when working with small sample sizes, where parametric tests might not be reliable.
4. **Less Affected by Outliers** – Since these tests rely on rankings or frequency counts rather than exact numerical values, extreme values have little impact on the results, ensuring greater robustness.
5. **Simple and Intuitive** – The computational methods involved in non-parametric tests are generally easier to understand and apply, making them accessible even for those with limited statistical knowledge.

1.3.7: Limitations of Non-Parametric Tests

While non-parametric tests offer flexibility and robustness, they also have certain drawbacks that limit their applicability in some research contexts.

1. **Lower Precision** – Unlike parametric tests, which use exact numerical values, non-parametric tests rely on rankings or frequency counts. As a result, they provide less precise estimates of central tendencies and variability.
2. **Reduced Statistical Power** – Non-parametric tests often require larger sample sizes to achieve the same level of statistical significance as parametric tests, making them less efficient in detecting small effects.
3. **Limited Scope for Certain Research Questions** – These tests are most effective for analyzing ordinal or categorical data but may not be suitable for research that requires detailed analysis of interval or ratio data, such as mean comparisons.
4. **Challenges in Interpretation** – Since non-parametric tests focus on medians and rank-based comparisons rather than mean differences, interpreting the results can sometimes be less intuitive, especially when trying to quantify the magnitude of an effect.

To sum up, non-parametric tests play a critical role in statistical analysis, especially when data do not meet the assumptions required for parametric methods. Their ability to handle small samples, categorical data, and skewed distributions makes them indispensable in many research fields.

Comparison: Advantages and Disadvantages of Parametric Test and Non-Parametric Test

Criteria	Parametric Test	Non-Parametric Test
Efficiency	More powerful when assumptions are met.	Less powerful but useful when assumptions are violated.
Applicability	Requires normal distribution and interval/ratio data.	Works with ordinal, nominal, and skewed data.
Interpretation	Easier to interpret as they provide specific measures like mean differences.	Harder to interpret as they rely on ranks and medians.
Sample Size	Requires a larger sample size.	Works well with small samples.

Both parametric and non-parametric tests have their own significance in statistical analysis. Parametric tests are preferred for their efficiency and statistical power when assumptions are met, whereas non-parametric tests provide flexibility and robustness when dealing with non-normal or small datasets. Non-parametric tests provide a powerful way to analyze data when traditional parametric methods do not apply. They are flexible, easy to use, and valuable in many real-world scenarios. Next time you have data that doesn't fit normal distribution assumptions, consider using a non-parametric test!

Check Your Progress: 1

1. Define non-parametric tests and explain why they are called distribution-free tests.
2. Describe a real-world scenario where a non-parametric test would be more appropriate than a parametric test.

3. Explain the differences between ranking-based and frequency-based non-parametric tests.

Non-parametric tests offer a robust, distribution-free approach to statistical analysis, particularly useful for small samples, ranked or categorical data, and datasets with outliers. Their flexibility and simplicity make them widely applicable in fields such as education, healthcare, business, and social sciences.

In the next section, we will focus specifically on the ***Chi-square test****,* one of the most widely used non-parametric tests for analyzing categorical data.

1.4: Chi-Square Test: Introduction:

We now turn to our final topic, the chi-square (χ^2) test, a specialized non-parametric test. While its structure differs from the parametric tests we have previously explored the fundamental principles of hypothesis testing remain unchanged. Among the various statistical techniques, the Chi-square test is one of the most widely used non-parametric tests for analyzing categorical data. It is particularly useful when dealing with nominal (categorical) data and when the assumptions of parametric tests (such as normality and equal variance) are not met. Let us discus the Chi-square test, its nature, types, applications, advantages, and limitations.

1.4.1: Meaning of Chi-Square Test

The **Chi-square (χ^2) test** is a statistical method used to assess the association between categorical variables. It examines whether the observed frequencies in a dataset differ significantly from the expected frequencies. The test determines whether the variation in data is due to chance or it also price to find out real deference existed if there is a meaningful relationship between variables. The primary purpose of the Chi-square test is to analyze the frequency distribution of a single categorical variable or to examine relationships between two categorical variables, particularly nominal variables, which are distinguished solely by their names and the frequencies with which they appear. This makes the test a valuable tool in various research contexts.

The Chi-square test is particularly useful for comparing multiple groups when both the independent and dependent variables are categorical. It helps determine whether a given frequency distribution arises due to a specific cause or is merely the result of chance. This is done by comparing the observed distribution with the expected distribution, assuming chance as the only influencing factor. If the difference between observed and expected frequencies is small, it suggests that chance may be the sole factor at play. However, if the difference is significant, it indicates the presence of an underlying factor influencing the outcome.

1.4.2: Categories and Frequency Tables

Unlike previous statistical tests that relied on measures such as means and standard deviations, the chi-square test utilizes frequency tables to represent and analyze data. In a chi-square test, data are categorized into nominal variables—categories that have no inherent order and are defined by their labels. Since nominal data cannot be summarized using means or standard deviations, they are instead represented in frequency tables.

A frequency table for a chi-square test consists of columns representing different categories of a single variable. For example, in a steam preference study, the categories could be "Arts," "Science," “Commerce,” and "Other." The chi-square test can accommodate as few as two categories, but having too many can complicate calculations and interpretation. The final column in the table represents the total number of observations (N). The test assumes that each observation comes from a different individual and that each person provides only one response. Thus, the total number of observations always equals the sample size.

Points to Remember

- The Chi-square test is a widely used non-parametric test, where the sampling distribution of the test statistic follows a chi-square distribution when the null hypothesis is true.
- Introduced by Karl Pearson, the test is primarily used to assess association between categorical variables. It is represented by the Greek letter χ^2.
- The test is applicable when there are minimal or no assumptions about the population parameters.
- It is used for categorical or qualitative data, typically analyzed using a contingency table.
- The Chi-square test is particularly useful for evaluating unpaired or unrelated samples and comparing proportions.

1.4.3: Observed and Expected Frequencies

A chi-square frequency table consists of two rows:

- **Observed Frequencies (fo):** These are the actual counts recorded in each category. For example, 14 people may prefer cats, 17 prefer dogs, and 5 prefer another pet.
- **Expected Frequencies (fe):** These represent the hypothetical values if all categories were equally

represented. The expected value for each category is calculated as:

$$fe = \frac{\boldsymbol{N}}{\mathbf{C}}$$

Where:

- ***N*** is the total sample size and
- ***C*** is the number of categories.

Since expected frequencies represent the assumption of equal distribution, the chi-square test determines whether the observed data significantly deviates from this expectation.

Importance of Assumption of Expected Frequencies in the Chi-Square Test

The assumption of expected frequencies is crucial because:

- **Ensures Validity of the Test:** The Chi-Square test approximates the true distribution well only when expected frequencies are sufficiently large (usually ≥ 5 in each cell).
- **Prevents Inaccurate Results:** When expected counts are too low, the Chi-Square approximation becomes unreliable, leading to an increased risk of Type I (false positive) or Type II (false negative) errors.
- **Justifies Use of the Chi-Square Distribution:** The test relies on the assumption that each expected frequency represents a sufficiently large sample size for statistical comparison.

1.4.4: Steps for Solving Problems Using the Chi-Square (χ^2) Test

Step 1: Calculate the expected frequencies using the formula:

- $fe = \frac{N}{C}$ (For Goodness of Fit Test)
- $fe = \frac{\text{Row Total} \times \text{Column Total}}{\text{Grand Total}}$ (For test of Independence)

Step 2: Find the difference between the observed (fo) and expected (fe) frequencies, then square the differences:

$$(fo - fe)^2$$

Step 3: Divide each squared difference by the corresponding expected frequency (fe) and sum up all the values to compute the Chi-square statistic **(χ^2)** using the formula:

$$\chi 2 = \sum [\frac{(fo - f\ )^2}{fe}]$$

Where:

- **χ^2** = Chi-square value
- **fo** = Observed frequency (actual data collected)
- **fe** = Expected frequency (hypothesized or theoretical frequency)

If the calculated χ^2 value is large, it suggests a significant difference between observed and expected values, indicating a possible relationship between the variables.

1.4.5: Nature of Chi-Square Test

- The **Chi-square test is non-parametric**, meaning that it does not assume a normal distribution or specific population parameters. Its nature can be understood through the following key characteristics:

- **Distribution-Free Test**: Unlike parametric tests, the Chi-square test does not require data to follow a normal distribution, making it applicable in a wide range of scenarios.
- **Categorical Data Analysis**: It is specifically designed for categorical data (nominal or ordinal), making it useful in social sciences, business, medical research, and education.
- **Comparison of Expected vs. Observed Frequencies**: The test compares the expected frequencies (hypothesized distribution) with the observed frequencies (actual data collected) to check for significant differences.
- **Based on Frequencies, Not Mean Values**: Unlike parametric tests that analyze mean values and variances, the Chi-square test works with frequency counts in different categories. The Chi-square test involves minimal mathematical calculations and does not require parameter values such as mean or standard deviation.
- **Requires a Sufficiently Large Sample Size**: For accurate results, the Chi-square test is most effective when sample sizes are moderate to large (expected frequencies should be at least 5 in each category).
- **Uses Degrees of Freedom (df) :** The test result depends on the degrees of freedom **(df)**, calculated as:
    - **df = (r - 1) times (c - 1)**
    - **$df=(r-1)\times(c-1)$**

Where **r** is the number of rows and **c** is the number of columns in a contingency table.

- **Hypothesis-testing**: It is primarily a hypothesis-testing technique and is not used for estimation purposes.

- **Additive property**: It exhibits the additive property, meaning the Chi-square values from independent data can be summed up.

1.4.6: Conditions for the Application of the Chi-Square (χ^2) Test

- There must be two sets of data: either two observed sets or one observed set and one expected set. Typically, the data is arranged in an n × c contingency table (with *n* rows and *c* columns).
- The sample size should be sufficiently large, generally 50 or more observations.
- The expected frequency in each category should not be too small and should ideally be at least 5.
- The data should be collected on a random basis, ensuring that observations are not biased.
- The items in the sample must be independent, meaning that one observation should not influence another.
- This test is particularly useful in research and is applied to complex contingency tables with multiple categories. In a contingency table, variables are organized in different formats to represent various attributes of a population.

Check Your Progress-2

1. Fill in the Blanks:

a) The Chi-square test is used to analyze __________ data.

b) The Chi-square formula compares __________ and __________ frequencies.

c) The Chi-square test is classified into two types: the __________ test and the ---test

d) The degrees of freedom for a Chi-square test is calculated using the formula—

e) If the expected frequency in any category is less than 5, the __________ test is recommended instead of the Chi-square test.

1.4.7: Advantages of the Chi-Square Test

The Chi-square test is a widely used statistical tool with several benefits, making it valuable for analyzing categorical data.

1. **Simple and Easy to Use** – The test requires only categorical data and involves straightforward calculations, making it accessible even for those with minimal statistical expertise.
2. **No Requirement for Normality** – Unlike parametric tests, the Chi-square test does not assume that the data follows a normal distribution, making it suitable for analyzing diverse datasets.
3. **Effective in Non-Parametric Analysis** – Since it does not require numerical values or specific assumptions about data distribution, it is useful when parametric conditions are not met.
4. **Applicable Across Multiple Fields** – The Chi-square test is used in various disciplines, including social sciences, healthcare, business, and marketing, to study relationships between categorical variables.

1.4.8: Limitations of the Chi-Square Test

While the Chi-square test is a useful statistical tool, it has certain limitations that researchers must consider when interpreting results.

1. **Highly Sensitive to Sample Size** – In very large samples, even small and practically insignificant differences may appear statistically significant, potentially leading to misleading conclusions.
2. **Does Not Measure the Strength of Association** – The test only determines whether a relationship exists between categorical variables but does not indicate how strong or meaningful the association is.
3. **Requires Independence of Observations** – The Chi-square test assumes that each observation is independent. If this assumption is violated (e.g., in repeated measures or clustered data), the results may not be valid.
4. **Not Reliable for Small Expected Frequencies** – If the expected frequency in any category is less than 5, the Chi-square test may yield inaccurate results. In such cases, alternative tests like Fisher's Exact Test should be considered.

Conclusion: The Chi-square test is a powerful statistical tool used for analyzing categorical data. It is particularly useful in examining relationships between variables and determining whether observed distributions differ from expected ones. Being a non-parametric test, it does not require normality assumptions, making it highly applicable in real-world research scenarios. By understanding the meaning, nature, types, applications, and limitations of the Chi-square test, researchers can effectively use this method to draw meaningful conclusions from categorical data.

1.5: Summing Up:

Parametric and non-parametric tests are statistical methods used for data analysis, differing in their assumptions. Parametric tests assume a normal distribution, require larger sample sizes, and analyze interval or ratio data (e.g., t-test, ANOVA, Pearson's correlation). Non-parametric tests are statistical methods that do not assume a specific data distribution, making them suitable for small samples, ordinal or categorical data, and cases where normality is not met. Common non-parametric tests include the Mann-Whitney U test, Kruskal-Wallis test, and Chi-square test, which are used to analyze differences, relationships, and associations in data. These tests serve as reliable alternatives when parametric assumptions are violated.

The Chi-square (χ^2) test is a non-parametric test used to analyze categorical data by comparing observed and expected frequencies. The test involves formulating hypotheses, organizing data in a contingency table, calculating expected frequencies, computing the Chi-square value, and interpreting significance. While widely used, it is sensitive to sample size and does not measure the strength of associations, requiring careful application in research. The Chi-square test follows a stepwise process that includes formulating hypotheses, organizing data in a contingency table, calculating expected frequencies, computing the Chi-square value, and interpreting results based on statistical significance. While the test is simple and widely applicable, it has limitations, such as sensitivity to sample size and the inability to measure the strength of association. Understanding these aspects ensures proper application and interpretation of results in research.

1.6. Questions and Exercises:

Fill in the Blanks:

1. The __________ test is used to determine the independence between categorical variables.
2. The Mann-Whitney U test is an alternative to the __________ test for comparing two independent groups.

3. A key advantage of non-parametric tests is that they make __________ assumptions about data distribution.
4. The __________ test is used to compare more than two independent groups when data is not normally distributed.
5. The Chi-square test is used to analyze __________ data.
6. The Chi-square formula compares __________ and __________ frequencies.
7. The Chi-square test is classified into two types: the __________ test and the __________ test.
8. The degrees of freedom for a Chi-square test is calculated using the formula __________.
9. If the expected frequency in any category is less than 5, the __________ test is recommended instead of the Chi-square test.

Short Answer Questions:

1. Define parametric and non-parametric tests.
2. Explain any two assumptions of parametric tests.
3. What are the advantages of using non-parametric tests?
4. Give two examples of parametric and non-parametric tests.
5. Describe the types of data for which non-parametric tests are most suitable, with examples.
6. Define the Chi-square test and explain its purpose in statistical analysis.
7. What are the two main types of Chi-square tests, and how do they differ?
8. Explain the formula used in the Chi-square test, including the meaning of each component.
9. Why the Chi-square test is considered a non-parametric test?

Essay-Type Questions:

1. Compare and contrast parametric and non-parametric tests with suitable examples.
2. Describe the types of data for which non-parametric tests are most suitable, with examples.
3. Explain when a researcher should use a non-parametric test instead of a parametric test.
4. Discuss the applicability of non-parametric tests in real-world research scenarios.
5. Discuss the key features of non-parametric tests with examples.
6. Define non-parametric tests and discuss their importance in statistical analysis.
7. Discuss the advantages and disadvantages of using non-parametric tests in research.
8. Discuss the steps involved in conducting a Chi-square test for independence with an example.
9. Explain the advantages and limitations of the Chi-square test.
10. Describe how the Chi-square Goodness of Fit Test is used in research, with an example.
11. Suppose a researcher wants to analyze whether students' preference for online or offline learning is related to their academic performance levels (High, Medium, Low). Describe how the Chi-square test would be applied in this study.

Application-Based Questions

6. A researcher is analyzing customer satisfaction ratings (ranked from 1 to 5) across three different retail stores. Which non-parametric test should they use, and why?

7. A psychologist wants to compare stress levels in two independent groups of students but finds that the data is not normally distributed. Suggest an appropriate test and justify your choice.
8. In a clinical trial, researchers measure patients' responses to a treatment at three different time points. Which non-parametric test would be suitable for analyzing the data?

1.7. Answers to Check Your Progress

Check Your Progress-1

1. Non-parametric tests are statistical methods that do not assume a specific probability distribution. They are called **distribution-free tests** because they do not rely on parameters like mean and standard deviation.

2. A non-parametric test would be suitable for analyzing customer satisfaction ratings (e.g., "Very Satisfied" to "Very Dissatisfied") since the data is ordinal and does not require a normal distribution.

3. Ranking-based tests, such as the Mann-Whitney U test, order data points without considering their exact values, whereas frequency-based tests, like the Chi-square test, count occurrences in different categories.

Check Your Progress-2

a) Categorical
b) Observed, Expected
c) Goodness of Fit, Independence
d) $df=(r-1)\times(c-1)$
e) Fisher's Exact Test

1.8. References and Suggested Readings

1. Agresti, A. (2018). *Statistical Methods for the Social Sciences* (5th Ed.). Pearson Education.
2. Garret H.E-Statistics inPsychology and Education .Bombay .Alied Pacific private Ltd.,1962
3. Gupta, S. P. (2021). *Statistical Methods* (45th ed.). Sultan Chand & Sons.
4. Kothari, C. R., & Garg, G. (2019). *Research Methodology: Methods and Techniques* (4th ed.). New Age International Publishers.
5. Levin, R. I., & Rubin, D. S. (2020). *Statistics for Management* (8th ed.). Pearson India.
6. Siegel, S., & Castellan, N. J. (1988). *Nonparametric Statistics for the Behavioral Sciences.* McGraw-Hill.
7. Srivastava, U. K., & Sharma, G. V. (2021). *Quantitative Techniques for Managerial Decisions*. New Age International Publishers.
8. Verma, J. P. (2013). *Data Analysis in Management with SPSS Software*. Springer.
9. Yamane, T. (1967). *Statistics: An Introductory Analysis* (2nd ed.). Harper & Row.

*****

UNIT-2
ASSUMPTIONS OF CHI-SQUARE TESTS

Unit Structure:

2.1 Introduction
2.2 Objectives
2.3 Introduction: Assumption of Chi-Square Test
2.4 Assumption 1: The Data Should Be Categorical
2.5 Assumption 2: Observations Should Be Independent
2.6 Assumption 3: The Sample Size Should Be Sufficiently Large
2.7 Assumption 4: The Expected Frequency in Each Cell Should Not Be Too Small
2.8 Assumption 5: The Data Should Be Based on Random Sampling
2.9 Summing Up
2.10 Answers to Check Your Progress
2.11 Questions and Exercises
2.12 References and Suggested Readings

2.1 Introduction:

In the previous unit, we explored the meaning and nature of non-parametric tests, with a focus on the Chi-Square Test and its significance in analyzing categorical data. We learned that the Chi-Square Test is a powerful statistical tool used to determine associations between variables by comparing observed and expected frequencies. It helps researchers determine whether there is a significant association between two or more groups based on observed and expected frequencies. However, like any statistical method, the Chi-Square Test is most effective when applied under certain **assumptions**. In this unit, we will explore the key assumptions of the Chi-Square Test, which ensure the test results are valid and reliable. Understanding these assumptions is essential for correctly applying the test in research and avoiding errors in interpretation of results. We will discuss factors such as the type of data required, sample size considerations, expected frequency conditions, and the independence of observations.

2.2 Objectives:

In this unit, we will:

- Examine the fundamental assumptions underlying the Chi-Square Test.
- Determine the type of data necessary for conducting a Chi-Square Test.
- Analyze the significance of sample size and expected frequency conditions in ensuring test accuracy.
- Explore the assumption of independence of observations and its role in maintaining the validity of results.
- Understand the importance of these assumptions in statistical analysis.
- Apply the Chi-Square Test appropriately in real-world research contexts.

By the end of this unit, learners will gain a clear understanding of these fundamental assumptions, allowing them to apply the Chi-Square Test appropriately in research and practical scenarios. This knowledge will serve as a critical step in mastering the correct use of non-parametric statistical methods.

2.3 Introduction: Assumption of Chi-Square Test:

The Chi-Square Test is a widely used non-parametric statistical test for analyzing categorical data. It helps determine whether there is a significant association between two categorical variables or whether the observed frequencies differ from expected frequencies.

However, like all statistical tests, the Chi-Square Test has certain assumptions that must be satisfied for its results to be valid and reliable. If these assumptions are violated, the conclusions drawn from the test may be misleading. Now, you will be introduced with the key assumptions underlying the Chi-Square Test, with examples and illustrations are discussed below to make them easy to understand.

2.4 Assumption 1: The Data Should Be Categorical:

The Chi-Square Test is specifically designed to analyze categorical data, meaning data that can be grouped into distinct categories. It cannot be used for numerical data measured on an interval or ratio scale (such as height, weight, or income). Categorical data is classified into two types:

- **Nominal Data**: Categories with no inherent order (e.g., gender: Male, Female, Other).
- **Ordinal Data**: Categories with a meaningful rank order but unequal intervals (e.g., customer satisfaction: Poor, Average, Good, Excellent).

Illustration

- **Correct Usage:** A researcher wants to analyze whether gender (Male/Female/Other) is associated with preference for a particular sport (Football, Cricket, Basketball, and Tennis).Since both variables are categorical, the Chi-Square Test is appropriate.
- **Incorrect Usage**: A researcher wants to analyze whether height (measured in cm) is associated with **exam scores (**out

of 100). Since both variables are numerical, the Chi-Square Test cannot be applied. Instead, a different statistical test, like correlation or regression, would be more suitable.

Check Your Progress: I

Which type of data is suitable for the Chi-Square Test?

a) Categorical (Nominal/Ordinal)

b) Continuous (Interval/Ratio)

c) Both a & b

2.5 Assumption 2: Observations Should Be Independent:

The Chi-Square Test assumes that each observation in the dataset is independent. This means that the occurrence of one event should not influence the occurrence of another. Each data point should represent a separate and independent observation. If data points are related (e.g., measuring the same individuals multiple times), the Chi-Square Test may give misleading results.

Illustration

- **Correct Usage**: A survey is conducted among 500 randomly selected individuals, asking them to choose their preferred social media platform (Facebook, Instagram, Twitter, and LinkedIn). Since each person's response is independent, the Chi-Square Test is valid.
- **Incorrect Usage:** A study examines students' performance before and after a training program using the **same** group of students. Since data is collected twice from the same individuals, the observations are not independent and a different test (such as the McNemar Test) should be used instead.

Check Your Progress: 2

What should a researcher do if observations are not independent?

a) Ignore the violation

b) Use a different test like McNemar's Test

c) Reduce the sample size

2.6 Assumption 3: The Sample Size Should Be Sufficiently Large:

The Chi-Square Test works best when the sample size is large enough to provide meaningful results. If the sample is too small, the test may lack statistical power, making it difficult to detect meaningful associations leading to unreliable conclusions.

A common rule of thumb is:

- The expected frequency in each category should be at least 5.
- If any category has an expected count below 5, the Chi-Square Test may not be appropriate. It may yield inaccurate results.

Illustration

- **Correct Usage**: A researcher surveys 200 students about their preference for different teaching methods (Lecture, Group Discussion, Online Learning). Since each category has a reasonable number of responses, the Chi-Square Test can be applied.

- **Incorrect Usage:** A researcher surveys only 10 students and finds that some categories have only 1 or 2 responses. This violates the assumption of sufficient sample size, making the results unreliable. If the expected frequency is too low, researchers can increase the sample size by collecting more data or combine categories or use an alternative test.

Check Your Progress: 3
What is the minimum expected frequency required for each category in a Chi-Square Test? a) 2 b) 5 c) 10

2.7 Assumption 4: The Expected Frequency in Each Cell Should Not Be Too Small :

In a contingency table (which displays the distribution of categorical variables), each cell should have an **expected frequency** of at least **5**. This ensures that the Chi-Square calculation is meaningful. If too many cells have very low expected frequencies, the Chi-Square approximation may become inaccurate.

Illustration

Correct Usage: A **3×3 table** comparing age groups (Under 20, 21-40, 41 and above) and favorite job (Teaching, Government Office job, Private Office Job) where each cell has an expected frequency of **5 or more**, satisfies this assumption.

Incorrect Usage: A **4×4 table** where some cells contain expected counts **less than 5**, leading to unreliable Chi-Square results. If some expected values are too small, researchers can:

- Increase the **sample size** to ensure more observations in each category.

- Combine categories with small counts (e.g., merge two age groups if their expected frequencies are very low).
- Use an alternative test **Fisher's Exact Test**, which is designed for small sample sizes.

Check Your Progress: 4

What should be done if expected frequencies in some cells are too low?

a) Increase sample size

b) Combine categories

c) Both a & b

2.8. Assumption 5: The Data Should Be Based on Random Sampling

The Chi-Square Test assumes that the data is collected through a random sampling process so that the results are representative of the larger population, meaning every individual in the population had an equal chance of being selected. If the sample is biased, the test results may not be generalizable to the entire population.

Illustration

- **Correct Usage**: A researcher randomly selects 200 students from different colleges to analyze their preference for online learning. Since the selection is random, the findings can be generalized to a larger population.
- **Incorrect Usage**: A researcher surveys only students from one college and assumes the results apply to all students in the country. This violates the assumption of random sampling, leading to biased results. It is desirable to use

random sampling techniques to avoid selection bias and to ensure the sample represents different groups within the population.

Check Your Progress: 5
Why is random sampling important in the Chi-Square Test?
a) To avoid bias
b) To make results generalizable
c) Both a & b

Conclusion

The **Chi-Square Test** is an essential statistical tool for analyzing categorical data. However, its effectiveness depends on meeting key **assumptions**. Understanding these assumptions helps researchers **avoid errors**, select **appropriate tests**, and obtain **reliable conclusions**. By applying these principles, researchers can **confidently use the Chi-Square Test** in real-world studies and make meaningful interpretations of their findings.

2.9 Summing Up:

Summary of Key Assumptions

Assumption	Description
1. Data Should Be Categorical	Variables must be in categories (e.g., gender, preferences).
2. Observations Should Be Independent	Each observation should be separate from others.

3. Sample Size Should Be Sufficiently Large	Each category should have enough responses.
4. Expected Frequency Should Not Be Too Small	Each cell in a contingency table should have ≥5 observations.
5. Data Should Be Randomly Sampled	The sample should represent the entire population.

2.10 Answers to Check Your Progress

Check Your Progress – Answer Key

- Check Your Progress -1 a) Categorical (Nominal/Ordinal)
- Check Your Progress -2 b) Use a different test like McNemar's Test
- Check Your Progress – 3 b) 5
- Check Your Progress – 4 c) Both a & b
- Check Your Progress – 5 c) Both a & b

2.11 Questions and Exercises

A. Short Answer Questions:

1. What are the key assumptions of the Chi-Square Test?
2. Why is it important for the data to be categorical in a Chi-Square Test?
3. What is the minimum expected frequency required in each cell of a contingency table?
4. How does a small sample size affect the Chi-Square Test?

B. True/False Statements:

1. The Chi-Square Test can be applied to continuous data. (T/F)
2. If observations are dependent, the Chi-Square Test may produce misleading results. (T/F)

3. Expected frequencies should be at least 5 in each category. (T/F)
4. Random sampling is not necessary for the Chi-Square Test. (T/F)

C. Case Study-Based Questions:

A researcher conducts a study to analyze the relationship between dietary habits (Vegetarian, Non-Vegetarian, Vegan) and physical activity levels (Low, Medium, High).

- Identify whether this study meets the assumptions of the Chi-Square Test.
- What should the researcher do if some categories have expected frequencies less than 5?

2.12 References and Suggested Readings

1. Agresti, A. (2018). *Statistical Methods for the Social Sciences* (5th Ed.). Pearson Education.
2. Garret H.E-Statistics in Psychology and Education .Bombay .Alied Pacific private Ltd.,1962
3. Gupta, S. P. (2021). *Statistical Methods* (45th ed.). Sultan Chand & Sons.
4. Kothari, C. R., & Garg, G. (2019). *Research Methodology: Methods and Techniques* (4th ed.). New Age International Publishers.
5. Levin, R. I., & Rubin, D. S. (2020). *Statistics for Management* (8th ed.). Pearson India.
6. Siegel, S., & Castellan, N. J. (1988). *Nonparametric Statistics for the Behavioral Sciences.* McGraw-Hill.

7. Srivastava, U. K., & Sharma, G. V. (2021). *Quantitative Techniques for Managerial Decisions*. New Age International Publishers.

8. Verma, J. P. (2013). *Data Analysis in Management with SPSS Software*. Springer.

9. Yamane, T. (1967). *Statistics: An Introductory Analysis* (2nd ed.). Harper & Row.

*****

UNIT-3
USES AND SIGNIFICANCE OF CHI-SQUARE TEST

Unit Structure:

3.1 Introduction

3.2 Objectives

3.3 Understanding Practical Uses of the Chi-Square Test

3.4 Chi-Square as a Test of Goodness of Fit

3.4.1 Purpose of the Goodness-of-Fit Test

3.4.2 Application of the Chi-Square Goodness-of-Fit Test

3.4.3 Steps of Using Goodness of Fit Test

3.4.4 How to Use a Chi-Square Table

3.4.5 Example of Chi-Square Goodness-of-Fit Test

3.5 Chi-Square Test for Independence (Association Test)

3.5.1 Difference between Goodness of Fit Test and Test of Independence

3.5.2 Steps of Using a Chi-Square Test of Independence

3.5.3 Chi-Square Test of Independence: Example

3.6 Chi-Square Test of Equality (Homogeneity)

3.6.1 Independence vs. Homogeneity

3.6.2 Chi-Square Test of Homogeneity Example

3.7 Summing Up

3.8 Answers to Check Your Progress

3.9 Questions and Exercises

3.10 References and Suggested Readings

3.1 Introduction:

The Chi-Square test is a non-parametric test used to determine if there is a significant association between categorical variables. In the previous units, we explored the concept of the Chi-Square test and its assumptions, which form the foundation for understanding its applications. Building on this foundation, this unit focuses on the uses and significance of the Chi-Square test, highlighting its practical relevance in statistical analysis. It will help you understand its applications, calculations, and interpretations in different contexts. Specifically, we will discuss its applications in three key areas including Goodness of Fit, Equality, and Independence.

3.2 Objectives:

By the end of this unit, you will be able to effectively apply the Chi-Square test in real-world scenarios, interpret its results, and critically evaluate its implications in research. This unit will guide you through:

- Explaining the significance and practical applications of the Chi-Square test.
- Differentiating between various types of Chi-Square tests.
- Applying the Chi-Square test for Goodness of Fit to assess how well an observed distribution aligns with an expected distribution.
- Conducting the Chi-Square test for Equality to determine whether multiple groups have equal distributions.
- Utilizing the Chi-Square test for Independence to analyze the relationship between two categorical variables.
- Interpreting and analyzing Chi-Square test results to draw meaningful conclusions in research.

3.3 Understanding Practical Uses of the Chi-Square Test:

The Chi-Square test is a statistical method used to determine whether there is a significant association between categorical variables. It is primarily used in:

1. **Goodness of Fit Test** – To check how well observed data matches expected data. It is used to test the single variable i.e., it is fit to the population or not. It means Sample is supporting the population or not.
2. **Test of Equality (Homogeneity)** – To determine if two or more categorical distribution or groups are equal. It is used to test a single categorical variable from two different populations. It is used to determine whether frequency counts are distributed identically across different populations.
3. **Test of Independence** – To examine whether two independent variables are independent of each other or not. the association between two categorical variables. If we wish to know the relationship between traits, attributes having two or more categories.

Let's begin our journey into the practical applications of the Chi-Square test!

3.4 Chi-Square as a Test of Goodness of Fit:

The Chi-Square Goodness-of-Fit test, developed by Karl Pearson in 1900, is a statistical method used to determine whether the deviation between observed and expected values is due to chance. This test helps researchers assess whether a given frequency distribution follows a hypothesized probability distribution.

3.4.1 Purpose of the Goodness-of-Fit Test:

The Goodness-of-Fit test is an inferential statistical procedure used to evaluate how well observed data aligns with an expected distribution. Researchers often use this test to determine whether an observed frequency pattern fits an expected frequency pattern. The test assesses whether differences between observed and expected frequencies are due to random variation or indicate a significant deviation from the hypothesized distribution.

This method is particularly useful when:

- Testing whether an observed frequency distribution follows a specific probability distribution.
- Evaluating the conformity between observed and expected frequencies across different categories.
- Determining if the sample data is representative of a larger population.
- Assessing whether all categories are equally represented in the sample.

3.4.2 Application of the Chi-Square Goodness-of-Fit Test:

The test requires a set of observed data values and an assumed probability distribution. It is applied when dealing with categorical variables and their frequency counts. The **Chi-Square statistic (χ^2)** helps determine whether the observed data significantly deviates from the expected distribution.

A key aspect of this test is assessing whether the sample data provides a "good enough" fit to the expected distribution. If the sample does not align well with the assumed population characteristics, conclusions drawn from the sample may not be reliable for the larger population.

Thus, the Chi-Square Goodness-of-Fit test serves as an essential tool in verifying whether a sample distribution matches an expected theoretical distribution, ensuring the validity of statistical inferences.

3.4.3 Steps of Using Goodness of Fit Test:

1. State the null hypothesis (H_O): The sample data follows the expected distribution. Chi-square tests are non-parametric, meaning they do not estimate or compare population parameters. Instead, hypotheses are stated in verbal form. For example**:** A researcher wants to know whether the distribution of students choosing different career options (Engineering, Medicine, Arts, Commerce) follows a predicted distribution.

- **Null Hypothesis (H_0):** There is no significant difference between the observed and expected frequencies.
- **Alternative Hypothesis (H_a):** There is a significant difference between the observed and expected frequencies.

Since chi-square tests do not use mathematical statements for hypotheses (as they do not involve parameters like means or variances), the hypotheses remain descriptive.

2. Compute the expected frequencies: Compute the expected frequencies based on the assumed distribution

$$E = \frac{N}{C}$$

Where:

N is the total sample size and

C **i**s the number of categories.

For example if sample size is 60 students and numbers of career options are are 4. , the expected frequencies for each category will be 15(60 ÷ 4 = 15)

3. Apply formula: Apply the Chi-Square formula.

$$\chi 2 = \sum \left[\frac{(fo - fe)^2}{fe}\right]$$

Where:

χ^2 = Chi-square value

fo = Observed frequency (actual data collected)

fe = Expected frequency (hypothesized or theoretical frequency)

4. Find Degrees of Freedom: The degrees of freedom (df) for a chi-square test depend on the number of categories (C), rather than the number of observations or participants. The formula for degrees of freedom is:

$$\mathbf{df = (r-1)(c-1)}$$

For instance, in the above example we have four career options i. e four categories in preference career, the degrees of freedom would be 4 - 1 = 3

4. **Find out Critical value /Table value of (χ^2):** A Chi-Square table is used to find the critical value for a given significance level (α) and degrees of freedom (df) when conducting a Chi-Square test. A Chi-Square Table is a statistical table that provides critical values for the chi-square (χ^2) distribution**,** which is used in hypothesis testing to determine the significance of observed data compared to expected data. The table helps researchers and statisticians determine whether to accept or reject a null hypothesis based on the chi-square test. The table helps determine whether to accept or reject the null hypothesis based on the calculated chi-Square statistic (χ^2).The chi-square table consists of degrees of freedom (df) in the rows and significance levels (α) in the columns**.**

Chi-Square (χ^2) Critical Values Table

Degree of Freedom	Probability of Exceeding the Critical Value								
	0.99	0.95	0.90	0.75	0.50	0.25	0.10	0.05	0.01
1	0.000	0.004	0.016	0.102	0.455	1.32	2.71	3.84	6.63
2	0.020	0.103	0.211	0.575	1.386	2.77	4.61	5.99	9.21
3	0.115	0.352	0.584	1.212	2.366	4.11	6.25	7.81	11.34
4	0.297	0.711	1.064	1.923	3.357	5.39	7.78	9.49	13.28
5	0.554	1.145	1.610	2.675	4.351	6.63	9.24	11.07	15.09
6	0.872	1.635	2.204	3.455	5.348	7.84	10.64	12.59	16.81
7	1.239	2.167	2.833	4.255	6.346	9.04	12.02	14.07	18.48
8	1.647	2.733	3.490	5.071	7.344	10.22	13.36	15.51	20.09
9	2.088	3.325	4.168	5.899	8.343	11.39	14.68	16.92	21.67
10	2.558	3.940	4.865	6.737	9.342	12.55	15.99	18.31	23.21
11	3.053	4.575	5.578	7.584	10.341	13.70	17.28	19.68	24.72
12	3.571	5.226	6.304	8.438	11.340	14.85	18.55	21.03	26.22
13	4.107	5.892	7.042	9.299	12.340	15.98	19.81	22.36	27.69
14	4.660	6.571	7.790	10.165	13.339	17.12	21.06	23.68	29.14
15	5.229	7.261	8.547	11.037	14.339	18.25	22.31	25.00	30.58
16	5.812	7.962	9.312	11.912	15.338	19.37	23.54	26.30	32.00
17	6.408	8.672	10.085	12.792	16.338	20.49	24.77	27.59	33.41
18	7.015	9.390	10.865	13.675	17.338	21.60	25.99	28.87	34.80
19	7.633	10.117	11.651	14.562	18.338	22.72	27.20	30.14	36.19
20	8.260	10.851	12.443	15.452	19.337	23.83	28.41	31.41	37.57
22	9.542	12.338	14.041	17.240	21.337	26.04	30.81	33.92	40.29
24	10.856	13.848	15.659	19.037	23.337	28.24	33.20	36.42	42.98
26	12.198	15.379	17.292	20.843	25.336	30.43	35.56	38.89	45.64
28	13.565	16.928	18.939	22.657	27.336	32.62	37.92	41.34	48.28
30	14.953	18.493	20.599	24.478	29.336	34.80	40.26	43.77	50.89
40	22.164	26.509	29.051	33.660	39.335	45.62	51.80	55.76	63.69
50	27.707	34.764	37.689	42.942	49.335	56.33	63.17	67.50	76.15
60	37.485	43.188	46.459	52.294	59.335	66.98	74.40	79.08	88.38
	Not Significant							*Significant*	

Point to Remember:

A Chi-Square table is a tool used in hypothesis testing to determine if there is a statistically significant association between categorical variables. It provides critical values based on significance levels and degrees of freedom to help evaluate test results.

Degrees of freedom, along with the chosen significance level (*α = 0.05* by default), determine the critical value from the chi-square table. Unlike t-tests, chi-square tests are always one-tailed because they measure deviation from an expected distribution in any direction.

3.4.4 How to Use a Chi-Square Table

The columns of the chi-square distribution table indicate the significance level of the critical value.

1. Find the degrees of freedom (df)

Choose the significance level (α): By convention, Common significance levels are.05 (α = 1 – 95%confidence level **or** .01(α = 1 – 99%confidence level). So the columns for.05 and .01 are highlighted in the table. In rare situations, you may want to increase α to decrease your Type II error rate or decrease α to decrease your Type I error rate.

2. Find the Critical Value in the Chi-Square Table

You now have the two numbers you need to find your critical value in the chi-square distribution table:

- The degrees of freedom (*df*) are listed along the left-hand side of the table. Find the table row corresponding to the degrees of freedom you calculated.
- The significance levels (α) are listed along the top of the table. Find the column corresponding to your chosen significance level **(α)**.
- Find the column that matches the chosen **significance level (α)**. In other words, the table cell where the row and column

meet i.e. the intersection of the row and column gives the critical value (χ^2 critical).

- Example: Finding the critical value in the table below - Where the row for *df* = 3 and the column for α = .05 meet, the critical value is 7.815.

Degrees of freedom (*df*)	Significance level (α)							
	.99	.975	.95	.9	.1	.05	.025	.01
1	-------	0.001	0.004	0.016	2.706	3.841	5.024	6.635
2	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210
3	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345
4	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277
5	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086
6	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812
7	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475
8	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090
9	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666

3. Comparison of Calculated χ^2 Value with Critical Value

Now, compare this chi-square critical value to the chi-square calculated for the sample. If the calculated χ^2 is larger than critical value, then the null hypothesis will be rejected. On the other hand, if the calculated χ^2 is smaller than critical value the null hypothesis **(**H_0**)** will be accepted.

- If χ^2 calculated $\geq$ χ^2 critical**,** reject the null hypothesis **(**H_0**).**
- If χ^2 calculated < χ^2 critical**,** fail to reject the null hypothesis **(**H_0**).**

5. **Taking Decisions:** If the calculated χ^2 value **is** large**,** it suggests a significant difference between observed and expected values, indicating a possible relationship between the variables. It indicates that the observed distribution differs from the expected distribution, suggesting that the assumed theoretical distribution

may not be valid. It suggests that the observed distribution does not match the expected distribution.

Example Use of Chi-Square Table

Suppose we conduct a Chi-Square Test for a distribution having four categorical data and our calculated χ^2 = **9**.5,

Here df will **be = 4-1 =**3

- From the Chi-Square table, χ^2 critical at df = 3 and $\alpha = 0.05$ is 7.815
- As our calculated $\chi^2 = 9.5$, then $9.5 > 7.815 \rightarrow$ Reject H_0
- There exists significant difference between the observed and expected frequencies.

3.4.5. Example of Chi-Square Goodness-of-Fit Test

Problem Statement:

A school administrator wants to determine whether students in a school are evenly distributed across four extracurricular activities: **Sports, Music, Drama, and Science Club**. The administrator assumes that students should be equally interested in each activity, meaning each category should have an equal proportion of students.

A random sample of **200 students** is surveyed, and the following data is collected:

Activity	**Observed Frequency (f_o)**
Sports	50
Music	60
Drama	30
Science Club	60
Total	**200**

The administrator conducts a Chi-Square Goodness-of-Fit test **at a** 5% significance level (α = 0.05) to check if student participation is uniformly distributed across these activities.

Step 1: State the Hypotheses

- **Null Hypothesis (H_0):** The students are evenly distributed across the four activities.
- **Alternative Hypothesis (H_1):** The students are not evenly distributed across the activities.

Step 2: Find out Expected Frequencies.

If all activities are equally preferred:

- Total Number of Students =200
- Categories=4
- Expected frequencies (f_e) = 200 /4= 50

Activity	Observed Frequency (*fo*)	Expected Frequency(f_e)
Sports	50	50
Music	60	50
Drama	30	50
Science Club	60	50
Total	**200**	**200**

Step 2: Calculate the Chi-Square Test Statistic

To compute the chi-square statistic, we compare the observed (*O*) and expected (*E*) frequencies for each category. The formula is:

$$\chi 2 = \sum \left[\frac{(fo-fe)^2}{fe}\right]$$

Where:

χ^2 = Chi-square value

fo = Observed frequency (actual data collected)

f_e = Expected frequency (hypothesized or theoretical frequency)

Now, we calculate for each category:

Sl. No.	Categories	Chi-Square Test Statistic $\frac{(fo-fe)^2}{fe}$
1	**Sports**	$\frac{(50-50)^2}{50} = \frac{0}{50} = 0$
2	**Music**	$\frac{(50-60)^2}{50} = \frac{100}{50} = 2$
3	**Drama**	$\frac{(50-30)^2}{50} = \frac{400}{50} = 8$
4	**Science Club**	$\frac{(50-60)^2}{50} = \frac{100}{50} = 2$

Summing expected frequencies values:

- **χ2=0+2+8+2=12**

Step 3: Determine the Degrees of Freedom (df)

- df=(Number of categories)−1=4−1=3df

Step 4: Find the Critical Value from the Chi-Square Table

- For **df = 3** and **α = 0.05**, the **critical value** from the Chi-Square table is **7.815**.

Step 5: Compare χ^2 Calculated with χ^2 Critical

- **χ^2 Calculated** = 12
- **χ^2 Critical** = 7.815

Since **12 > 7.815**, we **reject the null hypothesis (H_0)**.

Step 6: Conclusion

The test statistic (12) exceeds the critical value (7.815), leading to the rejection of the null hypothesis. This indicates a significant difference in students' preferences for extracurricular activities. The statistical evidence suggests that students are not evenly distributed across various activities, with some being more popular than others.

This example illustrates the application of the Chi-Square Goodness-of-Fit test in evaluating whether observed data conforms to an expected distribution. The chi-square test is an essential tool for analyzing categorical data, helping researchers determine if observed distributions align with expectations. Specifically, the Goodness-of-Fit test assesses whether a single categorical variable follows a presumed distribution. By computing observed and expected frequencies, determining degrees of freedom, and comparing the chi-square statistic to critical values, researchers can draw meaningful conclusions from the data.

3.5 Chi-Square Test for Independence (Association Test):

The Chi-Square Test of Independence is an inferential (derivable) statistical test used to determine whether two categorical variables are related or independent of each other. This non-parametric test is applied when we have frequency counts for two nominal (categorical) variables. It is similar to the Chi-Square Goodness-of-Fit Test but instead examines relationships between two variables rather than the distribution of a single variable.

1. Purpose of the Test

The Chi-Square Test of Independence assesses whether two variables influence each other. It helps determine whether one variable can be used to predict the other or if they are independent.

Unlike tests that measure frequencies along a single dimension, this test evaluates frequency distributions for two variables simultaneously within a contingency table.

2. Key Requirements

To perform the test, certain conditions must be met:

1. **Sufficiently Large Sample Size** – The test requires a reasonably large sample to ensure reliable results.
2. **Independence of Observations** – Each observation in the dataset must be independent of the others.
3. **Application Example: Gender and Course Choice**

Suppose a researcher wants to analyze whether a student's gender influences their course selection. The researcher would collect data on students' gender and the courses they choose, then organize this information into a contingency table. The Chi-Square test would then be used to compare the observed and expected frequencies of course selections among male and female students.

4. Interpreting the Results

- If the calculated Chi-Square value is less than the critical value from the Chi-Square table at a given level of significance (e.g., 0.05) and degrees of freedom, we fail to reject the null hypothesis (H_0). This means that the two variables are independent, indicating no significant association between them.
- If the calculated Chi-Square value is greater than the critical value, we reject the null hypothesis, concluding that the variables are not independent and have a significant relationship.

In summary, the Chi-Square Test of Independence helps determine whether there is a meaningful association between two categorical variables based on observed data.

3.5.1. Difference between Goodness of Fit Test and Test of Independence

Feature	Goodness of Fit Test	Test of Independence
Purpose	Checks if an observed distribution matches an expected one.	Determines if two categorical variables are related.
Number of Variables	One categorical variable.	Two categorical variables.
Hypotheses	H_0: Observed data follows the expected distribution. H_A: Observed data does not follow the expected distribution.	H_0: The two variables are independent. H_A: The two variables are dependent.
Example	Checking if the number of students in different majors follows a national trend.	Checking if gender and teaching method preference are related.
Data Representation	Observed vs. Expected Frequencies in one variable.	Contingency Table comparing two variables.

In summary, the **Goodness of Fit Test** is used to compare a sample distribution with an expected distribution, while the **Test of Independence** examines the relationship between two categorical variables.

3.5.2. Steps of Using a Chi-Square Test of Independence:

The Chi-Square Test of Independence is used to determine whether there is a significant association between two categorical variables. The steps are as follows:

Step 1: Define Hypotheses: "There is no association between the variables."

Null Hypothesis (H_0): There is no association between the two categorical variables (for independence- they are independent)

- **Example:** A school administrator wants to examine whether gender (male, female) is associated with participation in extracurricular activities (Yes, No). Here -
- **Null Hypothesis (H_0):** There is no relationship between gender and participation in extracurricular activities.

Alternative Hypothesis (H_1): There is a significant association between the variables (they are not independent).

- **Alternative Hypothesis (H_1):** There is a significant relationship between gender and participation in extracurricular activities.

Steps-2: Create a contingency table.

This contingency table organizes the observed frequency data for different categories. A contingency table is a two-way table showing the contingency between two variables where the variables have been classified into mutually exclusive categories and the cell entries are frequencies. It displays the counts of occurrences for each combination of the categorical variables.

Step 2: Create a Contingency Table

	Category 1	**Category 2**	**Category 3**	**Total**
Group 1	O_{11}	O_{12}	O_{13}	Row Total
Group 2	O_{21}	O_{22}	O_{23}	Row Total
Total	Column Total	Column Total	Column Total	Grand Total

Steps-3: Calculate expected frequencies (fe)

The expected frequency for each cell in the table is calculated using the formula:

$$fe = \frac{\text{Row Total} \times \text{Column Total}}{\text{Grand Total}}$$

- **Row Total** = Sum of observations in that row
- **Column Total** = Sum of observations in that column
- **Grand Total** = Total number of observations

Step 4: Calculate Chi-Square Value

Using the formula:

$$\chi 2 = \sum \left[\frac{(fo - fe)^2}{fe} \right]$$

Where:

- χ^2 = Chi-square value
- fo = Observed frequency (actual data collected)
- fe = Expected frequency (hypothesized or theoretical frequency)

Step 5: Determine the Degrees of Freedom (df)

The degree of freedom for the test is calculated as:

- df = (r - 1) (c - 1)

∴ **$df=(r-1)\times(c-1)$**

Where:

- **r** = Number of rows
- **c** = Number of columns

Step 6: Compare with Critical Value

- The calculated χ^2 value is compared with the critical value from the Chi-Square table at a chosen significance level (α = 0.05 or 0.01) using the degrees of freedom as previous test of Goodness of Fit.

Step 7: Make a Decision

- **If** the **calculated χ^2 value** is greater than the critical value **or** the **p-value** is less than **α (0.05)**, **reject** the null hypothesis (**H_0**) → This indicates a significant association between the variables.
- **If not**, fail to reject the null hypothesis (**H_0**) → This means there is no significant association between the variables.

Step 7: Interpretation of Chi-Square Results

- If the calculated χ^2 value is greater than the critical value, we reject the null hypothesis, meaning there is a significant relationship between the variables.
- If the χ^2 value is smaller than the critical value, we fail to reject the null hypothesis, indicating no significant relationship.

Step 8: Interpret the Results

- Summarize the findings and explain whether the categorical variables are independent or dependent based on the test results.

3.5.3. Chi-Square Test of Independence: Example

Problem:

A researcher wants to determine whether there is a relationship between gender (Male/Female) and preferred teaching methods (Lecture, Discussion, Activity-Based) among students. The researcher surveys **110 students** and records their preferences:

Step 1: Define Hypotheses

- **Null Hypothesis (H_0):** There is no association between **Gender** and **Teaching Method Preference** (Lecture, Discussion, Activity-Based).
- **Alternative Hypothesis (Ha):** There is an association between **Gender** and **Teaching Method Preference**.

Step 2: Given Observed Frequency (O) Table

Gender	Lecture (f_{O1})	Discussion (f_{O2})	Activity-Based (f_{O3})	Row Total
Male	20	15	10	**45**
Female	25	30	10	**65**
Column Total	**45**	**45**	**20**	**110**

Step 3: Compute Expected Frequencies (*fe*)

The expected frequency for each cell is calculated as:

$$f_e = \frac{(\text{Row Total} \times \text{Column Total})}{\text{Grand Total}}$$

- **Row Total** = Sum of observations in that row
- **Column Total** = Sum of observations in that column
- **Grand Total** = Total number of observations

Expected Frequencies Table (f_e)

Gender	Lecture (E_1)	Discussion (E_2)	Activity-Based (E_3)
Male	$\frac{(45\times45)}{110}$=18.41	$\frac{(45\times45)}{110}$=18.41	$\frac{(45\times20)}{110}$=8.18
Female	$\frac{(65\times45)}{110}$=26.59	$\frac{(65\times45)}{110}$= 26.59	$\frac{(65\times20)}{110}$=11. 82

Step 4: Compute Chi-Square Statistic

The Chi-Square formula is:

$$\chi2 = \sum \frac{(fo - fe)2}{fe}$$

Calculating chi square statistics for Each Cell

<u>1. For Male - Lecture:</u>

$$\frac{(20-18.41)2}{18.41} = \frac{(1.59)2}{18.41} = \frac{2.53}{18.41} = 0.137$$

2. For Male - Discussion:

$$\frac{(15-1\ \ .41)2}{18.41} = \frac{(-3.41)2}{18.41} = \frac{11.63}{18.41} = 0.632$$

3. For Male - Activity-Based:

$$\frac{(10-8.18)2}{8.18} = \frac{(1.82)2}{8.18} = \frac{3.31}{8.18} = 0.405$$

4. For Female - Lecture:

$$\frac{(25-2\ \ .59)2}{26.59} = \frac{(-1.59)2}{26.59} = \frac{2.53}{26.59} = 0.095$$

5. For Female - Discussion:

$$\frac{(30-26.59)2}{26.59} = \frac{(3.41)2}{26.59} = \frac{11.63}{26.59} = 0.437$$

6. For Female - Activity-Based:

$$\frac{(10-11.82)2}{11.82} = \frac{(-1.82)2}{11.82} = \frac{3.31}{11.82} = 0.280$$

Step 5: Compute Total Chi-Square Value

$\chi 2$=0.137+0.632+0.405+0.095+0.437+0.280=1.986

Step 6: Determine Degrees of Freedom

- df = (Rows - 1)x(Columns - 1)
- df=(2−1)×(3−1)=1×2=2

Step 7: Compare with Critical Value

- Using a **Chi-Square table**, at **df = 2** and **α = 0.05**, the critical value is **5.991**.
- Since calculated **χ^2 = 1.986 < 5.991(less than 5.991)**, we fail to reject the null hypothesis(H_0), meaning there is **no**

significant association between gender and teaching method preference.

Conclusion: Since the Chi-Square value is smaller than the critical value, we conclude that gender and teaching method preference are independent, meaning that a student's gender does not significantly influence their choice of teaching method.

Chi Square Test of Independence

- Uses the frequency data from a sample to evaluate the relationship between two variables in the population.
- Test of Independence deals with questions like -
- Are two variables of interest independent of each other?
- Examples: Is starting salary of fresh graduates independent of graduates' field of study?
- Is job preference independent of the gender of the students?
- Variables independent of each other when there is no consistent, predictable relationship between two variables
- Each individual in the sample is classified on both of the two variables, creating a two-dimensional frequency-distribution matrix.

Check Your Progress -1

1. What are the main uses of the Chi-Square test?
2. Explain the difference between the Goodness of Fit test and the Test of Independence.

3.6 Chi-Square Test of Equality (Homogeneity):

The Chi-Square Test of Homogeneity **is a** statistical test used to determine whether different populations have the same distribution of a categorical variable. It helps in assessing whether the proportions of categories in multiple groups are uniform (homogeneous) or significantly different from each other. This test can also be used to test whether the occurance of events follow uniformity or not e.g. the admission of patients in government hospital in all days of week is uniform or not can be tested with the help of chi square test. In a chi-square test for homogeneity of proportions, we test the claim that different populations have the same proportion of individuals with some characteristic.

1. Purpose of the Test

The test is used to compare two or more independent groups to see if they have the same distribution of a categorical variable. It differs from the Chi-Square Test of Independence, which checks for an association between two categorical variables within a single population.

2. Assumptions of the Test

1. **Categorical Data**: The variable being analyzed must be categorical (nominal or ordinal).
2. **Independent Samples**: The groups being compared should be independent of each other.
3. **Random Sampling**: Data should be collected through random sampling to ensure validity.
4. **Expected Frequency Rule**: Each expected frequency should be at least 5 for the Chi-Square approximation to be valid.

3. Example of Chi-Square Test of Homogeneity

A researcher wants to compare the preference for three teaching methods (Lecture, Discussion, Activity-Based) **across three different schools** to determine if students' preferences are the same across these schools.

1. The researcher collects data and organizes it into a contingency table showing how many students from each school prefer each method.
2. The expected frequencies are calculated.
3. The Chi-Square formula is applied to compute **χ2.**
4. The test statistic is compared with the critical value to determine if the distribution of preferences is the same across the schools.

Conclusion

- If the test **fails to reject** the null hypothesis, it suggests that student preferences for teaching methods **do not significantly differ** across the schools, meaning the distribution is homogeneous.
- If the null hypothesis is **rejected**, it indicates that student preferences **vary significantly** across the schools, meaning the distributions are not homogeneous.

3.6.1. **Independence vs. Homogeneity:**

The distinction between the Chi-Square Test of Independence and the Chi-Square Test of Homogeneity is important. If the column totals (e.g., the number of boys and girls) are predetermined, the test is one of homogeneity, as it compares distributions across pre-defined groups. However, if these totals are not fixed, the test is one of independence. For example, examining whether there is a relationship between gender and career preference among students is a test of independence. Here, the null hypothesis would be: ***"There is no relationship between gender and ice cream flavor preference."*** *But in case of* homogeneity test the null hypothesis would be: ***"Boys and girls have the same ice cream flavor preferences***." The statistical test then assesses whether there is sufficient evidence to reject this assumption.

Difference between Chi-Square Test of Homogeneity and Independence

Feature	Chi-Square Test of Homogeneity	Chi-Square Test of Independence
Purpose	Compares distributions of a categorical variable across different groups	Examines whether two categorical variables are related
Number of Populations	Two or more	One
Hypothesis	Tests if distributions are the same across groups	Tests if there is an association between variables
Data Collection	Different samples from different populations	Single sample with two variables
Example	Checking if the preference for a detergent brand is the same across three different cities.	Checking if gender and preference for a teaching method are related.

Key Points to Remember:

- The Test of Homogeneity compares distributions across multiple groups to see if they are the same.
- The Test of Independence examines whether two categorical variables within a single population are related.

3.6.2 Chi-Square Test of Homogeneity Example:

A researcher wants to determine whether the preference for a new brand of product is the same across three cities (**A, B, and C**). The researcher collects the following data:

City	Prefer	Neutral	Do Not Prefer	Total
A	30	20	50	100
B	40	25	35	100
C	25	30	45	100
Total	95	75	130	300

By calculating the Chi-Square statistic, we determine if detergent preference is homogeneous across the cities.

We will now perform the **Chi-Square Test of Homogeneity** following these steps:

Step 1: State the Hypotheses

- **Null Hypothesis (H_0)**: The preference for the detergent is **the same** across all three cities (i.e., the distribution of responses is homogeneous).
- **Alternative Hypothesis (H_A)**: The preference for the detergent is **different** across the cities (i.e., the distribution of responses is not homogeneous).

Step 2: Compute the Expected Frequencies

The expected frequency (fe) for each cell is calculated using the formula:

Using this formula, we calculate the expected frequencies:

Using the formula:

$$fe = \frac{\text{Row Total} \times \text{Column Total}}{\text{Grand Total}}$$

The expected frequency table is:

City	**Preference**			**Total**
	Prefer	**Neutral**	**Do Not Prefer**	
A	$fe = \frac{100\times95}{300} = \frac{9500}{300}$=31.67	$fe = \frac{100\times75}{300} = \frac{7500}{300}$=25.00	$fe = \frac{100\times130}{300} = \frac{13000}{300}$=43.33	100
B	$fe = \frac{100\times95}{300} = \frac{9500}{300}$=31.67	$fe = \frac{100\times75}{300} = \frac{7500}{300}$=25.00	$fe = \frac{100\times130}{300} = \frac{13000}{300}$=43.33	100
C	$fe = \frac{100\times95}{300} = \frac{9500}{300}$=31.67	$fe = \frac{100\times75}{300} = \frac{7500}{300}$=25.00	$fe = \frac{100\times130}{300} = \frac{13000}{300}$=43.33	100
Total	**95**	**75**	**130**	**300**

Step 3: Compute the Chi-Square Statistic

The **Chi-Square statistic (χ2)** is calculated using the formula:

1. For City A

Prefer $= \frac{(30-31.67)2}{31.67} = \frac{(-1.67)2}{31.67} = \frac{2.79}{20.25} = 0.088$

Neutral$= \frac{(20-25)2}{25} = \frac{(-5)2}{25} = \frac{25}{25} = 1.00$

$$\textbf{Do Not Prefer} = \frac{(50-43.33)2}{43.33} = \frac{(6.67)2}{43.33} = \frac{44.49}{43.33} = 1.03$$

2. For City B

$$\textbf{Prefer} = \frac{(40-31.67)2}{31.67} = \frac{(8.33)2}{31.67} = \frac{69.39}{31.67} = 2.19$$

$$\textbf{Neutral} = \frac{(25-25)2}{25} = \frac{0}{25} = 0$$

$$\textbf{Do Not Prefer} = \frac{(35-43.33)2}{43.33} = \frac{(-8.33)2}{43.33} = \frac{69.39}{43.33} = 1.60$$

3. For City C

$$\textbf{Prefer} = \frac{(25-31.67)2}{31.67} = \frac{(-6.67)2}{31.67} = \frac{44.49}{31.67} = 1.41$$

$$\textbf{Neutral} = \frac{(30-25)2}{25} = \frac{(5)2}{25} = \frac{25}{25} = 1.00$$

$$\textbf{Do Not Prefer} = \frac{(45-43.33)2}{43.33} = \frac{(1.67)2}{43.33} = \frac{2.79}{43.33} = 0.064$$

Step 4: Calculate Total Chi-Square Value

Summing these values:

$$\chi 2 = \sum \left[\frac{(fo - fe)^2}{fe} \right]$$

χ2=0.088+1.00+1.03+2.19+0+1.60+1.41+1.00+0.064=8.39

Step 5: Determine the Degrees of Freedom

The formula for degrees of freedom (df) is:

df=(r−1)×(c−1)

Where:

- **r**= Number of rows (cities) = 3
- **c** = Number of columns (preference categories) = 3

Step 6: Compare with Critical Value

Using a **Chi-Square table** at a significance level (alpha-α) of 0.05 and **df = 4**, the **critical value is 9.49**.

- If $\chi 2 > 9.49$, reject H_0 (preferences are significantly different).
- If $\chi 2 < 9.49$, fail to reject H_0 (preferences are homogeneous).

Since our calculated $\chi 2 = 8.39$ is **less than** 9.49, we **fail to reject H_0.**

Step 7: Conclusion

Since the Chi-Square value (8.39) is less than the critical value (9.49), we fail to reject the null hypothesis. This means that there is no significant difference in detergent preference across the three cities. The preference distribution is homogeneous.

Thus, the Chi-Square Test of Homogeneity is useful for comparing multiple groups to determine if they share a common distribution for a categorical variable. This test is used to compare multiple groups to see if they follow the same distribution.

Check Your Progress -2

1. What is a chi square test of homogeneity?
2. Explain the difference between the test of Homogeneity and the Test of Independence
3. How Do You Interpret a Significant Chi-Square Result?

3.7 Summing Up:

The Chi-Square test is a fundamental statistical tool used to analyze categorical data. It helps researchers assess whether observed frequencies significantly differ from expected ones. Understanding the different applications of the test – **Goodness of Fit, Test of Equality, and Test of Independence** – enhances our ability to make data-driven decisions in various fields.

This module has equipped you with the theoretical understanding and practical steps to apply the Chi-Square test effectively. To deepen your understanding, try solving real-life problems and analyzing the results critically.

3.8 Answer to Check Your Progress:

Check your progress -1

1. Main Uses of the Chi-Square Test

The **Chi-Square test** is a statistical method used to determine whether there is a significant association between categorical variables. The main uses of the Chi-Square test include:

a) Test of Goodness of Fit

- Determines if a sample distribution matches an expected distribution.
- Used when a researcher wants to compare observed data with a theoretical model.
- Example: Checking if the number of students preferring different teaching methods follows an expected distribution.

b) Test of Independence

- Determines if two categorical variables are independent or related.
- Used to examine relationships between two variables in a contingency table.
- Example: Checking if gender and preference for a teaching method are related.

c) Test of Homogeneity

- Similar to the test of independence but applied to different populations.
- Used to check if the distribution of one categorical variable is the same across multiple groups.
- Example: Comparing detergent brand preferences across different cities.

2. The **Goodness of Fit Test** is used to compare a sample distribution with an expected distribution, while the **Test of Independence** examines the relationship between two categorical variables.

Check Your Progress -2

1. Chi-Square Test of Homogeneity

The **Chi-Square Test of Homogeneity** is a statistical test used to determine whether the distribution of a categorical variable is the same across multiple independent groups or populations. It helps compare different groups to see if they share similar characteristics regarding a categorical variable.

Key Features:

- Used when comparing two or more independent populations.
- Checks if the proportions of a categorical variable are the same across different groups.
- Requires a contingency table to compare observed and expected frequencies.

Difference between test of Homogeneity and Independence

- The Test of Homogeneity compares distributions across multiple groups to see if they are the same.

- The Test of Independence examines whether two categorical variables within a single population are related.

3. Interpretation of significant Chi-Square result

When the **Chi-Square test** result is significant (i.e., p-value < chosen significance level, such as 0.05), it means:

- **For a Test of Independence:** There is a relationship between the two categorical variables, meaning they are not independent. Example: If gender and preferred teaching method are significantly associated, we conclude that gender influences the choice of teaching method.
- **For a Test of Goodness of Fit:** The observed data does not match the expected distribution. Example: If a company's expected sales distribution of three product variants does not match actual sales, adjustments may be needed.
- **For a Test of Homogeneity:** The distribution of the categorical variable is **not the same** across different groups. Example: If detergent preference differs significantly across three cities, companies may need to use region-specific marketing strategies.

A significant Chi-Square test tells us that there is a difference, but it does not indicate the strength or direction of the association. Post-hoc tests may be required for further analysis.

3.9 Questions and Exercises:

1. What are the main uses of the Chi-Square test?
2. Solve a practical problem using the Chi-Square formula and interpret the results.
3. Select a research paper that applies the Chi-Square test and critically analyze how the test was used. Summarize the key findings and discuss whether the conclusions drawn are justified.

UNIT- 4

TESTING NULL HYPOTHESIS OF INDEPENDENCE IN 2×2 CONTINGENCY TABLE

Unit Structure:

4.1 Introduction

4. 2 Objectives

4.3 2×2- Contingency Table

4.4 Example: Calculation of Chi Square Test of Independence in 2 X2 Contingency Table

4.5 Yates' Correction *(For Small Sample Sizes)*

4.6 Example: Yates' Correction *(For Small Sample Sizes)*

4.7 Summing Up

4.8 Answers to Check Your Progress

4.9 Questions and Exercises

4.10 References and Suggested Readings

4.1 Introduction:

In the previous units, we explored the concept of the Chi-Square test, its assumptions, its significance and its various applications in statistical analysis. Now, we focus on one of its most important uses: testing the independence of two categorical variables using a 2×2 contingency table.

4.2 Objectives:

By the end of this unit, you will be able to:

- Construct a 2×2 contingency table and compute expected frequencies.
- Formulate and test the null hypothesis of independence

- Use the Chi-Square test of independence to analyze relationships.
- Interpret the results of the Chi-Square test and apply them to real-world situations.

4.3 2×2 Contingency Table:

A contingency table is used to examine the relationship between two categorical variables. A 2×2 contingency table is a simple way to examine the relationship between two categorical variables, each having only two categories. The 2 X 2 contingency chi-square is used for the comparison of two groups with a dichotomous dependent variable. For instance we might compare males and females on a yes/no response scale. Thus, a 2×2 contingency table presents the frequency distribution of two categorical variables. It is structured as follows:

	Category 1	**Category 2**	**Row Total**
Group 1	a	b	a + b
Group 2	c	d	c + d
Column Total	a + c	b + d	N (Grand Total)

Where:

- **a, b, c, and d** are the observed frequencies (actual counts).
- **N** is the total number of observations
- N = (a+b+c+d).

The Chi-Square test of independence helps determine whether these two variables are statistically independent or whether an association exists between them. Thus, we use the Chi-Square test of independence to determine whether these two variables are related or if their association is due to random chance.

- **Independence:** Two variables are independent if the occurrence of one does not affect the occurrence of the other.
- **Dependence:** If a relationship exists, meaning that changes in one variable correspond to changes in another, the variables are dependent.

2. Formulating the Null and Alternative Hypothesis

When testing for independence, we set up the following hypotheses:

- **Null Hypothesis (H_0):** There is no association between the two categorical variables (they are independent).
- **Alternative Hypothesis (H_1):** There is an association between the two categorical variables (they are dependent).

3. Expected Frequencies Calculation

The contingency chi-square is based on the same principles as the simple chi-square analysis in which we examine the expected vs. the observed frequencies. The computation is quite similar, except that the estimate of the expected frequency is a little harder to determine.

Just like other tests for independence using the Chi-Square test, the same procedure is followed in the case of a 2×2 contingency table. To test the null hypothesis, we compare the observed frequencies (f_o) with the expected frequencies (*fe*) in each cell of the table. The expected frequency (f_e) for each cell is calculated using the formula:

$$fe = \frac{\text{Row Total} \times \text{Column Total}}{\text{Grand Total}}$$

This formula ensures that we estimate the expected count for each category under the assumption that the two variables are independent. If the observed and expected frequencies differ significantly, we reject the null hypothesis, indicating a relationship between the two variables.

The expected frequency (fe) for each cell is calculated as:

For each cell in the table:

$fe_{11}= \frac{(a+b)(b+c)}{N}$ $fe_{12}= \frac{(a+b)(b+d)}{N}$

$fe_{21}= \frac{(c+d)(a+c)}{N}$ $fe_{21}= \frac{(c+d)(b+d)}{N}$

Step 3: Computing the Chi-Square Statistic

$$\chi 2=\sum [\frac{(fo-fe)^2}{fe}]$$

Where:

- f_o = observed frequency in each cell
- fe = expected frequency in each cell

Short method or Alternate formula to find out chi square value:

When the contingency table is 2x2 fold, the chi-square may be calculated without first calculating the four expected frequencies. In this case following formula is used to find out chi square directly.

$$\chi 2= \frac{N(ad-bc)^2}{(a+c)(b+d)(a+b)(c+d)}$$

Step 4: Computing the degrees of freedom

For a **2×2 table**, the degrees of freedom (df) is:

- df=(rows−1) × (columns−1)=(2−1)×(2−1)=1

Step 5: Interpretation and conclusion

If the computed **χ^2 value is greater than the critical value**, we reject the null hypothesis,

Example Interpretation:

If a researcher finds **χ^2 = 6.63** at **df = 1**, and the critical value at **0.01 significance level** is **6.635**, we **fail to reject H_0** at 0.01 level but **reject H_0 at 0.05 level**, suggesting moderate evidence of association.

4.4 Example: Calculation of Chi Square Test of Independence in 2 X2 Contingency Table

Problem: A researcher wants to examine whether teaching method (Traditional vs. Activity-Based) influences student performance (Pass vs. Fail). The study collects data from 100 students and organizes it into a 2×2 contingency table:

Teaching Method	Pass	Fail	Total
Traditional	30	20	50
Activity-Based	40	10	50
Total	70	30	100

Step 1: Formulating the Hypotheses

- **Null Hypothesis (H_0):** Teaching method and students performance are independent (i.e., teaching method does not influence performance).
- **Alternative Hypothesis (Ha):** Teaching method and students performance are not independent (i.e., teaching method affects performance).

Step 2: The Chi-Square statistic is calculated using formula

$$\chi2=\frac{N(ad-bc)^2}{(a+c)(b+d)(a+b)(c+d)}$$

Teaching Method	Pass	Fail	Total
Traditional	30 (a)	20(b)	50 (a+ b)
Activity-Based	40(c)	10(d)	50(b+ c)
Total	70(a+c)	30(b+d)	100(N)

$\chi 2 = \frac{100(30x10-20x40)^2}{70x30x50x50)}$

$= \quad \frac{100(300-800)^2}{70x30x50x50)}$

$= \frac{100(-500)^2}{5250000}$

$= \frac{25000000}{5250000} = 4.762$

Step 4: Decision Rule and Interpretation

- Degrees of Freedom (df) = (rows - 1) × (columns - 1) = (2-1) × (2-1) = 1
- Critical value at $\alpha = 0.05$ for df = 1: 3.841
- Calculated $\chi^2 = 4.76$

Since **4.76 > 3.841**, we **reject the null hypothesis**.

Conclusion: There is a significant relationship between the teaching method and student performance. The activity-based method leads to better student performance compared to the traditional method.

Check Your Progress -1

A psychologist investigates whether exam anxiety (High vs. Low) affects exam results (Pass vs. Fail). The data from 200 students is:

Anxiety Level	Pass	Fail	Total
High Anxiety	50	50	100
Low Anxiety	90	10	100
Total	140	60	200

Following the same steps as above, conduct the Chi-Square test.

Check Your Progress -1

1. Which of the following assumptions is required for applying the 2×2 Chi-Square Test?
   a) The variables must be continuous
   b) The expected frequency in each cell should be greater than 5
   c) The samples must be dependent
   d) The test is used for paired data
2. Write the formula to directly calculate Chi square in2x2 contingency

4.5 Yates' Correction *(For Small Sample Sizes)*

If in the 2*2 contingency table, the expected frequencies are small say less than 5, then2 x2 test cannot be used. In that case, the direct formula of the chi square test is modified and given by Yate's correction for continuity. The correct rewritten formula for the Chi-Square test of independence in a 2×2 contingency table using Yates' correction for continuity is:

$$\chi2 = \frac{N(|ad-bc|-0.5xN)^2}{\mathrm{R1R2C1C2}}$$

Where:

- N = Grand total (sum of all observations)
- a, b, c, d = Frequencies in the 2×2 contingency table
- R_1, R_2 = Row totals
- C_1, C_2 = Column totals

This correction is applied to improve the accuracy of the Chi-Square test when dealing with small sample sizes.

4.6 Example: Yates' Correction *(For Small Sample Sizes):*

Problem: A researcher wants to examine whether the mode of teaching (Online vs. Offline) affects student performance (Pass vs. Fail).

Step 1: Observed Frequencies (O)

The study collects data from 100 students and organizes it into a **2×2 contingency table**:

Teaching Mode	Pass (P)	Fail (F)	Total (R)
Online	40(a)	10(b)	50 (R_1)
Offline	30(c)	20(d)	50(R_2)
Total (C)	70(C_1)	30(C_2)	100(N)

Here:

- a=40 b=10 c = 30 d=20
- R_1=50 (Total for Online)
- R_2=50 (Total for Offline)
- C_1=70(Total Passed)
- C_2=30 (Total Failed)
- N=100 (Grand Total)

Step 2: Compute Chi-Square with Yates' Correction

Applying the formula:

$$\chi 2 = \frac{100(|40x20 - 10x30| - 0.5x100)^2}{50x50x70x30} \qquad \chi 2 = \frac{100(|800-300|-500)^2}{5250000}$$

$$\chi 2 = \frac{100(500-50)^2}{5250000} \qquad \chi 2 = \frac{100\,x450^2}{5250000}$$

$$\chi 2 = \frac{100x202500}{5250000} \qquad \chi 2 = \frac{20250000}{5250000}$$

$$\chi 2 = 3.857$$

Step 3: Compare with Critical Value

- Degrees of Freedom (df) = (rows - 1) × (columns - 1) = (2-1) × (2-1) = 1
- Critical value for χ^2 at $\alpha = 0.05$ (df = 1) = 3.841

Since **3.857 > 3.841**, we **reject the null hypothesis**.

Step 4: Conclusion

Since the calculated **χ^2 (3.857)** is greater than the critical value (3.841), we conclude that there is a significant relationship between teaching mode and student performance. This suggests that teaching mode (Online vs. Offline) has an impact on whether students pass or fail.

Key Takeaways

✓ Chi-Square Test of Independence checks if two categorical variables are related.

✓ Yates' Correction is applied for small sample sizes to adjust for continuity. For small sample sizes (N < 50), Yates' correction is used to prevent overestimation of the Chi-Square value.

- ✓ The test follows four steps: creating the contingency table, computing χ^2, comparing with the critical value, and drawing a conclusion.
- ✓ This correction slightly reduces the Chi-Square value, making the test more conservative.

Check Your Progress -2

1. Yates' Correction for continuity is applied to the Chi-Square test to:
   a) Increase the Chi-Square value
   b) Adjust for small sample sizes and reduce bias
   c) Convert categorical data into numerical data
   d) Improve the normality of the data distribution

2. When should Yates' Correction be applied in a 2×2 Chi-Square test?
   a) When all expected frequencies are greater than 10
   b) When at least one expected frequency is less than 5
   c) When the sample size is above 100
   d) When using a one-tailed hypothesis test

4.7 Summing Up:

- The **2×2 contingency table** helps analyze relationships between two categorical variables.
- The **Chi-Square test of independence** determines if two variables are **related** or **independent**.
- The **Chi-Square test for independence** compares **observed vs. expected frequencies** to check for associations.

- **Expected frequencies** help compare observed data with what is expected under independence.
- **Hypothesis testing** determines whether variables are **independent or dependent**.
- The **decision to reject or accept H_0** is based on comparing **χ^2 calculated** with the **χ^2 critical value**
- **Yates' correction** is used for small sample sizes to adjust for continuity.
- **Decision-making** is based on comparing the **calculated χ^2** value with the **critical value** from the table.

4.8 Answers to Check your Progress:

Check Your Progress-1

1. The expected frequency in each cell should be greater than 5
2. $\chi 2 = \frac{N(ad-bc)^2}{(a+c)(b+d)(a+b)(c+d)}$

Check Your Progress-2

1. Adjust for small sample sizes and reduce bias
2. When at least one expected frequency is less than 5

4.9 Questions and Exercises:

1. What is the purpose of a 2×2 contingency table?
2. What is the formula to calculate expected frequency in a 2×2 contingency table?
3. How many degrees of freedom does a 2×2 contingency table have?
4. Why do we apply Yates' correction in small samples?
5. When should Yates' correction be applied?

6. If a Chi-Square test result shows χ^2 calculated = 4.2 and the critical value at $\alpha = 0.05$ is 3.84, what should you conclude?
7. Construct a 2×2 contingency table using hypothetical data for gender (Male/Female) and preference for online vs. offline shopping. Calculate expected frequencies.
8. Given the following observed data, test the null hypothesis of independence at **$\alpha = 0.05$** using the Chi-Square test:

	Prefers Online	**Prefers Offline**	**Total**
Male	30	20	50
Female	40	60	100
Total	70	80	150

- Formulate hypotheses
- Compute expected frequencies
- Calculate the **Chi-Square statistic**
- Compare with the **critical value at df = 1**
- Interpret the result

4.10 References and Suggested Readings

1. Agresti, A. (2018). *Statistical Methods for the Social Sciences* (5th ed.). Pearson Education.
2. Gupta, S. P. (2021). *Statistical Methods* (45th ed.). Sultan Chand & Sons.
3. Kothari, C. R., & Garg, G. (2019). *Research Methodology: Methods and Techniques* (4th ed.). New Age International Publishers.
4. Levin, R. I., & Rubin, D. S. (2020). *Statistics for Management* (8th ed.). Pearson India.

5. Siegel, S., & Castellan, N. J. (1988). *Nonparametric Statistics for the Behavioral Sciences.* McGraw-Hill.

6. Srivastava, U. K., & Sharma, G. V. (2021). *Quantitative Techniques for Managerial Decisions*. New Age International Publishers.

7. Verma, J. P. (2013). *Data Analysis in Management with SPSS Software*. Springer.

8. Yamane, T. (1967). *Statistics: An Introductory Analysis* (2nd ed.). Harper & Row.

9. Zikmund, W. G., Babin, B. J., Carr, J. C., & Griffin, M. (2021). *Business Research Methods* (10th ed.). Cengage Learning.

*****

APPENDICES

Appendix-1

Table 'S' of Square and Square Roots for the Nos. 1 to 500

Number	*Square*	*S.R.*	*Number*	*Square*	*S.R.*
1	1	1.0000	35	1225	5.9161
2	4	1.4142	36	1296	6.0000
3	9	1.7321	37	1369	6.0828
4	16	2.0000	38	1444	6.1644
5	25	2.2361	39	1521	6.2450
6	36	2.4495	40	1600	6.3246
7	49	2.6458	41	1681	6.4031
8	64	2.8284	42	1764	6.4807
9	81	3.0000	43	1849	6.5574
10	100	3.1623	44	1936	6.6332
11	121	3.3166	45	2025	6.7082
12	144	3.4641	46	2116	6.7823
13	169	3.6056	47	2209	6.8557
14	196	3.7417	48	2304	6.9282
15	225	3.8730	49	2401	7.0000
16	256	4.0000	50	2500	7.0711
17	289	4.1231	51	2601	7.1414
18	324	4.2426	52	2704	7.2111
19	361	4.3589	53	2809	7.2801
20	400	4.4721	54	2916	7.3485
21	441	4.5826	55	3025	7.4162
22	484	4.6904	56	3136	7.4833
23	529	4.7958	57	3249	7.5498
24	576	4.8990	58	3364	7.6158
25	625	5.0000	59	3481	7.6811
26	676	5.0990	60	3600	7.7460
27	729	5.1962	61	3721	7.8102
28	784	5.2915	62	3844	7.8740
29	841	5.3852	63	3669	7.9373
30	900	5.4772	64	4096	8.0000
31	961	5.5678	65	4225	8.0623
32	1024	5.6569	66	4356	8.1240
33	1089	5.7446	67	4489	8.1854
34	1156	5.8310	68	4624	8.2462

S.R. = Square Root

Table 'S' - (Continued)

Number	*Square*	*S.R.*	*Number*	*Square*	*S.R.*
69	4761	8.3066	107	11449	10.3441
70	4900	8.3666	108	11664	10.3923
71	5041	8.4261	109	11881	10.4403
72	5184	8.4853	110	12100	10.4881
73	5329	8.5440	111	12321	10.5357
74	5476	8.6023	112	12544	10.5830
75	5625	8.6603	113	12769	10.6301
76	5776	8.7178	114	12996	10.6771
77	5929	8.7750	115	13225	10.7238
78	6084	8.8318	116	13456	10.7703
79	6241	8.8882	117	13689	10.8167
80	6400	8.9443	118	13924	10.8628
81	6561	9.0000	119	14161	10.9087
82	6724	9.0554	120	14400	10.9545
83	6889	9.1104	121	14641	11.0000
84	7056	9.1652	122	14884	11.0454
85	7225	9.2195	123	15129	11.0905
86	7396	9.2736	124	15376	11.1355
87	7569	9.3274	125	15625	11.1803
88	7744	9.3808	126	15876	11.2250
89	7921	9.4340	127	16129	11.2694
90	8100	9.4868	128	16384	11.3137
91	8281	9.5394	129	16641	11.3578
92	8464	9.5917	130	16900	11.4018
93	8649	9.6437	131	17161	11.4455
94	8836	9.6954	132	17424	11.4891
95	9025	9.7468	133	17689	11.5326
96	9216	9.7980	134	17956	11.5758
97	9409	9.8489	135	18225	11.6190
98	9604	9.8995	136	18496	11.6619
99	9801	9.9499	137	18769	11.7047
100	10000	10.0000	138	19044	11.7473
101	10201	10.0499	139	19321	11.7898
102	10404	10.0995	140	19600	11.8322
103	10609	10.1489	141	19881	11.8743
104	10816	10.1980	142	20164	11.9164
105	11025	10.2470	143	20449	11.9583
106	11236	10.2956	144	20736	12.0000

Table 'S' - (Continued)

Number	*Square*	*S.R.*	*Number*	*Square*	*S.R.*
145	21025	12.0416	183	33489	13.5277
146	21316	12.0830	184	33856	13.5647
147	21609	12.1244	185	34225	13.6015
148	21904	12.1655	186	34596	13.6382
149	22201	12.2066	187	34969	13.6748
150	22500	12.2474	188	35344	13.7113
151	22801	12.2882	189	35721	13.7477
152	23104	12.3288	190	36100	13.7840
153	23409	12.3693	191	36481	13.8203
154	23716	12.4097	192	36864	13.8564
155	24025	12.4499	193	37249	13.8924
156	24336	12.4900	194	37636	13.9284
157	24649	12.5300	195	38025	13.9642
158	24964	12.5698	196	38416	14.0000
159	25281	12.6095	197	38809	14.0357
160	25600	12.6491	198	39204	14.0712
161	25921	12.6886	199	39601	14.1067
162	26244	12.7279	200	40000	1401421
163	26569	12.7671	201	40401	14.1774
164	26896	12.8062	202	40804	14.2127
165	27225	12.8452	203	41209	14.2478
166	27556	12.8841	204	41616	14.2829
167	27889	12.9228	205	42025	14.3178
168	28224	12.9615	206	42436	14.3527
169	28569	13.0000	207	42849	14.3875
170	28900	13.0384	208	43264	14.4222
171	29241	13.0767	209	43681	14.4568
172	29584	13.1149	210	44100	14.4914
173	29929	13.1529	211	44521	14.5258
174	30276	13.1909	212	44944	14.5602
175	30625	13.2288	213	45369	14.5945
176	30976	13.2665	214	45796	14.6287
177	31329	13.3041	215	46225	14.6629
178	31684	13.3417	216	46656	14.6969
179	32041	13.3791	217	47089	14.7309
180	32400	13.4164	218	47524	14.7648
181	32761	13.4536	219	47961	14.7986
182	33124	13.4907	220	48400	14.8324

Table 'S' - (Continued)

Number	*Square*	*S.R.*	*Number*	*Square*	*S.R.*
221	48841	14.8661	259	67081	16.0935
222	49284	14.8997	260	67600	16.1245
223	49729	14.9332	261	68121	16.1555
224	50176	14.9666	262	68644	16.1864
225	50625	15.0000	263	69169	16.2173
226	51076	15.0333	264	69696	16.2481
227	51529	15.0665	265	70225	16.2788
228	51984	15.0997	266	70756	16.3095
229	52441	15.1327	267	71289	16.3401
230	52900	15.1658	268	71824	16.3707
231	53361	15.1987	269	72361	16.4012
232	53824	15.2315	270	72900	16.4317
233	54289	15.2643	271	73441	16.4621
234	54756	15.2971	272	73984	16.4924
235	55225	15.3297	273	74529	16.5227
236	55696	15.3623	274	75076	16.5529
237	56169	15.3948	275	75625	16.5831
238	56644	15.4272	276	76176	16.6132
239	57121	15.4596	277	76729	16.6433
240	57600	15.4919	278	77284	16.6733
241	58081	15.5242	279	77841	16.7033
242	58564	15.5563	280	78400	16.7332
243	59049	15.5885	281	78961	16.7631
244	59536	15.6205	282	79524	16.7929
245	60025	15.6525	283	80089	16.8226
246	60516	15.6844	284	80656	16.8523
247	61009	15.7162	285	81225	16.8819
248	61504	15.7480	286	81796	16.9115
249	62001	15.7797	287	82369	16.9411
250	62500	15.8114	288	82944	16.9706
251	63001	15.8430	289	83521	17.0000
252	63504	15.8745	290	84100	17.0294
253	64009	15.9060	291	84681	17.0587
254	64516	15.9374	292	85264	17.0880
255	65025	15.9687	293	85849	17.1172
256	65536	16.0000	294	86436	17.1464
257	66049	16.0312	295	87025	17.1756
258	66564	16.0624	296	87616	17.2047

Table 'S' - (Continued)

Number	*Square*	*S.R.*	*Number*	*Square*	*S.R.*
297	88209	17.2337	335	112225	18.3030
298	88804	17.2627	336	112896	18.3303
299	89401	17.2916	337	113569	18.3576
300	90000	17.3205	338	114244	18.3848
301	90601	17.3494	339	114921	18.4120
302	91204	17.3781	340	115600	18.4391
303	91809	17.4069	341	116281	18.4662
304	92416	17.4356	342	116964	18.4932
305	93025	17.4642	343	117649	18.5203
306	93636	17.4929	344	118336	18.5472
307	94249	17.5214	345	119025	18.5742
308	94864	17.5499	346	119716	18.6011
309	95481	17.5784	347	120409	18.6279
310	96100	17.6068	348	121104	18.6548
311	96721	17.6352	349	121801	18.8615
312	97344	17.6635	350	122500	18.7083
313	97969	17.6918	351	123201	18.7350
314	98596	17.7200	352	123904	18.7617
315	99225	17.7482	353	124609	18.7883
316	99856	17.7764	354	125316	18.8149
317	100489	17.8045	355	126025	18.8414
318	101124	17.9326	356	126736	18.8680
319	101761	17.8606	357	127449	18.8944
320	102400	17.8885	358	128164	18.9209
321	103041	17.9165	359	128881	18.9473
322	103684	17.9444	360	129600	18.9737
323	104329	17.9722	361	130321	19.0000
324	104976	18.0000	362	131044	19.0263
325	105625	18.0278	363	131769	19.0526
326	106276	18.0555	364	132496	19.0788
327	106929	18.0831	365	133225	19.1050
328	107584	18.1108	366	133956	19.1311
329	108241	18.1384	367	134689	19.1572
330	108900	18.1659	368	135424	19.1833
331	109561	18.1934	369	136161	19.2094
332	110224	18.2209	370	136900	19.2354
333	110889	18.2483	371	137641	19.2614
334	111556	18.2757	372	138384	19.2873

Table 'S' - (Continued)

Number	*Square*	*S.R.*	*Number*	*Square*	*S.R.*
373	139129	19.3132	411	168921	20.2731
374	139876	19.3391	412	169744	20.2978
375	140625	19.3649	413	170569	20.3224
376	141376	19.3907	414	171396	20.3470
377	142129	19.4165	415	172225	20.3715
378	142884	19.4422	416	173056	20.3961
379	143641	19.4679	417	173889	20.4206
380	144400	19.4936	418	174724	20.4450
381	145161	19.5192	419	175561	20.4695
382	145924	19.5448	420	176400	20.4939
383	146689	19.5704	421	177241	20.5183
384	147456	19.5959	422	178284	20.5426
385	148225	19.6214	423	178929	20.5670
386	148996	19.6469	424	179776	20.5913
387	149769	19.6723	425	180625	20.6155
388	150544	19.6977	426	181476	20.6398
389	151321	19.7231	427	182329	20.6640
390	152100	19.7484	428	183184	20.6882
391	152881	19.7737	429	184041	20.7123
392	153664	19.7990	430	184900	20.7364
393	154449	19.8242	431	185761	20.7605
394	155236	19.8494	432	186624	20.7846
395	156025	19.8746	433	187489	20.8087
396	156816	19.8997	434	188356	20.8327
397	157609	19.9249	435	189225	20.8567
398	158404	19.9499	436	190096	20.8806
399	159201	19.9750	437	190969	20.9045
400	160000	20.0000	438	191844	20.9284
401	160801	20.0250	439	192721	20.9523
402	161604	20.0499	440	193600	20.9762
403	162409	20.0749	441	194481	21.0000
404	163216	20.0998	442	195364	21.0238
405	164025	20.1246	443	196249	21.0476
406	164836	20.1494	444	197136	21.0713
407	165649	20.1742	445	198025	21.0950
408	166464	20.1990	446	198916	21.1187
409	167281	20.2237	447	199809	21.1424
410	168100	20.2485	448	200704	21.1660

Table 'S' - (Concluded)

Number	*Square*	*S.R.*	*Number*	*Square*	*S.R.*
449	201601	21.1896	475	225625	21.7945
450	202500	21.2132	476	226576	21.8174
451	203401	21.2368	477	227529	21.8403
452	204304	21.2603	478	228484	21.8632
453	205209	21.2838	479	229441	21.8861
454	206116	21.3073	480	230400	21.9089
455	207025	21.3307	481	231361	21.9317
456	207936	21.3542	482	232324	21.9545
457	208849	21.3776	483	233289	21.9773
458	209764	21.4009	484	234256	22.0000
459	210681	21.4243	485	235225	22.0227
460	411600	21.4476	486	236196	22.0454
461	212521	21.4709	487	237169	22.0681
462	213444	21.4942	488	238144	22.0907
463	214369	21.5174	489	239121	22.1133
464	215296	21.5407	490	240100	22.1359
465	216225	21.5639	491	241081	22.1585
466	217156	21.5870	492	242064	22.1811
467	218089	21.6102	493	243049	22.2036
468	219024	21.6333	494	244036	22.2261
469	219961	21.6564	495	245025	22.2486
470	220900	21.6795	496	246016	22.2711
471	221841	21.7025	497	247009	22.2935
472	222784	21.7256	498	248004	22.3159
473	223729	21.7486	499	249001	22.3383
474	224676	21.7715	500	250000	22.3607

❖ ❖ ❖

Appendix-2

Table 'N' of Percentage of Area Under Normal Probability Curve

Percentage of area under the Normal Probability Curve falling between Mean and Successive points of Standard Deviation.

Sigma distance $\frac{x}{\sigma}$	*Percentage of Area* .00	*Sigma distance* $\frac{x}{\sigma}$	*Percentage of Area* .00
0.0	0.00	1.6	44.52
0.1	03.98	1.7	45.54
0.2	07.93	1.8	46.41
0.3	11.79	1.9	47.13
0.4	15.54	**2.0**	47.72
0.5	19.15	2.1	48.21
0.6	22.57	2.2	48.61
0.7	25.80	2.3	48.93
0.8	28.81	2.4	49.18
0.9	31.59	2.5	49.38
1.0	34.13	2.6	49.53
1.1	36.43	2.7	49.65
1.2	38.49	2.8	49.74
1.3	40.32	2.9	49.81
1.4	41.92	3.0	49.87
1.5	43.32		

Note: 0.13% of scores will occure after 3σ (upto ∞) (50-49.87=0.13). Exactly the same percent of scores will occure in the second half (part) of the curve.

Example - for sigma distance 1

34.13% scores will fall between

0 to $\pm 1\sigma$.

Appendix-3

Table 'T' Showing Significance of C.R./t-value
Degrees of Freedom = (N-2)

Degrees of Freedom df	Level .05	Level .01	Degrees of Freedom df	Level .05	Level .01
1	12.71	63.66	24	2.06	2.80
2	4.30	9.92	25	2.06	2.79
3	3.18	5.84	26	2.06	2.78
4	2.78	4.60	27	2.05	2.77
5	2.57	4.03	28	2.05	2.76
6	2.45	3.71	29	2.04	2.76
7	2.36	3.50	**30**	**2.04**	**2.75**
8	2.31	3.36	35	2.03	2.72
9	2.26	3.25	40	2.02	2.71
10	**2.23**	**3.17**	45	2.02	2.69
11	2.20	3.11	50	2.01	2.68
12	2.18	3.06	60	2.00	2.66
13	2.16	3.01	70	2.00	2.65
14	2.14	2.98	80	1.99	2.64
15	2.13	2.95	90	1.99	2.63
16	2.12	2.92	**100**	**1.98**	**2.63**
17	2.11	2.90	125	1.98	2.62
18	2.10	2.88	150	1.98	2.61
19	2.09	2.86	200	1.97	2.60
20	**2.09**	**2.84**	300	1.97	2.59
21	2.08	2.83	400	1.97	2.59
22	2.07	2.82	500	1.96	2.59
23	2.07	2.81	1000	1.96	2.58
			∞	1.96	2.58

Appendix-4

Table 'F' (For Significance of F-ratio)

Smaller Figures for 0.05 Level - Boldface for 0.0

Degrees of freedom for smaller mean square (Denominator)	Degrees of freedom for greater mean square (Numerator) 1	2	3	4	5	6	8	12	24	∞
1	161.45	199.50	215.72	224.57	230.17	233.97	238.89	243.91	249.04	254.32
	4052.10	4999.03	5403.49	5625.14	5764.08	5859.39	5981.34	6105.83	6234.16	6366.48
2	18.51	19.00	19.16	19.25	19.30	19.33	19.37	19.41	19.45	19.50
	98.49	99.01	99.17	99.25	99.30	99.33	99.36	99.42	99.46	99.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.84	8.74	8.64	8.53
	34.12	30.81	29.46	28.71	28.24	27.91	27.49	27.05	26.60	26.12
4	7.71	6.94	6.59	6.39	6.26	6.16	6.04	5.91	5.77	5.63
	21.20	18.00	16.69	15.98	15.52	15.21	14.80	14.37	13.93	13.46
5	6.61	5.79	5.41	5.19	5.05	4.95	4.82	4.68	4.53	4.36
	16.26	13.27	12.06	11.39	10.97	10.67	10.27	9.89	9.47	9.02
6	5.99	5.14	4.76	4.53	4.39	4.28	4.15	4.00	3.84	3.67
	13.74	10.92	9.78	9.15	8.75	8.47	8.10	7.72	7.31	6.88

Example for df 4/11

Significant value of F-ratio at 0.05 level = 3.36

Significant value of F-ratio at 0.01 level = 5.67

Table F - (Continued)

Degrees of freedom for smaller mean square (Denominator)	Degrees of freedom for greater mean square (Numerator)									
	1	2	3	4	5	6	8	12	24	∞
7	5.59	4.74	4.35	4.12	3.97	3.87	3.73	3.57	3.41	3.23
	12.25	9.55	8.45	7.85	7.46	7.19	6.84	6.47	6.07	5.65
8	5.32	4.46	4.07	3.84	3.69	3.58	3.44	3.28	3.12	2.93
	11.26	8.65	7.59	7.01	6.63	6.37	6.03	5.67	5.28	4.86
9	5.12	4.26	3.86	3.63	3.48	3.37	3.23	3.07	2.90	2.71
	10.56	8.02	6.99	6.42	6.06	5.80	5.47	5.11	4.73	4.31
10	4.96	4.10	3.71	3.48	3.33	3.22	3.07	2.91	2.74	2.54
	10.04	7.56	6.55	5.99	5.64	5.39	5.06	4.71	4.33	3.91
11	4.84	3.98	3.59	3.36	3.20	3.09	2.95	2.79	2.61	2.40
	9.65	7.20	6.22	5.67	5.32	5.07	4.74	4.40	4.02	3.60
12	4.75	3.88	3.49	3.26	3.11	3.00	2.85	2.69	2.50	2.30
	9.33	6.93	5.95	5.41	5.06	4.82	4.50	4.16	3.78	3.36
13	4.67	3.80	3.41	3.18	3.02	2.92	2.77	2.60	2.42	2.21
	9.07	6.70	5.74	5.20	4.86	4.62	4.30	3.96	3.59	3.16
14	4.60	3.74	3.34	3.11	2.96	2.85	2.70	2.53	2.35	2.13
	8.86	6.51	5.56	5.03	4.69	4.46	4.14	3.80	3.43	3.00

Table F - (Continued)

Degrees of freedom for smaller mean square (Denominator)	Degrees of freedom for greater mean square (Numerator) 1	2	3	4	5	6	8	12	24	∞
15	4.54	3.68	3.29	3.06	2.90	2.79	2.64	2.48	2.29	2.07
	8.68	6.36	5.42	4.89	4.56	4.32	4.00	3.67	3.29	2.87
16	4.49	3.63	3.24	3.01	2.85	2.74	2.59	2.42	2.24	2.01
	8.53	6.23	5.29	4.77	4.44	4.20	3.89	3.55	3.18	2.75
17	4.45	3.59	3.20	2.96	2.81	2.70	2.55	2.38	2.19	1.96
	8.40	6.11	5.18	4.67	4.34	4.10	3.79	3.45	3.08	2.65
18	4.41	3.55	3.16	2.93	2.77	2.66	2.51	2.34	2.15	1.92
	8.28	6.01	5.09	4.58	4.25	4.01	3.71	3.37	3.01	2.57
19	4.38	3.52	3.13	2.90	2.74	2.63	2.48	2.31	2.11	1.88
	8.18	5.93	5.01	4.50	4.17	3.94	3.63	3.30	2.92	2.49
20	4.35	3.49	3.10	2.87	2.71	2.60	2.45	2.28	2.08	1.84
	8.10	5.85	4.94	4.43	4.10	3.87	3.56	3.23	2.86	2.42
21	4.32	3.47	3.07	2.84	2.68	2.57	2.42	2.25	2.05	1.81
	8.02	5.78	4.87	4.37	4.04	3.81	3.51	3.17	2.80	2.36
22	4.30	3.44	3.05	2.82	2.66	2.55	2.40	2.23	2.03	1.78
	7.94	5.72	4.82	4.31	3.99	3.75	3.45	3.12	2.75	2.30

Table F - (Continued)

Degrees of freedom for smaller mean square (Denominator)	Degrees of freedom for greater mean square (Numerator) 1	2	3	4	5	6	8	12	24	∞
23	4.28	3.42	3.03	2.80	2.64	2.53	2.38	2.20	2.00	1.76
	7.88	5.66	4.76	4.26	3.94	3.71	3.41	3.07	2.70	2.26
24	4.26	3.40	3.01	2.78	2.62	2.51	2.36	2.18	1.98	1.73
	7.82	5.61	4.72	4.22	3.90	3.67	3.36	3.03	2.66	2.21
25	4.24	3.38	2.99	2.76	2.60	2.49	2.34	2.16	1.96	1.71
	7.77	5.57	4.68	4.18	3.86	3.63	3.32	2.99	2.62	2.17
26	4.22	3.37	2.98	2.74	2.59	2.47	2.32	2.15	1.95	1.69
	7.72	5.53	4.64	4.14	3.82	3.59	3.29	2.96	2.58	2.13
27	4.21	3.35	2.96	2.73	2.57	2.46	2.30	2.13	1.93	1.67
	7.68	5.49	4.60	4.11	3.78	3.56	3.26	2.93	2.55	2.10
28	4.20	3.34	2.95	2.71	2.56	2.44	2.29	2.12	1.91	1.65
	7.64	5.45	4.57	4.07	3.75	3.53	3.23	2.90	2.52	2.06
29	4.18	3.33	2.93	2.70	2.54	2.43	2.28	2.10	1.90	1.64
	7.60	5.42	4.54	4.04	3.73	3.50	3.20	2.87	2.49	2.03
30	4.17	3.32	2.92	2.69	2.53	2.42	2.27	2.09	1.89	1.62
	7.56	5.39	4.51	4.02	3.70	3.47	3.17	2.84	2.47	2.01

Table F - (Continued)

Degrees of freedom for smaller mean square (Denominator)	Degrees of freedom for greater mean square (Numerator)									
	1	2	3	4	5	6	8	12	24	∞
35	4.12	3.26	2.87	2.64	2.48	2.37	2.22	2.04	1.83	1.57
	7.42	5.27	4.40	3.91	3.59	3.37	3.07	2.74	2.37	1.90
40	4.08	3.23	2.84	2.61	2.45	2.34	2.18	2.00	1.79	1.52
	7.31	5.18	4.31	3.83	3.51	3.29	2.99	2.66	2.29	1.82
45	4.06	3.21	2.81	2.58	2.42	2.31	2.15	1.97	1.76	1.48
	7.23	5.11	4.25	3.77	3.45	3.23	2.94	2.61	2.23	1.75
50	4.03	3.18	2.79	2.56	2.40	2.29	2.13	1.95	1.74	1.44
	7.17	5.06	4.20	3.72	3.41	3.19	2.89	2.56	2.18	1.68
60	4.00	3.15	2.76	2.52	2.37	2.25	2.10	1.92	1.70	1.39
	7.08	4.98	4.13	3.65	3.34	3.12	2.82	2.50	2.12	1.60
70	3.98	3.13	2.74	2.50	2.35	2.23	2.07	1.89	1.67	1.35
	7.01	4.92	4.07	3.60	3.29	3.07	2.78	2.45	2.07	1.53
80	3.96	3.11	2.72	2.49	2.33	2.21	2.06	1.88	1.65	1.31
	6.96	4.88	4.04	3.56	3.26	3.04	2.74	2.42	2.03	1.47
90	3.95	3.10	2.71	2.47	2.32	2.20	2.04	1.86	1.64	1.28
	6.92	4.85	4.01	3.53	3.23	3.01	2.72	2.39	2.00	1.43

Table F - (Continued)

Degrees of freedom for smaller mean square (Denominator)	Degrees of freedom for greater mean square (Numerator)									
	1	2	3	4	5	6	8	12	24	∞
100	3.94	3.09	2.70	2.46	2.30	2.19	2.03	1.85	1.63	1.26
	6.90	4.82	3.98	3.51	3.21	2.99	2.69	2.37	1.98	1.39
125	3.92	3.07	2.68	2.44	2.29	2.17	2.01	1.83	1.60	1.21
	6.84	4.78	3.94	3.47	3.17	2.95	2.66	2.33	1.94	1.32
150	3.90	3.06	2.66	2.43	2.27	2.16	2.00	1.82	1.59	1.18
	6.81	4.75	3.91	3.45	3.14	2.92	2.63	2.31	1.92	1.27
200	3.89	3.04	2.65	2.42	2.26	2.14	1.98	1.80	1.57	1.14
	6.76	4.71	3.88	3.41	3.11	2.89	2.60	2.28	1.88	1.21
300	3.87	3.03	2.64	2.41	2.25	2.13	1.97	1.79	1.55	1.10
	6.72	4.68	3.85	3.38	3.08	2.86	2.57	2.24	1.85	1.14
400	3.86	3.02	2.63	2.40	2.24	2.12	1.96	1.78	1.54	1.07
	6.70	4.66	3.83	3.37	3.06	2.85	2.56	2.23	1.84	1.11
500	3.86	3.01	2.62	2.39	2.23	2.11	1.96	1.77	1.54	1.06
	6.69	4.65	3.82	3.36	3.05	2.84	2.55	2.22	1.83	1.08
1000	3.85	3.00	2.61	2.38	2.22	2.10	1.95	1.76	1.53	1.03
	6.66	4.63	3.80	3.34	3.04	2.82	2.53	2.20	1.81	1.04
∞	3.84	2.99	2.60	2.37	2.21	2.09	1.94	1.75	1.52	
	6.64	4.60	3.78	3.32	3.02	2.80	2.51	2.18	1.79	

Appendix-5

Areas of Normal Probability Curve - Z-Table

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990

Reading a Z table

Z tables are typically composed as follows:

- The label for rows contains the integer part and the first decimal place of Z.
- The label for columns contains the second decimal place of Z.
- The values within the table are the probabilities corresponding to the table type. These probabilities are calculations of the area under the normal curve from the starting point (0 for *cumulative from mean*, negative infinity for *cumulative* and positive infinity for *complementary cumulative*) to Z.

Example: To find **0.69**, one would look down the rows to find **0.6** and then across the columns to **0.09** which would yield a probability of **0.25490** for a *cumulative from mean* table or **0.75490** from a *cumulative* table.

Because the normal distribution curve is symmetrical, probabilities for only positive values of Z are typically given. The user has to use a complementary operation on the absolute value of Z, as in the example below.

Appendix-6

Table 'R' Significance of Co-efficient of Correlation (r/p) at 0.05 and 0.01 levels Degrees of Freedom = (N-2)

Degrees of Freedom	Level .05	Level .01	Degrees of Freedom	Level .05	Level .01
1	.997	1.000	24	.388	.496
2	.950	.990	25	.381	.487
3	.878	.959	26	.374	.478
4	.811	.917	27	.367	.470
5	.754	.874	28	.361	.463
6	.707	.834	29	.355	.456
7	.666	.798	30	.349	.449
8	.632	.765	35	.325	.418
9	.602	.735	40	.304	.393
10	.576	.708	45	.288	.372
11	.553	.684	50	.273	.354
12	.532	.661	60	.250	.325
13	.514	.641	70	.232	.302
14	.497	.623	80	.217	.283
15	.482	.606	90	.205	.267
16	.468	.590	100	.195	.254
17	.456	.575	125	.174	.228
18	.444	.561	150	.159	.208
19	.433	.549	200	.138	.181
20	.423	.537	300	.113	.148
21	.413	.526	400	.098	.128
22	.404	.515	500	.088	.115
23	.396	.505	1000	.062	.081

Appendix-7

Table 'C' Showing Significance of Chi-square (χ^2) Values

Degrees of Freedom df	Level .05	Level .01	Degrees of Freedom df	Level .05	Level .01
1	3.841	6.635	16	26.296	32.000
2	5.991	9.210	17	27.587	33.409
3	7.815	11.345	18	28.869	34.805
4	9.488	13.277	19	30.144	36.191
5	11.070	15.086	**20**	**31.410**	**37.566**
6	12.592	16.812	21	32.671	38.932
7	14.067	18.475	22	33.924	40.289
8	15.507	20.090	23	35.172	41.638
9	16.919	21.666	24	36.415	42.980
10	**18.307**	**23.209**	25	37.652	44.314
11	19.675	24.725	26	38.885	45.642
12	21.026	26.217	27	40.113	46.963
13	22.362	27.688	28	41.337	48.278
14	23.685	29.141	29	42.557	49.588
15	24.996	30.578	30	43.773	50.892

Appendix-8

Probability Values (t-Table)

This table enables the t-value from a t-test to be converted to a statement about significance.

- Select the column with probability that you want.
  - e.g. 0.05 means '95% chance'
- Select the row for degrees of freedom.
  - For two values, number of degrees of freedom is $(n_1 + n_2) - 2$
- Compare the value in the cell with your t-value.
- The results are significant if the t-value is *greater* than the value in the cell.

Degrees of Freedom	*Probability, p*		
	0.10	**0.05**	**0.01**
1	6.34	12.71	63.66
2	2.92	4.30	9.92
3	2.35	3.18	5.84
4	2.13	2.78	4.60
5	2.02	2.57	4.03
6	1.94	2.45	3.71
7	1.89	2.37	3.50
8	1.86	2.31	3.36
9	1.83	2.26	3.25
10	1.81	2.23	3.17
11	1.80	2.20	3.11
12	1.78	2.18	3.06
13	1.77	2.16	3.01
14	1.76	2.14	2.98
15	1.75	2.13	2.95

16	1.75	2.12	2.92
17	1.74	2.11	2.90
18	1.73	2.10	2.88
19	1.73	2.09	2.86
20	1.72	2.09	2.85
21	1.72	2.08	2.83
22	1.72	2.07	2.82
23	1.71	2.07	2.82
24	1.71	2.06	2.80
25	1.71	2.06	2.79
26	1.71	2.06	2.78
27	1.70	2.05	2.77
28	1.70	2.05	2.76
29	1.70	2.05	2.76
30	1.70	2.04	2.75
35	1.69	2.03	2.72
40	1.68	2.02	2.71
45	1.68	2.02	2.69
50	1.68	2.01	2.68
60	1.67	2.00	2.66
70	1.67	2.00	2.65
80	1.66	1.99	2.64
90	1.66	1.99	2.63
100	1.66	1.98	2.63
125	1.66	1.98	2.62
150	1.66	1.98	2.61
200	1.65	1.97	2.60
300	1.65	1.97	2.59
400	1.65	1.97	2.59
500	1.65	1.96	2.59
1000	1.65	1.96	2.58
infinity	1.65	1.96	2.58

❖ ❖ ❖

Appendix-9

Confidence Intervals

df	$\alpha = 0.1$	0.05	0.025	0.01	0.005	0.001	0.0005
∞	t_{α}=1.282	1.645	1.960	2.326	2.576	3.091	3.291
1	3.078	6.314	12.706	31.821	63.656	318.289	636.578
2	1.886	2.920	4.303	6.965	9.925	22.328	31.600
3	1.638	2.353	3.182	4.541	5.841	10.214	12.924
4	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	1.476	2.015	2.571	3.365	4.032	5.894	6.869
6	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	1.383	1.833	$\boxed{2.262}$				

❖ ❖ ❖

4. Conduct a Chi-Square test using real-world data. Choose a dataset, apply the test (either manually or using statistical software), and write a report on your findings.
5. How do you calculate expected frequencies?
6. When should you reject the null hypothesis in a Chi-Square test?

3.10 References and Suggested Readings:

1. Agresti, A. (2018). *Statistical Methods for the Social Sciences* (5th Ed.). Pearson Education.
2. Garret H.E-Statistics in Psychology and Education .Bombay. Alied Pacific private Ltd.,1962
3. Gupta, S. P. (2021). *Statistical Methods* (45th ed.). Sultan Chand & Sons.
4. Kothari, C. R., & Garg, G. (2019). *Research Methodology: Methods and Techniques* (4th ed.). New Age International Publishers.
5. Levin, R. I., & Rubin, D. S. (2020). *Statistics for Management* (8th ed.). Pearson India.
6. Siegel, S., & Castellan, N. J. (1988). *Nonparametric Statistics for the Behavioral Sciences.* McGraw-Hill.
7. Srivastava, U. K., & Sharma, G. V. (2021). *Quantitative Techniques for Managerial Decisions*. New Age International Publishers.
8. Verma, J. P. (2013). *Data Analysis in Management with SPSS Software*. Springer.
9. Yamane, T. (1967). *Statistics: An Introductory Analysis* (2nd ed.). Harper & Row.

*****