ਜਗਤ ਗੁਰੂ ਨਾਨਕ ਦੇਵ
ਪੰਜਾਬ ਸਟੇਟ ਓਪਨ ਯੂਨੀਵਰਸਿਟੀ
ਪਟਿਆਲਾ

The Motto of Our University
(SEWA)
SKILL ENHANCEMENT
EMPLOYABILITY
WISDOM
ACCESSIBILITY

JAGAT GURU NANAK DEV
PUNJAB STATE OPEN UNIVERSITY, PATIALA

(Established by Act No. 19 of 2019 of the Legislature of State of Punjab)

CERTIFICATE/ DIPLOMA IN STATISTICAL ANALYSIS AND RESEARCH METHODOLOGY

SEMESTER II

SARM 5: STATISTICAL INFERENCE

Head Quarter: C/28, The Lower Mall, Patiala-147001
Website: www.psou.ac.in

The Study Material has been prepared exclusively under the guidance of Jagat Guru Nanak Dev Punjab State Open University, Patiala, as per the syllabi prepared by the Committee of experts and approved by the Academic Council.

COURSE COORDINATOR AND EDITOR:

Dr. Pinky Sra
Assistant Professor
JGND PSOU, Patiala.

ਜਗਤ ਗੁਰੂ ਨਾਨਕ ਦੇਵ
ਪੰਜਾਬ ਸਟੇਟ ਓਪਨ ਯੂਨੀਵਰਸਿਟੀ
ਪਟਿਆਲਾ

AT GURU NANAK DEV PUNJAB STATE OPEN UNIVERSITY,PATIALA
(Established by Act No. 19 of 2019 of the Legislature of State of Punjab)

PREFACE

Jagat Guru Nanak Dev Punjab State Open University, Patiala was established in December 2019 by Act 19 of the Legislature of State of Punjab. It is the first and only Open University of the State, entrusted with the responsibility of making higher education accessible to all, especially to those sections of society who do not have the means, time or opportunity to pursue regular education.

In keeping with the nature of an Open University, this University provides a flexible education system to suit every need. The time given to complete a programme is double the duration of a regular mode programme. Well-designed study material has been prepared in consultation with experts in their respective fields.

The University offers programmes which have been designed to provide relevant, skill-based and employability-enhancing education. The study material provided in this booklet is self- instructional, with self-assessment exercises, and recommendations for further readings. The syllabus has been divided in sections, and provided as units for simplification.

The University has a network of 110 Learner Support Centres/Study Centres, to enable students to make use of reading facilities, and for curriculum-based counseling and practicals. We, at the University, welcome you to be a part of this institution of knowledge.

Prof. G.S. Batra
Dean Academic Affairs

CERTIFICATE/ DIPLOMA IN STATISTICAL ANALYSIS AND RESEARCH METHODOLOGY

SEMESTER II

SARM 5: STATISTICAL INFERENCE

Max. Marks: 100
External: 70
Internal: 30
Pass: 40%
Credits: 6

OBJECTIVE:

- To provide core knowledge required for statistical applications.
- To equip learners with the skills of using appropriate statistical techniques for applications in various fields.

INSTRUCTIONS FOR THE PAPER SETTER/ EXAMINER:

1. The syllabus prescribed should be strictly adhered to.
2. The Question Paper will have 70 Multiple Choice questions (MCQs) and four choices of answers will be there covering the entire syllabus. Each question will carry 1 mark. All questions will be compulsory; hence candidates will attempt all the questions.
3. Paper-setters/Examiners are requested to distribute the questions from Section A and Section B of the syllabus equally i.e., 35 questions from Section A and 35 questions from Section B.
4. The examiner shall give clear instructions to the candidates to attempt questions.
5. The duration of each paper will be two hours.

INSTRUCTIONS FOR THE STUDENTS

The question paper shall consist of 70 Multiple-choice questions. All questions will be compulsory and each question will carry 1 mark. There will be no negative marking. Students are required to answer using OMR (Optimal Mark Recognition) sheets.

SECTION A

Unit 1: Theory of Estimation: Point estimation and Interval estimation

Unit 2: Sampling distributions of a Statistics- Small Sample test or student-t test and its applications: t-test for single mean, difference of means, Paired t-test

Unit 3: Large Sample test: Introduction, Sampling of Attributes- test for Single Proportion, test for difference in proportion

SECTION B

Unit 4: F-statistics: meaning, equity of population variances

Unit 5: Chi-square test- goodness of fit, independent of attributes, test of variance (for population), equality of several population proportions

Unit 6: Analysis of Variance: One-way and Two-way

Unit 7: Interpretation of data and Report writing.

Note: Statistical analysis should also be taught with the help of MS Excel, SPSS or any other related software tool.

Suggested Readings

- A.M Goon, M.K Gupta and B. Dasgupta, fundamental of statistics Vol-I, World press Calcutta
- Anderson, D.R.; Sweeney, D.J. and Williams, T.A., “Statistics for Business and Economics”, 2nd edition (2011), Thompson, New Delhi.
- Cooper, D. R., and Schindler, P.S., “Business Research Methods”, 9th Edition,Tata McGraw-Hill, New Delhi.
- Gupta SC: Fundamental of statistics, S. Chand & Company. New Delhi
- Gupta, SP: Statistical Methods, S. Chand & Company. New Delhi
- Lehmann, E.L. (1986): Testing Statistical hypotheses (Student Edition).
- Monga, GS: Mathematics and Statistics for Economics, Vikas Publishing house, New Delhi.
- Zacks, S. (1971): Theory of Statistical Inference, John Wiley and Sons. New York.

ਜਗਤ ਗੁਰੂ ਨਾਨਕ ਦੇਵ
ਪੰਜਾਬ ਸਟੇਟ ਓਪਨ ਯੂਨੀਵਰਸਿਟੀ
ਪਟਿਆਲਾ

CERTIFICATE/ DIPLOMA IN STATISTICAL ANALYSIS AND RESEARCH METHODOLOGY

SEMESTER II

SARM 5: STATISTICAL INFERENCE

EDITOR AND COURSE CO-ORDINATOR- DR. PINKY SRA

SECTION A

UNIT NO.	UNIT NAME
Unit 1	Theory of Estimation
Unit 2	Sampling distributions of a Statistics- Small Sample test
Unit 3	Large Sample test: Introduction, Sampling of Attributes

SECTION B

UNIT NO.	UNIT NAME
Unit 4	F-statistics: meaning, equity of population variances
Unit 5	Chi-square test- goodness of fit, independent of attributes, test of variance (for population), equality of several population proportions
Unit 6	Analysis of Variance: One-way and Two-way
Unit 7	Interpretation of data and Report writing

CERTIFICATE/ DIPLOMA IN STATISTICAL ANALYSIS AND RESEARCH METHODOLOGY

SEMESTER II

SARM 5: STATISTICAL INFERENCE

UNIT 1: THEORY OF ESTIMATION: POINT ESTIMATION AND INTERVAL ESTIMATION

STRUCTURE

1.0 LEARNING OBJECTIVES

After studying the Unit, the learner will be able to:

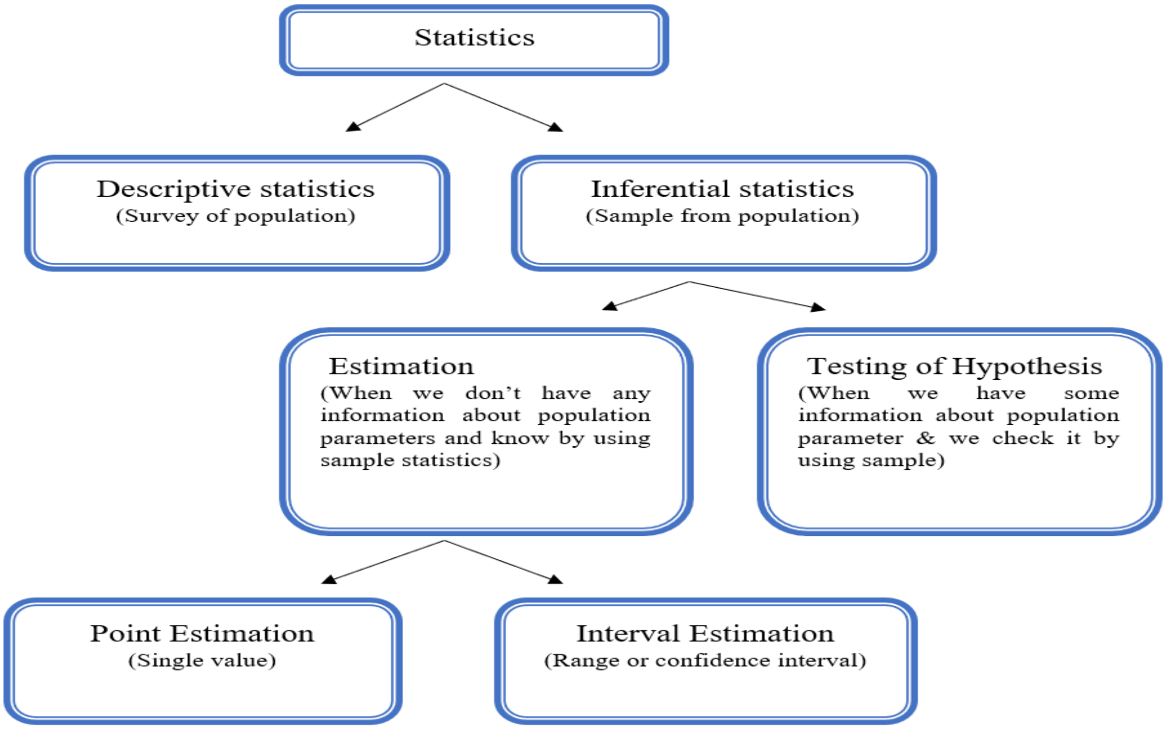
- Understand the meaning of estimation

- Important terminologies
- Various characteristics of the theory of estimation
- Know about the various types and applications of the theory of Estimation
- Properties of a good estimator
- how these properties are important in point Estimation and Interval Estimation
- types of Errors and different types of statistical tests.

1.1 INTRODUCTION

We live in an environment of information explosion. We obtain information from various resources and different formats. Information can be in numerical or non-numerical nature. The sources of information include newspapers, television, books, journal articles, governmental or nongovernmental organization publications, and international institutions, private and established institutions. Even in some cases, we used the primary data, which means data generated on our own. Statistics is a subfield of mathematics that deals with collecting, processing, and arranging data. It offers techniques for making inferences and choosing options to face uncertainty. In statistics, two techniques are used for the calculation of population parameter values. The first one is descriptive statistics, in which data is collected from the whole population like the census method. This method includes summarizing and presenting data in a meaningful way. Mean, median, mode, range and standard deviation all come under descriptive statistics and all these measures help to provide a compressed outline or abstract of the main features of the data set. The second one is inferential statistics, in this technique, the sample is selected from the population which represents the whole population and data is collected only from the sample. Hypothesis testing, regression analysis and confidence interval all are included in the inferential statistics. Statistical inference is a powerful tool for generalization and decision-making because it allows researchers to make inferences beyond the specific data they have collected. Statistical Inference or inferential statistics is further categorized into two parts:

- Estimation
- Testing of Hypothesis


Statistics
Descriptive statistics
(Survey of population)
Inferential statistics
(Sample from population)
Estimation
(When we don't have any information about population parameters and know by using sample statistics)
Testing of Hypothesis
(When we have some information about population parameter & we check it by using sample)
Point Estimation
(Single value)
Interval Estimation
(Range or confidence interval)

Theory of Estimation: Estimation is the procedure in which through sample statistics like- sample mean($\bar{x}$), sample variance(s^2), sample median(M), etc., we estimate the population parameters like- Population means (μ), Population variance (σ^2), etc. For example: If we want to know about the interests of university students and which games they like, so for, we will only take the data from the sample rather than collect the data from the whole population. The sample will represent the whole population.

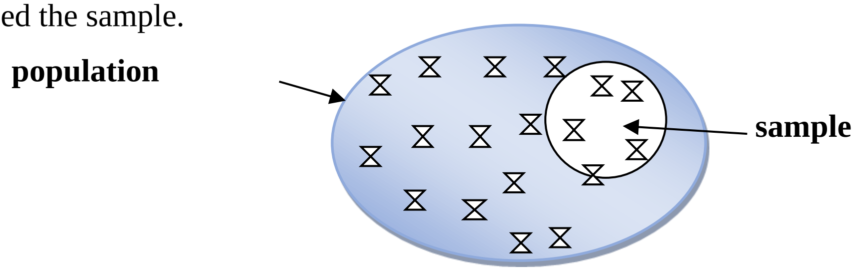
1.2 CONCEPT OF THEORY OF ESTIMATION

1.2.1. IMPORTANT TERMINOLOGY

1. **Estimators**: For the Estimation of the unknown population parameters some sample statistics are used. These sample statistics are called Estimators. In other words, Estimators are the formula or rule to make guesses of the population parameters through the sample. For example: Any function says A (a_1, a_2, a_3, a_4………..) of samples a_1, a_2, a_3, a_4………..are called as an estimator.
2. **Estimates**: The quantified value from the Estimation is called Estimates. In the other way, the result of Estimation is called Estimates. For example: The mean ($\bar{a}$) and the variance of random sample a_1, a_2, a_3, a_4………are called estimates. If Estimated parameters represented by γ (read

as gamma) then$\widehat{\gamma}$(read as gamma hat) will be represent the Estimator. It means for the population parameters (γ), an Estimated is $\hat{\gamma}$

3. **Population**: In statistics, A set of homogeneous items or events that are concerned with some problem or question is called a population. For example: let us suppose we want to study the average consumption of milk in a boy's hostel. So, for this purpose, the boys of the hostel who consumed the milk are our population for study.
4. **Parameter**: Any number or sign that represents the whole population is called a parameter. For example: for the normal distribution μ and σ^2 are called parameters that show the population mean and population variance respectively.
5. **Sample**: The subset or small part /portion that shows all the characteristics of the population called the sample.



1.2.2. CHARACTERISTICS OF THE THEORY OF ESTIMATION

1. **Estimation from Sample Data**– The Theory of estimation concentrates on estimating unknown population parameters by using sample data. It gives the method for making inferences about the population using a subset of the population.
2. **Statistical Properties**– The estimation theory is based on some statistical properties. These properties are unbiasedness, efficiency, sufficiency, and consistency.
3. **Sampling Distributions**– Estimation theory depends upon the concept of sampling distributions. A sampling distribution of statistics made from multiple samples. A random sample is drawn from a specific population.
4. **Hypothesis Testing**– A method of statistical inference that is used to determine whether the data is enough to support the hypothesis or not is called hypothesis testing. In which various types of tests are used to determine the best parameters.
5. **Efficiency and Precision**–This theory aims to develop estimators that are efficient and precise. The word efficiency means the capability of estimating small variances, making it more precise compared to other estimators. The efficient estimators give the balance between bias and variance.

6. **Methods of Selection**– This theory provides a range of methods like a method of moments, maximum likelihood estimation and Bayesian estimation. The selection of the estimation method depends upon the problem, the available data and any earlier knowledge or assumptions.
7. **Application and Generalization**– In various fields such as economics, social science, engineering and medical research, the estimation theory is applicable. In different contexts, the principles and techniques developed in estimation theory can be generalized to find a wide range of estimation problems.

1.2.3. KINDS OF ESTIMATION

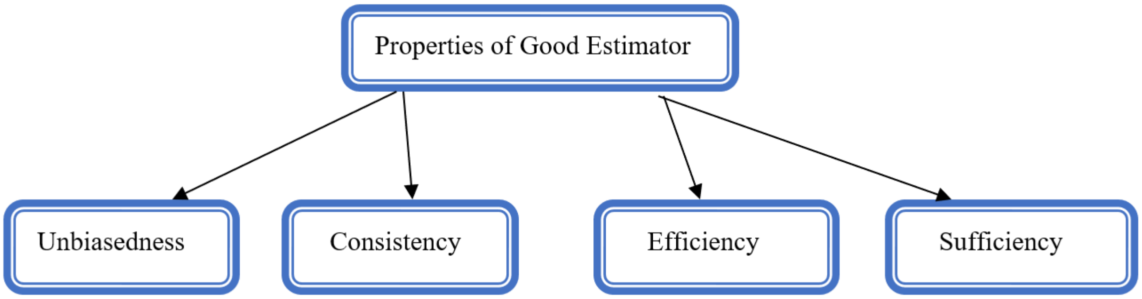
There are two manners through which population parameters can be estimated.

1. **Point Estimation**: Point Estimation is the type of estimation in Which single values of Statistics are used for the estimate of unknown population parameters. For example, the sample mean ($\bar{X}$) is the point Estimator of the population mean (μ) and statistics (s^2). Sample variance is the point estimate of population variance (σ^2). For Example: Suppose a student wants to be admitted to the Central University of Punjab then he estimates the charges that will be paid by him in the time of admission fees, medical fees, tuition fees, examination fees, etc. All these values when put together to calculate the entire cost. Then the calculated cost of admission is known as the point estimator.
2. **Interval Estimation**: In the point estimator only, a single value is used for estimation, this shows the dissatisfying estimate because the point estimate may deviate from the true value of the parameter. So that another method of estimation is interval estimation. It gives the probable range of values. In which the true value of the parameter is expected to lie. This probable range is called the confidence interval and the two extreme values of the range are called confidence limits. For example: Based on the sample we estimate the average marks of students in the monthly test of economics is between 60 and 80. This is the case of interval estimation in which 60 and 80 are extreme values or confidence limits.

1.3 PROPERTIES OF A GOOD ESTIMATOR

Population parameters can be estimated by various estimators like: sample mean or sample median or sample mode may be used to estimate the population mean and sample variance, sample S.D., and sample mean deviation used for population variance. Out of the various accessible or available

Estimators, it is important to determine a good Estimator. ‘‘A Good Estimator is as close to the true value of the parameter as possible.’’



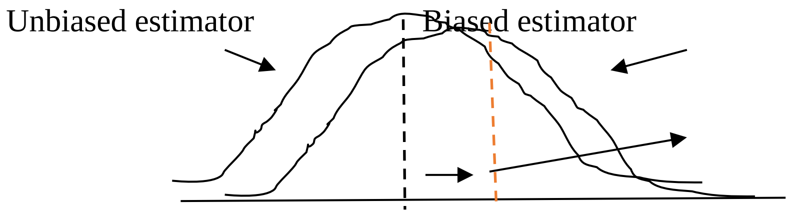
Unbiasedness: : If the mean of the sampling distribution of the estimator $\hat{\gamma}$ is equal to the population parameter γ then an Estimator is called an unbiased estimator of the population parameter.

A statistical term, an unbiased Estimator is an estimator in which the expected value of the sample estimate is equal to the population parameter. Symbolically,

$$E(\hat{\gamma}) = \gamma$$

Example: The expected mean of the sampling distribution is equal to the population mean then the sample mean ($\bar{x}$) is an unbiased estimator of a population mean (μ).

$E(\bar{x}) = \mu$



Deviation from the true value

True value / Original value

Another example: If sample variance &population variance is not equal, at that point

$E(s^2) \neq \sigma^2$

Though, remould sample variance ($\hat{s}^2$) is an unbiased Estimator of population variance

$$\hat{s}^2 = n/n-1*s^2$$

Consistent Estimator: When the sample size increases, the Estimator is likely to become near or close to the parameter then the Estimator is told as a consistent Estimator. A consistent estimator doesn't need to be unbiased.

The probability that $\hat{\gamma}$ near toγ is 1 as some sample increases, An Estimator $\hat{\gamma}$ is called a consistent Estimator of population parameter γ.

$P(\hat{\gamma}\rightarrow\gamma) \rightarrow 1 \quad \text{as } n\rightarrow \infty$

There are two main conditions for a consistent Estimator:

$E\ (\hat{\gamma})\rightarrow\gamma$

Variance $\hat{\gamma}\rightarrow 0$ as $n\rightarrow\infty$

Efficient Estimator: When parameters have more than one unbiased Estimator, then all unbiased Estimators are compared to each other based on variance. Efficiency is a comparative term. An Estimator will be selected as an efficient Estimator which has a minimum variance. An estimator with lesser variance than other variances is more efficient as well as more reliable than the other.

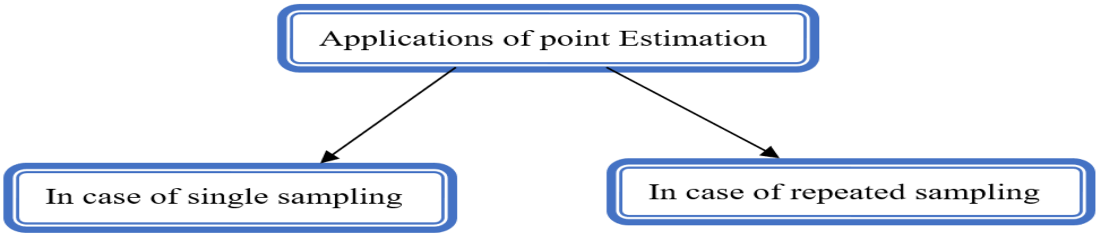
Variance $(\hat{\gamma}_1) \ < \ $ variance $(\hat{\gamma}_2)$

The above shows that the Estimator $\hat{\gamma}_2$ has more variance than the $\hat{\gamma}_1$, so that the $\hat{\gamma}_1$ is more efficient than the $\hat{\gamma}_2$.

Sufficient Estimator: If the estimator $\hat{\gamma}$ hold all information related to the population parameter γ then that Estimator should be called a sufficient Estimator. In other words, for a sufficient estimator, a sample should have all information regarding the population. This is the last property of the good Estimator. A sufficient Estimator is consistent and can or can't be an unbiased Estimator. For example, A sample mean ($\bar{x}$) is said to be the Sufficient Estimator for the population mean(μ) if the sample holds all information related to the population.

1.4 APPLICATIONS OF POINT ESTIMATION

There are two applications are studies related to point estimation:


Applications of point Estimation
In case of single sampling
In case of repeated sampling

1. **Point estimation in case of single sampling**: When a single independent random sample is drawn from an unknown population known as point estimation of single sampling.

Example 1: The following five observations constitute a random sample from an unknown population: 2,4,6,8,10

Find out unbiased and efficient estimates of (a) true mean (b) true variance.

Solution:(a) The unbiased and efficient estimator of the true mean is given by the value of

$$\overline{X}= \frac{\sum X}{n} =\frac{2+4+6+8+10}{5} = \frac{30}{5} = 6$$

(b) The unbiased and efficient estimate of the true variance is: $\hat{s}^2 = \frac{\sum(X-\overline{X})^2}{n-1}$

where $\hat{s}^2$ = modified sample variance.

$$= \frac{(2-6)^2+ (4-6)^2+ (6-6)^2+ (8-6)^2+ (10-6)^2}{5-1} = \frac{40}{4} = 10$$

2. **Point estimation in case of repeated sampling**: when a large number of random samples of the same size from the population with or without replacement are drawn, called point estimation of repeated sampling.

Example 2: A population contains five values 2,3,4,5,6. List all possible samples of size three without replacement. Calculate the mean x̄ of each sample. Check that sample mean x̄ is an unbiased estimate of the population.

Solution: Total number of possible samples of size 3 without replacement 5_{c3} =10

Sampling no.	Sample values	Sample Mean (x̄)
1.	(2, 3, 4)	1/3(2+3+4) =9/3=3
2.	(2, 3, 5)	1/3(2+3+5) =10/3=3.33
3.	(2, 3, 6)	1/3(2+3+6) =11/3= 3.67
4.	(2, 4, 5)	1/3(2+4+5) =11/3=3.67
5.	(2, 4, 6)	1/3(2+4+6) =12/3=4
6.	(2, 5, 6)	1/3(2+5+6) =13/3=4.33
7.	(3, 4, 5)	1/3(3+4+5) =12/3=4
8.	(3, 4, 6)	1/3(3+4+6) =13/3=4.33

9.	(3, 5, 6)	1/3(3+5+6) =14/3=4.67
10.	(4, 5, 6)	1/3(4+5+6) =15/3=5
Total	T=10	$\sum \bar{x}$=40

Mean of sampling distribution of mean = $\mu_{\bar{x}} = \sum \bar{x} / T = 40/10 = 4$

Population mean =μ= (2+3+4+5+6)/5 = 20/5= 4

Because $\mu_{\bar{x}} = \mu$, the sample mean is $\bar{x}$ is an unbiased estimate of the population mean μ.

Example 3. The population contains three values 2,3,4. Draw all possible samples of size two with replacement calculate the mean and variance for each sample examine both the statistics are unbiased and efficient for parameters. Solution: The total number of possible samples of size 2 with replacement is $N^n = 3^2 = 9$

(a) Mean of the sampling distribution of means= $\mu_{\bar{y}} = \sum \bar{y}/T$= 27/ 9 = 3, (T= no. of samples)

Population mean μ = (2+3+4)/3 = 3

The sample mean $\bar{y}$ is an unbiased estimate of the population mean $\mu_{\bar{y}}$, because $\mu_{\bar{y}}=\bar{y}$

Sample No.	Sample Variance	sample mean ($\bar{x}$)	sample variance $s^2=1/2[(Y_1-\bar{y})^2+(Y_2-\bar{y})^2]$	modified sample variance ($\hat{s}^2 = n/n-1*s^2$)
1.	(2,2)	½ (2+2) =2	½ $[(2-2)^2 + (2-2)^2]$= 0	0
2.	(2,3)	½ (2+3) = 2.5	½ $[(2-2.5)^2 + (3-2.5)^2]$=0.25	0.5
3.	(2,4)	½ (2+4) =3	½ $[(2-3)^2 + (4-3)^2]$=1	2
4.	(3,2)	½ (3+2) =2.5	½ $[(3-2.5)^2 + (2-2.5)^2]$= 0.25	0.5
5.	(3,3)	½ (3+3) =3	½ $[(3-3)^2 + (3-3)^2]$=0	0
6.	(3,4)	½ (3+4) =3.5	½ $[(3-3.5)^2 + (4-3.5)^2]$=0.25	0.5
7.	(4,2)	½ (4+2) =3	½ $[(4-3)^2 + (2-3)^2]$=1	2
8.	(4,3)	½ (4+3) =3.5	½$[(4-3.5)^2+(3-3.5)^2]$=0.25	0.5
9.	(4,4)	½ (4+4) = 4	½$[(4-4)^2+(4-4)^2]$=0	0
Total	T=9	$\sum \bar{y}$= 27	$\sum s^2$= 3	$\sum \hat{s}^2$=6

(b) Mean of the sampling distribution of variance =$\mu_{s^2} = \sum s^2 / T = 3/9 = 1/3$

Population variance $\sigma^2 = \frac{(2-3)^2+(3-3)^2+(4-3)^2}{3} = \frac{2}{3}$

$\mu_{s^2} \neq \sigma^2$, sample variance s^2 is not an unbiased estimate of the population variance σ^2.

but modified sample variance $\hat{s}^2$ will be an unbiased estimate of the population variance σ^2 because:

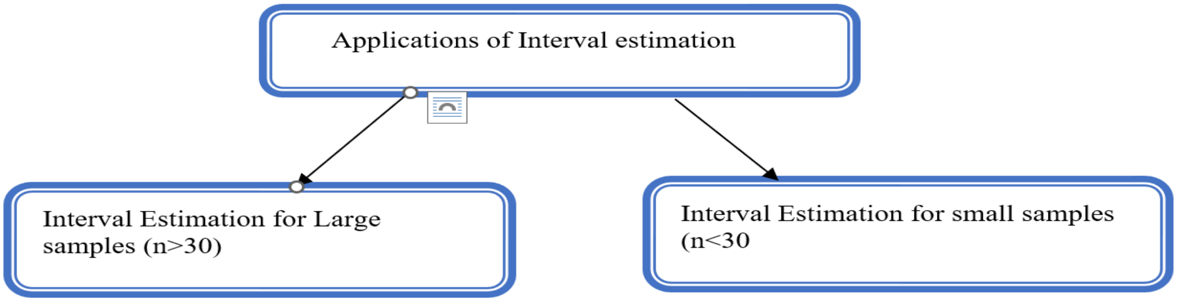
$$\hat{s}^2 = n/n\text{-}1 * s^2$$

$$\mu_{s^2} = \frac{\sum \hat{s}^2}{T} = \frac{6}{9} = \frac{2}{3}$$

$$\mu_{\hat{s}^2} = \sigma^2$$

So, the modified sample variance is an unbiased estimate of the population variance.

1.5 APPLICATIONS OF INTERVAL ESTIMATION (CONFIDENCE INTERVAL)



Applications of Interval estimation

Interval Estimation for Large samples (n>30)

Interval Estimation for small samples (n<30

1. Confidence interval for population mean

2. Confidence interval for population proportion

3. Confidence interval for population standard deviation

1. Confidence interval for Population Mean μ

Note: - I. The sample S.D (s) is used for large samples If the population S.D is not Known.

II. Values of $Z_{\alpha/2}$ (for large samples)[2]

Confidence level(1-α)	90%	95%	96%	98%	99%	Without any reference confidence level
Z- value	±1.64	±1.96	±2.06	±2.33	±2.58	±3

1. **Confidence interval or Limits for the population mean μ (when n > 30)**: In this the use of normal distribution required: (1-α)100% Confidence limits for μ are given by:

   $\bar{x} \pm Z_{\alpha/2}$ S.E or $\bar{x} \pm Z_{\alpha/2}\sigma/\sqrt{n}$ (where σ is known)

   $\bar{x} \pm Z_{\alpha/2}\, s/\sqrt{n}$ (where σ is not known.)

   for a large sample, σ=s

Example 4: sample mean ($\bar{x}$) =300, sample size (n)=400, sample variance (s^2) =1600 calculate the 95% confidence interval for the population mean.

Solution: here, $\bar{x}$=300, n=400, s^2=1600 or s=40

$S.E = s/\sqrt{n} = 40/\sqrt{400} = 40/20 = 2$

$S.E = 2$

The value of $Z_{\alpha/2} =$ 1.96 at 95% Confidence level $\overline{x} \pm 1.96$ S. E $_{\overline{x}}$

Put the values, 300 ± 1.96 × 2 = 300 ± 3.92 = 296.08or 303.92

Thus, $296.08 < \mu < 303.92$

2. **Confidence Interval or limits for population proportion (when > 30):** Also, in this case, the normal distribution can be used however in the proportion the binomial distribution is used as a sampling distribution.

(1-α) 100% confidence limits for P are given by:

$P \pm Z_{\alpha/2}$ S.E$_{(P)}$ OR $p \pm Z_{\alpha/2}\sqrt{PQ/n}$ where P is known

$P \pm Z_{\alpha/2}\sqrt{pq/n}$ where P is unknown

(n= sample size, p= proportion of success, q=1-p).

Example 5: A coin is tossed 1000 times and it gives 400 heads and 600 tails. Find the 99% confidence interval for the heads.

Solution: Given n=1000 total heads (np) =400

P = sample proportion heads = 400/1000 = 0.4

P= proportion of success (head) = 1/2 = 0.5

Q = 1-P = 1-0.5 = 0.5

S.E$_{(p)}$ = $\sqrt{PQ/n} = \sqrt{0.5 \times 0.5 /1000} = 0.05$

For a 99% confidence interval, the value of $Z_{\alpha/2} = 2.58$

Confidence interval for P = p± 2.58 S. E $_{P}$

Put the values: 0.4 ± 2.58 × 0.05

0.4 ±0.129

$= 0.271 < P < 0.529$

3. **Confidence Interval or limits for population standard deviation:**

$s \pm Z_{\alpha/2}\, S.E_S$ OR $s \pm Z_{\alpha/2}\, \sigma/\sqrt{2n}$ where σ is known

$s \pm Z_{\alpha/2}\, s/\sqrt{2n}$ where σ is not known

Example 6: A sample of 72 observations has an S.D equal to 20. Calculate the 90% confidence interval for population standard deviation σ.

Solution: Given, n=72 s= 20

$S.E(s) = \sigma/\sqrt{2n} = 20/\sqrt{2\times72} = 20/\sqrt{144} = 20/12 = 1.667$

90% confidence level, the value of $Z_{\alpha/2} = 1.64$

Putting the values: $= 20 \pm 1.64 \times 1.667$

$= 20 \pm 2.734$

$= 17.266$ to 22.734 or $17.266 < \sigma < 22.734$

Interval estimation (or confidence interval) for small samples n ≤ 30

Confidence interval or limits for the population mean ($n \leq 30$): In the case of small samples, the t-values are used instead of Z values.

$(1-\alpha)$ 100% confidence limit for population mean μ is given by:

$\overline{x} \pm t_{\alpha/2.}\, \hat{s}/\sqrt{n}$ where $\hat{s}$ is modified sample S.D = $\sqrt{\sum(x-\overline{x})^2/n-1}$.

or

$\hat{s} = \sqrt{n/n-1.\, s^2}$

Process:

a) Calculate or take $\overline{x}$ and calculate the modified sample S.D. using the formula
b) Calculate the Degree of freedom (d.f = v = n-1)
c) Choose the desired confidence level corresponding to that specified level of confidence and for given d.f, we note the value of that $t_{\alpha/2}$ from the table.
d) Calculate the confidence interval by substituting the values of $\overline{x}$, $\hat{s}$, and $t_{\alpha/2}$in the formula.

Example 7: A sample of 9 has 25 as the mean with S.D of 4. Obtain 99% confidence limits of the mean of the population.

Solution: Given $\overline{x} = 25$, $n=9$, $s= 4$ or $s^2 = 16$

$\hat{s}= \sqrt{n/n-1}.\ s^2 = \sqrt{9/9-1} \times 16 = 4.242$

Degree of freedom $=n-1 = v = 9-1 = 8$

For 99% confidence level $\alpha = 0.01$, so that $\alpha/2 = 0.01/2 = 0.005$

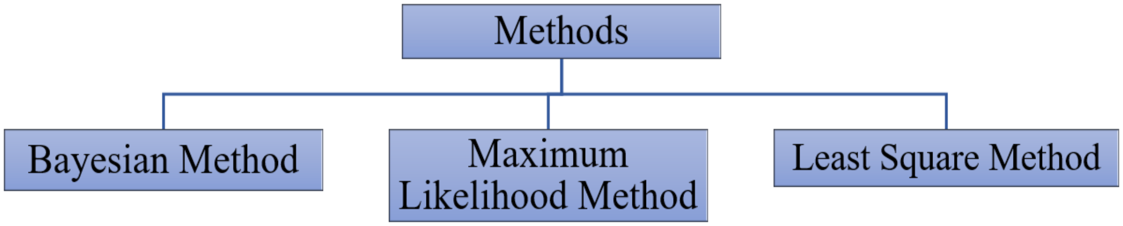
Using t-table the value of $t_{0.005}$ for 8 degree of freedom = 3.355

99% confidence limits for μ:

$\overline{x} \pm t_{0.005.}\ \hat{s}/\sqrt{n}$

put the values $25 \pm 3.355\times 4.242/\sqrt{9} = 25 \pm 4.744 =20.256$ to 29.744

1.6 METHODS OF POINT ESTIMATION



1. Bayesian Method: The Bayesian method of estimation in statistics fundamentally differs from classical or frequentist approaches by treating model parameters as probability distributions rather than fixed, unknown constants. It begins with prior beliefs about these parameters, representing existing knowledge or assumptions before observing new data. The likelihood function, describing the probability of observing the data given different parameter values, is then combined with prior beliefs using Bayes' theorem. This integration yields a posterior distribution, reflecting updated beliefs after considering the new evidence. Bayesian inference involves summarizing this posterior distribution to make statistical inferences about the parameters.

Notably, Bayesian methods allow for sequential updating as new data becomes available, providing a dynamic approach to incorporating information over time. By representing parameters as probability distributions, Bayesian analysis inherently captures and quantifies uncertainty, offering a more comprehensive view of the model's unpredictability. The flexibility of this approach in handling complex models and incorporating subjective information contributes to its appeal, providing a coherent framework that integrates both prior knowledge and new evidence for decision-making.

2. Maximum Likelihood Method: The point in the set of all possible values that a parameter can take in a statistical model, that maximizes the likelihood function is called the maximum likelihood estimate. Maximum Likelihood Estimation (MLE) is a statistical technique that determines the parameters of a presumed probability distribution based on observed data. This is accomplished by optimizing a likelihood function to make the observed data the most likely under the presumptive statistical model. The Maximum Likelihood Estimation (MLE) method is used to find the values of parameters that maximize the probability of the observed data. If the likelihood function, which shows the probability of the observed data given alternative parameter values, is differentiable, calculus can be used to find the maximum. The Ordinary Least Squares (OLS) estimator is acquired logically by solving the first-order requirements in situations such as linear regression with normally distributed errors. The goal of MLE, a unique kind of immoderate estimator in frequentist statistics, is to maximize the likelihood function given the observed data by finding the values of the parameters.

3. Least Square Method: The least squares method is a statistical technique used to figure out the best-fit line for a given set of data points. In regression analysis, we turn to the least squares method when dealing with an equation that involves both a dependent and an independent variable. The idea behind this method is to minimize the sum of squares of residuals, essentially the differences between the actual values and the fitted values in the model.

This method is commonly applied in data fitting, to achieve the best fit by reducing the sum of squared errors or residuals. There are essentially two main types of problems within the least squares method:

1. Linear or Ordinary Least Squares Method
2. Non-linear Least Squares Method

The distinction between these two types is rooted in whether the residuals display linearity or nonlinearity. Linear problems are frequently encountered in statistical regression analysis, whereas non-linear problems tend to find use in iterative refinement methods. In these methods, the model is progressively approximated toward linearity with each iteration.

1.7 QUESTIONS FOR PRACTICE

(a) Short Answer Type

Q1 What is the meaning of the theory of Estimation?

Q2 Explain Z-Test.

Q3 What are the types of the Estimation?

Q4 Name the various properties of a good estimator.

(b) Long Answer Type

Q1 Explain the concept of the theory of estimation in detail.

Q2 What are the various types of Estimation? Explain with Examples.

Q3 Explain the process of Testing of Hypothesis.

1.8 MCQs

Q 1: The standard error of an estimator Increases when:

a) The sample size increases.

b) The sample size decreases.

c) The estimator becomes biased.

d) The value of S.D is unknown.

Q 2: Which of the following is not a part of the properties of a good estimator:

a) Unbiasedness

b) Consistency

c) Variability

d) Sufficiency

Q 3: An estimator's efficiency is quantified by:

a) Its biasedness

b) The range of Confidence interval

c) Sample size

d) standard error

Q 4: A confidence interval is used to:

a) Provide an estimate of the population parameter.

b) Determine the standard error of the sample mean.

c) Test the hypothesis about the population mean.

d) Calculate the range of possible sample values.

Q 5: What is the estimated standard error of the sample mean if the population standard deviation of the 36 observations is 8.2 and the mean is 12?

a) 1.36

b) 1.92

c) 1.71

d) 0.78

Q 6: In point estimation the term 'Point' means:

a) The highest value in the estimation

b) A specific numerical value used as an estimate for a parameter.

c) The center of a distribution.

d) The range of possible values for a parameter.

Q 7: A 90% confidence interval for a population parameter means:

a) There is a 90% probability that the parameter falls within the interval.

b) The interval contains 90% of the population data.

c) With multiple sampling and estimate intervals, about 95% of them will contain the true parameter value.

d) The interval represents the range of uncertainty in the data.

Q 8: In which situations is the z-test most appropriate?

a) When the sample size is small.

b) When the population variances are unknown.

c) When the population means are significantly different.

d) When the sample size is large ($n \geq 30$) and variances are known.

Q 9: What does a null hypothesis (H_0) in a z-test generally assume?

a) The sample statistic is significantly different from the population parameter.

b) There is no significant difference between the sample statistic and the population parameter.

c) The sample size is small.

d) The sample is normally distributed.

Q 10: What is the critical region in a hypothesis test?

a) The region where the sample statistic falls.

b) The region where the null hypothesis is true.

c) The region where the alternative hypothesis is true.

d) The region beyond the critical value(s) associated with the chosen significance level.

Q 11: What is the purpose of a two-tailed z-test?

a) To test whether the sample statistic is significantly greater than the population parameter.

b) To test whether the sample statistic is significantly less than the population parameter.

c) To test whether the sample statistic is significantly different from the population parameter.

d) To test whether the sample statistic follows a normal distribution.

Q 12: What is the level of significance (α) in hypothesis testing?

a) The probability of making a type I error.

b) The probability of making a type II error.

c) The probability of correctly accepting the null hypothesis.

d) The probability of correctly rejecting the null hypothesis.

Q 13: In a Z-test, if the calculated Z-statistic is 2.45, and the critical value for a two-tailed test at $\alpha = 0.05$ is ±1.96, what is the appropriate decision?

a) Reject the null hypothesis.

b) Fail to reject the null hypothesis.

c) Not enough information to decide.

d) The sample size is insufficient.

Q 14: A researcher conducts a Z-test with a sample of 100 observations, a sample mean of 75, a population mean of 70, and a population standard deviation of 12. What is the Z-statistic?

a) 0.42

b) 1.67

c) 2.08

d) 4.17

Q 15: Who developed the t-test, and under what circumstances did he publish his findings?

a) John Smith, a mathematician, published his findings under a pseudonym.

b) Willian Sealy Gosset published his findings under a pen name due to company policies.

c) Student Pearson collaborated with Guinners Son & Company to develop the t-test.

d) Marie Curie introduced the t-test to the scientific community.

Q 16: What is the primary purpose of the t-test in statistics?

a) To calculate the total sum of a dataset

b) To determine the median of a dataset

c) To compare the means of two groups or samples

d) To analyze categorical data distributions

Q 17: When is the t-test commonly used?

a) When the sample size is large ($n > 30$)

b) When the sample size is small ($n < 30$)

c) When the population standard deviation is known

d) When the data distribution is negatively skewed

Q 18: What formula is used to calculate the t-value in a one-sample t-test?

a) $t = (x - \mu) / s$

b) $t = (x - \mu) * n$

c) $t = (x - \mu) * \sqrt{n}$

d) $t = (x - \mu) / \sqrt{s}$

Q 19: In an unpaired t-test, what does "unpaired" refer to?

a) Observations in the two groups are matched in some way.

b) The samples are taken from the same population.

c) Observations in each group are dependent on each other.

d) The samples are taken at different times of the day.

Q 20: What is the formula for calculating the t-value in a paired t-test?

a) $t = (x1 - x2) / s$

b) $t = (x1 - x2) * n$

c) $t = (x1 - x2) * \sqrt{n}$

d) t = (x1 - x2) / √s

Q 21: Which type of t-test would you use to compare the means of two independent samples?

a) One-sample t-test

b) Unpaired t-test

c) Paired t-test

d) Two-sample t-test

Answers

Q 1: B) The sample size decreases.

Q **2:** D) Sufficiency.

Q**3:** D) standard error.

Q4: A) Provide an estimate of the population parameter.

Q5: C) 1.71.

Q 6: B) A specific numerical value used as an estimate for a parameter.

Q 7: C) With multiple sampling and estimate intervals, about 95% of them will contain the true parameter value.

Q 8: d) When the sample size is large (n ≥ 30) and variances are known.

Q 9: b) There is no significant difference between the sample statistic and the population parameter.

Q 10: d) The region beyond the critical value(s) associated with the chosen significance level.

Q 11: c) To test whether the sample statistic is significantly different from the population parameter.

Q 12: a) The probability of making a type I error.

Q 13: a) Reject the null hypothesis.

Q 14: b) 1.67.

Q 15: b) Willian Sealy Gosset published his findings under a pen name due to company policies.

Q 16: c) To compare the means of two groups or samples.

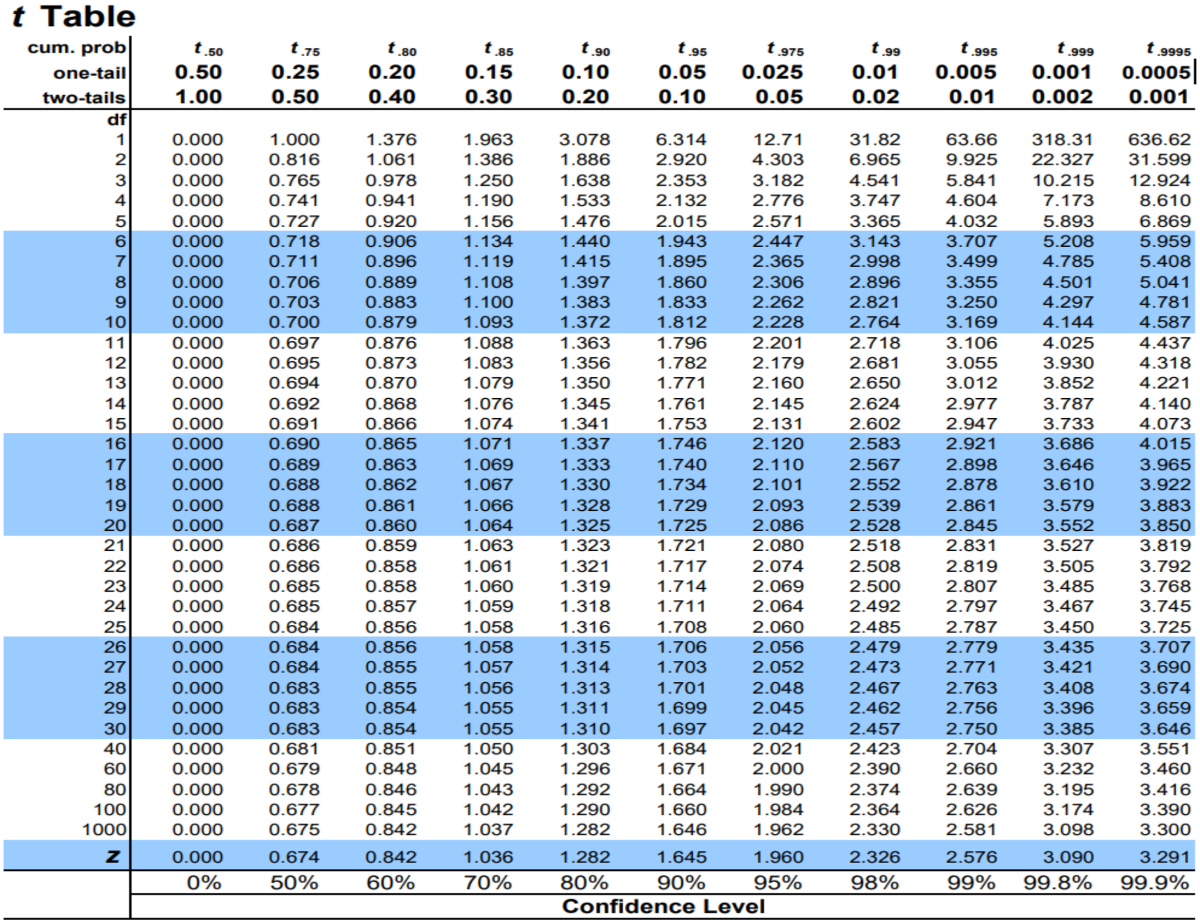
Q 17: b) When the sample size is small (n < 30).

Q 18: a) t = (x - µ) / s.

Q 19: a) Observations in the two groups are matched in some way.

Q 20: d) t = (x1 - x2) / √s.

Q 21: b) Unpaired t-test.



***t* Table**

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
						Confidence Level					

1.9 SUGGESTED READINGS

- Dr. S.C Aggarwal, Dr. R.K Rana; Statistics for Economists (Quantitative Methods for Economists), VK Global publications Pvt. Ltd. (p 372 - p 384, p529 - p555)
- www.investopedia.com
- www.cuemath.com
- www.analyticsvidya.com
- www.educba.com.
- Dr. Kailash Chandra Pradhan, Statistics for economics, Mahatma Gandhi central university

CERTIFICATE/ DIPLOMA IN STATISTICAL ANALYSIS AND RESEARCH METHODOLOGY

SEMESTER II

SARM 5: STATISTICAL INFERENCE

UNIT 2: SMALL SAMPLE TEST

STRUCTURE

2.0 OBJECTIVES

After studying this unit, the learner should be able to:

- know the procedure of t-test for testing a hypothesis
- describe testing of the hypothesis for the population mean for using a t-test
- explain the testing of the hypothesis for the difference between two population means
- when samples are independent using a t-test
- describe the procedure for paired t-test for testing of hypothesis
- difference of two populations means when samples are dependent or paired
- testing of the hypothesis for the population correlation coefficient using a t-test.

2.1 INTRODUCTION

t-test was developed in 1908 by “Willian Sealy Gosset” he was working with “Guinners Son & Company- A Dublin Brewery, in Ireland” and company did not permit employees to publish their research findings under their names, so he published his findings under the pen name “Student”. So, it is also called as “Student” t-test. It is a statistical hypothesis testing tool that is used to determine whether there is a significant difference between the means of two groups or samples. t-test is commonly used when the sample size is small or n<30 (n, means number of observations) and population standard deviation is unknown. It is based on t-distribution which is similar to the normal distribution but less peaked than normal distribution and has a higher tail than normal distribution.

The shape of the t-distribution varies with the change in the degree of freedom, it is less peaked than the normal distribution at centre and more peaked in the tails. The value of t-distribution ranges from -∞ to +∞ ($-\infty < t < +\infty$).

The following are the standard t-tests:

- One-sample: Compares a sample mean to a reference value.
- Two-sample: Compares two sample means.
- Paired: Compares the means of matched pairs, such as before and after scores.

To choose the correct t-test, you must know whether you are assessing one or two group means. If you’re working with two groups means, do the groups have the same or different items/people? Use the table below to choose the proper analysis.

Number of Group Means	**Group Type**	**t-test**
One	--------	One sample t-test
Two	Different items in each group	Two sample t-test
Two	The same items in both groups	Paired t-test

2.3 PROCEDURE OF T-TEST FOR TESTING A HYPOTHESIS

Let us give you similar details here. For this purpose, let X_1, X_2, …, Xn be a random sample of small size n (< 30) selected from a normal population having parameter of interest, say, θ which is unknown but its hypothetical value, say, θ_0 estimated from some previous study or some other way is to be tested.

t-test involves the following steps for testing this hypothetical value:

Step I: First of all, we have to set up null hypothesis H_0 and alternative hypothesis H_1.

Suppose, we want to test the hypothetical / Testing of the Hypothesis assumed value μ_0 of parameter μ.

So, we can take the null and alternative hypotheses as H_0: $\mu = \mu_0$

H_1: $\mu \neq \mu_0$ (for the two-tailed test)

While one- tail test as:

H_0: $\mu = \mu_0$ and H_1: $\mu > \mu_0$ (Right-tailed)

H_0: $\mu = \mu_0$ and H_1: $\mu < \mu_0$ (Left-tailed)

In case of comparing the same parameter of two populations of interest, say, μ_1 and μ_2, then our null and alternative hypotheses would be

H_0: and $\mu_1 = \mu_2$ and H1: $\mu_1 \neq \mu_2$ (for two-tailed test)

While one- tail test as:

H_0: $\mu_1 \leq \mu_2$ and H_1: $\mu_1 > \mu_2$ (Right-tailed)

H_0: $\mu_1 \geq \mu_2$ and H_1: $\mu_1 < \mu_2$ (Left-tailed)

Step II: After setting the null and alternative hypotheses, we establish a criterion for rejection or non-rejection of null hypothesis, that is, decide the level of significance (α), at which we want to test our hypothesis. The most common value of α is 0.05 or 5%. Other popular choices are 0.01 (1%) and 0.1 (10%).

Step III: The third step is to choose an appropriate test statistics form like t-test of any application.

Step IV: Obtain the critical value(s) in the sampling distribution of the test statistic and construct the rejection (critical) region of size α. Generally, critical values for various levels of significance are put in the form of a table for various standard sampling distributions of test statistics such as t-table of respective d.f ,etc.

Step V: After that, compare the calculated value of test statistic obtained from Step IV, with the critical value(s) obtained in Step V and locate the position of the calculated test statistic, that is, it lies in the rejection region or non-rejection region.

Step VI: ultimately testing the hypothesis, we have to conclude.

It is done as explained below:

(i) If the calculated test statistic value lies in the rejection region at the significance level, then we reject the null hypothesis. It means that the sample data provide us sufficient evidence against the null hypothesis and there is a significant difference between hypothesized value and observed value of the parameter.

(ii) If the calculated test statistic value lies in the non-rejection region at the significance level, then we do not reject the null hypothesis. It means that the sample data fails to provide sufficient evidence against the null hypothesis and the difference between hypothesized value and observed value of the parameter due to sample fluctuation.

Note: The decision about the null hypothesis is taken with the help of the p-value. The concept of p-value is very important because computer packages and statistical software such as SPSS, STATA, MINITAB, EXCEL, etc., all provide p-value.

2.4 TESTING OF HYPOTHESIS FOR POPULATION MEAN USING T-TEST

ASSUMPTIONS

When the standard deviation of a population is not known and the sample size is small so in this situation, we use a t-test provided the population under study is normal. Virtually every test has some assumptions which must be met before the application of the test. This t-test needs the following assumptions to work:

(i) The characteristic under study follows a normal distribution. In other words, populations from which a random sample is drawn should be normal for the characteristic of interest

(ii) Sample observations are random and independent.

(iii) Population variance σ 2 is unknown.

For describing this test, let X_1, X_2, ……, Xn be a random sample of small size n (< 30) selected from a normal population with mean μ and unknown variance σ_2. Now, follow the same procedure as we have discussed, that is, first of all, we set up the null and alternative hypotheses. Here, we want to test the claim about the specified value μ_0 of population means μ so we can take the null and alternative hypotheses as

take the null and alternative hypotheses as H_0: $\mu = \mu_0$

H_1: $\mu \neq \mu_0$ (for the two-tailed test)

While one- tail test as:

H_0: $\mu = \mu_0$ and H_1: $\mu > \mu_0$ (Right-tailed)

H_0: $\mu = \mu_0$ and H_1: $\mu < \mu_0$ (Left-tailed)

For testing the null hypothesis, the test statistic t is given by

$$t = \frac{\bar{X} - \mu_0}{s \, / \, \sqrt{n}}$$

where $\bar{X} = \frac{\sum X}{n}$ is the sample mean

$\bar{X}$ can be solved by A + $\frac{1}{n}\sum d$

$\sum$d = $\sum$ (X-A)

A= assumed mean

$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ is the sample variance

s^2 can be solved by $\frac{1}{n-1}\left(\sum d^2 - \frac{(\sum d)^2}{n}\right)$

Here, t-distribution with (n − 1) degrees of freedom.

After substituting values of X, S and n, we get the calculated value of test statistic t. Then we look for the critical value of test statistic t from the t-table. On comparing the calculated value and critical value(s), we decide on the null hypothesis.

Example 1: An electric tube producer claims that the average life of a particular category of electric tubes is 18000 km when used under normal driving conditions. A random sample of 16 electric tubes was tested. The mean and SD of life of the electric tubes in the sample were 20000 km and 6000 km respectively. Assuming that the life of the electric tubes is normally distributed, test the claim of the producer at a 1% level of significance.

Solution: Here, we are given that

n= 16, μ_0 = 18000, $\bar{X}$ = 20000, s = 6000

Here, we want to test that the producer's claim is true that the average life (µ) of electric tubes is 18000 km.

Step 1: Set up the null and alternative hypotheses as

H_0: μ_0 = 18000 (average life of electric tubes is 18000 km)

H_1: $\mu_0 \neq$ 18000 two-tailed (average life of electric tubes is not 18000 km)

Step 2: level of significance: As this is a one-tailed test,

α = 5%. This can be used to determine the critical value.

1 - α = 1 - 0.05 = 0.95, df= N-1= 6-1=5

Step 3: It is a small sample test; therefore t-test is to be determined as

$t = \frac{\bar{X}-\mu_0}{s\,/\,\sqrt{n}}$

Step 4: Determine the critical value

The critical value of test statistic t for two-tailed test corresponding (n-1) = 15 df at 1% level of significance are $\pm t_{(15), 1\,\%} = \pm 2.947$. Here, t-distribution with (n −1) degrees of freedom.

Step 5: Test Statistics

$t = \frac{20000-18000}{6000/\sqrt{16}}$

$t = \frac{2000}{1500} = 1.33$

Step 6: Conclusion

Since the calculated value of test statistic t (=1.33) is less than the critical (tabulated) value (= 2.947) and greater than the critical value (= − 2.947), that means a calculated value of test statistic lies in the non-rejection region, thus we do not reject the null hypothesis i.e. we support the producer's claim at 1% level of significance. Therefore, we conclude that the sample fails to provide sufficient evidence against the claim so we may assume that the producer's claim is true.

Example 2: (left-tail) The average score of a class is 90. However, a teacher believes that the average score might be lower. The scores of 6 students were randomly measured. The mean was 82 with a standard deviation of 18. With a 0.05 significance level use hypothesis testing to check if this claim is true.

Solution: Step 1: Set up the null and alternative hypotheses as

H_0: μ = 90,

Alternative hypothesis

H_1: μ < 90 (Left-tailed)

n = 6, s = 18

Step 2: level of significance:

As this is a one-tailed test,

α = 5%. This can be used to determine the critical value.

1 - α = 1 - 0.05 = 0.95, df= N-1= 6-1=5

Step 3: It is a small sample test; therefore t-test is to be determined as

$t = \frac{\bar{X}-\mu_0}{\frac{s}{\sqrt{n}}}$

Step 4: Determine the critical value

The critical value from the t table is -2.015

Step 5: Test Statistics

$t = \frac{\bar{X}-\mu_0}{\frac{s}{\sqrt{n}}}$

$t = \frac{82-90}{\frac{18}{\sqrt{6}}}$

t = -1.088

Step 6: Conclusion

As Cal t > Tab t (-1.088 > -2.015), therefore, fail to reject the null hypothesis. There is not enough evidence to support the claim.

2.5 TESTING OF HYPOTHESIS FOR DIFFERENCE OF TWO POPULATION MEANS USING T-TEST

When standard deviations of both populations are not known, in real-life problems t-test is more suitable compared to the Z-test.

Assumptions

This test works under the following assumptions:

a) It follows a normal distribution in both populations. Both populations from which random samples are drawn should be normal for the characteristics of interest.

b) Samples and their observations both are independent of each other.

c) Population variances σ_1^2 and σ_2^2 are both unknown but equal.

Let's we have to draw two independent random samples, X_1, X_2, ..., X_n and Y_1, Y_2, ..., Yn of sizes n1 and n2 from these normal populations. Let $\bar{X}$ and $\bar{Y}$ be the means of first and second sample respectively. Further, suppose the variances of both the populations are unknown but are equal, i.e., $\sigma_1^2 = \sigma_2^2 = \sigma$. In this case, σ^2 is estimated by value of pooled sample variance S^2

$$s_p^2 = \frac{1}{n_1+n_2-2}[(n_1-1)s_1^2+(n_2-1)s_2^2]$$

$$s_1^2 = \frac{1}{n_1-1}\sum(X-\overline{X})^2 \text{ and } s_2^2 = \frac{1}{n_2-1}\sum(Y-\overline{Y})^2$$

$$s_p^2 = \frac{1}{n_1+n_2-2}[\sum(X-\overline{X})^2+\sum(Y-\overline{Y})^2]$$

$$\bar{X} = A + \frac{\sum d_1}{n_1}$$

$$\bar{Y} = A + \frac{\sum d_2}{n_2}$$

d_1= (X-A_1)

d_2= (Y-A_2)

A_1 is assumed mean from series X

A_2 is assumed mean from series Y

$$s_p^2 = \frac{1}{n_1+n_2-2}[\sum d_1^2 - \frac{(\sum d_1)2}{n_1} + \sum d_2^2 - \frac{(\sum d_2)2}{n_2}]$$

Here, the hypotheses for a difference in two population means are similar to those for a difference in two population proportions. The null hypothesis, H_0, is again a statement of “no effect” or “no difference.”

- H_0: $\mu_1 - \mu_2 = 0$, which is the same as H_0: $\mu_1 = \mu_2$

The alternative hypothesis, H_a, can be any one of the following.

- H_a: $\mu_1 - \mu_2 < 0$, which is the same as H_a: $\mu_1 < \mu_2$
- H_a: $\mu_1 - \mu_2 > 0$, which is the same as H_a: $\mu_1 > \mu_2$
- H_a: $\mu_1 - \mu_2 \neq 0$, which is the same as H_a: $\mu_1 \neq \mu_2$

For testing the null hypothesis, the test statistic t is given by

$$t = \frac{\bar{X} - \bar{Y}}{S\sqrt{\frac{1}{n_1} - \frac{1}{n_2}}}$$

After substituting values of X, Y, S, n_1 and n_2 we get the calculated value of test statistic t. Then we look for critical (or tabulated) value(s) of test statistic t from the t-table. On comparing calculated value and critical value(s), we decide the null hypothesis either to accept or reject.

Example 3: Two different types of drugs A and B were tried on some patients for increasing their weights. Six persons were given drug A and other 7 persons were given drug B. The gain in weights (in ponds) is given below:

Drug A:	**5**	**8**	**7**	**10**	**9**	**6**	**–**
Drug B:	**9**	**10**	**15**	**12**	**14**	**8**	**12**

Assuming that increase in the weights due to both drugs follow normal distributions with equal variances, do the both drugs differ significantly with regard to their mean weights increment at 5% level of significance?

Solution: If μ_1 and μ_2 denote the mean weight increment due to drug A and drug B respectively then our claim is $\mu_1 = \mu_2$ and its complement is $\mu_1 \neq \mu_2$.

Since the claim contains the equality sign so we can take the claim as the null hypothesis and complement as the alternative hypothesis. Thus,

H_0: $\mu_1 = \mu_2$ [effect of both drugs is same]

H_1: $\mu_1 \neq \mu_2$ [effect of both drugs is not same]

Since the alternative hypothesis is two-tailed so the test is two-tailed test. Since it is given that increments in the weight due to both drugs follow normal distributions with equal and unknown variances and other assumptions of t-test for testing a hypothesis about difference of two population means also meet. So, we can go for this test. For testing the null hypothesis, the test statistic t is given by

$$t = \frac{\bar{X} - \bar{Y}}{S\sqrt{\frac{1}{n_1} - \frac{1}{n_2}}}$$

Assume, a = 8, b = 12 and use short-cut method to find X, Y and S

Drug A (X)			Drug B (Y)		
X	d_1 = (X- A_X) A_X=8		Y	d_2 = (Y- A_y) A_y= 12	
5	-3	9	9	-3	9
8	0	0	10	-2	4
7	-1	1	15	3	9
10	2	4	12	0	0
9	1	1	14	2	4
6	-2	4	8	-4	16
			12	0	0
$\sum X$ = 45	$\sum d_1$ = -3	$\sum d_1^2$ =19	$\sum Y$ =80	$\sum d_2$ = -4	$\sum d_2^2$ =42

$$\bar{X} = \frac{\sum X}{n} = \frac{45}{6} = 7.5$$

$$\bar{Y} = \frac{\sum Y}{n} = \frac{80}{7} = 11.43$$

$$S^2 = \frac{1}{n_1+n_2-2}\left(\sum {d_1}^2 - \frac{(\sum d1)^2}{n_1}\right) + \left(\sum {d_2}^2 - \frac{(\sum d2)^2}{n_1}\right)$$

$$= \frac{1}{6+7-2}\left(19 - \frac{(-3)^2}{6}\right) + \left(42 - \frac{(-4)^2}{7}\right)$$

$$= \frac{1}{11}\,(17.5 - 39.71)$$

$$= \sqrt{5.20}$$

S = 2.28

$$t = \frac{\bar{X} - \bar{Y}}{S\sqrt{\frac{1}{n_1} - \frac{1}{n_2}}}$$

$$t = \frac{7.5 - 11.43}{2.28\sqrt{\frac{1}{6} - \frac{1}{7}}}$$

$$= = \frac{-3.93}{2.28 \times 0.56}$$

$$= \frac{-3.93}{1.28}$$

= -3.07

The critical values of test statistic t for two-tailed test corresponding $(n_1 + n_2 - 2) = 11$ df at 5% level of significance are $\pm t_{(11),0.025} = \pm 2.201$. Since calculated value of test statistic t (= −3.07) is less than the critical values (= ± 2.201) that means calculated value of test statistic t lies in rejection region, so we reject the null hypothesis i.e. we reject the claim at 5% level of significance. Thus, we conclude that samples provide us sufficient evidence against the claim so drugs A and B differ significantly. Any one of them is better than other.

2.6 PAIRED t-TEST

Paired t-test gives a hypothesis examination of the difference between population means for a set of random samples whose variations are almost normally distributed. Subjects are often tested in a before-after situation or with subjects as alike as possible. The paired t-test is a test that the differences between the two observations are zero. For instance, a pharmaceutical company might create a new blood pressure-lowering medication. Twenty individuals had their blood pressure taken both before and after the medicine is administered for a month. To determine whether there is a statistically significant difference between pressure readings taken before and after taking the medication, analysts utilize a paired t-test.

Let us assume two paired sets, such as Xi and Yi for i = 1, 2, …, n such that their paired difference is independent which is identically and normally distributed. Then the paired t-test concludes whether they notably vary from each other. Here,

- Null hypothesis: The mean difference between pairs equals zero in the population ($\mu_D = 0$).
- Alternative hypothesis: The mean difference between pairs does not equal zero in the population ($\mu_D \neq 0$).

Assumptions

This test works under following assumptions:

- The population of differences follows normal distribution.
- Samples are not independent.
- Size of both the samples is equal.
- Population variances are unknown but not necessarily equal.

Let $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ be a paired random sample of size n and the difference between paired observations Xi & Yi be denoted by D_i, that is, $D_i = X_i - Y_i$ for I = 1, 2, …. n. Hence, we can assume that $D_1, D_2, \ldots, D_n$ be a random sample from normal population of differences with mean μ_D and unknown variance σ^2_D. This is same as the case of testing of hypothesis for population mean when population variance is unknown.

Here, we want to test that there is an effect of a diet, training, treatment, medicine, etc.

A paired samples t-test always uses the following null hypothesis:

- H_0: $\mu_1 = \mu_2$ or H_0**:** $\mu_{D} = \mu_1 - \mu_2$ (the two-population means are equal)

The alternative hypothesis can be either two-tailed, left-tailed, or right-tailed:

- H_1 (two-tailed): $\mu_1 \neq \mu_2$ or $\mu_D \neq 0$ (the two-population means are not equal)
- H_1 (left-tailed): $\mu_1 < \mu_2$ (population 1 mean is less than population 2 mean)
- H_1 (right-tailed): $\mu_1 > \mu_2$ (population 1 mean is greater than population 2 mean)

For testing null hypothesis, paired t statistic is given as:

$$t = \frac{\bar{D}}{s_D / \sqrt{n}}$$

$$\bar{D} = \frac{1}{n}\sum_{i=1}^{n} D_i$$

$$S_D^2 = \frac{1}{n-1}\sum_{i=1}^{n}(D_i - \bar{D})^2 = \frac{1}{n-1}\left[\sum D^2 - \frac{(\sum D)2}{n}\right]$$

After substituting values of D, S and n D we get calculated value of test statistic t. Then we look for critical (or cut-off or tabulated) value(s) of test statistic t from the t-table. On comparing calculated value and critical value(s), we take the decision about the null hypothesis.

Example 4: To verify whether the training programme improved performance of the laborers, a similar test was given to 10 laborers both before and after the programme. The original marks out of 100 (before training) recorded in an alphabetical order of the participants are

before training:	**42,**	**46,**	**50,**	**36,**	**44,**	**60,**	**62,**	**43,**	**70**	**53**
After training:	**45,**	**46,**	**60,**	**42,**	**60,**	**72,**	**63,**	**43,**	**80**	**65**

Assuming that performance of the before training and after follows normal distribution. Test whether the training programme has improved the performance of the laborers at 5% level of significance?

Solution: Here, we want to test whether the training programme has improved the performance of the laborer. Thus,

H_0: $\mu_1 = \mu_2$

H_1: $\mu_1 < \mu_2$ (left-tailed)

Since the alternative hypothesis is left-tailed so the test is left-tailed test.

It is a situation of before and after. Also, the marks of the students before and after the training programme follows normal distributions. Therefore, population of differences will also be normal. Also, all the assumptions of paired t-test meet. So, we can go for paired t-test. For testing the null hypothesis, the test statistic t is given by

$$t = \frac{\bar{D}}{S_D\ /\ \sqrt{n}}$$

Labours	X	Y	D = (X-Y)	D^2
1	42	45	-3	9
2	46	46	0	0
3	50	60	-10	100
4	36	42	-6	36
5	44	60	-16	256
6	60	72	-12	144
7	62	63	-1	1
8	43	43	0	0
9	70	80	-10	100
10	53	65	-12	144
			$\sum D$ = -70	$\sum D^2 = 790$

$$\bar{D} = \frac{1}{n}\sum_{i=1}^{n} D_i$$

$$= \frac{1}{10}\ (-70)$$

$$= -7$$

$$S_D^2 = \frac{1}{10-1}\left[\sum D^2 - \frac{(\sum D)2}{n}\right]$$

$$= \frac{1}{9}\left[790 - \frac{(-70)2}{10}\right]$$

$$= \frac{1}{9}\ [300]$$

$$=33.33$$

$$S_D = \sqrt{33.33}$$

$$S_D = 5.77$$

Putting the values in t-test, we have

$$t = \frac{\bar{D}}{S_D\ /\ \sqrt{n}}$$

$$t = \frac{-7}{5.77 / \sqrt{10}}$$

$$t = -3.38$$

The critical value of test statistic t for left-tailed test corresponding (n-1) = 9 df at 5% level of significance is – t (9), 0.05 = −1.833. Since calculated value of test statistic t (= −3.83) is less than the critical (tabulated) value (= −1.833), that means calculated value of test statistic t lies in rejection region, so we reject the null hypothesis and support the alternative hypothesis i.e. we support our claim at 5% level of significance. Thus, we conclude that samples fail to provide us sufficient evidence against the claim so we may assume that the participants have significant improvement after training programme.

2.7 TESTING OF HYPOTHESIS FOR POPULATION CORRELATION COEFFICIENT USING T-TEST

if two variables are related in such a way that change in the value of one variable affects the value of another variable then the variables are said to be correlated or there is a correlation between these two variables. Correlation can be positive, which means the variables move together in the same direction, or negative, which means they move in opposite directions. And correlation coefficient is used to measure the intensity or degree of linear relationship between two variables. The value of correlation coefficient varies between −1 and +1, where −1 representing a perfect negative correlation, 0 Small Sample Tests representing no correlation, and +1 representing a perfect positive correlation. Sometime, the sample data indicate for non-zero correlation but in population they are uncorrelated ($\rho = 0$). For example, price of tomato in Delhi (X) and in London (Y) are not correlated in population ($\rho = 0$). But paired sample data of 20 days of prices of tomato at both places may show correlation coefficient (r) $\neq 0$. In general, in sample data $r \neq 0$ does not ensure in population $\rho \neq 0$ holds. here, we will know how we test the hypothesis that population correlation coefficient is zero.

Assumptions

This test works under following assumptions:

(i) The characteristic under study follows normal distribution in both the populations. In other words, both populations from which random samples are drawn should be normal with respect to the characteristic of interest.

(ii) Samples observations are random.

Let us consider a random sample $(X_1, Y_1), (X_2, Y_2), \ldots, (Xn, Yn)$ of size n taken from a bivariate normal population. Let ρ and r be the correlation coefficients of population and sample data respectively. Here, we wish to test the hypothesis about population correlation coefficient (ρ), that is, linear correlation between two variables X and Y in the population, so we can take the null hypothesis as

H_0: $\rho = 0$ and H_1: $\rho \neq 0$ (two-tailed)

H_0: $\rho \leq 0$ and H_1: $\rho > 0$ (right -tailed)

H_0: $\rho \geq 0$ and H_1: $\rho < 0$ (left -tailed)

Here t statistic is as: $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$

here n-2 degrees of freedom

After substituting values of r and n, we find out calculated value of t-test statistic. Then we look for critical (or cut-off or tabulated) value(s) of test statistic t from the t-table. On comparing calculated value and critical value(s), we take the decision about the null hypothesis. Let us do some examples of testing of hypothesis that population correlation coefficient is zero.

Example 5: 20 families were selected randomly from Area A group to determine that correlation exists between family income and the amount of money spent per family member on food each month. The sample correlation coefficient (r) was computed as r = 0.40. By follow the normal distributions, test that there is a positive linear relationship between the family income and the amounts of money spent per family member on food each month in Area A group at 1% level of significance.

Solution: here, n = 20, r = 0.40, and to test that there is a positive linear relationship between the family income and the amounts of money spent per family member on food each month in area A group. If ρ denote the correlation coefficient between the family income and the amounts of money spent per family member then the claim is $\rho > 0$ and its complement is $\rho \leq 0$. Since complement contains the equality sign so we can take the complement as the null hypothesis and the claim as the alternative hypothesis.

Thus,

H_0: $\rho \leq 0$

H_1: $\rho > 0$ (right -tailed)

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

$$t= \frac{0.40\sqrt{20-2}}{\sqrt{1-(0.40)^2}}$$

$$t= \frac{0.40 \times 4.24}{0.92} = 1.84$$

The critical value of test statistic t for right-tailed test corresponding (n-2) = 18 df at 1% level of significance is $t_{(n-2),\alpha} = t_{(18),0.01} = 2.552$. Since calculated value of test statistic t (=1.84) is less than the critical value (= 2.552), it calculated value of test statistic t lies in non-rejection region, so we do not reject the null hypothesis and reject the alternative hypothesis *i.e.* we reject our claim at 1% level of significance. Thus, we conclude that sample provide us sufficient evidence against the claim so there is no positive linear correlation between the family income and the amounts of money spent per family member on food each month in area A group.

2.8 SUM UP

Since in many of the problems it becomes necessary to take a small size sample, considerable attention has been paid in developing suitable tests for dealing with problems of small samples. The greatest contribution to the theory of small samples is that of Sir William Gosset and Prof. R.A. Fisher. Sir William Gosset published his discovery in 1905 under the pen name 'Student' and later on developed and extended by Prof. R.A. Fisher. He gave a test popularly known as 't-test'. The t-distribution has a number of applications in statistics, t-test for significance of single mean, t-test for significance of the difference between two sample means, independent samples, paired t-test.

2.9 QUESTIONS FOR PRACTICE

Q1. Explain the need of small sample tests.

Q2. List out the assumptions of t-test.

Q3. List out the Procedure of testing a hypothesis for t-test.

Q4. Explain the testing of hypothesis for population mean using t-test.

Q5. Describe testing of hypothesis for difference of two population means when samples are independent using t-test.

Q6. Explain the procedure of paired t-test for testing of hypothesis for difference of two population means when samples are dependent or paired.

2.10 SUGGESTED READINGS

- C.R. Kothari (1990) Research Methodology. Vishwa Prakasan. India.
- Gupta SC: Fundamental of statistics, S. Chand & Company. New Delhi.

- Gupta, SP: Statistical Methods, S. Chand & Company. New Delhi.
- A.M Goon, M.K Gupta and B. Dasgupta, fundamental of statistics Vol-I, World press Calcutta.
- Anderson, D.R.; Sweeney, D.J. and Williams, T.A., “Statistics for Business and Economics”, 2nd edition (2011), Thompson, New Delhi.

CERTIFICATE/ DIPLOMA IN STATISTICAL ANALYSIS AND RESEARCH METHODOLOGY

SEMESTER II

SARM 5: STATISTICAL INFERENCE

UNIT 3: LARGE SAMPLE TEST

STRUCTURE

3.0 OBJECTIVES

After studying this unit, learners should be able to know:

- Meaning of large sample test
- Applying the Z-test to test the hypothesis about the population mean and the difference between the two population means
- proportion and difference of two population proportions
- Applying the Z-test to test the hypothesis about the population variance and two population variances.

3.1 INTRODUCTION

Sometimes in our studies in economics, psychology, medicine, etc., we take a sample of objects/units/participants/patients, etc. such as 70, 500, 1000, 10,000, etc. This situation comes under the category of large samples. As a thumb rule, a sample of size n is treated as a large sample only if it contains more than 30 units (or observations, $n > 30$). We know that for large samples ($n > 30$), one statistical fact is that almost all sampling distributions of the statistic(s) are closely approximated by the normal distribution. Therefore, the test statistic, a function of sample observations based on $n > 30$, could be assumed to follow the normal distribution approximately (or exactly).

In other words, we have seen that for large values of n, the number of trials, almost all the distributions e.g., Binomial, Poisson, etc. are very closely approximated by Normal distribution and in this case, we apply Normal Deviate test (Z-test). In cases where the population variance (s) is/are known, we use Z-test. The distribution of Z is always normal with mean zero and variance one. In statistics, a sample is said to be large if its size exceeds 30.

ASSUMPTIONS FOR A SINGLE SAMPLE Z-TEST

Every statistical method has assumptions. Assumptions mean that your data must satisfy certain properties for statistical method results to be accurate.

The assumptions for the Single Sample Z-Test include:

1. Continuous
2. Normally Distributed
3. Random Sample
4. Enough Data
5. Known Population

5.2 STEPS OF T-TEST TESTING OF HYPOTHESIS

Suppose X_1, X_2, …, X_n is a random sample of size n (> 30) selected from a population having unknown parameter θ and we want to test the hypothesis about the hypothetical / claimed/assumed value θ_0 of parameter θ. For this, a test procedure is required. We discuss it step by step as follows: t-test involves the following steps for testing this hypothetical value:

Step I: First of all, we have to set up null hypothesis H_0 and alternative hypothesis H_1.

Suppose, we want to test the hypothetical / Testing of the Hypothesis assumed value μ_0 of parameter θ.

So, we can take the null and alternative hypotheses as H_0: $\theta = \theta_0$

H_1: $\theta \neq \theta_0$ (for the two-tailed test)

While one- tail test as:

H_0: $\theta = \theta_0$ and H_1: $\theta > \theta_0$ (Right-tailed)

H_0: $\theta = \theta_0$ and H_1: $\theta < \theta_0$ (Left-tailed)

In case of comparing the same parameter of two populations of interest, say, θ_1 and θ_2, then our null and alternative hypotheses would be

H_0: and $\theta_1 = \theta_2$ and H1: $\theta_1 \neq \theta_2$ (for two-tailed test)

While one- tail test as:

H_0: $\theta_1 \leq \theta_2$ and H_1: $\theta_1 > \theta_2$ (Right-tailed)

H_0: $\theta_1 \geq \theta_2$ and H_1: $\theta_1 < \theta_2$ (Left-tailed)

Step II: After setting the null and alternative hypotheses, we establish a criterion for rejection or non-rejection of null hypothesis, that is, decide the level of significance (α), at which we want to test our hypothesis. The most common value of α is 0.05 or 5%. Other popular choices are 0.01 (1%) and 0.1 (10%).

Step III: Third step is to determine an appropriate test statistic, say, Z in case of large samples. Suppose Tn is the sample statistic such as sample mean, sample proportion, sample variance, etc. for the parameter θ then for testing the null hypothesis, test statistic is given by

$$Z= \frac{t-E\ (t)}{S.E.(t)}$$

Step IV: Obtain the critical value(s) in the sampling distribution of the test statistic and construct the rejection (critical) region of size α. Generally, critical values for various levels of significance are put in the form of a table for various standard sampling distributions of test statistics such as Z-table.

Step V: After that, we obtain the critical (cut-off or tabulated) value(s) in the sampling distribution of the test statistic Z corresponding to α assumed in Step II. These critical values are given in Table A (Z-table) this course corresponding to different levels of significance (α). For convenience, some useful critical values at α = 0.1, 0.01 and 0.05 for Z-test are given in Table A (mentioned below). After that, we construct a rejection (critical) region of size α in the probability curve of the sampling distribution of test statistic Z.

Step VI: ultimately testing the hypothesis, we have to conclude.

(i) Case I: When H_0: $\theta \leq \theta_0$ and H_1: $\theta > \theta_0$ (right-tailed test)

Now, if z (calculated value) $\geq Z\alpha$ (tabulated value), that means the calculated value of test statistic Z lies in the rejection region, then we reject the null hypothesis H_0 at α the level of significance. Therefore, we conclude that sample data provides us sufficient evidence against the null hypothesis and there is a significant difference between hypothesized or specified value and the observed value of the parameter. If $Z < Z\alpha$, that means the calculated value of test statistic Z lies in non-rejection region, then we do not reject the null hypothesis H_0 at α level of significance. Therefore, we conclude that the sample data fails to provide us sufficient evidence against the null hypothesis and the difference between hypothesized value and observed value of the parameter due to the fluctuation of a sample.

so, the population parameter θ may be θ_0.

Case II: When H_0: $\theta \geq \theta_0$ and H_1: $\theta < \theta_0$ (left-tailed test)

In this case, the rejection (critical) region falls under the left tail of the probability curve of the sampling distribution of test statistic Z. If $Z \leq -Z\alpha$, that means the calculated value of test statistic Z lies in the rejection region, then we reject the null hypothesis H_0 at α level of significance. If $Z > -Z\alpha$, that means the calculated value of test statistic Z lies in the non-rejection region, then we do not reject the null hypothesis H_0 at α level of significance.

In case of two-tailed test: H_0: $\theta = \theta_0$ and H_1: $\theta \neq \theta_0$

In this case, the rejection region falls under both tails of the probability curve of the sampling distribution of the test statistic Z. Half the area (α) i.e. $\alpha/2$ will lie under the left tail and the other half under the right tail. If $Z \geq Z_{\alpha/2}$ or $Z \leq - Z_{\alpha/2}$, that means the calculated value of test statistic Z lies in the rejection region, then we reject the null hypothesis H_0 at α level of significance. If $-Z < Z < - Z_{\alpha/2}$, that means the calculated value of test statistic Z lies in the non-rejection region, then we do not reject the null hypothesis H0 at α level of significance.

Table A: Critical Values of Z-test

Level of Significance (%)	Two-tailed	Right-tailed	Left-tailed
1	2.58	2.33	-2.33
5	1.96	1.645	-1.645
10	1.645	1.28	-1.28

5.3 APPLICATIONS OF LARGE SAMPLE TEST

- Test for a single proportion.
- Test for significance of difference of proportions.
- Test of significance for a single mean.
- Test of significance for difference of means.

5.4 TEST FOR SINGLE PROPORTION

For this purpose, let X_1, X_2, ..., X_n be a random sample of size n taken from a population with population proportion P. Also, let X denote the number of observations or elements that possess a certain attribute (number of successes) out of n observations of the sample then sample proportion p can be defined as

$$p = \frac{X}{n}$$

here mean and variance of the sampling distribution of sample proportion are E(p) = P and Var(p) $= \frac{\mathrm{PQ}}{\mathrm{n}}$ where, Q = 1– P.

Now, two cases arise: Large Sample Tests

Case I: When the sample size is not sufficiently large i.e. either of the condition's np > 5 or nq > 5 does not meet, then we use the exact binomial test. However, an exact binomial test is beyond the scope of this course.

Case II: When the sample size is sufficiently large, such that np > 5 and nq > 5 then by central limit theorem, the sampling distribution of sample proportion p is approximately normally distributed with mean and variance as

$$E(p) = P \text{ and } Var(p) = \frac{\mathrm{PQ}}{\mathrm{n}}$$

But we know that standard error = Variance

$$SE(p) = \sqrt{\frac{\mathrm{PQ}}{\mathrm{n}}}$$

Now, first of all, we set up null and alternative hypotheses. Here we want to test the hypothesis about the specified value P_0 of the population proportion. So, we can take the null and alternative hypotheses as

For two-tailed

H_0: P= P_0

H_1: $P \neq P_0$

For one-tailed

H_0: P = P_0

H_1: P > P_0

Or

H_1: P < P_0

$$Z = \frac{p - P_0}{\sqrt{\frac{PQ}{n}}}$$

After that, we calculate the value of the test statistic and compare it with the critical value(s) given in below Table A at a prefixed level of significance α.

Table A: Critical Values of Z-test

Level of Significance (%)	Two-tailed	Right-tailed	Left-tailed
1	2.58	2.33	-2.33
5	1.96	1.645	-1.645
10	1.645	1.28	-1.28

Example 1: A die is thrown 9000 times and a draw of 2 or 5 is observed 3100 times. Can we regard that die as unbiased at a 5% level of significance?

Solution: Let getting a 2 or 5 be our success, and getting a number other than 2 or 5 be a failure then in usual notions, we have n = 9000, X = number of successes = 3100, p = 3100/9000 = 0.3444 Here, we want to test that the die is unbiased and we know that if the die is Large Sample Tests unbiased then the proportion or probability of getting 2 or 5 is

P = Probability of getting a 2 or 5

= Probability of getting 2 + Probability of getting 5

$\frac{1}{6} + \frac{1}{3} + \frac{1}{3} = 0.3333$

So, our claim is P = 0.3333 and its complement is P ≠ 0.3333. Since the claim contains the quality sign. Thus, we can take the claim as the null hypothesis and complement as the alternative hypothesis. Thus, H_0: P = P_0= 0.3333 and

H_1 :P ≠ 0.3333

Since the alternative hypothesis is two-tailed so the test is two-tailed. Before proceeding further, first, we have to check whether the condition of normality meets or not.

np = 9000 ×0.3444 = 3099.6 > 5

nq = 9000 × (1 - 0.3444) = 9000 × 0.6556 = 5900.4 > 5

We see that the condition of normality meets, so we can go for Z-test. So, for testing the null hypothesis, the test statistic Z is given by

$$Z = \frac{p-P_0}{\sqrt{\frac{PQ}{n}}}$$

$$Z = \frac{0.3444-0.3333}{\sqrt{\frac{0.3333 \times 0.6667}{9000}}}$$

$$= \frac{0.0111}{0.005} = 2.22$$

Since the test is two-tailed so the critical values at 5% level of significance are ± $Z_{\alpha/2}$ = ± $Z_{0.025}$ = ±1.96. Since calculated value of Z (= 2.22) is greater than the critical value (= 1.96), that means it lies in the rejection region, so we reject the null hypothesis i.e. we reject our claim.

Example 2: A shop claims that only, 1% of its products are imperfect. Out of sample of 500 units 10 are found to be imperfect. Check the claim of the shop.

Solution: Let us set up the null hypothesis that the proportion of defective products is j against the alternative hypothesis that it is not equal to 1%. i.e.

H_0: P=0.01

H_1: P ≠ 0.01

Q =1-P=1-0.01= 0.99

Size of sample = n = 500

Number of imperfect items found =X = 10

Therefore, sample proportion of imperfect items = p = $\frac{X}{n}$ = $\frac{10}{500}$ = = 0.02

The test statistic is,

$$Z = \frac{p-P_0}{\sqrt{\frac{PQ}{n}}}$$

$$= \frac{0.02-0.01}{\sqrt{\frac{0.01 \times 0.99}{500}}}$$

$$= \frac{0.01}{0.00445}$$

= 2.25

The calculated value of Z is more than 1.96 (Critical value at 5% level), and H_0 is rejected. Thus, the claim of the shop is rejected.

3.5 TEST FOR SIGNIFICANCE OF DIFFERENCE OF PROPORTIONS

If we have two populations and each item of a population belongs to either of the two classes C_1 and C_2. A person is often interested to know whether the proportion of items in class C_1 in both the populations is the same or not that is we want to test the hypothesis.

H_0: $P_1=P_2$

H_1: $P_1 \neq P_2$ or $P_1>P_2$ or $P_1<P_2$

where P_1 and P_2 are the proportions of items in the two populations belonging to class C_1.

Let X_1, X_2 be the number of items belonging to class C_1 in random samples of sizes n_1 and n_2 from the two populations respectively. Then the sample proportion

$$p_1 = \frac{X_1}{n_2}$$

$$p_2 = \frac{X_2}{n_2}$$

If P_1 and P_2 are the proportions then

$E\ (P_1) = P_1$, $E\ (P_2) = P_2$

$$\text{Var}\ (p_1) = \frac{P_1 Q_1}{n_1}$$

$$\text{Var}\ (p_2) = \frac{P_2 Q_2}{n_2}$$

Since $P_1 = P_2 = P$ and $Q_1 = Q_2 = Q$, therefore

$$Z = \frac{p_1 - p_2}{\sqrt{P \times Q\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

If the population proportion P_1 and P_2 are given to be distinctly different that is $P_1 \neq P_2$, then

$$Z= \frac{(p_1-p_2)-(P_1-P_2)}{\sqrt{\frac{P_1 Q_2}{n_1}+\frac{P_1 Q_2}{n_2}}}$$

In general P, the common population proportion (under H_o) is not known, then an unbiased estimate of population proportion P based on both the samples is used and is given by

$$\hat{p} = \frac{X_1+X_2}{n_1+n_2} = \frac{n_1 p_2+n_1 p_2}{n_1+n_2}$$

Example 3. A machine turns out 16 imperfect items in a sample of 500. After overhauling it turns 3 imperfect articles in a batch of 100. Has the machine improved after overhauling?

Solution: We are given n_1=500and n_2=100

p_1= Proportions of imperfect items before overhauling of machine =16/500=0.032

p_2= Proportions of imperfect items after overhauling of machine =3/100=0.03

H_0: $P_1=P_2$ i.e. the machine has not improved after overhauling.

H_1: $P_1>P_2$ or $P_2<P_1$

Here,

$$\hat{p} = \frac{X_1+X_2}{n_1+n_2} = \frac{n_1p_2+n_1p_2}{n_1+n_2}$$

$$= \frac{16+3}{500+100} = 0.032$$

$$Z = \frac{p_1 - p_2}{\sqrt{P \times Q\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

$$Z = \frac{0.032-0.03}{\sqrt{0.032 \times 0.968}\left(\frac{1}{500}+\frac{1}{100}\right)}$$

$$Z = \frac{0.002}{\sqrt{0.031}\left(\frac{1+5}{500}\right)}$$

$$= \frac{0.002}{0.01878}$$

$$= 0.106$$

Since Z<1.645 (Right-tailed test), it is not significant at 5% level of significance. Hence, we accept the null hypothesis and conclude that the machine has not improved after overhauling.

3.6 TEST OF SIGNIFICANCE FOR A SINGLE MEAN

We have seen that if X_i (i=1, 2, ..., n) is a random sample of size n from a normal population with mean and variance σ^2, then the sample mean $\bar{X}$ is distributed normally with mean μ and variance σ^2/n i.e., $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$.Thus for large samples normal variate corresponding to $\bar{X}$ is

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

In test of significance for a single mean we deal the following situations

1) To test if the mean of the population has a specified value (μ_0) and null hypothesis in this case will be H_0: $\mu=\mu_0$ i.e., the population has a specified mean value.
2) To test whether the sample mean differs significantly from the hypothetical value of population mean with null hypothesis as there is no difference between sample mean ($\bar{X}$) and population mean (μ).
3) To test if the given random sample has been drawn from a population with specified mean μ_0 and variance σ^2 with null hypothesis the sample has been drawn from a normal population with specified mean μ_0 and variance σ^2

In all the above three situations the test statistic is given by

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

If |Z| <1.96, H_o is not rejected at 5% level of significance which implies that there is no significant difference between sample mean and population mean and whatever difference is there, it exists due to fluctuation of sampling.

|Z| >1.96, H_o is rejected at 5% level of significance which implies that there is a significant difference between sample mean and population mean.

Example 4. A random sample of 100 workers gave a mean weight of 64 kg with a standard deviation of 16 kg. Test the hypothesis that the mean weight in the population is 60 kg.

Solution: H_0: μ =60 kg., i.e. the mean weight in the population is 60 kg.

H_1: μ ≠ 60 kg., i.e. the mean weight in the population is not 60 kg.

here, n=100, μ=60 kg., $\bar{X}$ =64 kg.,

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

$$= \frac{64 - 60}{16 / \sqrt{100}}$$

$$= 2.5$$

Since calculated value of Z statistic is more than 1.96, it is significant at 5% level of significance. Therefore, H_0 is rejected at all levels of significance which implies that mean weight of population is not 60 kg.

Example 5: A sample of 900 rods has a mean length 3.4 cm. Is the sample regarded to be taken from a large population of rods with mean length 3.25 cm and S.D 2.61 cm at 5% level of significance?

Solution: Here, n = 900, $\bar{X}$ =3.4 cm, μ =3.25 cm and σ = 2.61 cm

Thus, H_0: μ =μ_0 = 3.25

H_1: μ ≠ 3.25 (two-tailed)

Here, we want to test the hypothesis regarding population mean when Large Sample Tests population SD is unknown, so we should use t-test if the population of rods known to be normal. But it is not the case. Since the sample size is large (n > 30) so we can go for Z-test instead of t-test as an approximate. So, test statistic is given by

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

$$= \frac{3.40 - 3.25}{2.61 / \sqrt{900}}$$

$= \frac{0.15}{0.087}$

$= 1.72$

The critical (tabulated) values for two-tailed test at 5% level of significance are ± $Z_{\alpha/2}$ = ± $Z_{0.02}$5 = ± 1.96. Since calculated value of test statistic Z (=1.72) is less than the critical value (=1.96) and greater than critical value (= −1.96), that means it lies in non-rejection region, so we do not reject the null hypothesis i.e. we support the claim at 5% level of significance.

Thus, we conclude that sample does not provide us sufficient evidence against the claim so we may assume that the sample comes from the population of rods with mean 3.25cm.

3.7 TEST OF SIGNIFICANCE FOR DIFFERENCE OF MEANS

Let $\overline{X_1}$ be the mean of a sample of size n_1 drawn from a population with mean μ_1 and variance σ_1^2 and let $\overline{X_2}$ be the mean of an independent sample of size n_2 drawn from another population with mean μ_2 and variance σ_2^2. Since sample sizes are large.

The co-variance terms vanish, since the sample means $\overline{X_1}$, $\overline{X_2}$ are independent.

Thus, under H_o: $\mu_1 = \mu_2$, the Z statistic is given by

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

Here σ_1^2 and σ_2^2 are assumed to be known. If they are unknown then their estimates provided by corresponding sample variances s_1^2 and s_2^2 respectively are used, i.e., $\widehat{\sigma_1^2}= s_1^2$ and $\widehat{\sigma_2^2}= s_2^2$, thus, in this case the test statistic becomes

$$Z= \frac{\bar{X}_1-\bar{X}_2}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$$

Remarks: If we want to test whether the two independent samples have come from the same population i.e., if $\sigma_1^2 = \sigma_2^2 = \sigma^2$ (with common S.D. σ), then under H_o : $\mu_1 = \mu_2$

$$Z= \frac{\bar{X}_1-\bar{X}_2}{\sqrt{\sigma^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

If the common variance σ^2 is not known, then we use its estimate based on both the samples which is given by

$$\widehat{\sigma}^2 = \frac{n_1 s_1^2+ n_2 s_2^2}{n_1+n_2}$$

Example 6: A university conducts both face to face and regular classes for a particular course indented both to be identical. A sample of 50 students of face-to-face mode yields examination results mean and SD respectively as: $\bar{X}_1$ = 80.4, S_1 = 12.8 and other sample of 100 regular students yields mean and SD of their examination results in the same course respectively as: $\bar{X}_2$= 74.3, S_2 = 20.5, Are both educational methods statistically equal at 5% level?

Solution: Here, we are given that

$n_1 = 50, \bar{X}_1=80.4, S_1 =12.8$

$n_2 = 100, \bar{X}_2=74.3, S_2 =20.5$

We wish to test that both educational methods are statistically equal. If μ_1 and μ_2 denote the average marks of face to face and distance mode students respectively then our claim is $\mu_1 = \mu_2$ and its complement is $\mu_1 \neq \mu_2$. Since the claim contains the equality sign so we can take the claim as the null hypothesis and complement as the alternative hypothesis. Thus,

$H_0: \mu_1 = \mu_2$

$H_1: \mu_1 \neq \mu_2$ (two-tailed)

We want to test the null hypothesis regarding two population means when σ standard deviations of both populations are unknown. So, we should go for t-test if population of difference is known to be normal. But it is not the case.

Since sample sizes are large (n_1, and $n_2 > 30$) so we go for Z-test. For testing the null hypothesis, the test statistic Z is given by

$$Z= \frac{\bar{X}_1-\bar{X}_2}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$$

$$= \frac{\bar{X}_1-\bar{X}_2}{\sqrt{\frac{12.8^2}{50}+\frac{20.5^2}{100}}}$$

$$= \frac{6.1}{\sqrt{3.28+4.20}}$$

$$= 2.23$$

The critical (tabulated) values for two-tailed test at 5% level of significance are $\pm Z_{\alpha/2} = \pm Z_{0.025} = \pm 1.96$. Since calculated value of Z (=2.23) is greater than the critical values (= ±1.96), that means it lies in rejection region, so we reject the null hypothesis i.e. we reject the claim at 5% level of significance.

Example 7: Two research laboratories have identically produced medicines that provide relief to thyroid patients. The first medicine was tested on a group of 50 thyroid patients and produced an average 8.3 hours of relief with a standard deviation of 1.2 hours. The second medicine was tested on 100 patients, producing an average of 8.0 hours of relief with a standard deviation of 1.5 hours. Do the first medicines provide a significant longer period of relief at a significant level of 5%?

Solution: $n_1 = 50$, $\bar{X}_1 = 8.3$, $S_1 = 1.2$

$n_2 = 100$, $\bar{X}_2 = 8.0$, $S_2 = 1.5$

Here, we want to test that the first medicines provide a significant longer period of relief than the other. If μ_1 and μ_2 denote the mean relief time due to first and second medicines respectively then our claim is $\mu_1 > \mu_2$ and its complement is $\mu_1 \le \mu_2$. Since complement contains the equality sign so we can take the complement as the null hypothesis and the claim as the alternative hypothesis.

Thus, $H_0: \mu_1 = \mu_2$

$H_1: \mu_1 > \mu_2$

Since the alternative hypothesis is right-tailed so the test is right-tailed test.

We want to test the null hypothesis regarding equality of two population means. The standard deviations of both populations are unknown. So, we should go for t-test if population of difference is known to be normal. But it is not the case. Since sample sizes are large (n_1, and $n_2 > 30$) so we go for Z-test. So, for testing the null hypothesis, the test statistic Z is given by

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

$$= \frac{8.3 - 8.0}{\sqrt{\frac{1.2^2}{50} + \frac{1.5^2}{100}}}$$

$$= \frac{0.3}{\sqrt{0.0288 + 0.0255}}$$

$$= \frac{0.3}{0.2265}$$

$$Z = 1.32$$

The critical (tabulated) value for right-tailed test at 5% level of significance is $Z_\alpha = Z_{0.05} = 1.645$. Since calculated value of test statistic Z (=1.32) is less than the critical value (=1.645), that means it lies in non-rejection region, so we do not reject the null hypothesis and reject the alternative hypothesis i.e. we reject the claim at 5% level of significance.

Thus, the samples provide sufficient evidence against the claim so the first medicines do not have longer period of relief than the other.

3.8 COMPARISON OF Z-TEST (Large Sample Test) AND t-TEST (Small Sample Test)

Basis for comparison	t-test	Z-test
Definition	When the population's standard deviation is unknown, the t-test is a statistical test that is used to evaluate hypotheses about the mean of a small sample taken from the population.	A statistical technique called the z-test is used to compare or assess the significance of various statistical measures, most notably the mean in a sample taken from a population that is normally distributed or between two independent samples.
Sample size	n ≤ 30	n > 30
Assumptions	A t-test is not based on the assumption that all key points on the sample are independent.	z-test is based on the assumption that all key points on the sample are independent.
Population variance	Unknown	known
Variance or standard deviation	Variance or standard deviation is not known in the t-test.	Variance or standard deviation is known in z-test.
Distribution	The sample values are to be recorded or calculated by the researcher.	In a normal distribution, the average is considered 0 and the variance as 1.
Population parameters	In addition, to the mean it compares partial or simple correlations among two samples.	In addition, to mean, it compares the population proportion.

3.9 SUM UP

Z-test is a statistical test that is used to determine whether the mean of a sample is significantly different from a known population mean when the population standard deviation is known. It is particularly useful when the sample size is large (>30). Z-test can also be defined as a statistical method that is used to determine whether the distribution of the test statistics can be approximated using the normal distribution or not. It is the method to determine whether two sample means are approximately the same or different when their variance is known and the sample size is large (should be >= 30). The Z-test compares the difference between the sample mean and the population means by considering the standard deviation of the sampling distribution.

3.10 SUGGESTED READINGS

- C.R. Kothari (1990) Research Methodology. Vishwa Prakasan. India.
- Gupta SC: Fundamental of statistics, S. Chand & Company. New Delhi
- Gupta, SP: Statistical Methods, S. Chand & Company. New Delhi
- A.M Goon, M.K Gupta and B. Dasgupta, fundamental of statistics Vol-I, World press Calcutta
- Anderson, D.R.; Sweeney, D.J. and Williams, T.A., "Statistics for Business and Economics", 2nd edition (2011), Thompson, New Delhi.

CERTIFICATE/ DIPLOMA IN STATISTICAL ANALYSIS AND RESEARCH METHODOLOGY

SEMESTER II

SARM 5: STATISTICAL INFERENCE

UNIT 4: F-STATISTICS: MEANING, EQUITY OF POPULATION VARIANCES

STRUCTURE

5.0 LEARNING OBJECTIVES

After reading this unit, the learner should know about:

- Assumptions of F-test
- Applications of F-test
- describe the testing of the hypothesis for population variance
- explain the testing of the hypothesis for two population variances.

5.1 INTRODUCTION

F-statistics, a foundation in statistical methodology, is significant in scientific inquiry, particularly within the context of Analysis of Variance (ANOVA) and regression analyses. This serves as a powerful tool for researchers aiming to explain the presence of meaningful differences among multiple groups.

F-statistic is a ratio that compares the variability among group means to the variability within groups. In ANOVA, a statistical technique widely used for comparing means across more than two groups, the F-statistic provides a comprehensive assessment of whether observed differences in means are statistically significant. The formula for the F-statistic involves calculating the ratio of the variance between groups to the variance within groups. Spontaneously, a high F-statistic suggests that the differences among group means are more substantial than what could be attributed to random chance alone.

The process of interpreting F-statistics involves null hypothesis testing, a fundamental concept in statistical inference. Researchers collect data, perform the necessary calculations to derive the F-statistic, and then compare it to a critical (called p-value). The p-value represents the probability of obtaining the observed results, or more extreme results, under the assumption that the null hypothesis is true. A low p-value (typically below a predetermined significance level, such as 0.05) leads to the rejection of the null hypothesis, indicating that there are statistically significant differences among the groups. In statistical analyses such as ANOVA or regression, the F-test assesses the equality of variances or means, as it is derived from the F-distribution. In hypothesis testing, the rejection zone is determined by critical values derived from the F-distribution, considering the degrees of freedom and significance level. An F-distribution can be found when two independent chi-square variables are split by their corresponding degrees of freedom.

The F-test is a statistical test used to compare variances or test the equality of means among different groups. There are two main F-test types: one-way analysis of variance (ANOVA) and two-way ANOVA. The objectives and assumptions of the F-test can vary slightly depending on the specific context. The objectives of the F-test are

- Testing Equality of Variances (One-Way ANOVA): to determine whether there are significant differences in the variances among multiple groups.

For Example, testing if the variances of test scores are equal across different teaching methods.

- Testing Equality of Means (One-Way or Two-Way ANOVA): to assess if there are significant differences in means among multiple groups.

For Example: Comparing the average scores of students across different schools or treatment groups.

5.2 ASSUMPTIONS OF F-TEST

1. **Normality**: The data within each group or sample should be approximately normally distributed. This assumption is more critical when sample sizes are small.
2. **Homogeneity of Variances**: The variances of the populations from which the samples are drawn should be equal. This is crucial for the validity of the F-test results. In one-way ANOVA, this assumption is specifically about the equality of variances across groups.
3. **Independence**: Observations within each group should be independent of each other. The values in one group should not be dependent on or related to the values in another group.
4. **Random Sampling**: The data should be collected through a random sampling process to ensure that the sample is representative of the population.
5. **Independence of Errors**: this implies that the variation of each item around the group should be independent for each value.

Note:

- Violation of assumptions can impact the reliability of F-test results. In cases where normality or homogeneity of variances assumptions are not met, alternative tests or transformations of the data may be considered.
- The F-test is sensitive to outliers, and non-parametric alternatives may be more appropriate in such cases.

It's important to tailor the objectives and assumptions to the specific context of the F-test being conducted, whether it's a one-way ANOVA, two-way ANOVA, or another variant of the test.

5.3 MAIN PROPERTIES OF F DISTRIBUTION

The F-distribution depends on the degrees of freedom and is usually defined as the ratio of variances of two populations normally distributed and therefore it is also called Variance Ratio

Distribution**.**

1. The F-distribution is positively skewed and with the increase in the degrees of freedom m and n, its skewness decreases.
2. The value of the F-distribution is always positive, or zero since the variances are the square of the deviations and hence cannot assume negative values. Its value lies between 0 and ∞.
3. The shape of the F-distribution depends on its parameters ν_1 and ν_2 degrees of freedom.
4. Mean =$\frac{n}{n-2}$

   Mean is defined for n>2 and is independent of m. Mean of F is always positive
5. Variance =$\frac{2n^2(m+n+2)}{m\ (n-2)^2\ (n-4)}$, n > 4

   Variance is always positive and is defined for n > 4.
6. Mode = $\frac{(\mathrm{m}-2)}{m\ (n+2)}$

   Mode is defined for n > 2 and is always less than 1.
7. Karl Pearson's coefficient of skewness

   S_k= $\frac{\text{Mean}-\text{Mode}}{S.D}$ > 0

   Since mean > 1 and Mode < 1.
8. F-distribution is positively skewed.
9. The probability curve of F firstly increases rapidly and reaches its maximum at mode (which is less than 1). Then it falls slowly and becomes an asymptote to the X-axis.
10. If a statistic F follows F distribution with degrees of freedom (m, n), then its reciprocal, 1/F follows F distribution with (n, m) degrees of freedom.
11. distribution tends to be the normal distribution for large (m, n).
12. Critical region of F distribution: The F-test is always a right-tailed test and as such the critical region always lies on the right tail of the distribution.

5.4 THE PROCEDURE FOR F-TEST FOR EQUALITY OF POPULATION VARIANCE

Let X_1, X_2, …….. X_n is a random sample of size n1 from a normal population with mean μ_1 and variance σ_1^2. Similarly, Y_1, Y_2…….. Y_n is a random sample of size n_2 from another normal population with mean μ_2 and variance σ_2^2. Here, we want to test the hypothesis about the two population variances so we can take our alternative null and hypotheses as

1. Set up the null hypothesis

H_0: $\sigma_1^2 = \sigma_2^2$

2. Set up alternative hypothesis

In case of one tailed:

H_1: $\sigma_1^2 > \sigma_2^2$ (Right-tailed)

In this case, the rejection (critical) region falls at the right side of the probability curve of the sampling distribution of test statistic F.

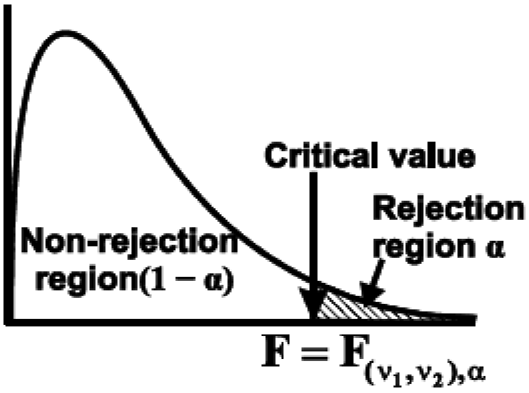

Critical value
Rejection region α
Non-rejection region(1 − α)
$\mathbf{F = F_{(\nu_1,\nu_2),\alpha}}$

Fig. 1: Right-tailed

H_1: $\sigma_1^2 < \sigma_2^2$ (Left-tailed)

In this case, the rejection (critical) region falls at the left side of the probability curve of the sampling distribution of test statistic F.

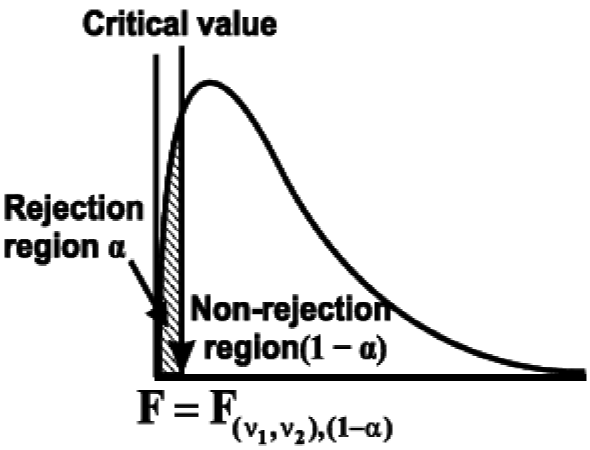


Fig. 2. Left-tailed

In case of two tailed

H1: $\sigma_1^2 \neq \sigma_2^2$

In this case, the rejection (critical) region falls at both sides of the probability curve of the sampling distribution of test statistic F and half the area(α) i.e. α/2 of rejection (critical) region lies at left tail and other half on the right tail.

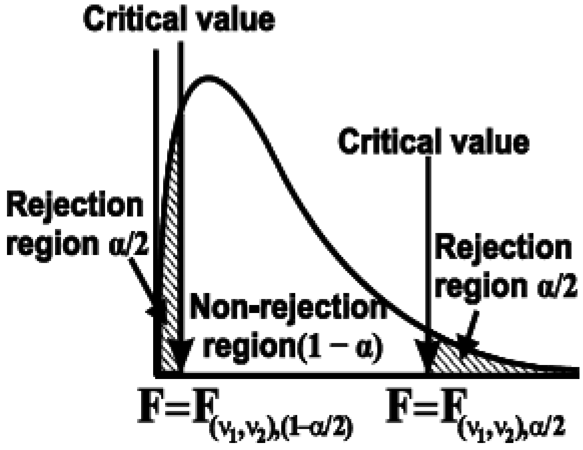

Critical value
Critical value
Rejection region α/2
Rejection region α/2
Non-rejection region(1 − α)
$F=F_{(v_1,v_2),(1-\alpha/2)}$
$F=F_{(v_1,v_2),\alpha/2}$

Fig 3. Two-tailed

3. Compute the test statistic

$$F=\frac{s_1^2}{s_2^2}$$

Where, $s_1^2 = \frac{1}{n_1-1}\sum(X-\overline{X})^2$

and $s_2^2 = \frac{1}{n_2-1}\sum(Y-\overline{Y})^2$

Note: Always take larger variance in the numerator of F. If sample standard deviations are given, then

$$s_1^2 = \frac{n_1\, s_1^2}{n_1 - 1}$$

$$s_2^2 = \frac{n_1\, s_2^2}{n_2 - 1}$$

4. Choose the appropriate level of significance (α) 90 %, 95% and 99%

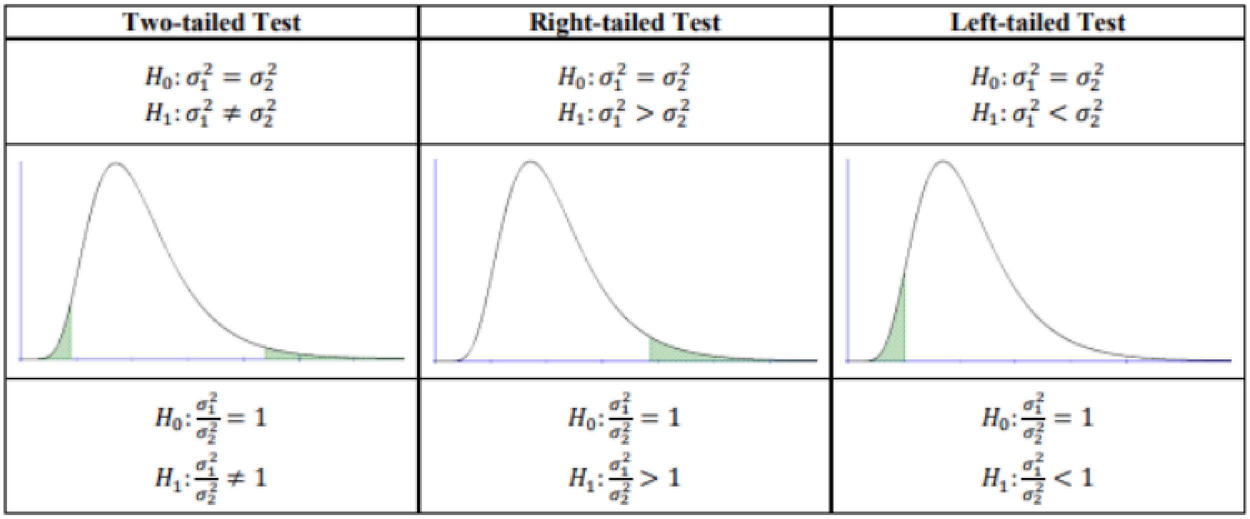
5. Procedure of deciding the null hypothesis based on p-value.

To decide on the null hypothesis based on p-value, the p-value is compared with the level of significance (α). Compare the calculated value of F with the critical value. If F > F_α, then reject H_0 where F_α is the critical value of F at a level of significance.

Note: With the help of computer packages and software such as SPSS, SAS, MINITAB, EXCEL, etc. we can find the exact p-value for F-test.

5.5 HYPOTHESIS TEST FOR TWO VARIANCES

Sometimes we will need to compare the variation or standard deviation between two groups. For example, let’s say that the average delivery time for two locations of the same company is the same but we hear complaint of inconsistent delivery times for one location. We can use an F-test to see if the standard deviations for the two locations was different.



Two-tailed Test	Right-tailed Test	Left-tailed Test
$H_0: \sigma_1^2 = \sigma_2^2$ $H_1: \sigma_1^2 \neq \sigma_2^2$	$H_0: \sigma_1^2 = \sigma_2^2$ $H_1: \sigma_1^2 > \sigma_2^2$	$H_0: \sigma_1^2 = \sigma_2^2$ $H_1: \sigma_1^2 < \sigma_2^2$
$H_0: \frac{\sigma_1^2}{\sigma_2^2} = 1$ $H_1: \frac{\sigma_1^2}{\sigma_2^2} \neq 1$	$H_0: \frac{\sigma_1^2}{\sigma_2^2} = 1$ $H_1: \frac{\sigma_1^2}{\sigma_2^2} > 1$	$H_0: \frac{\sigma_1^2}{\sigma_2^2} = 1$ $H_1: \frac{\sigma_1^2}{\sigma_2^2} < 1$

5.6 F-TEST FOR EQUALITY OF POPULATION VARIANCES FORMULA

Suppose we have two random samples of size n_1 and n_2 from two independent populations. We want to test whether the variances of two populations are equal or not. variances of two s are same i.e.,

H_0: $\sigma_1^2 = \sigma_2^2$

H_1: $\sigma_1^2 \neq \sigma_2^2$

Under H_0 the F-statistic is

$$F= \frac{s_1^2}{s_2^2}$$

When $s_1^2 \ and \ s_2^2$ are unbiased estimates of population variances

$$s_1^2 = \frac{1}{n_1-1} \sum(X - \overline{X})^2$$

$$s_2^2 = \frac{1}{n_2-1} \sum(Y - \overline{Y})^2$$

F follows F-distribution with n_1-1 and n_2-1, d.f It should be noted that the alternative hypothesis in this case is $\sigma_1^2 > \sigma_2^2$ (Right tail). numerical problems we take greater of the variances s_1^2 or s_2^2 in the numerator and adjust the d.f accordingly.

5.7 APPLICATION OF F-DISTRIBUTION

F-distribution has the following applications in statistical theory:

1. F-test for equality of population variances
2. F-test for testing the significance of an observed multiple correlation coefficient
3. F-test for equality of several means.

1. F-test for equality of population variances:

Suppose we are interested to find if two normal populations have same variance. Let $X_1, X_2,...,X_{n1}$ be a random sample of size n_1, from the first normal population with variance $\sigma_1^2$and $Y_1,Y_2,...,Y_{n2}$ be a random sample of size n_2 from the second normal population with variance σ_2^2 Obviously the two samples are independent. Set up the Null Hypothesis H_0: $\sigma_1^2 = \sigma_2^2 = \sigma^2$, population variances are same. In other words, H_0 is that the two independent estimates of the common population variance do not differ significantly.

Therefore, Under H_0 the test statistic is

$$F=\frac{s_1^2}{s_2^2}$$

F (n_1-1, n_2-1)

where s_1^2, and s_2^2 are unbiased estimates of the common population variance σ^2 and are given by:

$$s_1^2 = \frac{1}{n_1-1}\sum(X-\overline{X})^2$$

$$s_2^2 = \frac{1}{n_2-1}\sum(Y-\overline{Y})^2$$

F distribution with v_1= n_1-1, v_2= n_2-1 d.f F (v_1, v_2),

2. F-test for testing the significance of an observed multiple correlation coefficient

In multiple regression, the coefficient of determination, R^2, is the squared correlation between the observed values of the outcome variable y, and its predicted values. To test whether the population coefficient of determination, denoted ρ2, is 0, an F-test is used. Suppose k is the number of predictors in the regression model, and N is the sample size, the F-test is computed as

$$F=\frac{R^2/k}{(1-R^2)/(N-k-1)} \quad \text{………… (1)}$$

which, under the assumption of normality of the errors, has an F-distribution with k numerator degrees of freedom, and N−k−1 denominator degrees of freedom.

When investigators want to test a large model with k_2 predictors against a smaller model with k_1 (k1 < k_2) predictors, an F-test may be used for testing the change in R^2 for significance, denoted ΔR^2. Suppose that R_1^2 is the R^2 of the smaller model and R_2^2 is the R^2 of the larger model. The F-test for testing ΔR^2 for significance is given by

$$F=\frac{(R_2^2-R_1^2\Big)/\Big(k_2-k_1\Big)}{(1-R_2^2\Big)/N-k_2-1}. \quad \text{…………(2)}$$

For an overview of regression and its statistical tests, see Chatterjee and Hadi (Citation1999).

The computation of both (Δ)R^2 and the F-tests may be complicated by missing data. A highly recommended technique to handle missing data is multiple imputation.

The complete multiple imputation process consists of three steps:

- the missing data are estimated several times (M) using a stochastic model that accurately describes the data, creating M plausible complete versions of the incomplete data set,
- each completed data set is analyzed using the same statistical analysis, resulting in M different outcomes of this analysis, and
- the M analyses are combined into one analysis, using specific formulas that take into account the additional uncertainty due to the missing data in the standard errors and statistical tests. Such formulas for obtaining overall statistics from multiply imputed data sets are henceforth denoted combination rules.

3. F-test for equality of several means

Analysis of variance (ANOVA) can determine whether the means of three or more groups are different. ANOVA uses F-tests to statistically test the equality of means.

Example 1: In a sample of 8 observations, the sum of the squared deviations of items from their mean was 94.5. ln another sample of 10 observations, the value was found to be 101.7 Test whether the difference is significant at 5% level. (Given that at 5% level, critical value of F for v_1=7 and v_2=9 degrees of freedom are 3.29 and for v_1=8 and v_2=10 degrees of freedom, its value is 3.07)

Solution: Since we are given the critical values of F-statistic we shall apply F-test for equality of population variances.

Null Hypothesis H_0 = $\sigma_1^2 = \sigma_2^2$, i.e., the sample variances do not differ significantly

Alternative Hypothesis H_1 = $\sigma_1^2 \neq \sigma_2^2$, i.e, the sample variances differ significantly

n_1= 8, n_2= 10

$\sum(x - \overline{x})^2$= 94.5

$\sum(y - \overline{y})^2$ = 101.7

$s_x^2 = \frac{1}{n_1 - 1} \sum(X - \overline{X})^2$

$= \frac{94.5}{7} = 13.5$

$s_y^2 = \frac{1}{n_2 - 1} \sum(Y - \overline{Y})^2$

$= \frac{101.7}{9} = 11.3$

Now, $s_x^2 > s_y^2$

$F = \frac{s_x^2}{s_y^2} = 1.195$

f-distribution with d.f is (8-1, 10-1) i.e., (7,9) d.f tab $F_{0.05}$ = (7,9) = 3.29.

Since the calculated value is less than the table value (Cal F < Tab F), it is not significant. Hence H_0 is accepted and concluded that the difference in sample variability is not significant and may be due to sample fluctuations.

Example 2: In a study of wheat productivity in a sample of common ten subdivisions of equal area of agricultural plots it was seen that $\sum(x - \overline{x})^2$= 0.92 and $\sum(y - \overline{y})^2$= 0.26. Test at 5% significance level whether samples taken from two random populations have the same variance.

Solution: H_0: $\sigma_1^2 = \sigma_2^2$, null hypothesis states that there is no difference between the variance of two populations.

H_1: $\sigma_1^2 \neq \sigma_2^2$, alternative hypothesis states that there is a difference between the variance of two populations.

F-test is calculated as

$F = \frac{s_x^2}{s_y^2}$

$s_x^2 = \frac{1}{n_1 - 1} \sum(X - \overline{X})^2$

$= \frac{0.92}{10-1} = .102$

$s_y^2 = \frac{1}{n_2 - 1} \sum(Y - \overline{Y})^2$

$= \frac{0.26}{10-1} = .028$

$F = \frac{.102}{.028} = 3.64$

Degree of freedom for sample 1 = (n-1) = 9

Degree of freedom for sample 2 = (n-1) = 9

The table value of F for $v_1 = 9$, $v_2 = 9$ at 5% significance level is 3.18.

Since the calculated value is more than the table value (Cal F> Tab F), hence the null hypothesis is rejected. i.e. the samples have been drawn from populations having different variance.

Example 3: Two random sample has been drawn from two normal populations:

Sample A: 75 68 65 70 84 66 55

Sample B: 42 44 56 52 46

Test using variance ratio of 5 % level of significance that whether two populations have same variance

Solution: H_0: $\sigma_1^2 = \sigma_2^2$, null hypothesis states that there is no difference between the variance of sample A and B.

H_1: $\sigma_1^2 \neq \sigma_2^2$, alternative hypothesis states that there is a difference between the variance of sample A and B.

x	(x− $\bar{x}$)	$(x - \overline{x})^2$	(y- $\bar{y}$)	$(y - \overline{y})^2$	
75	6				
68	-1	1	42	-6	36
65	-4	16	44	-4	16
70	1	1	56	8	64
84	15	225	52	-4	16
66	-3	9	46	-2	4
55	14	196			
$\sum x = 483$		$\Sigma(x - \overline{x})^2$= 484	$\sum y = 240$		$\Sigma(y - \overline{y})^2$ = 136

$\bar{x} = \frac{\sum x}{n1} = \frac{483}{7} = 69$

$\bar{y} = \frac{\sum y}{n2} = \frac{240}{5} = 48$

$s_x^2 = \frac{1}{n_1 - 1} \sum(X - \overline{X})^2$

$= \frac{484}{7-1} = 80.67$

$s_y^2 = \frac{1}{n_2 - 1} \sum(Y - \overline{Y})^2$

$= \frac{136}{5-1} = 34$

Now, $s_x^2 > \ s_y^2$

$F = \frac{s_x^2}{s_y^2} = \frac{80.67}{34} = 2.37$

Degree of freedom for sample 1 = $(n_1\text{-}1) = 6$

Degree of freedom for sample 2 = $(n_2\text{-}1) = 4$

The table value of F for $v_1 = 6$, $v_2 = 4$ at 5% significance level is 6.16

Since the calculated value is less than the table value (Cal F < Tab F), hence the null hypothesis is accepted. i.e. null hypothesis accepted and samples have been drawn from populations having same variance.

5.8 SUM UP

As, before applying t-test for difference of two population means, one of the requirements is to check the equality of variances of two populations. This assumption can be checked with the help of F-test for two population variances. For example, an economist may want to test whether the variability in incomes differ in two populations, a quality controller may want to test whether the quality of the product is changing over time, etc.

5.9 QUESTIONS FOR PRACTICE

Q1. Two sources of raw materials are under consideration by a tubes manufacturing company. Both sources seem to have similar characteristics but the company is not sure about their respective uniformity. A sample of 12 lots from source A yields a variance of 125 and a sample of 10 lots from source B yields a variance of 112. Is it likely that the variance of source A significantly differs to the variance of source B at significance level $\alpha = 0.01$?

Ans: F-test= 0.28. (do not reject the null hypothesis and reject the alternative hypothesis i.e. we reject the claim at 5% level of significance.)

Q2. Two random samples drawn from two normal populations gave the following results:

	Size	Mean	Sum of Squares of Deviation from the Mean
Sample A	9	59	26
Sample B	11	60	32

Test whether both samples are from the same normal populations?

Ans: F-test = 0.88 (do not reject the null hypothesis)

Q3. In a sample of 10 observations, the sum of square of observations is 120 and in another sample of 12 observations it is 314. Test the significance defence at 5 % level.

Ans: F= 2.14, not significant at 5%

5.10 SUGGESTED READINGS

- C.R. Kothari (1990) Research Methodology. Vishwa Prakasan. India.
- Gupta SC: Fundamental of statistics, S. Chand & Company. New Delhi
- Gupta, SP: Statistical Methods, S. Chand & Company. New Delhi
- A.M Goon, M.K Gupta and B. Dasgupta, fundamental of statistics Vol-I, World press Calcutta
- Anderson, D.R.; Sweeney, D.J. and Williams, T.A., "Statistics for Business and Economics", 2nd edition (2011), Thompson, New Delhi.

CERTIFICATE/ DIPLOMA IN STATISTICAL ANALYSIS AND RESEARCH METHODOLOGY

SEMESTER II

SARM 5: STATISTICAL INFERENCE

UNIT 5: CHI-SQUARE TEST

STRUCTURE

5.0 LEARNING OBJECTIVES

After reading this unit, learners can able to know about:

- To explain and interpret interaction among attributes

- use the chi-square distribution to see if two classifications of the same data are independent of each other
- use the chi-square statistic in developing and conducting tests of goodness of fit
- analyse the independence of attributes by using the chi-square test.

5.1 INTRODUCTION

The Chi-Square Test is based on chi-square distribution and is a non-parametric test (or distribution-free test) as it does not require any assumptions for population parameters. This test helps to determine the difference between observed and expected data. The main objective of the Chi-Square test is to find out whether a difference between given categorical (qualitative) variables is due to chance or any link between them. As categorical variables can be nominal or ordinal, having few particular values so cannot be expressed with the help of normal distribution. For example: a tea-selling firm wants to find out the relationship between consumer's gender, location, and flavor of tea. Here difference between two categorical variables can be due to chance or because of some specific relationship.

Conditions for the validity of Chi-square test:

1. Sample observations should be independent which means all individual items are included only once in the sample.
2. The cell frequencies should be linear only i.e., $\sum O = \sum E = N$.
3. N, the total frequency should be reasonably large (greater than 50).
4. Theoretical frequency should not be less than 5, if so, then use the pooling technique (adding preceding or succeeding frequency or frequencies in theoretical frequency which is less than 5) and accordingly adjust degrees of freedom also.
5. The data should be given in original units only and not in relative or proportion form.

5.2 APPLICATIONS

Chi-square tests are commonly used to test null hypotheses related to the size of inconsistency between the expected results and actual results by using the degree of freedom. There are the following common Chi-Square tests which we will discuss in detail:

1. χ^2 test of the goodness of fit.
2. χ^2 test for independence of attributes.
3. χ^2 test if the population has a specified value of the variance σ^2.

4. χ^2 test of equality of several population proportions.

5.3 CHI-SQUARE TEST OF GOODNESS OF FIT

This test is first developed by Karl Pearson in 1900 to check the significant differences between experimental values and the theoretical values obtained under some theory or hypothesis. It helps to find out whether the difference between observation and theory is because of fluctuations of sampling or because of the inadequacy of theory to fit the observed data. We can decide whether data values are a good fit for our idea (theory) or whether sample data values represent the entire population.

Steps to compute $\boldsymbol{\chi^2}$ value for concluding:

1. Set the Null Hypothesis as follows:

   There is no significant difference between theory and experiment.

   Or

   There is no significant difference in observed (experimental) and theoretical (or hypothetical) values.
2. Calculate the expected frequencies (E_1, E_2,….E_n) corresponding to given observed frequencies (O_1, O_2,…..O_n).
3. Calculate the difference between each observed and expected frequency and then square them i.e., $(O – E)^2$.
4. Divide each square of difference of observed and expected frequency obtained in step 3 by the corresponding expected frequency i.e., $(O – E)^2 / E$.
5. Add the values calculated in step 4 to get $\boldsymbol{\chi^2 = \sum [}(O – E)^2 / E]$
6. Here null hypothesis follows Chi-Square distribution with v = (n-1) degrees of freedom (d.f)
7. Check critical (tabulated) values from the table of Chi-Square distribution for $\boldsymbol{\chi^2}$ for (n-1) d.f at a certain level of significance.
8. Compare calculated and critical values of $\boldsymbol{\chi^2}$. If calculated value is greater than tabulated value then it is significant and we reject null hypothesis which also means there is a significant difference between the experimental and theoretical values. In other words, differences between observed and expected frequencies cannot be because of fluctuations in sampling.

Example 5.1 The number of accidents per month in town A given as follows:

Month:	1	2	3	4	5	6	7	8	9	10
No. of accidents:	4	9	6	15	10	14	2	20	8	12

Test the null hypothesis that accident conditions were the same during all given months.

Solution: Set Null Hypothesis (H_0): accident conditions were the same during all the given months

Or

There is no significant difference between accident conditions during the given months.

Here total number of accidents during 10 months is 100 (4+9+6+15+10+14+2+20+8+12)

So Expected number of accidents (E) = 100/10 = 10

Table 5.1: Calculation of Chi-Square

Month	Observed no. of accidents (O)	Expected no. of accidents (E)	(O – E)	$(O – E)^2$	$(O – E)^2 / E$
1	4	10	-6	36	3.6
2	9	10	-1	1	0.1
3	6	10	-4	16	1.6
4	15	10	-5	25	2.5
5	10	10	0	0	0
6	14	10	4	16	1.6
7	2	10	-8	64	6.4
8	20	10	10	100	10
9	8	10	-2	4	0.4
10	12	10	2	4	0.4
Total	100	100			26.6

$\boldsymbol{\chi^2} = \sum [(O – E)^2 / E] = 26.6$ (calculated value)

d.f. = n-1= 10 – 1= 9 d.f. at 5% level of significance = 16.919 (tabulated value)

Here calculated value is greater than the tabulated value, it is significant and the null hypothesis is rejected. Hence it is concluded that accident conditions were not the same in the given 10-month period.

Example 5.2 An analysis of the results of 600 students was made and found that 230 had failed, 170 passed with third class, 150 secured second class and 50 got first class. Do these figures support the general examination results which are in the ratio of 4:3:2:1 for the given groups respectively? (the table value at 3 d.f at 55 Level of Significance is 7.81)

Solution: Null Hypothesis: There is no significant difference in general and sample examination results.

Following are the expected frequencies for different groups:

Groups:	Failed	Third class	Second class	First class
Exp. Freq.:	4/10 × 600	3/10 × 600	2/10 × 600	1/10 × 600
	= 240	= 180	= 120	= 60

Computation of Chi-Square

Groups	**Observed Freq.(O)**	**Expected Freq.(E)**	**O-E**	**$(O-E)^2$**	**$(O-E)^2$ / E**
Failed	230	240	-10	100	0.416
Third Class	170	180	-10	100	0.555
Second Class	150	120	30	900	7.5
First Class	50	60	-10	100	1.666
Total	600	600	0	1200	10.137

$\chi^2 = \sum [(O - E)^2 / E] = 10.137$, degrees of freedom = 3 at which tabulated value = 7.81

Here calculated value is greater than the tabulated value, it is significant and the null hypothesis is rejected. Thus, there is a significant difference in general and sample results.

5.4 CHI-SQUARE TEST FOR INDEPENDENCE OF ATTRIBUTES

The dictionary meaning of attributes is quality or characteristic for example health, employment, honesty, beauty, gender etc. Now suppose there are two attributes A and B, divided into m and n classes respectively, such that A_1, A_2, …., A_m classes to the attribute A and B_1, B_2, …., B_n classes for attribute B. This type of classification is also known as manifold classification. The frequency distribution of A and B attributes can be expressed in the following m × n manifold contingency table.

Table 5.2: m × n Manifold Contingency Table

Attributes	**B_1**	**B_2**	**…**	**Bj**	**…**	**B_n**	**Total**
A_1	**$(A_1 B_1)$**	**$(A_1 B_2)$**	**…**	**$(A_1 B_j)$**	**…**	**$(A_1 B_n)$**	**(A_1)**
A_2	**$(A_2 B_1)$**	**$(A_2 B_2)$**	**…**	**$(A_2 B_j)$**	**…**	**$(A_2 B_n)$**	**(A_2)**
…	**…**	**…**	**…**	**…**	**…**	**…**	**…**
Ai	**$(A_i B_1)$**	**$(A_i B_2)$**	**…**	**$(A_i B_j)$**	**…**	**$(A_i B_n)$**	**(A_i)**
…	**…**	**…**	**…**	**…**	**…**	**…**	**…**

A_m	**(A_m B_1)**	**(A_m B_2)**	**…**	**(A_m B_j)**	**…**	**(A_m B_n)**	**(A_m)**
Total	**(B_1)**	**(B_2)**	**…**	**(B_j)**	**…**	**(B_n)**	**N**

In the above table given population consisting of N items is divided into m mutually exclusive and exhaustive classes A_1, A_2, …., A_m of attribute A and n mutually exclusive and exhaustive classes B_1, B_2,….., B_n of attribute B.

5.4.1 2*2 CONTINGENCY TABLE

As we have discussed above manifold contingency classification of attributes, similarly we can have two attributes A and B, divided into 2 classes each, such that A_1 and A_2 classes to the attribute A and B_1 and B_2 classes for attribute B. This type of classification is also known as the 2 × 2 classification. The frequency distribution of A and B attributes can be expressed in the following 2 × 2 contingency table. Here population consisting of N items is divided into 2 mutually exclusive and exhaustive classes A_1 and A_2 of attribute A and 2 mutually exclusive and exhaustive classes B_1, and B_2 of attribute B.

Table 5.3: 2 × 2 Contingency Table

Attributes	**B_1**	**B_2**	**Total**
A_1	**(A_1 B_1)**	**(A_1 B_2)**	**(A_1)**
A_2	**(A_2 B_1)**	**(A_2 B_2)**	**(A_2)**
Total	**(B_1)**	**(B_2)**	**N**

Steps to compute χ^2 value for drawing a conclusion

1. Set Null Hypothesis as follows

   The two attributes are independent

2. Calculate the expected frequencies (E) corresponding to all observed frequencies (O) for example: for (A_1 B_2) = $\frac{(A1)\ (B2)}{N}$.
3. Calculate $(O – E)^2 / E$.
4. Add the values calculated in step 3 to get $\boldsymbol{\chi^2} = \sum [(O – E)^2 / E]$
5. Compare the calculated value χ^2 with its tabulated value at a certain significance level for (no. of rows-1) (no. of columns -1) = (2-1) (2-1) = 1 d.f and conclude.

Example 5.3 From the following table test whether the colour of the sons' eyes is associated with that of the fathers'

Father eye colour	**Son eye colour**		**Total**
	Not Black	**Black**	
Not Black	**230**	**148**	**378**
Black	**151**	**471**	**622**
Total	**381**	**619**	**1000**

Solution: Set null hypothesis, i.e., two attributes are independent.

Or attributes father eye color and the son's eye colour are independent.

Now calculate expected frequencies corresponding to all observed frequencies. The expected frequency for 230 can be written as E (230) = $\frac{378 \times 381}{1000}$ = 144.018

Similarly, E (148) = $\frac{378 \times 619}{1000}$ = 233.982, E (151) = $\frac{622 \times 381}{1000}$ = 236.982,

E (471) = $\frac{622 \times 619}{1000} = 385.018$

Table 5.4: Calculation Of χ^2

O	**E**	**(O-E)**	**(O-E)^2**	**(O-E)^2/E**
230	144.081	85.919	7382.0745	51.2355
148	233.982	-85.982	7392.9043	31.5960
151	236.982	-85.982	7392.9043	31.1960
471	385.081	85.919	7382.0745	19.1701
1000	1000	0		133.1976

χ^2 = ∑ [(O – E)^2 / E] =133.1976 (calculated value)

d.f. = (2-1) (2-1) = 1d.f. at 5% level of significance = 3.841 (tabulated value)

Here calculated value is greater than tabulated value so it is highly significant and we reject the null hypothesis. In other words, the colour of sons' eyes is associated with that of fathers' eyes.

5.4.2 YATES CORRECTION

If in the 2 × 2 table any cell frequency is less than 5 then for application of the Chi-Square pooling technique will not be useful as it will lead to a loss of degree of freedom. After adjustment with the help of the pooling technique to make cell frequency greater than 5, d.f. will be zero only. So here we use Yates Correction for 'continuity', under which 0.5 is added to the cell frequency which is less than 5 and adjusting remaining cell frequencies so that totals remain the same.

Suppose we have following 2 × 2 table

Attributes	B_1	B_2	Total
A_1	2	10	12
A_2	6	6	12
Total	8	16	24

In the above table A_1B_1 cell frequency is less than 5 so as per yates' correction it can be adjusted as follows:

Attributes	B_1	B_2	Total
A_1	2.5	9.5	12
A_2	5.5	6.5	12
Total	8	16	24

Rest of the procedure is same for calculating Chi-Square.

Example 5.4 The following table provides information on marks of English and economics of 30 students

Economics Marks	**English Marks**			
		30-60	60-100	total
	30-60	7	2	09
	60-100	10	11	21
	total	17	13	30

Use the Chi-Square test at 5% level of significance to know whether the marks in two subjects are related.

Solution: Null Hypothesis: Marks in English and economics are independent.

In the above table, one of the frequencies is less than 5 (i.e., 3) so we apply Yates correction to adjust frequencies. Following is the adjusted frequencies table:

Economics Marks	English Marks			
		30-60	60-100	total
	30-60	6.5	2.5	09
	60-100	10.5	10.5	21
	total	17	13	30

Now calculate expected frequencies as follows:

E (6.5) = $\frac{17 \times 09}{30}$ = 5.1, E (2.5) =$\frac{13 \times 09}{30}$ = 3.9, E (10.5) = $\frac{17 \times 21}{30}$ = 11.9, E (10.5) = $\frac{13 \times 21}{30}$ = 9.1

Computation of Chi-Square

O	E	O-E	$(O-E)^2$	$(O-E)^2$ /E
6.5	5.1	1.4	1.96	0.384
2.5	3.9	-1.4	1.96	0.502
10.5	11.9	-1.4	1.96	0.164
10.5	9.1	1.4	1.96	0.215
TOTAL =30	30	0		1.265

$\boldsymbol{\chi^2}$ = ∑ [$(O – E)^2$ / E] = 1.265, tabulated value at (2-1) (2-1) = 1d.f is 3.841.

The calculated value is less than tabulated value so null hypothesis is not rejected. Thus, Marks in English and Economics are independent.

5.5 CHI-SQUARE TEST IF THE POPULATION HAS A SPECIFIED VALUE OF THE VARIANCE σ^2

Under this we can test if the given population has a specified variance ($\sigma^2 = \sigma_0^2$). Now suppose we have a random sample (x_1, x_2, x_3,…..x_n) of size n from the given population then to test about the specified value of population variance following steps are given

1. Set null hypothesis

   H_0: $\sigma^2 = \sigma_0^2$

2. calculated value of $\boldsymbol{\chi^2}$ = $\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{\sigma_0^2}$ = $\frac{ns^2}{\sigma_0^2}$ which follows Chi-Square distribution with (n-1) d.f where $s^2 = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x_i} - \bar{\boldsymbol{x}})^2$**,** is a sample variance.
3. Find out the tabulated value at (n-1) d.f at a certain level of significance.
4. Compare calculated and tabulated values to draw a conclusion.

Example 5.5 A sample of 15 values shows the standard deviation to be 6.4. Does this agree with the hypothesis that the population standard deviation is 5, the population being normal?

Solution: set null hypothesis (H_0): population standard deviation is 5.

Here $\chi^2 = \frac{ns^2}{\sigma^2} = \frac{15 \times 40.96}{5 \times 5}$ = 24.576 (calculated value)

At (15-1) d.f for 5% level of significance, the tabulated value of Chi-Square is 23.685

The calculated value is greater than the tabulated value so null hypothesis is rejected.

In other words, population s.d. is not 5.

Example 5.6 A random sample of size 10 from a normal population gave the following values:

65, 72, 68, 74, 77, 61, 63, 69, 73, 71

Test the hypothesis that population variance is 32.

Solution: Null Hypothesis (H_0): $\sigma^2 = 32$

Sample Mean = 693/10 = 69.3

X	65	72	68	74	77	61	63	69	73	71	$\sum$**x = 693**
x- $\bar{x}$ **=** **x – 69.3**	-4.3	2.7	-1.3	4.7	7.7	-8.3	-6.3	-0.3	3.7	1.7	
$(\mathbf{x_i} - \bar{\mathbf{x}})^2$	18.49	7.29	1.69	22.09	59.29	68.89	39.69	.09	13.69	2.89	$\sum(\mathbf{x_i} - \bar{\mathbf{x}})^2$ **= 234.1**

$\chi^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{\sigma^2}$ = 234.1/32 = 7.31 (calculated value)

At (10-1) = 9 d.f, for 5% level of significance Chi-Square value = 16.92 (tabulated value)

Here calculated value is less than the tabulated value so null hypothesis is not rejected so population variance is 32.

If the sample size is large then use Fisher's Normal approximation to Chi-Square distribution as follows:

$z = \sqrt{2\chi^2} - \sqrt{2n-1}$, which follows normal distribution with zero mean and one variance.

5.6 CHI-SQUARE TEST OF EQUALITY OF SEVERAL POPULATION PROPORTIONS

The Z test of equality of two population proportions for large samples is extended to the Chi-Square test for several population means. Here we take independent random samples of pre-determined size from several populations and write down the population proportions (i.e., $P_1 = P_2 = P_3 = P_4 = \ldots = P_n$) into two categories (for example: married and unmarried). The null hypothesis is $P_1 = P_2 = P_3 = P_4 = \ldots = P_n$, which is the Chi-Square statistic for testing the independence of two variables for the 2 × n contingency table.

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

d.f = (r-1) (n-1) = (2-1) (n-1), where n is no. of columns

The expected frequency will be calculated in the same way as we have calculated in earlier tests i.e.,

$$\frac{\text{Row Total} \times \text{Coloumn Total}}{\text{Grand total}}$$

Example 5.7 In a survey it is found that 77 of 220 housewives in city A, 260 of 650 housewives

in City B, 72 of 225 housewives in City C, and 120 of 315 house wives in City D watch a daytime popular T.V serial. At 5% level of significance, test if there is no difference between the true proportions of housewives who watch the TV serial in these cities.

Solution: Firstly, write down the information in the table as follows

Serial Watch/Cities	**A**	**B**	**C**	**D**	**Total**
House wives watching	77	260	72	120	529
House wives not watching	143	390	153	195	881
Total	220	650	225	315	1410

Null Hypothesis (H_0): all population proportions are equal or $P_1 = P_2 = P_3 = P_4$

Alternative Hypothesis (H_1): all population proportions are not equal or P_1, P_2, P_3, P_4 are not equal.

Now calculate expected frequencies for all observed frequencies

E (77) = $\frac{529 \times 220}{1410}$ = 82.539, E (260) = $\frac{529 \times 650}{1410}$ = 243.865, E (72) = $\frac{529 \times 225}{1410}$ = 84.414

E (120) = $\frac{529 \times 315}{1410}$ = 118.180, E (143) = $\frac{881 \times 220}{1410}$ = 137.460, E (390) = $\frac{881 \times 650}{1410}$ = 406.134,

E (153) = $\frac{881 \times 225}{1410}$ = 140.585 E (195) = $\frac{881 \times 315}{1410}$ = 196.819

Calculation of Chi-Square

O	E	O - E	$(O – E)^2$	$(O – E)^2/E$
77	82.539	-5.539	30.680	0.371
260	243.865	16.135	260.338	1.067
72	84.414	-12.414	154.107	1.825
120	118.180	1.82	3.312	0.028
143	137.460	5.54	30.691	0.223
390	406.134	-16.134	260.305	0.640
153	140.585	12.415	154.132	1.096
195	196.819	-1.819	3.308	0.016
1410 (total)	1410 (total)			5.266 (total)

$\chi^2 = \sum [(O – E)^2 / E]$ = 5.266 (calculated value)

d.f = (r-1) (n-1) = (2-1) (4-1) = 3 d.f

tabulated value at 3 d.f at 5% level of significance = 7.815

here calculated value is less than the tabulated value so H_0 is not rejected. In other words, all population proportions are equal or there is no difference between the true proportions of housewives who watch the T.V serial in four given cities.

5.7 SUM UP

In this unit, we have learned about non parametric test (where assumptions for population parameters are not required) known as Chi-Square Test which is based on Chi-Square distribution. It helps to determine the difference between observed and expected data of categorical (qualitative) variables. We have discussed four main applications of Chi-Square test. First, Chi-Square test to know the compatibility between theory and experiment. Second, to know the independence of attributes. Third, to know if the population has a specified value of variance. Fourth, to test the equality of several population proportions where in case of large sample size, use of Fisher's Normal approximation to Chi-Square distribution is used. To solve all types of applications, the null hypotheses are set and computation of $\boldsymbol{\chi^2}$ is done as per required formula followed by rejection or acceptance of null hypotheses. Following is the summary of formulas discussed in unit where O and E are observed and expected frequencies respectively.

- **Chi-Square Test of Goodness of Fit**

  $\boldsymbol{\chi^2} = \sum$ [(O – E)^2 / E] with d.f. = n-1

- **Chi-Square test for independence of attributes**

$\boldsymbol{\chi^2} = \sum$ [(O – E)^2 / E]

  d.f. = (no. of rows -1) (no. of columns-1)

- **Chi-Square test if the population has a specified value of the variance** σ^2

$$\boldsymbol{\chi^2} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{\sigma^2} \text{ with (n-1) d.f,}$$

- **Chi-square test of Equality of Several Population Proportions**

  $\boldsymbol{\chi^2} = \sum$ [(O – E)^2 / E, with d.f = (r-1) (n-1) where n is no. of columns

In case of large sample size, use Fisher's Normal approximation to Chi-Square distribution as follows:

$z = \sqrt{2\chi^2} - \sqrt{2n-1}$, which follows a normal distribution with zero mean and one variance.

5.8 KEY WORDS

- **Chi-Square Test:** It is a non-parametric test applied to know the difference between observed and expected frequencies of categorical variables.

- **Cells Pooling:** When a contingency table contains one or more cells with an expected frequency of less than 5, we combine two rows or columns before calculating χ2. We combine these cells to get an expected frequency of 5 or more in each cell.
- **Observed Frequency:** These frequencies are the results that are collected during an experiment.
- **Expected Frequency:** These frequencies are calculated by using probability theory or statistical formula.
- **Contingency Table:** The frequency distribution of two attributes into m × n classes is a manifold contingency table. The frequency distribution of two attributes into 2 × 2 classes is 2 × 2 contingency table.
- **Pooling Technique:** If any theoretical frequency is less than 5, then to apply the Chi-Square test we use the pooling technique which consists of adding the frequencies that are less than 5 with preceding or succeeding frequencies or frequencies so that the resulting sum is greater than 5.
- **Goodness of Fit:** The chi-square test procedure used for the validation of our assumption about the probability distribution is called goodness of fit.
- **Yates correction:** In the case of a 2 × 2 contingency table, the use of the pooling technique results in a loss of degrees of freedom so here use of yates correction is recommended. This consists of adding 0.5 to cell frequency which is less than 0.5 and adjusting the remaining frequencies accordingly, since row and column totals are fixed and then applying Chi-Square test.

5.9 QUESTIONS FOR PRACTICE

A. Short Answer Type Questions

Q1. Define Chi-Square Test?

Q2. What is Categorical variable?

Q3. Differentiate between observed and expected frequencies.

Q4. What are the degrees of freedom for Chi-Square test for goodness of fit and independence of attributes?

Q5. What is Yates Correction?

Q6. What is 2 × 2 contingency table?

Q7. How to calculate expected frequency in case of independence of attributes to apply the Chi-Square test?

Q8. Give the null hypotheses for applying the Chi-Square test.

Q9. What is non parametric test? Is Chi-Square test non parametric?

Q10. What is Fisher's Normal Approximation?

B. Long Answer Type Questions

Q1. What do you mean by the Chi-Square test? Give the conditions and applications of the Chi-Square test.

Q2. Discuss the steps of applying Chi-Square tests of Goodness of Fit and Independence of Attributes.

Q3. State the conditions of validity of the Chi-Square test. Also, discuss the applications with their null hypothesis and degrees of freedom of Chi-Square.

Q4. In a set of random numbers, the digits 0 to 9 were found to have the following frequencies:

Digits:	0	1	2	3	4	5	6	7	8	9
Freq:	43	32	38	27	38	52	36	31	39	24

Test whether they are significantly different from those expected on the hypothesis of uniform distribution. (Calculated Chi-Square = 16.33, tabulated for 9 d.f. at 5% Level of Significance = 16.92)

Q5. Out of a sample of 120 persons in a village, 76 persons were administered a new drug for preventing corona and out of them 24 persons were attacked by corona. Out of those who were not administered the new drug, 12 persons were not affected by corona:

(a) Prepare a 2 × 2 table showing actual and expected frequencies.

(b) Use the Chi-Square test to find out whether the new drug is effective or not.

(Calculated Chi-Square = 18.968, tabulated at 5% Level of Significance for 1 d.f. = 3.84)

Q6. It is found that 35 of 250 housewives in Patiala, 22 out of 220 in Amritsar and 39 out of 300 in Jalandhar watch at least one talk show every day. At the 0.05 level of significance, test that there is no difference between the true proportions of housewives who watch talk shows in these cities. (Calculated Chi-Square = 1.822, tabulated value at 2 d.f = 5.991)

Q7. A random sample of size 20 from a normal population gives the sample standard deviation 6. Test the hypothesis that the population s.d is 9. (Calculated value = 8.89, tabulated value at 5% Level of Significance for 19 d.f. = 30.144)

Q8. 1000 students at the college level were graded according to their I.Q and the economic conditions of their homes. Use Chi-Square test to find out whether there is any association between economic conditions at home and IQ.

Economic Conditions	High IQ	Low IQ	Total
Rich	460	140	600
Poor	240	160	400
Total	700	300	1000

(Calculated Chi-Square = 31.74, tabulated value for 5% Level of Significance for 1 d.f. = 3.84)

Q9. A company surveyed employees to see whether they preferred a large increase in retirement benefits or a smaller salary increase. From a group of 1000 male employees, 850 supported the retirement benefits. Of 500 female employees, 400 supported the retirement benefits. Test the null hypothesis that the proportion of men and women supporting retirement benefits are equal. (calculated value of Chi-Square = 6, Tabulated value at 5 % Level of Significance for 1, 2 and 3 d.f = 3.80, 5.99 and 7.81 respectively)

Q10. Discuss the steps of applying Chi-Square tests if the population has a specified value of the variance σ^2 and Equality of Several Population Proportions.

5.10 MCQs

1. Which of the following tests determines the difference between observed and expected frequency?

a) Z test

b) F test

c) **Chi-Square test**

d) t-test

2. Theoretical frequency should not be less than-------in case of the Chi-Square Test

a) 4

b) **5**

c) 7

d) 6

3. Which of the following Chi-Square tests is applied to know whether data values are a perfect fit to our theory?

a) Independence of attributes

b) **Goodness of fit**

c) Specified value of population variance

d) Equality of population proportions

4. Which Chi-Square test is applied to know whether two characteristics are associated or not

a) **Independence of attributes**

b) Goodness of fit

c) Specified value of population variance

d) Equality of population proportions

5. The Z test of equality of two population proportions for large samples is extended to which of the following Chi-Square test

a) Independence of attributes

b) Fisher's Normal approximation to Chi-Square distributions

c) Specified value of population variance

d) **Equality of population proportions**

6. If in the 2 × 2 table any cell frequency is less than 5 then for application of Chi-Square which technique is useful

a) Pooling Technique

b) Fisher's Normal approximation

c) **Yates Correction**

d) Both a) and b)

7. (Row total × Column total)/ grand total is used to calculate____in case of Chi-Square test

a) Observed Frequency

b) Total frequency

c) **Expected frequency**

d) None of the above

8. The tabulated value of Chi-Square at a 5% level of significance for 2 degrees of freedom

a) 4.99

b) 5.99

c) 4.69

d) 5.69

9. The degree of freedom in case of 2 × 2 contingency table in the case of Chi-Square test is

a) n – 1

b) no. of rows – 1

c) no. of columns – 1

d) (no. of rows – 1) (no. of columns – 1)

10. A Bombay film director claims that his films are liked equally by males and females. An opinion survey of a random sample of 1000 filmgoers revealed the following results:

	Liked	Disliked
Males	402	193
Females	245	160

What would be the expected frequencies for males and females who liked films?

a) 380.965 and 260.035

b) 388.965 and 268.035

c) 384.965 and 262.035

d) 382.965 and 262.035

11. Which of the following is nonparametric test

a) T-test

b) z test

c) F test

d) Chi-Square test

12. In case of large sample size, which of the following is applied to test the population variance

a) Fisher's normal approximation to Chi-Square distribution

b) Chi-Square test for population variance

c) Chi-square test for population proportion

d) Chi-Square test of independence of attributes

13. If in the case of Chi-Square test, the calculated value is greater than the tabulated value the null hypothesis is

a) Accepted

b) Rejected

c) No conclusion is finalized

d) Chi-Square value is recalculated

14. Sum of square of deviations from mean is calculated in case of which of the following Chi-Square test

a) Independence of attributes

b) Fisher's Normal approximation to Chi-Square distributions

c) **Specified value of population variance**

d) Equality of population proportions

15. In the following table one of the cell frequencies is 2, which is less than 5. To apply Chi-Square test which technique is suitable to adjust cell frequency:

	died	**survived**
Vaccinated	**2**	**10**
Not vaccinated	**6**	**6**

a) Pooling Technique

b) Yates Correction

c) Both a) and b)

d) None of the above

16. Test the hypothesis that population s.d is 8 and it is given that for random sample of size 51, the s.d is 10. Here sample size is large so which of the following is suitable:

a) Chi-Square test for several population proportion

b) Chi-Square test for goodness of fit

c) Fisher normal approximation to Chi-Square distribution

d) Chi-Square test for population variance

17. If H_0: $\sigma^2 = \sigma_0^2$, then which of the following Chi-Square test is applied

a) Goodness of fit

b) Independence of attributes

c) Specified value of population variance

d) Equality of population proportions

18. If Null Hypothesis (H_0): all population proportions are equal or P1 = P2 = P3 = P4 then which of the following Chi-Square test is applied

a) Goodness of fit

b) Independence of attributes

c) Specified value of population variance

d) **Equality of population proportions**

19. A die is thrown 120 times and frequencies of various faces are as follows:

Face no.:	1	2	3	4	5	6
Frequency:	10	15	25	25	18	27

Test whether the die was fair. Here the calculated value of Chi-Square is:

a) 10.40

b) 11.40

c) 9.40

d) 12.40

20. From the following data, use the Chi-Square test and conclude whether inoculation is effective in preventing tuberculosis:

	Attacked	Not attacked	Total
Inoculated	31	469	500
Not Inoculated	185	1315	1500
Total	216	1784	2000

What is calculated value of Chi-Square?

a) 13.64

b) 15.64

c) 14.64

d) 13.60

5.11 SUGGESTED READINGS

- Gupta S.C., Fundamental of Statistics, Himalaya Publishing House, New Delhi (7^{Th} Edition)
- Gupta S.P., Statistical Methods, S Chand & Company, New Delhi.
- Kothari, C.R. (1985) Research Methodology Methods and Techniques, Wiley Eastern, New Delhi.

5.12 APPENDIX

The steps to apply the Chi-Square test for goodness of fit in SPSS:

1. Open the data file in SPSS and select the variable you want to test.
2. Go to "Analyze" and select "Descriptive Statistics" and "Frequencies".
3. Select the variable you want to test and click "Statistics".

4. Select "Chi-square" and click "Continue".
5. Click "Charts" and select "Bar charts" and "Expected values". Click "Continue".
6. Click "OK" to run the frequency analysis.
7. Go to "Analyze" in the top menu and select "Nonparametric Tests" and "One-Sample Chi-Square".
8. In the "One-Sample Chi-Square" dialog box, select the variable you want to test and click "Define Range".
9. In the "Define Range" dialog box, select the values you want to include in the analysis and click "Continue".
10. In the "One-Sample Chi-Square" dialog box, select the expected values you generated from the frequency analysis and click "Continue".
11. Click "OK" to run the Chi-Square test for goodness of fit.
12. The results will appear in the output window. Check "Chi-Square Tabulated value and compare it with the calculated value to make a decision.

The steps to apply the Chi-Square test for independence of attributes in SPSS:

1. Open the data file in SPSS and select the variables you want to test.
2. Go to "Analyze" in the top menu and select "Descriptive Statistics" and "Crosstabs".
3. In the "Crosstabs" dialog box, select the variable for the rows and the variable for the columns.
4. Click the "Statistics" button and check the box for "Chi-square".
5. Click the "Cells" button and select the desired output format, such as expected or observed.
6. Click "Continue" to return to the "Crosstabs" dialog box and click "OK" to run the analysis.
7. The results will appear in the output window. Look for the Chi-Square Tabulated value.
8. Look at the "Expected Values" table to see the expected frequencies for each cell.
9. Compare these values to the observed frequencies to see if there are any discrepancies.
10. You can also examine the "Cell Counts" table to see the frequency of each combination

The steps to apply the Chi-Square test for specified value of population variance in SPSS:

1. Open the data file in SPSS and select the variable you want to test.
2. Go to "Analyze" in the top menu and select "Descriptive Statistics" and "Explore".
3. In the "Explore" dialog box, select the variable you want to test and move it to the "Dependent List" box.

4. Click the "Statistics" button and select "Descriptives". Check the box for "Variances" and enter the specified value of the population variance that you want to test against.
5. Click "Continue" to return to the "Explore" dialog box and click "OK" to run the analysis.
6. The results will appear in the output window. Look for the "Descriptives" table, which contains the sample variance and the specified value of the population variance.
7. Check tabulated value to take decision.
8. Go to "Analyze" in the top menu and select "Nonparametric Tests" and "Chi-Square Test".
9. In the "Chi-Square Test" dialog box, select the variable you want to test and move it to the "Test Variable List" box.
10. Enter the hypothesized value of the population variance in the "Value" box and click "OK" to run the analysis.
11. The results will appear in the output window.

The steps to apply the Chi-Square test for several population proportions in SPSS:

1. Open the data file in SPSS and select the variables you want to test.
2. Go to "Analyze" in the top menu and select "Descriptive Statistics" and "Frequencies".
3. In the "Frequencies" dialog box, select the variables you want to test and move them to the "VariablE (s)" box.
4. Click the "Statistics" button and check the box for "Chi-square". Also, select any other statistics that you would like to see, such as percentages.
5. Click "Continue" to return to the "Frequencies" dialog box and click "OK" to run the analysis.
6. The results will appear in the output window. Check tabulated value to make decision.
7. Look at the "Expected" table to see the expected frequencies for each cell. Compare these values to the observed frequencies to see if there are any discrepancies.

CERTIFICATE/ DIPLOMA IN STATISTICAL ANALYSIS AND RESEARCH METHODOLOGY

SEMESTER II

SARM 5: STATISTICAL INFERENCE

UNIT 6: ANALYSIS OF VARIANCE: ONE-WAY AND TWO-WAY

STRUCTURE

6.0 OBJECTIVES

After reading this unit, learners should be able to know:

- the meaning of the analysis of variance technique
- various types of analysis of variance
- describe the one-way analysis of the variance model
- various assumptions involved in one and two-way analysis of variance

- test the hypothesis under one-way and two-way analysis of variance
- meaning of two-way ANOVA Model
- procedure of constructing the ANOVA Table for two-way classification

6.1 INTRODUCTION: MEANING OF ANOVA

It involves the calculation of several measures of variability, all of which come down to one or another version of the measure of variability such as the sum of squared deviations or mean sum of squared deviations. The statistical technique known as "Analysis of Variance", commonly referred to by the acronym ANOVA was developed by Professor R. A. Fisher in the 1920's. Variation is inherent, so analysis of variance means examining the variation present in data or parts of data. In other words, analysis of variance means to find out the cause of variation in the data.

The variation in the data due to assignable causes can be detected, measured and controlled whereas the variation due to chance causes is not in the control of human beings and cannot be traced or found separately. The reason, this analysis is called analysis of variance rather than multi-group mean analysis (or something like that), is because it compares group means by analyzing comparisons of variance estimates. Analysis of variance facilitates the analysis and interpretation of data from field trials and laboratory experiments in agriculture and biological research. Today, it constitutes one of the principal research tools of biological scientists, and its use is spreading rapidly in the social sciences, the physical sciences, engineering, management, etc. The total variation present in the data is divided into two components of variation one is due to assignable causes (between the group's variability) or the other is variation due to chance causes (within-group variability)

Basic Assumptions in Analysis of Variance

- Assumption of Randomness
- Assumption of Additivity
- Equality of Variances or Homoscedasticity and Zero Correlation
- Assumptions of Normality

6.2 USES AND APPLICATIONS OF ANOVA

The following are some of the uses of ANOVA:

1. To Test the Homogeneity of Several Means (k groups) or

H_0: $\mu_1 = \mu_2 = . . . = \mu_k$

If H_0 is rejected then we can say that there is a significant difference between these k groups or there is a significant effect of these k independent variables.

2. To Test the Relationship between Two Variables

This test provides evidence that the dependent variable Y_{ij} and independent variable X_{ij} are related in their movements. If Y_{ij} does not relate with X_{ij} then we expect H_0: $\mu_1 = \mu_2 = \dots = \mu_k$, which is the null hypothesis for testing the absence of a relationship.

3. Test for Linearity of Regression

After the relationship is established, the next step will be to find the appropriate regression function. In the first stage, we try to find out whether the linear regression fits the observed data. So, the null hypothesis is now

H_0: $\mu_i = \alpha + \beta X_i$

with the sample model $Y_{ij} = \mu_i + e_{ij}$, when α and β are the parameters.

4. Test for Polynomial Regression

The test procedure for testing the null hypothesis

H_0: $\mu i = \alpha + \beta 1\ Xi + \beta 2\ Xi\ ^{2}+...+ \beta k\ Xi^{k}$

The relationship between X and Y can be explained by a polynomial of degree k.

5. Some Other Uses of ANOVA:

- To test of homogeneity of a group of regression coefficients.
- To test for equality of regression equations from p groups.
- To test for multiple linear regression model.

APPLICATIONS OF ANOVA

ANOVA was primarily designed to analyse the data relating to Agriculture. However, it has a wide variety of applications these days. Some of the applications of ANOVA are:

(i) The yield of three or more varieties of crops is the same or not.

(ii) The effect of different drugs in controlling pain is the same or not?

(iii) Is there any significant difference between the board results of different schools?

(iv) Is there any significant difference in the yield of crops by applying different fertilizers?

(v) (Are the sales of different salesmen differing significantly or not?

(vi) Does the different methods of advertising have any bearing on the sales of a company

6.3 TYPES OF ANOVA

- One-way ANOVA
- Two-way ANOVA

The analysis of variance is one of the most powerful techniques of statistical analysis. Analysis of variance is used for testing of equality of means of several populations. It tests the variability of the means of the several populations. One-way analysis of variance is a technique where only one independent variable at different levels is considered which affects the response variable.

6.4 ONE-WAY ANALYSIS OF VARIANCE: ASSUMPTIONS

One-factor analysis of variance or one-way analysis of variance is a special case of ANOVA, for one factor of variable of interest and a generalization of the two-sample t-test. The two-sample t-test is used to decide whether two groups (two levels) of a factor have the same mean. One-way analysis of variance generalizes this to k levels (greater than two) of a factor. In the following, subscript i refers to the i_{th} level of the factor and subscript j refers to the j_{th} observation within a level of factor. For example, y_{23} refers to the third observation of the second level of a factor.

One-way Analysis of Variance (ANOVA) makes several key assumptions to ensure the validity and reliability of the statistical analysis.

These assumptions are crucial for obtaining accurate results and meaningful interpretations:

- **Normality**: The data within each group should follow a roughly normal distribution. This assumption is more critical when sample sizes are small, as larger sample sizes tend to be more robust to deviations from normality.
- **Homogeneity of Variance:** The variances within each group should be approximately equal. This assumption, also known as homoscedasticity, ensures that the groups have similar levels of variability. Violations of this assumption can lead to unreliable results and affect the overall interpretation of group differences.
- **Independence:** Observations within each group must be independent of each other. This means that the value of one observation should not be influenced by or related to the value of another observation within the same group. Independence is crucial for the accurate estimation of within-group and between-group variances.
- **Random Sampling:** Data should be collected through a random sampling process to ensure that the sample is representative of the population. This assumption facilitates the generalization of the results to the broader population from which the samples were drawn.

- **Interval or Ratio Scale:** The dependent variable (the one being measured) should be measured on an interval or ratio scale. ANOVA is less appropriate for categorical variables or ordinal variables with uneven intervals.

If these assumptions are violated, it may impact the reliability and validity of the ANOVA results. Researchers often use diagnostic tools and statistical tests to assess the fulfillment of these assumptions before interpreting the findings of a one-way ANOVA. Additionally, non-parametric alternatives may be considered if the assumptions cannot be met.

6.5 ONE-WAY ANOVA: COMPUTATION PROCEDURE

One-way Analysis of Variance (ANOVA) is a statistical method used to assess whether there are any significant differences among the means of three or more independent groups. It is an extension of the t-test, which compares means between two groups, and is applicable when dealing with multiple groups or levels of a categorical independent variable. ANOVA examines the variance within each group and compares it to the variance between the groups to determine if the observed differences are likely due to genuine group effects or simply random variability.

The basic idea behind ANOVA is to partition the total variability in the data into two components: variance within each group and variance between the groups. If the between-group variance is significantly greater than the within-group variance, it suggests that there are real differences among the groups. ANOVA provides a more efficient and powerful analysis compared to conducting multiple t-tests, helping to reduce the risk of Type I errors (false positives) associated with multiple comparisons. This method is widely used in experimental and observational studies across various disciplines to assess group differences and make informed conclusions about population means. Therefore, the objective is to determine whether these differences are significant or in other words, are the difference more than what might be expected to occur by chance? If the differences are more than what might be expected to occur by chance, you have sufficient evidence to conclude that there are differences between the population means of different levels of a factor.

A one-way ANOVA uses the following null and alternative hypotheses:

- H_0 (null hypothesis): $\mu_1 = \mu_2 = \mu_3 = \ldots\ldots\ldots = \mu_k$ (all the population means are equal)
- H_1 (alternative hypothesis): $\mu_1 \neq \mu_2 \neq \mu_3 \ldots\ldots\ldots \neq \mu_k$ (at least one population mean is different from the rest)

Example of One-way ANOVA

Consider an example where a researcher wants to examine whether there are any significant differences in the average test scores among students taught by three different teaching methods: Method A, Method B, and Method C. The test scores (a continuous dependent variable) are collected from independent samples of students exposed to each teaching method.

PROCEDURE: ONE-WAY ANOVA

Set up a Hypotheses:

- Null Hypothesis (H_0): There is no significant difference in the average test scores among the three teaching methods.
- Alternative Hypothesis (H_1): There is a significant difference in the average test scores among the three teaching methods.
- Compute the total of each of K treatments as T_1, T_2………. T_k
- Compute the Grand Total as

  $G= T_1+T_2+.......+ T_K$

  $N=n_1+n_2+...........+ n_K$
- Compute the Total Sum of Squares as:

  $TSS =\sum X_{ij}^2 - \frac{G^2}{N}$

  Where $\frac{G^2}{N}$ = Correction factor
- Compute the sum of squares between samples.

  $SSB = \frac{T_1^2}{n_1} + \frac{T_1^2}{n_2} + \ldots\ldots\ldots\ldots\ldots + \frac{T_k^2}{n_k} - \frac{G^2}{N} = \sum \frac{T_i^2}{n_i} - \frac{G^2}{N}$
- Compute sum of squares due to errors (SSE) by subtracting SSB from SSE

  SSE = TSS-SSB
- Construct ANOVA Table.

ANOVA Table for One-way Classified Data

Source of Variation	Degrees of Freedom (df)	Sum of Squares (SS)	Mean Sum of Squares (MSS)	Variance Ratio
Between	$df_r = k-1$	$SSB= \sum \frac{T_i^2}{n_i} - \frac{G^2}{N}$	MSSB = SSB/(k−1)	F=MSST/MSSE With{(k−1), (N−k)}df

Within (Error)	df_e =N−k	SSE=TSS-SSB	MSSE= SSE/(N−k)	
Total	df_t =N−1	TSS= $\sum X_{ij}^2$ - $\frac{G^2}{N}$		

Here**,** SSB: regression sum of squares

SSE: error sum of squares

TSS: total sum of squares (TSS = SSR + SSE)

df_r: regression degrees of freedom (df_r = k-1)

df_e: error degrees of freedom (df_e = n-k)

df_t: total degrees of freedom (df_t = n-1)

k: total number of groups

n**:** total observations

MSSB**:** regression mean square (MSST = SST/df_r)

MSSE**:** error mean square (MSSE = SSE/df_e)

F: The F test statistic (F = MSR/MSE)

The calculated value of F in the last column is compared to the Tabulated value of F at (k-1, N-k) degrees of freedom at a specified level of significance.

If the calculated value of F is greater than the critical or tabulated value, then H_0 of no significant difference between different treatments is rejected. i.e., we conclude that the different treatments differ significantly.

It should be noted that the ANOVA technique described above enables us to test the equality of several population means. It is not designed to test the equality of several population variances Rather based on decomposing the variations in the experimental data it aims at testing the significance of the difference between means of 3 or more samples.

Since the sum of squares and the mean sum of squares are independent of change of origin, we can make our calculations by subtracting a constant value from the entire set of data, which may be short cut method.

Example 1: Three varieties A, B and C of wheat are shown in five plots each of the following fields per acre obtained as shown in the table

Plots	**A**	**B**	**C**
1	8	7	12

2	10	5	9
3	7	10	13
4	14	9	12
5	11	9	14

Solution: Null hypothesis H_0: $\mu_1 = \mu_2 = \mu_3$ i.e. the mean fields of three variables is the same,

Alternative hypothesis H_1: $\mu_1 \neq \mu_2 \neq \mu_3$

The calculation is done based on the given data and the results are as follows:

G = $\sum\sum y_{ij}$ = Sum of all observations

G = 8+10+7+14+11+7+5+10+9+9+12+9+13+12+14 = 150

N = Total number of observations = 15

Correction factor (CF) = $\frac{G^2}{N} = \frac{150 \times 150}{15}$ =1500

Raw Sum of Squares (RSS) = $\sum\sum y_{ij}^2$

= $8^2 + 10^2 + 7^2 + 14^2 + 11^2 + 7^2 + 5^2 + 10^2 + 9^2 + 9^2 + 12^2 + 9^2 + 13^2 + 12^2 + 14^2 = 1600$

Total Sum of Squares (TSS) = Raw Sum of Squares – CF = 1600-1500 = 100

Sum of Squares due to Treatments (SST) = $\frac{T_A^2}{5} + \frac{T_B^2}{5} + \frac{T_C^2}{5} - \text{CF}$

= $\frac{50^2}{5} + \frac{40^2}{5} + \frac{60^2}{5}$ - CF

= $\frac{1}{5}$[2500 1600 3600] - 1500

1540-1500 = 40

Sum of Squares due to Error (SSE) = TSS-SST

= 100-40 = 60

Mean Sum of Squares due to Treatments (MSST) = $\frac{\text{SST}}{\text{df}}$ = 20

Mean Sum of Squares due to Error (MSSE) = $\frac{\text{MSSE}}{\text{df}} = \frac{60}{12}$ = 5

F = $\frac{\text{MSST}}{MSSE} = \frac{20}{5}$ = 4

ANOVA Table for One-way Classified Data

Source of Variation	Degrees of Freedom (df)	Sum of Squares (SS)	Mean Sum of Squares (MSS)	Variance Ratio	

Treatments	df$_r$ =k−1 =2	SST =40	MSST= SST/(k−1)=20	F=MSST/MSSE = $\frac{20}{5}$ = 4	F{(k−1),(N−k)}df F (2,12)$_{.05\%}$ = 3.88
Error	df$_e$ =n−k=12	SSE =60	MSSE= SSE/(N−k)= 5		
Total	df$_t$ =n−1=14	TSS= 100			

For 2, 12, $\nu_1 = \nu_2$ = the table value of F at 5% level of significance is 3.88 which can be seen from the statistical table. Since the calculated value is greater than the table value of F at 5% level of significance. So, we reject the null hypothesis and hence we conclude that the difference between the mean field of three varieties is significant. Since the null hypothesis is rejected, then a pairwise comparison test may be applied to test the null hypothesis of equality of two population means. For this, critical difference (CD) will be calculated by using the formula.

$$CD = \sqrt{\frac{2MSSE}{5}} \times t_{0.05} \text{ at error df}$$

$$CD = \sqrt{\frac{2\times5}{5}} \times 2.571$$

$$= \sqrt{2} \times 2.571 = 1.41\times 2.571 = 3.625$$

$$|\bar{T}1 - \overline{T2}| = |10 - 8| = 2$$

$$|\bar{T}1 - \overline{T3}| = |10 - 12| = 2$$

$$|\bar{T}2 - \overline{T3}| - |8 - 12| = 4$$

Since $|\bar{T}1 - \overline{T2}|$ and $|\bar{T}1 - \overline{T3}|$, are less than CD

So, we accept the null hypothesis which means that if we are interested in taking out of A and B varieties then we can take any of these two. Similarly, between A and C, we can take any of the varieties. But if we conclude to take out of B and C then we should prefer C because the hypothesis of equality two mean is rejected and the mean value corresponding to C varieties is higher than B varieties.

Merits or Advantages of Analysis of Variance

- It is a better method than the "t" or "z" tests since it evaluates variations that are both "between" and "within."
- This method is employed to determine the distinction between many groups or treatments concurrently.
- It is a low-cost gadget that can analyse the primary effects and interaction effects of multiple variables.

- One-way analysis of variance provides the foundation for certain experimental designs, such as levels X treatment designs and simple random designs.
- The F test must be used to assess the difference between two means if "t" is not significant.

Demerits or Limitations of Analysis of Variance

- It is obvious that the application of analysis of variance techniques relies on some presumptions, including the normality and homogeneity of the group variances. The conclusions could suffer if the data differs from these presumptions.
- While the F value can reveal general differences across groups, it is unable to define the conclusion. Consequently, the "t" test is used to describe the statistical inference for a comprehensive analysis of variance.
- The statistical table of the "F" value is necessary for the use of the "F" test; without it, the results cannot be interpreted.

CHECK YOUR PROGRESS (A)

Q1. What is a one-way ANOVA test?

Ans: --

Q2. What are the assumptions of one-way ANOVA?

Ans: --

Q3: Mention the procedure to find out one-way ANOVA?

Ans: --

6.6 MEANING OF TWO-WAY ANOVA

Two-way Analysis of Variance (ANOVA) is a statistical technique used to investigate the simultaneous influence of two categorical independent variables on a continuous dependent variable. This method extends the principles of one-way ANOVA to situations where there are two factors or sources of variation that may affect the outcome. The primary goal is to determine whether there are significant main effects of each independent variable and if there is an interaction effect between them.

The main effects represent the individual impact of each independent variable on the dependent variable, ignoring the presence of the other variable. Meanwhile, the interaction effect explores whether the combined influence of both variables is greater or less than the sum of their individual effects. Two-way ANOVA is valuable in understanding complex relationships and interactions within experimental or observational designs, where multiple factors may contribute to the observed variation.

Researchers typically formulate hypotheses regarding the main effects and interaction, and statistical tests are employed to assess the significance of these effects. If the interaction effect is significant, it suggests that the influence of one variable depends on the level of the other, providing deeper insights into the factors affecting the dependent variable in a multifaceted manner. Two-way ANOVA is when a researcher like to know how two factors affect a response variable and whether or not there is an interaction effect between the two factors on the response variable. For example, suppose a botanist wants to explore how sunlight exposure and watering frequency affect plant growth. She plants 40 seeds and lets them grow for two months under different conditions for sunlight exposure and watering frequency. After two months, she records the height of each plant.

In this case, we have the following variables:

- Response variable: plant growth
- Factors: sunlight exposure, watering frequency

And we would like to answer the following questions:

- Does sunlight exposure affect plant growth?
- Does watering frequency affect plant growth?
- Is there an interaction effect between sunlight exposure and watering frequency? (e.g. the effect that sunlight exposure has on the plants is dependent on watering frequency).

We would use a two-way ANOVA for this analysis because we have **two** factors. If instead, we wanted to know how only watering frequency affected plant growth, we would use a one-way ANOVA since we would only be working with one factor.

6.7 TWO-WAY ANOVA ASSUMPTIONS

For the results of a two-way ANOVA to be valid, the following assumptions should be met:

1. **Normality** – The response variable is approximately normally distributed for each group.
2. **Homoscedasticity:** In a two-way ANOVA test, the variance should be homogenous. The variation around the mean for each set of data should not vary significantly for all the groups.

3. **Independence of variables:** The two variables for testing should be independent of each other. One should not affect the other, or else it could result in skewness. This means that one cannot use the two-way ANOVA test in settings with categorical variables.

6.8 TWO WAY ANOVA: COMPUTATION PROCEDURE

For analysing the two-way classified data, with one observation per cell, one has to follow the following procedure:

In two-way classified data, we can test two hypotheses, one for levels of factor A that is equality of different levels of factor A

- H_{0A}: $\alpha_1 = \alpha_2 = \dots = \alpha_p = 0$

  H_{1A}: $\alpha_1 \neq \alpha_2 \neq \dots \neq \alpha p \neq 0$

  and for equality of different levels of factor B

  H_{0B}: $\beta_1 = \beta_2 = \dots = \beta_q = 0$

  H_{1B}: $\beta_1 \neq \beta_2 \neq \dots \neq \beta_q \neq 0$

- Calculate G = Grand Total = Total of all observations (Row and Column wise)

  $G= R_1+ R_2 +\dots..+ R_k$

  $G= C_1+ C_2+\dots\dots+ C_h$

  and $N = k.h$

- Find Correction Factor (CF) or TSS=

  $TSS= \sum X_{ij}^2 - \frac{G^2}{N}$

- Calculate Row Sum Squares (RSS) = $\frac{R_1^2+R_2^2+ \dots..+ R_K^2}{h}$ - $\frac{G^2}{N}$

  $SSR = \frac{\sum R_i^2}{h}$ - $\frac{G^2}{N}$

- Calculate Row column Squares (SSC) =$\frac{C_1^2+C_2^2+ \dots..+ C_K^2}{K}$ - $\frac{G^2}{N}$

  $SSC = \frac{\sum c_J^2}{h}$ - $\frac{G^2}{N}$

- Compute Sum of Squares due to Error (SSE) =TSS – SSR- SSC
- Compute, MSSR = SSR/df,

  MSSC = SSC/df

  MSSE = SSE/df

- Find F_R = MSSA/MSSE and F_C = MSSB/MSSE

At the end, Compare the calculated value of F_A to tabulated value of F_A; if calculated value is greater than the tabulated value then rejects the hypothesis H_{0A}, otherwise, it may be accepted. Compare the calculated value of F_B to tabulated value of F_B; if calculated value is greater than the tabulated value then rejects the hypothesis H_{0B}, otherwise, it may be accepted.

ANOVA Table for Two-way Classified Data

Source of Variation	Degrees of Freedom (df)	Sum of Squares (SS)	Mean Sum of Squares (MSS)	F- Ratio
Between Levels of A (ROW)	k−1	SSR= $\frac{\sum R_i^2}{h}$ - $\frac{G^2}{N}$	MSSR = SSR/(k−1)	F_R = MSSR/MSSE
Between Levels of B (COLUMN)	h-1	SSC= $\frac{\sum c_j^2}{h}$ - $\frac{G^2}{N}$	MSSC =SSC/(h-1)	F_C = MSSC/MSSE
Error	(k-1) (h-1)	SSE=TSS–SSR-SSC	MSSE = SSE/(k-1) (h-1)	
Total	N-1	TSS= $\sum X_{ij}^2 - \frac{G^2}{N}$		

Example 2: **An experiment was conducted to determine the effect of different data of planting and different methods of planting on the field of sugar-cane. The data below show the fields of sugar-cane for four different data and the methods of planting:**

Data of Planting

Method of Planting	**Oct.**	**Nov**	**Feb**	**March**
A	7.10	3.69	4.70	1.90
B	10.29	4.79	4.50	2.64
C	8.30	3.58	4.90	1.80

Solution: H_{0A}: There is no difference among the different method of planting.

H_{0A}: $\alpha_1 = \alpha_2 = \alpha_3$

Against, H_{1A}: $\alpha_1 \neq \alpha_2 \neq \alpha_3$

H_{0B}: There is no any difference among the different data of planting.

H_{0B}: $\beta_1 = \beta_2 = \beta_3 = \beta_4$

Against, H_{1B}: $\beta_1 \neq \beta_2 \neq \beta_3 \neq \beta_4$

G = $\sum\sum y_{ij}$ = Grand Total = Total of all observations

= 7.10 + 3.69 + 4.70 + 1.90 + 10.29 + 4.79 + 4.58 + 2.64 + 8.30 + 3.58 + 4.90 + 1.80

= 58.28 N = No. of observations = 12

Correction Factor (CF) =$\frac{G^2}{N}$ =$\frac{58.28 \times 58.28}{12} =$ 283.0465

Raw Sum of Squares (RSS)

=$(7.10)^2 + (3.69)^2 + (4.70)^2 + (1.90)^2 + (10.29)^2 + (4.79)^2 + (4.58)^2 + (2.64)^2 + (8.30)^2 + (3.58)^2 +$ $(4.90)^2 + (1.80)$

= 355.5096

Total Sum of Squares (TSS) = RSS – CF = 355.5096-283.0465= 72.4631

Sum of Squares due to Data of Planting (SSD)

$= \frac{D_1^2}{3} + \frac{D_2^2}{3} + \frac{D_3^2}{3} + \frac{D_4^2}{3}$ - CF

= $\frac{25.69^2}{3} + \frac{12.06^2}{3} + \frac{14.18^2}{3} + \frac{6.35^2}{3} - 283.0465$

= 65.8917

Sum of Squares due to Method of Planting (SSM)

= $\frac{M_1^2}{4} + \frac{M_2^2}{4} + \frac{M_3^2}{4}$ – CF

= $\frac{17.39^2}{4} + \frac{22.31^2}{4} + \frac{15.58^2}{4} - 283.0465$

= 286.3412-283.0465 = 3.2947

Sum of Squares due to Error (SSE) = TSS-SSD-SSM

= 72.4631-3.2947-65.8917

= 3.2767

ANOVA Table

Source of Variation	Degrees of Freedom (df)	Sum of Squares (SS)	Mean Sum of Squares (MSS)	Variance Ratio

Between the Levels of A (planting)	p−1 =2	SSA = 3.2947	MSSA = SSA/(p−1)
			$\frac{\mathbf{3.2947}}{\mathbf{2}}$ **= 1.6473**
				F1= MSSA/MSSE
				F1 = $\frac{\mathbf{1.6473}}{\mathbf{0.5461}}$ **= 3.02**

Between the Levels of B	q-1 =3	SSB= 65.8917	MSSB =SSA/(q-1)
			$\frac{\mathbf{65.8917}}{\mathbf{3}}$ **=21.963**
				F2= MSSB/MSSE
				F2= $\frac{\mathbf{21.963}}{\mathbf{0.5461}}$ **= 40.22**

Error	(p-1) (q-1) = 6	SSE =3.2767	MSSE = SSE/(p-1) (q-1) $\frac{\mathbf{3.2767}}{\mathbf{6}}$ **= 0.5461**	
Total	pq-1 = 11	TSS = 72.4631		

The tabulated value of $F_{2,6}$ at 5% level of significance is 5.14 which is greater than the calculated value of FM (3.02) so H_{0A} is accepted. So, we conclude that there is no significant difference among the different methods of planting. The tabulated value of $F_{3,6}$ at 5% level of significance is 4.76 which is less than calculated value of FD (40.22).

So, we reject the null hypothesis H_{0B}. Hence there is a significant difference among the data of planting. In all, we conclude that the different methods of planting affect the mean field of sugar-cane in the same manner. But the mean field differs with different data of planting.

6.9 SUM UP

Analysis of variance is used to test the significance of the difference between the means of a number of different populations say two or more than two. Analysis of variance deals with variance rather to deal with means and their standard error of the difference exist between the means. The variance is the most important measure of variability of a group. It is simply the square of S.D. of the group i.e. v = σ^2.

The problem of testing the significance of the differences between the number of means results from experiments designed to study the variation in a dependent variable with variation in independent variable. Therefore, analysis of variance is used when difference in the means of two or more groups is found insignificant. While determining the significance of calculated or obtained ratio, we consider two types of degrees of freedom. One greater i.e. degree of freedom between the groups and second smaller i.e. degree of freedom within the groups.

CHECK YOUR PROGRESS (B)

Q1. What is two-way ANOVA test?

Ans: --

Q2. What are the assumptions of two-way ANOVA

Ans: --

Q3: When to use two-way ANOVA test?

Ans: --

6.10 QUESTIONS FOR PRACTICE

Q1. A manufacturing company wishes to test whether its three salesmen X, Y and Z tend to make of the same size or whether they differ in their selling ability as measured by the average sales. Following is the weekly record:

SALES		
X	Y	Z
300	600	700
400	300	300
300	300	400
500	400	600
0		500

Test whether the salesmen differ significantly in their performance, as far as sales concerned.

Ans: F= 1.82, F (2, 11) = 3.98, H_0 not rejected

Q2. Three varieties of wheat shown in 12 plots yield output:

A	B	C
12	10	9
18	16	16
17	22	15
	21	23
	25	14
		18

Is there any significant difference between output of different varieties.

Ans: F= 0.55, F (2,11) =3.98, H_0 not reject

Q 3. Summary of analysis of variance is given below:

Source of variance	Df	SS	MSS	F
Between sets	2	180	90.00	17.11
Within sets	27	142	5.26	
Total	29			

Interpret the result obtained.

Note: Table F values are

$F_{.05}$ for 2 and 27 df = 3.35, H_0 rejected

$F_{.01}$ for 2 and 27 df = 5.49, H_0 rejected

Q4. Perform 2-way ANOVA for the data set

Treatment I						
		A	B	C	D	E
Treatment II	A	30	24	33	36	27
	B	26	29	24	31	35

	C	38	28	35	30	35

Sol: F (Treatment I) Column =0.68, $F_{0.05}(2,8) = 4.46$, H_0 not rejected

F (Treatment II) Row= 1.12, H_0 not rejected

Q5. Data represents the number of units of a product produced by 3 different workers using three different types of machines

	Machines			
Workers		A	B	C
	X	8	32	20
	Y	28	36	38
	Z	6	28	14

Test whether mean productivity is the same for the three different machines

Whether the three workers differ with respect to mean productivity

Sol: F (Between Workers) Rows= 10.38, $F_{0.05}$ (2,4) = 6.94, H_0 rejected

F (Within Machines) Column= 9.38, $F_{0.05}$ (2,4) =6.94, H_0 rejected

6.11 SUGGESTED READINGS

- Aggarwal, Y.P. (1990). Statistical Methods – Concept, Applications, and Computation. New Delhi: Sterling Publishers Pvt. Ltd.
- Walker, H.M. & Lev. J. (1965). Statistical Inference. Calcutta: Oxford & I.B.H. Publishing Co.

CERTIFICATE/ DIPLOMA IN STATISTICAL ANALYSIS AND RESEARCH METHODOLOGY

SEMESTER II

SARM 5: STATISTICAL INFERENCE

UNIT 7: INTERPRETATION OF DATA AND REPORT WRITING

STRUCTURE

7.0 OBJECTIVES

After reading this unit, learners can able to know about:

- Importance and precautions of interpretation
- Research report writing

- Qualities of a good research report
- Significance of report writing
- Types of report
- Referencing styles
- Evaluation of the research report

7.1 INTRODUCTION

Interpretation and reporting are important processes in various fields, including research, data analysis, business, and more. They involve analysing data, information, or results and presenting them in a clear and meaningful way to make informed decisions, convey findings, or communicate a message. In other words, Data Interpretation is the process of making sense of numerical data that has been collected, analyzed, and presented. It is the process of attaching meaning to the data and establishing meaning out of the data. Interpretation demands fair and careful judgments as it reflects the theoretical and analytical ability of the researcher to penetrate the data and identify the variables exhibiting a relationship.

Definition: "Interpretation refers to the process of making sense of numerical data that has been collected, analyzed, and presented".

An example of a social media platform generates vast amounts of data every second, and businesses can use this data to analyse customer behavior, track sentiment, and identify trends. Data interpretation in social media analytics involves analysing data in real-time to identify patterns and trends that can help businesses make informed decisions about marketing strategies and customer engagement.

7.2 BENEFITS/ IMPORTANCE OF INTERPRETATION

- Interpretation plays a crucial role in highlighting the practical applications of research findings and opening them up for utilisation. Utilising the results in practical settings makes the research worthwhile. For the following reasons, interpretation is crucial:
- It is a necessary and significant step at the end of research projects because it allows the researcher to explain the general idea that underlies his findings.

- Research continuity is aided by interpretation. By utilising the same basic principles, the results of this study can be connected to those of previous studies, offering an updated and more insightful understanding of preexisting notions.
- Exploratory interpretation frequently acts as a roadmap for developing hypotheses for framing hypotheses for successive research efforts. It provides a base for carrying forward the research efforts.
- Interpretation gives the researcher the chance to better understand the outcomes of his research endeavour while also enabling him to communicate the findings to others in a more relevant way.
- It is common for interpretation to result in the identification of novel connections and conceptual frameworks. Thus, it creates more opportunities for intellectual pursuits.

7.3 PRECAUTIONS IN INTERPRETATION

While interpreting the results the researcher should be alert and take certain precautions like:

1. Firstly, the researcher should ensure that the data available for interpretation has been collected through reliable sources and is valid. It should be adequate to render itself to meaningful interpretation.
2. The statistical methods used for data analysis must be looked into and the researcher should be satisfied that the appropriate methods have been used. An expert's help should be sought to check the reliability and appropriateness of statistical methods since the output of statistical analysis will serve as an input to interpretation.
3. The sampling and non-sampling errors that have crept into the research process must be listed. Generalizations made on a very small sample size or errors in the editing, coding, or tabulating process can severely damage the authenticity of data. Such errors would lead to errors of interpretation and which if committed would nullify the findings of even the best research.
4. A researcher should have a clear understanding of the problem. An inadequate grasp of the broader perspective of the problem and focusing all attention on the immediate aspect may lead to a narrow and limited interpretation.
5. At the other extreme broad generalisations should also be avoided. The endeavour should be to make a correct interpretation of the observed occurrences within a specified time, place and conditions and at the same time bring out the latent relationships and occurrences in the open.

6. It has been seen that it is easier to interpret results that conform to existing theories. However, when the results are negative or inconclusive then interpretation becomes difficult. In such a situation the entire research process and methodologies should be carefully scrutinized.
7. Interpretation should not be left in the hands of inexperienced people. Objectivity, dexterity, and professionalism are the qualities that a researcher interpreting the results should possess. Further, to the extent possible, interpretation should be done by people who have been associated with the project right from the beginning.
8. Lastly, interpretation should consider dynamism. The data of which interpretation has been done relates to a single point in time or period. in the past and by the time the interpretation is done, things might have changed. Hence interpretation should be done keeping in mind the past, as well as the changed conditions.

7.4 SHAPE OF RESEARCH REPORT

The research to be successful should usually follow a pattern that includes the following items:

- Prelims
- Introduction of the topic
- Literature review
- Research aim(s)
- The methods used
- Research findings/results
- Implications of the results
- End matter (which includes references, index, appendices, etc.)

1. Prelims: A contents page can be included behind the title page, and, at the foot of this, you should acknowledge those who have assisted you – including your teacher and supervisor (even though it's their job to help you). If you are writing for publication, you will usually be expected to lead off with an abstract that provides a very brief summary of your work, and some journals will ask you to indicate four or five keywords intended to be of value in any computerized database system that might accommodate your report.

2. Introduction of the topic: The purpose of the introduction is to introduce the research report to the readers. It should contain a clear statement of the objectives of the research i.e., enough background should be given to make clear to the reader why the problem was considered worth investigating. A summary of other relevant research may also be stated so that the present study can be seen in that context. The hypothesis of the study, if any, and the definitions of the major concepts employed in the study should be explicitly stated in the introduction of the report.

3. Literature review: One of the essential preliminary tasks, when you undertake a research study, is to go through the existing literature to acquaint yourself with the available body of knowledge in your area of interest. Reviewing the literature can be time-consuming, daunting, and frustrating, but it is also rewarding. The literature review is an integral part of the research process and makes a valuable contribution to almost every operational step. It is important even before the first step; that is when you are merely thinking about a research question that you may want to find answers to through your research journey. In the initial stages of your research, it helps to clarify your ideas, establish the theoretical roots of your study and develop your research methodology.

Later in the process, the literature review serves to enhance and consolidate your knowledge base in your subject area and helps you to examine your findings in the context of the existing body of knowledge. Since an important responsibility in research is to compare your findings with those of others, it is here that the literature review plays an extremely important role. During the write-up of your report, it helps you to integrate your findings with the existing knowledge – that is, to either support or contradict earlier research. About your study, the literature review can help in four ways. It can:

- Bring clarity and focus to your research problem
- Improve your research methodology
- Broaden your knowledge base in your research area; and
- Contextualize your findings, that is, integrate your findings with the existing body of knowledge.

Searching for existing literature

To search effectively for the literature in your field of inquiry, you must have at least some ideas of the broad subject area and the problem you wish to investigate, to set parameters for your search. You must also have some idea of the study population.

Some sources to conduct a literature review are: Books, Journals, Conference papers and the internet

Though books are a central part of any bibliography, they have their advantages as well as disadvantages. The main advantage is that the material published in books is usually important and of good quality, and the findings are integrated with other research to form a coherent body of knowledge. The main disadvantage is that the material is not completely up to date, as a year or more may pass between the completion of a work and its publication in the form of a book. One of the best ways to look for books is in library catalogs. Next, you need to go through journals similarly relating to your research. Journals can provide you with the most up-to-date information, even though there is often a gap of 2-3 years between the completion of a research project and its publication in a journal. You should select as many journals as you possibly can, though the number of journals available depends upon the field of study. Another important source for the literature review is the papers presented at professional conferences. These can provide you with the most recent research in your area. Last but not least, the internet has become an important tool for finding published literature. Through an Internet search, you can identify published material in books, journals, and other sources with immense ease and speed.

4. Research aims: The section on research aims should be shooting, precise and accurate. Quantitative research will normally reflect the aims you identified at the outset, emphasizing the research question you set out to answer. In qualitative research, your aims and research questions may have evolved during the project; if so, you should indicate the nature and outcomes of this process.

5. Methods used: This section in the research report should give the readers an idea about the methodology that has been used in conducting the research. The scientific reader would like to know in detail about a such thing: How was the study carried out? What was its basic design?

- If the study was an experimental one, then what were the experimental manipulations?
- If the data were collected using a questionnaire or interviews, then exactly what questions were asked (The questionnaire or interview schedule is usually given in an appendix)?
- If measurements were based on observation, then what instructions were given to the observers?

Regarding the sample used in the study, the reader should be told about the subjects, the number, and the procedure for selection. All the above questions are crucial for estimating the probable limits of the generalizability of the findings. The statistical analysis adopted must also be clearly stated. In addition to all this, the scope of the study should be stated and the boundary lines demarcated. The various limitations, under which the research project was completed, must also be narrated.

6. Research findings/results: The findings constitute the heart of a research report. It must contain a statement of findings and recommendations in non-technical language so that it can be easily understood by all concerned. If the findings happen to be extensive, at this point they should be put in the summarized form. All the findings should be presented in a logical sequence and split into readily identifiable sections.

7. Implications of the results: Towards the end of the main text, the researcher should again put down the results of his research clearly and precisely. S(he) should state the implications that flow from the results of the study, for the general reader is interested in the implications for understanding human behavior. Such implications may have three aspects as stated below:

- A statement of the inferences drawn from the present study which may be expected to apply in similar circumstances.
- The conditions of the present study may limit the extent of legitimate generalizations of the inferences drawn from the study.
- The relevant questions that remain unanswered or new questions raised by the study along with suggestions for the kind of research that provide answers for them.

7.5 REPORT WRITING

Report writing is most crucial as it is through the report that the findings of the study and their implications are communicated to the readers. Even the most brilliant hypothesis, highly well-designed and conducted research study, and the most striking generalizations and findings are of little value unless they are effectively communicated to readers. Most people will not be aware of the amount and quality of work that has gone into the study, while much hard work and care might have been put in at every stage of the research, what all readers see is the report. Therefore, the whole purpose of working so hard is defeated if appropriate justice is not done in writing the report. The very purpose of writing the report is to be able to communicate to the readers the nature of methodology followed in doing research and in deriving the findings at which one has arrived.

Definition: A Research Report is a "Systematic, articulate and orderly presentation of research work in a written form”.

7.6 QUALITIES OF A GOOD RESEARCH REPORT

A research report is documentary evidence of the research effort. It serves as a guide for the authorities to evaluate the quality of the entire research effort and decisions are guided by the results presented in the report. Hence a research report should possess certain basic qualities as discussed below:

- **Communicate with the reader**: A report is of some worth only if its contents are easily understood by the readers. The report should be specific to port the readers in terms of details mentioned, presentation and technical language used. "The readers of your report are busy people, and very few of them can balance a research report, a cup of coffee, and a dictionary at once". Technical terms of unavoidable, then should be explained clearly in the glossary.
- **Completeness**: A report is considered complete if it provides all the relevant information and answers the problem adequately. It is not the length but the content which determines the completeness. Too short a report may omit necessary information whereas the same may be lost in the cluster of a lengthy report. The length of the report is generally determined by the characteristics of the reader.
- **Brief**: To be concise is to express a thought completely and clearly in the fewest words possible. However, shortness should not be achieved at the expense of completeness. The researcher is unable to find the right phrase or word to express his idea, he tends to write around it, restating it several times hoping that repetition will hide the lack of poor expression.
- **Accurate**: It is possible that despite the input being accurate, the output i.e., the report may develop inaccuracies. These inaccuracies generally result from carelessness in handling data, grammatical errors, concept phrasing, etc. Hence a good report should be free from such inaccuracies.
- **Objective**: Objectivity requires courage to present and defend the results as they are, rather than twisting them to suit somebody's preferences.
- **Logical**: A good report should be properly structured and there should be logical. A logically written report means a clear report. The sequence of various sections should be logical and an outline of the major points should be clear. Generally, clarity demands writing and rewriting till the time the ambiguity is removed. Use of headings and subheadings wherever required. Here think what you want to say. Write your sentence. Then strip it of all adverbs and adjectives.
- **Professional appearance**: quality of the paper print and cover should be good. Standardization should be maintained throughout the report. Tables, graphs, pictures should be added wherever required.

Therefore, the quality of the report is dependent on:

- written communication skills,
- clarity of thoughts,
- ability to express thoughts rationally and sequentially,
- knowledge of the subject area.

The use of statistical procedures enhances the quality of work done by the researcher. It also reinforces the validity of one's conclusion and arguments. The use of graphs, tables, and diagrams at appropriate places is likely to make the report more attractive and easier to understand for its readers. The most important point to be kept in mind while writing a research report is its intended readers. A report directed to fellow social scientists will be different in certain respects from a report which is meant for laypersons. Whoever is the audience, two general considerations must weigh heavily in the minds of the researchers engaged in writing their research reports:

- What does the audience (or readers) want or need to know about the study?
- How can this information be best presented?

Writing for research is regulated in that you must exercise great caution in what you write, the words you use, the way you present your thoughts, and the reliability and validity of the data you use to support your conclusions. The greatest difference between research writing and other types of writing is the level of intellectual rigour that is necessary. Writings for research projects ought to be precise, understandable, rational, crisis-free, and ambiguity-free. Avoid making assumptions about what readers already know about the study. The researcher should be able to provide solid, scientific evidence to back up any claims they make in the report, and their arguments should be convincing to the reader. One needs to avoid superficial language in the report. Even the best researchers make a number of drafts before writing up their final report. The success will depend on how intelligently it is planned to meet all the objectives it has to fulfill.

A research report should essentially satisfy the following requirements:

- It should include all material of information relevant to the research.
- It should show the information by a predetermined plan that is based on a classification system and logical analysis of the relevant material.
- It should make this idea so clear that readers may understand it with ease.

- It should be expressed in a clear, uncomplicated manner that eliminates any chance of misunderstanding.
- You should use tables and graphs to supplement it.
- It must include references and appendices.

7.7 SIGNIFICANCE OF REPORT WRITING

- Major component
- Findings are brought into light
- Medium to communicate research work with relevant people.
- Contributes to the body of knowledge
- effective way of conveying the research work
- Reference material
- Aid for decision making

7.8 TYPES OF REPORTS

- Written Report
- Oral Report

7.8.1 THE WRITTEN REPORT

The most popular type of reports is written ones, which can be classified in a variety of ways. For example, a report's length determines whether it qualifies as a "long report" or a "short report," similar to a progress report. Reports can also be classified as analytical and informational based on their functional purpose. The informational report contains only the facts; it does not analyse or draw any conclusions. The examination report examines the facts, but the analytical report goes one step further and provides recommendations and findings in addition to the analysis. The reports can be divided into three categories based on the relationship between the writer and reader: administrative, professional, and independent. A professional report is prepared by someone from outside the organisation to which it is to be submitted, whereas an administrative report is made by a member of the same organisation. A report that is released for the public's benefit is considered independent. On this basis, a report may be a technical report or a popular report, both of which are long reports.

a) **Technical Report:** A technical report is a long report containing full documentation. Since everything is described in detail it serves as a source document for future research.

b) **Management Report:** There are situations when the target audience is more interested in the findings presented in a simple manner rather than the methodology. The approach encourages rapid reading, a quick understanding of the findings, and a proper grasp of the implications of the study. A management report should stress on the following things:

   - Objectives should be simple with clear.
   - Pictures and graphs should be used.
   - Findings should be clear.
   - The paragraphs should be short and direct.
   - Important points should be highlighted.

STEPS IN WRITING THE REPORT

Writing a report is time-consuming work and goes through numerous drafts before the final presentation is ready. The various steps involved are as follows:

(i) Report Format

The report format can be of three types: logical, Psychological, and Chronological

- **Logical pattern**: It follows an inductive approach where associations are developed between one thing and another using analysis.
- **Psychological pattern**: This approach is the opposite of the logical pattern where the most critical information is stated first and thereafter the findings are stated in a manner that justifies the conclusion.
- **Chronological pattern**: The reporting of events is done along the time dimension i.e., the events that happened first are stated first and others are stated in the order of their occurrence.

(ii) Preparing a Report Outline

Once the approach to be used to analyse the data has been finalized, the next stage is of preparing an outline. An outline can be prepared according to two styles.

- **Topic outline**: This uses a few keywords, on the assumption that the writer knows its significance and later while writing the detailed report will recall the entire argument.

- **Sentence outline**: This expresses the essential thoughts in a sentence form and is the preferred method for new inexperienced researchers. Since most of the thoughts are outlined initially and only elaboration and explanation are left for subsequent stages, there is less chance of jumbling and error.

(iii) Preparation of Rough Draft

When the report outline has been decided, it is time to make decisions on the settlement of graphics, tables and charts. The outline prepared in the earlier section is elaborated and the research objective, methodology, findings and recommendations are written out. It is time to write down the account of the entire research process.

(iv) Rewriting and Refining the Report

This stage is a slow and careful step where the writer has to spend more time locating and correcting the errors. Weak points need to be identified in the report format, report outline, grammatical errors in tables and graphs and writing. While rewriting, the writer should ensure comprehensibility, continuity and accuracy. The writer should not submit the report hurriedly but should show patience and attention to every detail.

(v) Preparing Bibliography

A bibliography documents the sources used by the researcher in writing the report in an alphabetic order. It contains a list of all the work that the researcher has consulted in the course of his study. In certain situations, the organisation to which the report has to be submitted also states the format of giving a bibliography. Citation, style and format are unique to the purpose of the report.

(vi) Writing the Final Proof

Setting the earlier draught aside for a day or two is a good idea before you start writing the final proof. The researcher can approach the revised draught with a new and critical perspective because of this gap. A clear, objective statement with a high level of consistency and clarity should be the final proof. The final product should be eye-catching enough to entice readers to take it up and captivating enough to hold their interest as they turn the pages. The final proof or report ought to pique readers' intellectual curiosity and broaden their understanding as well as that of the researcher. As said in the previous section, your report needs to embody every quality that makes an excellent report.

7.8.2 ORAL REPORT

An oral report is the presentation of information through spoken word. He/ She has the advantage of creating an interactive environment where they share the information. However, it has the disadvantage that there is no permanent record of the report and the reader does not have the advantage of controlling the pace of presentation. In the case of a written report, if the reader has any doubt or confusion he can refer back to the document and read it as many times and at the place he desires. Hence it becomes important that the presentation of the oral report is of utmost importance.

Key to Effective Oral Presentation

- A written report may not be necessary for an oral report, but the researcher should nonetheless create a thorough description of the format beforehand. After that, the presentation needs to be practiced in front of the audience multiple times before it is given.
- The use of graphs, tables, and diagrams might help to partially compensate for the oral report's shortcoming—namely, the absence of a hard copy. To display the information, the presenter can use audiovisual devices like screens or overhead projectors.
- an essential aspect of an effective presentation is to maintain eye contact with the audience. It makes the targets feel participative and they are more receptive to the information being presented by the researcher. The presentation can be made more interesting with the help of examples, stories and quotations wherever appropriate.
- Another point to remember is the body language. Each stance and gesture of the body conveys a specific meaning e.g., emphatic gestures are used to emphasize a point, and encouraging gestures can be used to elicit a response from the audience. The volume, pitch, and tone all contribute to the effectiveness of the presentation.
- While delivering the report the two extremes, memorization and reading should be avoided. Memorization creates a speaker-centered approach whereas reading makes the presentation dull and listless. The presentation should involve brief notes and while delivering the way of expressing, it evolves naturally and the presenter conversationally develops phrases. This approach is audience-centered and more effective.

7.9 REFERENCING STYLES - ELEMENTS

- Name of author
- Title of article/research paper
- Name of journal /periodical
- Volume number
- Date/month of issuance
- Year of publication
- Page number
- Name of book (in case of a book)
- Book publisher and place (in case of a book)

Referencing Styles APA (American Psychological Association) Referencing Style

The style to reference is 1. For a Research paper/Article, the Author's surname followed by his (their) initials (Year of publication) Article title—name of Journal, Volume, Page no.

For Example: Tomaszewski, S. & Showerman, S. (2010). IFRS in the United States: Challenges and opportunities, Review of Business,30,59-71

Referencing Styles APA (American Psychological Association)

Referencing Style, the style to reference is For a Book is

Author's surname, Initials (Year of publication) Book title. Place: Publisher Example Kumar, R (2014).

Research Methodology: A Step-by-Step Guide for Beginners (4th ed). Thousand Oaks, California: Sage Publications Referencing Styles Harvard Referencing Style The broad style to reference is the Author's Surname and initials. (Publication Year) 'Article title, Newspaper/Magazine Name, Day Month Published, Page(s). Available at: URL or DOI (Accessed date)

7.10 EVALUATION OF THE RESEARCH REPORT

A research report should be critically evaluated to obtain new insights and knowledge. The researcher could evaluate the report or get it done by somebody else. In certain situations, feedback comes automatically whereas in others the researcher may have to ask for it. Evaluation should be done objectively and preferably by people not associated with the research activities. A report can be evaluated against the following parameters:

(a) **Does it address the problem correctly**: A research report should be checked to see if it provides the right background to the problem and if it has successfully identified and located the problem. A report that does not provide such information is weak.

(b) **Is the research design appropriate**: This question evaluates the report on two criteria; firstly, on a choice of research design and secondly on the description of the research design. The report should describe the method of drawing the sample collecting data and analysing it. However, the description should have been simple and non-technical. If the audience is unable to understand the research design, then the report falls a notch on the scale of a good report.

(c) **Have appropriate numbers and statistics been used**: Almost all reports make use of numbers and statistics to support discussion. These should be carefully examined to see if the appropriate statistics have been used and if they serve the purpose of the study well.

(d) **Are the interpretations and conclusion objective:** The report is evaluated for the objectivity and candidness of the interpretation of results. Any assumptions used while interpreting the results should have been clearly stated and the conclusions should be evaluated in the light of limitations faced by the researcher in his research process. This demands a complete and honest disclosure by the researcher.

(e) **Are the results generalizable:** The research report should be evaluated in terms of the generalizability of results. A good report provides sufficient evidence regarding the reliability, validity and generalizability of the findings. The target population to which the findings apply should be identified and factors that limit the generalizability of results should be stated.

Thus, a good report would satisfy these questions that crop up in the mind of the reader to a great extent. An evaluator might develop a few other criteria of his own, but these questions are a good starting point to start evaluating the report.

7.11 QUESTIONS FOR PRACTICE

A. Short Answer Type Questions

Q1. What does interpretation mean?

Q2. Meaning of report writing in statistics.

Q3. Explain the meaning of a written report.

Q4. What is an oral report?

Q5. Give two precautions of interpretation.

B. Long Answer Type Questions

Q1. Explain the benefits/ importance of interpretation.

Q2. Discuss the precautions in interpretation.

Q3. What are the qualities of a good research report?

Q4. Explain the significance of report writing.

Q5. What are the different types of reports?

Q6. What are the steps in writing the report

Q7. Explain the referencing styles.

Q8. Explain the steps to evaluate the research report

7.12 SUGGESTED READINGS

- Goode, W.J. and Hatt, P.K. (1952). Methods of Social Research. New York: McGraw Hill.
- Kothari, L.R. (1985). Research Methodology, New Delhi: Vishwa Prakashan.
- Anastas, J.W. (1999). Research Design for Social Work and The Human Services (2nd ed.) New York: Columbia University Press
- Burns, R.B. (2000). Introduction to Research Methods. New Delhi: Sage Publications.
- Ruane, J.M. (2005). Essentials of Research Methods: A Guide to Social Science Research. Melbourne: Blackwell Publishing.