ESTIMATION

M.Sc., STATISTICS First Year

Semester – I, Paper-III

Lesson Writers

Dr. R. Vishnu Vardhan
Associate Professor
Department of Statistics
Pondicherry University

Dr. O. V. Rajasekharam
Assistant Professor,
Guest Faculty
Department of Statistics,
Acharya Nagarjuna University

Dr. K. Kalyani
Assistant Professor
Dept. of Mathematics & Statistics,
Vignan's Foundations For
Science, Technology & Research,
Guntur

Dr. U. Ramkiran
Tech. Asst.
Department of Statistics
Acharya Nagarjuna University

Editor:
Prof. G. V. S. R. Anjaneyulu
Professor of Statistics (Retd.)
Acharya Nagarjuna University

Director, I/c
Prof. V.VENKATESWARLU
MA., M.P.S., M.S.W., M.Phil., Ph.D.

CENTRE FOR DISTANCE EDUCATION
ACHARAYANAGARJUNAUNIVERSITY
NAGARJUNANAGAR – 522510
Ph:0863-2346222,2346208,
0863-2346259(Study Material)
Website: www.anucde.info
e-mail:anucdedirector@gmail.com

M.Sc., STATISTICS - Estimation

First Edition 2025

No. of Copies :

Published by:

Prof. V.VENKATESWARLU,

Director, I/C

Centre for Distance Education, Acharya Nagarjuna University

Printed at:

FOREWORD

Since its establishment in 1976, Acharya Nagarjuna University has been forging ahead in the path of progress and dynamism, offering a variety of courses and research contributions. I am extremely happy that by gaining 'A^{+}' grade from the NAAC in the year 2024, Acharya Nagarjuna University is offering educational opportunities at the UG, PG levels apart from research degrees to students from over 221 affiliated colleges spread over the two districts of Guntur and Prakasam.

The University has also started the Centre for Distance Education in 2003-04 with the aim of taking higher education to the doorstep of all the sectors of the society. The centre will be a great help to those who cannot join in colleges, those who cannot afford the exorbitant fees as regular students, and even to housewives desirous of pursuing higher studies. Acharya Nagarjuna University has started offering B.Sc., B.A., B.B.A., and B.Com courses at the Degree level and M.A., M.Com., M.Sc., M.B.A., and L.L.M., courses at the PG level from the academic year 2003-2004 onwards.

To facilitate easier understanding by students studying through the distance mode, these self-instruction materials have been prepared by eminent and experienced teachers. The lessons have been drafted with great care and expertise in the stipulated time by these teachers. Constructive ideas and scholarly suggestions are welcome from students and teachers involved respectively. Such ideas will be incorporated for the greater efficacy of this distance mode of education. For clarification of doubts and feedback, weekly classes and contact classes will be arranged at the UG and PG levels respectively.

It is my aim that students getting higher education through the Centre for Distance Education should improve their qualification, have better employment opportunities and in turn be part of country's progress. It is my fond desire that in the years to come, the Centre for Distance Education will go from strength to strength in the form of new courses and by catering to larger number of people. My congratulations to all the Directors, Academic Coordinators, Editors and Lesson-writers of the Centre who have helped in these endeavors.

Prof. K.GangadharaRao
M.Tech.,Ph.D.,
Vice-Chancellor I/c
Acharya Nagarjuna University

M.Sc. – Statistics Syllabus

SEMESTER-I

103ST24: ESTIMATION

UNIT-I:
Concepts of population, parameter (scalar, vector), parametric space, sample, sample space, statistic, estimator, estimate, sampling distribution, standard error, etc. Problem of point estimation, properties of good estimator, sufficiency- concept with examples, distinction between joint density and likelihood function. Fisher Neyman Factorization theorem. Complete sufficiency-examples, Exponential class, Minimal sufficiency.

UNIT-II:
Unbiasedness-concept, examples, properties, LMVUE, UMVLIE, regularity conditions, Cramer-Rao Inequality and condition(s) for existence equality, examples of construction of UMWE using Cramer-Rao Inequality, Rao-Blackwell Theorem, Lehmann-Scheffe Theorem, Necessary and Sufficient condition for the existence of MVUE. Minimum Mean Square Error (MMSE) Estimation. Consistency-Concept and examples, necessary condition for the existence of consistent estimator, efficiency, asymptotic relative Efficiency (ARE)' CAN, CAUN estimators.

UNIT-III:
Moment method of Estimation, ML method of Estimation, Percentile estimation, Minimum Chi-square and Modified Minimum Chi- square.

UNIT-IV:
Interval Estimation, Confidence level, Construction of Confidence intervals using pivots, shortest expected length, UMA, UMAU Confidence sets. Relationship between confidence estimation and testing of hypothesis. Priori and posteriori distributions, loss function, risk function. Minmax & Bayes Estimator.

UNIT-V:
Censored and truncated distributions: Type I and Type 2 Censoring for normal and exponential distributions and their MLE's. Interval estimation: Confidence lntervals, using pivots; shortest expected length confidence intervals.

BOOKS FOR STUDY:
1) Statistical Inference by H.C,. Saxena & Surendran.
2) An introduction to Probability and Statistics by V.K. Rohatgi and A.K.Md.E. Saleh(2001).
3) Mathematical Statistics- Parimal Mukopadhyay (1996), New Central Book Agency (P) Ltd., Calcutta.

BOOKS FOR REFERENCES:
1) An Outline of Statistical Theory, Vol.[by A.M.Goom, M.K. Gupta and B. Dasgupta (1980), World Press, Calcutta.
2) Linear Statistical lnference and its Application by C.R. RAO (1973), John Wiley.
3) A First Course on Parametric Inference by B.K. kale(l999) Narosa Publishing Co.,
4) Lehman, E. L., and Cassella, G. (1998). Theory of Point Estimation, Second Edition, Springer, NY.

CODE:103ST24

M.Sc DEGREE EXAMINATION

First Semester

Statistics :: Paper III- Estimation

MODEL QUESTION PAPER

Time: Three hours Maximum:70 Marks

Answer ONE question from each unit (5x14=70)

UNIT-I

1. (a) Distinguish between estimator and estimate. Explain (i) Sampling distribution and (ii) Standard error and its utility.

   (b) Explain (i) Sufficiency and minimal sufficiency. Let $X_i(i=1,2,\dots n)$ be a random Sample from an exponential distribution with p.d.f $f_\theta(x)=\exp[-(x-\theta)]$, $\theta<x<\infty$, $-\infty<\theta<\infty$. Obtain sufficient statistics for θ.

(or)

2. (a) State and prove Fisher- Neyman Factorization theorem.

   (b) Suppose that $X_1, X_2, ..., X_n$ be a random sample from the p.d.f. $f(x,\theta)=\frac{\theta}{(1+x)^{1+\theta}}, 0<x<\infty$ and $\theta>0$

   =0, otherwise

   Show that $\prod_{i=1}^{n}(1+X_i)$ is minimal sufficient statistic for θ.

UNIT-II

3. (a) Construct UMVUE using Cramer- Rao Inequality.

   (b) Explain (i) Consistency (ii) relative efficiency and (iii) CAN and CAUN estimators with suitable examples.

(or)

4. (a) State and prove Lehmann- Scheffe theorem.

   (b) State and prove necessary and sufficient conditions for the existence of MVUE.

UNIT-III

5. (a) Let $X_1 \, and \, X_2$ be a sample from a distribution with the pdf $f(x,\theta)=\frac{2}{(\theta^2)}(\theta-x), 0<x<\theta.$

   Find the maximum likelihood estimator (MLE) of θ. Based on single observation, show that $2X$ is the MLE of θ but not unbiased.

   (b) Explain minimum chi-square and modified minimum chi-square methods of Estimation.

(or)

6. (a) Prove that the maximum likelihood estimate of the parameter α of a population having density function: $\frac{2}{\alpha}(\alpha-x), 0<x<\alpha.$

   (b) Describe Percentile Estimation with example.

UNIT-IV

7. (a) Define UMA and UMAU confidence sets. Describe the relationship between Confidence estimation and testing of hypothesis.
   (b) Let $X_1, X_2, \ldots, X_n$ be n independent N (μ, σ^2) variables when μ is unknown but σ^2 is known. Let prior distribution of μ be N (θ, σ^2). Find Bayes estimate of μ.

(or)

8. (a) Define (i) confidence bounds and (ii) UMA confidence bounds. State and prove a Sufficient condition for a pivot to yield a confidence interval for a real valued Parameter.
   (b) Obtain 100 $(1-\alpha)\%$ confidence limits (for large samples) for the parameter λ of the Poisson distribution.

UNIT-V

9. (a) Define type I censoring and discuss the ML estimation of the parameters of censored $N(\mu, \sigma^2)$.
   (b) Explain one parameter exponential distribution.

(or)

10. (a) Describe the Truncated exponential distribution.
    (b) Let $X_1, X_2, \ldots, X_n$ be n independent N (μ, σ^2) variables when μ is unknown but σ^2 is Known. Obtain the shortest expected length confidence interval for θ.

CONTENTS

LESSON-1

THEORY OF POPULATION

OBJECTIVES:

This unit is an attempt to have the basic ideas of Estimation. After going through this Lesson you will be able to:

- Understand the concept of estimation
- Understand point estimation
- Explain the characteristics of a good estimator
- Discuss the techniques of solving practical problems

STRUCTURE:

1.1 INTRODUCTION:

The theory of statistical inference is based on sampling theory for making inferences about a population. The primary aim of sampling is to study the features of a population or to estimate the values of its parameter(s). It may be pointed out that it is possible to get reliable information about a population on the basis of sample information even if nothing is known about the population.

Estimation of population parameters by means of sample statistic is one of the important problems of statistical inference. The Theory of estimation was founded by Prof. R. A. Fisher in his research papers in 1930.This is often unavoidable for economic and business decisions and research studies. Thus, we can define the term estimation as follows. In Statistical Inference Theory of estimation is one of the branches.

"It is a procedure by which sample information is used to estimate the numerical magnitude of one or more parameters of the population. A function of sample values is called on estimator (or statistic) while its numerical value is called an estimate". For example $\bar{x}$ is an estimator of population mean μ. On the other hand, if $\bar{x}$ =50 for a sample, the estimate of population mean is said to be 50.

1.2 CONCEPTS OF POPULATION:

The collection of all the observations under study in any statistical investigation is called population or universe for that specific study. The number of observations included in a population is termed as the size of the population or population size.

1.2.1 Parameter (Scale, Vector):

A statistic can be used to estimate a scale parameter as long as it:

- Is location-invariant,
- Scales linearly with the scale parameter, and
- Converges as the sample size grows.

Various measures of statistical dispersion satisfy these. In order to make the statistic a consistent estimator for the scale parameter, one must in general multiply the statistic by a constant scale factor. This scale factor is defined as the theoretical value of the value obtained by dividing the required scale parameter by the asymptotic value of the statistic. Note that the scale factor depends on the distribution in question.

A 'Parameter Vector' refers to a vector that contains parameters used in optimization models, such as the Yield parameter indexed by different crops like Corn, Wheat, and Sugar Beets as described in the provided example.

1.2.2 Definition of Parametric Space:

The set of all possible values that the parameter θ or parameters $\theta_1, \theta_2, \ldots, \theta_k$ can assume is called the parameter space. It is denoted by Θ and is read as "big theta". For example, if parameter θ represents the average life of electric bulbs manufactured by a company, then parameter space of θ is $\Theta=\{\theta;\ \theta \geq 0\}$, that is, the parameter average life θ can take all possible values greater than or equal to 0, Similarly, in normal distribution (μ, σ^2), the parameter space of parameters μ and σ^2 is $\Theta=\{(\mu, \sigma^2): -\infty < \mu < \infty; 0 < \sigma < \infty\}$.

Simply the space of admissible values of parameters in the problem is called the parametric space.

1. Specification : $x \sim P(\lambda), 0 < \lambda < \infty$

Here, $\lambda\ is\ a\ parameter.$
Parametric space for λ: $(0, \infty)$

2. $x \sim N(\mu, \sigma^2)$
Here, parameters:μ and σ^2(unknown)
Parametric space:μ : $(-\infty < \mu < \infty)$
Parametric space:σ: $(0, \infty)$
Parametric space $(\mu, \sigma^2) = (-\infty, 0)X(0, \infty)$

1.3 SAMPLE AND SAMPLE SPACE:

A sample refers to a smaller, manageable version of a larger group. It is a subset containing the characteristics of a larger population.

A sample space is usually denoted using set notation, and the possible ordered outcomes, or sample points are listed as elements in the set. It is common to refer to a sample space by the labels *S*, Ω, or *U* (for "universal set"). The elements of a sample space may be numbers, words, letters, or symbols. They can also be finite, countable infinite, or uncountable.

1.4 STATISTIC:

A mathematical function of random sample observations is called a statistic.
$(x_1, x_2, \dots \dots, x_n) \rightarrow T(x_1, x_2, \dots \dots, x_n)$; T is a statistic.

Example:

$$(x_1, x_2, \dots \dots, x_n) \rightarrow \bar{x}(x_1, x_2, \dots \dots, x_n) = \frac{(x_1 + x_2 + \cdots + x_n)}{n}$$

$$(x_1, x_2, \dots \dots, x_n) \rightarrow s^2(x_1, x_2, \dots \dots, x_n) = \frac{\sum(x_i - \bar{x})^2}{n}$$

A static $T(x_1, x_2, \dots, x_n)$ is a function of sample observation.

1.5 THEORY OF INFERENCE:

In this topic, we will generalize the results of sample to the population to find how far these generalizations are valid, and also to estimate the population parameters along with degree of confidence.

To answer to these are provided by the statistical inference and it is classified in to the following two.

(i) Theory of estimation
(ii) Testing of Hypothesis

1.5.1 Theory of Estimation:

The theory of estimation was founded by Prof .R.A. Fisher in a series of fundamental papers round about 1930 and is divided into two groups.

(i) Point Estimation
(ii) Interval Estimation

In point estimation, a sample statistic is used to provide an estimate of the population parameter where as in interval estimation, probable range is specified within which the true value of parameter might be expected to lie.

1.5.2 Estimator:

An "estimator" is a statistical function that calculates an estimate of a population parameter based on a sample of data; essentially, it's a rule that uses observed data to approximate an unknown value of interest, like the population mean, using a calculated value from a sample (e.g., the sample mean is a common estimator for the population mean).

The mathematical function of random sample observations is called an estimator. Any statistics proposed to estimate unknown parameters is called an estimator of θ.

$$(x_1, x_2, \ldots\ldots x_n) \rightarrow T(x_1, x_2, \ldots\ldots x_n) \text{ to estimate } \theta$$

$$\downarrow$$

Estimator of θ

Estimating statistic is nothing but an estimator.

Example: $x \sim N(\mu, 1)$, μ is unknown

$\downarrow$

Random sample

$(x_1, x_2, \ldots\ldots, x_n)$

$An\ estimator\ of\ \mu\ is\ T_1 = \bar{x}(sample\ mean\ )$

$T_2 = x^{\sim}(sample\ mean)$ Infinitely estimators of μ can be proposed.

1.5.3 Characteristics of Estimators:

It is to be noted that a large number of estimators can be proposed for an unknown parameter. For example, if we want to estimate the average income of the persons living in a city then the sample mean, sample median, sample mode, etc. can be used to estimate the average income. Now, the question arises, "Are some of possible estimators better, in some sense, than the others?" Generally, an estimator can be called good for two different situations:

(i) When the true value of parameter is being estimated is known− An estimator might be called good if its value close to the true value of the parameter to be estimated. In other words, the estimator whose sampling distribution concentrates as closely as possible near the true value of the parameter may be regarded as the good estimator.

(ii) When the true value of the parameter is unknown− An estimator may be called good if the data give good reason to believe that the estimate will be closed to the true value.

In the whole estimation, we estimate the parameter when the true value of the parameter is unknown. Hence, we must choose estimates not because they are certainly close to the true value, but because there is a good reason to believe that the estimated value will be close to the true value of parameter.

1.5.4 Estimate:

An "estimate" refers to a calculated value derived from a sample of data, which is used to approximate the true value of an unknown population parameter, like the mean or proportion, based on the information available from that sample.

The realized particular value of an estimator obtained on calculating with a given realization is called estimate.

$(x_1, x_2, \ldots\ldots, x_n) \rightarrow T(x_1, x_2, \ldots, x_n)$ (Theory before taking random sample)

↓ ↓ ↓

R.S estimator statistic of random variable

Random sample was taken

$(x_1, x_2, \ldots\ldots, x_n) \rightarrow T(x_1, x_2, \ldots\ldots x_n)$

↓

Actual observations

- ➔ Real number proposed are estimated.
- ➔ Estimate a real of θ value.

Example:

$n = 3$

x_1, x_2, x_3

$$\hat{\hat{H}}(\text{unknown}) = \frac{x_1 + x_2 + x_3}{3} = \bar{x}(\text{estimator})$$

$$x_1, x_2, x_3$$

$$(70,70,67)$$

$$\bar{x} = \frac{70 + 70 + 67}{3} = 69(\text{estimate})$$

1.6 SAMPLING DISTRIBUTION:

If it is possible to obtain the values of a statistic (t) from all the possible samples of a fixed sample size along with the corresponding probabilities, then we can arrange the values of the statistic, which is to be treated as a random variable, in the form of a probability distribution. Such a probability distribution is known as the sampling distribution of the statistic.

Starting with a population of N units, we can draw many a sample of a fixed size n. In case of sampling with replacement, the total number of samples that can be drawn is N^n and consequently we shall get N^n different values of any statistic (t) like mean, median, S.D. etc. computed for N^n samples. Again, when sampling is done without replacement of the sampling units, the total number of samples that can be drawn is $N_{c_n - m}$= (say). We can compute any statistic (t) like mean, median, Standard deviation etc. for these m samples resulting in m values of the statistic. These N^n values of the statistics (t) in case of sampling with replacement and $N_{c_n - m}$ values in case of samplingwithout replacement may be arranged in the form of a probability distribution known as the sampling distribution of the statistic.

The sampling distribution, just like a theoretical probability distribution possesses different properties. One of these is the 'Law of Large Numbers' which asserts that a positive integer n can be determined such that if a random sample of size n or large is drawn from a population having mean m, the probability that the sample mean $\bar{x}$will deviate from μ by less than any arbitrarily small quantity can be made to be as close to 1. This implies that a fairly reliable inference can be made about an infinite population by taking only a finite sample of sufficiently large size.

Since our estimators are statistics (particular functions of random variables), their distribution can be derived from the joint distribution $X_1, X_2, ..., X_n$. It is called the sampling distribution because it is based on the joint distribution of the random sample.

Given a sampling distribution, we can

- Calculate the probability that an estimator will not differ from the parameter θ by more than a specified amount.
- Obtain interval estimates rather than point estimates after we have a sample- an interval estimate is a random interval such that the true parameter lies within this interval with a given probability (say 95%).
- Choose between to estimators- we can, for instance, calculate the mean-squared error of the estimator, $E_\theta\left[\left(\hat{\theta}-\theta\right)^2\right]$ using the distribution of $\hat{\theta}$.

Sampling distributions of estimators depend on sample size, and we want to know exactly how the distribution changes as we change this size so that we can make the right trade-offs between cost and accuracy.

1.7 STANDARD ERROR:

The standard deviation of a statistic is termed as standard error. We know that, the population standard deviation describes the variation among values of members of the population, whereas the standard deviation of sampling distribution measures the variability among the values of the statistic (such as mean values, median values, etc) due to sampling errors. Thus, knowledge of sampling distribution of a statistic enables us to find the probability of sampling error of the given magnitude. Consequently, standard deviation of sampling distribution of a sample statistic measures sampling error and is also known as standard error of the statistic. If t be any statistic calculated for different samples, then the standard error of the statistic t is generally denoted by S.E. (t).

The S.E (t) measures not only the amount of chance error in the sampling process but also the accuracy desired in estimation of population parameters.

Standard error is a mathematical tool used in statistics to measure variability. It enables one to arrive at an estimation of what the standard deviation of a given sample is. It is commonly known by its abbreviated form SE. Standard error is used to estimate the efficiency, accuracy, and consistency of a sample. Some of the common results of standard error of different statistic are given below:

1. $S.E\left(\bar{x}\right)=\frac{\sigma}{\sqrt{n}}$: Sample drawn with replacement

   $S.E\left(\bar{x}\right)=\frac{\sigma}{\sqrt{n}}\left[\frac{N-n}{N-1}\right]$: Sample drawn without replacement

2. $S.E\left(p\right)=\sqrt{\frac{PQ}{n}}$: Sample drawn with replacement

$$S.E(p) = \sqrt{\frac{PQ}{n}}\sqrt{\frac{N-n}{N-1}}$$ Sample drawn without replacement

3. $S.E(s) = \dfrac{\sigma}{\sqrt{2n}}$

4. $S.E(Samplemedian) = \sqrt{\dfrac{\Pi}{2n}}\sigma$ where $\prod = 3.1416$

5. $S.E(r) = \dfrac{1-y^2}{\sqrt{n}}$

6. $S.E(\bar{x}_1 - \bar{x}_2) = \sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}$

7. $S.E(p_1 - p_2) = \sqrt{\dfrac{P_1Q_1}{n_1} + \dfrac{P_2Q_2}{n_2}}$

8. $S.E(S_1 - S_2) = \sqrt{\dfrac{\sigma_1^2}{2n_1} + \dfrac{\sigma_2^2}{2n_2}}$

Example:

A population comprises 3 members 1, 5, 3. Draw all possible samples of size two i) with replacement ii) without replacement Find the sampling distribution of sample mean in both cases.

Solution: i) with replacement:

Since n=2 and N=3, the total number of possible samples of size 2 with replacement = 3^2 =9. These are exhibited along with the corresponding sample mean in the following table:

Sl. No	Sample	Sample Mean ($\bar{x}$)
1	1, 1	1
2	1, 5	3
3	1, 3	2
4	5, 1	3
5	5, 5	5
6	5, 3	4
7	3, 1	2
8	3, 5	4
9	3, 3	3

This sampling distribution of the sample mean is given as follows:

$\bar{x}$	1	2	3	4	5	Total
p	1/9	2/9	3/9	2/9	1/9	1

(ii) Without replacement:

As N = 3 and n = 2, the total number of possible samples without replacement $= N_{c_n} = 3_{c_2} = 3$. Possible samples of size 2 and corresponding sample means are given below:

Serial No.	Sample	Sample Mean ($\bar{x}$)
1	1, 3	2
2	1, 5	3
3	3, 5	4

The sampling distribution of the sample mean is given as follows:

$\bar{x}$	2	3	4	Total
p	1/3	1/3	1/3	1

1.8 POINT ESTIMATION:

There can be more than one estimator of a population parameter. So, it is necessary to determine a good estimator out of a number of available estimators. We know that, a function of random variables$(x_1, x_2, \ldots\ldots x_n)$, is a random variable. Therefore, a good estimator is one whose distribution is more concentrated around the population parameter. Thus, we may define point estimation as follows:

A particular value of a statistic which is used to estimate a given parameter is known as point estimate or estimator of the parameter.

The following are some of the criteria of a good estimator:

1. Unbiasedness
2. Consistency
3. Efficiency
4. Sufficiency

We wish to determine the function of the sample observation,
$\hat{\theta}_1(x_1, x_2, \ldots., x_n)$, $\hat{\theta}_2(x_1, x_2, \ldots\ldots, x_n)$, ... $\hat{\theta}_k(x_1, x_2, \ldots. x_n)$ such that this distribution is concentrated as closely as possible near the true value of the "Parameter".The estimating functions are then prefer to as "Estimators".

1.9 THEORY OF PARAMETRIC POINT ESTIMATION:

We know the form of distribution (Mathematical form of the density function) completely and it involves certain parameters and they are not known. That is specification of the distribution is known.

Example:

1. $X \sqcup$ N(μ,σ^2) → Specification normality but μ & σ^2 are known.
2. X □ Poissionp (λ) → Specification Poisson but λ is not unknown.

1.10 CRITERIA OF POINT ESTIMATION IN A STATISTICS:

Suppose T_1 & T_2 are two computing estimators of a parameter "θ"

Example:
To estimate mean (μ) of normal distribution $(\bar{x}, x^{\sim})$ which is better?
To compareT_1&T_2, we have to propose a measure of closeness of the estimator of the parameter.

"D" is a measure of closness.

$$D_1(T_1,\theta), D_2(T_2,\theta)$$

$$if\, D_1 < D_2, conclude\ T_1 is\ better\ than\ T_2.$$

- ➔ θ is unknown
- ➔ T_1 is a random variable
- ➔ T_2 is random variable

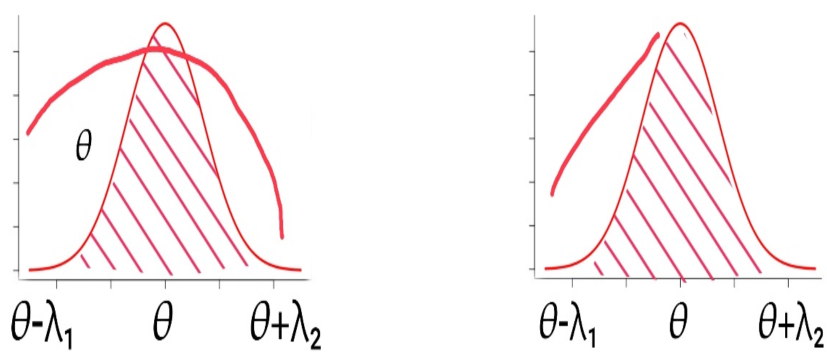
In the case of this, we have given a criterion for comparing T_1 & T_2 a measure of closeness D.

Find 'D'$(T_i,\theta)t/i$,choose a estimator 'T' for estimation for which $D(T,\theta)$ is minimum. "T" is better (or) optimum estimator of θ. A theory of point estimation need selection of a 'D' (A measure of closeness of 'T' to 'θ') and a restriction on 'r' (on type of estimates) to obtain the best estimator for 'θ'

In choice of 'R' and 'D', there may be some controversy.

1.11 SOME MEASURES OF CLOSENESS:

$T_1\ and\ \ T_2\ are\ two\ estimators\ of\ '\theta'$



SD of T_1 SD of T_2

1.12 SUMMARY:

This document explores key concepts in statistical inference, focusing on population and sample-based analysis. It begins with an introduction to population parameters and parametric space, followed by an explanation of sample and sample space. The document then delves into statistical measures such as estimators, estimation theory, sampling distribution, and standard error. Furthermore, it discusses the challenges associated with point estimation and the criteria used for parametric point estimation. Finally, it highlights measures of closeness in statistical estimations.

1.13 SELF-ASSESSMENT QUESTIONS:

1. What is the difference between a population parameter and a statistic?
2. Define parametric space and explain its significance in statistical inference.
3. What are the main characteristics of an estimator?
4. Explain the concept of sampling distribution and its role in statistical analysis.
5. What is the standard error, and why is it important?
6. Describe the challenges associated with point estimation.
7. What are the criteria used to evaluate a point estimator?
8. How do measures of closeness impact statistical inference?
9. Compute the standard deviation of sample mean for the last problem. Obtain the SE of sample mean and show that they are equal.
10. Construct a sampling distribution of the sample mean for the following population when random samples of size 2 are taken from it (a) with replacement and (b) without replacement. Also find the mean and standard error of the distribution in each case.

Population Unit	1	2	3	4
Observation	22	24	26	28

1.14 SUGGESTED READINGS:

1. Casella, G., & Berger, R. L. (2002). *Statistical Inference*. Duxbury.
2. Lehmann, E. L., & Casella, G. (1998). *Theory of Point Estimation*. Springer.
3. Hogg, R. V., McKean, J. W., & Craig, A. T. (2012). *Introduction to Mathematical Statistics*. Pearson.
4. Wasserman, L. (2004). *All of Statistics: A Concise Course in Statistical Inference*. Springer.
5. Mood, A. M., Graybill, F. A., & Boes, D. C. (1974). *Introduction to the Theory of Statistics*. McGraw-Hill.

Dr. K. Kalyani

LESSON- 2

LIKELIHOOD FUNCTION

OBJECTIVES:

- To know about the concept of definition of likelihood function
- To understand about the properties of likelihood function
- To learn about the exponential family of distribution.

STRUCTURE:

2.1 DEFINITION OF LIKELIHOOD FUNCTION:

The joint density function of the random sample $x_1, x_2, \ldots, x_n$ with parameters θ treated as a function of θ for the given random sample observation is called likelihood function of parameter.

$$\pi_{i=1}^{n} f(x_i, \theta)$$

$\phi(x_1, x_2, \ldots, x_n|\theta)$
joint density funciton of
$x_1, x_2, \ldots, x_n$ given by θ

$L(\theta|x_1, x_2, \ldots, x_n)$
joint density funciton of θ
given the random sample observations.

2.1.1 Problems based on Likelihood Function:

Example 1: Likelihood Function for a Poisson distribution

Scenario:
A shop receives an average of λ customers per hour, and we observe customer arrivals for n hours. The number of customers arriving per hour follows a **Poisson distribution** with:

$$P(X = x/\lambda) = \frac{e^{-\lambda}\lambda^{x}}{x!}$$

Given n independent observations$x_1, x_2, \dots, x_n$, the likelihood function is:

$$L = \pi_{i=1}^{n} P(X = x_i)$$
$$= \pi_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}$$
$$= \frac{e^{-\lambda}\lambda^{\Sigma x_i}}{\pi_{i=1}^{n} x_i!}$$

Taking the log-likelihood:

$$\log L(\lambda) = \sum_{i=1}^{n} \left[x_i \log \lambda - \lambda - \log(x_i!) \right]$$

Maximizing this function helps us estimate λ, typically using the Maximum Likelihood Estimation (MLE) approach.

Example 2: Likelihood Function for a Bernoulli distribution

Scenario:

Suppose we are observing a sequence of independent coin flips, where each flip results in either heads (1) or tails (0). The probability of getting heads is p, and the probability of getting tails is 1−p. The coin flips follow a Bernoulli distribution.

Let's say we flip the coin n times and observe $x_1, x_2, \dots, x_n$where each x_i is either 0 or 1. The probability mass function (PMF) of a Bernoulli distribution is:

$$P(X = x/p) = p^x (1-p)^{1-x}$$

Since the flips are independent, the joint probability (likelihood function) for all observations is:

$$L(p/x_1, x_2, \dots, x_n) = \prod_{i=1}^{n} p^{x_i} (1-p)^{1-x_i}$$

If we observe k heads and n−k tails, then the likelihood function simplifies to:

$$L(p) = p^k (1-p)^{n-k}$$

This function describes how likely the observed data is, given a particular value of p.

Example 3: Likelihood Function for a Normal Distribution

Scenario:

Suppose we collect n independent observations $x_1, x_2, \dots, x_n$ from a normal distribution with an unknown mean μ and known variance σ2. The probability density function (PDF) of a normal distribution is:

$$f\left(x/\mu,\sigma^2\right)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{\left(x-\mu\right)^2}{2\sigma^2}\right)$$

The likelihood function is the product of individual probabilities:

$$L\left(\mu/x_1,x_2,...,x_n\right)=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{\left(x_i-\mu\right)^2}{2\sigma^2}\right)$$

Taking the natural logarithm (log-likelihood function):

$$\log L\left(\mu\right)=-\frac{n}{2}\log\left(2\pi\sigma^2\right)-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(x_i-\mu\right)^2$$

To estimate μ, we maximize this likelihood function using calculus techniques such as differentiation.

2.2 ASSUMPTIONS:

- We assume that the range of random variables under consideration is independent of θ.

  Counter example:

  1) $A(x,\theta)$ [rectangular distribution or uniform distribution]density function

  $$f(x,\theta)=\frac{1}{\theta};0\le x\le\theta.$$

  2) Differentiation under integral is permitted with respect to θ one or two types.

2.3 PROPERTIES OF LIKELIHOOD FUNCTION:

1) Let $L(\theta|x_1,x_2,\ldots,x_n)$ be the likelihood function L is treated as function of $x_1,x_2,..,x_n$and is the joint density function of $x_1,x_2,\ldots,x_n$

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}.\ldots..\int_{-\infty}^{\infty}L(dx_1,dx_2,\ldots\ldots,dx_n)=1$$

$$=>\int Ldv=1-------\rightarrow(1)(\text{since}, dv=dx_1,dx_2,\ldots,dx_n)$$

diffrentiating $Eq^n(1)$w.r.t θ

$$=>\frac{d}{d\theta}\int Ldv=0-------\rightarrow(2)$$

assumptions $Eq^{n's}(1)$ and (2),

$$(2)=>\frac{\int dL}{d\theta}dv=0$$

$$\text{i.e,}\int\left(\frac{1}{L}.\frac{dL}{d\theta}\right)Ldv=0$$

$$=\int\left(\frac{dlogL}{d\theta}\right)Ldv=0\qquad\left[\text{since},\frac{1}{L}.\frac{dL}{d\theta}=\frac{dlogL}{d\theta}\right]$$

$$=L\left(\frac{dlogL}{d\theta}\right)=0$$

$2.\, L\left(\frac{d\log L}{d\theta}\right)^2 = -E\left(\frac{d^2\log L}{d\theta^2}\right)$

Proof:

$$\int L dx = 1$$
$$\text{differentiating w.r.t } \theta$$
$$= \frac{d}{d\theta}\int L dv = 0$$
$$= \int \left(\frac{1}{L}.\frac{dL}{d\theta}\right)^2 dv = 0$$
$$= \int \frac{d\log L}{d\theta}. L dv = 0 - - - - - - - - \rightarrow (1)$$
$$\text{differentiating eq}^{n}(1) \text{ w.r.t } \theta \text{ again}$$
$$= \int L\frac{d^2\log L}{d\theta} dv + \int \frac{d\log L}{d\theta}.\frac{dL}{d\theta} dv = 0$$
$$=> E\left(\frac{d^2\log L}{d\theta^2}\right) + \int \frac{d\log L}{d\theta}\left(\frac{1}{L}.\frac{dL}{d\theta}\right) L dv = 0$$
$$=> E\left(\frac{d^2\log L}{d\theta^2}\right) + \int \frac{d\log L}{d\theta}.\frac{d\log L}{d\theta} L dv = 0$$
$$=> E\left(\frac{d^2\log L}{d\theta^2}\right) + \int E\left(\frac{d\log L}{d\theta}\right)^2 = 0$$
$$=> E\left(\frac{d\log L}{d\theta}\right)^2 > -E\left(\frac{d^2\log L}{d\theta^2}\right)$$

$$\left[\text{since}, \int \left(\frac{d\log L}{d\theta}\right)^2 L dx = \int x. f(x) dx\right]$$

NOTE:

$$V\left(\frac{d\log L}{d\theta}\right) = E\left(\frac{d^2\log L}{d\theta^2}\right) - E\left(\frac{d\log L}{d\theta}\right)^2 \left[\text{since}, V(x) = E(x^2) - E(x)\right]$$
$$= E\left(\frac{d^2\log^2}{d\theta^2}\right)\left[\text{since}, E\left(\frac{d\log L}{d\theta}\right) = 0\right]$$

2.4 DISTINCTION BETWEEN JOINT DENSITY AND LIKELY HOO FUNCTION:

In statistics, the **joint density function** and the **likelihood function** are both derived from the joint probability distribution of a set of random variables, but they serve different purposes and are functions of different variables.

Joint Density Function:

The joint density function, $f_{X,Y}(x, y)$, describes the probability distribution of two continuous random variables X and Y. It gives the probability that X falls within a particular range and Y falls within another range simultaneously. This function is defined over the possible values of X and Y and integrates to 1 over the entire space, ensuring it represents a valid probability distribution.

Likelihood Function:

The likelihood function, $L(\theta, y)$, arises when we have observed data y and seek to infer the parameters θ of the statistical model that likely produced this data. While it is mathematically similar to the joint density function, the key distinction is in its role: the likelihood function is viewed as a function of the parameters θ, with the observed data y held constant. It does not integrate to 1 over θ and is not a probability density function itself.

Key Distinctions:

- **Function of Different Variables:** The joint density function is a function of the random variables X and Y, whereas the likelihood function is a function of the parameters θ, given fixed observed data.
- **Purpose:** The joint density function describes the probability distribution of random variables. In contrast, the likelihood function is used to estimate the parameters of a statistical model that are most likely to have produced the observed data.
- **Normalization:** The joint density function integrates to 1 over all possible values of the random variables, ensuring it represents a valid probability distribution. The likelihood function does not have this property with respect to the parameters θ.

Understanding these distinctions is crucial for correctly applying statistical methods, especially in parameter estimation and inferential statistics.

2.4.1 Example:

Suppose we have a dataset consisting of the heights and weights of individuals, and we want to model the relationship between these two continuous variables.

1. **Joint Density Function:**

   The joint density function, denoted as $f_{H,W}(h,w)$, represents the probability density of observing a specific height h and weight w simultaneously. This function characterizes the underlying distribution of the data.

   For instance, if we assume that heights and weights follow a bivariate normal distribution, the joint density function can be expressed as:

$$f_{H,W}(h,w) = \frac{1}{2\pi\sigma_H\sigma_W\sqrt{1-\rho^2}}$$
$$*\left\{\frac{1}{2(1-\rho^2)}\left[\left(\frac{h-\mu_H}{\sigma_H}\right)^2 - 2\rho\left(\frac{h-\mu_H}{\sigma_H}\right)\left(\frac{w-\mu_W}{\sigma_W}\right)+\left(\frac{w-\mu_W}{\sigma_W}\right)^2\right]\right\}$$

   Here, μ_h and μ_w are the mean height and weight, σ_h and σ_w are the standard deviations, and ρ is the correlation coefficient between height and weight.

2. **Likelihood Function:**

   The likelihood function is used when we have observed data and aim to estimate the parameters of our statistical model. It's a function of the parameters, given the observed data.

   Continuing with our example, suppose we have a sample of n individuals with observed heights $h_1, h_2, ..., h_n$ and weights $w_1, w_2, ..., w_n$. Assuming these observations are independent, the likelihood function $L(\mu_H, \mu_w, \sigma_H, \sigma_w, \rho; \{(h_i, w_i)\}_{i=1}^{n})$ is

$$L(\mu_H, \mu_w, \sigma_H, \sigma_w, \rho; \{(h_i, w_i)\}_{i=1}^{n}) = \prod_{i=1}^{n} f_{H,W}(h_i, w_i)$$

   This product represents the joint probability of observing the entire dataset given the parameters. We can obtain the parameter estimates that are most likely to have produced the observed data.

Key Distinctions:

- **Perspective:**
  - *Joint Density Function:* Focuses on the probability distribution of the data itself.
  - *Likelihood Function:* Focuses on how likely specific parameter values are, given the observed data.
- **Function of:**
  - *Joint Density Function:* Variables (e.g., height and weight).
  - *Likelihood Function:* Parameters of the model (e.g., means, standard deviations, correlation).
- **Purpose:**
  - *Joint Density Function:* Describes the distribution and relationship between variables.
  - *Likelihood Function:* Used for parameter estimation to find the model that best explains the observed data.

Understanding these distinctions is crucial for tasks such as statistical modeling and inference, where accurately interpreting data and estimating model parameters are essential.

2.5 EXPONENTIAL FAMILY OF DISTRIBUTION:

Let x be random variable with probability density function $f(x, \theta)$. The family of distribution $[f(x, \theta); \theta \in r]$statisfies the following regularity conditions.

1. The sample space $S = \{x; f(x, \theta) > \theta\}$ of random variable 'x' does not depend on θ.
2. The parameter space Ω is an open interval R (set of real number)
3. $f(x, \theta)$is in the from $\log f(x, \theta) = v(\theta)f(x) + v(\theta) + w(x)$for each$x \in s$ and $\theta \in \Omega$where .
   i) $v(0)$ is twice differentially with $v'(\theta) \neq 0$
   ii) $[1, t(x)]$ are linearly independent i.e., $c_0 + c_1 t(x) = 0 <=> c_0 = c_1 = 0$, then $[f(x, \theta), \theta \in \Omega]$ is called an one parameter exponential family of distribution.

EX: $f(x,\theta) = \theta(1-\theta)^n; x = 0,1,2,\ldots\ldots$
$0 < \theta < 1$. This is geometricpmf.
Sample space S; $[X: x = 0,1,2,\ldots] = [0,1,2,\ldots\ldots]$
$\Omega = [0: 0 < \theta < 1] = [0,1]$
→ S does nto depend on θ.
→Ω is an open interval in R.
→ $\log f(x,\theta) = \log\theta + x\log(1-\theta)$
Where, $U(\theta) = \log(1-\theta), T(X) = x, T(\theta) = \log\theta$ and $W(X) = 0$
→ $U(\theta) = \log(1-\theta)$
→ $U'(\theta) = \frac{-1}{1-\theta} \neq 0$ for $\theta \in \Omega$
→ $U''(\theta) = \frac{(-1)(-1)(-1)}{(1-\theta)^2} = \frac{-1}{(1-\theta)^2} < \infty \; for \;\; \theta \in \Omega$
Then, $U(\theta) \neq 0$ and $U'(\theta)$exists.
Consider,
→ $C_0 + C_1 x = 0, \forall\, x \in S$, if $x = 0$, then $C_0 = 0$
if $x = 1,2,3,\ldots$ and $c_0 = 0$ then $C_1 = 0$
→ $C_0 = C_1 = 0 => 1$ *and x are linearly independent* .
Hence, $[f(x,\theta); \theta \in \Omega]$ is a one parameter exponential family of distribution.

2.6 K - PARAMETER EXPONENTIAL FAMILY:

Let x be a real valued random variable with probability density function $f(x: \theta_1, \theta_2, .., \theta_k)$,
where $(\theta_1, \theta_2, .., \theta_k) \in \Omega$ family of distribution $\{P(x: \theta_1, \theta_2, \ldots, \theta_k), (\theta_1, \theta_2, \ldots., \theta_k) \in \Omega\}$ satisfying
1) $S = \{x: f(\theta_1, \theta_2, \ldots., \theta_k) > 0\}$does not depend on θ.
2) Ω is open set in R^k
3)For each $x \in S$ and $(\theta_1, \theta_2, \ldots, \theta_k) \in \Omega \log f(x: \theta_1, \theta_2, \ldots, \theta_k)$

$$= \sum_{i=1}^{k} \mu(\theta_1, \theta_2, .., \theta_k) t_j + V(\theta_1, \theta_2, \ldots, \theta_k) + w(x)$$

i) $\mu_j(\theta_1, \theta_2, \ldots, \theta_k)$ has partial deviation with respect to $\theta_1, \theta_2, \ldots, \theta_k$ of oder "2"and jacobian

$$\left|\frac{\partial(\mu_1, \mu_2, \ldots, \mu_k)}{\partial(\theta_1, \theta_2, \ldots., \theta_k)}\right| \neq 0 \; \forall \theta_1, \theta_2, \ldots, \theta_k \in \Omega$$

ii) $\Omega = \{\mu: -\infty < \mu < \infty, \sigma^2 = \sigma > 0\} = (-\infty, \infty)X(0, \infty)$is an open interval in R^2

iii) $$\log f(x: \mu, \sigma^2) = \frac{-1}{2}\log(2\pi) - \frac{1}{2}\log\sigma^2$$
$$= \frac{-1}{2\sigma^2}x^2 + \pi\frac{\mu}{\sigma^2} - \frac{\mu}{2\sigma^2}$$
$$= \left(\frac{-1}{2\sigma}\right)x^2 + \left(\frac{\mu}{\sigma^2}\right)x + \left[\frac{-\mu^2}{2\sigma^2} - \frac{1}{2}\log\sigma^2 - \frac{1}{2}\log(2\pi)\right] + 0$$
$$= \mu_i(\mu, \sigma^2)t_i(x) + \mu_2(\mu, \sigma^2)t_2(x) + v(\mu, \sigma^2)$$
$$\text{Here, } \mu_1(\mu, \sigma^2) = \frac{-1}{2\sigma^2}, t_1(x) = x^2$$
$$\mu_2(\mu, \sigma^2) = \frac{\mu}{\sigma^2}, t_2(x) = x$$

2.7 SUMMARY:

The likelihood function is a fundamental concept in statistical inference, used to estimate parameters of a probability distribution given observed data. It expresses the probability of the observed data as a function of the parameters. Unlike a probability distribution, which describes the likelihood of data occurring under given parameters, the likelihood function treats the parameters as variables and the data as fixed.

A key distinction exists between likelihood and joint probability. The joint probability function describes the probability of all observed data points occurring together, considering them as random variables. In contrast, the likelihood function is not a probability function in itself; rather, it measures how well a given set of parameters explains the observed data.

In essence, while joint probability is used to define the probability distribution of data, likelihood functions are used to estimate parameters by maximizing the likelihood of the observed data under a given model. This leads to methods such as Maximum Likelihood Estimation (MLE), which is widely used in statistical modeling.

2.8 KEY WORDS:

- Likelihood Function
- Joint Probability
- Probability Distribution
- Maximum Likelihood Estimation (MLE)
- Parameter Estimation
- Statistical Inference
- Bayesian Analysis
- Conditional Probability
- Probability Density Function (PDF)
- Data Likelihood

2.9 SELF-ASSESSMENT QUESTIONS:

1. What is a likelihood function, and how does it differ from a probability function?
2. How is joint probability different from likelihood?
3. What is the importance of likelihood in statistical inference?
4. Give an example where likelihood estimation is useful in real-world applications.
5. How does the likelihood function relate to Bayesian statistics?
6. Why is the likelihood function not a probability distribution?
7. Can the likelihood function be greater than 1? Why or why not?

2.10 SUGGESTED READINGS:

1. Casella, G., & Berger, R. L. (2001). *Statistical Inference*. Duxbury Press.
2. Wasserman, L. (2004). *All of Statistics: A Concise Course in Statistical Inference*. Springer.
3. Bickel, P. J., & Doksum, K. A. (2015). *Mathematical Statistics: Basic Ideas and Selected Topics*. CRC Press.
4. Rice, J. A. (2006). *Mathematical Statistics and Data Analysis*. Cengage Learning.
5. Lehmann, E. L., & Casella, G. (1998). *Theory of Point Estimation*. Springer.

Dr. K. Kalyani

LESSON- 3
SUFFICIENCY

OBJECTIVES:

- Understand the concept of **sufficiency** in statistical estimation.
- Define and explain the **Factorization Theorem** for sufficient statistics.
- Identifying and derive **sufficient statistics** for different probability distributions.
- Solve problems involving sufficiency in **binomial, normal, Poisson** etc and other distributions.

STRUCTURE:

3.1 SUFFICIENCY:

An estimator is said to be sufficient for a parameter, if it contains all the information in the sample regarding the parameter.

If $T = t(x_1, x_2, ..., x_n)$ is an estimator of a parameter θ, based on a sample $x_1, x_2, ..., x_n$ of size n from the population with density $f(x, \theta)$ such that the conditional distribution of $x_1, x_2, ..., x_n$ given T, is independent of θ, then T is sufficient estimator for θ.

Note:

- Sufficiency is another important concept in statistical estimation that relates to the amount of information an estimator captures about the population parameter.
- A sufficient estimator contains all the relevant information about the parameter that is available in the sample data.
- Sufficiency is desirable because it ensures that no information is lost when using the estimator to make inferences about the population.

3.1.1 Remarks on Sufficiency:

- A sufficient statistic contains all the information in the sample about the parameter, but it does **not** necessarily provide the best estimator (additional properties like unbiasedness and efficiency are needed).

- Minimal sufficient statistics provide the most compact form of sufficiency, removing redundancy in the data.
- Not all statistics are sufficient; an arbitrary function of the sample may lose information about the parameter.
- The **factorization theorem** provides an easy way to check sufficiency mathematically.
- While sufficiency is widely used in classical statistics, in **Bayesian analysis**, the concept of sufficiency is complemented by the idea of conjugacy in prior distributions.

3.1.2 Applications of Sufficiency:

1. **Parameter Estimation**
   - Helps in constructing estimators that retain all the information about a parameter.
   - Used in deriving the **Minimum Variance Unbiased Estimator (MVUE)**.
2. **Data Reduction**
   - Reduces the sample data to a lower-dimensional statistic without losing essential information about the parameter.
   - Essential in large-sample inference to simplify calculations.
3. **Rao-Blackwell Theorem Application**
   - Improves an unbiased estimator by conditioning on a sufficient statistic, leading to a lower variance estimator.
4. **Exponential Family Distributions**
   - Many common probability distributions (normal, binomial, Poisson, gamma, etc.) have natural sufficient statistics, making analysis easier.
5. **Bayesian Inference**
   - In Bayesian estimation, sufficiency helps in defining the posterior distribution effectively, reducing computational complexity.
6. **Hypothesis Testing**
   - Sufficient statistics play a crucial role in **Neyman-Pearson Lemma** for finding most powerful tests.
7. **Information Theory**
   - Links with **Fisher Information**, showing that a sufficient statistic retains all information about the parameter in the data.

3.2 NEYMAN FACTORIZATION THEOREM:

STATEMENT :

Let $X = (x_1, x_2, \ldots, x_n)$ be a random sample drawn from a population $f(x,\theta), \theta \in \Omega$,then the statistic *T(x)* is said to be sufficient for θ if and only if

$$L(x,\theta) = g[T(x),\theta]h(x) ----\rightarrow (1)$$

where,

$L(x,\theta) = \Pi_{i=1}^{n} f(x_i,\theta)$ is a likelihood function of the random variable X.

$g[T(x),\theta]$ is a non-negative function of *T(x)* and$\theta, and\ h(x)$ is a function of x and is independent of parameter θ .

PROOF:

We assume that random variable x is discrete.

Necessary Condition:
Given $T(x)$ is sufficient estimator for θ.
Then, We Prove that $Eq^n(1)$ holds good.
condider L. H. S

$$L(x,\theta) = P(X = x) = P(X = x) = P(X = x \cap T(x) = t)$$
$$= P\big(T(X = t)\big)P[X = x/T(x) = t]$$
$$L(x,\theta) = g[T(x),\theta].h(x) \longrightarrow (2)$$

Where $P(T(x) = t) = g[T(x),\theta]$
$P[X = x/T(x) = t] = h(x)$

consider $P(T(x) = t) = g[T(x),\theta] \longrightarrow (3)$
which is function of $T(x)$ and θ
consider $P[X = x/T(x) = t] = h(x) \longrightarrow (4)$
Which is a function of random variable x and independent of θ.
From $eq^n(2),(3)\&(4)$

$$L(x,\theta) = g[T(x),\theta].h(x)$$

Sufficiency condition:
Given, $L(x,\theta) = g[T(x),\theta].h(x)$
To Prove that $T(x)$ is sufficient for θ.
To Show that $P[X = x/T(x) = t]$ is independent of θ.
Let $B_t[x: T(x) = t]$

$$\text{Consider, } P_\theta[T(x) = t] = \sum_{x\in B_t} P(X = x) = \sum_{x\in B_t} L(x,\theta)$$

$$P_\theta[T(x) = t] = \sum_{x\in B_t} g[T(x),\theta].h(x) \longrightarrow (5)$$

$$\text{Consider, } P[X = x/T(x) = t] = \frac{P(X = x \cap T(x) = t)}{P[T(x) = t]}$$
$$= \frac{P(X = x)}{P(T(x) = T)}$$
$$= \frac{L(x,\theta)}{\sum_{x\in B_t} g[T(x),\theta].h(x)}$$
$$= \frac{g[T(x),\theta].h(x)}{g[T(x),\theta]\sum_{x\in B_t} h(x)}$$
$$= \frac{h(x)}{\sum_{xB_t} h(x)}$$

which is function of x and independent prameter θ.
Therefore, $T(x)$ is a sufficient parameter θ.

3.3 PROBLEMS ON SUFFICIENCY:

1. Let $x_1, x_2, \dots \dots, x_n \sim P(\lambda)$, $T = \sum_{i=1}^n x_i$, whether it is Sufficient estimator of λ.
Solution: Given that $x_1, x_2, \dots \dots, x_n \sim \mathbf{P(\lambda)}$, then its probability mass function is given by

$$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}; x = 0,1,2,\dots \dots, \text{infi}$$
$$= 0, \text{otherwise}$$

here, λ is the parameter

w.k.t, $t = \sum x_i \sim P(n\lambda)$

$$P(T=t) = \frac{e^{-n\lambda}(n\lambda)}{t!}; t = 0,1,2, \dots \dots ....$$

$$\text{Then, } P(X = x_1, x_2, \dots, x_n | T = t) = \frac{P(x_1, x_2, \dots \dots, x_n)}{P(T=t)}$$

$$= \frac{P(X=x_1)P(X=x_2) \dots \dots \dots .. P(X=x_n)}{P(T=t)}$$

$$= \frac{\frac{e^{-\lambda}\lambda^{x_1}}{x_1!} \cdot \frac{e^{-\lambda}\lambda^{x_2}}{x_2!} \cdot \frac{e^{-\lambda}\lambda^{x_3}}{x_3!} \dots \dots \dots \frac{e^{-\lambda}\lambda^{x_n}}{x_n!}}{\frac{e^{-n\lambda}(n\lambda)^t}{t!}}$$

$$= \frac{e^{-n\lambda}\lambda^{\sum x_i}}{\pi_{i=1}^n x_i!} \cdot \frac{t!}{e^{-n\lambda}(n\lambda)^t}$$

$$= \frac{\lambda \sum x_i}{\pi_{i=1}^n x_i!} \cdot \frac{\sum x_i!}{(n\lambda)\sum x_i}$$

$$= \frac{\lambda^{\sum x_i}}{\pi_{i=1}^n x_i!} \cdot \frac{\sum x_i!}{n^{\sum x_i}\lambda^{\sum x_i}}$$

$$= \frac{\sum x_i}{\pi_{i=1}^n x_i!} \cdot \frac{1}{n \sum x_i!}$$

Thus, $t = \sum x_i$ is a sufficient statistic for the parameters.

2. Let x_1, x_2 follows $P(\lambda)$, $T = x_1 + 2x_2$ whether it is sufficient estimator for λ

Solution: Given that $x_1, x_2 \sim P(\lambda)$

P. m. f of Poission distribution is $P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}$

$T = x_1 + 2x_2 \sim P(2\lambda)$

$$P(X = x_1, X = 2x_2 / T(X) = t) = \frac{P(X = x_1, X = 2x_2)}{P(T=t)}$$

$$= \frac{\frac{e^{-\lambda}\lambda^{x_1}}{x_1!} \cdot \frac{e^{-\lambda}\lambda^{2x_2}}{2x_2!}}{\left(\frac{e^{-2\lambda}(2\lambda)^{x_1+x_2}}{(x_1+2x-2)!}\right)}$$

$$= \frac{e^{-2\lambda}\lambda^{x_1+2x_2}}{2x_2!\, x_1!} \cdot \frac{(x_1+2x_2)!}{e^{-2\lambda}\lambda^{x_1+2x_2} 2^{x_1+2x_2}}$$

$$= \frac{t!}{2^{t+1} x_1!\, x_2!}$$

Which is independent of $'\lambda'$.

Therefore, $T = x_1 + 2x_2$ is a sufficient statisitc for λ.

3. Find sufficient estimator for θ an exponential distribution.

Solution: P.d.f of exponential distribution is $f(x,\lambda) = \lambda e^{-\lambda x}, x > 0, \lambda > 0$

Let $x_1, x_2, \ldots.. x_n$ be random sample drawn from exponential distribution likelyhood function is,

$$L = \pi_{i=1}^{n} f(x_i, \lambda)$$
$$= f(x_1,\lambda) f(x_2,\lambda) \ldots\ldots\ldots f(x_n,\lambda)$$
$$= \lambda e^{-\lambda x_1} \lambda e^{-\lambda x_2} \ldots\ldots\ldots \lambda e^{-\lambda x_n}$$
$$= \lambda^n e^{-\lambda}(x_1 + x_2 + \cdots + x_n)$$
$$= \lambda^n \left[e^{-\lambda \Sigma x_i}\right]$$
$$= \lambda^n e^{-\lambda \Sigma x_i}$$
$$= g\theta[t(x)]h(x)$$
$$= \lambda^n e^{-\lambda \Sigma x_i}(1)$$
$$= \lambda^n e^{-\lambda \Sigma x_i}$$

Hence by neyman factorization $t(x) = \sum_{i=1}^{n} x_i$ is sufficient estimator for λ .

4. **Let $x_1, x_2, \ldots.., x_n$ be a random sample from a Bernoulli population with parameter $'p'$, 0<p<1 i.e.,**

$$x_i = \begin{cases} 1, \text{ with probability p} \\ 0, \text{ with probability q}=(1-p) \end{cases}$$

Find the Sufficient Statistics.

Solution: $T = t(x_1, x_2, \ldots.. x_n) = x_1 + x_2 + \cdots + x_n \sim B(n,p)$

$$P(T = k) = \binom{n}{k} p^k \ (1-p)^{n-k} ; x = 0,1,2, \ldots\ldots., n$$

The conditional distribution of $x_1, x_2, \ldots.., x_n$ given T is:

$$P(x_1 \cap x_2 \cap \ldots.. \cap x_n / T = k) = \frac{P(x_1 \cap x_2 \cap \ldots.. \cap x_n \cap T = k)}{P(T = k)}$$

$$= \begin{cases} \dfrac{p^k (1-p)^{n-k}}{\binom{n}{k} p^k \ (1-p)^{n-k}} = \dfrac{1}{\binom{n}{k}} \\ 0, \text{ if} \sum_{i=1}^{n} x_i \neq k \end{cases}$$

Since this does not depend on $'p'$, $T = \sum_{I=1}^{n} x_i$,is sufficient for ‘p’.

5. Find sufficient estimator of parameter P in binomial distribution.

Solution: P. d. f of binomial distribution is, $f(x, P) = n_{C_x} p^x q^{n-x}$

Let $x_1, x_2, \ldots, x_n$ be a random sample drawn from binomial distribution

the likelihood function is,

$$L = \pi_{i=1}^{n} p(x_i)$$
$$= p(x_1)p(x_2) \ldots \ldots p(x_n)$$
$$= n_{C_x} p^{\Sigma x_n} q^{n-\Sigma x_i}$$
$$= g\theta[t(x)]h(x)$$

Where, $g\theta[t(x)] = n_{C_x} p^{\Sigma x_i} q^{n-\Sigma x_i}, h(x) = 1$

Hence, by neyman factorization,

$t(x) = \sum x_i$ is sufficient estimator for λ.

6. Let $x_1, x_2, \ldots, x_n$ random sample from a population with pdf $f(x, \theta) = \theta x^{\theta-1}; 0 < x < 1, \theta > 0$. Show that $t = \pi_{i=1}^{n} x_i$ is sufficient for θ.

Solution: pdf of exponential distribution, $f(x_i, \theta) = \theta x^{\theta-1}$.

Let $x_1 . x_2, \ldots, x_n$ be random sample drawn from exponential distribution

The likelihood function is,

$$L = (x, \theta) = \pi_{i=1}^{n} f(x_i, \theta)$$
$$= \pi_{i=1}^{n} \theta x_i^{\theta-1}$$
$$= \theta^n \pi_{i=1}^{n} x_i^{\theta-1}$$
$$= \theta^n \pi_{i=1}^{n} x_i^{\theta} . \frac{1}{\pi_{i=1}^{n} x_i}$$
$$= g[T(x), \theta]h(x)$$

Hence, by neyman factorization, $t = \pi_{i=1}^{n} x_i$ is a sufficient statistic for θ.

7. Let $X_1, X_2, \ldots, X_n$ be a random Sample from a distribution with p.d.f $f(x,\theta) = e^{-(x-\theta)}; \theta < x < \infty, -\infty < \theta < \infty$ Obtain the sufficient statistics for θ.

Solution: Here $L = \sum_{i=1}^{n} f(x_i, \theta) = \sum_{i=1}^{n} e^{-(x_i-\theta)} = \exp\left(\sum_{i=1}^{n} x_i\right) X \exp(n\theta)$ (1)

$Y_1, Y_2, \ldots, Y_n$ denote the ordered statistics of the random sample such that $Y_1 < Y_2 < \ldots < Y_n$. The p.d.f of the smallest observation Y_1 is given by:

$g_1(y_1, \theta) = n[1 - F(y_1)]^{n-1} f(y_1, \theta)$, where $F(.)$ is the distribution function corresponding to p.d.f $f(.)$.

Now $F(x) = \int_{\theta}^{x} e^{-(x-\theta)} dx = \left|\frac{e^{-(x-\theta)}}{-1}\right|_{\theta}^{x} = 1 - e^{-(x-\theta)}$

$\therefore g_1(y_1, \theta) = n\left[e^{-(y_1-\theta)}\right]^{n-1} e^{-(y_1-\theta)} = \{n\, e^{-n(y_1-\theta)}, \theta < y_1 < \infty$

$0, otherwise$

Thus the likelihood function (1) of $X_1, X_2, \ldots, X_n$ may be expressed as

$$L = e^{n\theta}\exp\left(-\sum_{i=1}^{n} x_i\right) = n\exp\{-n(y_i-\theta)\}\left\{\frac{\exp\left(-\sum_{i=1}^{n} x_i\right)}{n\exp(-ny_1)}\right\}$$

$$=g_1(\min x_i,\theta)\left(\frac{\exp\left(-\sum_{i=1}^{n} x_i\right)}{n\exp(-n\min x_i)}\right)$$

Hence by Fisher- Neymann criterion, the first order statistics $Y_1 = \min(X_1, X_2, ..., X_n)$ is a sufficient statistic for θ.

8. Let $X_1, X_2, ..., X_n$ be a random Sample from Cauchy population: $f(x,\theta) = \frac{1}{\pi}\frac{1}{1+(x-\theta)^2}; -\infty < x < \infty, -\infty < \theta < \infty$ Examine if there exists a sufficient statistic for θ.

Solution: $L = \prod_{i=1}^{n} f(x_i,\theta) = \frac{1}{\pi^n}\prod_{i=1}^{n}\left\{\frac{1}{1+(x_i-\theta)^2}\right\}$

Hence by Factorisation theorem, there is no single statistic, which alone, is sufficient estimator of θ.

However, $L(x,\theta) = k_1(X_1, X_2, ..., X_n, \theta).k_2(X_1, X_2, ..., X_n, \theta)$

The whole set $(X_1, X_2, ..., X_n)$ is jointly sufficient for θ.

9. If X_1, X_2 are i.i.d from P.d(λ). Prove that $T(X_1, X_2) = X_1 + 3X_2$ is not sufficient for θ.

Solution:

$T(x_1, x_2) \sim p(4\lambda)\left[\text{ since}, x_1, x_2 \text{are Id } T(x_1, x_2) = \alpha x_1 + \beta x_2 + (\alpha+\beta)^{\lambda}\right]$

Consider $P(x_1 = 1, x_2 = 1)/T(x_1, x_2) = 4$

$$\frac{P(x_1 = 1), (x_2 = 1)(T(x_1, x_2) = 4)}{P(T(x_1, x_2) = 4)} - - - - - \rightarrow (A)$$

$$P(x_1 = 1)(x_2 = 1)T(x_1, x_2 = 4) = P(x_1 = 1)P(x_2 = 1), x_1 + 3x_2 = 4$$
$$= P(x_1 = 1\ x_2 = 1)$$
$$= \frac{e^{-\lambda}\lambda^x}{x!}\frac{e^{-\lambda}\lambda^x}{x!}$$
$$= \frac{e^{-\lambda}\lambda^1}{1!}\frac{e^{-\lambda}\lambda^1}{1!} = e^{-2\lambda}\lambda - - - - - \rightarrow (2)$$

$$\text{Consider}, P(T(x_1, x_2) = 4) = P(x_1 + 3x_2 = 4)$$
$$= P(x_1 = 1, x_2 = 1)$$
$$= P(x_1 = 4, x_2 = 0)$$
$$= \frac{e^{-\lambda}\lambda^x}{x!}\frac{e^{-\lambda}\lambda^x}{x!} + \frac{e^{-\lambda}\lambda^x}{x!}\frac{e^{-\lambda}\lambda^0}{0!}$$
$$= e^{-\lambda}\lambda . e^{-\lambda}\lambda + \frac{e^{-\lambda}\lambda^4}{4!}.(e)^{-\lambda}$$

$$= e^{-2\lambda}\lambda^2\left(1+\frac{\lambda^2}{4}\right) ------\rightarrow (3)$$

from (2)&(3), substituting the value in (A), we get;

$$\frac{P(x_1 = 1), (x_2 = 1)(T(x_1, x_2) = 4)}{P(T(x_1, x_2) = 4)} = \frac{e^{-2\lambda}\lambda^2}{e^{-2\lambda}\lambda\left(1+\frac{\lambda^2}{4!}\right)}$$

$$= \frac{1}{1+\frac{\lambda^2}{4!}}$$

$$= \left(1+\frac{\lambda^2}{4!}\right)^{-1}$$

It is not independent of θ

So, the statistic $T(x_1, x_2) = x_1 + 3x_2$ is not sufficient.

3.4 SUMMARY:

Sufficiency is a fundamental concept in statistical estimation that concerns how much information a statistic retains about an unknown parameter. A statistic T(X) is said to be **sufficient** for a parameter θ if it captures all the information about θ present in the sample X. Mathematically, sufficiency is defined using the **Factorization Theorem**, which states that a statistic T(X) is sufficient for θ if the joint probability (or likelihood) function of the sample can be factorized as:

$$f(X|\theta)=g(T(X),\theta)h(X)$$

where g(T(X),θ) depends on θ only through T(X) and h(X) is independent of θ.

Sufficient statistics help in reducing data without losing essential information about the parameter. They are particularly useful in constructing minimum variance unbiased estimators (MVUEs) and in the Rao-Blackwell theorem, which improves an estimator using a sufficient statistic.

3.5 KEY WORDS:

- Sufficient Statistic
- Factorization Theorem
- Likelihood Function
- Minimal Variance Unbiased Estimator (MVUE)

3.6 SELF-ASSESSMENT QUESTIONS:

1. Define a sufficient statistic and explain why it is important in estimation theory.
2. State and prove the Factorization Theorem for sufficiency.
3. What is the relationship between sufficiency and completeness in statistics?
4. How can the Rao-Blackwell theorem be used to improve an estimator using a sufficient statistic?
5. Explain how sufficiency is applied in the exponential family of distributions.
6. Find a sufficient statistic for the parameter θ in a binomial distribution X~Bin(n,θ).

7. What is the difference between minimal sufficiency and sufficiency?
8. How does the concept of sufficiency contribute to finding an MVUE?
9. In what ways do sufficient statistics help in reducing data dimensionality without losing estimation accuracy?
10. Illustrate sufficiency with an example from normal distribution.

3.7 SUGGESTED READINGS:

1. **Casella, G., & Berger, R. L.** - *Statistical Inference*
2. **Hogg, R. V., McKean, J. W., & Craig, A. T.** - *Introduction to Mathematical Statistics*
3. **Mood, A. M., Graybill, F. A., & Boes, D. C.** - *Introduction to the Theory of Statistics*
4. **Lehmann, E. L., & Casella, G.** - *Theory of Point Estimation*
5. **Bickel, P. J., & Doksum, K. A.** - *Mathematical Statistics: Basic Ideas and Selected Topics*
6. **Rohatgi, V. K., & Saleh, A. K. Md. E.** - *An Introduction to Probability and Statistics*
7. **Davidson, J.** - *Statistical Inference for Engineers and Scientists*
8. **Dudewicz, E. J., & Mishra, S. N.** - *Modern Mathematical Statistics.*

Dr. K. Kalyani

LESSON- 4
COMPLETENESS AND MINIMAL SUFFICIENCY

OBJECTIVES:

After learning about **Complete Sufficiency and Minimal Sufficiency**, students will be able to:

- **Identify Minimal Sufficient Statistics** – Find the minimal sufficient statistic for a given probability distribution.
- **Differentiate Between Sufficiency Types** – Explain the difference between **sufficiency, minimal sufficiency, and complete sufficiency** with examples.
- **Explain the Role of Completeness** – Understand why a **complete** statistic ensures no nontrivial function has an expected value of zero.
- **Analyze Distributions for Sufficiency** – Determine sufficient and complete statistics for common distributions like **Normal, Exponential, and Bernoulli**.
- **Use Likelihood Ratio to Find Minimal Sufficiency** – Apply the likelihood ratio method to derive minimal sufficient statistics.
- **Interpret the Role of Sufficiency in Estimation** – Understand how sufficient statistics reduce data complexity while retaining parameter information.

STRUCTURE:

4.1 COMPLETENESS:

Completeness in statistics ensures a statistic captures all available information about a parameter. It's crucial for determining optimal estimators and plays a key role in hypothesis testing and parameter estimation by guaranteeing uniqueness in statistical procedures.

This concept interacts closely with sufficiency, forming a powerful framework for inference. Completeness prevents the existence of multiple unbiased estimators with the same expectation, acting as a "maximal" property to ensure no information is lost when using a statistic.

4.1.1 Definition of Completeness:

- Completeness serves as a fundamental concept in theoretical statistics enabling statisticians to determine optimal estimators
- Plays a crucial role in hypothesis testing and parameter estimation by ensuring uniqueness of certain statistical procedures
- Relates closely to sufficiency, another key concept in statistical theory, forming a powerful framework for inference.

4.1.2 Formal Mathematical Definition:

- Defined for a family of probability distributions and a statistic T(X)
- A statistic T(X) is complete if for any measurable function $E\left[g\left(T\left(x\right)\right)\right]=0$ for all distributions in the family $\Rightarrow g(T(X))=0$ almost surely Implies no non-zero function of T(X) has zero expectation for all distributions in the family
- Ensures T(X) captures all available information about the parameter of interest

4.1.3 Intuitive Explanation:

- Completeness indicates a statistic contains all relevant information about a parameter
- Prevents the existence of two different unbiased estimators with the same expectation
- Acts as a "maximal" property, ensuring no information is lost when using the statistic
- Allows for unique determination of optimal estimators in many statistical problems

4.1.4 Relationship to Sufficiency:

- Sufficiency reduces data without loss of information, while completeness ensures uniqueness
- Complete sufficient statistics combine both properties, providing powerful tools for inference
- Not all sufficient statistics are complete, and not all complete statistics are sufficient
- Completeness often "complements" sufficiency in statistical theory and practice

4.1.5 Properties of Complete Statistics:

- Complete statistics form the backbone of many optimal estimation procedures in theoretical statistics

- Enable the development of uniformly minimum variance unbiased estimators (UMVUEs)
- Provide a foundation for proving uniqueness and optimality in statistical inference

4.2 TYPES OF COMPLETENESS:

- Different notions of completeness exist to address various statistical scenarios and requirements
- Each type of completeness offers unique properties and applications in theoretical statistics
- Understanding these variations helps in selecting appropriate techniques for specific problems

BOUNDED COMPLETENESS:

- Relaxes the completeness condition to hold only for bounded functions
- A statistic T is boundedly complete if $E[g(T(X))]=0$ for all distributions $\Rightarrow g(T(X))=0$ a.s. for all bounded g.
- Often easier to verify than full completeness
- Sufficient for many practical applications in statistical inference

Sequential completeness:

- Applies to sequences of statistics rather than a single statistic
- A sequence of statistics {Tn} is sequentially complete if: $E[g(Tn(X))]\to 0$ for all distributions $\Rightarrow g\left(T_n\left(x\right)\right)\xrightarrow{p}0$
- Useful in asymptotic theory and sequential analysis
- Allows for the study of limiting behavior of estimators and test statistics

4.3 LIMITATIONS OF COMPLETENESS:

- While powerful, completeness has certain limitations and challenges in statistical theory
- Understanding these limitations helps in proper application and interpretation of results
- Encourages the development of alternative approaches for scenarios where completeness fails.

4.4 COMPLETENESS OF A FAMILY OF DENSITIES (OR) DISTRIBUTIONS:

A family of densities $f(x, \theta): \theta \in \Theta$ is said to be complete (or said to be complete family of densities) if h(x) is only function of whose expectation is identically zero $\forall \theta \in \Theta$ implies h(x) is identically zero (almost surely),

$$x \sim f(x, \theta); \theta \in \Theta$$

Take any function z(x) of a random variable x such that $E[z(x)] = o \ \forall \ \theta$
$\Rightarrow z(x) = 0$ (almost surely) .

P is complete, if $E[z(x)] = 0 \; \forall \, \theta \Rightarrow z(x) = 0$. In other words, the only function of the random variables x whose expectation is zero.

4.5 PROBLEMS BASED ON COMPLETE SUFFICIENCY:

1) **Binomial distribution** : $\mathbf{P(X = x) = (n_x)P^x(1 - P)^{n-x}}$;
$\mathbf{x = 0, 1, 2, 3, \dots, n}$ **is binomial distribution**.

Proof:

$$\text{Let } z(x) = E\{z(x)\} = \sum z(x)p(x)$$

$$\Rightarrow \sum_{x=0}^{n} z(x)(n_x)P^x(1-P)^{n-x} = 0$$

$$\Rightarrow (1-P)^n \sum_{x=0}^{n} z(x)(n_x)\left(\frac{P}{1-P}\right) = 0$$

$$\Rightarrow (1-P)^n \sum_{x=0}^{n} z(x)(n_x)\pi^x = 0 \left(\text{Say} \frac{P}{1-P} = \pi\right)$$

$$\Rightarrow (1-P)^n \sum_{x=0}^{n} ax\pi^x = 0$$

$$\Rightarrow (1-P)^n\{a_0\pi^0 + a_1\pi^1 + a_2\pi^2 + \cdots + a_n\pi^n\}$$

$$E\{z(x)\} = 0$$

$$\Rightarrow a_0\pi^0 + a_1\pi^1 + a_2\pi^2 + \cdots + a_n\pi^n = 0[\text{since}, (1-P)^n \neq 0]$$

$$\Rightarrow a_i = 0 \; \forall \, i = 1,2,\dots, mmn$$

$$\text{But } a_i = z(i)$$

$$\Rightarrow z(i) = 0 \; \forall \, i$$

$$\Rightarrow z(x) = 0$$

Therefore, Binomial distribution is complete.

2) **Poisson distribution** : $\mathbf{P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}}$,

$\mathbf{x = 0, 1, 2, 3, \dots \dots, \infty}$ **is poisson complete**.

Proof:

$$\text{Let } z(x) \Rightarrow E\{z(x)\} = \sum z(x)p(x)$$

$$\Rightarrow \sum_{x=0}^{\infty} z(x)\frac{e^{-\lambda}\lambda^x}{x!} = 0$$

$$\Rightarrow e^{-\lambda}\sum_{x=0}^{\infty} z(x)\frac{\lambda^x}{x!} = 0$$

$$\Rightarrow e^{-\lambda}\sum_{x=0}^{\infty} a_x\pi^x = 0 \; (\text{since}, \frac{z(x)}{x!} = a_x, \lambda^x = \pi^x$$

$$\Rightarrow e^{-\lambda}[a_0\pi^0 + a_1\pi^1 + \cdots] = 0$$

$$E[z(x)] = 0$$

$$\Rightarrow a_0\pi^0 + a_1\pi^1 + a_2\pi^2 + \cdots += 0 \; (\text{since}, e^{-\lambda} \neq 0)$$

$$\Rightarrow a_i = 0 \; \forall \, i = 1,2,3, \dots, n.$$

But $a_i = z(i)\lambda' = 0; \rightarrow z(x) = 0.$

Therefore, Poisson distribution is complete.

3) Exponential distribution :

$$\mathbf{f(x) = \theta e^{-\theta x}}\ ; \mathbf{x} > 0$$
$$\boldsymbol{\theta} > 0$$

Is exponential complete.

Proof:

$$\text{Let } z(x) \ni E[z(x)] = \int z(x)f(x)dx$$
$$= \int_0^\infty z(x)\theta e^{-\theta x}dx$$
$$= \theta \int_0^\infty z(x)e^{-\theta x}dx$$
$$= \theta e^{\theta} \int_0^\infty z(x)e^{-x}dx$$
$$= \theta e^{\theta} \int_0^\infty ax\frac{1}{e^x}dx$$
$$= \theta e^{\theta} \int_0^\infty ax\left(\frac{1}{e}\right)^x dx$$
$$= \theta e^{\theta} \int_0^\infty ax\pi^x dx \left(\text{since}, z(x) = ax, \frac{1}{e^x} = \pi\right)$$

This is so form the theory of laplace transformation,

$$e[z(x) = 0$$
$$\Rightarrow a_i = 0$$
$$z[x] = 0$$

Therefore, Exponential distribution is complete.

4.6 MINIMAL SUFFICIENCY:

Intuitively, a minimal sufficient statistic for parameter θ is the one that collects the useful information in the sample about θ*but only the essential one*, excluding any superfluous information on the sample that does not help on the estimation of θ.

Observe that, if T is a sufficient statistic and T′=φ(T) is also a sufficient statistic, being φ a non-injective mapping, then T′ condenses more the information. That is, the information in T can't be obtained from that one in T′ because φ can'tbe inverted, yet still both collect the sufficient amount of information. In this regard, φ acts as a one-way compressor of information.

A minimal sufficient statistic is a sufficient statistic that can be obtained by means of (not necessarily injective but measurable) functions of *any* other sufficient statistic.

A statistic T(x) is said to be minimal sufficient for θif it is a function of all other sufficient estimator of θ.

(OR)

T(X) is a minimal sufficient statistic if and only if [f(x)/f(y)] is independent of θ.

$\Rightarrow T(x) = T(y)$ as a consequence of fisher's factorization theorem.

4.6.1 Definition (Minimal Sufficient Statistic):

A sufficient statistic T for θ is *minimal sufficient* if, for any other sufficient statistic $\tilde{T}$, there exists a measurable function φ such that

$$T=\varphi(\tilde{T})$$

The factorization criterion of Theorem provides an effective way of obtaining sufficient statistics that *usually* happen to be minimal. A guarantee of minimality is given by the next theorem.

4.7 MINIMAL SUFFICIENCY VS COMPLETENESS:

- Minimal sufficient statistics reduce data to the smallest possible set without losing information
- Completeness ensures uniqueness but doesn't necessarily imply minimal sufficiency
- A statistic can be complete without being minimally sufficient (over complete)
- Minimal sufficient statistics that are also complete provide the most concise and informative summaries of data

4.8 PROBLEMS BASED ON COMPLETE SUFFICIENCY:

1) Binomial Distribution:

$$P(X = x) = (n_x)p^x p(1-p)^{n-x}; x = 0,1,2,3 \dots, n, p > 0.$$

$$\frac{p(x)}{p(y)} = \frac{p^x(1-p)^{n-x}}{p^y(1-p)^{n-y}}$$

$$\pi_{i=1}^{n}\frac{p(x)}{p(y)} = \frac{p\sum_{i=1}^{n} x_i(1-p)^{n-\sum_{i=0}^{n} x_i}}{p\sum_{(i=1)}^{n} y_i(1-p)^{n-\sum_{i=1}^{n} y_i}}$$

$$\text{for,} \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \text{ and } n - \sum x_i = n - \sum y_i$$

$$\text{Therefore, Minimal sufficient statistic } (\sum x_i, n)$$

2) **Poisson Distribution :**

$$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}$$

$$\frac{p(x)}{p(y)} = \frac{e^{-\lambda}\lambda^x}{x!} \cdot \frac{y!}{e^{-\lambda}\lambda^y} = \frac{\lambda^x y!}{x!\,\lambda^y}$$

$$\pi_{i=1}^{n}\frac{p(x)}{p(y)} = \frac{\lambda^{\sum x_i}\pi_{i=1}^{n} y!}{\lambda^{\sum y_i}\pi_{i=1}^{n} x!}$$

$$= \lambda^{\sum x_i - \sum y_i}\frac{\Pi_{i=1}^{n} y_i!}{\Pi_{i=1}^{n} x_i!}$$

$$\text{for } \sum x_i = \sum y_i, \text{ the above equation is independent of } \lambda.$$

$$\text{Therefore, Minimal sufficient statistic for } \lambda \text{ is } \sum x_i.$$

3) Normal Distribution :

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left[\frac{x-\mu}{\sigma}\right]^2}, -\infty<x<\infty; -\infty<\mu<\infty; \sigma>0$$

$$\frac{f(x)}{f(y)} = \frac{\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left[\frac{x-\mu}{\sigma}\right]^2}}{\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left[\frac{y-\mu}{\sigma}\right]^2}}$$

$$\Pi_{i=1}^{n} \frac{f(x)}{f(y)} = \frac{e^{-\frac{1}{2}\left(\frac{\sum x_i-\mu}{\sigma}\right)^2}}{e^{-\frac{1}{2}\left(\frac{\sum y_i-\mu}{\sigma}\right)^2}}$$

$$\Pi_{i=1}^{n} \frac{f(x)}{f(y)} = \frac{e^{-\frac{1}{2\sigma^2}}(\sum x_i - \mu)^2}{e^{-\frac{1}{2\sigma^2}}(\sum y_i - \mu)^2}$$

$$\Pi_{i=1}^{n} \frac{f(x)}{f(y)} = \frac{e^{-\frac{1}{2\sigma^2}}(\sum x_i^2 + \mu^2 - 2\sum x_i\mu)}{e^{-\frac{1}{2\sigma^2}}(\sum y_i^2 + \mu^2 - 2\sum y_i\mu)}$$

$$\Pi_{i=1}^{n} \frac{f(x)}{f(y)} = \frac{e^{-\frac{1}{2\sigma^2}}\sum x_i^2 . e^{-\frac{1}{2\sigma^2}}\mu^2 . e^{-\frac{1}{2\sigma^2}}(-2\sum x_i\mu)}{e^{-\frac{1}{2\sigma^2}}\sum y_i^2 . e^{-\frac{1}{2\sigma^2}}\mu^2 . e^{-\frac{1}{2\sigma^2}}(-2\sum y_i\mu)}$$

$$\Pi_{i=1}^{n} \frac{f(x)}{f(y)} = \frac{e^{-\frac{1}{2\sigma^2}}(\sum x_i^2 - 2\sum x_i\mu)}{e^{-\frac{1}{2\sigma^2}}(\sum y_i^2 - 2\sum y_i\mu)}$$

Minimal sufficient statisitics for $\mu = \sum x_i, \sigma^2$ is $\sum x_i, \sum x_i^2$

(OR)

Normal Distribution :

$$\frac{f(x)}{f(y)} = \frac{e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}}{e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2}}$$

$$\Pi_{i=1}^{n} \frac{f(x_1)}{f(y_1)} = e\left\{\frac{\sum_{i=1}^{n}(y_i - \mu)^2 - \sum_{i=1}^{n}(x_i - \mu)^2}{2\sigma^2}\right\}$$

$$= e\left\{\frac{\sum y_i^2 + \mu^2 - 2\sum y_i\mu - \sum x_i^2 - \mu^2 + 2\mu\sum x_i}{2\sigma^2}\right\}$$

for constand μ if f $\sum_{i=1}^{n} y_i^2 = \sum_{i=0}^{n} x_i^2$ is MSS for μ

for contant σ^2 if f $\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} x_i$ and $\sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n} x_i^2$

$\Rightarrow \sum x_i, \sum x_i^2$ is MSS for σ^2

4.9 SUMMARY:

In statistical inference, sufficiency is a crucial concept related to data reduction. A statistic is said to be **sufficient** if it captures all the information about a parameter contained in the sample. Two important types of sufficiency are **Complete Sufficiency** and **Minimal Sufficiency**:

- **Complete Sufficiency:** A statistic T(X) is **complete** if no nonzero function of T(X) has an expected value of zero for all parameter values. This ensures that the statistic is not only sufficient but also provides the most information possible about the parameter.
- **Minimal Sufficiency:** A statistic T(X) is **minimally sufficient** if it retains all the information about the parameter while eliminating redundancy. It is the smallest function of the data that remains sufficient.

The **Fisher-Neyman Factorization Theorem** helps determine sufficiency, and completeness often leads to unique estimators, such as the **Uniformly Minimum Variance Unbiased Estimator (UMVUE)**.

4.10 KEY WORDS:

- Sufficient Statistic
- Minimal Sufficient Statistic
- Complete Statistic
- Fisher-Neyman Factorization Theorem
- Uniformly Minimum Variance Unbiased Estimator (UMVUE).

4.11 SELF-ASSESSMENT QUESTIONS:

1. Define a sufficient statistic and explain its significance in parameter estimation.
2. How do you determine if a statistic is minimally sufficient?
3. What is the role of the Fisher-Neyman Factorization Theorem?
4. Provide an example where a statistic is complete but not minimal.
5. How does completeness relate to finding UMVUE estimators?
6. Explain the difference between minimal sufficiency and completeness.
7. Why is a complete statistic useful in unbiased estimation?

4.12 SUGGESTED READINGS:

1. **Casella, G., & Berger, R. L.** (2002). *Statistical Inference*. Duxbury Press.
2. **Lehmann, E. L., & Casella, G.** (1998). *Theory of Point Estimation*. Springer.
3. **Mood, A. M., Graybill, F. A., & Boes, D. C.** (1974). *Introduction to the Theory of Statistics*. McGraw-Hill.
4. **Bickel, P. J., & Doksum, K. A.** (2001). *Mathematical Statistics: Basic Ideas and Selected Topics*. Prentice Hall.
5. **Hogg, R. V., McKean, J. W., & Craig, A. T.** (2012). *Introduction to Mathematical Statistics*. Pearson.

Dr. K. Kalyani

LESSON- 5
UNBIASED ESTIMATORS AND MVUE

OBJECTIVES:

After studying this lesson, the student is able to:

- Understand the Concept of Unbiasedness
- Examine Properties of Unbiased Estimators
- Learn About Minimum Variance Unbiased Estimators (MVUE)
- Calculate UMVUE in Practical Scenarios

STRUCTURE:

5.1 INTRODUCTION:

Unbiasedness is a fundamental concept in statistics, referring to the property of an estimator whose expected value equals the true parameter it aims to estimate. This means that, on average, the estimator neither overestimates nor underestimates the parameter, leading to accurate and reliable estimations across multiple samples.

UNBIASEDNESS:

A statistic $t = (t_1, t_2, \dots, t_n)$, a function of the sample observations $X_1, X_2, \dots, X_n$ is said to be an unbiased estimate of the corresponding population parameter θ, if

$$E(t) = \theta$$

This means the mean value of the sampling distribution of a statistics equal to the parameter.

For example, the sample mean $\bar{x}$ is an unbiased estimate of the population mean μ. The sample proportion of p is an unbiased estimate of the population proportion P.

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

$$E(p) = P$$

Why Unbiasedness Matters in Estimation:

Unbiasedness is an important property in statistics because it ensures that, on average, an estimator gives the correct value of the parameter it is estimating. Here's why it matters:

1. Accuracy in the Long Run

- If an estimator is unbiased, its expected value is equal to the true parameter. This means that if we were to take many repeated samples and compute the estimator each time, the average of those estimates would be correct.
- Example: If we estimate the population mean (μ) using the sample mean ($\bar{X}$), we know that:

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

So, across many samples, ($\bar{X}$) will, on average, give the true value of μ.

2. Avoiding Systematic Errors

- A biased estimator systematically overestimates or underestimates the true parameter, leading to incorrect conclusions.
- Example: If a weighing scale consistently adds 2 kg to every measurement, it is biased because it does not reflect the true weight.

3. Foundation for Other Desirable Properties

- Unbiasedness is a fundamental property that is often combined with **efficiency** (minimum variance) to form the **best unbiased estimator** (e.g., the Minimum Variance Unbiased Estimator, or MVUE).
- Example:

  The sample variance $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ is an unbiased estimator of the population variance σ^2, while the alternative $\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ is biased.

When Unbiasedness Is Not Enough

- While unbiasedness is a good property, it is not the only consideration.
- Some unbiased estimators may have **high variance**, meaning their estimates fluctuate too much from sample to sample. In such cases, a slightly biased but lower-variance estimator might be preferred.

5.2 PROBLEMS BASED ON UNBIASED ESTIMATORS:

1. If X_i is a Bernoulli random variable with parameter p, then: $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is the maximum likelihood estimator (MLE) of p. Is the MLE of p an unbiased estimator of p?

Answer: Recall that if Xi is a Bernoulli random variable with parameter p, then $E(X_i) = p$. Therefore:

$$E(\hat{p}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)$$

$$E(\hat{p}) = \frac{1}{n}\sum_{i=1}^{n} E(X_i)$$

$$E(\hat{p}) = \frac{1}{n}\sum_{i=1}^{n} P$$

$$E(\hat{p}) = \frac{1}{n}(nP) = P$$

The first equality holds because we've merely replaced $\hat{p}$with its definition. The second equality holds by the rules of expectation for a linear combination. The third equality holds because$E(X_i) = p$. The fourth equality holds because when you add the value p up n times, you get np. And, of course, the last equality is simple algebra.

In summary, we have shown that:

$$E(\hat{p}) = P$$

Therefore, the maximum likelihood estimator is an unbiased estimator of *p*.

2. If $X_i$are normally distributed random variables with mean μ and varianceσ^2, then:

$$\hat{\mu} = \frac{\sum_{i=1}^{n} X_i}{n} = \bar{X} \text{ and } \hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n}$$

are the maximum likelihood estimators of μ and σ^2, respectively. Are the MLEs unbiased for their respective parameters?

Answer:

Recall that if $X_i$is a normally distributed random variable with mean μ and variance σ^2, then $E(X_i) = \mu$and$V(X_i) = \sigma^2$. Therefore:

$$E(\bar{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)$$

$$E(\bar{X}) = \frac{1}{n}\sum_{i=1}^{n} E(X_i)$$

$$E(\bar{X}) = \frac{1}{n}\sum_{i=1}^{n} \mu$$

$$E(\bar{X}) = \frac{1}{n}(n\mu) = \mu$$

The first equality holds because we've merely replaced $\bar{X}$ with its definition. Again, the second equality holds by the rules of expectation for a linear combination. The third equality holds because $E(X_i) = \mu$. The fourth equality holds because when you add the value μ up n times, you get $n\mu$. And, of course, the last equality is simple algebra.

In summary, we have shown that:

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

Therefore, the maximum likelihood estimator of μ is unbiased. Now, let's check the maximum likelihood estimator ofσ^2. First, note that we can rewrite the formula for the MLE as:

$$\hat{\sigma}^2 = \left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\right) - \bar{X}^{\,2}$$

Because

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

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i^2 - 2x_i\bar{x} + \bar{x}^2)$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - 2\bar{x}\frac{1}{n}\sum_{i=1}^{n} x_i + \frac{1}{n}(n\bar{x}^2)$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2$$

Then, taking the expectation of the MLE, we get:

$$E(\hat{\sigma}^2) = \frac{(n-1)\sigma^2}{n}$$

as illustrated here:

$$E(\hat{\sigma}^2) = E\left[\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2\right]$$

$$E(\hat{\sigma}^2) = \frac{1}{n}\sum_{i=1}^{n} E(x_i^2) - E(\bar{x}^2)$$

$$E(\hat{\sigma}^2) = \frac{1}{n}\sum_{i=1}^{n}(\sigma^2 + \mu^2) - \left(\frac{\sigma^2}{n} + \mu^2\right)$$

$$E(\hat{\sigma}^2) = \sigma^2 - \frac{\sigma^2}{n} = \frac{(n-1)\sigma^2}{n}$$

The first equality holds from the rewritten form of the MLE. The second equality holds from the properties of expectation. The third equality holds from manipulating the alternative formulas for the variance, namely:

$$V(X) = \sigma^2 = E(X^2) - \mu^2 \text{ and } V(\bar{X}) = \frac{\sigma^2}{n} = E(\bar{X}^2) - \mu^2$$

The remaining equalities hold from simple algebraic manipulation. Now, because we have shown:

$$E(\hat{\sigma}^2) \neq \sigma^2$$

the maximum likelihood estimator of σ^2 is a biased estimator.

5.3 PROPERTIES OF UNBIASED ESTIMATORS:

An Unbiased estimator is a statistical estimator whose expected value is equal to the true parameter it's estimates. If $\hat{\theta}$ is an estimator for a population parameter θ then $\hat{\theta}$is unbiased if:

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

Below are the key properties of unbiased estimators.

1. Expected Value Equals the True Parameter
   - The most fundamental property of an unbiased estimator is that its expected value equals the parameter being estimated.
   - This ensures that, on average, the estimator does not systematically overestimate or underestimate the true value.
   - Example: The sample mean $\bar{X}$ is an unbiased estimator of the population mean μ, since:
   $$E(\bar{X}) = \mu$$
2. No Systematic Bias
   - An unbiased estimator does not introduce any systematic error in estimation
   - This means that if multiple samples are taken, the average of the estimates will be correct.
   - Example: If estimating the probability of success *p* in a binomial distribution using the sample proportion $\hat{p} = \frac{X}{n}$, we get:
   $$E(\hat{p}) = P$$
   Ensuring there is no bias in estimating *p*.

3. Unbiasedness Does not Guarantee Minimum Variance
   - An estimator can be unbiased but still have high variance, meaning it fluctuates significantly from sample to sample.
   - In such cases, a biased but lower-variance estimator might be preferable.
   - The best unbiased estimator is the one with the minimum variance (called the minimum variance unbiased estimator, MVUE).

4. Linear Unbiasedness
   - If $\hat{\theta}_1$ and $\hat{\theta}_2$ are two unbiased estimators of θ, then any linear combination of them is also unbiased.
   - That is, for constants a and b:
   $$E[a\hat{\theta}_1 + b\hat{\theta}_2] = aE[\hat{\theta}_1] + bE[\hat{\theta}_2] = a\theta + b\theta = (a + b)\theta$$
   Making it an unbiased estimator.

5. Consistency and Asymptotic Unbiasedness
   - A consistent estimator approaches the true parameter values as the size increases.
   - Some estimators are asymptotically unbiased, meaning they become unbiased as the sample size *n* grows.
   - Example: The maximum likelihood estimator (MLE) of variance σ^2, given by:

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

is biased, but its bias decreases as $n \to \infty$, making it asymptotically unbiased. Unbiased estimators are essential in statistics because they provide accurate estimates on average. However, unbiasedness alone does not ensure the best estimator – other properties like low variance and consistency must also be considered when choosing an estimator.

5.4 MINIMUM-VARIANCE UNBIASED ESTIMATOR (MVUE):

Ultimately, we would like to be able to argue that a given estimator is (or is not)optimal in some sense.Usually this is very difficult, but in certain cases it is possible to make precise statements about optimality. MVUE estimators are one class where this is sometimes the case.

The statistic “T” will be called a MVUE of $V(\theta)$. If it is unbiased and the smallest variance (for each θ) among all unbiased estimator of $V(\theta)$. Thus “T” is an MVUE for $V(\theta)$.

$$E_\theta(T) = V(\theta); \forall\theta \in \Omega \text{ and}$$
$$V_\theta(T) \le V_\theta(T'); \forall\theta \in \Omega$$

How to Find the MVUE:

To find the MVUE for a parameter, we typically follow these steps:

1. Find an Unbiased estimator of the parameter.
2. Check for Minimum Variance using the Cramer Rao Lower Bound (CRLB).
3. Use the Lehmann-Scheffe theorem, which states that the MVUE can be found using a sufficient and complete statistic.

Necessary and Sufficient condition for the estimator of MVUE:

Statement:

Let $X = (X_1, X_2, ..., X_n)$ be a random sample from $f(x,\theta):\theta \in \Omega$. Let $T(X)$ and $U(X)$ be unbiased estimator of $\varphi(\theta)$ and θ respectively then $T(X)$ is MVUE of $\varphi(\theta)$ iff $E_\theta(T.U) = 0 \forall\theta \in \Omega$.

Proof:

Necessary:

$$\text{Given } E[T(X)] = \varphi(\theta)$$
$$E[U(X)] = 0 \text{ and}$$
$$E_\theta(T.U) = 0$$

We prove by condition

Suppose $E_{\theta_0}(T_0.U_0) \ne 0$ for some $\theta_0 \in \Omega$ and $E_{\theta_0}(U_0) = 0$.

Let $T^* = T + \alpha U_0$ for some constant 'α'. If $P_{\theta_0}(U_0 = 0) = 1$ (Degenerate random variables)

then $E_{\theta_0}(U_0) = 0, E_{\theta_0}(U_0^2) = 0$ and $V_{\theta_0}(U_0) = E_{\theta_0}(U_0^2) - [E_{\theta_0}(U_0)]^2 = 0$

But $E_{\theta_0}(T_0.U_0) \ne 0 \Rightarrow U_0 \ne 0 \Rightarrow E_{\theta_0}(U_0) > 0$

Let $\alpha_0 = \frac{-E_{\theta_0}\left(T_0.U_0\right)}{E_{\theta_0}\left(U_0^2\right)}$ (by regression estimator) then

$$E_{\theta_0}\left(T^*\right)^2 = E_{\theta_0}\left(T+\alpha U_0\right)^2$$

$$E_{\theta_0}\left(T^*\right)^2 = E_{\theta_0}\left(T^2\right)+\alpha^2 E_{\theta_0}\left(U_0^2\right)+2\alpha E\left(TU_0\right)$$

$$E_{\theta_0}\left(T^*\right)^2 = E_{\theta_0}\left(T^2\right)+\frac{E_{\theta_0}^2\left(T_0.U_0\right)}{E_{\theta_0}\left(U_0^2\right)}-2\frac{E_{\theta_0}^2\left(T_0.U_0\right)}{E_{\theta_0}\left(U_0^2\right)}$$

$$E_{\theta_0}\left(T^*\right)^2 = E_{\theta_0}\left(T^2\right)-\frac{E_{\theta_0}^2\left(T_0.U_0\right)}{E_{\theta_0}\left(U_0^2\right)}$$

$$E_{\theta_0}\left(T^*\right)^2 < E_{\theta_0}\left(T^2\right)$$

$$V_{\theta_0}\left(T^*\right) < V_{\theta_0}\left(T\right)$$

$\Rightarrow T^*$ is having less variance than ‘T’ which is MVUE this is contradiction.

$\therefore$ Our assumption is wrong.

Hence $E_\theta\left(T.U\right)=0$

Sufficiency:

$$\text{Given } E_\theta\left(T.U\right)=0 \qquad (1)$$

Show that ‘T’ is MVUE of $\varphi(\theta)$.

Let T^* be another unbiased estimator of $\varphi(\theta)$.

$\Rightarrow E_\theta\left(T^*\right)=\varphi(\theta)$ but $E_\theta\left(T\right)=\varphi(\theta)$

Let $U = T^* - T$ then $E_\theta\left(U\right)=0$ (2)

Then from (1), we have

$$E_\theta\left[T\left(T^*-T\right)\right]=0$$

$$E_\theta\left(TT^*\right)-E_\theta\left(T^2\right)=0$$

$$E_\theta\left(T^2\right)=E_\theta\left(TT^*\right)$$

$$\leq\left[E_\theta\left(T^2\right)E_\theta\left(T^*\right)^2\right]^{1/2}$$

Using Cauchy Schwartz inequality,

$$\Rightarrow E_\theta\left(T^2\right)\leq\left[E_\theta\left(T^2\right)E_\theta\left(T^*\right)^2\right]^{1/2}$$

$$\Rightarrow E_\theta^2\left(T^2\right)\leq E_\theta\left(T^2\right)E_\theta\left(T^*\right)^2$$

$$E_\theta\left(T^2\right)\leq E_\theta\left(T^*\right)^2$$

$$V_\theta\left(T\right)\leq V_\theta\left(T^*\right)$$

$\therefore$ T is MVUE of $\varphi(\theta)$.

5.5 UNIFORMLY MINIMUM VARIANCE UNBIASED ESTIMATOR (UMVUE):

An estimator is a function of sample data used to estimate an unknown population parameter. Among unbiased estimators, the **Uniformly Minimum Variance Unbiased Estimator (UMVUE)** is the one with the **least variance** for all possible values of the parameter.

The UMVUE is an MVUE that holds for **all values** of θ, meaning that no other unbiased estimator has a lower variance for any possible value of the parameter.

Finding the UMVUE:

To find the UMVUE of a parameter θ\thetaθ, the following steps are generally followed:

1. **Find an Unbiased Estimator**: Identify an unbiased estimator of θ
2. **Check Sufficiency Using the Factorization Theorem**:
   Use the **Factorization Theorem** to determine if the statistic is sufficient. A statistic T(X) is **sufficient** for θ if the joint density $f\left(X_1, X_2, ..., X_n/\theta\right)$ can be factorized as:

$$f\left(X_1, X_2, ..., X_n/\theta\right) = g\left(T\left(X\right), \theta\right)h\left(X\right)$$

3. **Apply the Lehmann-Scheffé Theorem**:
   - If an unbiased estimator is a function of a complete, sufficient statistic, then it is the UMVUE.
   - The Lehmann-Scheffé theorem states that if $\hat{\theta}$ is an unbiased estimator based on a complete, sufficient statistic, then $\hat{\theta}$ is UMVUE.

5.6 PROPERTIES OF UMVUE:

- **Uniqueness**: If an estimator is UMVUE, it is unique.
- **Efficiency**: The UMVUE has the lowest variance among unbiased estimators.
- **Admissibility**: In decision theory, UMVUE is not necessarily admissible under all loss functions.

Example:

Let $X_1, X_2, ..., X_n$ be a random sample from an exponential distribution:

$$f\left(x/\lambda\right) = \lambda e^{-\lambda x}; x > 0, \lambda > 0$$

We want to estimate $\theta = 1/\lambda$.

Step 1:

The likelihood function for an exponential distribution is:

$$L\left(\lambda\right) = \prod_{i=1}^{n} \lambda e^{-\lambda X_i}$$

Taking the log-likelihood

$$\log L(\lambda) = n\log\lambda - \lambda\sum_{i=1}^{n} X_i$$

Differentiating w.r.t λ and solving $\frac{\partial}{\partial\lambda}\log L(\lambda) = 0$, we get:

$$\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} X_i}$$

Since $\theta = \frac{1}{\lambda}$, the MLE for θ is:

$$\hat{\theta} = \frac{1}{\hat{\lambda}} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Step 2: Find an unbiased estimator for θ

$$E\left[\frac{\sum_{i=1}^{n} X_i}{n}\right] = \frac{1}{n}E\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n}(n\theta) = \theta$$

Since $\hat{\theta}$ is biased (it underestimates θ), we apply a correction factor

$$\hat{\theta} = \frac{\sum_{i=1}^{n} X_i}{n-1}$$

Now checking its expectation

$$E(\hat{\theta}) = E\left(\frac{\sum_{i=1}^{n} X_i}{n-1}\right) = \frac{1}{n-1}E\left(\sum_{i=1}^{n} X_i\right) = \frac{n\theta}{n-1} = \theta$$

So, $\hat{\theta}$ is an unbiased estimator of θ.

Step 3: Check sufficiency and completeness:
The statistic $T = \sum X_i$ is sufficient for λ (Factorization Theorem)
Since the exponential family is part of the one-parameter exponential family, T is also complete.
Step 4: Apply Lehmann-Scheffé Theorem

Since $\hat{\theta} = \frac{\sum_{i=1}^{n} X_i}{n-1}$ is unbiased and a function of the sufficient and complete statistic T.

It is the UMVUE for $\theta = \frac{1}{\lambda}$.

5.7 KEY WORDS:

- **Estimator:** A rule or formula that provides an estimate of a parameter based on observed data.
- **Bias:** The difference between the expected value of an estimator and the true value of the parameter.
- **Variance:** A measure of the dispersion of the estimator's sampling distribution.
- **Sufficient Statistic:** A statistic that captures all necessary information from the data regarding the parameter.
- **Complete Statistic:** A statistic is complete if no non-zero function of it has an expected value of zero for all parameter values.

Understanding these concepts is essential for constructing optimal estimators in statistical inference, ensuring both unbiasedness and minimal variance.

5.8 SUMMARY:

Unbiased estimators are fundamental in statistical inference, ensuring accurate parameter estimation without systematic errors. Among these, MVUEs and UMVUEs are particularly valuable due to their efficiency and uniformly minimal variance properties for identifying and deriving these optimal estimators, enhancing the reliability and precision of statistical analyses.

5.9 SELF ASSESSMENT QUESTIONS:

1. What distinguishes an unbiased estimator from a biased one?
2. State and prove necessary and sufficient conditions for the existence of MVUE.
3. Show that the minimum variance unbiased estimator if it exists, is unique.
4. Provide an example of a UMVUE in a common statistical distribution.
5. Why is achieving uniformly minimum variance important in estimation?

5.10 SUGGESTED READINGS:

1. Goon, A. M., Gupta, M. K., and Dasgupta, B. (1989). An Outline of Statistical Theory- Vol.II, World Press, Calcutta.
2. Casella, G. & Berger, R.L. (2013). Statistical Inference, 2nd Edition. Cengage Learning
3. Kale, B. K. (1999). A First Course on Parametric Inference, Narosa Publishing House, New Delhi.
4. Lehman, E. L., and Cassella, G. (1998). Theory of Point Estimation, Second Edition, Springer, NY.

Dr. U. Ramkiran

LESSON- 6

MINIMUM VARIANCE UNBIASED ESTIMATORS

OBJECTIVES:

After studying this lesson, the student is able to:

- Learn the necessary regularity conditions required for the application of statistical inference methods, particularly for unbiased estimation.
- Understand the derivation and significance of the Cramér-Rao Lower Bound (CRLB) and the conditions under which equality holds, leading to the Minimum Variance Unbiased Estimator (MVUE).
- Solve estimation problems using the Cramér-Rao inequality to determine the efficiency of estimators and identify the MVUE.
- Study how the Rao-Blackwell theorem helps improve an unbiased estimator by conditioning on sufficient statistics to achieve better efficiency.
- Learn the conditions under which an estimator is the unique Uniformly Minimum Variance Unbiased Estimator (UMVUE) using the combination of sufficiency and unbiasedness principles.
- Understand the concept of MMSE estimators and their applications in reducing estimation error in various statistical problems.

STRUCTURE:

6.1 REGULARITY CONDITIONS:

For the Cramer-Rao inequality to hold, the following regularity conditions must be satisfied:

1. Differentiability: The likelihood function $f(X;\theta)$ must be differentiable with respect to θ.

2. Finite Fisher Information: The Fisher information *I(θ)* must be finite.
3. Exchange of Expectation and Differentiation: The expectation and differentiation operations must be interchangeable. i.e.,

$$E\left[\frac{\partial}{\partial\theta}\log f\left(X;\theta\right)\right]=0$$

These conditions ensure that the inequality holds properly.

6.2 CRAMER-RAO INEQUALITY AND CONDITIONS FOR EXISTENCE EQUALITY:

Theorem Statement:

Let $X_1, X_2, ..., X_n$ be a random sample from a probability distribution with probability density function $f\left(X;\theta\right)$, where θ is an unknown parameter. Let $T\left(X\right)$ be an unbiased estimator of θ then the variance of $T\left(X\right)$ satisfies the inequality:

$$V\left[T\left(X\right)\right]\geq\frac{1}{I\left(\theta\right)}$$

Where $I\left(\theta\right)$ is the Fisher information, given by:

$$E\left[\left(\frac{\partial}{\partial\theta}\ln f\left(X;\theta\right)\right)^2\right]$$

Proof:

Step 1: Define the Score Function

The score function $U\left(X\right)$ is defined as:

$$U\left(X\right)=\frac{\partial}{\partial\theta}\ln f\left(X;\theta\right)$$

By definition of Fisher information:

$$I\left(\theta\right)=E\left[U^2\left(X\right)\right]$$

Also, since $f\left(X;\theta\right)$ is a probability density function, we have:

$$E\left[U\left(X\right)\right]=E\left[\frac{\partial}{\partial\theta}\ln f\left(X;\theta\right)\right]=0$$

Step 2: Consider Covariance between Estimator and Score Function

Since $T\left(X\right)$ is an unbiased estimator of θ, we have:

$$E\left[T\left(X\right)\right]=\theta$$

Taking the derivative with respect to θ,

$$E\left[\frac{\partial}{\partial\theta}T\left(X\right)\right]=1$$

Now, consider the covariance between $T\left(X\right)$ and $U\left(X\right)$:

$$Cov\left(T\left(X\right),U\left(X\right)\right)=E\left[T\left(X\right)U\left(X\right)\right]-E\left[T\left(X\right)\right]E\left[U\left(X\right)\right].$$

Since $E\left[U\left(X\right)\right]=0$, this simplifies to:

$$Cov\big(T(X),U(X)\big)=E\big[T(X)U(X)\big].$$

Using the Cauchy-Schwarz inequality:

$$Var\big(T(X)\big)I(\theta)\geq\big(Cov\big(T(X),U(X)\big)\big)^2.$$

Step 3: Evaluate the Covariance Expression

From integration by parts, we can show:

$$E\big[T(X)U(X)\big]=1$$

Thus,

$$\big(Cov\big(T(X),U(X)\big)\big)^2=1$$

Step 4: Derive the Cramer-Rao Bound

From the inequality:

$$Var\big(T(X)\big)I(\theta)\geq 1$$

Rearranging,

$$Var\big(T(X)\big)\geq\frac{1}{I(\theta)}$$

Hence the proof.

KEY POINTS:

- The Cramér-Rao Bound gives a fundamental limit on the precision of unbiased estimators.
- If an estimator attains the bound, it is efficient**.**
- The Maximum Likelihood Estimator (MLE) often asymptotically achieves this bound.

6.3 PROBLEMS BASED ON MVUE FOR BY USING CRAMÉR-RAO INEQUALITY:

1. Find MVUE for binomial distribution by using Cramér-Rao inequality

Solution: The p.m.f of binomial distribution is

$$P(x)=\begin{cases}\binom{n}{x}p^x q^{n-x}; x=0,1,\ldots,n\\ 0 \qquad ;\ Otherwise\end{cases}$$

The Fisher information $I(p)$ is given by

$$I(p)=-E\left[\frac{\partial^2}{\partial p^2}\log P(X)\right]$$

First, compute the log-likelihood function:

$$L(p)=\log P(X)=\log\binom{n}{X}+X\log p+(n-X)\log(1-p)$$

Taking the first derivative

$$\frac{\partial}{\partial p}L(p)=\frac{X}{p}-\frac{n-X}{1-p}$$

Taking the second derivative

$$\frac{\partial^2}{\partial p^2}L(p)=\frac{X}{p^2}-\frac{n-X}{(1-p)^2}$$

Taking the expectation

$$E\left[\frac{\partial^2}{\partial p^2}L(p)\right]=-\left(\frac{np}{p^2}+\frac{n(1-p)}{(1-p)^2}\right)=-\frac{n}{p(1-p)}$$

Thus, the Fisher information is:

$$I(p)=\frac{n}{p(1-p)}$$

The variance of any unbiased estimator $\hat{p}$ must satisfy

$$V(\hat{p})\geq\frac{1}{I(p)}=\frac{p(1-p)}{n}$$

Cosider the sample proportion estimator.

The natural estimator for p is $\hat{p}=\frac{X}{n}$.

Compute its variance

$$V(\hat{p})=V\left(\frac{X}{n}\right)=\frac{1}{n^2}Var(X)$$

Since $V(X)=np(1-p)$ for a binomial distribution,

$$V(\hat{p})=\frac{np(1-p)}{n^2}=\frac{p(1-p)}{n}$$

This matches the Cramér-Rao Lower Bound, meaning that $\hat{p}=X/n$ achieves the bound, making it the MVUE.

2. Find MVUE for Poisson distribution by using Cramér-Rao inequality.

Solution:

Let $X \sim P(\lambda)$, where

- X is a Poisson distributed random variable with mean λ.
- The probability mass function (PMF) is:

$$P(X=x)=\frac{e^{-\lambda}\lambda^x}{x!};x=0,1,2,...$$

The Fisher information $I(\lambda)$ is given by

$$I(\lambda)=-E\left[\frac{\partial^2}{\partial p^2}\log P(X)\right]$$

First, compute the log-likelihood function:

$$L(\lambda)=\log P(X)=-\lambda+x\log\lambda-\log x!$$

Taking the first derivative

$$\frac{\partial}{\partial\lambda}L(\lambda)=-1+\frac{x}{\lambda}$$

Taking the second derivative

$$\frac{\partial^2}{\partial\lambda^2}L(\lambda) = -\frac{x}{\lambda^2}$$

Taking the expectation

$$E\left[\frac{\partial^2}{\partial\lambda^2}L(\lambda)\right] = -E\left(\frac{x}{\lambda^2}\right) = -\frac{\lambda}{\lambda^2} = -\frac{1}{\lambda}$$

Thus, the Fisher information is:

$$I(\lambda) = \frac{1}{\lambda}$$

The variance of any unbiased estimator $\hat{\lambda}$ must satisfy

$$V\left(\hat{\lambda}\right) \geq \frac{1}{I(\lambda)} = \lambda$$

Consider the sample mean estimator.
A natural estimator for λ is $\hat{\lambda} = X$.
Its variance is $V\left(\hat{\lambda}\right) = V(X) = \lambda$

Since $V\left(\hat{\lambda}\right)$ achieves the Cramér-Rao Lower Bound, $\hat{\lambda} = X$ is an efficient estimator.

Since X is also unbiased for λ, if follows that X is the MVUE.

3. Find MVUE for Normal distribution by using Cramér-Rao inequality.

Solution: To find the Minimum Variance Unbiased Estimator (MVUE) for parameters of parameters of a Normal distribution using the Cramer-Rao inequality, we analize in two cases:

- Estimating the mean μ when σ^2 is known.
- Estimating the mean σ^2 when μ is known.

Case 1: Estimating the mean μ when σ^2 is known

Let $X_1, X_2, ..., X_n$ be a random sample from $X_i \sim N\left(\mu, \sigma^2\right)$.
The probability density function is:

$$f(X_i;\mu) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(X_i-\mu)^2}{2\sigma^2}\right)$$

Our goal is to find the MVUE of μ.

Step 2: Compute the Fishers Information
The likelihood function for n observations is:

$$L(\mu) = \prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(X_i-\mu)^2}{2\sigma^2}\right)$$

Taking the log-likelihood

$$\log L(\mu) = -\frac{n}{2}\log\left(2\pi\sigma^2\right) - \sum_{i=1}^{n}\frac{\left(X_i-\mu\right)^2}{2\sigma^2}$$

First derivative

$$\frac{\partial}{\partial\mu}\log L(\mu) = \sum_{i=1}^{n}\frac{\left(X_i-\mu\right)}{\sigma^2}$$

Second derivative

$$\frac{\partial^2}{\partial\mu^2}\log L(\mu) = -\frac{n}{\sigma^2}$$

Taking expectation

$$I(\mu) = -E\left[\frac{\partial^2}{\partial\mu^2}\log L(\mu)\right] = \frac{n}{\sigma^2}$$

Step 3: Cramer-Rao Lower Bound (CRLB)
For any unbiased estimator $\hat{\mu}$,

$$V(\hat{\mu}) \geq \frac{1}{I(\mu)} = \frac{\sigma^2}{n}$$

Step 4: Consider the Sample Mean Estimator
A natural estimator μ is

$$\hat{\mu} = \overline{X} = \frac{1}{n}\sum_{i=1}^{n}X_i$$

And its variance is

$$V(\hat{\mu}) = \frac{1}{n^2}\sum_{i=1}^{n}Var(X_i) = \frac{n}{n^2} = \frac{\sigma^2}{n}$$

Since $Var(\hat{\mu})$ attains the CRLB, $\hat{\mu} = \overline{X}$ is the MVUE.

Thus, the MVUE for μ in a normal distribution (when σ^2 is known) is:

$$\hat{\mu} = \overline{X}$$

Case 2: Estimating the mean σ^2 when μ is known
Now, we assume μ is known and find the MVUE for σ^2.

Step 1: Compute the Fisher Information

The likelihood function depends on σ^2, so we differentiate:

$$\log L(\mu) = -\frac{n}{2}\log\left(2\pi\sigma^2\right) - \sum_{i=1}^{n}\frac{\left(X_i-\mu\right)^2}{2\sigma^2}$$

First derivative

$$\frac{\partial}{\partial\sigma^2}\log L\left(\sigma^2\right) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}\left(X_i-\mu\right)^2$$

Second derivative

$$\frac{\partial}{\partial\sigma^4}\log L\left(\sigma^2\right) = \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}\sum_{i=1}^{n}\left(X_i-\mu\right)^2$$

Taking expectation

$$I\left(\sigma^2\right)=\frac{n}{2\sigma^4}$$

Step 2: Cramer-Rao Lower Bound (CRLB)

For any unbiased estimate $\hat{\sigma}^2$,

$$V\left(\hat{\sigma}^2\right)\geq\frac{1}{I\left(\sigma^2\right)}=\frac{2\sigma^4}{n}$$

Step 3: Consider the sample variance Estimator

A common estimator for σ^2 is:

$$S^2=\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\mu\right)^2$$

Compute its expectation

$$E\left[S^2\right]=\sigma^2$$

Compute its variance

$$Var\left[S^2\right]=\frac{2\sigma^4}{n}$$

This attains the CRLB.

Thus, the MVUE for σ^2 in a normal distribution (when μ is known) is:

$$\hat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\mu\right)^2$$

For a normal distribution $X_i\sim N\left(\mu,\sigma^2\right)$

- If σ^2 is known, the MVUE for μ is:

$$\hat{\mu}=\overline{X}$$

- If μ is known, the MVUE for σ^2 is:

$$\hat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\mu\right)^2$$

6.4 RAO - BLACKWELL THEOREM:

Statement:

If $f(x,\theta)$ is a family of densities admitted to a sufficient statistic T for θ and T_1 is any unbiased estimator of θ then $E(T_1/T)=h(T)$ is also an unbiased estimator of θ and $V\{h(T)\}\leq V(T_1)$.

Proof:

Part-I:

Given T is a sufficient statistic for θ. T_1 is an unbiased estimator of θ.

$$E(T_1/T)=h(T)$$

Since T is sufficient for θ. Obviously $E(T_1/T)$ is independent of θ.

Therefore put $E(T_1/T)=h(T)$

Consider $E_T\big(h(T)\big) = E_T\{E(T_1/T)\}$

$= E(T_1)$ [since $E_y\left(\frac{X}{Y}\right) = E(X)$, T_1 is Unbiased estimate of θ]

$$E_T\big(h(T)\big) = \theta$$

$\therefore$ $h(t) = E(T_1/T)$ is unbiased estimator of θ.

Part-II:

To prove $V\{h(T)\} \leq V(T_1)$

$$V(T_1) = E[T_1 - \theta]^2$$

$$= E[(T_1 - h(T)) + (h(T) - \theta)]^2$$

$$= E\{(T_1 - h(T)\}^2 + E\{h(T) - \theta\}^2 + 2E\{(T_1 - h(T))(h(T) - \theta)\} \quad (1)$$

Consider $E\{(T_1 - h(T))(h(T) - \theta)\} = E_T\{E((T_1 - h(T))(h(T) - \theta)/T)\}$

$=E_T\{(h(T) - \theta)\}E\{T - h(t)/T\}$

$=0$ [since $E(T_1/T) = h(T)$] (2)

From (1) and (2)

$$V(T_1) = E[T_1 - \theta]^2 + V(h(T))$$

$$V(T_1) \geq V(h(T))$$

$\therefore$ Hence Proved.

6.5 LEHMANN-SCHEFFE THEOREM:

Statement:

Let $X = (X_1, X_2, \dots, X_n)$ be a random sample from $f(x, \theta)$: $\theta \in \Omega$. Let T(X) be sufficient and complete estimator of $\varphi(\theta)$. Let U(x) be an unbiased estimator of $\varphi(\theta)$. Let $\varphi(\theta) = E\left(\frac{U}{T} = t\right)$ then $\varphi(T)$ is unique UMVUE of $\varphi(\theta)$.

Proof:

By Rao- Black well theorem $\emptyset(T)$ is UMVUE of $\varphi(\theta)$.

Here, we proved that $\emptyset(T)$ is unique UMVUE of $\varphi(\theta)$.

Let $\emptyset_1(T)$, $\emptyset_2(T)$ be two UMVUE of $\varphi(\theta)$.

Here $E[\emptyset_1(T)] = E[\emptyset_2(T)] = \varphi(\theta)$

$E[\emptyset_1(T)] - E[\emptyset_2(T)] = 0$

$E[\emptyset_1(T) - \emptyset_2(T)] = 0$

$E[\tau(T)] = 0$

Since T is a complete estimator of $\varphi(\theta)$ then

$$\tau(T) = 0$$

$$\emptyset_1(T) - \emptyset_2(T) = 0$$

$$\emptyset_1(T) = \emptyset_2(T)$$

Therefore, UMVUE is unique.

6.6 CRITERION FOR MINIMUM MEAN SQUARE ERROR:

The **criterion for Minimum Mean Square Error (MMSE)** is a fundamental principle in estimation theory that seeks to minimize the expected squared difference between an estimator and the true parameter value.

Definition of MMSE criterion:

Given an estimator $\hat{\theta}$ for a parameter θ, the Mean square error (MSE) is defined as:

$$MSE(\hat{\theta}) = E\left[(\hat{\theta} - \theta)^2\right]$$

The MMSE estimator is the estimator $\hat{\theta}$ that minimizes this MSE.

Decomposition of MSE:

The MSE can be decomposed into bias and variance components:

$$MSE(\hat{\theta}) = Var(\hat{\theta}) + \left[Bias(\hat{\theta})\right]^2$$

Where

- $Var(\hat{\theta}) = E\left[\left(\hat{\theta} - E(\hat{\theta})\right)^2\right]$ is the variance of the estimator.
- $Bias(\hat{\theta}) = E(\hat{\theta}) - \theta$ is the bias of the estimator.

The MMSE estimator aims to balance bias and variance to achieve the smallest possible MSE.

Criterion for MMSE estimation:

To achieve Minimum mean square error, an estimate $\hat{\theta}$ should satisfy:

1. Unbiased Estimation (if possible):
   - If an unbiased estimator with minimum variance exists, it is preferred.
   - If no unbiased estimator with low variance is available, a slightly biased estimator with lower variance may be used.
2. Trade-off Between Bias and Variance:
   - A biased estimator with significantly lower variance may have a lower MSE than an unbiased estimator with high variance.
   - Regularization techniques (like Ride Regression in statistics) exploit this trade-off.
3. Bayesian MMSE Estimator (conditional Expectation):
   - In Bayesian estimation, the MMSE estimator is given by:

$$\hat{\theta}_{MMSE} = E[\theta / X]$$

- This means the posterior mean is the best estimator under the MMSE criterion.

Example: Finding Minimum Mean Square Error (MMSE) Estimator

Problem Statement:

Suppose we have a signal *X* corrupted by additive noise *N*, and we observe: $Y=X+N$

where:

- $X \sim N(3,4)$ (mean = 3, variance = 4),
- $N \sim N(0,1)$ (mean = 0, variance = 1),
- *X* and *N* are independent.

Find the **MMSE estimator** $\hat{X}_{MMSE}$ and the **minimum mean square error (MSE)**.

Step 1: Compute the MMSE Estimator

We know that the **MMSE estimator** is the conditional expectation:

$$\hat{X}_{MMSE} = E\left[X/Y\right]$$

For jointly Gaussian X and Y, the MMSE estimator is given by:

$$\hat{X}_{MMSE} = \mu_X + \frac{\sigma_X^2}{\sigma_X^2 + \sigma_N^2}\left(Y - \mu_X\right)$$

Substituting the given values:

- $\mu_X = 3$
- $\sigma_X^2 = 4$
- $\sigma_N^2 = 1$

$$\hat{X}_{MMSE} = 3 + \frac{4}{5}\left(Y - 3\right)$$

$$\hat{X}_{MMSE} = \frac{4}{5}Y + \frac{3}{5}*3$$

$$\hat{X}_{MMSE} = \frac{4}{5}Y + \frac{9}{5}$$

Step 2: *Compute the Minimum Mean Square Error (MSE)*

The MMSE is given by:

$$MSE = \frac{\sigma_X^2 \sigma_N^2}{\sigma_X^2 + \sigma_N^2}$$

$$MSE = \frac{(4)(1)}{4+1} = 0.8$$

Special Case: Linear MMSE (LMMSE) Criterion:

If the estimator is restricted to be linear, the LMMSE estimator is:

$$\hat{X}_{LMMSE} = aY + b$$

Where a and b are chosen to minimize MSE. The solution is:

$$a = \frac{Cov(X,Y)}{Var(Y)}$$

$$b = E(X) - aE(Y)$$

If X and Y are jointly Gaussian LMMSE=MMSE, meaning the best linear estimator is also the best overall estimator.

Example: Estimating a Normal mean known

Consider a random sample $X_1, X_2, ..., X_n$ from $N(\mu, \sigma^2)$, where μ is unknown and σ^2 is known.

Case 1: Unbiased Estimator (MVUE)

The sample mean: $\hat{\mu} = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

- This is unbiased because $E(\hat{\mu}) = \mu$
- The variance is $\frac{\sigma^2}{n}$
- If no better estimator exists, then $\hat{\mu}$ is OMSE.

Case 2: Biased Estimator with lower MSE

Suppose we use a shrinkage estimator

$$\hat{\mu}_\lambda = \lambda \overline{X}, 0 < \lambda < 1$$

- This introduces bias: $E(\hat{\mu}_\lambda) = \lambda\mu \neq \mu$.
- However, it reduces variance, making it a better choice when the sample size is small.

If reducing variance dominates the increase in bias, this biased estimator may have lower MSE than the unbiased estimator, making it OMSE.

Applications of MMSE:

- Signal Processing**:** Noise reduction and filtering.
- Machine Learning: Regression models, Bayesian learning.
- Communications**:** Channel estimation, equalization.
- Econometrics**:** Predictive modelling and forecasting.

The **Optimum Mean Square Error (OMSE) estimator** is the estimator that achieves the **lowest possible MSE**, balancing bias and variance. It may be:

- The **MVUE**, if an unbiased minimum variance estimator exists.

- A **biased estimator** with significantly reduced variance.

The **Bayesian posterior mean** in a Bayesian framework.

6.7 KEY TERMS:

- **Regularity Conditions –** Conditions ensuring unbiased and efficient estimation.
- **Cramér-Rao Inequality** – Provides a lower bound on the variance of an unbiased estimator.
- **MVUE (Minimum Variance Unbiased Estimator)** – Best unbiased estimator, often found using Cramér-Rao.
- **Rao-Blackwell Theorem** – Improves an unbiased estimator by conditioning on a sufficient statistic.
- **Lehmann-Scheffé Theorem** – Ensures MVUE exists if the unbiased estimator is based on a sufficient and complete statistic.
- **MMSE (Minimum Mean Square Error)** – Estimator minimizing the expected squared error.

6.8 SUMMARY:

The topics covered provide a fundamental understanding of statistical estimation theory, particularly focusing on unbiased estimators and efficiency measures. The Cramér-Rao bound sets theoretical limits, while the Rao-Blackwell and Lehmann-Scheffé theorems offer practical methods to attain optimal estimators. The MMSE criterion extends the discussion by considering both bias and variance, ensuring minimal estimation error in practical scenarios.

6.9 SELF ASSESSMENT QUESTIONS:

1. What are the regularity conditions required for the Cramér-Rao inequality to hold? Why are regularity conditions necessary in estimation theory?
2. Let $X_1, X_2, ..., X_n$ be a random sample from a point binomial distribution with parameter p. Find the cramer-Rao lower bound for the estimator of p.
3. State and prove Lehman- Scheffe theorem. Give the importance of this theorem.
4. State and prove necessary and sufficient conditions for the existence of MVUE.
5. Let $X_1, X_2, ..., X_n$ be a random sample from the distribution with p.d.f. $f_\theta(x) = \frac{1}{\beta - \alpha}, \alpha < x < \beta$. Where $\theta = (\alpha, \beta)$ and $0 < \alpha < \beta < \infty$. Obtain MVU estimator of $\alpha + \beta/2$ and β-α.
6. How do you find the minimum variance unbiased estimator for $e^{-\lambda}$, where λ is the mean parameter of poisson population based on a random sample of size 'n'.
7. State and prove Rao-Blackwell theorem.
8. Describe MMSE with an example.

6.10 SUGGESTED READINGS:

1. Kale, B. K. (1999). A First Course on Parametric Inference, Narosa Publishing House, New Delhi.
2. Lehman, E. L., and Cassella, G. (1998). Theory of Point Estimation, Second Edition, Springer, NY.
3. An Introduction to Probability and Statistics by V.K. Rohatgi and.K.Md.E.Saleh(2001).
4. Goon, A. M., Gupta, M. K., and Dasgupta, B. (1989). An Outline of Statistical Theory-Vol.II, World Press, Calcutta.
5. Casella, G. & Berger, R.L. (2013). Statistical Inference, 2nd Edition. Cengage Learning

Dr. U. Ramkiran

LESSON – 7
CONSISTENCY ESTIMATORS

OBJECTIVES:

After studying this lesson, the student is able to:

- Understand the Notion of Consistency
- Apply the Study properties and methods of statistical estimation theory
- Obtain Consistency of Common Estimators
- Evaluate consistent and asymptotic behavior of estimators

STRUCTURE:

7.1 INTRODUCTION:

In statistical inference, consistency is a key property of an estimator that ensures it produces increasingly accurate estimates of the true parameter as the sample size grows. Intuitively, an estimator is consistent if, given a sufficiently large sample, it converges in probability to the true parameter it is meant to estimate. Estimation plays a crucial role in statistical inference, where the objective is to derive meaningful conclusions about a population from a given sample. A good estimator should possess desirable properties such as unbiasedness, efficiency, and consistency. Among these, consistency is a fundamental property that ensures an estimator converges to the true parameter value as the sample size increases. This section delves into the concept of consistency, its types, conditions, and implications.

7.2 CONSISTENCY:

Definition:

An estimator $\hat{\theta}_n$ of a parameter θ is consistent if it converges in probability to the true parameter value as the sample size n increases. This means that as we collect more data, the estimator gets arbitrarily close to θ.

Mathematically, an estimator $\hat{\theta}_n$ is consistent if:

$$\hat{\theta}_n \xrightarrow{P} \theta \text{ as } n \to \infty$$

Which means that for any small $\in > 0$:

$$P\left(\left|\hat{\theta}_n - \theta\right| > \in\right) \to 0 \text{ as } n \to \infty$$

Remarks: 1. If $X_1, X_2, ..., X_n$ is a random sample from population with finite mean $E(X_i) = \mu < \infty$, then by Khinchine's weak law of large numbers (W.L.L.N), we have

$$\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{p} E(X_i) = \mu, as \text{ n} \to \infty$$

Hence, sample mean $\left(\overline{X_n}\right)$ is always a consistent estimator of the population mean (μ).

2. Obviously consistency is a property concerning the behaviour of an estimator for indefinitely large values of the sample size n, as $\text{n} \to \infty$. Nothing is regarded of its behaviour for finite n.

Moreover, if there exists a consistent estimator, say, T_n of $\gamma(\theta)$, then infinitely many such estimators can be constructed, e.g.,

$$T_n^{'} = \left(\frac{n-a}{n-b}\right)T_n = \left[\frac{1-a/n}{1-b/n}\right]T_n \to T_n \xrightarrow{p} \gamma(\theta), as \text{ n} \to \infty$$

And hence, for different values of a and b, $T_n^{'}$ is also consistent for $\gamma(\theta)$.

7.2.1 Types of Consistency:

1. Weak Consistency: if $\hat{\theta}_n$ converges to θ in probability:
   $$\lim_{n \to \infty} P\left(\left|\hat{\theta}_n - \theta\right| < \in\right) = 1$$
2. Strong Consistency: : if $\hat{\theta}_n$ converges to θ almost surely:

$$P\left(\lim_{n \to \infty} \hat{\theta}_n = \theta\right) = 1$$

Strong consistency is a stronger condition than weak consistency.

7.2.2 Conditions for Consistency:

An estimator $\hat{\theta}_n$ is consistent if the following conditions hold:

1. Unbiased (or Asymptotic Unbiasedness):
   $$\lim_{n \to \infty} E\left(\hat{\theta}_n\right) = \theta$$
   The expected value of the estimator should approach θ.
2. Variance Goes to Zero:
   $$\lim_{n \to \infty} Var\left(\hat{\theta}_n\right) = 0$$
   This ensures that the estimator does not fluctuate much as n increases.
3. Law of Large Numbers (LLN):

   Many consistent estimators follow from the LLN, which states that, the sample mean of a large number of independent and identically distributed (i.i.d) random variables converges to the true mean.

7.3 INVARIANCE PROPERTY OF CONSISTENT ESTIMATORS:

Theorem: If T_n is a consistent estimator of $\gamma(\theta)$ and $\psi\{\gamma(\theta)\}$ is a continuous function of $\gamma(\theta)$, then $\psi\{T_n\}$ is a consistent estimator of $\psi\{\gamma(\theta)\}$.

Proof: Since T_n is a consistent estimator of $\gamma(\theta)$, $T_n \xrightarrow{p} \gamma(\theta)$ *as* n $\to \infty$, i.e., for every $\varepsilon > 0, \eta > 0, \exists$ appositive integer $n \geq m(\varepsilon, \eta)$ such that

$$P\left\{\left|T_n - \gamma(\theta)\right| < \varepsilon\right\} > 1 - \eta, n \geq m \quad (1)$$

Since $\psi(\cdot)$ is a continuous function, for every $\in > 0$, however small, $\exists$ a positive number $\in_1$ such that $\left|\psi(T_n) - \psi\{\gamma(\theta)\}\right| < \varepsilon_1$, whenever $\left|T_n - \gamma(\theta)\right| < \varepsilon$, i.e.,

$$\left|T_n - \gamma(\theta)\right| < \varepsilon \Rightarrow \left|\psi(T_n) - \psi\{\gamma(\theta)\}\right| < \varepsilon_1 \quad (2)$$

For two events A and B, if $A \Rightarrow B$, then

$$A \subseteq B \Rightarrow P(A) \leq P(B) \text{ or } P(B) \geq P(A) \quad (3)$$

From (1) and (2), we get

$$P\left[\left|\psi(T_n) - \psi\{\gamma(\theta)\}\right| < \varepsilon_1\right] \geq P\left\{\left|T_n - \gamma(\theta)\right| < \varepsilon\right\}$$

$\Rightarrow P\left[\left|\psi\left(T_n\right)-\psi\left\{\gamma(\theta)\right\}\right|<\varepsilon_1\right]\geq 1-\eta;\forall n\geq m$ (using (1))

$\Rightarrow \psi\left(T_n\right)-\psi\left\{\gamma(\theta)\right\}$, as $n\to\infty$ or $\psi\left(T_n\right)$ is a consistent estimator of $\gamma(\theta)$.

7.3.1 Sufficient Condition for Consistency:

Theorem: Let $\{T_n\}$ be a sequence of estimators such that for all $\theta\in\Theta$. (i) $E_\theta\left(T_n\right)\to\gamma(\theta), n\to\infty$ and (ii) $V_\theta\left(T_n\right)\to 0, n\to\infty$. Then T_n is a consistent estimator of $\gamma(\theta)$.

Proof: We have to prove that T_n is a consistent estimator of $\gamma(\theta)$.

i.e., $T_n \xrightarrow{p} \gamma(\theta)$ *as* $\text{n}\to\infty$

i.e., $P\left\{\left|T_n-\gamma(\theta)\right|<\varepsilon\right\}>1-\eta, n\geq m\ (\varepsilon,\eta)$ (4)

where ε and η are arbitrarily small positive numbers and m is some large value of n.

Applying Chebychev's inequality to the statistics T_n, we get

$$P\left[\left|T_n-E_\theta\left(T_n\right)\right|\leq\delta\right]\geq 1-\frac{Var\left(T_n\right)}{\delta^2} \quad (5)$$

We have

$$\left|T_n-\gamma(\theta)\right|=\left|T_n-E\left(T_n\right)+E\left(T_n\right)-\gamma(\theta)\right|\leq\left|T_n-E_\theta\left(T_n\right)\right|+\left|E_\theta\left(T_n\right)-\gamma(\theta)\right| \quad (6)$$

Now $\left|T_n-E_\theta\left(T_n\right)\right|\leq\delta\Rightarrow\left|T_n-\gamma(\theta)\right|\leq\delta+\left|E_\theta\left(T_n\right)-\gamma(\theta)\right|$ (7)

Hence, on using (3) of Theorem Invariance Property of Consistent Estimators, we get

$$\begin{aligned} P\left\{\left|T_n-\gamma(\theta)\right|\leq\delta+\left|E_\theta\left(T_n\right)-\gamma(\theta)\right|\right\}&\geq P\geq P\left\{\left|T_n-E_\theta\left(T_n\right)\right|\leq\delta\right\} \\ &\geq 1-\frac{Var_\theta\left(T_n\right)}{\delta^2} \qquad [From\ 5] \end{aligned} \quad (8)$$

We are given : $E_\theta\left(T_n\right)-\gamma(\theta)\forall\theta\in\Theta\ as\ \text{n}\to\infty$

Hence, for every $\delta_1>0,\exists$ a positive integer $\text{n}\geq\text{n}_0\left(\delta_1\right)$ such that

$$\left|E_\theta\left(T_n\right)-\gamma(\theta)\right|\ \leq\delta_1\forall\text{n}\geq\text{n}_0\left(\delta_1\right) \quad (9)$$

Also $Var_\theta\left(T_n\right)\to 0\ as\ \text{n}\to\infty,(Given)\ \ \therefore\frac{Var_\theta\left(T_n\right)}{\delta^2}\leq\eta,\ \forall\text{n}\geq\text{n}_0'(\eta)$ (10)

Where η is arbitrarily small positive number.

Substituting from (9) and (10) in (8), we get

$$P\left[\left|T_n-\gamma(\theta)\right|\leq\delta+\delta_1\right]\geq 1-\eta;\ \text{n}\geq \text{m}(\delta_1,\eta)$$

$$\Rightarrow \quad P\left[\left|T_n-\gamma(\theta)\right|\leq\varepsilon\right]1-\eta;\ \text{n}\geq\text{m},$$

where $\text{m=ma}_0\text{x}\left(m_0,n_0^{'}\right)$ and $\varepsilon=\delta+\delta_1>0.$

$$\Rightarrow \quad T_n\xrightarrow{p}\gamma(\theta)\, as\ \text{n}\to\infty \qquad \text{(using (4))}$$

$\therefore$ $\quad \text{T}_n$ is a consistent estimator of $\gamma(\theta)$.

7.4 PROBLEMS ON CONSISTENT ESTIMATOR:

Example 1: (a) Prove that in sampling from a $N(\mu,\sigma^2)$ population, the sample mean is a consistent estimator of μ.

(b) Prove that for Cauchy's distribution not sample mean but sample median is a consistent estimator of the population mean.

Solution: In sampling from a $N(\mu,\sigma^2)$ population, the sample mean $\bar{x}$ is also normally distributed as $N(\mu,\sigma^2/n)$, i.e., $E(\bar{x})=\mu$ and $V(\bar{x})=\sigma^2/n$

Thus $as\ \text{n}\to\infty$, $E(\bar{x})=\mu$ and $V(\bar{x})=0.$

Hence, by Theorem - Sufficient Condition for consistency, $\bar{x}$ is a consistent estimator for μ.

(b) The Cauchy's population is given by the probability function

$$dF(x)=\frac{1}{\pi}\cdot\frac{dx}{1+(x-\mu)^2},-\infty\leq x\leq\infty$$

The mean of the distribution, if we conventionally agree to assume that it exists, is at $x=\mu$. If $\bar{x}$, the sample mean is taken as an estimator of μ, then the sampling distribution of $\bar{x}$ is given by:

$$dF(\bar{x})=\frac{1}{\pi}\cdot\frac{d\bar{x}}{1+(\bar{x}-\mu)^2},-\infty\leq\bar{x}\leq\infty \qquad (1)$$

Because in Cauchy's distribution, the distribution of $\bar{x}$ is same as the distribution of x.

Since in this case, the distribution of $\bar{x}$ is same as the distribution of any single sample observation, it does not increase in accuracy with increasing n. In other words

$E(\bar{x}) = \mu$ but $V(\bar{x}) = V(x) \neq 0, \text{as } n \to \infty$

Hence, by Theorem - Sufficient Condition for consistency $\bar{x}$ is not a consistent estimator of μ in this case.

Consideration of symmetric of (1) is enough to show that the sample median Md is an unbiased estimate of the population mean, which of course is same as the population median.

Therefore $E(Md) = \mu$ (2)

For large n, the sampling distribution of median is asymptotically normal and is given by

$$dF \propto \exp\left\{-2nf_1^2 (x-\mu)^2\right\} dx,$$

where f_1 is the median ordinate of the parent population. i.e.,

$$dF \propto \exp\left\{-\frac{(x-\mu)^2}{1/2nf_1^2}\right\} \tag{3}$$

But f_1=Median ordinate of (1)=Model ordinate of (1) [Because of symmetry]

$$\left[f(x)\right]_{x=\mu} = \frac{1}{\pi}$$

Hence, from (3), the variance of the sampling distribution of median is:

$$V(Md) = \frac{1}{4nf_1^2} = \frac{1}{4n(1/\pi)^2} = \frac{\pi^2}{4n} \to 0 \text{ as n} \to \infty \tag{4}$$

Hence form (2) and (4), using Theorem - Sufficient Condition for consistency, we conclude that for in Cauchy's distribution, median is a consistent estimator for μ.

Example 2: If $X_1, X_2, ..., X_n$ are random observations on a Bernoulli variate X taking the value 1 with probability p and the value 0 with probability $(1-p)$, show that: $\frac{\sum x_i}{n}\left(1-\frac{\sum x_i}{n}\right)$ is a consistent estimator of $p(1-p)$.

Solution: Since $X_1, X_2, ..., X_n$ are i.i.d Bernoulli variates with parameter '*p*'.

$$T = \sum_{i=1}^{n} x_i \sim B(n,p) \Rightarrow E(T) = np \text{ and } \operatorname{var}(T) = npq \tag{1}$$

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{T}{n} \Rightarrow E(\bar{X}) = \frac{1}{n}E(T) = \frac{1}{n}\cdot np = p$$ [From (1)]

and $Var(\bar{X}) = Var\left(\frac{T}{n}\right) = \frac{1}{n^2} \cdot Var(T) = \frac{pq}{n}$ as $n \to \infty$. [From (1)]

Since $E(\bar{X}) \to p$ and $V(\bar{X}) \to 0$, as $n \to \infty$; $\bar{X}$ is a consistent estimator of *p*. Also $\frac{\sum x_i}{n}\left(1 - \frac{\sum x_i}{n}\right) = \bar{X}(1-\bar{X})$, being a polynomial in $\bar{X}$, is a continuous function of $\bar{X}$.

Since $\bar{X}$ is consistent estimator of *p*, by the invariance property of consistent estimators. $\bar{X}(1-\bar{X})$ is a consistent estimator of $p(1-p)$.

Note:

Suppose we have a random sample $X_1, X_2, ..., X_n$ from a distribution with mean μ and variance σ^2. The sample mean

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

Is a consistent estimator of μ, because:

- $E[\bar{X}_n] = \mu$ (unbiased)
- $Var[\bar{X}_n] = \frac{\sigma^2}{n} \to 0$ as $n \to \infty$
- By the Law of Large Numbers (LLN), $\bar{X}_n \xrightarrow{P} \mu$.

Thus, as the sample size increases, the sample mean gets closer to the true mean.

7.5 INCONSISTENCY:

Definition:

An estimator $\hat{\theta}_n$ of a parameter θ is said to be inconsistent if it does not converge to θ in probability as the sample size $n \to \infty$.

Mathematically, an estimator $\hat{\theta}_n$ is consistent if :

$$\lim_{n \to \infty} P\left(\left|\hat{\theta}_n - \theta\right| > \in\right) \neq 0 \text{ for some } \in > 0$$

This means that, even with large sample sizes, the estimator does not get arbitrarily close to the true parameter θ.

7.5.1 Causes of Inconsistency:

An estimator can be inconsistent for several reasons:

1. **Biased Estimator with Nonzero Limit:**
   - If an estimator has a bias that does not vanish as $n \to \infty$, it may not converge to θ.
   - Example: Estimating a parameter using a constant value (e.g., always using 5 to estimate the mean of a distribution).
2. **Increasing Variance:**
   - If the variance of $\hat{\theta}_n$ does not shrink as n grows, the estimator may fluctuate too much to converge.
   - Example: If an estimator has variance that increases with n, it will not stabilize around θ.
3. **Wrong Model Assumption:**
   - If the model used for estimation is incorrect, the estimator might not capture the true parameter.
   - Example: Using a Poisson model to estimate parameters of a normal distribution will give inconsistent estimates.
4. **Dependence or Non-Identifiability:**
   - If observations are not independent or the parameter is not well-defined, estimation may fail.
   - Example: In regression, if there is perfect multicollinearity, estimates of regression coefficients can be inconsistent.

7.5.2 How to Address Inconsistency

- **Use correct model specification**: Ensure functional form and variable selection are appropriate.
- **Handle endogeneity**: Use Instrumental Variables (IV) or Generalized Method of Moments (GMM).
- **Correct for measurement errors**: Use latent variable models or reliability correction methods.
- **Account for heteroskedasticity**: Use robust standard errors or weighted least squares.
- **Ensure identifiability**: Avoid perfect multicollinearity by ensuring sufficient variation in data.
- **Check for stationarity in time series**: Use tests like ADF and apply transformations if needed.
- **Avoid selection bias**: Use techniques like Heckman's correction or randomized sampling.

7.5.3 Implications of Inconsistency

- Estimates remain biased even with large sample sizes.
- Standard inference methods (e.g., confidence intervals, hypothesis testing) become unreliable.
- Results cannot be used for meaningful predictions or decision-making.

7.6 KEY TERMS:

- **Estimator**: A rule for calculating an estimate of a given quantity.
- **Consistency**: Property of an estimator to converge to the true parameter.

- **Weak Consistency:** An estimator $\hat{\theta}_n$ is weakly consistent if it converges in probability θ.
- **Strong Consistency:** An estimator is strongly consistent if it converges almost surely θ, meaning : $P \lim_{n \to \infty} E\left(\hat{\theta}_n\right) = \theta$
- **Asymptotic Bias**: The bias that remains even as the sample size grows.

7.7 SUMMARY:

Consistency is a fundamental property of an estimator, ensuring its reliability as sample size increases. Understanding its conditions and implications is crucial in estimation theory. While inconsistency can arise due to various factors, it can often be mitigated through proper model selection, unbiased estimators, and careful sampling techniques.

7.8 SELF ASSESSMENT QUESTIONS:

1. What is the definition of a consistent estimator?
2. Explain the difference between weak and strong consistency.
3. What are the key conditions for an estimator to be consistent?
4. State and explain the invariance property of consistent estimators.
5. Provide an example of a consistent estimator and justify why it is consistent.
6. Explain about consistency criterion of estimation in large samples.
7. Define (i) Consistency (ii) relative efficiency and (iii) Consistent Estimators with suitable examples.
8. If $\hat{\theta}_n$ is a consistent estimator of θ, prove that $g\left(\hat{\theta}_n\right)$ is a consistent estimator of $g(\theta)$, Where g is a continuous function.

7.9 SUGGESTED READINGS:

1. Kale, B. K. (1999). A First Course on Parametric Inference, Narosa Publishing House, New Delhi.
2. Lehman, E. L., and Cassella, G. (1998). Theory of Point Estimation, Second Edition, Springer, NY.
3. An Introduction to Probability and Statistics by V.K. Rohatgi and.K.Md.E.Saleh(2001).
4. Goon, A. M., Gupta, M. K., and Dasgupta, B. (1989). An Outline of Statistical Theory-Vol.II, World Press, Calcutta.
5. Dudewicz, E. J., and Mishra, S. N. (1988). Modern Mathematical Statistics, John Wiley & Sons, NY.
6. Rajagopalan, M., and Dhanavanthan, P.(2012). Statistical Inference, PHI Learning Pvt., Ltd., New Delhi.
7. Linear Statistical Inference and its Application by C.R. RAO (1973), John Wiley.

Dr. U. Ramkiran

LESSON- 8
CAN AND CUAN ESTIMATORS

OBJECTIVES:

After studying this lesson, the student is able to:

- Understand Efficiency in Estimation and its role in statistical inference.
- Analyze Asymptotic Relative Efficiency (ARE)
- Apply ARE to compare the performance of estimators in large samples.
- Explore Consistent and Asymptotically Normal (CAN) Estimators and their properties.
- Analyze consistency and asymptotic normality of estimators.
- Evaluate asymptotic distributions for common statistical estimators.
- Identify conditions for an estimator to be CAUN.
- Evaluate practical applications of CAUN estimators in statistical inference.
- Observe consistent and asymptotic behavior of BAN estimators

STRUCTURE:

8.1 INTRODUCTION:

Statistical inference is the process of drawing conclusions about a population based on data from a sample. It involves estimating parameters, testing hypotheses, and making predictions. The efficiency of an estimator is a crucial concept in statistical inference, helping determine how well an estimator performs relative to others.

8.2 EFFICIENCY:

Definition:

If T_1 is the most efficient estimator with variance V_1 and T_2 is any other estimator with variance V_2, then the efficiency E of T_2 is defined as:

$$E = \frac{V_1}{V_2}$$

Obviously E cannot exceed unity.

8.3 RELATIVE EFFICIENCY:

- The efficiency of an estimator is often compared to another estimator or a benchmark estimator.
- **Relative Efficiency** between two estimators $\hat{\theta}_1$ and $\hat{\theta}_2$.

$$RE = \frac{MSE\left(\hat{\theta}_2\right)}{MSE\left(\hat{\theta}_1\right)}$$

where Mean Squared Error (MSE) is:

$$MSE\left(\hat{\theta}\right) = Var\left(\hat{\theta}\right) + \left[Bias\left(\hat{\theta}\right)\right]^2$$

8.3.1 Efficiency and the Cramér-Rao Lower Bound (CRLB):

- The Cramér-Rao Lower Bound (CRLB) sets a theoretical lower bound on the variance of any unbiased estimator.
- An estimator achieving this bound is called efficient or minimum variance unbiased estimator (MVUE).
- Mathematically, if $I(\theta)$ is the Fisher Information, the lower bound for an unbiased estimator $\hat{\theta}$ is:

8.3.2 Asymptotic Efficiency:

- For large samples, the efficiency of an estimator is defined as the limit:

$$e\left(\hat{\theta}\right) = \underset{n\to\infty}{Lim}\frac{Var\left(Bestestimator\right)}{Var\left(\hat{\theta}\right)}$$

- **Maximum Likelihood Estimators (MLEs)** are often asymptotically efficient, meaning they attain the CRLB as sample size increases.

8.3.3 Trade-Off Between Efficiency and Other Properties:

- **Efficiency vs. Bias**: Sometimes a biased estimator (like ridge regression) has lower variance than an unbiased estimator, leading to a better overall MSE.
- **Efficiency vs. Robustness**: Highly efficient estimators may be sensitive to outliers, whereas more robust estimators (e.g., median instead of mean) may sacrifice some efficiency.

8.4 EFFICIENCY IN DIFFERENT ESTIMATORS:

Let's consider specific examples of efficiency by comparing different estimators.

Example 1: Sample Mean vs. Sample Median (Estimating μ of a Normal Distribution)

- Suppose we want to estimate the population mean μ of a normal distribution $N\left(\mu,\sigma^2\right)$.
- Two common estimators:
  1. **Sample Mean**: $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
  2. **Sample Median**: The median of the X_i's.
- **Efficiency Comparison**:
  - The variance of $\bar{X}$ is $\frac{\sigma^2}{n}$.
  - The variance of the **median** is approximately $\frac{1.57\sigma^2}{n}$
  - The relative efficiency is:

$$\frac{Var(median)}{Var(\bar{X})} = \frac{1.57\sigma^2/n}{\sigma^2/n} = 1.57$$

  - This means the sample mean is **more efficient** than the median for normal data.
  - However, for heavy-tailed distributions (e.g., Cauchy), the **median is more robust and has lower variance than the mean**.

Example 2: Estimating Variance σ^2

- Suppose we have a sample from $N\left(\mu,\sigma^2\right)$, and we estimate variance using:
    1. **Unbiased Estimator** (sample variance):

$$S^2 = \frac{1}{n-1}\sum\left(X_i - \bar{X}\right)^2$$

    2. **Maximum Likelihood Estimator (MLE)**:

$$S^2_{MLE} = \frac{1}{n}\sum\left(X_i - \bar{X}\right)^2$$

- **Efficiency Comparison**:
    - S^2 is **unbiased**, but has slightly higher variance.
    - S^2_{MLE} is **biased**, but has lower variance, making it more efficient in some cases.

8.4.1 Efficiency in Large Samples (Asymptotic Efficiency):

- As sample size $n \to \infty$, some estimators become asymptotically efficient**,** meaning their variance approaches the Cramér-Rao Lower Bound.
- MLE (Maximum Likelihood Estimators) are typically asymptotically efficient**.**
- For example, for a Poisson-distributed sample with mean λ, the MLE $\hat{\lambda} = \bar{X}$ is asymptotically efficient.

8.4.2 Problems on Efficiency:

Example 1: A random sample $\left(X_1, X_2, X_3, X_4, X_5\right)$ of size 5 is drawn from a normal population with known mean μ. Consider the following estimators to estimate μ:

(i) $t_1 = \dfrac{X_1 + X_2 + X_3 + X_4 + X_5}{5}$, (ii) $t_2 = \dfrac{X_1 + X_2}{2} + X_3$, (iii) $t_3 = \dfrac{2X_1 + X_2 + \lambda X_3}{3}$,

Where λ is such that t_3 is an unbiased estimator of μ. Find λ. Are t_1 and t_2 unbiased? State giving reasons, the estimator which is best among t_1, t_2 and t_3.

Solution: We are given

$$E\left(X_i\right) = \mu, Var\left(X_i\right) = \sigma^2 \text{(say)}; Cov\left(X_i, X_j\right) = 0 \left(i = 1, 2, ..., n\right) \qquad (1)$$

(i) $E\left(t_1\right) = \frac{1}{5}\sum_{i=1}^{5} E\left(X_i\right) = \frac{1}{5}\sum_{i=1}^{5}\mu = \frac{1}{5}.5\mu = \mu \Rightarrow t_1$ is an unbiased estimator of μ.

(ii) $E\left(t_2\right) = \frac{1}{2}E\left(X_1 + X_2\right) + E\left(X_3\right) = \frac{1}{2}\left(\mu + \mu\right) + \mu = 2\mu$

$\Rightarrow t_2$ is not an unbiased estimator of μ.

(iii) $E(t_3)=\mu \Rightarrow \frac{1}{3}E(2X_1+X_2+\lambda X_3)=\mu$

$(\because t_3 \text{ is unbiased estimator of } \mu)$

$\therefore\ 2E(X_1)+E(X_2)+\lambda E(X_3)=3\mu$

$\therefore\ 2\mu+\mu+\lambda\mu=3\mu \Rightarrow \lambda=0$

Using(1), we get

$$V(t_1)=\frac{1}{25}\{V(X_1)+V(X_2)+V(X_3)+V(X_4)+V(X_5)\}=\frac{1}{5}\sigma^2$$

$$V(t_2)=\frac{1}{4}\{V(X_1)+V(X_2)\}+V(X_3)=\frac{1}{2}\sigma^2+\sigma^2=\frac{3}{2}\sigma^2$$

$$V(t_3)=\frac{1}{9}\{4V(X_1)+V(X_2)\}=\frac{1}{9}(4\sigma^2+\sigma^2)=\frac{5}{9}\sigma^2 \qquad (\because \lambda=0)$$

Since $V(t_1)$ is least t_1 is the best estimator (in the sense of least variance) of μ.

Example 2: $X_1, X_2,$ and X_3 is a random sample of size 3 from a population with mean value μ and variance σ^2. T_1, T_2, T_3 are the estimators used to estimate mean value μ, where

$T_1 = X_1 + X_2 - X_3,\ T_2 = 2X_1 + 3X_3 - 4X_2,$ and $T_3 = \frac{1}{3}(\lambda X_1 + X_2 + X_3)/3$

$E(t_1)=\frac{1}{5}\sum_{i=1}^{5}E(X_i)=\frac{1}{5}\sum_{i=1}^{5}\mu=\frac{1}{5}.5\mu=\mu \Rightarrow t_1$ is an unbiased estimator of μ.

(i) Are T_1 and T_2 unbiased estimators?

(ii) Find the value of λ such that T_3 is unbiased estimator for μ.

(iii) With this value of λ is T_3 a consistent estimator?

(iv) Which is the best estimator?

Solution: Since X_1, X_2, X_3 is a random sample from a population with mean μ and variance σ^2, $E(X_i)=\mu,\ \text{Var}(X_i)=\sigma^2$ and $\text{Cov}(X_i, X_j)=0,\ (i=1,2,.....,n)$ (1)

(i) We have [on using (1)],

$E(T_1)=E(X_1)+E(X_2)-E(X_3)=\mu \Rightarrow T_1$ is an unbiased estimator of μ

$E(T_2)=2E(X_1)+3E(X_3)-4E(X_2)=\mu \Rightarrow T_2$ is an unbiased estimator of μ

(ii) We are given : $E(T_3)=\mu \Rightarrow \frac{1}{3}\{\lambda E(X_1)+E(X_2)+E(X_3)\}=\mu$

$\frac{1}{3}(\lambda\mu+\mu+\mu)=\mu \Rightarrow \lambda+2=3 \Rightarrow \lambda=1$.

(iii) With $\lambda=1,\ T_3=\frac{1}{3}(X_1+X_2+X_3)=\overline{X}$. Since sample mean is a consistent estimator of population mean μ, by Weak Law of Large Numbers, T_3 is a consistent estimator of μ.

(iv) We have [on using (1)],

$Var(T_1)=Var(X_1)+Var(X_2)+Var(X_3)=3\sigma^2$

$Var(T_2)=4\ Var(X_1)+9\ Var(X_3)+16Var(X_2)=29\sigma^2$

$$Var(T_3) = \frac{1}{9}\left[Var(X_1) + Var(X_2) + Var(X_3)\right] = \frac{1}{3}\sigma^2 \qquad (\therefore \lambda = 1)$$

Since $Var(T_3)$ is minimum, T_3 is the best estimator of μ in the sense of minimum variance.

8.5 ASYMPTOTICALLY EFFICIENT ESTIMATORS:

Definition:

An estimator $\hat{\theta}_n$ is said to be asymptotically efficient if it achieves the lowest possible asymptotic variance among all consistent estimators of θ.

This means that as the sample size $n \to \infty$, the estimator becomes optimal in terms of variance, approaching the Cramer -Rao Lower Bound (CRLB).

Mathematically, an estimator $\hat{\theta}_n$ is asymptotically efficient if its asymptotic variance equals the inverse of the Fisher information:

$$\lim_{n\to\infty} n\,Var(\hat{\theta}_n) = \frac{1}{I(\theta)}$$

Where $I(\theta)$ is the Fisher Information.

Key Properties of Asymptotically Efficient Estimators:

- Consistency: The estimator $\hat{\theta}_n$ must converge in probability to θ.

$$\hat{\theta}_n \xrightarrow{p} \theta$$

- Asymptotic Normality: The estimator follows a normal distribution asymptotically:

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, I(\theta)^{-1})$$

- Achieves Cramer-Rao Lower Bound (CRLB): Asymptotically efficient estimators attain the minimum possible variance for large *n*.

8.5.1 Examples of Asymptotically Efficient Estimators:

1. **Maximum Likelihood Estimator (MLE)**: Under regularity conditions, the MLE is asymptotically efficient.
2. **Least Squares Estimator (LSE) in Linear Regression**: Efficient when errors are normally distributed.
3. **Bayesian Estimators**: Certain Bayesian estimators achieve asymptotic efficiency under regularity conditions.

Example Maximum Likelihood estimation:

MLE as an Asymptotically Efficient Estimator

Maximum Likelihood Estimators (MLEs) are often asymptotically efficient under regularity conditions.

For example, suppose $X_1, X_2, ..., X_n \sim N(\mu, \sigma^2)$ where σ^2 is known. The MLE for μ is:

$$\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- $\hat{\mu}_n$ is consistent because $E(\hat{\mu}_n) = \mu$.
- It follows an asymptotic normal distribution:

$$\sqrt{n}(\hat{\mu}_n - \mu) \xrightarrow{d} N(0, \sigma^2)$$

- The Fisher Information for μ is $I(\mu) = \frac{1}{\sigma^2}$, so the CRLB is $\frac{1}{I(\mu)} = \sigma^2$.
- Since the variance of $\hat{\mu}_n$ achieves this bound, it is asymptotically efficient.

8.5.2 Importance of Asymptotic Efficiency:

- **Optimality in Large Samples:** Ensures the best possible variance performance for large *n*.
- **MLEs are Often Asymptotically Efficient:** Many practical estimation methods (e.g., logistic regression, normal mean estimation) rely on this property.
- **Guides Estimator Selection:** If multiple estimators are available, choosing an asymptotically efficient one ensures optimal long-run behaviour.

8.6 CAN ESTIMATORS:

In estimation theory, we often deal with a parameter θ that we wish to estimate based on a sample $X = (X_1, X_2, ..., X_n)$. An estimator $\hat{\theta}_n$ is a function of the sample that provides an estimate of θ.

Two important properties of estimators in large samples are:

- **Consistency** (ensuring that the estimator converges to the true parameter value).
- **Asymptotic Normality** (ensuring that the estimator follows a normal distribution in large samples).

These properties define two key classes of estimators: **CAN (**Consistent and Asymptotically Normal) estimators and **CAUN (**Consistent, Asymptotically Unbiased, and Normal) estimators.

8.6.1 Can Estimators (Consistent and Asymptotically Normal):

Definition:

An estimator sequence $\hat{\theta}_n$ is said to be **CAN** if it satisfies:

(i) **Consistency:**

$$\hat{\theta}_n \xrightarrow{P} \theta \text{ (Convergence in Probability)}$$

This means that for $\in > 0$:

$$P\left(\left|\hat{\theta}_n - \theta\right| > \in\right) \to 0 \text{ as } n \to \infty$$

This ensures that as the sample size increases, the estimator $\hat{\theta}_n$ gets arbitrarily close to the true parameter θ.

(ii) Asymptotic Normality:

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N\big(0, V(\theta)\big) \text{ (convergence in distribution)}$$

Where $V(\theta)$ is some variance function depending on θ.

This means that $n \to \infty$, the scaled difference $\sqrt{n}(\hat{\theta}_n - \theta)$ follows an asymptotic normal distribution with mean 0 and variance $V(\theta)$.

Interpretation:

- Consistency ensures that the estimator gives accurate results as *n* grows.
- Asymptotic normality allows us to construct confidence intervals and perform hypothesis testing.

8.6.2 Example: Maximum Likelihood Estimator (MLE)

Under regularity conditions, the Maximum Likelihood Estimator (MLE) is a CAN estimator.

Proof:

Let $L(\theta; X_1, X_2, ..., X_n)$ be the log-likelihood function. The MLE $\hat{\theta}_n$ maximizes $L(\theta)$, so it satisfies the first order condition:

$$\left|\frac{\partial L}{\partial \theta}\right|_{\theta=\hat{\theta}_n} = 0$$

Using central limit theorem and Taylor series expansion, we obtain

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, I(\theta)^{-1}),$$

where $I(\theta)$ is the Fisher Information. This proves that MLE is **CAN.**

2. Explanation with an Example

Example: Sample Mean as a CAN Estimator

Consider a random sample $X_1, X_2, ..., X_n$ from a population with mean μ and variance σ^2. The sample mean is:

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

Step 1: Check Consistency

By the law of Large Numbers (LLN):

$$\bar{X}_n \xrightarrow{P} \mu$$

This shows that $\bar{X}_n$ is a consistent estimator of μ.

Step 2: Check Asymptotic Normality

By the Central Limit Theorem (CLT):

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} N(0, \sigma^2)$$

Thus, $\bar{X}_n$ is an asymptotically normal estimator of μ, making it a CAN estimator.

8.6.3 Importance of Can Estimators:

- CAN estimators are useful in large-sample inference because they provide approximate normality, allowing us to construct confidence intervals and hypothesis tests.
- Many common estimators, such as Maximum Likelihood Estimators (MLEs), are CAN under regularity conditions.
- The asymptotic normality property allows us to approximate the distribution of an estimator, even when the exact finite-sample distribution is unknown.

Conclusion

A CAN estimator is an estimator that is both consistent (converges to the true value) and asymptotically normal (follows a normal distribution after scaling). The sample mean is a common example of a CAN estimator.

8.7 CAUN ESTIMATORS (CONSISTENT, ASYMPTOTICALLY UNBIASED AND NORMAL):

Definition:

An estimator $\hat{\theta}_n$ is said to be CAUN if it satisfies:

1. Consistency: Same as in CAN estimators.
2. Asymptotic Normality: Same as CAN estimators.
3. Asymptotic Unbiasedness

   $$E\left(\hat{\theta}_n\right) = \theta + o(1),$$

   Where $o(1)$ represents a term that approaches 0 as $n \to \infty$.

Interpretation:

- A **CAN estimator** may have some bias in large samples.
- A **CAUN estimator** ensures that this bias disappears asymptotically.

8.7.1 Example - Method of Moments Estimator (MME):

Under suitable conditions, a Method of Moments Estimator (MME) is a CAUN estimator.

Proof:

Suppose we have a random variable X with mean $E(X) = g(\theta)$. The method of moments estimator is defined as:

$$\hat{\theta}_n = g^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)$$

By the law of large numbers, $\frac{1}{n}\sum_{i=1}^{n} X_i$ converges in probability to $g(\theta)$, implying consistency. Using a Taylor expansion, it can be shown that:

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N\big(0, V(\theta)\big)$$

Proving asymptotic normality, under suitable conditions, the bias vanishes as $n \to \infty$, making $\hat{\theta}_n$ CAUN.

Comparison of CAN and CAUN Estimators:

Property	**CAN Estimator**	**CAUN Estimator**
Consistency	Yes	Yes
Asymptotic Normality	Yes	Yes
Asymptotic Unbiasedness	Not necessarily	Yes
Example	MLE	Method of Moments Estimator (MME)

Key Takeaways:

- A **CAN estimator** is not necessarily asymptotically unbiased, but it is consistent and asymptotically normal.
- A **CAUN estimator** is a CAN estimator that is also **asymptotically unbiased**.

Summary and Applications:

- **CAN estimators** (e.g., MLE) are widely used because they enable large-sample statistical inference.
- **CAUN estimators** provide an additional guarantee of asymptotic unbiasedness, which is useful when bias reduction is critical.
- These estimators are essential in **econometrics, machine learning, and signal processing** for estimating unknown parameters in statistical models.

8.8 BAN ESTIMATORS:

Definition:

A Best Asymptotically Normal (BAN) estimator is

1. **Introduction to BAN Estimators**

- A **Best Asymptotically Normal (BAN) estimator** is a type of estimator in statistics that is both **asymptotically normal** and **efficient** in the sense that it achieves the lowest possible asymptotic variance.
- BAN estimators arise in situations where Maximum Likelihood Estimators (MLEs) may be difficult to compute directly but can still be approximated with similar efficiency properties.

2. Properties of BAN Estimators

- **Asymptotic Normality**: An estimator $\hat{\theta}_n$ is asymptotically normal if:

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N\big(0, V(\theta)\big)$$

where V is the asymptotic variance.

- **Efficiency**: BAN estimators achieve the **Cramér-Rao lower bound** asymptotically, meaning they have the smallest possible variance among consistent estimators.
- **Consistency**: They converge to the true parameter as the sample size increases:

$$\hat{\theta}_n \xrightarrow{P} \theta$$

- **Asymptotic Equivalence to MLEs**: If an estimator is BAN, it behaves like an MLE for large samples in terms of efficiency and distribution.

3. Difference between BAN and MLE

- **MLE**: Derived directly from maximizing the likelihood function.
- **BAN Estimators**: Can be constructed in cases where the likelihood is complex, but the estimator still retains the same desirable asymptotic properties.

4. Applications of BAN Estimators

- Used in **generalized estimating equations (GEE)** and **robust estimation** when dealing with complex models.
- Common in **survey sampling, econometrics, and biostatistics** where exact MLEs are difficult to compute.

5. Example of a BAN Estimator

Consider a regression model where the likelihood is not explicitly solvable. Instead of the MLE, a **quasi-likelihood estimator** may be used, which is still BAN under regularity conditions.

8.8.1 Importance of BAN Estimators:

- **Optimal for Large Samples**: Provides the best performance among asymptotically normal estimators.
- **MLEs are Often BAN**: Many practical estimation methods rely on this property.
- **Used in Hypothesis Testing**: Asymptotic normality makes them useful for confidence intervals and statistical inference.

8.9 KEY TERMS:

- **Efficiency**: Measures how well an estimator uses the available data. An efficient estimator has the smallest possible variance among all unbiased estimators.

- **Asymptotic Relative Efficiency (ARE)**:
  - Compares the performance of two estimators as the sample size approaches infinity.
  - Given two estimators T_1 *and* T_2, the ARE is typically defined as:

$$ARE\left(T_1, T_2\right) = \frac{MSE\left(\hat{\theta}_2\right)}{MSE\left(\hat{\theta}_1\right)}$$

  - **CAN (Consistent and Asymptotically Normal) Estimator**:
- An estimator T_n is **consistent** if it converges in probability to the true parameter as n→∞
- It is **asymptotically normal** if its distribution, when properly normalized, approaches a normal distribution.
- **CAUN (Consistent and Asymptotically Unbiased) Estimator**:
- It is consistent (approaching the true value as sample size increases).
- It is asymptotically unbiased, meaning its bias tends to zero as n→∞.
- **BAN (Best Asymptotically Normal) Estimator**:

8.10 SUMMARY:

Understanding efficiency and asymptotic efficiency helps in selecting the best estimator for statistical problems. The properties of CAN, CAUN, and BAN estimators provide a structured approach to estimation, ensuring consistency, unbiasedness, and optimal asymptotic performance. Among these, BAN estimators are often preferred for their efficiency in large samples.

8.11 SELF ASSESSMENT QUESTIONS:

1. What is the efficiency of an estimator, and how is it measured? How does the Cramér-Rao lower bound relate to efficiency?
2. If two estimators are unbiased, how do you determine which one is more efficient? Define Asymptotic Relative Efficiency (ARE).
3. Explain CAN and CAUN estimators with suitable examples.

4. Describe the Fisher's definition of asymptotic efficiency. Give the definition of CAN, CUAN and the best CAN estimators.
5. Explain the Maximum Likelihood Estimator (MLE) is a CAN estimator.
6. Derive the Method of Moments Estimator (MME) is a CAUN estimator.
7. What is the significance of the asymptotic variance of a BAN estimator?
8. Discuss about the consistency criterion of estimation in large samples. Define the best CAUN estimator.

8.12 SUGGESTED READINGS:

1. Kale, B. K. (1999). A First Course on Parametric Inference, Narosa Publishing House, New Delhi.
2. Lehman, E. L., and Cassella, G. (1998). Theory of Point Estimation, Second Edition, Springer, NY.
3. Dudewicz, E. J., and Mishra, S. N. (1988). Modern Mathematical Statistics, John Wiley & Sons, NY.
4. Rajagopalan, M., and Dhanavanthan,P.(2012). Statistical Inference, PHI Learning Pvt., Ltd.,New Delhi.
5. Linear Statistical Inference and its Application by C.R. RAO (1973), John Wiley.
6. Goon, A. M., Gupta, M. K., and Dasgupta, B. (1989). An Outline of Statistical Theory - Vol.II, World Press, Calcutta.

Dr. U. Ramkiran

LESSON - 9
METHODS OF ESTIMATION

OBJECTIVES:

After studying the lesson the students will be conversant with some methods of point estimation.

- Understand the fundamental concepts of parameter estimation in statistics.
- Learn the importance of different estimation techniques in statistical inference.
- Compare various estimation methods based on their efficiency and applicability.

STRUCTURE:

9.1 INTRODUCTION:

So, far we have been discussing the requisites of a good estimator. Now we shall briefly outline some important methods for obtaining such estimators. The methods to estimate the parameters of the population using its representative called sample are many in number but we have in our syllabi, Moment method of estimation, ML method of estimation, Percentile estimation, Minimum Chi-square estimation, Modified minimum Chi-square estimation.

9.2 MOMENT METHOD OF ESTIMATION:

This method is introduced by Karl Pearson and it is the simplest method of finding estimator to the parameter by using the moments.

Let $f\left(x,\theta_{1,}\theta_{2,}.......,\theta_k\right)$ be the density function of the parent population with K parameters $\theta_{1,}\theta_{2,}.......,\theta_k$.

If μ_r^1 denotes the r[th] moment about origin, then

$$\mu_r^1 = \int x^r f\left(x;\theta_{1,}\theta_{2,}.......,\theta_k\right) dx, (r = 1,2,.....,K).$$

In general, $\mu_1^1, \mu_2^1,.....,\mu_k^1$ will be functions of the parameters $\theta_{1,}\theta_{2,}.......,\theta_k$.

Let $x_i = 1,2,........,n$ be a random sample of size n from a given population. This method consists in solving the K equations for $\theta_{1,}\theta_{2,}.......,\theta_k$ in terms of $\mu_1^1, \mu_2^1,.....,\mu_k^1$ and then replacing these moments $\mu_r^1, r = 1,2,........K$ by sample moments.

$$\text{i.e., } \hat{\theta}_i = \theta_i\left(\hat{\mu}_1^1, \hat{\mu}_2^1,.......,\hat{\mu}_K^1\right)$$

$$= \theta_i\left(m_1^1, m_2^1,.....,m_k^1\right), i = 1,2,........K$$

Where m_i^1 is the i[th] moment about origin in the sample.

Then by the method of moments $\hat{\theta}_1, \hat{\theta}_2,.........., \hat{\theta}_K$ are the required estimators of $\theta_{1,}\theta_{2,}.......,\theta_k$ respectively.

9.2.1 Remarks:

1. Let $\left(x_1, x_2,..., x_n\right)$ be a random sample of size n from a population with p.d.f $f\left(x,\theta\right)$. Then $x_i, \left(i = 1,2,...,n\right)$ are i.i.d $\Rightarrow x_i^r, \left(i = 1,2,...,n\right)$ are i.i.d. Hence if $E\left(x_i^r\right)$ exists, then by week law of large numbers, we get

$$\frac{1}{n}\sum_{i=1}^{n} x_i^r \xrightarrow{p} E\left(x_i^r\right) \Rightarrow m_r^{'} \xrightarrow{p} \mu_r^{'}$$

Hence the sample moments are consistent estimators of the corresponding population moments.

2. It has been shown that under quite general conditions, the estimates obtained by the method of moments are asymptotically normal but not, in general, efficient.
3. Generally the method of moments yields less efficient estimators than those obtained from the principle of maximum likelihood. The estimators obtained by the method of moments are identical with those given by the method of maximum likelihood if the probability mass function or probability density function is of the form:

   $$f\left(x,\theta\right) = \exp\left(b_0 + b_1 x + b_2 x^2 + ...\right) \qquad (1)$$

   where b's are independent of x but may depend on $\theta = \left(\theta_1, \theta_2,...\right)$. (1) implies that: $L\left(x_1, x_2,..., x_n;\theta\right) = \exp\left(nb_0 + b_1\sum x_1 + b_2\sum x_1^2 + ...\right)$

$$\Rightarrow \quad \frac{\partial}{\partial\theta}\log L = a_o + a_1\sum x_i + a_2\sum x_i^2 + a_3\sum x_1^3 + \ldots. \qquad (2)$$

Where $a_i = \frac{\partial}{\partial\theta}(b_i), (i = 1, 2, \ldots\ldots)$ and $a_0 = n\frac{\partial b_0}{\partial\theta}$

Thus both the methods yield identical estimators if MLE's are obtained as linear functions of the moments.

9.2.2 Properties:

The following are the properties of moment estimators:

1. The moment estimators can be obtained easily.
2. The moment estimators are not necessarily unbiased.
3. The moment estimators are consistent because by the law of large numbers a sample moment (raw or central) is a consistent estimator for the corresponding population moment.
4. The moment estimators are generally less efficient than maximum likelihood estimators.
5. The moment estimators are asymptotically normally distributed.
6. The moment estimators may not be function of sufficient statistics.
7. The moment estimators are not unique.

9.2.3 Problems Based on Moment Method of Estimation:

1. By the method of moments, find the estimators to the parameters in Normal population $N(\mu, \sigma^2)$.

Solution: From the given population $\mu_1^1 = \mu$

$$\mu_2^1 = \sigma^2 + \mu^2.$$

If the sample moments with a sample size 'n' are m_1^1, m_2^1, then

$$m_1^1 = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$$

$$m_2^1 = \frac{1}{n}\sum_{i=1}^{n} x_i^2.$$

If we equate sample moments with population moments, we get

$$\bar{x} = \mu$$

$$\frac{1}{n}\sum_{i=1}^{n} x_i^2 = \sigma^2 + \mu^2$$

Solving the above equations, we get $\hat{\mu} = \bar{x}$

$$\hat{\sigma}^2 = \frac{\sum x_i^2}{n} - \bar{x}^2 = s^2$$

Therefore, $\bar{x}$ and s^2 are the moment estimators μ and σ^2 respectively.

2. Let $X_1, X_2, \ldots\ldots, X_n$ be Bernoulli random variables with parameter p. What is the method of moments estimator of p?

Answer:
Here, the first theoretical moment about the origin is:

$$E(X_i) = p$$

We have just one parameter for which we are trying to derive the method of moments estimator. Therefore, we need just one equation. Equating the first theoretical moment about the origin with the corresponding sample moment, we get:

$$p = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Now, we just have to solve for p. In this case, the equation is already solved for p. We just need to put a hat (^) on the parameter to make it clear that it is an estimator. We can also subscript the estimator with an "MM" to indicate that the estimator is the method of moment's estimator:

$$\hat{p}_{MM} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

So, in this case, the method of moment's estimator is the same as the maximum likelihood estimator, namely, the sample proportion.

9.3 MAXIMUM LIKELIHOOD METHOD OF ESTIMATION:

This method was introduced by Prof. R. A. Fisher and later on developed by him in a series of papers.

The principle of maximum likelihood consists in finding an estimator for the unknown parameter θ which the likelihood function $L(\theta)$ for variations in parameter.

We wish to find $\hat{\theta} = \left(\hat{\theta}_1, \hat{\theta}_2, \ldots\ldots\ldots, \hat{\theta}_K\right)$ so that

$$L\left(\hat{\theta}\right) = Sup\ L(\theta) \forall \theta.$$

Thus if there exists a function $\hat{\theta} = \hat{\theta}(x_1, x_2, \ldots\ldots., x_n)$ of sample values which maximizes L for variations in θ, then $\hat{\theta}$ is to be taken as an estimator of θ. $\hat{\theta}$ is called maximum likelihood estimator (M.L.E.) of θ.

Thus $\hat{\theta}$ is the solution, if any, of

$$\frac{\partial L}{\partial \theta} = 0 \text{ and } \frac{\partial^2 L}{\partial \theta^2} < 0 \quad .$$

Since $L > 0$, and log L is a non-decreasing function of L, L and log L attain their extreme values (maxima and minima) at the same value of $\hat{\theta}$. The above two equations can be written as

$$\frac{\partial \log L}{\partial \theta} = 0 \text{ and } \frac{\partial^2 \log L}{\partial \theta^2} < 0,$$

A form which is much more convenient from practical point of view.

9.3.1 Regularity Conditions:

a) The first and second order derivatives $\frac{\partial \log L}{\partial \theta}, \frac{\partial^2 \log L}{\partial \theta^2}$ exists and are continuous functions of θ in a range R.

b) Third order derivative $\frac{\partial^3 \log L}{\partial \theta^3}$ exists such that $\left|\frac{\partial^3 \log L}{\partial \theta^3}\right| < M(X)$ where $E\left[M(X)\right] < k$, a positive quantity.

c) For every θ in R $E\left(\frac{\partial^2 \log L}{\partial \theta^2}\right) = \int_{-\infty}^{\infty}\left(\frac{\partial^2 \log L}{\partial \theta^2}\right) L dx = I(\theta)$ is finite and non-zero.

d) The range of integration is independent of θ. But if the range of integration depends on θ, then $f(x,\theta)$ vanishes at the extremes depending on θ.

9.3.2 Properties of Maximum Likelihood Estimators:

The following are the properties of maximum likelihood estimators:

1. A ML estimator is not necessarily unique.
2. A ML estimator is not necessarily unbiased.
3. A ML estimator may not be consistent in rare case.
4. If a sufficient statistic exists, it is a function of the ML estimators.
5. If $T = t(X_1, X_2, ..., X_n)$ is a ML estimator of θ and $\Upsilon(\theta)$ is a one to one function of θ, then $\Upsilon(T)$ is a ML of $\Upsilon(\theta)$. This is known as invariance property of ML estimator.
6. When ML estimator exists, then it is most efficient in the group of such estimators. It is now time for you to try the following exercises to make sure that you get the concept of ML estimators.

9.3.3 Problems Based on ML Method of Estimation:

Example 1: For random sampling from normal population $N(\mu,\sigma^2)$, find the maximum likelihood estimators for μ and σ^2.

Solution: Let $X_1, X_2, ..., X_n$ be a random sample of size n taken from normal population $N(\mu,\sigma^2)$, whose probability density function is given by

$$f\left(x,\mu,\sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}; -\infty < x < \infty, -\infty < x < \infty, \sigma > 0 \qquad (1)$$

Therefore, the likelihood function for parameters μ and σ^2 can be obtained as

$$L\left(\mu,\sigma^2\right) = L = f\left(x_1,\mu,\sigma^2\right).f\left(x_2,\mu,\sigma^2\right)...f\left(x_n,\mu,\sigma^2\right)$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_1-\mu)^2}{2\sigma^2}} . \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_2-\mu)^2}{2\sigma^2}} ... \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_n-\mu)^2}{2\sigma^2}}$$

$$= \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2} \qquad (2)$$

Taking log on both sides of equations (2), we get

$$\log L=\frac{n}{2}\log 1-\frac{n}{2}\log 2\pi-\frac{n}{2}\log\sigma^2-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(x_i-\mu\right)^2 \qquad (3)$$

Differentiate equation (3) partially with respect to μ and σ^2 respectively, we get

$$\frac{\partial}{\partial\mu}(\log L)=-\frac{1}{2\sigma^2}\sum_{i=1}^{n}2\left(x_i-\mu\right)(-1) \qquad (4)$$

$$\frac{\partial}{\partial\sigma^2}(\log L)=-\frac{n}{2}\frac{1}{\sigma^2}-\frac{1}{2\left(\sigma^2\right)^2}(-1)\sum_{i=1}^{n}\left(x_i-\mu\right)^2 \qquad (5)$$

For finding ML estimate of μ, we put

$$\frac{\partial}{\partial\mu}(\log L)=-\frac{1}{2\sigma^2}\sum_{i=1}^{n}2\left(x_i-\mu\right)(-1)=0$$

$$\sum_{i=1}^{n}\left(x_i-\mu\right)=0$$

$$\sum_{i=1}^{n}x_i-n\mu=0\Rightarrow\hat{\mu}=\frac{\sum_{i=1}^{n}x_i}{n}=\bar{x}$$

Thus, the ML estimate for μ is the observed sample mean $\bar{x}$. For ML estimate of σ2, we put

$$\frac{\partial}{\partial\sigma^2}(\log L)=-\frac{n}{2}\frac{1}{\sigma^2}-\frac{1}{2\left(\sigma^2\right)^2}(-1)\sum_{i=1}^{n}\left(x_i-\mu\right)^2=0$$

$$\Rightarrow\frac{-n\sigma^2+\sum_{i=1}^{n}\left(x_i-\mu\right)^2}{2\sigma^4}=0$$

$$\Rightarrow -n\sigma^2+\sum_{i=1}^{n}\left(x_i-\mu\right)^2=0$$

$$\Rightarrow\hat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\mu\right)^2=s^2$$

Thus, the ML estimates for μ and σ^2 are $\bar{x}$ and s^2 respectively.

Example 2: Find the maximum likelihood estimate for the parameter λ of a Poisson distribution on the basis of a sample size n.

Solution: The probability function of Poisson distribution with parameter λ is given

by $P\left(X=x\right)=\dfrac{e^{-\lambda}\lambda^x}{x!};x=0,1,2,....$

Likelihood function of random sample $x_{1,}x_{2,}.......,x_n$ of n observations from this population is

$$L=\prod_{i=1}^{n}P\left(X=x_i\right)=\frac{e^{-n\lambda}\lambda^{\prod_{i=1}^{n}x_i}}{x_1!x_2!.......,x_n!}$$

$$\log L=-n\lambda+n\bar{x}\ \log\lambda-\sum_{i=1}^{n}\log\left(x_i!\right)$$

The likelihood equation for estimating, λ is

$$\frac{\partial \log L}{\partial \lambda} = 0 \Rightarrow -n + \frac{n\overline{x}}{\lambda} = 0$$

$$\therefore \hat{\lambda} = \overline{x}$$

Thus the MLE for λ is the sample mean $\overline{x}$.

Example 3: Obtain the maximum likelihood estimated of P in the case of sample observations $x_{1,} x_{2,} ...., x_n$ of size n drawn from binomial distribution $B(n, P)$.

Solution: If $x_{1,} x_{2,} ...., x_n$ is a random sample drawn from binomial distribution the likelihood function of sample $x_{1,} x_{2,} ...., x_n$ is

$$L = \prod_{i=1}^{n} \binom{n}{x_i} p^{\sum x_i} q^{\sum (n - x_i)} where\ q = 1 - p.$$

$$\log L = \sum_{i=1}^{n} \log \binom{n}{x_i} + \sum x_i \log P + \left(n^2 - \sum x_i\right) \log\ q$$

$$= \sum_{i=1}^{n} \log \binom{n}{x_i} + \sum x_i \log P + \left(n^2 - \sum x_i\right) \log(1 - p)$$

Consider $\frac{\partial \log L}{\partial P} = 0 \Rightarrow \frac{\sum_{i=1}^{n} x_i}{P} + \frac{n^2 - \sum_{i=1}^{n} x_i}{(1 - P)}(-1) = 0$

$$(1 - P)\sum_{i=1}^{n} x_i - P\left(n^2 - \sum_{i=1}^{n} x_i\right) = 0$$

$$\sum_{i=1}^{n} x_i - P\sum_{i=1}^{n} x_i - n^2 P + P\sum_{i=1}^{n} x_i = 0$$

$$\sum_{i=1}^{n} x_i - n^2 P = 0$$

$$n^2 P = \sum_{i=1}^{n} x_i$$

$$P = \frac{\sum_{i=1}^{n} x_i}{n^2} = \frac{\overline{x}}{n}$$

$$\hat{P} = \frac{\overline{x}}{n}$$

Thus $\frac{\overline{x}}{n}$ is the MLE for the parameter P in binomial distribution.

9.4 PERCENTILE ESTIMATION:

The term percentile plays an important role in descriptive statistics and percentiles can be used with the cumulative distribution function. By solving the resulting equations simultaneously, we get estimates for the unknown parameters.

In our usual estimation procedure, we suppose to a suitable statistic which is a function of the observations to estimate the regarding parameter $'\theta'$.

Instead of using the statistic which is a function of sample observations, if possible, use one observation instead of total sample observations which are representatives of samples i.e., estimation based on one or two ordered statistics X(k), X(n). This process of estimation based on the one or two ordered statistics is known as percentile estimation. But it is not an efficient method of estimation.

Suppose the sample size is 100. If the 67th percentile is suitable to estimate the parameter than to proceed. The further estimation of the 67th percentile may already discuss that not be a sufficient statistic and it's not an efficient procedure to estimate by the method of percentile having some asymptotic variance. The user wishes to tolerate this minimum asymptotic variance then we can suggest the estimation of parameters both on percentiles.

Example: Find the estimate the parameters of the exponential distribution by using percentile estimation.

Solution: The Cumulative distribution of the exponential distribution is

$$\text{Cdf}=\ F(x) = 1 - e^{-\theta x}; x \geq 0, \theta > 0$$

$$F(x) = \text{P}$$

$$1 - e^{-\theta x} = P$$

$$e^{-\theta x} = 1 - P$$

$$\ln\,(e^{-\theta x}) = \ln\,(1 - P)$$

$$-\theta\text{x} = \ln\,(1 - P)$$

$$\text{x} = \frac{-\ln\,(1 - P)}{\theta}$$

To get different percentile values replace P with 0.10, 0.50 to get the 10th and 50th percentile respectively.

9.5 MINIMUM CHI-SQUARE ESTIMATION:

The method of minimum Chi-square makes use of the Pearson's Chi-square statistic. This method can be used in case of discrete distributions or for grouped data from a continuous distribution.

Let $f_1, f_2, ..., f_k$ be the observed frequencies in K grouped or classes and unknown probabilities that f_i elements belong to the i^{th} group or class be $p_i\,(i = 1, 2, ..., k)$. P's are the functions of unknown parameters $\theta_{1,}\theta_{2,}......., \theta_m$.Thus $p_i = p_i(\theta)$ where $\underset{\sim}{\theta} = (\theta_{1,}\theta_{2,}......., \theta_m)$. Suppose the total sample size n. Therefore, $\sum f_i = n$. The expected frequencies are $np_1\left(\underset{\sim}{\theta}\right), np_2\left(\underset{\sim}{\theta}\right), ..., np_k\left(\underset{\sim}{\theta}\right)$. We know, Personian Chi-square statistic is

$$\chi^2 = \sum_{i=1}^{k} \frac{\left[f_i - np_i\left(\underset{\sim}{\theta}\right)\right]^2}{np_i\left(\underset{\sim}{\theta}\right)}$$

$$\sum_{i=1}^{k} \frac{f_i^2}{np_i\left(\underset{\sim}{\theta}\right)} - n$$

Under the method of minimum Chi-square one has to choose $(\theta_1, \theta_2, ..., \theta_m)$ which minimize χ^2. This will be minimum when $np_i(\theta)$ is as close as possible to f_i. So to obtaon the estimates of θ_i's, partially differentiate χ^2- statistic. W.r.t. $\theta_i(i = 1, 2, ..., m)$ successively and equate them to zero. Also check that the second derivatives and non-negative, i.e.,

$$\frac{\partial \chi^2}{\partial \theta_i} = 0 \text{ for } i = 1, 2, ..., m$$

and

$$\frac{\partial^2 \chi^2}{\partial \theta_i^2} \geq 0$$

$\frac{\partial \chi^2}{\partial \theta_i} = 0$ provides m simultaneous equations in m unknowns. Solving these m- equations for m unknown parameters, one gets the estimated values of $\theta_1, \theta_2, ..., \theta_m$ respectively.

9.5.1 Problems Based on Minimum CHI-Square Estimation:

1. A method for generating uniformly distributed random integers in the range 0–9 has been devised and tested by generating 1000 digits with results shown below.

Digit	0	1	2	3	4	5	6	7	8	9
Frequency	106	89	86	110	123	93	82	110	91	111

Do these results support the idea that the method of generation is suitable?

Solution:

If the digits were uniformly distributed, then the expected frequencies would all be 100. So, using

$$\chi^2 = \sum_{i=1}^{k} \frac{\left[f_i - np_i(\theta)\right]^2}{np_i(\theta)}$$

We find $\chi^2 = 16.86$ and this is for 9 degrees of freedom. From χ^2 table $\chi^2 \geq 16.86$ for 9 degrees of freedom is 0.05. So although it cannot be ruled out, as this is a fairly low probability, it raises some doubt that the method really is producing uniformly distributed integers.

2. Is gender independent of education level? A random sample of 395 people was surveyed and each person was asked to report the highest education level they obtained. The data that resulted from the survey are summarized in the following table:

	High School	**Bachelors**	**Masters**	**Ph.D.**	**Total**
Female	60	54	46	41	201
Male	40	44	53	57	194
Total	100	98	99	98	395

Are gender and education level dependent at a 5% level of significance? In other words, given the data collected above, is there a relationship between the gender of an individual and the level of education that they have obtained?

Solution: Here's the table of expected counts:

	High School	Bachelors	Masters	Ph.d.	Total
Female	50.886	49.868	50.377	49.868	201
Male	49.114	48.132	48.623	48.132	194
Total	100	98	99	98	395

So, working this out,

$$\chi^2 = \frac{(60-50.886)^2}{50.886} + \ldots + \frac{(57-48.132)^2}{48.132} = 8.006$$

The critical value of χ^2 with 3 degrees of freedom is 7.815. Since 8.006 > 7.815, we reject the null hypothesis and conclude that the education level depends on gender at a 5% level of significance.

9.6 MODIFIED MINIMUM CHI-SQUARE ESTIMATION:

Expected frequency $np_i(\theta)$ in the denominator of χ^2- statistic causes certain difficulties. Hence a modification has been suggested which makes the process of differentiation easier. The modified Chi - square statistic is,

$$\chi^2 = \sum_{i=1}^{k} \frac{\left[np_i(\theta) - f_i\right]^2}{f}$$

$$\sum_{i=1}^{k} \frac{n^2 p_i^2(\theta)}{f_i} - n$$

Rest of the procedure of modified minimum. Chi-Square remains same as that minimum Chi-Square.

9.7 APPLICATIONS OF CHI-SQUARE ESTIMATION:

Minimum Chi-Square Estimation (MCSE) is a statistical estimation method used in various fields where observed frequencies are compared to expected frequencies under a given model. Below are some key applications of this method:

1. Contingency Table Analysis
   - Used to estimate parameters in categorical data models.
   - Helps in assessing relationships between categorical variables in surveys and social sciences.
2. Goodness-of-Fit Testing
   - Estimates parameters while ensuring the model fits the data well.
   - Applied in testing whether a sample follows a theoretical distribution (e.g., Poisson, binomial, normal).
3. Reliability Engineering & Quality Control
   - Used in life-testing experiments to estimate failure rates.
   - Helps in analyzing discrete failure-time data for product reliability assessment.

4. Epidemiology & Medical Statistics
   - Applied in disease modeling and case-control studies.
   - Helps estimate risk factors when comparing observed and expected frequencies in health data.
5. Genetics & Bioinformatics
   - Used in Hardy-Weinberg equilibrium testing to estimate allele frequencies.
   - Applied in genetic linkage studies where observed genotype distributions are compared with expected ones.
6. Econometrics & Market Research
   - Used in discrete choice models to analyze consumer behavior.
   - Helps in modeling economic variables based on frequency-based observations.
7. Actuarial Science
   - Applied in insurance risk modeling and loss distribution estimation.
   - Helps actuaries estimate claim frequencies and policyholder behavior patterns.
8. Physics & Experimental Sciences
   - Used in fields like particle physics and astrophysics where observed event counts are compared with theoretical models.
   - Helps fit models in experimental physics where data follow probability distributions.
9. Machine Learning & Data Science
   - Applied in classification problems involving categorical data.
   - Helps in model selection and validation by comparing observed class distributions with expected ones.

9.7.1 Difference Between Minimum Chi-Square and Modified Minimum Chi-Square Estimation:

Feature	**Minimum Chi-Square Estimation (MCSE)**	**Modified Minimum Chi-Square Estimation (MMCSE)**
Definition	Estimates parameters by minimizing the chi-square statistic between observed and expected frequencies.	A refined version of MCSE that improves efficiency and robustness, particularly for small samples or sparse data.
Bias & Efficiency	Can be biased, especially in small sample sizes or when expected frequencies are very small.	Provides more efficient and less biased estimates, improving performance in small samples.
Formula Difference	Uses the standard chi-square formula: $\chi^2 = \sum_{i=1}^{k} \frac{\left[f_i - np_i(\theta)\right]^2}{np_i(\theta)}$	Adjusts the chi-square formula by modifying the denominator or applying weight adjustments to improve estimation.
Small Sample Performance	May not perform well when expected frequencies are small or	More stable and reliable in small sample cases, reducing

	zero, leading to instability.	sensitivity to small expected frequencies.
Computational Complexity	Simpler to compute, but less robust in certain scenarios.	Slightly more complex but yields better estimates in challenging conditions.
Robustness	Can be influenced by outliers.	More robust against outliers and extreme observations.

9.8 KEY WORDS:

Here are some key terms related to the methods of estimation you mentioned:

1. **Moment Method of Estimation**: A technique for estimating parameters of a probability distribution by equating sample moments (e.g., mean, variance) to theoretical moments.
2. **Maximum Likelihood Method of Estimation (MLE)**: A method that estimates parameters by maximizing the likelihood function, which is the probability of observing the given data under different parameter values.
3. **Percentile Estimation**: Estimation based on specific percentiles of the distribution (such as the median or other quantiles) to infer parameters.
4. **Minimum Chi-square Estimation**: An estimation method that minimizes the chi-square statistic, which measures the difference between observed and expected frequencies in a distribution.
5. **Modified Minimum Chi-square Estimation**: A variation of the minimum chi-square method, often adjusting for biases or adding weights to the estimation process.

9.9 SUMMARY

This chapter explored various estimation techniques used in statistical inference. The moment method of estimation provides parameter estimates by equating sample and theoretical moments. Maximum likelihood estimation identifies parameters by maximizing the likelihood function. Percentile estimation uses empirical percentiles to estimate distribution parameters. Minimum chi-square estimation minimizes the discrepancy between observed and expected frequencies, and the modified minimum chi-square method refines this approach for improved robustness. Each method has its own strengths and applications, making them useful tools in statistical analysis.

9.10 SELF ASSESSMENT QUESTIONS:

1. What is the fundamental principle behind the moment method of estimation?
2. Explain the key idea of maximum likelihood estimation and how it differs from the moment method.
3. How is percentile estimation used to estimate distribution parameters?
4. If $x_{1,}x_{2,}.......,x_n$ denote a random sample from a population with p.d.f. $f(x,\theta)=\theta x^{\theta-1}, 0<x<1$. Show that $Y=x_{1,}x_{2,}.......,x_n$ is a sufficient statistic for θ.
5. Examine if the following distribution admits a sufficient statistic for the parameter θ. $f(x,\theta)=(1+\theta)x^{\theta}, 0\le x\le 1, \theta>0$
6. Describe the concept of minimum chi-square estimation and its applications.

7. How does modified minimum chi-square estimation improve upon the standard chi-square method?
8. Provide an example where the moment method of estimation might be preferred over MLE.
9. Discuss the advantages and limitations of maximum likelihood estimation.
10. (a) Prove that the maximum likelihood estimate of the parameter α of a population having density function: $\frac{2}{\alpha}(\alpha - x), 0 < x < \alpha$.

    (b) Describe Percentile Estimation with example.

9.11 SUGGESTED READINGS:

1. Goon, A. M., Gupta, M. K., and Dasgupta, B. (1989). An Outline of Statistical Theory-Vol.II, World Press, Calcutta.
2. Kale, B. K. (1999). A First Course on Parametric Inference, Narosa Publishing House, New Delhi.
3. Lehman, E. L., and Cassella, G. (1998). Theory of Point Estimation, Second Edition, Springer, NY.

Dr. K. Kalyani

LESSON -10
INTERVAL ESTIMATION

OBJECTIVES:

By the end of this lesson, you should be able to:

- Understand the concept and purpose of interval estimation in statistical inference
- Define interval estimation and its role in estimating population parameters
- Recognize real-world applications where interval estimation is essential
- Explain how increasing or decreasing the confidence level affects the width of the confidence interval.
- Construct confidence intervals for population parameters using appropriate pivot methods.

STRUCTURE:

10.1 INTRODUCTION:

In point estimation, when the random variable are $X_1, X_2, \ldots\ldots, X_n$ and θ is the unknown parameter (or the set of unknown parameters), we try to estimate a parametric function $\gamma(\theta)$ by means of a single value, say t, the value of a statistic (an estimator) T corresponding to the observed values $x_{1,} x_{2,} \ldots\ldots, x_n$ of the random variables. It is, however, far from our intention to take t to be exactly the value of $\gamma(\theta)$ in the given population. All that we mean by taking t to be the estimate of $\gamma(\theta)$ is that t is likely to differ from $\gamma(\theta)$ only by a small amount. This is why is has become customary in statistical practice to give, together with the estimate t, the standard error or an estimate of the standard error of the estimator T. Having the standard error or its estimate, one may suppose that $\gamma(\theta)$ is very likely to lie between $t \neq s.e.$, still likely to lie between $t \neq 2s.e.$, and so on.

This idea is given clearer expression in the estimation procedure. Here two limits, say t_1 and t_2 $(t_1 < t_2)$, are computed from the set of observations X and it is claimed with a certain degree of confidence (measured in probabilistic terms) that the true value of $\gamma(\theta)$ lies between t_1 and t_2. Thus, instead of trying to pinpoint the true value of $\gamma(\theta)$ by means of a single value t (a 'point), we put forward an interval $[t_1, t_2]$ which we expect would include the true value of $\gamma(\theta)$, That is why this type of estimation is called interval estimation. We shall see in the later part that the estimation of $\gamma(\theta)$ may be achieved more generally in terms of a set, which need not be an interval. The methods of obtaining such intervals or sets, as also the formulation of criteria for intervals or sets that may be regarded as he most desirable, are the subject – matter.

10.2 INTERVAL ESTIMATION:

If the point estimation does not give the correct and best estimation to the parameter θ, in this case interval estimation is more useful to estimate the parameter.

Let $x_{1,} x_{2,} \ldots\ldots, x_n$ be a random sample from a population with the density function $f(x, \theta)$. If we use two functions of the sample values as an interval to estimate the population parameter θ, this interval is called the confidence interval and this estimation procedure is called interval estimation to the unknown parameter θ.

10.3 CONFIDENCE LEVEL:

In interval estimation, the "confidence level" refers to the degree of certainty or probability that a calculated confidence interval actually contains the true population parameter, typically expressed as a percentage (like 95%) - meaning that if you repeated the sampling process multiple times, the calculated interval would capture the true value in that percentage of cases.

Key points about confidence level:

Interpretation:

A 95% confidence level indicates that if you were to take many random samples and calculate a confidence interval for each, approximately 95% of those intervals would contain the true population parameter.

Relationship with margin of error:

A higher confidence level results in a wider confidence interval, which means a larger margin of error.

Choosing a confidence level:

The appropriate confidence level depends on the context of the study and the level of certainty desired.

Example:

If you calculate a 95% confidence interval for the average height of adults in a city, it means that you are 95% confident that the true average height falls within the range specified by that interval.

10.4 LEVEL OF SIGNIFICANCE AND CONFIDENCE LEVEL:

Let us consider the following problem $H_0 : \theta = \theta_0$ vs $H_0 : \theta \neq \theta_0$. Here $X_1, X_2, ..., X_n$ be the random sample and W be the critical region.

Let the level of significance is assigned to be α

$P[Rejecting \text{ true } H_0] \leq \alpha$

$\Rightarrow P[\underset{\sim}{X} \in W \; H_0] \leq \alpha$

$\Rightarrow 1 - P[\underset{\sim}{X} \in W \; H_0] \geq 1 - \alpha$

$\Rightarrow P[\text{Accepting the null when it is true}] \geq 1 - \alpha$

$\Rightarrow P[\text{Containing the true value of the parameter}] \geq 1 - \alpha$

Hence the confidence coefficient is $100(1-\alpha)\%$.

Fundamental Notation of Confidence Estimation:

Considered a random variable on some function of it as the basis observable quantity. Let X be a random variable and a, b be two given positive real numbers then,

$$P(a < X < b) = P(a < X \text{ and } X<b) = P\left(b < \frac{bX}{a} \text{ and } X<b\right) = P\left(X < b < \frac{bX}{a}\right)$$

as if we know the distribution of X and quantities a and b. Then we can determine the probability $P(a < X < b)$. Consider the interval $I(X) = \left(X, \frac{bX}{a}\right)$. This is an interval with a random variable in the end points and hence it takes the values $\left(X, \frac{bX}{a}\right)$ whenever the random variable. X takes the values of x. Thus $I(X)$ is a random quantity and is n example of a random interval. Note that $I(X)$ includes the value b with a certain fixed probability. In general, larger the length of the interval, the larger the coverage probability.

10.5 CONFIDENCE INTERVAL :

Let $X_1, X_2, ..., X_n$ be a random sample of size n on a random variable X having distribution belonging to the family

$$H = \{f_\theta(x) : \theta \in \Theta\}$$

If $\underline{\theta}(\underline{X})$ and $\overline{\theta}(\overline{X})$ be two statistics ∋

$P_\theta\left[\underline{\theta}(X) < \theta < \overline{\theta}(X)\right] \geq 1 - \alpha$. Then $\underline{\theta}(\underline{X})$ and $\overline{\theta}(\overline{X})$ is called a confidence interval with confidence coefficient $(1-\alpha)$. Confidence interval means the region where the value of the parametric function lies.

10.5.1 EXAMPLES:

1. $X \sim N\left(\theta, \sigma^2\right); \sigma^2$ is known. Find a confidence interval of θ with confidence coefficient $(1-\alpha)$.

Answer:
$$P_\theta\left(\left|\frac{\sqrt{n}\left(\bar{X}-\theta\right)}{\sigma}\right| > \tau_{\alpha/2}\right) = \alpha$$

$$P_\theta\left[-\tau_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \bar{X}-\theta < \tau_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right] = 1-\alpha$$

$$P_\theta\left[\bar{X}-\tau_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \theta < \bar{X}+\tau_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right] = 1-\alpha$$

$$\therefore \underline{\theta}\left(\underline{X}\right) = \bar{X}-\tau_{\alpha/2}\frac{\sigma}{\sqrt{n}} \text{ and } \bar{\theta}\left(\bar{X}\right) = \bar{X}+\tau_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

Hence, $100(1-\alpha)$ confidence interval of θ is given by

$$\left[\bar{X}-\tau_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X}+\tau_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right].$$

Step to find out Confidence Interval:

i. Give the critical region of the both tailed test at level α
ii. Reverse the inequality sign and hence the RHS will be $(1-\alpha)$.
iii. From the inequality under probability solve for θ.

Example 2: Confidence Interval for the mean when the Variance of normal distribution is known
Answer: Let us assume that we have a both sides from Normal Population with mean μ and variance σ^2. As we know that the most efficient point estimation from the population mean μ is the sample mean $\bar{X}$, we can find a C.I. for μ by considering the sampling distribution of $\bar{X}$.

$$\bar{X} \sim N\left(\mu, \sigma^2/n\right)$$

and $Z = \dfrac{\bar{X}-\mu}{\sigma^2/n} \sim N(0,1)$

So, that $f(Z) = \dfrac{1}{\sqrt{2\Pi}} e^{-\frac{1}{2}Z^2}, Z \in R$

Now, let us assume that $Z_{\alpha/2}$ be the value of Z such that

$$P\left(Z \geq Z_{\alpha/2}\right) = \int_{Z_{\alpha/2}}^{\infty} f(Z)\,dz = \int_{Z_{\alpha/2}}^{\infty} \frac{1}{\sqrt{2\Pi}} e^{-\frac{1}{2}Z^2} dz = \alpha/2$$

and $Z_{1-\alpha/2} = -Z_{\alpha/2}$ the value of Z such that

$$P\left(Z \leq -Z_{\alpha/2}\right) = \int_{-\infty}^{-Z_{\alpha/2}} f(Z)\,dz = \int_{-\infty}^{-Z_{\alpha/2}} \frac{1}{\sqrt{2\Pi}} e^{-\frac{1}{2}Z^2} dz = \alpha/2$$

Thus, clearly $P\left(-Z_{\alpha/2} \le Z \le Z_{\alpha/2}\right) = 1-\alpha$

or, $P\left(-Z_{\alpha/2} \le \frac{\bar{X}-\mu}{\sigma^2/n} \le Z_{\alpha/2}\right) = 1-\alpha$

or, $P\left(-Z_{\alpha/2}.\frac{\sigma^2}{\sqrt{n}} \le \bar{X}-\mu \le Z_{\alpha/2}.\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$

$P\left(\bar{X}-Z_{\alpha/2}.\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X}+Z_{\alpha/2}.\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$

or, $P\left(\bar{X}-Z_{\alpha/2}.\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X}+Z_{\alpha/2}.\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$

Thus, the $(1-\alpha)\%$ confidence interval from μ in $N\left(\mu,\sigma^2\right)$ is

$$\bar{X}-Z_{\alpha/2}.\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X}+Z_{\alpha/2}.\frac{\sigma}{\sqrt{n}}.$$

Example 3: Confidence Interval for the mean when the Variance of normal population is not known.

Answer: If variance is not known, then σ is replaced by S,

Where $S^2 = \frac{1}{n-1}\sum\left(X_i-\bar{X}\right)^2$

In this case, we use the t-statistic defined as

$$t = \frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}$$

Now, $P\left(-t_{\alpha/2} \le \text{ t } \le \text{ t}_{\alpha/2}\right) = \int_{-\alpha/2}^{\alpha/2} f\left(t,n-1\right)dt = 1-\alpha$

$\Rightarrow P\left(-t_{\alpha/2} \le \frac{\bar{X}-\mu}{S/\sqrt{n}} \le \text{ t}_{\alpha/2}\right) = 1-\alpha$

or, $P\left(\bar{X}-t_{\alpha/2}.\frac{S}{\sqrt{n}} \le \mu \le \bar{X}+t_{\alpha/2}.\frac{S}{\sqrt{n}}\right) = 1-\alpha$

Thus, $(1-\alpha)100\%$ confidence interval for μ is

$$\bar{X}-t_{\alpha/2}.\frac{S}{\sqrt{n}} \le \mu \le \bar{X}+t_{\alpha/2}.\frac{S}{\sqrt{n}}$$

Problem 1: Obtain 95% confidence intervals for mean of a normal distribution with known variance σ^2 .

Answer: $\bar{X} \sim N\left(\mu,\sigma^2/n\right)$

and $Z = \frac{\bar{X}-\mu}{\sigma^2/n} \sim N\left(0,1\right)$

Also, $P\left(-Z_{\alpha/2} \le Z \le Z_{\alpha/2}\right) = P\left(\bar{X}-Z_{\alpha/2}.\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X}+Z_{\alpha/2}.\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$ (say) for $\alpha = 0.05,\ Z_{\alpha/2} = 1.96$

Then we have,

$$P\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

and thus,

$\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}$ is a confidence interval for μ, with a confidence coefficient 0.95.

Problem 2: Find 95% confidence interval for exponential distribution with p.d.f. $f(x) = \theta.e^{-\theta x},\ 0 \le x \le \infty,\ \theta > 0$.

Answer: $\mathrm{E}(X) = \frac{1}{\theta},\ \mathrm{V}(X) = \frac{1}{\theta^2}$

Thus, $\mathrm{E}(\bar{X}) = \frac{1}{\theta},$

$$\mathrm{V}(\bar{X}) = \frac{1}{n^2}V(X_1 + X_2 + \ldots\ldots + X_n)$$

$$= \frac{1}{n^2}.\,n\,V(X)$$

$$= \frac{1}{n\theta^2}$$

Using CLT for large n, we have

$$Z = \frac{\hat{\theta} - E(\hat{\theta})}{\sqrt{V(\hat{\theta})}} \sim N(0,1)$$

$$i.e.\ \mathrm{Z} = \mathrm{n}\left(\frac{1}{\theta} - \bar{X}\right) \sim N(0,1)$$

$$\Rightarrow \sqrt{n}(1 - \theta\bar{X}) \sim N(0,1)$$

Hence, 95% confidence limits for θ are given by

$$P\left[-1.96 \le \sqrt{n}(1 - \theta\bar{X}) \le 1.96\right] = 0.95$$

Now, $\sqrt{n}(1 - \theta\bar{X}) \le 1.96$

$$\Rightarrow \left(1 - \frac{1.96}{\sqrt{n}}\right)\frac{1}{\bar{X}} \le \theta \qquad (1)$$

and

$$-1.96 \le \sqrt{n}(1 - \theta\bar{X})$$

$$\Rightarrow \theta \le \left(1 + \frac{1.96}{\sqrt{n}}\right)\frac{1}{\bar{X}} \qquad (2)$$

Hence from (1) & (2), the 95% C.I. is given by

$$\theta = \left(1 \pm \frac{1.96}{\sqrt{n}}\right) \cdot \frac{1}{\overline{X}}$$

10.6 CONSTRUCTION OF CONFIDENCE LEVELS USING PIVOTS:

Definition:- For a random sample $(x1, \ldots, xn)$ from the distribution of a $r.\ v.\ x$ having þ, d, (x, θ) Let $L1L1(x1, \ldots, xn)$and $L2(x1, \ldots, xn)$ be two statistics such that $L1 \leq L2$. The interval $[L1, L2]$ is a confidence interval for θ with. Confidence coefficient $1-\propto$ $(0 <\propto< 1)$ if $P\theta\ [L1 \leq \theta \leq L2\ ] = 1-\propto$ for all $\theta \epsilon \Omega L1$ and $L2$ are called the lower and upper confidence limits, respectively at least one of them should not be a constant.

Construction of Confidence Intervals using Pivots

To construct a confidence interval using a pivot, follow these steps:

1. Choose a pivot: Select a pivot that is appropriate for the problem, such as the sample mean or the sample proportion.
2. Determine the distribution of the pivot: Determine the distribution of the pivot, such as the standard normal distribution or the t-distribution.
3. Specify the confidence level: Specify the desired confidence level, such as 95%.
4. Calculate the critical values: Calculate the critical values from the distribution of the pivot that correspond to the desired confidence level.
5. Construct the confidence interval: Use the pivot and the critical values to construct the confidence interval.

Example 1: Constructing a Confidence Interval for a Population Mean

Suppose we want to construct a 95% confidence interval for a population mean, μ, based on a sample of size n = 25.

1. Choose a pivot: We choose the sample mean, x̄, as the pivot.
2. Determine the distribution of the pivot: The sample mean has a normal distribution with mean μ and standard deviation σ/√n.
3. Specify the confidence level: We specify a confidence level of 95%.
4. Calculate the critical values: We calculate the critical values from the standard normal distribution, which are z = ±1.96.
5. Construct the confidence interval: We construct the confidence interval as:

$$\bar{x} - 1.96 * (\sigma/\sqrt{n}) \leq \mu \leq \bar{x} + 1.96 * (\sigma/\sqrt{n})$$

This confidence interval provides a range of values within which we expect the true population mean to lie with 95% confidence.

Advantages of using Pivots

Using pivots to construct confidence intervals has several advantages:

- Flexibility: Pivots can be used to construct confidence intervals for a wide range of

parameters, including population means, proportions, and regression coefficients.
- Accuracy: Pivots provide accurate confidence intervals, even for small sample sizes.
- Simplicity: Pivots are often easy to calculate and interpret.

Common Pivots

Some common pivots used in confidence interval construction include:
- Sample mean: Used to construct confidence intervals for population means.
- Sample proportion: Used to construct confidence intervals for population proportions.
- Regression coefficient: Used to construct confidence intervals for regression coefficients.

In conclusion, pivots provide a flexible and accurate method for constructing confidence intervals. By choosing an appropriate pivot and following the steps outlined above, you can construct confidence intervals for a wide range of parameters.

Suppose $T = T(X)$ is a statistics the function $g(T,\theta)$ is random variance and its distribution is independent of θ for example t-distribution and χ^2-distribution then the statistic $g(T,\theta)$ is called pivotal quantity (or) pivotal statistics.

$$P_\theta\left[g(T,\theta) \geq a\right] = \alpha_1 \forall \theta$$

$$P_\theta\left[g(T,\theta) \geq b\right] = \alpha_2 \forall \theta$$

So that

$$P_\theta\left[a \leq g(T,\theta) \geq b\right] = 1-\alpha \ \forall \theta$$

Suppose that the pivotal quantity $g(T,\theta)$ can be invited, such that

$$a \leq g(T,\theta) \geq b \Leftrightarrow C(T) \leq \theta \leq d(T)$$

Where a,b are two quantities $a \,\&\, b,$ which may depend on 'α', but independent of θ, ∋ where c,d are independent of θ. then,

$$P_\theta\left[C(T) \leq \theta \leq d(T)\right] = P_\theta\left[a \leq g(T,\theta) \geq b\right]$$

$$= 1-\alpha \ \forall \theta$$

Then the random interval (c,d) may be taken as $(1-\alpha)$ confidence interval for 'θ'.

10.6.1 EXAMPLE :

Problem : The sample $\bar{X} = (x_1, x_2, ..., x_n)$ from $N(\mu, \sigma^2)$ were both μ and σ^2 are unknown. Find the shortest confidence interval for μ using a pivotal statistic.

Answer: $\bar{X} \sim N(\mu, \sigma^2)$

Where 'σ^2' is unknown

The t-distribution is defined by $\dfrac{\bar{X}-\mu}{S/\sqrt{n}}$ (1)

Where $S^2 = \dfrac{1}{n-1}\sum\left(X_i - \bar{X}\right)^2$ is independent of μ and σ.

Here t-distribution is statistic may serve as a pivotal statistic

We have $P_u\left[t_{(n-1)},\ \alpha_1 \le \frac{\bar{X}-\mu}{S/\sqrt{n}} \le t_{(n-1)},\ \alpha_2\right] = 1-\alpha\ \forall\mu$ (2)

Where $\alpha = \alpha_1 + \alpha_2$

$$P_u\left[\bar{x} - t_1\frac{S}{\sqrt{n}} \le \mu \le \bar{x} + t_2\frac{S}{\sqrt{n}}\right] = 1-\alpha\ \forall\mu \quad (3)$$

Where

$$t_1 = t_{(n-1)}\alpha_1$$

$$t_2 = t_{(n-1)(1-\alpha_2)}$$

The confidence interval is

$$\left[\bar{x} - t\alpha_1, \bar{x} + t_{(1-\alpha_2)}\right] \quad (4)$$

In the above they have different values of α_1 and α_2 such that $\alpha_1 + \alpha_2 = \alpha$.

We shall get different confidence interval for given 'α'. But we may choose one with minimum length

$$L = \frac{S}{\sqrt{n}}\left(t_2 - t_1\right) \quad (5)$$

We are to minimize 't' w.r.t. t_1 & t_2 ∍

$$\int_{t_1}^{t_2} f\left(t\right)dt = 1-\alpha \quad (6)$$

Here $f\left(t\right)$ is the p.d.f. of 't'.

$\therefore t$ is symmetrical above origin. So minimize 'L' t_1 & t_2 should be symmetrical placed about origin.

from equation (6) 't_2' can be expressed as function 't_1'

differentiation w.r.t. 't_1' we get

$$f\left(t_2\right)\frac{dt_2}{dt_1} = f\left(t_1\right) = 0 \quad (7)$$

from liebnitz formula

$$\left[\frac{d}{dt}\int_{a(t)}^{b(t)} f\left(t\right)\,dt = f\left(b\right)\frac{db}{dt} - f\left(a\right)\frac{da}{dt}\right]$$

$$= \int_{a(t)}^{b(t)} f\left(t\right)\,dt$$

To minimize w.r.t. 't_1'

from equation (5) we have

$$\frac{d}{dt_1} = \frac{S}{\sqrt{n}}\left(\frac{dt_2}{dt_1} - 1\right) = 0 \quad (8)$$

Substituting the values of $\frac{dt_2}{dt_1}$ from (7) and (8)

We get $\frac{S}{\sqrt{n}}\left(\frac{f(t_1)}{f(t_2)}-1\right)=0$ (9)

$$\Rightarrow f(t_1)=f(t_2) \quad (10)$$

$i.e., t_1 = t_2$ Distribution of t is symmetrical about origin.

10.7 KEY WORDS:

- **Estimation** – The process of inferring the value of a population parameter based on sample data.
- **Point Estimate** – A single value used to estimate a population parameter.
- **Interval Estimate** – A range of values within which the parameter is expected to lie.
- **Parameter** – A numerical characteristic of a population (e.g., population mean, population proportion).
- **Statistic** – A numerical measure calculated from a sample.
- **Confidence Interval (CI)** – An interval estimate, derived from sample statistics, that is likely to contain the true population parameter.
- **Margin of Error** – The maximum expected difference between the point estimate and the actual population parameter.
- **Standard Error (SE)** – The standard deviation of the sampling distribution of a statistic.
- **Confidence Level** – The probability (usually expressed as a percentage, e.g., 95%) that the confidence interval contains the true population parameter.
- **Critical Value** – A point on the distribution that corresponds to the desired confidence level (e.g., Z or t value).
- **Alpha (α)** – The probability of error or the level of significance (α = 1 - confidence level).
- **Pivot (Pivot Quantity)** – A function of the sample data and unknown parameters that has a known distribution and is used to construct confidence intervals.
- **Z-Distribution** – The standard normal distribution used when population variance is known.
- **t-Distribution** – A distribution used when the population variance is unknown and sample size is small.
- **Degrees of Freedom (df)** – The number of independent values in a sample used to estimate a parameter.

10.8 SUMMARY:

In this lesson, we explored the essential concepts of interval estimation, a core part of statistical inference. Starting with an understanding of estimation methods, we distinguished between point and interval estimates, recognizing the value of providing a range that likely contains the true population parameter.

We then introduced the concept of the **confidence level**, which reflects the degree of certainty associated with an interval estimate. Higher confidence levels lead to wider intervals, illustrating the trade-off between precision and confidence.

Finally, we examined how **confidence intervals are constructed using pivotal quantities** - a technique that leverages known distributions (such as the Z and t distributions) to create reliable intervals for estimating population parameters.

By mastering these concepts, you gain the tools to interpret and construct confidence intervals effectively, making informed decisions based on sample data while acknowledging inherent uncertainty.

10.9 SELF ASSESSMENT QUESTIONS:

1. What is the difference between point estimation and interval estimation?
2. Why is interval estimation preferred over point estimation in many real-world situations?
3. What are some examples of parameters that might be estimated using interval estimation?
4. Define a confidence interval and its main components.
5. What happens to the confidence interval when the confidence level increases?
6. What is a pivot (or pivotal quantity), and why is it important in constructing confidence intervals?
7. When do you use a Z-distribution versus a t-distribution in interval estimation?
8. How do degrees of freedom impact the t-distribution used in confidence intervals?

10.10 SUGGESTED READINGS:

1. Kale, B. K. (1999). A First Course on Parametric Inference, Narosa Publishing House, New Delhi.
2. Lehman, E. L., and Cassella, G. (1998). Theory of Point Estimation, Second Edition, Springer, NY.
3. An Introduction to Probability and Statistics by V.K. Rohatgi and.K.Md.E.Saleh(2001).
4. Goon, A. M., Gupta, M. K., and Dasgupta, B. (1989). An Outline of Statistical Theory-Vol.II, World Press, Calcutta.
5. Linear Statistical Inference and its Application by C.R. RAO (1973), John Wiley.

Prof. G. V. S. R. Anjaneyulu

LESSON - 11
SHORTEST EXPECTED LENGTH, UMA, UMAU CONFIDENCE SETS

OBJECTIVES:

After studying the lesson the students will able to:

- Understand the concept of confidence sets and their role in parameter estimation.
- Define and construct confidence sets with the shortest expected length, and explain why minimizing expected length is desirable for precision in estimation.
- Identify and interpret Uniformly Most Accurate (UMA) confidence sets, recognizing their optimality in terms of maximum coverage probability across all parameter values.
- Understand the need for unbiasedness in inference and construct Uniformly Most Accurate Unbiased (UMAU) confidence sets.
- Explain the dual relationship between confidence sets and hypothesis testing, and use this relationship to derive one from the other.
- Apply these concepts through examples and exercises, gaining practical skills in constructing and interpreting optimal confidence sets in various statistical settings.

STRUCTURE:

11.1 INTRODUCTION:

In statistical inference, one of the fundamental goals is to estimate unknown parameters of a population using data from a sample. Confidence sets, particularly confidence intervals in one-dimensional settings, are a primary tool for this task. A confidence set provides a range of plausible values for an unknown parameter with a specified probability, known as the confidence level.

This chapter focuses on the theory and methods behind constructing *optimal* confidence sets. Among various criteria, one important goal is to find confidence sets with the **shortest expected length**, thereby offering more precise inference without sacrificing coverage probability. This is particularly relevant when minimizing estimation uncertainty is crucial.

Beyond expected length, statistical theory also seeks confidence sets that are **uniformly most accurate (UMA)**—sets that maximize the probability of containing the true parameter value across all possible values of the parameter. In some cases, **unbiasedness** is also a requirement, leading to **uniformly most accurate unbiased (UMAU)** confidence sets. These UMAU sets ensure not only high accuracy but also that the probability of covering false values does not exceed that of covering the true value, providing a balanced approach in estimation.

An important conceptual link exists between **confidence estimation and hypothesis testing**. Every confidence set corresponds to a family of hypothesis tests and vice versa. This duality allows the properties of one to inform the design of the other. For example, a test that minimizes the probability of type I and II errors can guide the construction of a confidence set with desirable coverage and precision characteristics.

In the following sections, we delve into the construction and properties of these optimal confidence sets, illustrated with practical examples, and explore their deep connection to hypothesis testing procedures.

11.2 SHORTEST EXPECTED LENGTH:

For any fixed $(1-\alpha)$ there are many possible pairs of numbers q1 and q2 that can be selected so that $P(q_1 < \theta < q_2) = 1-\alpha$. Different pairs of q_1 and q_2 will produce t_1 and t_2. We should want to select the pair of q_1 and q_2 that will make t_1 and t_2 close together in some sense. For instance, if $\mathrm{t}_2(X)-\mathrm{t}_1(X)$ which is the length of the confidence interval, is not random, then we might select that pair of q_1 and q_2 that makes the length of the interval smallest; or if the length of the confidence interval is random, then we might select that pair q_1 and q_2 that makes the average length of the interval smallest.

Thus a shortest length confidence (SLCI) interval is the confidence interval for which $\mathrm{t}_2(x)-\mathrm{t}_1(x)$ is minimized and a shortest expected length confidence interval (SELCI) is the confidence interval for which $E_\theta\{T_2(X)-T_1(X)\}$ is minimized $\forall\theta\in\Theta$.

11.2.1 EXAMPLES:

Example 1: Let $X_1, X_2, \ldots\ldots, X_n \sim N(\mu,1)$ To find shortest length confidence interval for μ, based on the pivot $Q(X,\mu)=\sqrt{n}(\bar{X}-\mu)$. We know that $Q(X,\mu)\sim N(0,1)$ for any μ.

Answer: $P_u\left[q_1 \le Q(X,\mu) \le q_2\right] = 1-\alpha$

or, $P_u\left[q_1 \le \sqrt{n}(\bar{X}-\mu) \le q_2\right] = 1-\alpha$

or, $P_u\left[\bar{X}-\frac{q_1}{\sqrt{n}}\le\mu\le\bar{X}+\frac{q_1}{\sqrt{n}}\right]=1-\alpha$

Therefore, $t_1(x)=\bar{X}-\frac{q_1}{\sqrt{n}}$ and $t_2(x)=\bar{X}+\frac{q_1}{\sqrt{n}}$, hence $T_2(x)-T_1(x)=(q_2-q_1)\frac{1}{\sqrt{n}}$.

is not random variable. Our objective is to minimize q_2-q_1 subject to,

$$P[q_1\le Z\le q_2]=1-\alpha, Z\sim N(0,1).$$

Let us define,

$$\psi(q_1,q_2)=(q_2-q_1)+\lambda\left[F_Z(q_2)-F_Z(q_1)-(1-\alpha)\right]$$

Hence, by differentiating $\psi(q_1,q_2)$ w.r.t. q_1 and q_2 we get the following set of equations, and solving for 0 gives us the shortest length confidence interval.

$$\frac{\partial}{\partial q_1}\psi(q_1,q_2)=-1-\lambda f_Z(q_1)=0$$

and,

$$\frac{\partial}{\partial q_1}\psi(q_1,q_2)=1+\lambda f_Z(q_2)=0$$

$$\Leftrightarrow f_Z(q_1)=f_Z(q_2)$$

Therefore, $q_2=-q_1$ and hence $q_2=\tau\frac{\alpha}{2}$, where $\tau\frac{\alpha}{2}$ does not upper $100\frac{\alpha}{2}$ point of $N(0,1)$ distribution.

$$SLCI:\left[\bar{x}-\frac{\tau\frac{\alpha}{2}}{\sqrt{n}},\bar{x}+\frac{\tau\frac{\alpha}{2}}{\sqrt{n}}\right].$$

Example 2: Let $X=X_1,X_2,.......,X_n\sim N(\mu,\sigma^2)$. To find shortest length confidence interval for μ, based on the pivot $Q(X,\mu)=\frac{\sqrt{n}(\bar{X}-\mu)}{s}$. We know that $Q(X,\mu)\sim t_{n-1}$ for any μ,σ.

Answer: $$P_{\mu,\sigma}\left[q_1\le Q(X,\mu)\le q_2\right]=1-\alpha$$

or, $P_{\mu,\sigma}\left[q_1\le\frac{\sqrt{n}(\bar{X}-\mu)}{s}\le q_2\right]=1-\alpha$

or, $P_{\mu,\sigma}\left[\bar{X}-\frac{q_1 s}{\sqrt{n}}\le\mu\le\bar{X}+\frac{q_1 s}{\sqrt{n}}\right]=1-\alpha$

Therefore, $t_1(x)=\bar{X}-\frac{q_1 s}{\sqrt{n}}$ and $t_2(x)=\bar{X}+\frac{q_2 s}{\sqrt{n}}$, hence $t_2(x)-t_1(x)=(q_2-q_1)\frac{s}{\sqrt{n}}$.

Our objective is to minimize q_2-q_1 subject to,

$$P[q_1\le t_{n-1}\le q_2]=1-\alpha$$

Let us define,

$$\psi(q_1,q_2)=(q_2-q_1)+\lambda\left[F_{t_{n-1}}(q_2)-F_{t_{n-1}}(q_1)-(1-\alpha)\right]$$

Hence, by differentiating $\psi(q_1,q_2)$ w.r.t. q_1 and q_2 we get the following set of equations, and solving for 0 gives us the shortest length confidence interval.

$$\frac{\partial}{\partial q_1}\psi(q_1,q_2) = -1-\lambda f_{t_{n-1}}(q_1) = 0$$

and,

$$\frac{\partial}{\partial q_1}\psi(q_1,q_2) = 1+\lambda f_{t_{n-1}}(q_2) = 0$$

$$\Leftrightarrow f_{t_{n-1}}(q_1) = f_{t_{n-1}}(q_2)$$

Therefore, $q_2 = -q_1$ and hence $q_2 = t_{\alpha/2,n-1}$ and,

$$SLCI:\left[\bar{X} - t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}, \bar{X} + t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}\right].$$

11.3 UNIFORMLY MOST ACCURATE (UMA) CONFIDENCE SETS:

Uniformly Most Accurate (UMA) Confidence Sets are a concept in statistical inference related to constructing confidence intervals or regions that have optimal accuracy properties across a class of probability distributions. Here's a structured breakdown to explain the concept clearly:

Definition: A Uniformly Most Accurate (UMA) confidence set is a confidence set (interval, region, etc.) that, among all confidence sets with the same confidence level (e.g., 95%), has the smallest probability of covering incorrect values (i.e., the highest precision or power) uniformly over all possible true parameter values.

In other words, a UMA confidence set for a parameter θ satisfies:

- For a given confidence level $1-\alpha,$ the confidence set has at least $1-\alpha$ coverage probability.
- Among all such sets, it has the minimum probability of including incorrect values (e.g., smallest expected volume or most power to reject false values of θ) uniformly over all $\theta \in \Theta$.

11.3.1 EXAMPLES:

1. UMA Confidence Interval for a Normal Mean (Known Variance)

Suppose $X_1, X_2, ..., X_n$ are i.i.d. samples from $N(\mu, \sigma^2)$, where σ^2 is known, The Uniformly Most Accurate (UMA) confidence interval for μ at level $1-\alpha,$ is given by

$$\left(\bar{X} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

Where $Z_{\alpha/2}$ is the critical value from the standard normal distribution. This interval is UMA due to the sufficiency and completeness of $\bar{X}$ as an estimator of μ in the normal case.

2. UMA Confidence Interval for the Mean of Normal Distribution (UnKnown Variance)

If $X_1, X_2, ..., X_n \sim N(\mu, \sigma^2)$, but σ^2 is unknown, then the UMA confidence interval for μ is the Student's t-interval**:**

$$\left(\bar{X}-t_{\alpha/2,n-1}\frac{S}{\sqrt{n}},\bar{X}+t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}\right)$$

where S is the sample standard deviation and $t_{\alpha/2,n-1}$ is the critical value from the Student's t-interval.

3. UMA Confidence Interval for the Variance of Normal Distribution

If $X_1, X_2, ..., X_n \sim N(\mu, \sigma^2)$, but σ^2 is unknown, then the UMA confidence interval for σ^2 is based on the chi-square distribution:

$$\left(\frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}},\frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}}\right)$$

where $\chi^2_{\alpha/2,n-1}$ and $\chi^2_{1-\alpha/2,n-1}$ are the critical values of the chi-square distribution with $n-1$ degrees of freedom.

11.4 UNIFORMLY MOST ACCURATE UNBIASED (UMAU) CONFIDENCE SETS:

Definition: A family $\{S(\underset{\sim}{x})\}$ of confidence sets for a parameter θ is said to be unbiased level $1-\alpha$, if

$$P_\theta\left[S(\underset{\sim}{X}) \ni \theta\right] \geq 1-\alpha \ \forall \theta \in \Theta$$

and $P_{\theta'}\left[S(\underset{\sim}{X}) \ni \theta\right] \leq 1-\alpha \ \forall \theta' \in \Theta$

If $S^*(\underset{\sim}{X})$ is a family of $(1-\alpha)$ unbiased confidence sets that minimizes $P_\theta\left[S(\underset{\sim}{X}) \ni \theta'\right] \forall \theta, \theta' \in \Theta$.

Then, $S^*(\underset{\sim}{X})$ is called Uniformly Most Accurate Unbiased (UMAU) family of confidence sets at level $(1-\alpha)$.

Discuss by theorem the relationship between UMP unbiased size -α acceptance region and UMAU family of confidence set at let $1-\alpha$.

Theorem: Consider the testing problem $H_0: \theta = \theta_0 \ Vs \ H_1: \theta \neq \theta_0$ for each $\theta_0 \in \Theta$. Let $A(\theta_0)$ be the UMP unbiased size α acceptance region for this problem. Then $S(\underset{\sim}{x}) = \{\theta / \underset{\sim}{x} \in A(\theta)\}$ is a UMP unbiased family of confidence sets at level $(1-\alpha)$.

Proof: Let the UMP unbiased size -α test be given by $\phi_0(\underset{\sim}{x})$.

Unbaisedness gives $E_{\theta'\phi_0}(\underset{\sim}{X}) \geq \forall \theta' \in H_1(\theta_0)$

$$\Rightarrow E_{\theta'}\left(1-\phi_0(\underset{\sim}{X})\right) \leq 1-\alpha \ \forall \theta' \in H_1(\theta_0)$$

$$\Rightarrow P_{\theta'}\left(\underset{\sim}{x} \in A(\theta)\right) \leq 1-\alpha$$

$$\Rightarrow P_{\theta'}\left(S(\underset{\sim}{X}) \ni \theta\right) \leq 1-\alpha.$$

Show that $S\left(\underset{\sim}{X}\right)$ is unbiased.

Next, consider any other unbiased size $-\alpha$ test $\phi^*\left(\underset{\sim}{X}\right)$, with acceptance region $A^*(\theta)$; we get a corresponding $(1-\alpha)$ level family of unbiased confidence sets $S^*\left(\underset{\sim}{X}\right)$, *i.e.*,

$$P_{\theta'}\left[S^*\left(\underset{\sim}{X}\right) \ni \theta\right] \leq 1-\alpha \;\; \forall \theta' \in H_1(\theta_0)$$

The test $\phi_0\left(\underset{\sim}{X}\right)$ has been given to be UMP, therefore

$$E_{\theta'}\left(\phi_0\left(\underset{\sim}{X}\right)\right) \geq E_{\theta'}\left(\phi_0^*\left(\underset{\sim}{X}\right)\right) \forall \theta' \in H_1(\theta_0)$$

or,
$$E_{\theta'}\left(1-\phi_0\left(\underset{\sim}{X}\right)\right) \leq E_{\theta'}\left(1-\phi^*\left(\underset{\sim}{X}\right)\right)$$

or,
$$P_{\theta'}\left(\underset{\sim}{x} \in A(\theta)\right) \leq P_{\theta'}\left(\underset{\sim}{x} \in A^*(\theta)\right)$$

or,
$$P_{\theta'}\left(S\left(\underset{\sim}{X}\right) \ni \theta\right) \leq P_{\theta'}\left(S^*\left(\underset{\sim}{X}\right) \ni \theta\right) \forall \theta' \in H_1(\theta_0)$$

This follows $S\left(\underset{\sim}{X}\right)$ is UMA Unbiased family of confidence sets at level $(1-\alpha)$.

11.4.1 EXAMPLES:

Example 1: Let $X_1, X_2, ..., X_n \sim N(\mu, \sigma^2)$, where σ^2 is known for testing $H_0 : \mu = \mu_0$ *against* $H_1 : \mu \neq \mu_0$. Find a UMA$(1-\alpha)$ level confidence sets for μ.

Solution: For testing of hypothesis $H_0 : \mu = \mu_0$ *against* $H_1 : \mu \neq \mu_0$
α

The UMP unbiased size $-\alpha$ test is given by

$$\phi\left(\underset{\sim}{x}\right) = \begin{cases} 1 & \text{if } \dfrac{\sqrt{\text{n}}\left|(\bar{x}-\mu_0)\right|}{\sigma} > C \\ 0 & \text{otherwise} \end{cases}$$

The test is known as Z=test. The constant C can be determined by the size condition

$$E_{\mu_0}\left[\phi\left(\underset{\sim}{X}\right)\right] = \alpha$$

or,
$$P_{\mu_0}\left\{\frac{\sqrt{\text{n}}\left|(\bar{x}-\mu_0)\right|}{\sigma} > C\right\} = \alpha/2$$

which gives $C = Z_{\alpha/2}$.

Thus, the acceptance region corresponding to this UMP unbiased size - α - test is given by

$$A(\mu_0) = \left\{\underset{\sim}{X} : \frac{\sqrt{\text{n}}\left|(\bar{x}-\mu_0)\right|}{\sigma} \leq Z_{\alpha/2}\right\}$$

By the above theorem, the UMA unbiased family of confidence sets $S\left(\underset{\sim}{X}\right)$ at level is finally given by

$$S(\underset{\sim}{x}) = \{\mu : \underset{\sim}{x} \in A(\mu)\}$$

$$= \left\{-Z_{\alpha/2} \frac{\sqrt{n}\,(\overline{x}-\mu)}{\sigma} \leq Z_{\alpha/2}\right\}$$

$$= \left\{-\frac{\sigma}{\sqrt{n}} Z_{\alpha/2} \leq (\mu - \overline{x}) \leq \frac{\sigma}{\sqrt{n}} Z_{\alpha/2}\right\}$$

$$= \left\{\overline{x} - \frac{\sigma}{\sqrt{n}} Z_{\alpha/2} \leq (\mu \leq \overline{x}) \leq \frac{\sigma}{\sqrt{n}} Z_{\alpha/2}\right\}$$

Example 2: Let X be a random variable with the density $f_X(x/\theta) = \begin{cases} \frac{1}{\theta}.e^{-x/\theta}; & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}$

$H_0 : \mu = \mu_0\, against\; H_1 : \mu \neq \mu_0$. Find a UMA$(1-\alpha)$ level confidence sets for μ.
where $\theta > 0$. Consider the testing problem $H_0 : \theta = \theta_0\; Vs\; H_1 : \theta \neq \theta_0$.
Find out a UMA $(1-\alpha)$ level family of confidence sets corresponding to size - α UMP test.

Solution: The UMP size - α acceptance region is given by

$$A(\theta) = \{x : T(x) \geq C(\theta)\}$$

$$A(\theta_0) = \{x : x \geq C(\theta)\}$$

Where, we choose $C(\theta)$ by

$$P_{\theta_0} A(\theta_0) = 1 - \alpha$$

or, $\displaystyle\int_0^{C(\theta_0)} \frac{1}{\theta}.\, e^{-x/\theta_0} dx = \alpha$

or, $\displaystyle\frac{1}{\theta_0}\left[-\theta_0\; e^{-x/\theta_0}\right]_0^{C(\theta_0)} = \alpha$

$\Rightarrow e^{-C(\theta_0)/\theta_0} + 1 = \alpha$

$\Rightarrow C(\theta_0) = \theta_0 . \log\dfrac{1}{1-\alpha}, 0 < \alpha < 1.$

Therefore, for corresponding UMA family of $(1-\alpha)$ level of confidence sets is given by

$$S(x) = \{\theta : x \in A(\theta)\} = \left\{\theta : x \geq \theta \log\frac{1}{1-\alpha}\right\}$$

$$S(x) = \left\{\theta : \theta \leq \frac{x}{\log\frac{1}{1-\alpha}}\right\} = \left(0, \frac{x}{\log\left(\frac{1}{1-\alpha}\right)}\right) (\sin ce\;\; \theta > 0)$$

11.5 RELATIONSHIP BETWEEN CONFIDENCE ESTIMATION AND TESTING OF HYPOTHESIS:

There's a **deep duality** between these two:

- A $(1-\alpha)$ confidence interval for a parameter can be thought of as the set of parameter values that are **not rejected** by a corresponding **level-α hypothesis test**.
- Conversely, a hypothesis test defines a rejection region, and the complement (the set of values not rejected) can be used to construct a confidence set.
- For example:
    - If you do a two-sided test at α = 0.05, the confidence interval consists of all parameter values that wouldn't be rejected by this test.
    - Confidence intervals are essentially **collections of plausible values** that pass hypothesis tests.

We investigate in this topic the relationship between confidence estimation and hypothesis testing. More precisely given a level of test $H_0:\theta=\theta_0$ is it possible to construct $\alpha(1-\alpha)$ level of confidence interval for θ_0 conversely. This is considered in the following theorems.

Theorem: For each $\theta\in\Theta$ let $A(\theta_0)$ be the acceptance region of an α - level test for testing $H_0:\theta=\theta_0$ against for each $x\in X$, let $s(x)$ denote the set of parameter values defined by $S(X)=\{\theta/x\in A(\theta),\theta\in\Theta\}$, the $S(X)$ is a family of confidence sets for θ at confidence level $(1-\alpha)$. If moreover, $A(\theta_0)$ is UMP for $H_0:\theta=\theta_0$ against $H_1:\theta\neq\theta_0$. Then $S(X)$ defined is for varying x over X, a family of UMA confidence sets of levels $(1-\alpha)$.

Proof: We have $S(X)\Rightarrow\theta$ iff $x\in A(\theta)$

Hence $P_\theta(S(X)\ni\theta)\leq P_\theta\{x\in A(\theta)\}\geq 1-\alpha$. Hence, $S(X)$ for varying $x\in X$, is a family of confidence sets for θ at confidence level $(1-\alpha)$. Let $S(X)$ be any other family of $(1-\alpha)$ level confidence sets. Let $A^*(\theta)=\{x:S*(x)\Rightarrow\theta\}$ then,

$$P_\theta\{x\in A^*(\theta)\}=P_\theta\{S^*(x)\ni\theta\}\geq 1-\alpha$$

Again, since $A(\theta_0)$ is UMP for $H_0:\theta=\theta_0$ against $H_1:\theta\neq\theta_0$;

It follows that

$$P_\theta\{x\in A^*(\theta_0)\}\geq P_\theta\{x\in A(\theta_0)\}$$
$$=P_\theta\{S(X)\ni\theta_0\}\forall\theta\neq\theta_0$$

$\therefore$ Hence the proof.

The above theorem is not of much use, since a UMP test against two sided hypothesis generally does not exist. The following theorem indicates row a family of most accurate unbiased confidence sets can be obtained from a family of UMP tests.

Theorem: Let $A(\theta_0)$ be the acceptance region for the level $(1-\alpha)$ UMP tests for $H_0(\theta=\theta_0)$ against $H_1(\theta\neq\theta_0)$ and $S(X)=\{\theta/x\in A_0(\theta_0)\}$. Then the sets $S_0(X)$ for varying, $x\in X$, are a family of UMAU confidence sets at the confidence levels.

Proof: Since $A(\theta_0)$ be the acceptance region of an unbiased test

$$P_\theta\{x\in A(\theta)\}\leq 1-\alpha, \forall\theta'\neq\theta$$

Hence $P_\theta'(S(X)\ni\theta)=P_\theta'\{x\in A(\theta)\}\leq 1-\alpha, \forall\theta'\neq\theta$

Then the sets $S(X)$ is a UMA, let $S^*(A)$ be any other unbiased $(1-\alpha)$ level family of UMAU confidence sets.

Let $A^*(\theta)=\{x\in S^*(x)\ni\theta\}$ then

$P_\theta'\{x\in A^*(\theta)\}.P_\theta'\{S^*(x)\ni\theta\}\leq 1-\alpha$ and therefore, $A^*(\theta)$ is the acceptance region of an unbiased size α - level test.

Hence,

$$P_\theta'\{S^*(x)\ni\theta\}=P_\theta'\{x\in A^*(\theta)\}\geq P_\theta'\{x\in A(\theta)\}$$
$$=P_\theta'\{S^*(x)\ni\theta\}$$

The inequality holds, since $A(\theta)$ be the acceptance region of an UMA unbiased test

$\therefore$ Hence the proof.

11.5.1 EXAMPLE:

Example : Let $X_1,X_2,\ldots,X_n\sim N(\mu,\sigma^2)$, suppose we are interested in the following testing hypothesis.

Solution: For testing of hypothesis $H_0:\mu=\mu_0 \; vs \; H_1:\mu\neq\mu_0$

We know that, we reject H_0 at level α if $\frac{\sqrt{n}|(\bar{x}-\mu_0)|}{s}>t_{\alpha/2,n-1}$. Therefore,

$$R(\mu_0)=\left\{x:\frac{\sqrt{n}\left(|(\bar{x}-\mu_0)|\right)}{s}>t_{\alpha/2,n-1}\right\}$$

$$R(\mu)=\left\{x:\frac{\sqrt{n}\left(|(\bar{x}-\mu)|\right)}{s}>t_{\alpha/2,n-1}\right\}$$

$$A(\mu)=\left\{x:\frac{\sqrt{n}\left(|(\bar{x}-\mu)|\right)}{s}>t_{\alpha/2,n-1}\right\}=\{x:\mu\in S(x)\},$$

where $S(x)=\{\mu:x\in A(\mu)\}$ Hence,

$$A(\mu)=\left\{x:\bar{x}-\frac{x}{\sqrt{n}}t_{\alpha/2,n-1}\leq\mu\leq\bar{x}+\frac{x}{\sqrt{n}}t_{\alpha/2,n-1}\right\},$$

and

$$S(x)=\left[\bar{x}-\frac{x}{\sqrt{n}}t_{\alpha/2,n-1}\leq\mu\leq\bar{x}+\frac{x}{\sqrt{n}}t_{\alpha/2,n-1}\right].$$

11.6 KEY WORDS:

- Confidence Set
- Confidence Interval
- Expected Length
- Shortest Expected Length Confidence Set
- Uniformly Most Accurate (UMA)
- Uniformly Most Accurate Unbiased (UMAU)
- Unbiasedness
- Hypothesis Testing

11.7 SUMMARY:

Confidence sets and hypothesis testing are two sides of the same inferential coin. UMA and UMAU confidence sets aim to optimize precision and reliability in estimation, paralleling the goals of optimal tests in hypothesis testing. Understanding the relationship between these concepts helps in developing statistical procedures that are both accurate and interpretable, especially in settings requiring stringent control over uncertainty and error.

11.8 SELF ASSESSMENT QUESTIONS:

1. What does it mean for a confidence interval to have the shortest expected length?
2. Why is minimizing the expected length of a confidence interval desirable in estimation?
3. What is the difference between a UMA and UMAU confidence set?
4. How is the concept of unbiasedness defined in the context of confidence sets?
5. Define UMA and UMAU confidence sets. Describe the relationship between Confidence estimation and testing of hypothesis.
6. How can a UMPU test be used to construct a UMAU confidence interval?
7. What is the interpretation of a confidence interval in terms of hypothesis testing?
8. Distinguish between the problems of point estimation and interval estimation.

11.9 SUGGESTED READINGS:

1. Goon, A. M., Gupta, M. K., and Dasgupta, B. (1989). An Outline of Statistical Theory-Vol.II, World Press, Calcutta.
2. Kale, B. K. (1999). A First Course on Parametric Inference, Narosa Publishing House, New Delhi.
3. Lehman, E. L., and Cassella, G. (1998). Theory of Point Estimation, Second Edition, Springer, NY.
4. Casella, G. & Berger, R.L. (2013). Statistical Inference, 2nd Edition. Cengage Learning

Prof. G. V. S. R. Anjaneyulu

LESSON- 12

PRIORI AND POSTERIORI DISTRIBUTIONS, LOSS FUNCTION, RISK FUNCTION, MINMAX & BAYES ESTIMATOR

OBJECTIVES:

By the end of this course, students will be able to:

- Explain the concepts of prior, likelihood, and posterior distributions in Bayesian inference.
- Apply Bayes' Theorem to update beliefs based on observed data.
- Define various types of loss functions (e.g., quadratic loss, 0-1 loss, absolute loss).
- Understand how loss functions influence decision-making in statistical inference.
- Define the risk function as the expected loss under a statistical decision rule.
- Analyze the performance of estimators and decision rules using risk functions.
- Understand and compute Bayes estimators under different priors and loss functions.
- Define the minimax criterion and identify minimax estimators for given statistical models.
- Contrast frequentist and Bayesian approaches to estimation and decision-making.
- Evaluate the trade-offs between different estimation methods in terms of risk and robustness.
- Use decision theory to formulate and solve problems in statistical estimation and hypothesis testing.
- Interpret results in context and justify chosen estimation strategies based on theoretical foundations.

STRUCTURE:

12.1 INTRODUCTION:

In statistics, especially Bayesian inference, the concepts of priori and posteriori distributions play a crucial role in updating beliefs based on evidence.

- **Prior Distribution (a priori)**: This represents our initial beliefs or assumptions about a parameter before observing any data. It is based on previous knowledge, expert opinion, or theoretical considerations. For example, if we are estimating the probability of rain tomorrow, our prior might be based on long-term weather patterns.
- **Posterior Distribution (a posteriori)**: This is the updated distribution of the parameter after observing new data. It combines the prior distribution with the likelihood of the observed data using **Bayes' Theorem**. The posterior distribution reflects our revised beliefs about the parameter.

Bayes' Theorem, which forms the foundation of this process, is expressed as:

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior}$$

This means the posterior distribution is proportional to the product of the likelihood (how likely the observed data is, given the parameter) and the prior.

These distributions allow us to incorporate both prior knowledge and new evidence, making Bayesian methods particularly powerful for learning and decision-making under uncertainty.

12.1 PRIORI AND POSTERIORI DISTRIBUTIONS:

Recall that if X, Y are two random variables having joint PDF or PMF $f_{X,Y}(x, y)$, then the marginal distribution of X is given by the PDF

$$f_X(x) = \int f_{X,Y}(x, y)\,dy$$

in the continuous case and by the PMF

$$f_X(x) = \sum_y f_{X,Y}(x, y)$$

in the discrete case; this describes the probability distribution of X alone. The conditional distribution of Y given X = x is denoted by the PDF or PMF

$$f_{Y/X}(y/x) = \frac{f_{X,Y}(x, y)}{f_X(x)},$$

and represents the probability distribution of Y if it is known that $X = x$. (This is a PDF or PMF as a function of y, for any fixed x.) Defining similarly the marginal distribution $f_Y(y)$ of Y and the conditional distribution $f_{X/Y}(x/y)$ of X given $Y = y$, the joint PDF $f_{X,Y}(x, y)$ factors in two ways as

$$f_{X,Y}(x, y) = f_{Y/X}(y/x)\, f_X(x) = f_{X/Y}(x/y)\, f_Y(y)$$

In Bayesian analysis, before data is observed, the unknown parameter is modeled as a random variable Θ having a probability distribution $f_\Theta(\theta)$, called the **prior distribution**.

This distribution represents our prior belief about the value of this parameter. Conditional on $\Theta = \theta$, the observed data X is assumed to have distribution $f_{X/\Theta}(x/\theta)$, where $f_{X/\Theta}(x/\theta)$ defines a parametric model with parameter θ. The joint distribution Θ of and X is then the product

$$f_{X,\Theta}(x, \theta) = f_{X/\Theta}(x/\theta)\, f_\Theta(\theta)$$

and the marginal X distribution of (in the continuous case) is

$$f_X(x) = \int f_{X,\Theta}(x,\theta)\,d\theta = f_{X/\Theta}(x/\theta)\,f_\Theta(\theta)\,d\theta.$$

The conditional distribution of Θ given $X = x$ is

$$f_{\Theta/X}(\theta/x) = \frac{f_{X,\Theta}(x,\theta)}{f_X(x)} = \frac{f_{X/\Theta}(x/\theta)\,f_\Theta(\theta)}{\int f_{X/\Theta}(x/\theta')\,f_\Theta(\theta')\,d\theta'}. \quad \text{.......... (1)}$$

This is called the **posterior distribution** of Θ: It represents our knowledge about the parameter Θ after having observed the data X. We often summarize the preceding equation simply as

$$f_{\Theta/X}(\theta/x) \propto f_{X/\Theta}(x/\theta)\,f_\Theta(\theta) \quad \text{.......... (2)}$$

Posterior density $\propto$ Likelihood X Prior density

where the symbol $\propto$ hides the proportionality factor $f_{X/\Theta}(x/\theta')\,f_\Theta(\theta')\,d\theta'$ which does not depend on θ.

12.2.1 EXAMPLES:

Example 1: Let $P \in (0,1)$ be the probability of heads for a biased coin, and let $X_1, X_2, ..., X_n$ be the outcomes of n tosses of this coin. If we do not have any prior information about P, we might choose for its **prior distribution** Uniform (0,1), having PDF $f_P(p) = 1$ *for* all $p \in (0,1)$. Given $P = p$, we model $X_1, X_2, ..., X_n \overset{IID}{\sim}$ Bernoulli(p). Then the joint distribution of $P, X_1, X_2, ..., X_n$ is given by

$$f_{X,P}(x_1, x_2, ..., x_n, p) = f_{X/P}(x_1, x_2, ..., x_n/p)\,f_P(p)$$

$$= \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} \times 1 = p^{\sum_{i=1}^{n} x_i}(1-p)^{n-\sum_{i=1}^{n} x_i}.$$

Let $s = x_1 + x_2 + ... + x_n$. The marginal distribution of $X_1, X_2, ..., X_n$ is obtained by integrating $f_{X,P}(x_1, x_2, ..., x_n, p)$ over p:

where $B(x,y)$ is the Beta function

$$B(x,y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}.$$

Hence the posterior distribution of P given $X_1 = x_1, ..., X_n = x_n$ has PDF

$$f_{P/X}(p/x_1, x_2, ..., x_n) = \frac{f_{X,P}(x_1, x_2, ..., x_n, p)}{f_X(x_1, x_2, ..., x_n)} = \frac{1}{B(s+1, n-s+1)}\,p^s(1-p)^{n-s}.$$

This is the PDF of the Beta$(s+1, n-s+1)$ distribution, so the posterior distribution of P given $X_1 = x_1, ..., X_n = x_n$ is Beta$(s+1, n-s+1)$, where $s = x_1 + x_2 + ... + x_n$.

We computed explicitly the marginal distribution $f_X(x_1, x_2, ..., x_n)$ above, but this was not necessary to arrive at the answer. Indeed, equation (2) gives

$$f_{P/X}\left(p/x_1,x_2,.......,x_n\right)\propto f_{X/P}\left(x_1,x_2,.......,x_n/p\right)f_P\left(p\right)=p^s\left(1-p\right)^{n-s}.$$

This tells us that the PDF of the posterior distribution of P is proportional to $p^s\left(1-p\right)^{n-s}$, as a function of p. Then it must be the PDF of the Beta$\left(s+1,n-s+1\right)$ distribution, and the proportionality constant must be whatever constant is required to make this PDF integrate to 1 over $p\in\left(0,1\right)$. We will repeatedly use this trick to simplify our calculations of posterior distributions.

Example 2: Suppose now we have a prior belief that P is close to $1/2$. There are various prior distributions that we can choose to encode this belief; it will turn out to be mathematically convenient to use the prior distribution $Beta\left(\alpha,\alpha\right)$, which has mean $1/2$ and variance $1/\left(8\alpha+4\right)$. The constant α may be chosen depending on how confident we are, a priori P, that is near $1/2$ choosing $\alpha=1$ reduces to the Uniform $\left(0,1\right)$ prior of the previous example, whereas choosing $\alpha>1$ yields a prior distribution more concentrated around $1/2$.

The prior distribution has $Beta\left(\alpha,\alpha\right)$ PDF $f_P\left(p\right)=\dfrac{1}{Beta\left(\alpha,\alpha\right)}p^{\alpha-1}\left(1-p\right)^{\alpha-1}$.

Then, applying equation (2), the posterior distribution of P given $X_1=x_1,.......,X_n=x_n$ has PDF

$$f_{P/X}\left(p/x_1,x_2,.......,x_n\right)\propto f_{X/P}\left(x_1,x_2,.......,x_n/p\right)f_P\left(p\right)$$
$$\propto p^s\left(1-p\right)^{n-s}\times p^{\alpha-1}\left(1-p\right)^{\alpha-1}=p^{s+\alpha-1}\left(1-p\right)^{n-s+\alpha-1}.$$

Where $s=x_1+x_2+.....+x_n$ as before, and where the symbol $\propto$ hides any proportionality constants that do not depend on p. This is proportional to the PDF of the distribution $Beta\left(s+\alpha,n-s+\alpha\right)$, so this Beta distribution is the posterior distribution of P.

In the previous example, the parametric form for the prior was (cleverly) chosen so that the posterior would be of the same form - they were both Beta distributions. This type of prior is called a **conjugate prior** for P in the Bernoulli model. Use of a conjugate prior is mostly for mathematical and computational convenience - in principle, any prior $f_P\left(p\right)$ on $\left(0,1\right)$.may be used. The resulting posterior distribution may be not be a simple named distribution with a closed-form PDF, but the PDF may be computed numerically from equation (1) by numerically evaluating the integral in the denominator of this equation.

Example 3: Let $A\in\left(0,\infty\right)$ be the parameter of the Poisson model $X_1,X_2,.......,X_n \overset{IID}{\sim} \text{Poisson}\left(\lambda\right)$. As a prior distribution for Λ, let us take the Gamma distribution $Gamma\left(\alpha,\beta\right)$. The prior and likelihood are given by

$$f_\Lambda\left(\lambda\right)=\frac{\beta^\alpha}{\Gamma\left(\alpha\right)}\lambda^{\alpha-1}e^{-\beta\lambda}$$

$$f_{X/\Lambda}\left(x_1,x_2,.......,x_n/\Lambda\right)=\prod_{i=1}^{n}\frac{\lambda^{x_i}e^{-\lambda}}{x_i\,!}.$$

Dropping proportionality constants that do not depend on λ, the posterior distribution of Λ given is then $X_1 = x_1, ..., X_n = x_n$ is then

$$f_{\Lambda/X}\left(\lambda/x_1, x_2, ..., x_n\right) \propto f_{X/\Lambda}\left(x_1, x_2, ..., x_n/\lambda\right) \propto \prod_{i=1}^{n}\left(\lambda^{x_i} e^{-\lambda}\right) \times \lambda^{\alpha-1} e^{-\beta\lambda} = \lambda^{s+\alpha-1} e^{-(n+\beta)\lambda}.$$

Where $s = x_1 + x_2 + ... + x_n$. This is proportional to the PDF of the Gamma $(s+\alpha, n+\beta)$ distribution, so the posterior of Λ must be Gamma $(s+\alpha, n+\beta)$.

As the prior and posterior are both Gamma distributions, the Gamma distribution is a conjugate prior for Λ in the Poisson model.

12.3 LOSS FUNCTION:

A Loss function quantifies the cost or penalty with making a decision α when the true parameter value is θ. It helps in assessing how good or bad an estimator is

1. Squared Error loss

$$L(\theta, a) = (a - \theta)^2$$

- Penalizes large deviations quadratically
- Common in regression problems

2. Absolute Error Loss

$$L(\theta, a) = |a - \theta|$$

- Less sensitive to outliers than squared error loss.

3. 0-1 Loss Function

$$L(\theta, a) = \begin{cases} 0, & \text{if } a = \theta \\ 1, & \text{if } a \neq \theta \end{cases}$$

- Used in classification problems.

4. Hinge Loss (for SVMs)

$$L(y, f(x)) = \max(0, 1 - y f(x))$$

- Used in support vector Machines (SVM).

12.3.1 EXAMPLE:

Suppose you're estimating the mean height of a population. The true value is θ, and your estimate is a.

- **Squared Error Loss**: $L(\theta, a) = (\theta - a)^2$
- If true height is 170 cm and you estimate 165 cm, the loss is:

$$L(170,165)=(170-165)^2 = 25.$$

12.4 RISK FUNCTION:

The risk function is the expected loss over the distribution of the data. I t measures the overall performance of an estimator $\delta(x)$.

$$R(\theta,\delta)=E\left[L(\theta,\delta(X))\right]$$

Where $\delta(X)$ is an estimator based on observed data X.

12.4.1 EXAMPLES:

Example 1: Estimating Mean with squared error loss

Let $X_1, X_2, ..., X_n \sim N(\theta,\sigma^2)$, where θ is unknown. Consider the estimator :

$$\delta(X)=\bar{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$$

The risk function is:

$$R\left(\theta,\bar{X}\right)=E\left[\left(\bar{X}-\theta\right)^2\right]=\frac{\sigma^2}{n}$$

Example 2: 0-1 Loss in classification

Suppose we classify an email as spam $(a=1)$ or not spam $(a=0)$, and the true class is θ. The loss function is:

$$L(\theta,a)=\begin{cases}0, & \text{if } a=\theta\\ 1, & \text{if } a\neq\theta\end{cases}$$

The risk function is the probability of misclassification:

$$R(\theta,\delta)=P(\delta(X)\neq\theta).$$

12.5 MINIMAX ESTIMATOR:

Finding minimax estimators is complicated and we cannot attempt a complete coverage of that theory here but we will mention a few key results. The main message to take away from this section is: Bayes estimators with a constant risk function are minimax.

Theorem 1: Let $\hat{\theta}$ be the Bayes estimator for some prior π. If

$$R\left(\theta,\hat{\theta}\right)\leq B_\pi\left(\hat{\theta}\right) \text{ for all } \theta \qquad (1)$$

then $\hat{\theta}$ is minimax and π is called a least favourable prior.

Proof: Suppose that $\hat{\theta}$ is not minimax. Then there is another estimator $\hat{\theta}_0$ such that $\sup_\theta \mathrm{R}\left(\theta,\hat{\theta}_0\right)<\sup_\theta R\left(\theta,\hat{\theta}\right)$. Since the average of a function is always less than or equal to its maximum, we have that $B_\pi\left(\hat{\theta}_0\right)\leq\sup_\theta \mathrm{R}\left(\theta,\hat{\theta}_0\right)$.

$$B_\pi\left(\hat{\theta}_0\right)\leq\sup_\theta \mathrm{R}\left(\theta,\hat{\theta}_0\right)<\sup_\theta R\left(\theta,\hat{\theta}\right)\leq B_\pi\left(\hat{\theta}\right) \qquad (2)$$

which is a contradiction.

Theorem 2: Suppose that $\hat{\theta}$ is the Bayes estimator with respect to some prior π. If the risk is constant then $\hat{\theta}$ is minimax.

Proof: The Bayes risk is $B_\pi\left(\hat{\theta}\right)=\int \mathrm{R}\left(\theta,\hat{\theta}\right)\,\pi(\theta)d\theta=c$ and hence $\mathrm{R}\left(\theta,\hat{\theta}\right)\leq B_\pi\left(\hat{\theta}\right)$ for all θ.

12.5.1 EXAMPLES:

Example 1: Consider the Bernoulli model with squared error loss. In example 4 we showed that the estimator

$$\hat{p}\left(X^n\right)=\frac{\sum_{i=1}^{n} X_i+\sqrt{n/4}}{n+\sqrt{n}}$$

has a constant risk function. This estimator is the posterior mean, and hence the Bayes estimator, for the prior $Beta(\alpha,\beta)$ with $\alpha=\beta=\sqrt{n/4}$. Hence, by the previous theorem, this estimator is minimax.

Example 2: Consider again the Bernoulli but with loss function

$$L(p,\hat{p})=\frac{(p-\hat{p})^2}{p(1-p)}$$

Let $\hat{p}\left(X^n\right)=\hat{p}=\sum_{i=1}^{n} X_i \Big/ n$. The risk is

$$R(p,\hat{p})=E\left(\frac{(\hat{p}-p)^2}{p(1-p)}\right)=\frac{1}{p(1-p)}\left(\frac{p(1-p)}{n}\right)=\frac{1}{n}$$

which, as a function of p, is constant. It can be shown that, for this loss function, $\hat{p}\left(X^n\right)$ is the Bayes estimator under the prior $\pi(p)=1$. Hence, $\hat{p}$ is minimax.

12.6 BAYES ESTIMATOR:

Let π be a prior distribution. After observing $X^n=\left(X_1,X_2,......,X_n\right)$, the posterior distribution is, according to Baye's theorem,

$$P\left(\theta\in A/X^n\right)=\frac{\int_A p\left(X_1,X_2,......,X_n/\theta\right)\pi(\theta)\,d\theta}{\int_\Theta p\left(X_1,X_2,......,X_n/\theta\right)\pi(\theta)\,d\theta}=\frac{\int_A L(\theta)\pi(\theta)\,d\theta}{\int_\Theta L(\theta)\pi(\theta)\,d\theta} \qquad (1)$$

where $L(\theta)=p\left(x^n;\theta\right)$ is the likelihood function. The posterior has density

$$\pi\left(\theta/x^n\right)=\frac{p\left(x^n/\theta\right)\pi(\theta)}{m\left(x^n\right)} \qquad (2)$$

where $m\left(x^n\right)=\int p\left(x^n/\theta\right)\pi(\theta)d\theta$ is the marginal distribution of X^n. Define the posterior risk of an estimator $\hat{\theta}\left(x^n\right)$ by

$$\pi\left(\hat{\theta}/x^n\right)=\int L\left(\theta,\hat{\theta}\left(x^n\right)\right)\pi\left(\theta/x^n\right)d\theta. \qquad (3)$$

Theorem 1: The Bayes risk $B_\pi\left(\hat{\theta}\right)$ satisfies

$$B_\pi\left(\hat{\theta}\right)=\int r\left(\hat{\theta}/x^n\right)m\left(x^n\right)dx^n.$$

Let $\hat{\theta}\left(x^n\right)$ be the value of θ that maximizes $r\left(\hat{\theta}/x^n\right)$. Then $\hat{\theta}$ is the Bayes estimator.

Proof: Let $p(x,\theta) = p(x/\theta)\pi(\theta)$ denote the joint density of X and θ. We can write the Bayes risk as follows.

$$B_\pi\left(\hat{\theta}\right) = R\left(\theta,\hat{\theta}\right)\pi(\theta)d\theta = \int\left(\int L\left(\theta,\hat{\theta}\left(x^n\right)\right)p(x/\theta)dx^n\right)\pi(\theta)d\theta$$
$$= \int\int L\left(\theta,\hat{\theta}\left(x^n\right)\right)p(x,\theta)dx^n d\theta = \int\int L\left(\theta,\hat{\theta}\left(x^n\right)\right)\pi\left(\theta/x^n\right)m\left(x^n\right)dx^n d\theta$$
$$= \int\left(\int L\left(\theta,\hat{\theta}\left(x^n\right)\right)\pi\left(\theta/x^n\right)d\theta\right)m\left(x^n\right)dx^n = \int r\left(\hat{\theta}/x^n\right)m\left(x^n\right)dx^n.$$

If we choose $\hat{\theta}\left(x^n\right)$ to the value of θ that minimizes $r\left(\hat{\theta}/x^n\right)$ then we will minimize the integrand at every x and thus minimize the integral $\int r\left(\hat{\theta}/x^n\right)m\left(x^n\right)dx^n$.

Now, we can find an explicit formula for the Bayes estimator for some specific loss functions.

Theorem 2: The $L\left(\theta,\hat{\theta}\right) = \left(\theta-\hat{\theta}\right)^2$ then the Bayes estimator is

$$B_\pi\left(\hat{\theta}\right) = \int r\left(\hat{\theta}/x^n\right)m\left(x^n\right)dx^n.$$

Let $\hat{\theta}\left(x^n\right)$ be the value of θ that maximizes $r\left(\hat{\theta}/x^n\right)$. Then $\hat{\theta}$ is the Bayes estimator is

$$\hat{\theta}\left(x^n\right) = \int\theta\pi\left(\theta/x^n\right)d\theta = E\left(\theta/X = x^n\right).$$

If $L\left(\theta,\hat{\theta}\right) = \left|\theta-\hat{\theta}\right|$ then the Bayes estimator is the median of the posterior $\pi\left(\theta/x^n\right)$. If $L\left(\theta,\hat{\theta}\right)$ is zero one loss, then the Bayes estimator is the mode of the posterior $\pi\left(\theta/x^n\right)$.

Proof: We will prove the theorem for squared error loss. The Bayes estimator $\hat{\theta}\left(x^n\right)$ minimizes $r\left(\hat{\theta}/x^n\right) = \int\left(\theta-\hat{\theta}\left(x^n\right)^2\right)\pi\left(\theta/x^n\right)d\theta.$ Taking the derivative of $r\left(\hat{\theta}/x^n\right)$ with respect to $\hat{\theta}\left(x^n\right)$ and setting it equal to zero yields the equation $2\int\left(\theta-\hat{\theta}\left(x^n\right)\right)\pi\left(\theta/x^n\right)d\theta = 0.$

12.6.1 EXAMPLE:

Let $X_1, X_2, \ldots\ldots, X_n \sim N\left(\theta,\sigma^2\right)$, where σ^2 is known. Suppose we use a $N\left(a,b^2\right)$ prior for μ. The Bayes estimator with respect to squared error loss is the posterior mean, which is

$$\hat{\theta}\left(X_1, X_2, \ldots\ldots, X_n\right) = \frac{b^2}{b^2+\frac{\sigma^2}{n}}\bar{X} + \frac{\frac{\sigma^2}{n}}{b^2+\frac{\sigma^2}{n}}a.$$

12.7 KEY WORDS:

- Prior Distribution (Prior)
- Posterior Distribution (Posterior)

- Loss Function
- Risk Function
- Bayes Estimator
- Minimax Estimator

12.8 SUMMARY:

Bayesian decision theory offers a coherent framework for making decisions under uncertainty. The interplay between prior beliefs, data (likelihood), and loss functions allows for flexible and context-aware estimation. The Bayes estimator is optimal under the assumed prior and loss function, while the minimax estimator offers protection against the worst-case scenario. Choosing between them depends on available prior information and the desired level of conservatism in the decision-making process.

12.9 SELF ASSESSMENT QUESTIONS:

1. Let $X_1, X_2, \ldots, X_n$ be n independent N (μ, σ^2) variables when μ is unknown but σ^2 is known. Let prior distribution of μ be N (θ, σ^2). Find Bayes estimate of μ.
2. Explain (i) Loss function (ii) Risk function (iii) Minimax Estimator and (iv) Bayes Estimator.
3. State Bayes' theorem and explain its use in updating beliefs.
4. Suppose you have a prior Beta(2, 2) and you observe 3 successes out of 5 Bernoulli trials. What is the posterior distribution?
5. Define a loss function in the context of statistical decision theory. What is the common form of loss function used for point estimation?
6. What is the risk function associated with a decision rule? How is the risk function computed under a squared error loss?
7. Explain the different types of a loss functions of $X_1, X_2, \ldots\ldots, X_n$ are n independent $N\left(\mu, \sigma^2\right)$ variables where σ^2 is known. Find minimum estimator of M.
8. Distinguish between minimax and Bayes estimators.
9. Explain minimax estimate. Under the conditions to be specified by you prove that a Bayes estimate is a minimax estimate.
10. Describe Priori and postpriori distributions with suitable examples.

12.10 SUGGESTED READINGS:

1. Lehman, E. L., and Cassella, G. (1998). Theory of Point Estimation, Second Edition, Springer, NY.
2. Casella, G. & Berger, R.L. (2013). Statistical Inference, 2nd Edition. Cengage Learning
3. Goon, A. M., Gupta, M. K., and Dasgupta, B. (1989). An Outline of Statistical Theory-Vol.II, World Press, Calcutta.
4. Kale, B. K. (1999). A First Course on Parametric Inference, Narosa Publishing House, New Delhi.
5. An Introduction to Probability and Statistics by V.K. Rohatgi and. K. Md. E. Saleh (2001).

Prof. G. V. S. R. Anjaneyulu

LESSON- 13
CENSORED AND TRUNCATED DISTRIBUTIONS

OBJECTIVES:

By the end of this course/module, students will be able to:

- **Understand the concepts of censoring and truncation** in statistical data, and distinguish between Type I, Type II, and random censoring.
- **Differentiate between censored and truncated data**, and explain their implications in statistical inference and data analysis.
- **Model lifetime and reliability data** using appropriate censored or truncated distributions, particularly the Normal and Exponential distributions.
- **Derive and interpret the likelihood functions** for Type I and Type II censored data from Normal and Exponential distributions.
- **Compute Maximum Likelihood Estimators (MLEs)** for parameters under both Type I and Type II censoring schemes.
- **Apply statistical techniques** to estimate parameters and assess goodness-of-fit for censored datasets using real-world reliability and survival data.
- **Use appropriate statistical software or programming tools** (e.g., R, Python) to perform analysis involving censored and truncated data.
- **Compare efficiency and bias** of estimators under censored and uncensored scenarios.
- **Recognize applications** of censored and truncated data analysis in fields such as biomedical research, engineering, and actuarial science.

STRUCTURE :

13.1 INTRODUCTION:

In statistical analysis, particularly in reliability engineering, survival analysis, and medical research, it is common to encounter incomplete data due to **censoring or truncation.** These phenomena arise when the full data on lifetimes or measurements are not available either due to design limitations or observational constraints. Understanding censored and truncated data is essential for accurate modeling and inference.

Censoring refers to situations where the value of an observation is only partially known. Two of the most widely studied censoring mechanisms are **Type I censoring**—where observations are censored at a predetermined time—and **Type II censoring**—where censoring occurs after a fixed number of failures or events have been observed. These types are frequently encountered in life-testing experiments and clinical trials.
On the other hand, **truncation** occurs when certain data points are not observed at all if they fall outside a specified range. Truncated distributions, therefore, require special consideration, as they involve conditional probability models that differ from standard distributions.

This work focuses on censored data from **normal** and **exponential distributions**, two of the most commonly used models in statistical inference. The primary objective is to develop and understand the **maximum likelihood estimators (MLEs)** for these distributions under both Type I and Type II censoring. MLEs are essential tools in estimating parameters, even when full data is not available, and their derivation under censoring provides deeper insights into robust estimation methods.

Through mathematical formulations and illustrative examples, this study aims to highlight how censoring affects estimation and how one can adjust standard inferential procedures to accommodate incomplete data while still achieving valid statistical conclusions.

13.2 TRUNCATION DEFINITION:

In Estimation, "Truncation" refers to the act of limiting a Data set by excluding values that fall outside A certain range, essentially "cutting off" the extreme ends of the data, resulting in a truncated sample where only values within the specified range are considered for analysis; This can lead to biased estimates if not properly accounted for **truncated distribution** is a probability distribution that results when data or observations are limited or "cut off" beyond a certain threshold (or truncation point). Essentially, it is a modified version of a standard probability distribution where values outside a specific range are excluded, or the probability is redistributed to the truncated range.

Truncation can happen on either end (from below or from above), or at both ends. A truncated distribution represents the conditional distribution of the original distribution, given that the values fall within a specified interval.

13.3 TYPES OF TRUNCATION:

- **Left truncation**: Data points below a certain value (threshold) are excluded.
- **Right truncation**: Data points above a certain value (threshold) are excluded.
- **Two-sided truncation**: Data points outside a specific range (both below and above certain thresholds) are excluded standard distribution, such as the normal distribution.

Type 1 Truncation refers to a specific case where the normal distribution is truncated at either the lower bound, upper bound, or both. In truncation, observations outside a specified range are entirely excluded from the analysis. For **Type 1 truncation** in a normal distribution, the truncation typically occurs at one end (either lower or upper), meaning that data outside the truncation limits is not considered or observed.
Here's a detailed explanation of **Type 1 truncation** in the context of the normal distribution:

Type 1 Truncation - Lower Bound:
In this case, any values below a certain threshold are excluded, and only values above this threshold are observed.

- **Example**: Suppose we have a normal distribution with mean μand standard deviation . If the distribution is truncated at a lower bound L, then any values less than L are removed from the sample. So, we only observe values X such that X≥L
- The truncated normal distribution for values X≥L can be expressed as

$$f_X(x|L) = \frac{f_X(x)}{P(X \geq L)} = \frac{\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{\sigma^2}\right)}{1 - \emptyset\left(\frac{L-\mu}{\sigma}\right)}$$

where:

- f_X(x)s the probability density function (PDF) of the normal distribution.
- Φ is the cumulative distribution function (CDF) of the standard normal distribution.
- The denominator ensures that the total probability under the truncated distribution is

Type 1 Truncation - Upper Bound:
In this case, any values above a certain threshold are excluded, and only values below this threshold are observed.

- **Example**: If the normal distribution is truncated at an upper bound U, then any values greater than are removed from the sample. So, we only observe values X such that X≤U
- The truncated normal distribution for values X≤U is given by:

$$f_X(x|U) = \frac{f_X(x)}{P(X \geq U)} = \frac{\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)}{\emptyset\left(\frac{U-\mu}{\sigma}\right)}$$

where:

- The denominator normalizes the distribution to ensure proper probability scaling.

Type 1 Truncation - Both Lower and Upper Bound:

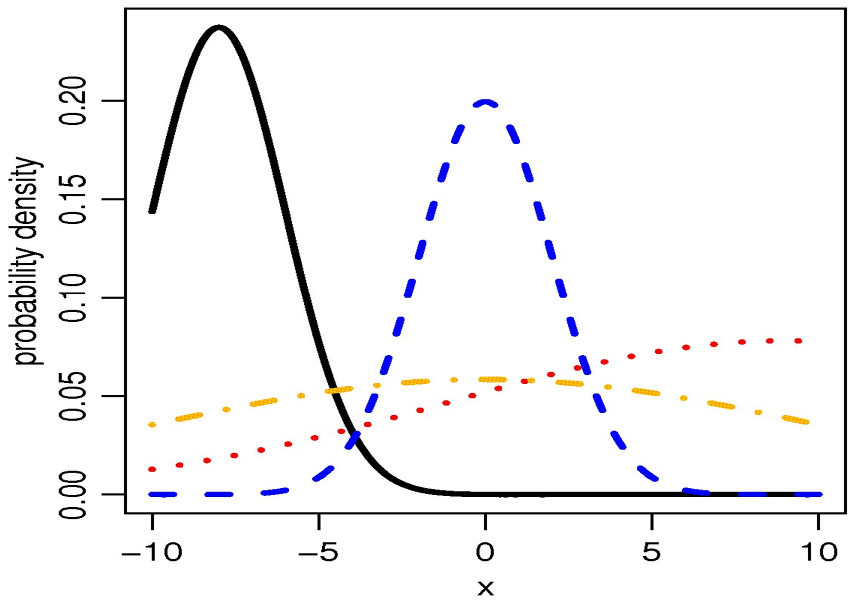
Truncation can also happen at both ends of the distribution. In this case, we only observe values X such that L≤X≤U , where L is the lower bound and U is the upper bound.

$$\text{The truncated normal distribution for values between } L \text{ and } U \text{ is:}$$

$$f_X(x|L,U) = \frac{f_X(x)}{P(L \le X \le U)} = \frac{\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)}{\emptyset\left(\frac{U-\mu}{\sigma}\right) - \emptyset\left(\frac{L-\mu}{\sigma}\right)}$$

where:

The denominator normalizes the probability within the truncation range.



EXAMPLE OF TYPE 1 TRUNCATION:

- Suppose we have a normal distribution with a mean μ=10\mu = 10μ=10 and a standard deviation σ=2\sigma = 2σ=2.
- If we truncate this distribution at a=8a = 8a=8 and b=12b = 12b=12, then the resulting truncated distribution will have:
  - No values below 8.
  - No values above 12.
  - The probability mass will be redistributed between 8 and 12, with the normal shape still present but truncated at both ends.

TYPE 2 TRUNCATION :

- **Type 2 truncation** in the context of a **normal distribution** refers to truncation based on **one-sided limits**, where values outside of the truncation limits are excluded, but only **one side** of the distribution is truncated. This differs from Type 1 truncation, where both ends of the distribution are truncated.
- In **Type 2 truncation**, truncation occurs either at the **lower** or the **upper bound** of the distribution, but not both.

Types of Type 2 Truncation:

1. **Upper truncation** (Right truncation):
   In this case, values above a certain threshold are excluded.
   - If the threshold is b, then values where X>bare excluded. The remaining values are X≤b
2. **Lower truncation** (Left truncation):
   In this case, values below a certain threshold are excluded.
   - If the threshold is a, then values where X<aare excluded. The remaining values are X≥a

Mathematical Formulation for Type 2 Truncation:

For a **normal distribution** with mean μand standard deviation σ, the truncated distribution's probability density function (PDF) changes depending on whether you have **upper** or **lower** truncation:

1. Upper truncation at b:

If the normal distribution is truncated above a threshold b, the PDF of the truncated distribution (for X≤b) is

$$f_{trunc}(x) = \frac{f_{X(x)}}{\text{P(X≤b)}} \text{ for X≤b}$$

- f_X(x) is the normal PDF:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

P(X≤b) is the cumulative probability for the normal distribution up to b:

$$\text{P(X≤b)}=\Phi\left(\frac{b-\mu}{\sigma}\right)$$

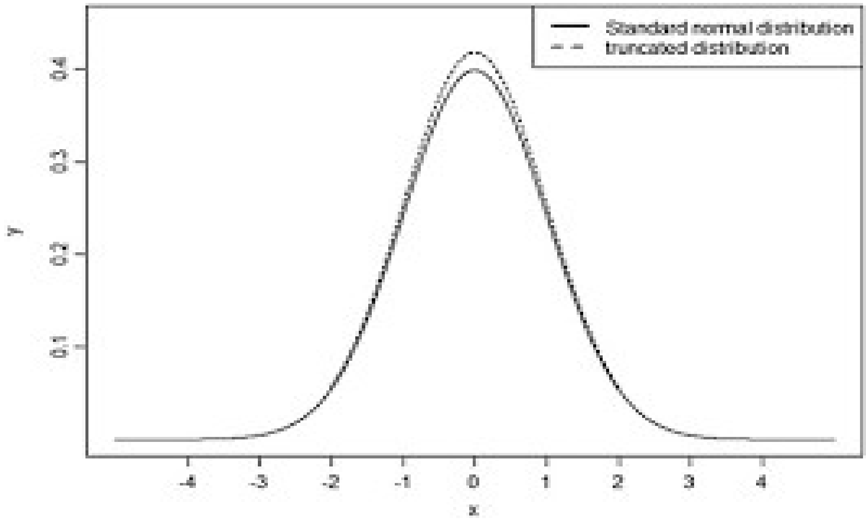
- where Φ(x) is the cumulative distribution function (CDF) of the standard normal distribution.

1. Lower truncation at a:

If the normal distribution is truncated below a threshold a, the PDF of the truncated distribution (for X≥a) is

$f_{trunc}(x) = \frac{f_{X(x)}}{P(X \geq a)}$ for X≥a

Where: $P(X \geq a) = 1 - \Phi(\frac{a-\mu}{\sigma})$



EXAMPLE OF TYPE 2 TRUNCATION:

Consider a normal distribution with:

- Mean μ=10
- Standard deviation σ=2

Upper truncation at b=12:

The distribution will include only values less than or equal to 12. So, if any data points are above 12, they are excluded.

Lower truncation at a=8:

The distribution will include only values greater than or equal to 8. Any data points below 8 will be excluded.

13.4 TRUNCATION IN EXPONENTIAL DISTRIBUTION:

Type 1 Truncation:

Type 1 truncation in the context of the **exponential distribution** refers to a situation where the distribution is truncated at both ends, meaning that only values within a certain range are observed. In other words, values below a lower threshold and above an upper threshold are excluded, and only values in between the thresholds are observed.

For an **exponential distribution**, truncation alters the shape and properties of the distribution. Typically, the **exponential distribution** is defined as

f_X(x)=$\lambda e^{-\lambda x}$ for x≥0

where:

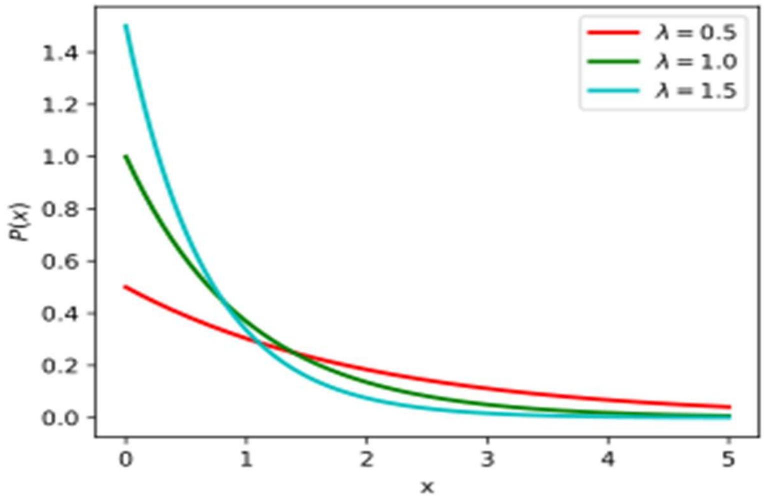
- λ>0is the rate parameter (inverse of the mean),
- x is the random variable.

EXAMPLE:

Let's say the original exponential distribution has a rate parameter λ=0.5 and the istribution is truncated at a=1 and b=3

- The **original exponential PDF** is fX(x)=0.5$e^{-0.5}$ for x≥0.
- The **truncated exponential PDF** within the range [1,3]

$$f_{trunc}(x)) = \frac{0.5e^{-0.5}\, r}{e^{0.5\times1} - e^{0.5\times1}}\ for\ x\geq 0$$



TYPE 2 TRUNCATION IN THE EXPONENTIAL DISTRIBUTION:

1. Right Truncation (Upper truncation):

- In this case, values **greater than** a threshold b are excluded, and only values x≤b are observed.
- The **exponential distribution** is truncated from the upper side.
- The PDF is adjusted to account for the truncation, and normalization occurs over the truncated range [0,b].

2.Left Truncation (Lower truncation):

- In this case, values **less than** a threshold a are excluded, and only values x≥a are observed.
- The **exponential distribution** is truncated from the lower side.
- The PDF is adjusted to account for the truncation, and normalization occurs over the truncated range(a,∞)

Mathematical Formulation for Type 2 Truncation:

Let Xbe a random variable that follows an **exponential distribution** with rate parameter λmeaning the original distribution has the PDF:

$$f_X(x)=\lambda e^{-\lambda x} \text{ for } x\geq 0$$

1. Right Truncation at bbb (values greater than bbb are excluded):

In this case, we only observe values in the range [0,b][0, b][0,b].

- The probability density function (PDF) of the truncated exponential distribution is:

$$f_{trunc}(x)=\frac{f_X(x)}{P(0\leq X\leq b)}=\frac{\lambda e^{-\lambda x}}{1-e^{-\lambda b}} \text{ for } 0\leq x\leq b$$

where

- $f_X(x)$ (x)=$\lambda e^{-\lambda x}$ is the PDF of the exponential distribution.
- P(0≤X≤b)=1−his normalization ensures that the total probability for x in the truncated range sums to 1.

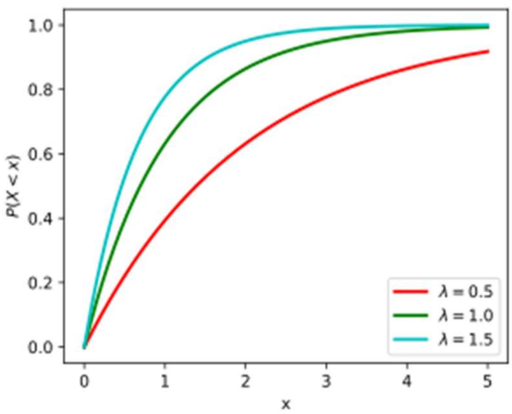
2. Left Truncation at a(values less than aare excluded):

- In this case, we only observe values in the range [a,∞]
- The PDF of the truncated exponential distribution is:
- $e^{-\lambda b}$ is the probability that X lies within the range [0,b]

$$f_{trunc}(x)=\frac{f_X(x)}{P(0\leq X\leq b)}=\frac{\lambda e^{-\lambda x}}{1-e^{-\lambda b}} \text{ for } 0\leq x\leq b$$

where

- $f_X(x)$ (x)=$\lambda e^{-\lambda x}$ is the PDF of the exponential distribution.
- P(≤X≤b)=1−his normalization ensures that the total probability for x in the truncated range



EXAMPLE OF TYPE 2 TRUNCATION:

Suppose we have an exponential distribution with a rate parameter λ=0.5 (i.e., the mean is 2).

1. Right Truncation at b=3b

For this right truncation, the truncated PDF will be:

$$f_{trunc}(x) = \frac{0.5e^{-0.5}\,r}{\mathrm{e}^{0.5\times1} - \mathrm{e}^{0.5\times1}}\; for 0 \leq x \leq 3$$

2. Left Truncation at a=1a

For this left truncation, the truncated PDF will be:

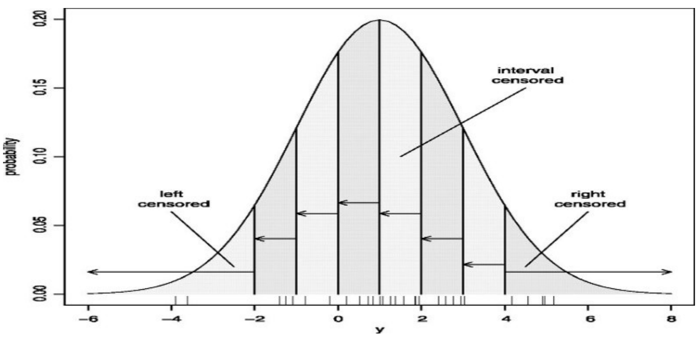
$$f_{trunc}(x) = \frac{0.5e^{-0.5}\,r}{1 - \mathrm{e}^{0.5\times1}}\; for\ x \leq 3$$

CENSORING IN NORMAL DISTRIBUTION:

13.5 CENSORING TYPE 1 IN NORMAL DISTRIBUTION:

Type 1 censoring in the context of a normal distribution refers to a scenario where the values of a random variable are partially observed. Specifically, the observed values are censored if they fall below or above a certain threshold. The exact value of the variable is not known for those censored observations, but it is known that they are either below or above the threshold.

In Type 1 censoring, you only know that the true value of the random variable is either greater than or less than a certain threshold, but you don't know the exact value.



1. Right Censoring

Right Censoring (a typical form of Type 1 censoring):
For example, let's say we are observing the lifetime of a product, and the lifetime is assumed to follow a normal distribution with mean μ and standard deviation σ. However, we can only

observe values up to a certain threshold, say c. If a product survives longer than c, we only know that its lifetime is greater than c, but we don't know the exact value. This is a form of **right-censoring** (a Type 1 censoring case).

Example of Right-Censoring:

Consider an experiment where the lifetime of a machine is measured and assumed to follow a normal distribution. Let's assume the lifetime is normally distributed with mean μ=1000hours and standard deviation σ=150hours. However, we cannot observe lifetimes beyond 1200 hours due to the experiment's constraints.

For machines that live longer than 1200 hours, we only know that their lifetime is greater than 1200 hours, but we don't know the exact value. This is an example of **right-censoring**.

2. Left censoring:

Left censoring: If an observation falls below a certain threshold, we know only that the value is less than that threshold, but not its exact value.

EXAMPLE OF LEFT CENSORING IN A NORMAL DISTRIBUTION:

Suppose we are conducting an experiment where we measure the **blood pressure** of a group of individuals, and assume that the blood pressure values follow a **normal distribution** with mean μ=120mmHg and standard deviation σ=15 mmHg

However, the measurement device we are using can only record **blood pressure readings above 100 mmHg**. If someone's blood pressure is below 100 mmHg, the device records it as "censored," and we only know that their blood pressure is less than 100 mmHg.

In this case, all blood pressure readings below 100 mmHg are considered **left-censored**, as we only know that their blood pressure is **less than 100 mmHg**, but not the exact value.

CENSORING TYPE 2 IN NORMAL DISTRIBUTION:

Type 2 censoring in the context of a **normal distribution** refers to a situation where a fixed number of data points are censored during an experiment or observation, rather than a specific threshold. This type of censoring typically involves censoring a pre-defined fraction of the data after the experiment has been conducted, not based on a fixed threshold value.

In **Type 2 censoring**, you observe a set of data points, and a specific number of them are censored at random (not based on a time or threshold limit). After the experiment, you may have only a partial view of the data, knowing which specific data points are censored, but not their exact values.

1. Left Censoring (Type 2 Censoring):

In **left censoring**, we know that some data points are **below** a certain threshold, but we do not observe their exact values. We only know that they are **less than** the threshold.

Example:

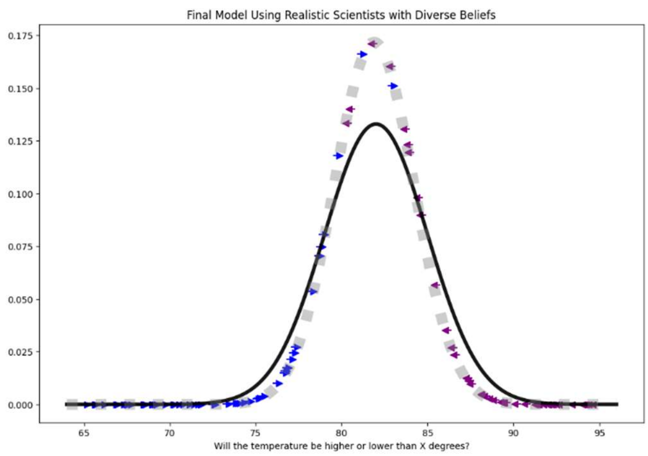
If you are studying the **time-to-failure** of machines and you are using a normal distribution to model the failure times, you may know that **some machines failed very early**, but their exact failure time is not recorded (only that it is below a certain value).

2. Right Censoring (Type 2 Censoring):

In **right censoring**, we know that some data points are **above** a certain threshold and we cannot observe their exact values. We only know that they are **greater than** the threshold.

Example:

In a clinical trial measuring **time to event (e.g., recovery time)**, if the trial ends before an individual experiences the event, you only know that the time-to-event is **greater than** the time of the last observation. This is **right-censoring**.



13.6 TYPE 1 CENSORING IN EXPONENTIAL DISTRIBUTION:

Type 1 censoring in the context of an **Exponential Distribution** is when data points are censored at a **fixed threshold**. In other words, you observe the exact values of the random variable **up to a certain time** or threshold, but you do not observe values **beyond that threshold**.

In the context of an **Exponential Distribution**, the censoring typically involves the **failure time** (or survival time) of an event. The Exponential Distribution is often used to model the **time between events** in a **Poisson process**, where events occur at a constant rate λ.

1. Right Censoring (Type 1 Censoring)

In **right censoring**, we observe data up until a certain threshold c, but we do not know the exact value of the observations that are **greater than c**. We only know that the event occurred after time c.

EXAMPLE:

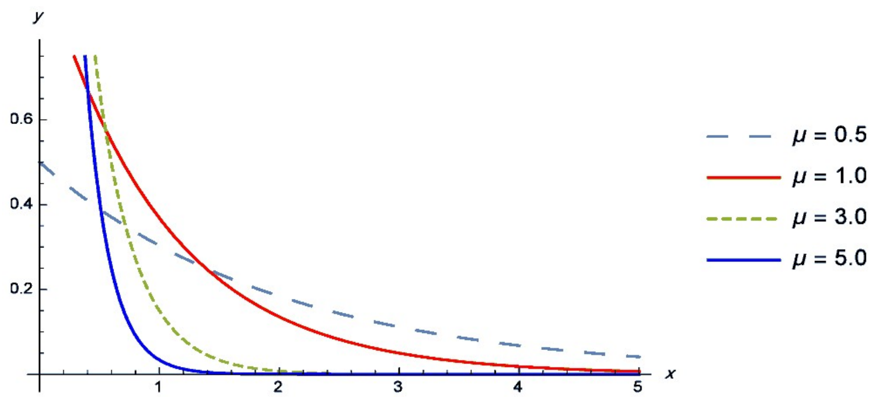
- You are studying the lifetime of light bulbs, and you have set a maximum observation time of **10 hours**. Any light bulb that doesn't fail within this time period is **right-censored**. If a bulb is still working after 10 hours, you only know that its failure time is **greater than 10 hours** but you don't know the exact failure time.

2. Left Censoring (Type 1 Censoring)

In **left censoring**, we know that some observations are **below** a threshold c, but we do not observe their exact values. We only know that the value of the random variable is **less than c**.

EXAMPLE:

- You are studying the time-to-failure of light bulbs, but you cannot observe failures that happen **before 5 hours**. Any light bulb that fails **before 5 hours** is **left-censored**. For these light bulbs, you only know that their failure time is **less than 5 hours**.



TYPE 2 CENSORING IN EXPONENTIAL DISTRIBUTION:

In **Type 2 Censoring**, the number of observations that are censored is fixed, rather than the time point at which the data is censored, as in Type 1 censoring. In this case, the censoring process stops after a specific number of events have been observed (uncensored) or after a fixed number of censored observations have been made. The key idea is that, rather than a predetermined time threshold, the censoring is determined by the number of uncensored events.

1. Right Censoring (Type 2 Censoring)

In **right censoring** with **Type 2 Censoring**, we are observing the failure times of some number of events until a fixed number of censoring events have occurred. This means that, after a specific number of **right-censored observations**, we stop collecting data.

EXAMPLE:

Let's say we have **10 light bulbs** and we are monitoring them until we have observed **2 right-censored events**. For each light bulb that **fails after 6 hours**, we do not know the exact failure time but only that it is greater than 6 hours.

- We stop the experiment after 2 light bulbs are **right-censored**.
- The failure times of the **remaining bulbs** are observed.

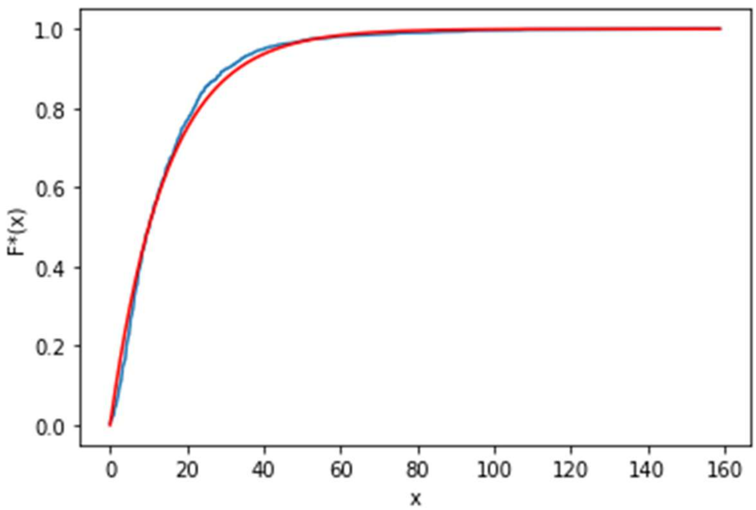
2. Left Censoring (Type 2 Censoring)

In **left censoring** with **Type 2 Censoring**, we observe failure times of some events but censor a fixed number of events where the failure time is less than the censoring threshold. Here, after **a specific number of left-censored observations**, we stop collecting data.

EXAMPLE:

Let's say we are studying the lifetime of light bulbs, and we have **10 bulbs**. We monitor them for a maximum time of **7 hours**. After observing **2 left-censored events** (bulbs that fail before 7 hours), we stop the experiment. For bulbs that fail before 7 hours, we only know that the failure time is **less than 7 hours**.

- We stop the experiment after 2 bulbs are **left-censored**.
- The failure times of the remaining bulbs are observed.



13.7 DISTRIBUTION AND THEIR MAXIMUM LIKELIHOOD ESTIMATION (MLE) & INTERVAL ESTIMATION:

Maximum Likelihood Estimation (MLE) and interval estimation are two key techniques used in statistical inference. MLE is used to estimate parameters of a distribution that best fit the observed data, while interval estimation helps provide a range of plausible values for the parameter of interest.
Below, I'll outline some common distributions, their MLE derivations, and methods of interval estimation for the estimated parameters.

1. Exponential Distribution
The **Exponential Distribution** is commonly used to model the time between events in a Poisson process. The probability density function (PDF) of the exponential distribution is:

$$f(x;\lambda)=\lambda e^{-\lambda x},\ x\geq 0, \lambda>0$$

Where

- λ is the rate parameter of the distribution.

Maximum Likelihood Estimation (MLE) for λ:
Given independent observations x1,x2,…,xn is the likelihood function is**:**

$$L(\lambda)=\prod_{n=1}^{i} \lambda e^{-\lambda xi}$$

Confidence Interval :
For large n, we can use the **asymptotic normality** of MLE for interval estimation. The variance of λ is approximately:

$$\mathrm{Var}(\lambda)=\frac{\lambda^2}{n}$$

2. Normal Distribution
The **Normal Distribution** is one of the most widely used distributions, characterized by two parameters: the mean μ and the standard deviation σ The probability density function (PDF) is:

$$(f(x;\mu,\sigma)) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \text{ for } -\infty<x<\infty$$

Maximum Likelihood Estimation (MLE) for μ and *σ*
Given nnn independent observations x1,x2,…,xn is the likelihood function is:

$$L(\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Confidence Interval for μ and σ:

For μ and ***σ***(knownσ): The 95% confidence interval for μ\muμ is given by:

$$\mu \pm z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

13.8 KEY WORDS:

- Censoring
- Type I Censoring
- Type II Censoring
- Right Censoring
- Left Censoring
- Truncated Distributions
- Truncation
- Right Truncation
- Left Truncation
- Maximum Likelihood Estimation (MLE)
- Likelihood Function
- Exponential Distribution
- Normal Distribution

13.9 SUMMARY:

In many real-world scenarios, complete data may not be available due to various limitations. Censoring and truncation are two mechanisms that describe such incomplete data, particularly in fields like survival analysis, reliability engineering, and biostatistics.

Censoring:
Censoring occurs when the exact value of an observation is unknown but is known to fall above or below a certain threshold. There are different types:

- Type I Censoring: The experiment ends at a pre-specified time. Units that haven't failed by this time are considered censored.
- Type II Censoring: The experiment continues until a pre-specified number of failures occur. Remaining units are censored at the time of the last observed failure.

Truncation:
Truncation occurs when certain data points are **not observed at all** because they fall outside a specified range.

- Left Truncation: Observations below a threshold are unrecorded.
- Right Truncation: Observations above a threshold are unrecorded.

MLE FOR CENSORED AND TRUNCATED DATA:
The Maximum Likelihood Estimation (MLE) method must be adjusted to account for the missing or censored data. For both the normal and exponential distributions:

- The likelihood function is modified based on the type of censoring or truncation.
- In exponential distributions, due to their memoryless property, MLEs under censoring are often simpler and analytically tractable.
- In normal distributions, MLEs often require numerical methods for censored data because the likelihood involves integration over the tails.

CONCLUSIONS:

1. Handling censored and truncated data correctly is essential to avoid biased parameter estimation.

2. MLE remains a powerful tool, but its implementation must be adapted to the data structure.
3. Exponential distributions offer more analytical tractability under censoring, while normal distributions often require computational approaches.
4. The type of censoring (Type I vs. Type II) significantly impacts the structure of the likelihood function and the inference process.
5. Understanding these methods is crucial in reliability testing, clinical trials, and survival analysis, where data is often incomplete.

13.10 SELF ASSESSMENT QUESTIONS:

1. What is the difference between censoring and truncation in statistical data?
2. Explain the difference between Type I and Type II censoring. Give a real-world example of each.
3. How does censoring affect the likelihood function used in maximum likelihood estimation (MLE)?
4. Why is MLE preferred over other estimation methods in the context of censored data?
5. Why is the exponential distribution particularly suitable for modeling censored lifetime data?
6. What is right censoring? How does it differ from left censoring?
7. Explain Type-1 censoring for normal distribution.
8. Derive MLE for Exponential distribution.
9. Describe the Truncated exponential distribution.
10. Discuss about the Type II censoring for exponential distribution.

13.11 SUGGESTED READINGS:

1. Lawless, J. F. (2003). Statistical Models and Methods for Lifetime Data – A comprehensive book covering lifetime distributions, censoring, truncation, and MLE methods.
2. Klein, J. P., & Moesch berger, M. L. (2003). Survival Analysis: Techniques for Censored and Truncated Data – Detailed explanations on survival data, types of censoring/truncation, and statistical estimation.
3. David, H. A., & Nagaraja, H. N. (2003). Order Statistics – Useful for understanding Type II censoring and its relationship with order statistics.
4. Cox, D. R., & Oakes, D. (1984). Analysis of Survival Data.
5. An Introduction to Probability and Statistics by V.K. Rohatgi and. K. Md. E. Saleh (2001).

Prof. G. V. S. R. Anjaneyulu

LESSON -14

INTERVAL ESTIMATION AND CONFIDENCE INTERVALS

OBJECTIVES:

By the end of this module, students will be able to:

- Understand the Concept of Interval Estimation
- Construct Confidence Intervals Using Pivotal Quantities
- Evaluate and Interpret Confidence Intervals
- Derive and Compare Confidence Intervals
- Understand and Construct Shortest Expected Length Confidence Intervals
- Apply Interval Estimation Techniques to Real Data
- Appreciate Theoretical Foundations and Limitations.

STRUCTURE :

14.1 INTRODUCTION:

In the previous unit, we have discussed the point estimation, under which we learn how one can obtain point estimate(s) of the unknown parameter(s) of the population using sample observations. Everything is fine with point estimation but it has one major drawback that it does not specify how confident we can be that the estimated close to the true value of the parameter. Hence, point estimate may have some possible error of the estimation and it does not give us an idea of how these estimates deviate from the true value of the parameter being estimated. This limitation of point estimation is over come by the technique of interval estimation. Therefore, instead of making the inference of estimating the true value of the parameter through a point estimate one should make the inference of estimating the true value of parameter by a pair of estimate values which are constituted an interval in which true value of parameter expected to lie with certain confidence. The technique of finding such interval is known as "Interval Estimation".

14.2 INTERVAL ESTIMATION:

Interval estimation is a statistical technique used to estimate a population parameter (like a mean or proportion) by specifying a range (or interval) of values within which the parameter is expected to fall. Unlike a single point estimate (like a sample mean), interval estimation gives a range that expresses the uncertainty around the estimate.

If the point estimation does not give the correct and best estimation to the parameter θ, in this case interval estimation is more useful to estimate the parameter. Let $x_{1,} x_{2,} ...., x_n$ be a random sample from a population with the density function $f(x,\theta)$. If we use two functions of the sample values as an interval to estimate the population parameter θ, this interval is called the confidence interval and this estimation procedure is called interval estimation to the unknown parameter θ.

Interval estimation is a statistical technique used to estimate a population parameter (such as the mean or proportion) by specifying a range (interval) of values within which the parameter is expected to lie, along with a certain level of confidence.

- **Point estimate**: A single value estimate of a parameter (e.g., sample mean).
- **Confidence interval (CI)**: A range constructed around the point estimate that likely contains the true population parameter.
- **Confidence level**: The probability (typically 90%, 95%, or 99%) that the interval contains the true parameter.

EXAMPLE: Estimating a Population Mean

Scenario: A company wants to estimate the average weight of a particular product. A random sample of 50 items is taken, and the sample mean is found to be 100 grams with a standard deviation of 10 grams.

We want to calculate a 95% confidence interval for the population mean.

Step-by-step Solution:

1. Sample mean $(\bar{x}) = 100 grams$
2. Standard deviation $(s) = 10$ grams
3. Sample size $(n) =$ 50
4. Confidence level = 95% → z-value = 1.96 (from Z-tables for normal distribution)
5. Standard error $(SE) = \frac{s}{\sqrt{n}} = \frac{10}{\sqrt{50}} \approx 1.414$

14.3 CONFIDENCE INTERVALS (CI):

Let $X_{1,} X_{2,}, \ldots\ldots, X_n$ be a random sample from the p.d.f. $f(x;\theta)$. *Let* $\mathrm{T}_1 = t_1(X_1, X_2, \ldots\ldots, X_n)$ and $\mathrm{T}_2 = t_2(X_1, X_2, \ldots\ldots, X_n)$ be two statistics satisfying $\mathrm{T}_1 \le \mathrm{T}_2$ for which $\mathrm{P}[\mathrm{T}_1 < \mathrm{g}(\theta) < \mathrm{T}_2] = 1-\alpha$, where $\mathrm{g}(\theta)$ is some real function of θ and $1-\alpha$ does not depend on θ. Then the random interval $(\mathrm{T}_1, \mathrm{T}_2)$ is called a $100(1-\alpha)$ percent confidence interval for $\mathrm{g}(\theta)$; $1-\alpha$ is called confidential coefficient; and T_1 and T_2 are called lower and upper confidence limits for $\mathrm{g}(\theta)$ respectively. A value (t_1, t_2) of the random interval $(\mathrm{T}_1, \mathrm{T}_2)$ is called a $100(1-\alpha)$ percent confidence interval for $\mathrm{g}(\theta)$.

If one or the other, not both, of the two statistics $t_1(X_1, X_2, \ldots\ldots, X_n)$ and $t_2(X_1, X_2, \ldots\ldots, X_n)$ is constant, then we have one-sided confidence interval as defined below.

EXAMPLE: $X \sim N(\theta, \sigma^2); \sigma^2$ is known. Find a confidence interval of θ with confidence coefficient $(1-\alpha)$.

Answer:

$$P_\theta\left(\left|\frac{\sqrt{n}(\bar{X}-\theta)}{\sigma}\right| > \tau_{\alpha/2}\right) = \alpha$$

$$P_\theta\left[-\tau_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \bar{X}-\theta < \tau_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right] = 1-\alpha$$

$$P_\theta\left[\bar{X}-\tau_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \theta < \bar{X}+\tau_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right] = 1-\alpha$$

$$\therefore \underline{\theta}(\underline{X}) = \bar{X}-\tau_{\alpha/2}\frac{\sigma}{\sqrt{n}} \text{ and } \bar{\theta}(\bar{\mathrm{X}}) = \bar{X}+\tau_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

Hence, $100(1-\alpha)$ confidence interval of θ is given by

$$\left[\bar{X}-\tau_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X}+\tau_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right].$$

Step to find out Confidence Interval:

i. Give the critical region of the both tailed test at level α
ii. Reverse the inequality sign and hence the R.H.S will be $(1-\alpha)$.
iii. From the inequality under probability solve for θ.

14.4 INTERVAL ESTIMATION and CONFIDENCE INTERVALS:

Interval Estimation refers to estimating a range of values (called an Interval) for an unknown parameter (like a population mean, variance or proportion) based on sample data. The interval estimation provides more information than a point estimate because it includes a measure of uncertainty.

Confidence Intervals are a popular method of interval estimation. A confidence interval (CI) is a range of values derived from sample data that is likely to contain the population parameter of interest. The interval is constructed so that it has a specified probability (confidence level) of containing the true population parameter.

For example, if we say we have a 95% confidence interval for population mean, it means that if we were to repeat the sampling process many times, about 95% of the calculated intervals would contain the true mean.

Steps to Construct Confidence Intervals:

1. **Choose the confidence Level:** Often 90%, 95% or 99% (the higher the confidence level, the wider the interval).
2. **Find the Sampling Distribution:** The shape of the sampling distribution of the statistic (Ex: sample mean) depends on the underlying population and sample size.
3. **Calculate the Standard Error:** This is the standard deviation of the sampling distribution of the statistic. It quantifies the variability of the estimate.
4. **Critical Value:** The critical value is based upon the confidence level and the sampling distribution. For example for a 95% confidence interval with a normal distribution, the critical value is approximately 1.96 (for large sample sizes)
5. **Construct the Interval:** Use the formula

$$\textit{Confidence Interval} = \hat{\theta} \pm \textit{(Critical value X Standard Error)}$$

Where

- $\hat{\theta}$ is the sample statistic (Ex: sample mean).
- Critical Value is based upon the Chosen confidence level (Ex: Z -value for large samples or t-value for small samples).
- Standard Error is the standard deviation of the statistic.

14.5 USING PIVOTS IN INTERVAL ESTIMATION:

A **pivot** is a statistic that has a known distribution, which is used to construct confidence intervals. The idea is to use a function of the sample statistic and the unknown parameter that has a known distribution, like the t-distribution or normal distribution, and solve for the parameter.

For example, consider the formula, for a confidence interval for the population mean when the population variance is unknown:

$$\frac{\overline{x} - \mu}{S/\sqrt{n}} \sim t_{(n-1)}$$

- $\overline{x}$ is the sample mean.
- S is the sample standard deviation.
- n is the sample size
- $t_{(n-1)}$ is the t-distribution with $n - 1$ degrees of freedom.

Using the pivot, we can set up the confidence interval by rearranging the formula:

$$\mu = \bar{x} \pm t_{(n-1)} \times \frac{S}{\sqrt{n}}$$

This equation provides the confidence interval for the population mean based on the sample data.

EXAMPLE:

A **pivot** is a function of the sample data and the parameter of interest whose distribution does **not** depend on the parameter.

This allows us to construct confidence intervals (CIs) for the parameter without needing to know its true value.

Let $X_1, X_2, \ldots\ldots, X_n \sim N\left(\mu, \sigma^2\right)$, with known variance σ^2

Define the pivot:

$$Z = \frac{\bar{X} - \mu}{\sigma/n} \sim N\left(0,1\right)$$

Then

$$P\left(-Z_{\alpha/2} \le \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \le Z_{\alpha/2}\right) = 1 - \alpha$$

Solving for μ

$$P\left(\bar{X} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

Confidence Interval**:**

$$\mu \in \left[\bar{X} \pm Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right]$$

14.6 SHORTEST CONFIDENCE INTERVALS:

1. **Definition**:
   A *shortest confidence interval* is the confidence interval with the smallest possible length that still maintains the required confidence level (e.g., 95%).
2. **Goal**:
   Minimize the width of the interval while ensuring the true parameter is captured with high probability.
3. **Asymmetric Intervals are often better**:
   For non-symmetric or skewed distributions, the shortest interval is typically not centered around the estimator (e.g., not symmetric around the mean or median).
4. **Bayesian Perspective**:
   In Bayesian statistics, Highest Posterior Density (HPD) intervals are often the shortest credible intervals for a given posterior distribution.

5. **Depend on the Distribution**:
   The shortest interval depends heavily on the distribution of the estimator or sample (e.g., Normal, Binomial, Poisson).
6. **Useful for Precision**:
   Especially valuable when precision is critical, such as in medical trials or quality control, where tighter bounds are needed.
7. **Computational Methods May be Needed**:
   Finding shortest intervals often requires numerical optimization or likelihood ratio methods, especially in discrete or complex models.

14.7 EXPECTED LENGTH OF CONFIDENCE INTERVALS:

We will focus on how to minimize the **expected length of confidence intervals** (CIs) for various distributions. The **expected length** is key in understanding the precision of our estimates, and minimizing it can make the estimation process more efficient. We'll focus on practical strategies, including sample size considerations, the choice of distribution, and how different factors affect the CI length.

Introduction to Confidence Interval Length

- **Confidence Interval (CI)**: A range of values derived from the sample data used to estimate an unknown parameter. The width of this range is important: a narrower CI provides a more precise estimate.
- **Expected Length of CI**: The average width of the CI. Minimizing it improves the precision of the estimate.

Key Factors Affecting Expected Length:

- **Sample Size (n)**: Larger sample sizes lead to narrower CIs.
- **Confidence Level (1−α)**: Higher confidence levels result in wider CIs.
- **Distribution**: Different distributions affect how the expected length behaves.

General Formula for Expected Length

For most confidence intervals, the expected length can be written as:
Expected Length = 2 × Critical Value × Standard Error text{Expected Length} = 2 times text{Critical Value} times text Standard Error}Expected Length=2 × Critical Value × Standard Error.

Where:

- **Critical Value**: Determines how wide the interval is for a given confidence level (e.g., $Z_{\alpha/2}$)
- **Standard Error (S.E.)**: A measure of variability, typically S.E.= $\sigma/\sqrt{n}$

Minimizing Expected Length - Sample Size

- ***Sample Size**: Larger samples result in smaller standard errors and, therefore, narrower confidence intervals.*

- *For most estimators, the expected length of the CI is inversely proportional to the square root of the sample size:*
- *Expected Length* $\alpha/\sqrt{n}$

- **Example**:
  If you increase your sample size from 100 to 400, the expected length of the CI will decrease by a factor of 2, making the estimate more precise.

  Choosing the Optimal Confidence Level
- **Confidence Level (1−α1-\alpha1−α)**: Higher confidence levels result in wider intervals, and lower levels result in narrower intervals.
  For instance:
  95% Confidence: The critical value $z_{\alpha/2} \approx 1.96$.
  99% Confidence: The critical value $z_{\alpha/2} \approx 2.576$

- **Trade-off**: While increasing the confidence level increases certainty, it also increases the CI's length. To minimize expected length, choose a confidence level that balances precision with confidence.

Minimizing Expected Length in Normal Distribution

- Normal Distribution (Known σ):

$$\mu \pm z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

- The expected length depends on the sample size and the **critical value**.

Strategy:
Z minimize the expected length:

- Use large sample sizes (nnn).
- Choose the desired confidence level based on the application needs.
  Theoretical Background
  Confidence Intervals and Their Properties
  A confidence interval for a parameter θ is an interval (L, U) constructed from sample data such that the probability of containing the true parameter value is at least $1 - \alpha$: $P(L \leq \theta \leq U) = 1 - \alpha$.
  Expected Length of a Confidence Interval
  The expected length of a confidence interval is given by: $E(U - L)$, where U and L are the upper and lower bounds, respectively. The expected length depends on factors such as sample size, confidence level, population variance, and estimation method.
  Factors Affecting Confidence Interval Length

1. **Sample Size (n):** Larger samples lead to narrower intervals due to reduced standard errors.
2. **Confidence Level** $(1 - \alpha)$**:** Higher confidence levels require wider intervals to maintain coverage probability.

3. **Population Variability:** Higher variance in data results in wider confidence intervals.
4. **Estimation Method:** Different estimation techniques impact the length and efficiency of confidence intervals.

Common Confidence Interval Methods

Normal Approximation Interval

For a population mean μ with known variance σ^2, the confidence interval is:

$\bar{X} \pm Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$, where $Z_{\alpha/2}$ is the critical value from the standard normal distribution.

t-Distribution Interval

When the population variance is unknown, the interval is: $\bar{X} \pm t_{\alpha/2,n-1} \cdot \frac{s}{\sqrt{n}}$, where s is the sample standard deviation and $t_{\alpha/2,n-1}$ is the critical value from the t-distribution.

Bootstrap Confidence Intervals

Non-parametric methods such as bootstrapping estimate confidence intervals without assuming normality, often leading to different expected lengths compared to parametric methods.

Applications of Confidence Intervals

1. **Medical Research:** Estimating treatment effects and survival rates.
2. **Economics:** Estimating population means and proportions in survey analysis.
3. **Engineering:** Quality control and reliability testing.
4. **Social Sciences:** Estimating public opinion metrics.

14.7.1 Practical Example:

To illustrate, consider a study estimating the average height of adult males. Given a sample size of 100, a standard deviation of 6 cm, and a 95% confidence level, the confidence interval is: $\bar{X} \pm 1.96 \cdot \frac{6}{\sqrt{100}} = \bar{X} \pm 1.176$. The expected length is approximately 2.352 cm.

Discussion and Limitations

The expected length of confidence intervals is influenced by model assumptions, sample characteristics, and real-world constraints. Over-reliance on asymptotic properties in small samples may lead to misleading inferences. Further research is needed to develop adaptive methods for interval estimation.

Conclusion

Understanding the expected length of confidence intervals is crucial for precise statistical inference. Factors such as sample size, confidence level, and data variability play key roles in determining interval length. Proper application of confidence intervals ensures reliable estimation across various domains.

14.6 KEY WORDS:

- Interval Estimation
- Confidence Interval (CI)
- Confidence Level
- Pivot (Pivotal Quantity)
- Margin of Error
- Standard Error (S.E.)
- Shortest Expected Length Confidence Interval

14.7 SUMMARY :

1. **CONFIDENCE INTERVALS (CIS)**

   Definition: A confidence interval provides a range of plausible values for an unknown population parameter (like a mean or proportion) with a specified confidence level (e.g., 95%).

   Interpretation: A 95% CI means that if we repeated the sampling and interval construction many times, approximately 95% of those intervals would contain the true parameter.

   CIs are a fundamental tool in inferential statistics for expressing estimation uncertainty.

2. USING PIVOTS TO CONSTRUCT CONFIDENCE INTERVALS

- A **pivot** is a function of the observed data and the parameter of interest whose probability distribution does not depend on unknown parameters.
- Because its distribution is known and parameter-free, pivots are extremely useful for creating exact confidence intervals.
- Pivot-based methods often yield exact confidence intervals, unlike approximate methods relying on asymptotic normality.

3. SHORTEST EXPECTED LENGTH CONFIDENCE INTERVALS

- Among all confidence intervals with the same confidence level, some intervals have shorter expected lengths (i.e., are more precise on average).
- **Shortest Expected Length CIs** minimize the average width, enhancing estimation precision without sacrificing coverage.
- These intervals are often derived using:
  - **Bayesian techniques** (e.g., Highest Posterior Density intervals)
  - **Neyman's construction** with ordering rules for acceptance regions
  - **Likelihood ratio methods** for optimality
- Finding these intervals can be more computationally intensive but provides the most efficient use of data.

14.8 SELF ASSESSMENT QUESTIONS:

1. Derive confidence Intervals for one parameter exponential distribution.
2. Describe shortest expected length confidence intervals.
3. Describe the pivotal quantity method of finding confidence intervals.
4. Explain the steps to construct a confidence interval with suitable example.
5. Let $X_1, X_2, ..., X_n$ be a random sample from $N(\theta, \sigma^2)$ where σ^2 is known. Obtain the shortest expected length confidence interval for θ.
6. How does the confidence interval change when the variance σ^2 is unknown and must be estimated from the sample?
7. Explain why pivots are useful in deriving confidence intervals.
8. Compare the efficiency of shortest expected length confidence intervals with equal-tailed confidence intervals.
9. Explain the trade-off between interval length and confidence level.
10. Explain how to use bootstrap methods to approximate confidence intervals when a pivot is not available.

14.9 SUGGESTED READINGS:

1. Lehman, E. L., and Cassella, G. (1998). Theory of Point Estimation, Second Edition, Springer, NY.
2. Casella, G. & Berger, R.L. (2013). Statistical Inference, 2nd Edition. Cengage Learning
3. Goon, A. M., Gupta, M. K., and Dasgupta, B. (1989). An Outline of Statistical Theory-Vol.II, World Press, Calcutta.
4. Kale, B. K. (1999). A First Course on Parametric Inference, Narosa Publishing House, New Delhi.
5. An Introduction to Probability and Statistics by V.K. Rohatgi and. K. Md. E. Saleh (2001).

Prof. G. V. S. R. Anjaneyulu