U.P. Rajarshi Tandon Open University, Prayagraj

MScSTAT – 102N / MASTAT – 102N
Statistical Inference

Course Design Committee

Prof. Ashutosh Gupta **Chairman**
Director, School of Sciences, U. P. Rajarshi Tandon Open University, Prayagraj

Prof. Anup Chaturvedi (Retd.) **Member**
Ex. Head, Department of Statistics, University of Allahabad, Prayagraj

Prof. S. Lalitha (Retd.) **Member**
Ex. Head, Department of Statistics, University of Allahabad, Prayagraj

Prof. Himanshu Pandey **Member**
Department of Statistics, D. D. U. Gorakhpur University, Gorakhpur

Prof. Shruti **Member-Secretary**
Professor, School of Sciences, U.P. Rajarshi Tandon Open University, Prayagraj

Course Preparation Committee

Prof. Shashi Bhushan (Unit 1, 2, 3, 4, 5, 6 & 7) **Writer**
Department of Statistics, University of Lucknow, Lucknow

Dr. Pratyasha Tripathi (Unit 8, 9, 10 & 11) **Writer**
Department of Statistics, Tilka Manjhi Bhagalpur University, Bhagalpur, Bihar

Dr. Prabhat Kr. Singh (Unit 12) **Writer**
Department of Statistics, Jamshedpur Co-Operative College, Jamshedpur

Dr. Sunit Kumar (Unit 13) **Writer**
Department of Statistics, Central University of South Bihar, Gaya, Bihar

Prof. S. K. Pandey (Unit 1, 2, 3, 4, 5, 6 & 7) **Editor**
Department of Statistics, University of Lucknow, Lucknow

Prof. Shruti **(**Unit 8, 9, 10, 11, 12 & 13) **Editor**
School of Sciences, U. P. Rajarshi Tandon Open University, Prayagraj

Prof. Shruti **Course Coordinator**
School of Sciences, U. P. Rajarshi Tandon Open University, Prayagraj

MScSTAT – 102N/ MASTAT – 102N STATISTICAL INFERENCE

First Edition: ***July 2024***
ISBN : 978-93-48987-21-1

Printed and Published by Mr. Vinay Kumar, Registrar, Uttar Pradesh Rajarshi Tandon Open University, 2025.

Printed By : Cygnus Information Solution Pvt.Ltd., Lodha Supremus Saki Vihar Road, Andheri East, Mumbai.

Blocks & Units Introduction

The present SLM on ***Statistical Inference*** consists of eleven units with three blocks.

The ***Block - 1 – Point Estimation,*** is the first block, which is divided into four units.

The ***Unit - 1 – Sufficiency, Minimal Sufficiency & Completeness,*** is the first unit of present self-learning material, which describes the concept of data reduction, sufficiency, minimal sufficient statistic, methods of finding sufficient and minimal sufficient statistic, concept of complete statistic and complete family of distributions along with relevant examples. This unit also introduces ancillary statistic and Basu's Theorem.

In ***Unit – 2 – Unbiased Estimation, Minimum Variance & CRLB,*** the main emphasis of this unit is on the unbiased estimators, how to identify the best unbiased estimators, the concept of minimum variance unbiased estimator and methods of obtaining it using Cramer Rao Inequality as well as by using Lehmann Scheffe & Rao Blackwell Theorems.

In ***Unit – 3 – Maximum Likelihood Estimation, Consistency & Other Lower Bounds,*** we have focussed mainly on maximum likelihood estimation, its importance and using numerical analysis-based solution if closed form cannot be easily obtained like Newton Raphson method. The concept of consistency and other large sample properties. The Bhattacharya and Chapman – Robbin – Kiefer Bounds have been discussed in this unit.

The ***Block - 2 – Testing of Hypotheses & Interval Estimation*** is the second block with four units.

In ***Unit – 4 – Basic Concepts*** discusses the fundamental concepts and definitions, concept of randomized test using test functions. The concept of UMP test using a BCR has been introduced.

In ***Unit – 5 – Generalized Neyman Pearson Lemma & UMPU*** has been discussed, Generalized NP Lemma has been derived and its utility in finding UMPU where UMP test

does not exist has been extensively discussed. The general result for UMPU Test for exponential family has been derived & discussed

In ***Unit – 6 – Likelihood Ratio Tests*** dealt with derivation of some important LR test.

In ***Unit – 7 – Interval Estimation*** dealt with Confidence estimation and the relationship between UMP test and UMA confidence intervals.

The ***Block - 3 – Testing of Hypotheses & Confidence Estimation*** is the second block with four units.

In ***Unit – 8 –Methods of Estimation – I*** & ***Unit – 9 –Methods of Estimation – II,*** deals with the Maximum likelihood estimation, method of moments, MVUE, necessary and sufficient conditions for MVUE, etc., Zehna theorem for invariance, Cramer theorem for weak consistency. Cramer-Huzurbazar theorem

In ***Unit – 10 – Criterion for Good Estimators – I*** & ***Unit – 11 – Criterion for Good Estimators - II*** discussed about the Criterion for Good Estimators, Bhattacharya bound, Chapman Robbins and Kiefer (CRK) bound, asymptotic normality, BAN and CAN estimators, asymptotic efficiency, equivariant consistency.

In ***Unit – 12 – Confidence Estimation*** discussed about the Confidence interval and confidence coefficient, shortest length confidence interval, relation between confidence estimation and hypotheses testing.

In ***Unit – 13 – Hypothesis Testing*** defined about the Generalized Neyman Pearson lemma, MP and UMP tests for distributions with MLR, LR tests and their properties, UMPU tests, similar regions, Neyman structure, Invariant tests.

At the end of every block/unit the summary, self-assessment questions and further readings are given.

U.P. Rajarshi Tandon Open University, Prayagraj

MScSTAT – 102N / MASTAT – 102N
Statistical Inference

Course Design Committee

Prof. Ashutosh Gupta **Chairman**
Director, School of Sciences
U. P. Rajarshi Tandon Open University, Prayagraj

Prof. Anup Chaturvedi (Retd.) **Member**
Ex. Head, Department of Statistics
University of Allahabad, Prayagraj

Prof. S. Lalitha (Retd.) **Member**
Ex. Head, Department of Statistics
University of Allahabad, Prayagraj

Prof. Himanshu Pandey **Member**
Department of Statistics, D. D. U. Gorakhpur University, Gorakhpur

Prof. Shruti **Member-Secretary**
Professor, School of Sciences
U.P. Rajarshi Tandon Open University, Prayagraj

Course Preparation Committee

Prof. Shashi Bhushan **Writer**
Department of Statistics
University of Lucknow, Lucknow

Prof. S. K. Pandey **Editor**
Department of Statistics
University of Lucknow, Lucknow

Prof. Shruti **Course Coordinator**
School of Sciences
U. P. Rajarshi Tandon Open University, Prayagraj

MScSTAT – 102N/ MASTAT – 102N **STATISTICAL INFERENCE**

First Edition: ***July 2024***
ISBN : 978-93-48987-21-1

Printed and Published by Mr. Vinay Kumar, Registrar, Uttar Pradesh Rajarshi Tandon Open University, 2025.

Block and Units Introduction

The ***Block - 1 – Point Estimation,*** is the first block, which is divided into three units.

The ***Unit - 1 – Sufficiency, Minimal Sufficiency & Completeness,*** is the first unit of present self-learning material, which describes the concept of data reduction, sufficiency, minimal sufficient statistic, methods of finding sufficient and minimal sufficient statistic, concept of complete statistic and complete family of distributions along with relevant examples. This unit also introduces ancillary statistic and Basu's Theorem.

In ***Unit – 2 – Unbiased Estimation, Minimum Variance & CRLB,*** the main emphasis of this unit is on the unbiased estimators, how to identify the best unbiased estimators, the concept of minimum variance unbiased estimator and methods of obtaining it using Cramer Rao Inequality as well as by using Lehmann Scheffe & Rao Blackwell Theorems.

In ***Unit – 3 – Maximum Likelihood Estimation, Consistency & Other Lower Bounds,*** we have focussed mainly on maximum likelihood estimation, its importance and using numerical analysis-based solution if closed form cannot be easily obtained like Newton Raphson method. The concept of consistency and other large sample properties. The Bhattacharya and Chapman – Robbin – Kiefer Bounds have been discussed in this unit.

At the end of every block/unit the summary, self-assessment questions and further readings are given.

References and Further Readings

- Kale B. K.: A first Course on Parametric Inference, Narosa Publishing House.
- Mukhopadhyay P.: Mathematical Statistics, New Central Book Agency, Kolkata.
- Zacks S.: Theory of statistical Inference, John Wiley & Sons, New York.
- Rao C.R.: Linear Statistical inference and its applications, John Wiley & Sons
- Rudin, W.: Principles of Mathematical Analysis, McGraw
- Mood, A.M., Graybill, F.A. and Boes, D.C.: Introduction to the Theory of Statistics, McGraw Hill
- Rohatgi, V.K.: An Introduction to Probability Theory and Mathematical Statistics, McGraw Hill Education
- Lehmann, E. L. and Casella, G.C: Theory of Point Estimation, Springer
- Lehmann, E. L. and Romano, J. P.: Testing of Statistical Hypotheses, Springer
- Dudewicz, E.J. and Mishra, S.N.: Modern Mathematics Statistics, Wiley.
- Cramer H.: Mathematical Methods of Statistics, Princeton.

UNIT 1: SUFFICIENCY, MINIMAL SUFFICIENCY & COMPLETENESS

Structure

1.1 Introduction

Let X be a random variable with pdf $f(.,\theta)$ of the known functional form but depending upon an unknown constant θ which is called a **parameter**. We let **Ω** stand for the set of all possible values of θ and call it the **parameter space.** By $\Im$ we denote the family of all pdf's we get by letting θ vary over **Ω**; that is $\Im=\{f(.,\theta): \theta \, \varepsilon \, \Omega\}$. Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from $f(.,\theta)$, that is, n independent random variables distributed as X above. One of basic problem of Statistics is that of making inference about the parameter θ (i.e. estimating θ by a point value or by an interval with known probability, testing hypotheses about θ etc.) on the basis of the observed values $x_1, x_2, \ldots, x_n$, the data, of the random variables $X_1, X_2, \ldots, X_n$.

1.2 Objectives

The objectives of this unit is to provide us a basic understanding of the concepts related to Sufficiency, Minimal Sufficiency & Completeness. The concepts of random sample, data reduction, sufficiency, minimal sufficient statistic, methods of finding sufficient and minimal sufficient statistic, concept of complete statistic, complete family of distributions, ancillary statistic and Basu's Theorem should be clear after reading this material.

1.3 Random Sample

Definition: The random variables $X_1, \ldots\ldots, X_n$ are called a *random sample of size n from the population* $f(x)$ if $X_1, \ldots\ldots, X_n$ are mutually independent random variables and the marginal pdf or pmf of each X_i is the same function$f(x)$. Alternatively, $X_1, \ldots., X_n$ are called *independent and identically distributed random variables with pdf or pmf* $f(x)$.This is commonly abbreviated to iid random variables.

$$f(x_1, \ldots\ldots, x_n) = f(x_1)f(x_2) \ldots\ldots f(x_n) = \prod_{i=1}^{n} f(x_i) \qquad (1)$$

$$f(x_1, \ldots\ldots, x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta), \qquad (2)$$

where the same parameter value θ is used in each of the terms in the product. If, in a statistical setting, we assume that the population we are observing is a member of a specified parametric family but the true parameter value is unknown, then a random sample from this population has a joint pdf or pmf of the above form with the value form with the value of θ unknown. By considering different possible values of θ, we can study how a random sample would behave for different populations.

Example (Sample pdf-Exponential): Let $X_1, \ldots\ldots, X_n$ be a random sample from an exponential(β) population. Specifically, $X_1, \ldots\ldots, X_n$ might correspond to the times until failure (measured in years) for n identical circuit boards that are put on test and used until they fail. The joint pdf of the sample is

$$f(x_1, \ldots\ldots, x_n|\beta) = \prod_{i=1}^{n} f(x_i|\beta) = \prod_{i=1}^{n} \frac{1}{\beta} e^{-x_i/\beta} = \frac{1}{\beta^n} e^{-(x_1+\cdots+x_n)/\beta}$$

This pdf can be used to answer questions about the sample. For example, what is the probability that all the boards last more than 2 years? We can compute

$$P(X_1 > 2, \ldots, X_n > 2) = \int_2^{\infty} \ldots\ldots \int_2^{\infty} \prod_{i=1}^{n} \frac{1}{\beta} e^{-x_i/\beta} dx_1 \ldots dx_n$$

$= e^{-2/\beta} \int_2^{\infty} \ldots\ldots \int_2^{\infty} \prod_{i=2}^{n} \frac{1}{\beta} e^{-x_i/\beta} dx_2 \ldots dx_n$ (integrate out x_1)

$\vdots$ (integrate out the remaining $x_i's$ successively)

$= \left(e^{-2/\beta}\right)^n$

$= e^{-2n/\beta}$

If β, the average lifelength of a circuit board, is large relative to n, we see that this probability is near 1.

$P(X_1 > 2, \ldots\ldots, X_n > 2) = P(X_1 > 2) \ldots\ldots P(X_n > 2)$ (independence)

$= [P(X_1 > 2)]^n$ (identical distributions)

$= \left(e^{-2/\beta}\right)^n$ (exponential calculation)

$= e^{-2n/\beta}$

The random sampling model in Definition 1 is sometimes called sampling from an *infinite* population. Think of obtaining the values of $X_1, \ldots\ldots, X_n$ sequentially. First, the experiment is performed and $X_1 = x_1$ is observed. Then, the experiment is repeated and $X_2 = x_2$ is observed. The assumption of independence in random sampling implies that the probability distribution for X_2 is unaffected by the fact that $X_1 = x_1$ was observed first. "Removing" x_1 from the infinite population does not change the population, so $X_2 = x_2$ is still a random observation from the same population.

When sampling is from a *finite* population, Definition 1 may or may not be relevant depending on how the data collection is done. A finite population is a finite set of numbers, $\{x_1, \ldots \ldots, x_N\}$. A sample $X_1, \ldots \ldots, X_n$ is to be drawn from this population.

Suppose a value is chosen from the population in such a way that each of the N values is equally likely (probability $= 1/N$) to be chosen. (Think of drawing numbers from a hat.) This value is recorded as $X_1 = x_1$. Then the process is repeated. Again, each of the N values is equally likely to be chosen. The second value chosen is recorded as $X_2 = x_2$. (If the same number is chosen, then $x_1 = x_2$.) This process of drawing from the N values is repeated n times , yielding the sample $X_1, \ldots \ldots, X_n$. This kind of sampling is called *with replacement* because the value chosen at any stage is "replaced" in the population and is available for choice again at the next stage. For this kind of sampling, the conditions of Definition 1 are met. Each X_i is a discrete random variable that takes on each of the values $x_1, \ldots \ldots, x_N$ with equal probability. The random variables $X_1, \ldots \ldots, X_n$ are independent because the process of choosing any X_i is the same, regardless of the values that are chosen for any of the other variables.

A second method for drawing a random sample from a finite population is called sampling *without replacement*. Sampling without replacement is done as follows. A value is chosen from $(x_1, \ldots \ldots, x_N)$ in such a way that each of the N values has probabaility $1/N$ of being chosen. This value is recorded as $X_1 = x_1$. Now a second value is chosen from the remaining $N - 1$ values have probability $1/(N - 1)$ of being chosen. The second chosen value is recorded as $X_2 = x_2$. Choice of the remaining values continues in this way, yielding the sample $X_1, \ldots \ldots, X_n$. But once a value is chosen, it is unavailable for choice at any later stage.

A sample drawn from a finite population without replacement does not satisfy all the conditions of Definition 1. The random variables $X_1, \ldots \ldots, X_n$ are not mutually independent. To see this, let x and y be the distinct elements of $\{x_1, \ldots x_N\}$. Then $P(X_2 = y | X_1 = y) = 0$, since the value y cannot be chosen at the second stage if it was already chosen at the first. However, $P(X_2 = y | X_1 = x) = 1/(N - 1)$. The probability distribution for X_2 depends on the values of X_1 that is observed and, hence, X_1 and X_2 are not independent. However, it is interesting to note that $X_1, \ldots \ldots X_n$ are identically distributed.

That is, the marginal distribution of X_i is same for each $i = 1, \ldots, n$. For X_1 it is clear that the marginal distribution is $P(X_1 = x) = 1/N$ for each $x \in \{x_1, \ldots, x_N\}$. To compute the marginal distribution for X_2, use the definition of conditional probability to write

$$P(X_2 = x) = \sum_{i=1}^{N} P(X_2 = x | X_1 = x_i) P(X_1 = x_i)$$

For one value of the index, say $k, x = x_k$ and $P(X_2 = x | X_1 = x_k) = 0$. For all other j$\neq k$, $P(X_2 = x | X_1 = x_j) = 1/(N-1)$. Thus,

$$P(X_2 = x) = (N-1)\left(\frac{1}{N-1}\frac{1}{N}\right) = \frac{1}{N} \qquad (3)$$

Similar arguments can be used to show that each of the $X_i's$ has the same marginal distribution.

Sampling without replacement from a finite population is sometimes called *simple random sampling*. It is important to realize that this is not the same sampling situations as that described in Definition 1. However, if the population size N is large compared to the sample size n, $X_1, \ldots, X_n$ are nearly independent and some approximate probability calculation can be made assuming they are independent. By saying they are "nearly independent" we simply mean that the conditional distribution of X_i given $X_1, \ldots X_{i-1}$ is not too different from the marginal distribution of X_i. For example, the conditional distribution of X_2 given X_1 is

$P(X_2 = x_1 | X_1 = x_1) = 0$ and $P(X_2 = x | X_1 = x_1) = \frac{1}{N-1}$ for $x \neq x_1$.

This is not too different from the marginal distribution of X_2 given in (3) if N is large. The nonzero probabilities in the conditional distribution of X_i given $X_1, \ldots, X_{i-1}$ are $1/(N-i+1)$, which are close to $1/N$ if $i \leq n$ is small compared with N.

When a sample $X_1, \ldots \ldots, X_n$ is drawn, some summary of the values is usually computed. Any well-defined summary of the value is usually computed. Any well-defined summary may be expressed mathematically as a function $T(x_1, \ldots, x_n)$ whose domain includes the sample space of the random vector $(X_1, \ldots, X_n)$. The function T may be real-valued or vector-valued; thus the summary is a random variable(or vector), $Y = T(X_1, \ldots, X_n)$.

This is used to describe the distribution of Y in terms of the distribution of the population from which the sample was obtained. Since the random sample $X_1, \ldots\ldots, X_n$ has a simple probabilistic structure (because the $X_i's$ are independent and identically distributed), the distribution of Y is particularly tractable. Because this distribution is usually derived from the distribution of the variable in the random sample, it is called the *sampling distribution* of Y. This distinguishes the probability distribution of Y from the distribution of the population of the population, that is, the marginal distribution of each X_i.

Definition: Let $X_1, \ldots\ldots, X_n$ be a random sample of size n from a population and let $T(x_1, \ldots\ldots, x_n)$ be a real-valued or vector-valued function whose domain includes the sample space of $(X_1, \ldots\ldots, X_n)$. Then the random variable or random vector $Y = T(X_1, \ldots\ldots, X_n)$ is called a *statistic*. The probability distribution of a statistic Y is called the *sampling distribution* of Y.

The definition of a statistic is very broad, with the only restriction being that a statistic cannot be the function of unknown parameter. The sample summary given by a statistic can include many types of information. For example, it may give the smallest or largest value in the sample, the average sample value, or a measure of the variability in the sample observations. Three statistics that are often used and provide good summaries of the sample are now defined.

1.4 Some Important Definitions and Results

Definition: The *sample mean* is the arithmetic average of the values in a random sample. It is usually denoted by

$$\bar{X} = \frac{X_1 + \cdots + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Definition: The *sample variance* is the statistic defined by

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n} (X_i - \bar{X})^2$$

The *sample standard deviation* is the statistic defined by $S = \sqrt{S^2}$

Theorem: Let $x_1, \ldots, x_n$ be any numbers and $\bar{x} = (x_1 + \cdots + x_n)/n$. Then

a. $\min_a \sum_{i=1}^{n} (x_i - a)^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2$,

b. $(n-1)s^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$

Proof: To prove part (a), add and subtract $\bar{x}$ to get

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

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

$$= \sum_{i=1}^{n} (x_i - \bar{x})^2 + \sum_{i=1}^{n} (\bar{x} - a)^2 \qquad \text{(cross term is 0)}$$

It is now clear that the right-hand side is minimized at $a = \bar{x}$.

To prove part (b), take $a = 0$ in the above.

Lemma: Let $X_1, \ldots, X_n$ be a random sample from a population and let $g(x)$ be a function such that $Eg(X_1)$ and $Var\ g(X_1)$ exist. Then

$$E\left(\sum_{i=1}^{n} g(X_i)\right) = n(Eg(X_1)) \qquad (1)$$

and

$$Var\left(\sum_{i=1}^{n} g(X_i)\right) = n\big(Var\ g(X_1)\big) \qquad (2)$$

Proof: To prove (1), note that

$$E\left(\sum_{i=1}^{n} g(X_i)\right) = \sum_{i=1}^{n} Eg(X_i) = n(Eg(X_1)).$$

Since the $X_i s$ are identically distributed, the second equality is true because $Eg(X_i)$ is the same for all i. Note that the independence of $X_1, \ldots\ldots, X_n$ is not needed for (1)

To prove (2), note that

$Var(\sum_{i=1}^{n} g(X_i)) = E[\sum_{i=1}^{n} g(X_i) - E(\sum_{i=1}^{n} g(X_i))]^2$(definition of variance)

$$= E\left[\sum_{i=1}^{n}\left(g(X_i) - Eg(X_i)\right)\right]^2 \quad \text{(expectation property)}$$

In this last expression there are n^2 terms. First, there are n terms $\left(g(X_i) - Eg(X_i)\right)^2, i = 1, \ldots\ldots, n,$ and for each, we have

$E\left(g(X_i) - Eg(X_i)\right)^2 = Var\ g(X_i)$ (definition of variance)

$= Var\ g(X_1).$ (identically distributed)

The remaining $n(n-1)$ terms are all of the form $\left(g(X_i) - Eg(X_i)\right)\left(g(X_j) - Eg(X_j)\right)$, with $i \neq j$. For each term,

$$E\left[\left(g(X_i) - Eg(X_i)\right)\left(g(X_j) - Eg(X_j)\right)\right] = Cov\left(g(X_i), g(X_j)\right)$$

$= 0$ (independence)

Thus, we obtain equation (2).

Theorem: Let $X_1, \ldots\ldots, X_n$ be a random sample from a population with mean μ and variance $\sigma^2 < \infty$. Then

a. $E\bar{X} = \mu$,

b. $Var\ \bar{X} = \frac{\sigma^2}{n}$,

c. $ES^2 = \sigma^2$.

Proof: To prove (a), let $g(X_i) = X_i/n,$ so $Eg(X_i) = \mu/n$. Then, by Lemma 1,

$$E\bar{X} = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}E\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n}nEX_1 = \mu.$$

Similarly, for (b), we have

$$Var\,\bar{X} = Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}Var\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}n\,Var\,X_1 = \frac{\sigma^2}{n}.$$

For the sample variance, using Theorem 1, we have

$$ES^2 = E\left(\frac{1}{n-1}\left[\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right]\right)$$

$$= \frac{1}{n-1}(nEX_1^2 - nE\bar{X}^2)$$

$$= \frac{1}{n-1}\left(n(\sigma^2 + \mu^2) - n\left(\frac{\sigma^2}{n} + \mu^2\right)\right) = \sigma^2,$$

establishing part(c) and proving the theorem.

Theorem: Let $X_1, \ldots\ldots, X_n$ be a random sample from a population with mgf $M_X(t)$. Then the mgf of the sample mean is

$$M_{\bar{X}}(t) = [M_X(t/n)]^n.$$

Example (Distribution of the mean): Let $X_1, \ldots\ldots, X_n$ be a random sample from a $n(\mu, \sigma^2)$ population. Then the mgf of the sample mean is

$$M_{\bar{X}}(t) = \left[exp\left(\mu\frac{t}{n} + \frac{\sigma^2(t/n)^2}{2}\right)\right]^n$$

$$= exp\left(n\left(\mu\frac{t}{n} + \frac{\sigma^2(t/n)^2}{2}\right)\right) = exp\left(\mu t + \frac{(\sigma^2/n)t^2}{2}\right)$$

Thus, $\bar{X}$ has a $n(\mu, \sigma^2/n)$ distribution.

Theorem: Suppose $X_1, \ldots \ldots, X_n$ is a random sample from a pdf or pmf $f(x|\theta)$ where

$$f(x|\theta) = h(x)c(\theta)exp\left(\sum_{i=1}^{k} w_i(\theta)t_i(x)\right)$$

is a member of an exponential family. Define statistics $T_1, \ldots \ldots, T_k$ by

$$T(X_1, \ldots \ldots, X_n = \sum_{j=1}^{n} t_i\left(X_j\right), \;\; i = 1, \ldots, k.$$

If the set $\{(w_1(\theta), w_2(\theta), \ldots \ldots, w_k(\theta)), \theta \in \Theta\}$ contains an open subset of R^{k}, then the distribution of $(T_1, \ldots \ldots, T_k)$ is an exponential family of the form

$$f_T(\mu_1, \ldots \ldots, \mu_k|\theta) = H(u_1, \ldots \ldots, u_k)[c(\theta)]^n exp\left(\sum_{i=1}^{k} w_i(\theta)u_i\right) \qquad (6)$$

The open set condition eliminates a density such as the $n(\theta, \theta^2)$ and, in general, eliminates curved exponential families from Theorem 4.

Example (Sum of Bernoulli random variables): Suppose $X_1, \ldots \ldots, X_n$ is a random sample from a Bernoulli(p) distribution. We know that a Bernoulli(p) distribution is an exponential family with $k = 1, c(p) = (1-p), w_1(p) = \log(p/(1-p))$, and $t_1(x) = x$. Thus, by previous theorem, $T_1 = T_1(X_1, \ldots \ldots, X_n) = X_1 + \cdots + X_n$. From the definition of the binomial distribution, we know that T_1 has a binomial (n, p) distribution. Also, we know that a binomial(n, p) distribution is an exponential family with the same $w_1(p)$ and $c(p) = (1-p)^n$. Thus expression (6) is verified for this example.

Theorem: If X and Y are two random variables, then

$$EX = E\big(E(X|Y)\big), \qquad (1)$$

provided that the expectation exists.

Proof: Let $f(x, y)$denote the joint pdf of X and Y. By definition, we have

$$EX = \iint xf(x,y)dxdy = \int [\int xf(x|y)dx] f_Y(y)dy, \qquad (2)$$

where $f(x|y)$ and $f_Y(y)$ are the conditional pdf of X and $Y = y$ and the marginal pdf of Y, respectively. Now, notice that the inner integral in (2) is the conditional expectation $E(X|y)$, and we have

$$EX = \int E(X|y) f_Y(y)dy = E(E(X|Y))$$

as desired. Replace integrals by sums to prove the discrete case.

Theorem (Conditional variance identity): For any two random variables X and Y,

$$Var\ X = E\big(Var(X|Y)\big) + Var\big(E(X|Y)\big) \qquad (4)$$

provided that the expectations exist.

Proof: By definition, we have

$$Var\ X = E([X - EX]^2) = E([X - E(X|Y) + E(X|Y) - EX]^2),$$

where in the last step we have added and subtracted $E(X|Y)$.Expanding the square in this last expectation now gives

$$Var\ X = E([X - E(X|Y)]^2) + E([E(X|Y) - EX]^2)$$

$$+2E([X - E(X|Y)][E(X|Y) - EX]). \qquad (5)$$

The last term in this expression is equal to 0, however, which can easily be seen by iterating the expectation:

$$E([X - E(X|Y)][E(X|Y) - EX]) = E(E\{[X - E(X|Y)][E(X|Y) - EX]|Y\}) \qquad (6)$$

In the conditional distribution $X|Y$, X is the random variable. So in the expression

$$E\{[X - E(X|Y)][E(X|Y) - EX]|Y\},$$

$E(X|Y)$ and EX are constants. Thus,

$$E\{[X - E(X|Y)][E(X|Y) - EX]|Y\} = (E(X|Y) - EX)(E\{[X - E(X|Y)]|Y\})$$

$$= (E(X|Y) - EX)(E(X|Y) - E(X|Y))$$

$$= (E(X|Y) - EX)(0)$$

$$= 0.$$

Thus, from (6) we have that $E\left(\left(X - E(X|Y)\right)(E(X|Y) - EX)\right) = E(0) = 0.$

Referring back to equation (5), we see that

$$E([X - E(X|Y)]^2) = E(E\{[X - E(X|Y)]^2|Y\})$$

$$= E(Var(X|Y))$$

and

$$E([E(X|Y) - EX]^2) = Var\big(E(X|Y)\big),$$

establishing

$$Var\ X = E\big(Var(X|Y)\big) + Var(E(X|Y))$$

Definition: Let $\boldsymbol{X_1}, \ldots, \boldsymbol{X_n}$ be random vectors with joint pdf or pmf $f(\boldsymbol{x_1}, \ldots, \boldsymbol{x_n})$. Let $f_{X_i}(\boldsymbol{x_i})$ denote the marginal pdf or pmf of $\boldsymbol{X}_i$. Then $\boldsymbol{X_1}, \ldots, \boldsymbol{X_n}$ are called *mutually independent random vectors* if, for every $(\boldsymbol{x_1}, \ldots, \boldsymbol{x_n})$,

$$f(\boldsymbol{x_1}, \ldots, \boldsymbol{x_n}) = f_{X1}(\boldsymbol{x_1}) \ldots \ldots f_{X_n}(\boldsymbol{x_n}) = \prod_{i=1}^{n} \boldsymbol{f_{X_i}}(\boldsymbol{x_i}).$$

If the $X_i's$ are all one-dimensonal, then $X_1, \ldots, X_n$ are called *mutually independent random variables.*

Theorem: Let $X_1, \ldots, X_n$ be mutually independent random variables. Let $g_1, \ldots, g_n$ be real-valued functions such that $g_i(x_i)$ is a function only $x_i,\ i = 1, \ldots, n$. Then

$$E\big(g_1(X_1) \ldots . g_n(X_n)\big) = \big(Eg_1(X_1)\big) \ldots . \big(Eg_n(X_n)\big).$$

Theorem: Let $X_1, \ldots, X_n$ be mutually independent random variables with mgfs $M_{X_1}(t), \ldots., M_{X_n}(t)$. Let $Z = X_1 + \cdots + X_n$. Then the mgf of Z is

$$M_Z(t) = M_{X_1}(t) \ldots \ldots M_{X_n}(t).$$

In particular, if $X_1 \ldots . X_n$ all have the same distribution with mgf $M_X(t)$, then

$$M_Z(t) = \big(M_X(t)\big)^n.$$

Lemma: Let (X, Y)be a bivariate random vector with joint pdf or pmf $f(x, y)$. Then X and Y are independent random variables if and only if there exist functions $g(x)$ and $h(y)$ such that, for every $x \in$ R and $y \in$ R,

$$f(x, y) = g(x)h(y)$$

Proof: The "only if" part is proved by defining $g(x) = f_X(x)$ and $h(y) = f_Y(y)$ and using (1). To prove the "if" part for continuous random variables, suppose that $f(x, y) = g(x)h(y)$. Define

$$\int_{-\infty}^{\infty} g(x)dx = c \quad and \quad \int_{-\infty}^{\infty} h(y)dy = d,$$

where the constants c and d satisfy

$$cd = \left(\int_{-\infty}^{\infty} g(x)dx\right)\left(\int_{-\infty}^{\infty} h(y)dy\right)$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x)h(y)\, dx\, dy \qquad (2)$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x, y)\, dx\, dy$$

$$= 1. \qquad (f(x, y) \text{ is a joint pdf})$$

Furthermore, the marginal pdfs are given by

$f_X(x) = \int_{-\infty}^{\infty} g(x)h(y)dy = g(x)d \qquad and$

$f_Y(y) = \int_{-\infty}^{\infty} g(x)h(y)dx = h(y)c.$ (3)

Thus, using (2) and (3), we have

$$f(x,y) = g(x)h(y) = g(x)h(y)cd = f_X(x)f_Y(y),$$

showing that X and Y are independent. Replacing integrals with sums proves the lemma for discrete random vectors.

Theorem: Let $\boldsymbol{X_1}, \ldots, \boldsymbol{X_n}$ be random vectors. Then $X_1, \ldots, X_n$ are mutually independent random vectors if and only if there exist functions $g_i(\boldsymbol{x_i}), i = 1, \ldots, n$, such that joint pdf or pmf of $(\boldsymbol{X_1}, \ldots, \boldsymbol{X_n})$ can be written as

$$f(x_1, \ldots, x_n) = g_1(x_1) \ldots \ldots g_n(x_n).$$

Theorem: Let $\boldsymbol{X_1}, \ldots, \boldsymbol{X_n}$ be independent random vectors. Let $g_i(\boldsymbol{x_i})$ be a function only of $\boldsymbol{x_i}, i = 1, \ldots, n$. Then the random variables $U_i = g_i(\boldsymbol{X_i}), i = 1, \ldots, n$, are mutually independent.

Central Limit Theorem: Let $X_1, X_2, \ldots$ be a sequence of iid random variables with $EX_i = \mu$ and $0 < Var\, X_i = \sigma^2 < \infty$. Define $\bar{X}_n = \left(1/n\right)\sum_{i=1}^{n} X_i$. Let $G_n(x)$ denote the cdf of $\sqrt{n}(\bar{X}_n - \mu)/\sigma$. Then, for any $x, -\infty < x < \infty$,

$$\lim_{n\to\infty} G_n(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-y^2/2}\, dy;$$

that is, $\sqrt{n}(\bar{X} - \mu)/\sigma$ has a limiting standard normal distribution.

Slutsky's Theorem: If $X_n \to X$ in distribution and $Y_n \to a$, a constant in probability, then

$\boldsymbol{a}. Y_n X_n \to aX$ in distribution.

$\boldsymbol{b}. X_n + Y_n \to X + a$ in distribution.

Example (Normal approximation with estimated variance): Suppose that

$\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma} \to N(0,1)$, but the value of σ is known.

$$P(|S_n^2 - \sigma^2| \geq \epsilon) \leq \frac{E(S_n^2 - \sigma^2)^2}{\epsilon^2} = \frac{Var\, S_n^2}{\epsilon^2}$$

If $\lim_{n\to\infty} Var\, S_n^2 = 0$, then $S_n^2 \to \sigma^2$ in probability. $\sigma/S_n \to 1$ in probability. Hence, Slutsky's Theorem tells us

$$\frac{\sqrt{n}(\bar{X} - \mu)}{S_n} = \frac{\sigma}{S_n}\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \to N(0,1).$$

1.5 The Delta Method

Example (Estimating the odds): Suppose we observe $X_1, X_2, \ldots, X_n$ independent Bernoulli(p) random variables. The typical parameter of interest is p, the success probability, but another popular parameter is $\frac{p}{1-p}$, the *odds*. For example, if the data represent the outcomes of a medical treatment with $p = 2/3$, then a person has odds $2:1$ of getting better. Moreover, if there were another treatment with success probability r, biostatisticians often estimate the *odds ratio* $\frac{p}{1-p}/\frac{r}{1-r}$, giving the relative odds of one treatment over another.

As we would typically estimate the success probability p with the observed success probability $\hat{p} = \sum_i X_i/n$, we might consider using $\frac{\hat{p}}{1-\hat{p}}$ as an estimate of $\frac{p}{1-p}$. But what are the properties of this estimator? How might we estimate the variance of $\frac{\hat{p}}{1-\hat{p}}$? Moreover, how can we approximate its sampling distribution?

Intuition abandons us, and exact calculation is relatively hopeless, so we have to rely on an approximation. The Delta Method will allow us to obtain reasonable, approximate answers to our questions.

Definition: If a function $g(x)$ has derivatives of order, that is $g^{(r)}(x) = \frac{d^r}{dx^r}g(x)$ exists, then for any constant a, the *Taylor polynomial of order r about a* is

$$T_r(x) = \sum_{i=0}^{r} \frac{g^{(i)}(a)}{i!}(x-a)^i$$

Taylor's major theorem, which we will not prove here, is that the *remainder* from the approximation, $g(x) - T_r(x)$, always tends to 0 faster than the highest-order explicit term.

Theorem (Taylor): *If* $g^{(r)}(a) = \frac{d^r}{dx^r} g(x)|_{x=a}$ *exists, then*

$$\lim_{x \to a} \frac{g(x) - T_r(x)}{(x-a)^r} = 0$$

In general, we will not be concerned with the explicit form of the remainder one useful one being

$$g(x) - T_r(x) = \int_a^x \frac{g^{r+1}(t)}{r!}(x-t)^r dt$$

For the statistical application of Taylor's Theorem, we are most concerned with the *first-order* Taylor series, that is, an approximation using just the first derivative

Let $T_1, \ldots\ldots, T_k$ be random variables with means $\theta_1, \ldots\ldots, \theta_k$, and define $\boldsymbol{T} = (T_1, \ldots\ldots, T_k)$ and $\boldsymbol{\theta} = (\theta_1, \ldots\ldots, \theta_k)$. Suppose there is a differential function $g(\boldsymbol{T})$ (an estimator of some parameter) for which we want an approximate estimate of variance. Define

$$g_i'(\theta) = \frac{\partial}{\partial t_i} g(t)|_{t_1=\theta_1, \ldots, t_k=\theta_k}$$

The first –order Taylor series expansion of g about θ is

$$g'(t) = g(\theta) + \sum_{i=1}^{k} g_i'(\theta)(t_i - \theta) + remainder$$

For our statistical approximation we forget about the remainder and write

$$g(t) \approx g(\theta) + \sum_{i=1}^{k} g_i'(\theta)(t_i - \theta_i) \qquad (7)$$

Now, take expectations on both sides of (7) to get

$$E_\theta g(T) \approx g(\theta) + \sum_{i=1}^{k} g_i'(\theta) E_\theta(T_i - \theta_i) \qquad (8)$$

$$= g(\theta) \qquad (T_i \text{ has mean } \theta_i)$$

We can now approximate the variance of $g(T)$ by

$$Var_\theta g(T) \approx E_\theta([g(T) - g(\theta)]^2) \qquad \text{(using (8))}$$

$$\approx E_\theta\left(\left(\sum_{i=1}^k g_i'(\theta)(T_i - \theta_i)\right)^2\right) \qquad \text{(using (7))}$$

$$= \sum_{i=1}^k [g_i'(\theta)]^2 Var_\theta T_i + 2\sum_{i>j} g_i'(\theta) g_j'(\theta) Cov_\theta(T_i, T_j), \qquad (9)$$

useful because it gives us a variance formula for a general function, using only simple variance and covariances

Example (Continuation): Recall that we are interested in the properties of $\frac{\hat{p}}{1-\hat{p}}$ as an estimate of $\frac{p}{1-p}$, where p is a binomial success probability. In our above notation, take $g(p) = \frac{p}{1-p}$ so $g'(p) = \frac{1}{(1-p)^2}$ and

$$Var\left(\frac{\hat{p}}{1-\hat{p}}\right) \approx [g'(p)]Var(\hat{p})$$

$$= \left[\frac{1}{(1-p)^2}\right]^2 \frac{p(1-p)}{n} = \frac{p}{n(1-p)^3}$$

giving us an approximation for the variance of our estimator.

Example (Approximate mean and variance): Suppose X is a random variable with $E_\mu g(X) = \mu \neq 0$. If we want to estimate a function $g(\mu)$, a first-order approximation would give us

$$g(X) = g(\mu) + g'(\mu)(X - \mu).$$

If we use $g(X)$ as an estimator of $g(\mu)$, we can say that approximately

$$E_\mu g(X) \approx g(\mu),$$

$$Var_\mu g(X) \approx [g'(\mu)]^2 Var_\mu X.$$

For a specific example, take $g(\mu) = 1/\mu$. We estimate $1/\mu$ with $1/X$, and we can say

$$E_\mu\left(\frac{1}{X}\right) \approx \frac{1}{\mu},$$

$$Var_\mu\left(\frac{1}{X}\right) \approx \left(\frac{1}{\mu}\right)^4 Var_\mu X$$

Using these Taylor series approximations for the means and variance, we get the following useful generalization of the Central Limit Theorem, Known as the *Delta method.*

Theorem (Delta Method): Let Y_n be a sequence of random variables that satisfies $\sqrt{n}(Y_n - \theta) \to N(0, \sigma^2)$ in distribution. For a given functions g and a specific value of θ, suppose that $g'(\theta)$ exists and is not 0. Then

$$\sqrt{n}[g(Y_n) - g(\theta)] \to N(0, \sigma^2[g'(\theta)]^2) \tag{10}$$

in distribution.

Proof: The Taylor expansion of $g(Y_n)$ around $Y_n = \theta$ is

$$g(Y_n) = g(\theta) + g'(\theta)(Y_n - \theta) + remainder, \tag{11}$$

where the remainder $\to 0$ as $Y_n \to \theta$ in probability it follows that the remainder $\to 0$ in probability. By applying Slutsky's Theorem

$$\sqrt{n}[g(Y_n) - g(\theta)] = g'(\theta)\sqrt{n}(Y_n - \theta)$$

the result now follows.

Example (Continuation): Suppose now that we have the mean of a random sample $\bar{X}$.For $\mu \neq 0$, we have

$$\sqrt{n}\left(\frac{1}{\bar{X}} - \frac{1}{\mu}\right) \to N\left(0, \left(\frac{1}{\mu}\right)^4 Var_\mu X_1\right)$$

in distribution.

If we do not know the variance of X_1, to use the above approximation requires an estimate, say S^2. Moreover, there is the equation of what to do with the $1/\mu$ term, as we also do not know μ. We can estimate everything, which give us the approximate variance $\widehat{Var}\left(\frac{1}{\bar{X}}\right) \approx \left(\frac{1}{\bar{X}}\right)^4 S^2$.

Furthermore, as both $\bar{X}$ and S^2 are consistent estimator, we can again apply Slutsky's Theorem to conclude that $\mu \neq 0$,

$$\frac{\sqrt{n}\left(\frac{1}{\bar{X}} - \frac{1}{\mu}\right)}{\left(\frac{1}{\bar{X}}\right)^2 S} \rightarrow N(0,1)$$

in distribution.

There are two extensions of the basic Delta Method that we need to deal with to complete our treatment. The first concerns the possibility that $g'(\mu) = 0$.

If $g'(\theta) = 0$, we take one more term in the Taylor expansion to get

$$g(Y_n) = g(\theta) + g'(\theta)(Y_n - \theta) + \frac{g''(\theta)}{2}(Y_n - \theta)^2 + Remainder.$$

If we do some rearranging (setting $g' = 0$), we have

$$g(Y_n) - g(\theta) = \frac{g''(\theta)}{2}(Y_n - \theta)^2 + Remainder. \qquad (12)$$

Now recall that the square of a $N(0,1)$ is a χ_1^2, which implies that

$$\frac{n(Y_n - \theta)^2}{\sigma^2} \rightarrow \chi_1^2$$

Theorem (second–order Delta Method): Let Y_n be a sequence of random variables that satisfies $\sqrt{n}(Y_n - \theta) \rightarrow N(0, \sigma^2)$ in distribution. For a given function g and a specific value of θ, suppose that $g'(\theta) = 0$ and $g''(\theta)$ exists and is not 0. Then

$$n[g(Y_n) - g(\theta)] \rightarrow \sigma^2 \frac{g''(\theta)}{2} \chi_1^2 \qquad (13)$$

in distribution.

Example (Moments of a ratio estimator): Suppose X and Y are random variables with nonzero means μ_X and μ_Y, respectively. The parametric function to be estimated is $g(\mu_X, \mu_Y) = \mu_X/\mu_Y$. It is straightforward to calculate

$$\frac{\partial}{\partial \mu_X} g(\mu_X, \mu_Y) = \frac{1}{\mu_Y}$$

and

$$\frac{\partial}{\partial \mu_Y} g(\mu_X, \mu_Y) = \frac{-\mu_X}{\mu_Y^2}$$

The first-order Taylor approximation (8) and (9) give

$$E\left(\frac{X}{Y}\right) \approx \frac{\mu_X}{\mu_Y}$$

and

$$Var\left(\frac{X}{Y}\right) \approx \frac{1}{\mu_Y^2} VarX + \frac{\mu_Y^2}{\mu_Y^4} Var\, Y - 2\frac{\mu_X}{\mu_Y^3} Cov(X, Y)$$

$$= \left(\frac{\mu_X}{\mu_Y}\right)^2 \left(\frac{Var\, X}{\mu_X^2} + \frac{Var\, Y}{\mu_Y^2} - 2\frac{Cov(X, Y)}{\mu_X \mu_Y}\right)$$

Thus, we have an approximation for the mean and variance of the ratio estimator, and then approximations use only the means, variances, and covariance of $\bar{X}$ and Y. Exact calculations would be quite hopeless, with closed –form expressions being unattainable.

Present a CLT to cover an estimator such as the ratio estimator.

Theorem (Multivariate Delta Method): Let $\boldsymbol{X_1}, \ldots, \boldsymbol{X_n}$ be a random sample with $E(X_{ij}) = \mu_i$ and $Cov(X_{ik}, X_{jk}) = \sigma_{ij}$. For a given function g with continuous first partial derivatives and a specific value of $\mu = (\mu_1, \ldots\ldots, \mu_p)$ for which $\sigma^2 = \sum\sum \sigma_{ij} \frac{\partial g(\mu)}{\partial \mu_i} . \frac{\partial g(\mu)}{\partial \mu_i} > 0$,

$$\sqrt{n}\left[g(\bar{X}_1, \ldots, \bar{X}_s) - g(\mu_1, \ldots, \mu_p)\right] \to N(0, \sigma^2)$$

in distribution.

Suppose the vector–valued random variable $\boldsymbol{X} = (X_1, \dots, X_p)$ has mean $\mu = (\mu_1, \dots, \mu_p)$ and covariances $Cov(X_i, X_j) = \sigma_{ij}$ and we observe an independent random sample $\boldsymbol{X_1}, \dots, \boldsymbol{X_n}$ and calculate the means $\bar{X}_i = \sum_{i=1}^{n} X_{ik}, i = 1, \dots, p$. For a function $g(x) = g(x_1, \dots, x_p)$ we can write

$$g(\bar{x}_1, \dots, \bar{x}_p) = g(\mu_1, \dots, \mu_p) + \sum_{k=1}^{p} g'_k(x)(\bar{x} - \mu_k),$$

1.6 Sufficiency

Any statistic, T(**X**), defines a form of data reduction or data summary. An experiment who uses only the observed value of the statistic, $T(x)$, rather than the entire observed sample x, will treat as equal two sample, $\boldsymbol{x}$, and $\boldsymbol{y}$, that satisfy $T(\boldsymbol{x}) = T(\boldsymbol{y})$ even though the actual sample values may be different in some ways.

Data reduction in terms of a particular statistic can be thought of as a partition of the sample space $\boldsymbol{\chi}$. Let $T = \{t: t = T(x) for\ some\ x \in \chi\}$ be the image of $\boldsymbol{\chi}$ under $T(x)$. Then $T(x)$ partitions the sample space into sets $A_t, t \in T$, defined $A_t = \{x: T(x) = t\}$. The statistic summarizes the data in that, rather than reporting the entire sample x, it reports only that $T(x) = t$ or, equivalently, $x \in A_t$. For example, if $T(x) = x_1 + \cdots + x_n$, then $T(x)$ does not report the actual sample values but only the sum. There may be many different sample points that have the same sum.

The Sufficiency

A *sufficient statistic* for a parameter θ is a statistic that, in a certain sense, captures all the information about θ contained in the sample. Any additional information in the sample, besides the value of the sufficient statistic, does not contain any more information about θ.

SUFFICIENCY *PRINCIPLE*: If $T(\boldsymbol{X})$ is a sufficient statistic for θ, then any inference about θ should depend on the sample $\boldsymbol{X}$ only through the value $T(\boldsymbol{X})$. That is, if $\boldsymbol{x}$ and $\boldsymbol{y}$ are

two sample points such that $T(\boldsymbol{x}) = T(\boldsymbol{y})$, then the inference about θ should be the same whether $\boldsymbol{X} = \boldsymbol{x}$ or $\boldsymbol{Y} = \boldsymbol{y}$ is observed.

Let us assume that X be a random variable with pdf $f(.,\theta)$ of the known functional form but depending upon an unknown constant θ which is called a parameter. We let $\boldsymbol{\Omega}$ stand for the set of all possible values of θ and call it the parameter space. By $\Im$ we denote the family of all pdf's we get by letting θ vary over $\boldsymbol{\Omega}$; that is $\Im=\{f(.,\theta):\theta \in \Omega\}$.

Definition: A statistic $T(\boldsymbol{X})$ is a *sufficient statistic* for θ if the conditional distribution of the sample **X** given the value of the $T(\boldsymbol{X})$ does not depend on θ.

Theorem: *If $p(\boldsymbol{x}|\boldsymbol{\theta})$ is the joint pdf or pmf of $\boldsymbol{X}$ and $q(t|\theta)$ is the pdf or pmf of $T(\boldsymbol{X})$, then $T(\boldsymbol{X})$ is a sufficient statistic for θ if, for every $\boldsymbol{x}$ in the sample space, the ratio $p(\boldsymbol{x}|\theta)/q(T(\boldsymbol{x})|\theta)$ is constant as a function of θ.*

Proof: Since $\{\boldsymbol{X} = \boldsymbol{x}\}$ is a subset of $\{\boldsymbol{T}(\boldsymbol{X}) = \boldsymbol{T}(\boldsymbol{x})\}$,

$$P_\theta\big(\boldsymbol{X} = \boldsymbol{x}\big|T(\boldsymbol{X}) = T(x) = T(x)\big) = \frac{P_\theta(\boldsymbol{X} = \boldsymbol{x}\ and\ T(\boldsymbol{X}) = T(\boldsymbol{x}))}{P_\theta(T(\boldsymbol{X}) = T(\boldsymbol{x}))}$$

$$= \frac{P_\theta(\boldsymbol{X} = \boldsymbol{x})}{P_\theta(T(\boldsymbol{X}) = T(\boldsymbol{x}))}$$

$$= \frac{p(\boldsymbol{x}|\theta)}{q(T(\boldsymbol{x})|\theta)},$$

where $p(x|\theta)$ is the joint pmf of the sample $\boldsymbol{X}$ and $q(t|\theta)$ is the pmf of $T(\boldsymbol{X})$. Thus, $T(\boldsymbol{X})$ is a sufficient statistic for θ if and only if, for every x, the above ratio of pmfs is constant as a function of θ.

Example (Binomial Sufficient Statistic): Let $\boldsymbol{X_1}, \dots, \boldsymbol{X_n}$ be iid Bernoulli random variables with parameter θ, $0 < \theta < 1$. We will show that $T(\boldsymbol{X}) = X_1 + \cdots + X_n$ is a sufficient statistic for θ. Note that $T(\boldsymbol{X})$ counts the number of $\boldsymbol{X_i's}$ that equal 1, so $T(\boldsymbol{X})$ has a binomial(n, θ) distribution. The ratio of pmfs is thus

$$\frac{p(x|\theta)}{q(T(x)|\theta)} = \frac{\prod \theta^{x_i}(1-\theta)^{1-x_i}}{\binom{n}{t}\theta^{t}(1-\theta)^{n-t}}$$

$$= \frac{\theta^{\sum x_i}(1-\theta)^{\sum(1-x_i)}}{\binom{n}{t}\theta^{t}(1-\theta)^{n-t}} \qquad \left(\text{define } t = \sum x_i\right)\left(\prod \theta^{x_i} = \theta^{\sum x_i}\right)$$

$$= \frac{\theta^{t}(1-\theta)^{n-t}}{\binom{n}{t}\theta^{t}(1-\theta)^{n-t}}$$

$$= \frac{1}{\binom{n}{t}} = \frac{1}{\binom{n}{\sum x_i}}$$

Since this ratio does not depend on θ, by previous theorem, $T(\boldsymbol{X})$ is a sufficient statistic for θ. The interpretation is this: The total number of $1s$ in this Bernoulli sample contains all the information about θ that is in the data. Other features of the data, such as the exact value of X_3, contain no additional information.

Example (Normal Sufficient Statistic): Let $X_1, \ldots, X_n$ be iid $n(\mu, \sigma^2)$, where σ^2 is known. We wish to show that the sample mean, $T(\boldsymbol{X}) = \bar{X} = (X_1 + \cdots + X_n)/n$, is a sufficient statistic for μ. The joint pdf of the sample $\boldsymbol{X}$ is

$$f(x|\mu) = \prod_{i=1}^{n}(2\pi\sigma^2)^{-\frac{1}{2}}\exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)$$

$$= (2\pi\sigma^2)^{-n/2}exp\left(-\sum_{i=1}^{n}(x_i-\mu)^2\Big/(2\sigma^2)\right)$$

$$= (2\pi\sigma^2)^{-n/2}exp(-\textstyle\sum_{i=1}^{n}(x_i-\bar{x}+\bar{x}-\mu)^2/(2\sigma^2))$$

(add and subtract $\bar{x}$)

$$= (2\pi\sigma^2)^{-n/2}exp(-\textstyle\sum_{i=1}^{n}(x_i-\bar{x})^2 + n(\bar{x}-\mu)^2/(2\sigma^2)) \qquad (1)$$

The last equality is true because the cross-product term $\sum_{i=1}^{n}(x_i - \bar{x})(\bar{x} - \mu)$ may be rewritten as $(\bar{x} - \mu)\sum_{i=1}^{n}(x_i - \bar{x})$, and $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$. Recall that the sample mean $\bar{X}$ has a $n(\mu, \sigma^2/n)$ distribution. Thus, the ratio of pdfs is

$$\frac{f(x|\theta)}{q(T(x)|\theta)} = \frac{(2\pi\sigma^2)^{-n/2}\exp\left(-(\sum_{i=1}^{n}(x_i - \bar{x})^2 + n(\bar{x} - \mu)^2)/(2\sigma^2)\right)}{(2\pi\sigma^2/n)^{-1/2}\exp\left(-n(\bar{x} - \mu)^2/(2\sigma^2)\right)}$$

$$= n^{-1/2}(2\pi\sigma^2)^{-n-1}\exp\left(-\sum_{i=1}^{n}(x_i - \bar{x})/(2\sigma^2)\right)$$

which does not depend on μ. Therefore, the sample mean is a sufficient statistic for μ.

Example: Let (X_1, X_2) be a random sample from $b(1, \theta)$. Let $T = X_1 + KX_2$, $K = 1, 2, \ldots$. Show that T is sufficient statistic for all K.

Consider $[T = 0] = \{(x_1, x_2): x_1 + K\,x_2 = 0\} = \{(0, 0)\}$

$$[T = 1] = \{(x_1, x_2): x_1 + K\,x_2 = 1\} = \{(1, 0)\}$$

$$[T = K] = \{(x_1, x_2): x_1 + K\,x_2 = K\} = \{(0, 1)\}$$

$$[T = K + 1] = \{(x_1, x_2): x_1 + K\,x_2 = K + 1\} = \{(1, 1)\}$$

Therefore, $P[X_1 = x_1, X_2 = x_2\,|T = t]$=1 or 0, for all (x_1, x_2) and t. Hence T is sufficient by definition.

Example: Let (X_1, X_2) be a random sample from Poisson distribution $P(\theta)$. Let $T = X_1 + KX_2$, $K = 1, 2, \ldots$. Show that T is not sufficient statistic for all K.

Consider $[T = 0] = \{(x_1, x_2): x_1 + K\,x_2 = 0\} = \{(0, 0)\}$

$$[T = 1] = \{(x_1, x_2): x_1 + K\,x_2 = 1\} = \{(1, 0)\}$$

$$[T = K] = \{(x_1, x_2): x_1 + K\,x_2 = K\} = \{(K, 0), (0, 1)\}$$

$$[T = K + 1] = \{(x_1, x_2): x_1 + K\,x_2 = K + 1\} = \{(1, 1),\ (K + 1, 0)\}$$

Then

$$P[X_1 = 0, X_2 = 1 \,/\, T = K] = \frac{P_\theta[X_1 = 0, X_2 = 1]}{P_\theta[T = K]}$$

$$= \frac{e^{-\theta}\,\theta e^{-\theta}}{e^{-\theta}\,\theta e^{-\theta} + \frac{\theta^k e^{-\theta}}{K!}}$$

which depends on θ. Hence T is not sufficient.

Example: Let $(X_1, X_2, \dots, X_n)$ be a random sample from Uniform distribution $U(0, \theta)$. Show that $\max_{1 \le i \le n} X_i$ is sufficient for the family of distributions $\Im = \{U(0, \theta);\ \theta > 0\}$.

Consider the joint pdf of $(X_1, X_2, \dots, X_n)$ is given by

$$f(x_1, x_2, \dots, x_n; \theta) = \frac{1}{\theta^n}, \quad if\ 0 < \max_{1 \le i \le n} X_i \le \theta$$

$$= 0 \qquad otherwise$$

Using order statistics theory, the pdf of $T = \max_{1 \le i \le n} X_i$ is given by

$$g(t; \theta) = \frac{n\, t^{n-1}}{\theta^n}, \qquad 0 \le t \le \theta$$

$$= 0, \quad otherwise$$

Thus, the conditional pdf of $(X_1, X_2, \dots, X_n)$ given T i.e. $f(x_1, x_2, \dots, x_n | t)\ is$

$$f(x_1, x_2, \dots, x_n | t) = \frac{f(x_1, x_2, \dots, x_n; \theta)}{g(t; \theta)}$$

$$= \frac{1}{n\, t^{n-1}}, \qquad 0 \le \max_{1 \le i \le n} x_i \le t$$

$$= 0\,, \qquad \text{otherwise}$$

which is independent of θ for given $T = t$. Hence, by definition of sufficiency, T is sufficient for the family $\Im = \{U(0, \theta);\ \theta > 0\}$.

Example: Let $(X_1, X_2, \dots, X_n)$ be a random sample from discrete Uniform distributionP_N with pmf

$$P_N[X = x] = \frac{1}{N}, x = 1,2, \dots, N$$

Show that $T = \max_{1\le i\le n} X_i$ is sufficient for the family $\Im = \{P_N;\ N \ge 1, \text{integers}\}$.

First, we will find the pmf of T given by

$$P_N[T = t] = P_N\left[\max_{1\le i\le n} X_i = t\right]$$

$$= P_N\left[\max_{1\le i\le n} X_i \le t\right] - P_N\left[\max_{1\le i\le n} X_i \le t-1\right]$$

$$= \{P_N[X \le t]\}^n - \{P_N[X \le t-1]\}^n$$

$$= \left(\frac{t}{N}\right)^n - \left(\frac{t-1}{N}\right)^n$$

$$= \frac{t^n-(t-1)^n}{N^n}, \text{ if } 1 \le t \le N$$

The joint pmf of $(X_1, X_2, \dots, X_n)$ is given by

$P_N[X_1 = x_1, X_2 = x_2, \dots, X_n = x_n] = \frac{1}{N^n}$, if $1 \le$ all $x_i \le N$, integers.

Hence

$P_N[X_1 = x_1, X_2 = x_2, \dots, X_n = x_n \mid T = t]$

$$= \frac{P_N[X_1 = x_1, X_2 = x_2, \dots, X_n = x_n]}{P_N[T = t]}$$

$$= \frac{1}{t^n - (t-1)^n}, \text{if } 1 \le \text{all } x_i \le t$$

which is independent of N. Hence T is sufficient of the family $\Im = \{P_N;\ N \ge 1, \text{integers}\}$.

Example: Let (X_1, X_2) be a random sample with joint pmf

X_1 \ X_2	0	1
0	$\frac{12-7\theta+\theta^2}{12}$	$\frac{4\theta-\theta^2}{12}$
1	$\frac{4\theta-\theta^2}{12}$	$\frac{\theta^2-\theta}{12}$

Then (i) Is (X_1, X_2) a random sample? (ii) Is $X_1 + X_2$ sufficient? (iii) Is X_1X_2 sufficient?

(i) For random sample, we should satisfy

$$P[X_1 = x_1, X_2 = x_2] = P[X_1 = x_1]P[X_2 = x_2].$$

Also, since

$$P[X_1 = 0, X_2 = 0] = \frac{12-7\theta+\theta^2}{12}$$

$$\neq P[X_1 = 0]\, P[X_2 = 0]$$

$$= \frac{(4-\theta)^2}{16}$$

Hence, (X_1, X_2) is not a random sample.

(ii) Let us consider $T = X_1 + X_2$ such that

$$P[X_1 = 0, X_2 = 1 \mid T = 1] = P[X_1 = 1, X_2 = 0 \mid T = 1] = \frac{1}{2}$$

$$P[X_1 = 1, X_2 = 1 \mid T = 2] = 1 = P[X_1 = 0, X_2 = 0 \mid T = 0] = 1$$

which are independent of θ for all (x_1, x_2) and all $T = t$. Hence T is sufficient.

(iii) Let us now consider $T = X_1X_2$ such that

$$P_\theta\,[T = 0] = P_\theta[X_1 = 0, X_2 = 1] + P_\theta[X_1 = 1, X_2 = 0]$$

$$+P_\theta[X_1 = 0, X_2 = 0]$$

$$= \frac{(3+\theta)(4-\theta)}{12}$$

Thus

$$P_\theta[X_1 = 0, X_2 = 0 \mid T = 0] = \frac{P_\theta[X_1 = 0, X_2 = 0]}{P_\theta[T = 0]}$$

$$= \frac{(3-\theta)(4-\theta)}{(3+\theta)(4-\theta)}$$

which depends on θ, hence T is not sufficient.

Remark: In the above examples it so happens that the dimensionality of the sufficient statistic is the same as the dimensionality of the parameter or to put it differently, the number of the real-valued statistics which are jointly sufficient for the parameter Θ coincides with the number of independent coordinates of Θ. However, this need not always be the case.

Example: Let $(X_1, X_2, \dots, X_n)$ be a random sample from Uniform distribution $U(-\theta, \theta), \theta > 0$. Show that $\left(\min_{1\le i\le n} X_i, \max_{1\le i\le n} X_i\right)$ is jointly sufficient for the family $\Im = \{U(-\theta, \theta);\ \theta > 0\}$. Also, show that $T = \max_{1\le i\le n}|X_i|$ is sufficient for the family $\Im = \{U(-\theta, \theta);\ \theta > 0\}$.

Consider, the joint pdf of $(X_1, X_2, \dots, X_n)$ is given by

$$f(x_{1,}x_{2,}, \dots, x_{n,}; \theta) = \frac{1}{(2\theta)^n}, -\theta < x_i < \theta,\ i = 1,2.., n;\ \theta > 0$$

The joint pdf of T= (T_1, T_2), $T_1 = \min_{1\le i\le n} X_i, T_2 = \max_{1\le i\le n} X_i$, is given by

$$g(t_1, t_2; \theta) = \frac{n(n-1)[t_2 - t_1]^{n-2}}{(2\theta)^n}, -\theta < t_1 < t_2 < \theta, \theta > 0$$

Hence,

$$f(x_{1,}x_{2,}, \dots, x_n \mid t_1, t_2) = \frac{f(x_{1,}x_{2,}, \dots, x_{n,}; \theta)}{g(t_1, t_2; \theta)}$$

$$= \frac{1}{n(n-1)[t_2 - t_1]^{n-2}}, \qquad t_1 < all\ x_i < t_2$$

so for a given T=$t = (t_1, t_2)$, the conditional pdf of $(X_1, X_2, \dots, X_n)$ is independent of θ.

Therefore, T is sufficient (jointly) for θ. Note that the dimension of T is two and the dimension of θ is one. Thus, the dimensionality of sufficient statistics is more than that of the parameter.

Next, the pdf of $T = \max_{1\le i\le n}|X_i|$ is given by

$$g(t; \theta) = \frac{n\, t^{n-1}}{\theta^n}, \qquad 0 \le t \le \theta$$

$$= 0, \quad otherwise$$

by assuming $Y_i = |X_i|, i = 1,2, \dots, n$.

Hence,

$$\frac{f(|x_1|, |x_2|, \dots., |x_n|; \theta)}{g(t;\theta)} = \frac{1}{nt^{n-1}}, 0 < all|x_i| \leq \theta$$

$$= 0, \qquad otherwise$$

showing that the conditional pdf of $(|X_1|, |X_2|, \dots., |X_n|)$ given $T = t$ is independent of θ. Therefore, by definition T is sufficient for θ.

Remark: Suppose m is the dimension of the sufficient statistics T= $(T_1, T_2, .., T_m)$ and r is the dimension of parameter $\Theta = (\theta_1, \theta_2, \dots., \theta_r)$. Further, if $m = r$ and that the conditional distribution of $(X_1, X_2, \dots, X_n)$ given $T_j = t_j$ is independent of θ_j. In a situation like this, one may be tempted to declare that T_j is sufficient for θ_j. This outlook, however, is not in conformity with the definition of a sufficient statistic. The notion of sufficiency is connected with the family of pdf's $\Im = \{f(., \Theta): \Theta = (\theta_1, \theta_2, \dots., \theta_r) \subseteq \Omega \subseteq R^r\}$, and we may talk about T_j being sufficient for θ_j , if all other $\theta_i, i \neq j$ are known; otherwise T_j is to be either sufficient for the family or not at all. To understand this let us consider an following illustrative example.

Example: Let $(X_1, X_2, \dots, X_n)$ be a random sample from $N(\theta_1, \theta_2)$. Then show that $(\overline{X}, S^2)$, $S^2 = \sum_{i=1}^{n}(X_i - \overline{X})^2$ is jointly sufficient for $\Theta = (\theta_1, \theta_2)$

$-\infty < \theta_1 < \infty, 0 < \theta_2 < \infty,$ thus $\Theta \subset \{(-\infty, \infty) \text{X} (0, \infty)\}$. Also, show that the conditional pdf of $(X_1, X_2, \dots, X_n)$ given $\overline{X} = \overline{x}$ is independent of θ_1.

We know that $\overline{X}$ is distributed as $N\left(\theta_1, \frac{\theta_2}{n}\right)$ and S^2 is distributed as $G\left(\frac{n-1}{2}, 2\theta_2\right)$. Also, we know that, $\overline{X}$ *and* S^2 are independent. Then, the joint pdf of $(\overline{X}, S^2)$ is given by

$$f_{(\overline{X}, S^2)}(\overline{x}, s^2; \theta_1, \theta_2) = \frac{\sqrt{n}}{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{n-1}{2}\right)(2\theta_2)^{\frac{n}{2}}} (s^2)^{\frac{n-3}{2}} \exp\left\{-\left[\frac{n(\overline{x}-\theta_1)^2 + (s^2)}{2\theta_2}\right]\right\}.$$

Then, the conditional pdf of $(X_1, X_2, \dots, X_n)$ given $(\overline{X}, S^2) = (\overline{x}, s^2)$ is given by

$$\frac{f(x_1,x_2,\ldots,x_n,;\theta_1,\theta_2)}{f_{(\overline{X},S^2)}(\overline{x},s^2;\theta_1,\theta_2)} = \frac{1}{2^{n-1}\pi^{\frac{n-1}{2}}\Gamma\left(\frac{n-1}{2}\right)}(s^2)^{\frac{n-3}{2}}$$

which is independent of (θ_1, θ_2) . Hence, by definition of sufficient statistic, $(\overline{X}, S^2)$ is jointly sufficient for the family $\Im=\{N(\theta_1, \theta_2); -\infty < \theta_1 < \infty, 0 < \theta_2 < \infty\}$. The conditional pdf of $(X_1, X_2, \ldots, X_n)$ given $\overline{X} = \overline{x}$ is as follows:

$$\frac{f(x_1,x_2,\ldots,x_n,;\theta_1,\theta_2)}{f_{\overline{X}}(\overline{x},;\theta_1,\theta_2)} = \frac{1}{\sqrt{n}(2\pi\theta_2)^{\frac{n-1}{2}}} exp - \frac{\sum_{i=1}^{n}(x_i-\overline{x})^2}{2\theta_2}$$

which depends on θ_2 but independent of θ_1. Thus, by definition $\overline{X}$ is not sufficient for the family $\Im=\{N(\theta_1, \theta_2); -\infty < \theta_1 < \infty, 0 < \theta_2 < \infty\}$. The concept of $\overline{X}$ being sufficient for θ_1 is not valid unless θ_2 is known.

Example (Sufficient Order Statistics): Let $\boldsymbol{X_1, \ldots, X_n}$ be iid from a pdf f, where we are unable to specify any more information about the pdf (as is the case in nonparametric estimation).It then follows that the sample density is given by

$$f(x) = \prod_{i=1}^{n} \boldsymbol{f(x_i)} = \prod_{i=1}^{n} \boldsymbol{f\left(x_{(i)}\right)}, \qquad (2)$$

Where $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$ are the order statistics. By theorem 2, we can show that the order statistics are a sufficient statistic. Of course, this is not much of a reduction, but we shouldn't expect more with so little information about the density f.

However, even if we do specify more about the density, we still may not be able to get much of a sufficiency reduction. For example, suppose that f is the Cauchy pdf $f(x|\theta) = \frac{1}{\pi(x-\theta)^2}$ or the logistic pdf$f(x|\theta) = \frac{e^{-(x-\theta)}}{\left(1+e^{-(x-\theta)}\right)^2}$. We then have the same reduction as in (), and no more. So reduction to the order statistic is the most we can get in these families.

It turns out that outside of the exponential family of distribution, it is rare to have a sufficient statistic of smaller dimension than the size of the sample, so in many cases it will turn out that the order statistics are the best that we can do.

To use the definition, we must guess a statistic $T(\boldsymbol{X})$ to be sufficient, find the pmf or pdf of $T(\boldsymbol{X})$, and check that the ratio of pdfs or pmfs does not depend on θ. The first step requires a good deal of intuition and the second sometimes requires some tedious analysis. Fortunately, the next theorem, due to Halmos and Savage (1949), allows us to find a sufficient statistic by simple inspection of the pdf or pmf of the sample.

Theorem (Factorization Theorem): *Let $f(x|\theta)$ denote the joint pdf or pmf of a sample X. A statistic $T(X)$ is a sufficient statistic for θ if and only if there exist functions $g(t|\theta)$ and $h(x)$ such that, for all sample points x and all parameter points θ,*

$$f(x|\theta) = g(T(x)|\theta)h(x) \qquad (3)$$

Proof: We give the proof only for discrete distributions.

Suppose $T(\boldsymbol{X})$ is a sufficient statistic. Choose $g(t|\theta) = P_\theta(T(\boldsymbol{X}) = t)$ and $h(\boldsymbol{x}) = P\big(\boldsymbol{X} = \boldsymbol{x}\big|T(\boldsymbol{X}) = T(\boldsymbol{x})\big)$. Because $T(\boldsymbol{X})$ is a sufficient, the conditional probability defining $h(\boldsymbol{x})$ does not depend on θ. Thus this choice of $h(\boldsymbol{x})$ and $g(t|\theta)$ is legitimate, and for this choice we have

$$f(x|\theta) = P_\theta(\boldsymbol{X} = \boldsymbol{x})$$

$$= P_\theta(\boldsymbol{X} = \boldsymbol{x}\ and\ T(\boldsymbol{X}) = T(x))$$

$$= P_\theta\big(T(\boldsymbol{X}) = T(\boldsymbol{x})\big)P(\ \boldsymbol{X} = \boldsymbol{x}|T(\boldsymbol{X}) = T(\boldsymbol{x}))\ \text{(sufficiency)}$$

$$= g(T(\boldsymbol{x})|\theta)h(\boldsymbol{x})$$

So factorization (3) has been exhibited. We also see from the last two lines above that

$$P_\theta\big(T(\boldsymbol{X})\big) = T(x)) = g(T(x)|\theta),$$

So $g(T(x)|\theta)$ is the pmf of $T(\boldsymbol{X})$.

Now assume the factorization (3) exists. Let $q(t|\theta)$ be the pmf of $T(X)$. To show that $T(X)$ is sufficient we examine the ratio $\frac{f(x|\theta)}{q\big(T(x)\big|\theta\big)}$. Define $A_{T(x)} = \{y: T(y) = T(x)\}$. Then

$$\frac{f(x|\theta)}{q(T(x)|\theta)} = \frac{g(T(x)|\theta)h(x)}{q(T(x)|\theta)} \qquad \text{(since (3) is satisfied)}$$

$$= \frac{g(T(x)|\theta)h(x)}{\sum_{A_{T(x)}} g(T(y)|\theta)h(y)} \quad \text{(definition of the pmf of } T\text{)}$$

$$= \frac{g(T(x)|\theta)h(x)}{g(T(x)|\theta)\sum_{A_{T(x)}} h(y)} \quad \text{(since } T \text{ is constant on } A_{T(x)}\text{)}$$

$$= \frac{h(x)}{\sum_{A_{T(x)}} h(y)}$$

Since the ratio does not depend on θ, by Theorem 2. $T(\boldsymbol{X})$ is a sufficient statistic for θ.

Example (Continuation): For the normal model described earlier, we saw the pdf could be factored as

$$f(x|\mu) = (2\pi\sigma^2)^{-\frac{n}{2}} exp(-\sum_{i=1}^{n}(x_i - \bar{x})^2/(2\sigma^2))\, exp(-n(\bar{x} - \mu)^2/(2\sigma^2)) \qquad (4)$$

We can define

$$h(x) = (2\pi\sigma^2)^{-n/2} exp(-\sum_{i=1}^{n}(x_i - \bar{x})^2/(2\sigma^2))$$

which does not depend on the unknown parameter μ. The factor in () that contains μ depends on the sample x only through the function $T(\boldsymbol{x}) = \bar{x}$, the sample mean. So we have

$$g(t|\mu) = exp(-n(t - \mu)^2/2\sigma^2))$$

and note that

$$f(x|\mu) = h(x)g(T(x)|\mu).$$

Thus, by the factorization Theorem, $T(X) = \bar{X}$ is a sufficient statistic for μ.

Example (Uniform Sufficient Statistic): Let $X_1, \ldots, X_n$ be iid observation from the discrete uniform distribution on $1, \ldots, \theta$.That is, the unknown parameter, θ, is a positive integer and the pmf of $X_i$is

$$f(x|\theta) = \begin{cases} \frac{1}{\theta} & x = 1,2 \ldots\ldots, \theta \\ 0 & otherwise \end{cases}$$

Thus, the joint pmf of $X_1, \ldots\ldots, X_n$ is

$$f(x|\theta) = \begin{cases} \theta^{-n} & x_i \in \{1,2, \ldots\ldots\} for\ i = 1, \ldots\ldots, n \\ 0 & otherwise \end{cases}$$

The restriction "$x_i \in \{1, \ldots, \theta\}$ for $i = 1, \ldots, n$" can be re-expressed as "$x_i \in \{1,2, \ldots\ldots\}$ for $i = 1, \ldots, n$(note that there is no θ inthis restriction) and $\max_i x_i \leq \theta$".If we define $T(X) = \max_i x_i$,

$$h(x) = \begin{cases} 1 & x_i \in \{1,2, \ldots\ldots\} for\ i = 1, \ldots, n \\ 0 & otherwise\ . \end{cases}$$

and

$$g(t|\theta) = \begin{cases} \theta^{-n} & t \leq \theta \\ 0 & otherwise \end{cases}$$

it is easily verified that $f(x|\theta) = g(T(x)|\theta)h(x)$ for all x and θ. Thus, the largest order statistic, $T(\boldsymbol{X}) = \max_i X_i$, is a sufficient statistic in this problem.

This type of analysis can sometimes be carried out more clearly and concisely using indicator functions. Recall that $I_A(x)$ is the indicator function of the set A; that is, it is equal to 1 if $x \in A$ and equal to 0 otherwise. Let $N = \{1,,2, \ldots\ldots\}$ be the set of positive integers and let $N_\theta = \{1,2,, \ldots, \theta\}$.Then the jnt pmf of $X_1, \ldots, X_n$ is

$$f(x|\theta) = \prod_{i=1}^{n} \theta^{-1} I_{N_\theta}(x_i) = \theta^{-n} \sum_{i=1}^{n} I_{N_\theta}(x_i).$$

Defining $T(x) = \max_i x_i$,we see that

$$\prod_{i=1}^{n} I_{N_\theta}(x_i) = \left(\prod_{i=1}^{n} I_N(x_i) \right) I_{N_\theta}\big(T(x)\big).$$

Thus, we have the factorization

$$f(x|\theta) = \theta^{-n} I_{N_\theta}\big(T(x)\big)\left(\prod_{i=1}^{n} I_N(x_i)\right)$$

The first factor depends on $x_1, \ldots, x_n$ only through the value of $T(x) = \max_i x_i$.

It is easy to find a sufficient statistic for an exponential family of distributions using the Factorization Theorem. The proof of the following important result is straight forward.

Theorem: Let $X_1, \ldots, X_n$ be iid observations from a pdf or pmf $f(x|\theta)$that belongs to an exponential family given by

$f(x|\theta)$=$h(x)c(\theta)\exp\big(\sum_{i=1}^{k} w_i(\theta)t_i(x)\big)$,

where $\theta = (\theta_1, \theta_2, \ldots., \theta_d), d \leq k$. Then

$$T(X) = \left(\sum_{j=1}^{n} t_1\,(X_j), \ldots, \sum_{j=1}^{n} t_k(X_j)\right)$$

is a sufficient statistic for θ.

Proof: The proof of theorem is straight forward using Neyman Fisher Factorization Theorem (left as exercise).

Example: Let *X* be a random variable with pdf $N(\theta_1, \theta_2)$. Show that $N(\theta_1, \theta_2)$ belongs to 2-parameter exponential family.

Note that here $\Theta = (\theta_1, \theta_2)$ and we may write

$$f(x;\Theta) = \frac{1}{\sqrt{2\pi\theta_2}}\, exp - \frac{1}{2\theta_2}[x - \theta_1]^2$$

$$= \frac{1}{\sqrt{2\pi\theta_2}} exp - \frac{1}{2\theta_2}\left[x^2 + {\theta_1}^2 - 2\theta_1 x\right]$$

$$= \frac{1}{\sqrt{2\pi\theta_2}}\, exp - \frac{{\theta_1}^2}{2\theta_2} exp\left[-\frac{x^2}{2\theta_2} + \frac{\theta_1 x}{\theta_2}\right]$$

If we put $c(\Theta) = c(\theta_1, \theta_2) = \frac{1}{\sqrt{2\pi\theta_2}}\, exp - \frac{{\theta_1}^2}{2\theta_2}$

$$w_1(\theta_1, \theta_2) = \frac{\theta_1}{\theta_2}\,, w_2(\theta_1, \theta_2) = -\frac{1}{2\theta_2}$$

$$t_1(x) = x, t_2(x) = x^2, and\ h(x) \equiv 1.$$

Then we can write

$f(x; \Theta) = c(\theta_1, \theta_2)\ exp\left[\sum_{j=1}^{2} w_j(\theta_1, \theta_2)\, t_j(x)\right] h(x)$.

Since $h(x) > 0$ and the set of positivity of $f(x; \Theta)$ does not depend on $\Theta = (\theta_1, \theta_2)$ and $C(\theta_1, \theta_2) > 0$. for all $\Theta = (\theta_1, \theta_2) \subset (-\infty, \infty)x(0, \infty) \subset R^2$ $f(x; \Theta)$ belongs to 2-parameter exponential family.

Remark: Thus, T being a sufficient for θ implies that for every $A\ \varepsilon\ \boldsymbol{\mathcal{B}}^n$, where $\boldsymbol{\mathcal{B}}^n$ is the n-dimensional Borel field corresponding to $\boldsymbol{R}^n$.

$P_\theta[(X_1, X_2, \ldots., X_n)\varepsilon A \mid T = t]$ is independent of θ for almost all t.

Actually, more is true, namely, if $T^* = (T_1, T_2, \ldots, T_k)$ is any k-dimensional statistic, then the conditional distribution of T^*, given T = t, is independent of θ for almost all t . To see this, let B be any set in $\boldsymbol{\mathcal{B}}^k$ and then

$P_\theta[T^*\varepsilon\ B \mid T = t] = P_\theta[(X_1, X_2, \ldots., X_n)\varepsilon A \mid T = t]$

and this is independent θ for almost all t. We finally remark that $\boldsymbol{X}$=$(X_1, X_2, \ldots., X_n)$ is always a sufficient statistic for θ. A sufficient statistic condenses the data in such a way that whatever information is contained in the sample about, all of the information is also contained in the statistic. Thus, a sufficient statistic condenses the data without losing any information about θ contained in the sample. We choose the sufficient statistic which maximizes the condensation of data.

1.7 Minimal Sufficient Statistics

In any problem there are, in fact, many sufficient statistics. It is always true that the complete sample, $\boldsymbol{X}$, is itself a sufficient statistic. We can factor the pdf or pmf of $\boldsymbol{X}$ as $f(x|\theta) = f(T(\boldsymbol{x})|\theta)h(\boldsymbol{x})$, where $T(\boldsymbol{x}) = \boldsymbol{x}$ and $h(\boldsymbol{x}) = 1$ for all $\boldsymbol{x}$. By the factorization

theorem, $\boldsymbol{T(X) = X}$ is a sufficient statistic. Also, it follows that any one- to- one function of a sufficient statistic is a sufficient statistic. Suppose $T(\boldsymbol{X})$ is a sufficient statistic and define $T^*(\boldsymbol{x}) = r(T(\boldsymbol{x}))$ for all $\boldsymbol{x}$, where r is one-to-one function with inverse r^{-1}. Then by the Factorization Theorem there exist g and h such that

$$f(\boldsymbol{x}|\theta) = g(T(\boldsymbol{x})|\theta)h(\boldsymbol{x})$$

$$= g\big(r^{-1}\big(T^*(\boldsymbol{x})\big)\big|\theta\big)h(\boldsymbol{x}).$$

Defining $g*(t|\theta) = g(r^{-1}(t)|\theta)$, we see that

$$f(x|\theta) = g^*(T^*(\boldsymbol{x})|\theta)h(\boldsymbol{x}).$$

So, by the Factorization Theorem, $T^*(\boldsymbol{X})$ is a sufficient statistic.

Recall that the purpose of a sufficient statistic is to achieve data reduction without loss of information about the parameter θ; thus, a statistic that achieves the most data reduction while still retaining all the information about θ might be considered preferable.

Definition: A sufficient statistic T(**X**) is called a *minimal sufficient statistic* if, for any other sufficient statistic $T'(\boldsymbol{X}), T(\boldsymbol{x})$ is a function of $T'(\boldsymbol{x})$.

To say that $T(\boldsymbol{x})$ is a function of $T'(\boldsymbol{x})$ simply means that if $T'(\boldsymbol{x}) = T'(\boldsymbol{y})$,then $T(\boldsymbol{x}) = T(\boldsymbol{y})$. In terms of the partition sets, if $\{B_{t'}: t' \in T'\}$ are the partition sets for $T'(\boldsymbol{x})$ and $\{A_t: t \in T\}$ are the partition sets for $T(\boldsymbol{x})$, then above definition states that every $B_t{}'$ is a subset of some A_t. Thus, the partition associated with a minimal sufficient statistic, is the *coarsest* possible partition for a sufficient statistic, and a minimal sufficient statistic achieves the greatest possible data reduction for a sufficient statistic.

Example (Two Normal Sufficient Statistics): Let $X_1, \ldots, X_n$ are iid $N(\mu, \sigma^2)$ with σ^2 known. Using Factorization Theorem, we conclude that $T(\boldsymbol{X}) = \overline{\boldsymbol{X}}$ is a sufficient statistic for μ. Instead, we could also correctly conclude that $T'(\boldsymbol{X}) = (\bar{X}, S^2)$ is a sufficient statistic for μ in this problem. Clearly $T(\boldsymbol{X})$ achieves a greater data reduction than $T'(\boldsymbol{X})$ since we do not know the sample variance if we know only $T(\boldsymbol{X})$.We can write $T(\boldsymbol{x})$ as a function of $T'(\boldsymbol{x})$ by defining the function $r(a, b) = a$.Then $T(\boldsymbol{x}) = \bar{x} = r(\bar{x}, s^2) = r(T', \boldsymbol{x})$).Since $T(\boldsymbol{X})$ and $T'(\boldsymbol{X})$ are both sufficient statistics, they both contain the same information about

μ.Thus the additional information about the value of S^2, the sample variance does not add to our knowledge of μ since the population variance σ^2 is known.

Of course, if σ^2 is unknown, $T(\boldsymbol{X}) = \bar{X}$ is not a sufficient statistic and $T'(\boldsymbol{X})$ contains more information about the parameter (μ, σ^2) than does $T(\boldsymbol{X})$.

Theorem: Let $f(\boldsymbol{x}|\theta)$ be the pmf or pdf of a sample $\boldsymbol{X}$. Suppose there exists a function $T(\boldsymbol{x})$ such that, for every two sample points $\boldsymbol{x}$ and $\boldsymbol{y}$, the ratio $f(\boldsymbol{x}|\theta)/f(\boldsymbol{y}|\theta)$ is constant as a function of θ if and only if $T(\boldsymbol{x}) = T(\boldsymbol{y})$.Then $T(\boldsymbol{X})$ is a minimal sufficient statistic for θ.

Proof: First, we show that $T(\boldsymbol{X})$ is a sufficient statistic. Let $T = \{t: t = T(\boldsymbol{x})$ for some $\boldsymbol{x} \in \chi\}$ be the image of $\boldsymbol{\chi}$ under $T(\boldsymbol{x})$.Define the partition sets induced by $T(\boldsymbol{x})$ as $A_t = \{\boldsymbol{x}: T(\boldsymbol{x}) = t\}$.For each A_t, choose and fix one element $\boldsymbol{x}_t \in A_t$.For any $\boldsymbol{x} \in \boldsymbol{\chi}, \boldsymbol{x}_{T(\boldsymbol{x})}$ is the fixed element that is in the same set,A_t, as $\boldsymbol{x}$. Since $\boldsymbol{x}$ and $\boldsymbol{x}_T(x)$ are in the same set $A_t, T(\boldsymbol{x}) = T\big(\boldsymbol{x}_T(x)\big)$ and, hence, $f(\boldsymbol{x}|\theta)|f(\boldsymbol{x}_{T(\boldsymbol{x})}|\theta)$ is constant as a function of θ .Thus, we can define a function on $\boldsymbol{\chi}$ by $h(\boldsymbol{x}) = f(\boldsymbol{x}|\theta/f(\boldsymbol{x}_{T(\boldsymbol{x})}|\theta)$ and h does not depend on θ. Define a function on T by $g(t|\theta) = f(\boldsymbol{x}_t|\theta)$. Then it can be seen that

$$f(\boldsymbol{x}|\theta) = \frac{f\big(\boldsymbol{x}_{T(\boldsymbol{x})}\big|\theta\big)f(\boldsymbol{x}|\theta)}{f(\boldsymbol{x}_{T(\boldsymbol{x})}|\theta)} = g(T(\boldsymbol{x})|\theta)h(\boldsymbol{x})$$

And, by the Factorization Theorem, $T(\boldsymbol{X})$ is a sufficient statistic for θ.

Now to show that $T(\boldsymbol{X})$ is minimal, let $T'(\boldsymbol{X})$ be any other sufficient statistic. By the Factorization Theorem, there exist functions g' and h' such that $f(\boldsymbol{x}|\theta) = g'(T'(\boldsymbol{x})|\theta)h'(\boldsymbol{x})$. Let $\boldsymbol{x}$ and $\boldsymbol{y}$ be any two sample points with $T'(\boldsymbol{x}) = T'(\boldsymbol{y})$. Then

$$\frac{f(\boldsymbol{x}|\theta)}{f(\boldsymbol{y}|\theta)} = \frac{g'(T'(\boldsymbol{x})|\theta)h'(\boldsymbol{x})}{g'(T'(\boldsymbol{y})|\theta)h'(\boldsymbol{y})} = \frac{h'(\boldsymbol{x})}{h'(\boldsymbol{y})}$$

Since this ratio does not depend on θ, the assumptions of the theorem imply that $T(\boldsymbol{x}) = T(\boldsymbol{y})$. Thus, $T(\boldsymbol{x})$ is a function of $T'(\boldsymbol{x})$ and $T(\boldsymbol{x})$ is minimal.

Example (Normal Minimal Sufficient Statistic): Let $X_1, \ldots, X_n$ be iid $n(\mu, \sigma^2), both\ \mu$ and σ^2 unknown.Let $\boldsymbol{x}$ and $\boldsymbol{y}$ denote two sample points, and let $(\bar{x}, s_x^2)$ and $\big(\bar{y}, s_y^2\big)$, be the

sample means and variances corresponding to the $\boldsymbol{x}$ and $\boldsymbol{y}$ samples,respectively. Then, we see that the ratio of densities is

$$\frac{f(\boldsymbol{x}|\mu,\sigma^2)}{f(\boldsymbol{y}|\mu,\sigma^2)} = \frac{(2\pi\sigma^2)^{-\frac{n}{2}}\exp(-[n(\bar{x}-\mu)^2+(n-1)s_x^2]/(2\sigma^2))}{(2\pi\sigma^2)^{-\frac{n}{2}}\exp\big(-[n(\bar{y}-\mu)^2+(n-1)s_y^2]/(2\sigma^2)\big)}$$

$$= \exp\left(\left[-n(\bar{x}^2-\bar{y}^2)+2n\mu(\bar{x}-\bar{y})-(n-1)\left(s_x^2-s_y^2\right)\right]/(2\sigma^2)\right)$$

This ratio will be constant as a function of μ and σ^2 if and only if $\bar{x}=\bar{y}$ and $s_x^2 = s_y^2$. Thus, $(\bar{X}, S^2)$ is a mnimal sufficient statistic for (μ,σ^2).

Example (Uniform Minimal Sufficient Statistic): Suppose $X_1,\dots,X_n$ are iid uniform observations on the interval $\theta,\theta+1), -\infty<\theta<\infty$. Then the joint pdf of $\boldsymbol{X}$ is

$$f(x|\theta) = \begin{cases} 1 & \theta < x_i < \theta+1, i=1,\dots,n, \\ 0 & otherwise, \end{cases}$$

which can be written as

$$f(x|\theta) = \begin{cases} 1 & \max_i x_i - 1 < \theta < \min_i x_i \\ 0 & otherwise. \end{cases}$$

Thus, for two sample points $\boldsymbol{x}$ and $\boldsymbol{y}$, the numerator and denominator of the ratio $f(\boldsymbol{x}|\theta)/f(\boldsymbol{y}|\theta)$ will be positive for the same values of θ if and only if $\min_i x_i = \min_i y_i$. and $\max_i x_i = \max_i y_i$. And, if the minima and maxima are equal, then the ratio is constant and, in fact, equals 1. Thus letting $X_{(1)} = \min_i X_i$ and $X_n = \max_i X_i$, we have that $T(\boldsymbol{X}) = \left(X_{(1)}, X_{(n)}\right)$ is a minimal sufficient statistic.This is a case in which the dimension of a minimal sufficient statistic does not match the dimension of the parameter.

A minimal sufficient statistic is not unique. Any one-to-one function of a minimal sufficient statistic is also a minimal sufficient statistic.

$(X_{(n)}-X_{(1)}, (X_{(n)}+X_{(1)}/2)$ is also a minimal sufficient statistic in uniform example. $(\sum_i^n X_i, \sum_i^n X_i^2)$ is also a minimal sufficient statistic in normal example.

1.8 Ancillary Statistic

Sufficient statistics contain all the information about θ that is available in the sample we introduce a different sort of statistic, one that has a complementary purpose.

Definition: A statistic $S(\boldsymbol{X})$ whose distribution does not depend on the parameter θ is called an ancillary statistic.

Alone, an ancillary statistic contains no information about θ. An ancillary statistic is an observation on a random variable whose distribution is fixed and known, unrelated to θ. Paradoxically, an ancillary statistic, when used in conjunction with other statistics, sometimes does contain valuable information for inferences about θ.

Example (Uniform Ancillary Statistic): Let $X_1, \ldots, X_n$ be iid uniform observations on the interval $(\theta, \theta+1), -\infty < \theta < \infty$. Let $X_{(1)} < \ldots < X_{(n)}$ be the order statistics from the sample. We show below that the range statistic, $R = X_{(n)} - X_{(1)}$ is an ancillary statistic by showing that the pdf of R does not depend on θ.Recall that the cdf of each X_i is

$$F(x|\theta) = \begin{cases} 0 & x \le \theta \\ x-\theta & \theta < x < \theta+1 \\ 1 & \theta+1 \le x. \end{cases}$$

Thus, the joint pdf of $X_{(1)}\ and\ X_{(n)}, is$

$$g\left(x_{(1)}, x_{(n)}|\theta\right) = \begin{cases} n(n-1)\left(x_{(n)} - x_{(1)}\right)^{(n-2)} & ;\ \theta < x_{(1)} < x_{(n)} < \theta+1 \\ 0 & ;\ otherwise \end{cases}$$

Making the transformation $R = X_{(n)} - X_{(1)}\ and\ M = \left(X_{(1)} + X_{(n)}\right)/2,$ which has the inverse transformation $X_{(1)} = (2M - R)/2$ and $X_{(n)} = (2M + R)/2$ with jacobian 1, we see that the joint pdf of R and M is

$$h(r, m|\theta) = \begin{cases} n(n-1)r^{n-2} & ;\ 0 < r < 1, \theta + (r/2) < m < \theta + 1 - (r/2) \\ 0 & ;\ \text{otherwise.} \end{cases}$$

(Notice the rather involved region of positivity for $h(r, m|\theta).$) Thus, the pdf for R is

$$h(r|\theta) = \int_{\theta+(r/2)}^{\theta+1-(r/2)} n(n-1)r^{n-2}dm$$

$$= n(n-1)r^{n-2}(1-r), \qquad 0 < r < 1.$$

This is a beta pdf with $\alpha = n-1 \; and \; \beta = 2$.More important, the pdf is the same for all θ.Thus, the distribution of R does not depend on θ, and R is ancillary.

The ancillarity of R does not depend on the uniformity of the $X_i s$, but rather on the parameter of the distribution being a location parameter.

Example (Location Family Ancillary Statistic): Let $X_1, \ldots, X_n$ be iid observations from a location parameter family with cdf $F(x-\theta), -\infty < \theta < \infty$.We will show that the range, $R = X_{(n)} - X_{(1)}$, is an ancillary statistic. We work with $Z_1, \ldots, Z_n$ iid observations from $F(x)$ (corresponding to $\theta = 0$) with $X_1 = Z_1 + \theta, \ldots, X_n = Z_n + \theta$.Thus the cdf of the range statistic, R, is

$$F_R(r|\theta) = P_\theta(R \le r)$$

$$= P_\theta(\max_i X_i - \min_i X_i \le r)$$

$$= P_\theta(\max_i (Z_i + \theta) - \min_i (Z_i + \theta) \le r)$$

$$= P_\theta(\max_i Z_i - \min_i Z_i + \theta - \theta \le r)$$

$$= P_\theta(\max_i Z_i - \min_i Z_i \le r).$$

The last probability does not depend on θ because the distribution of $Z_1, \ldots, Z_n$ does not depend on θ.Thus the cdf of R does not depend on θ and, hence, R is an ancillary statistic.

Example (Scale Family Ancillary Statistic): Scale parameter families also have certain kinds of ancillary statistics. Let $X_1, \ldots, X_n$ be iid observations from a scale parameter family with cdf $F(x/\sigma), \sigma > 0$. Then any statistic that depends on tha sample only through the $n-1$ values $X_1/X_n, \ldots X_{n-1}/X_n$ is an ancillary statistic. For example,

$$\frac{X_1 + \cdots + X_n}{X_n} = \frac{X_1}{X_n} + \cdots + \frac{X_{n-1}}{X_n} + 1$$

Is an ancillary statistic. To see this fact, let $Z_1, \ldots, Z_n$ be iid observations from $F(x)$(corresponding to $\sigma = 1$)with $X_i = \sigma Z_i$.The joint cdf of $X_1/X_n, \ldots X_{n-1}/X_n$ is

$$F(y_1, \ldots, y_{n-1}|\sigma) = P_\sigma(X_1/X_n \le y_1, .., X_{n-1}/X_n \le y_{n-1})$$

$$= P_\sigma(\sigma Z_1/\sigma Z_n) \le y_1, \ldots, \sigma Z_{n-1}/(\sigma Z_n) \le y_{n-1})$$

$$= P_\sigma(Z_1/Z_n \le y_!, .., Z_{n-1}/Z_n \le y_{n-1})$$

The last probability does not depend on σ because the distribution of $Z_1, .., Z_n$ does not depend on σ.So the distribution of $X_1/X_n, \ldots X_{n-1}/X_n$ is independent of σ, as is the distribution of any function of these quantities.

In particular, let $X_1$and X_2 be iid $n(0, \sigma^2)$observations.From the above result,we see that X_1/X_2 has a distribution that is the same for every value of σ.But, in we see that, if $\sigma = 1, X_1/X_2$ has a Cauchy(0,1) distribution.Thus,for any $\sigma > 0$,the distribution of X_1/X_2 is this same Cauchy distribution.

Definition: Let $f(x)$ be any pdf. Then the family of pdfs $f(x-\mu)$, indexed by the parameter $\mu, -\infty < \mu < \infty$, is called the location family with standard pdf $f(x)$ and μ is called the location parameter for the family.

Example (Exponential Location Family): Let $f(x) = e^{-x}, x \ge 0$, and $f(x) = 0, x < 0$. To form a location family we replace x with $x - \mu$ to obtain

$$f(x|\mu) = \begin{cases} e^{-(x-\mu)} & x - \mu \ge 0 \\ 0 & x - \mu < 0 \end{cases}$$

$$= \begin{cases} e^{-(x-\mu)} & x \ge \mu \\ 0 & x < 0. \end{cases}$$

Definition: Let $f(x)$ be any pdf. Then for any $\sigma > 0$,the family of pdfs $(1/\sigma), f(x/\sigma)$, indexed by the parameter σ, is called the scale family with standard pdf $f(x)$ and σ is called *scale parameter* of the family.

Definition: Let $f(x)$ be any pdf. Then for any $\mu, -\infty < \mu < \infty$, and any $\sigma > 0$, the family of pdfs$(1/\sigma)f((x-\mu)/\sigma)$, indexed by the parameter(μ, σ), is called the location-scale

family with standard pdf $f(x)$; μ is called the *location parameter* and σ is called the *scale parameter*.

The normal and double exponential families are examples of location-scale families. The Cauchy is also a location-scale family.

Theorem: Let $f(.)$ be any pdf. Let μ be any real number, and let σ be any positive real number. Then X is a random variable with pdf$(1/\sigma)f((x-\mu)/\sigma)$ if and only if there exists a random variable Z with pdf $f(z)$ and $X = \sigma Z + \mu$.

Proof: To prove the "if" part, define $g(z) = \sigma z + \mu$.Then $X = g(Z), g$ is a monotone function,$g^{-1}(x) = (x-\mu)/\sigma$, and $|(d/dx)g^{-1}(x)| = 1/\sigma$. Thus the pdf of X is

$$f_X(x) = f_Z\big(g^{-1}(x)\big)\left|\frac{d}{dx}g^{-1}(x)\right| = f\left(\frac{x-\mu}{\sigma}\right)\frac{1}{\sigma}.$$

To prove the "only if" part, define $g(x) = (x-\mu)/\sigma$ and let $Z = g(X)\,.\,g^{-1}(z) = \sigma z + \mu, |(d/dz)g^{-1}(z)| = \sigma$, and the pdf of Z is

$$f_Z(z) = f_X\big(g^{-1}(z)\big)\left|\frac{d}{dz}g^{-1}(z)\right| = \frac{1}{\sigma}f\left(\frac{(\sigma z+\mu)-\mu}{\sigma}\right)\sigma = f(z)$$

Also,

$$\sigma Z + \mu = \sigma g(X) + \mu = \sigma\left(\frac{(X-\mu)}{\sigma}\right) + \mu = X.$$

Theorem: Let Z be a random variable with pdf $f(z)$.Suppose EZ and Var Z exist. If X is a random variable with pdf$(1/\sigma)f((x-\mu)/\sigma)$,then

$$EX = \sigma EZ + \mu \quad \text{and} \quad Var\,X = \sigma^2 VarZ.$$

In particular, if $EZ = 0\ and\ Var\ Z = 1$ then $EX = \mu$ and $Var\ X = \sigma^2$.

Proof: By previous theorem, there is a random variable Z^* with pdf $f(z)$ and $X = \sigma Z^* + \mu$. So $EX = \sigma EZ^* + \mu$ and $Var\ X = \sigma^2 Var\ Z^* = \sigma^2 VarZ$.

Probabilities for any member of location-scale family may be computed in terms of the standard variable Z because

$$P(X \leq x) = P\left(\frac{X-\mu}{\sigma} \leq \frac{x-\mu}{\sigma}\right) = P\left(Z \leq \frac{x-\mu}{\sigma}\right).$$

Thus, if $P(Z \leq z)$ is tabulated or easily calculable for the standard variable Z, then probabilities for X may be obtained. Calculations of normal probabilities using the standard normal table are examples of this.

1.9 Complete Statistic

Definition: Let $f(t|\theta)$ be a family of pdfs or pmfs for a statistic $T(\boldsymbol{X})$.The family of probability distribution is called complete if $E_\theta g(T) = 0$ for all θ implies $P_\theta(g(T) = 0) = 1$ for all θ. Equivalently, $T(\boldsymbol{X})$ is called a complete statistic.

Notice that completeness is a property of a family of distributions, not of a particular distribution. For example, if X has a $N(0,1)$ distribution, then defining $g(x) = x$,we have that $Eg(X) = EX = 0$.But the function $g(x) = x$ satisfies $P(g(X) = 0) = P(X = 0) = 0$, not 1. However, this is a particular distribution, not a family of distributions. If X has a $N(\theta, 1)$ distribution, $-\infty < \theta < \infty$, we shall see that no function of X,except one that is 0 with probability 1 for all θ,satisfies $E_\theta g(X) = 0$ for all θ. Thus, the family of $N(\theta, 1)$ distributions, $-\infty < \theta < \infty$, is complete.

Example: Show that the family of binomial distributions $\{b(1,\theta): \theta\ \varepsilon\ (0,1) = \Omega\}$ is complete.

Let $g(X)$ be any random variable such that

$E_\theta[g(X)] = 0\ \ for\ all\ \theta\ \varepsilon\ \Omega$

$\Rightarrow g(0)P_\theta[X = 0] + g(1)P_\theta[X = 1] = 0$ for all $\theta\ \varepsilon\ \Omega$

$\Rightarrow\ g(0)(1-\theta) + g(1)\theta = 0$for all $\theta\ \varepsilon\ \Omega = (0,1)$

$\Rightarrow g(0) + [g(1) - g(0)]\theta = 0$ for all $\theta\ \varepsilon\ (0,1)$.

Let θ_1 and $\theta_2\ \varepsilon\ (0,1), \theta_1 \neq \theta_2$

$\Rightarrow\ g(0) + [g(1) - g(0)]\theta_1 = 0$ and $g(0) + [g(1) - g(0)]\theta_2 = 0$

$\Rightarrow [g(1) - g(0)](\theta_1 - \theta_2) = 0$

$\Rightarrow g(1) = g(0) = 0$

$\Rightarrow g(x) \equiv 0$ for all $x = 0,1$.

Hence, family of distributions $\{b(1,\theta): \theta\ \varepsilon\ (0,1) = \Omega\}$ is complete.

Example (Binomial Complete Sufficient Statistic): Suppose that T has a binomial (n,p) distribution, $0 < p < 1$. Let g be a function such that

$$E_p g(T) = 0.$$

Then

$$0 = E_p g(T) = \sum_{t=0}^{n} g(t) \binom{n}{t} p^t (1-p)^{n-t}$$

$$= (1-p)^n \sum_{t=0}^{n} g(t) \binom{n}{t} \left(\frac{p}{1-p}\right)^t$$

For all $p, 0 < p < 1$.The factor $(1-p)^n$ is not 0 for any p in this range. Thus it must be that

$$0 = \sum_{t=0}^{n} g(t) \binom{n}{t} \left(\frac{p}{1-p}\right)^t = \sum_{t=o}^{n} g(t) \binom{n}{t} r^t$$

For all $r, 0 < r < \infty$.But the last expression is a polynomial of degree n in r, where the coefficient of r^t is $g(t)\binom{n}{t}$.For the polynomial to be 0 for all r, each coefficient must be 0.Since none of the $\binom{n}{t}$ terms is 0, this implies that $g(t) = 0$ for $t = 0,1,\dots,n$.Since T takes on the values $0,1,..n$ with probability 1, this yields that $p_p(g(T) = 0) = 1$ for all p, the desired conclusion. Hence, T is a complete statistic.

Example (Uniform Complete Sufficient Statistic): Let $X_1, \dots, X_n$ be iid uniform$(0, \theta)$ observations,$0 < \theta < \infty$. We know that $T(\boldsymbol{X}) = \max_i X_i$ is a sufficient statistic and, the pdf of $T(\boldsymbol{X})$ is

$$f(t|\theta) = \begin{cases} nt^{n-1}\,\theta^{-n} & ;\ 0 < t < \theta \\ 0 & ;\quad otherwise \end{cases}$$

Suppose $g(t)$ is a function satisfying $E_\theta g(T) = 0$ for all θ. Since $E_\theta g(T)$ is constant as a function of θ, its derivative with respect to θ is 0. Thus we have that

$$0 = \frac{d}{d\theta} E_\theta g(T)$$

$$= \tfrac{d}{d\theta} \textstyle\int_0^\theta g(t) n t^{n-1} \theta^{-n} dt$$

$$= (\theta^{-n}) \tfrac{d}{d\theta} \textstyle\int_0^\theta n g(t) t^{n-1} dt + \left(\tfrac{d}{d\theta}\theta^{-n}\right) \textstyle\int_0^\theta n g(t) t^{n-1} dt$$

$$= \theta^{-n} n g(\theta) \theta^{n-1} + 0 \qquad \text{(applying the product rule for differentiation)}$$

$$= \theta^{-1} n g(\theta).$$

The first term in the next to last line is the result of an application of the Fundamental Theorem of Calculus. The second term is 0 because the integral is, except for a constant, equal to $E_\theta g(T)$, which is 0.Since $\theta^{-1} n g(\theta) = 0$ and $\theta^{-1} n \neq 0$, it must be that $g(\theta) = 0$. This is true for every $\theta > 0$; hence, T is a complete statistic.

Example: Let X be a random variable from discrete Uniform distributionP_N with pmf

$$P_N[X = x] = \frac{1}{N}, x = 1,2, \dots., N$$

Show that the family of pmf's $\Im = \{P_N;\ N \geq 1, \text{integers}\}$ is complete.

Let $g(X)$ be any random variable such that

$$E_N[g(X)] = 0 \text{ for all } N \text{ integers} \geq 1.$$

For $N = 1 \Rightarrow g(1) = 0$.

For $N = 2 \Rightarrow g(1) + g(2) = 0 \Rightarrow g(2) = 0$.

Thus, $g(1) = 0 \Rightarrow g(2) = 0 \Rightarrow g(3) = 0 \Rightarrow \cdots \Rightarrow g(N) = 0$

$$\Rightarrow g(1) = g(2) = \cdots g(N) = 0 \; i.e. \; g(x) = 0$$

For all $= 1,2, \ldots N \geq 1$.

Hence, the family $\mathfrak{I} = \{P_N;\ N \geq 1, \text{integers}\}$ is complete.

Example: Suppose that we exclude the value $N = n_0$ for some $n_0 \geq 1$ from the family in the previous Example. Let us write $\mathfrak{I}^* = \{P_N;\ N \geq 1, \text{integers}, N \neq n_0\}$. Show that the of pmf's $\mathfrak{I}^*$ is not complete.

To show that $\mathfrak{I}^*$ is not complete, it is sufficient to give a function $g(X)$ whose expectation is zero for all $N \geq 1$ but the function is not identically equal to zero i.e. $g(x) \neq 0$ for all $= 1,2, \ldots, N$.

Let us define a random variable $g(X)$ as

$$g(x) = 0 \;\; for \; x = 1,2, \ldots, n_0 - 1, n_0 + 2, n_0 + 3 \ldots, N$$

$$= c \; for \; x = \; n_0$$

$$= \; -c \;\; for \; x = \; n_0 + 1$$

where $c \neq 0$ real number.

Then,

$$E_N[g(X)] = \frac{1}{N}\sum_{x=1}^{N} g(x)$$

$$= \frac{1}{N} \sum_{x=1, x \neq n_0, n_0+1}^{N} g(x) + \frac{1}{N}\, g(n_0) + \frac{1}{N} g(n_0 + 1)$$

$$= 0 + c - c = 0$$

Thus,

$$E_N[g(X)] = 0 \text{ for all } N \text{ integers } \geq 1$$

but

$$g(x) \neq 0 \text{ for all } x = 1,2, \ldots, N.$$

Hence, $\mathfrak{I}^*$ is not complete by definition.

Remark: Completeness is a property of a family of distributions. We know that if a statistic is sufficient for a class of distributions, it is sufficient for any subclass of those distributions. Completeness works in the opposite direction. The above example shows that the exclusion of even one number from the family $\mathfrak{I} = \{P_N;\ N \geq 1, \text{integers}\}$ destroys completeness. If a family of distributions $\mathfrak{I}$ is complete, it is sometimes possible to conclude the completeness of larger class.

Theorem: Let $\{f_{\theta:}\ \theta\ \varepsilon\ \Omega \subset R\}$ be one-parameter exponential family given by

$$f_\theta(x) = C(\theta) \exp[Q(\theta)x]\ h(x)$$

Suppose that $Q(\theta)$ contains an open interval in real line $\boldsymbol{R}$. Then $\{f_{\theta:}\ \theta\ \varepsilon\ \Omega \subset R\}$ is a complete family of distributions.

Proof: Sufficiency part can be easily proved by Factorization Theorem. Now, we shall prove the completeness. Let us write $Q(\theta) = \beta$ and the range of β as $(a, b) \subset R$. We have to show that for any random variable $g(X)$ such that

$$E_\beta[g(X)] = 0 \quad \text{for all } \beta\ \varepsilon\ (a, b)$$

$$\Rightarrow g(x) = 0 \qquad \text{for almost all } x\ \varepsilon\ R.$$

We prove the Theorem for continuous case only. The discrete case can be proved in the similar way just by replacing the integral sign by summation sign. We assume that the moment generating function of the random variable X exists and is finite in an interval containing zero. Now, for all $\beta\ \varepsilon\ (a, b)$, consider

$$E_\beta[g(X)] = \int_{-\infty}^{\infty} g(x) f_\beta(x)\, dx$$

$$= \int_{-\infty}^{\infty} g(x)\, C(\beta)\, exp[\beta x]\, h(x)dx = 0$$

$\Rightarrow E_\beta[g(X)] = \int_{-\infty}^{\infty} g(x)\ exp[\beta x]\, h(x)dx = 0$ as $C(\beta) > 0$.

Let us write

$$g(x) = g^+(x) - g^-(x),$$

where $g^+(x) = g(x)$ if $g(x) \geq 0$, and

$$= 0 \qquad \text{otherwise}$$

and $g^-(x) = -g(x)$ if $g(x) \leq 0$, and

$$= 0 \qquad \text{otherwise.}$$

Then, in terms of $g^+(x)$ and $g^-(x)$ we have

$$E_\beta[g(X)] = \int_{-\infty}^{\infty} [g^+(x) - g^-(x)]\ exp[\beta x]\, h(x)dx = 0$$

$$\Rightarrow \int_{-\infty}^{\infty} [g^+(x)]\ exp[\beta x]\, h(x)dx = \int_{-\infty}^{\infty} [g^-(x)]\ exp[\beta x]\, h(x)dx, \text{ for all } \beta\ \varepsilon\ (a, b)$$

Let $\beta_0\ \varepsilon\ (a, b)$ be a fixed value. Then

$$\int_{-\infty}^{\infty} [g^+(x)]\, exp[\beta_0 x]\, exp[(\beta - \beta_0)x]\, h(x)dx$$

$$= \int_{-\infty}^{\infty} [g^-(x)]\, exp[\beta_0 x]\, exp[(\beta - \beta_0)x]\, h(x)dx \qquad (1)$$

for all $\beta\ \varepsilon\ (a, b)$. Let us write $\beta - \beta_0 = t$, and the range of t is $(c, d) \subset R$. Define

$$f_{\beta_0}^+(x) = \frac{g^+(x) exp[\beta_0 x] h(x)}{\int_{-\infty}^{\infty} g^+(x) exp[\beta_0 x] h(x) dx} \text{ and}$$

$$f_{\beta_0}^{-}(x) = \frac{g^{-}(x)exp[\beta_0 x]h(x)}{\int_{-\infty}^{\infty} g^{-}(x)exp[\beta_0 x]h(x)dx}$$

then $f_{\beta_0}^{+}(x)$ and $f_{\beta_0}^{-}(x)$ are pdf's. Then, from (1)

$$\int_{-\infty}^{\infty} \frac{g^{+}(x)exp[\beta_0 x]h(x)}{\int_{-\infty}^{\infty} g^{+}(x)exp[\beta_0 x]h(x)dx} exp[tx]\, dx$$

$$= \int_{-\infty}^{\infty} \frac{g^{-}(x)exp[\beta_0 x]h(x)}{\int_{-\infty}^{\infty} g^{-}(x)exp[\beta_0 x]h(x)dx} exp[tx]\, dx, \text{for all } t\ \varepsilon\ (c,d)$$

$$\Rightarrow \int_{-\infty}^{\infty} exp[tx]\, f_{\beta_0}^{+}(x)dx = \int_{-\infty}^{\infty} exp[tx]\, f_{\beta_0}^{-}(x)dx, \text{for all } t\ \varepsilon\ (c,d)$$

Hence, by uniqueness of distribution function and moment generating function that

$f_{\beta_0}^{+}(x) = f_{\beta_0}^{-}(x)$, for almost all $x\ \varepsilon\ R$

$\Rightarrow\ g^{+}(x) = g^{-}(x)$, for almost all $x\ \varepsilon\ R$

as other terms are positive

$\Rightarrow g(x) = 0$, for almost all $x\ \varepsilon\ R$.

Hence, by definition, the family $\{f_{\theta}.\ \theta\ \varepsilon\ \Omega \subset R\}\ is$ complete.

Example: Let $(X_1, X_2, \dots, X_n)$ be a random sample from discrete Uniform distribution P_N with pmf

$$P_N[X = x] = \frac{1}{N}, x = 1,2, \dots ., N$$

Show that $T = \max_{1 \le i \le n} X_i$ is complete statistic.

We know that from earlier example that

$$P_N[T = t] = \frac{t^n - (t-1)^n}{N^n}, \text{ if } 1 \leq t \leq N$$

Let $g(T)$ be any random variable such that

$$E_N[g(T)] = 0 \text{ for all } N \geq 1 \text{ integers}$$

or

$$\frac{1}{N^n}\sum_{t=1}^{N} g(t)\,[t^n - (t-1)^n] = 0 \text{ for all } N \geq 1 \text{ integers}$$

For *N=1*, $g(1).1 = 0 \Rightarrow g(1) = 0$

For $N = 2$, $g(1).1 + g(2)[2^n - 1] = 0 \Rightarrow g(2) = 0, \text{as } g(1) = 0, \text{and so on}$

$$\Rightarrow g(1) = g(2) = \cdots . = g(N) = 0$$

$$\Rightarrow g(t) = 0 \text{ for all } t = 1,2, \dots, N \geq 1 \text{ integers}$$

Hence, pmf's of T is complete.

Example: Let $(X_1, X_2, \dots, X_n)$ be a random sample from discrete Uniform distribution $U(0, \theta)$. Show that $T = \max\limits_{1 \leq i \leq n} X_i$ is complete.

We know that the pdf of T is

$$f_T\,(t, \theta) = \frac{nt^{n-1}}{\theta^n} \quad 0 < t < \theta$$

$$= 0 \qquad \text{otherwise}$$

Let $g(T)$ be any random variable such that

$$E_\theta[g(T)] = 0 \text{ for all } \theta > 0.$$

$$\Rightarrow \int_0^\theta g(t) n t^{n-1} = 0, \text{for all } \theta > 0$$

$$\Rightarrow \int_0^\theta g(t) t^{n-1} = 0, \text{for all } \theta > 0$$

Let

$$h(\theta) = \int_0^\theta g(t) t^{n-1} = 0, \text{for all } \theta > 0$$

$$\Rightarrow \frac{dh(\theta)}{d\theta} = 0$$

$$\Rightarrow g(\theta)\theta^{n-1} = 0 \text{ for all } \theta > 0$$

$$\Rightarrow g(\theta) = 0 \text{ for all } \theta > 0$$

$$\Rightarrow g(t) = 0 \; for \; all \; t > 0$$

$$\Rightarrow g(t) = 0 \text{ for all } 0 < t < \theta.$$

Example: Let (X_1, X_2) be a random sample with joint pmf

X_1 \ X_2	0	1
0	$\frac{12-7\theta+\theta^2}{12}$	$\frac{4\theta-\theta^2}{12}$
1	$\frac{4\theta-\theta^2}{12}$	$\frac{\theta^2-\theta}{12}$

Then, show that

(i) $T = X_1 + X_2$,

(ii) $T = X_1 X_2$

are complete statistic.

(i) Let $g(T)$ be any statistic such that, for all $\theta \varepsilon (0,1)$, we have

$E_\theta[g(T)] = 0$

$g(0)P_\theta[T=0] + g(1)P_\theta[T=1] + g(2)P_\theta[T=2] = 0$

$g(0)\frac{(12-7\theta+\theta^2)}{12} + g(1)\frac{(4\theta-\theta^2)}{6} + g(2)\frac{(\theta^2-\theta)}{12} = 0$

$g(0) + \theta[8g(1) - g(2) - 7g(0)] + \theta^2[g(0) - 2g(1) + g(2)] = 0$

$\Rightarrow g(0) = 0, g(2) = 8g(1), g(2) = 2g(1)$

$\Rightarrow g(0) = g(1) = g(2) = 0$.

Hence $T = X_1 + X_2$ is complete.

(ii) Let $g(T)$ be any statistic such that, for all $\theta \varepsilon (0,1)$, we have

$E_\theta[g(T)] = 0$

$\Rightarrow g(0)P_\theta[T=0] + g(1)P_\theta[T=1] = 0$

$\Rightarrow g(0)\frac{(12+\theta-\theta^2)}{12} + g(1)\frac{(\theta^2-\theta)}{12} = 0$

$\Rightarrow g(0) + \theta[g(0) - g(1)] + \theta^2[g(1) - g(0)] = 0$

$\Rightarrow g(0) = g(1) = 0.$

Hence $T = X_1X_2$ is complete.

Basu' s Theorem: If $T(\boldsymbol{X})$ is a complete and minimal sufficient statistic, then $T(\boldsymbol{X})$ is independently of every ancillary statistic.

Proof: We give the proof for only discrete distributions.

Let $S(\boldsymbol{X})$ be any ancillary statistic. Then$P(S(\boldsymbol{X}) = s)$ does not depend on θ since $S(\boldsymbol{X})$ is ancillary. Also the conditional probability,

$$P(S(\boldsymbol{X}) = s|T(\boldsymbol{X}) = t) = P(\boldsymbol{X} \in \{\boldsymbol{x}: S(\boldsymbol{x}) = s\}|T(\boldsymbol{X}) = t),$$

does not depend on θ because $T(\boldsymbol{X})$ is a sufficient statistics (recall the definition!).Thus, to show that $S(\boldsymbol{X})$ and $T(\boldsymbol{X})$ are independent, it suffices to show that

$$P(S(\boldsymbol{X}) = s|T(\boldsymbol{X}) = t) = P(S(\boldsymbol{X}) = s) \qquad (1)$$

for all possible values $t \in T$. Now,

$$P(S(\boldsymbol{X}) = s) = \sum_{t\in T} P(S(\boldsymbol{X}) = s|T(\boldsymbol{X}) = t)P_\theta(T(\boldsymbol{X}) = t).$$

Furthermore, since $\sum_{t\in T} P_\theta(T(\boldsymbol{X}) = t) = 1$, we can write

$$P(S(\boldsymbol{X}) = s) = \sum_{t\in T} P(S(\boldsymbol{X}) = s)P_\theta(T(\boldsymbol{X}) = t).$$

Therefore, if we define the statistic

$$g(t) = P(S(\boldsymbol{X}) = s|T(\boldsymbol{X}) = t) - P(S(\boldsymbol{X}) = s),$$

The above two equation show that

$E_{\theta g}(T) = \sum_{t\in T} g(t)\, P_\theta(T(\boldsymbol{X}) = t) = 0$ for all θ.

Since $T(\boldsymbol{X})$ is a complete statistic, this implies that $g(t) = 0$ for all possible values $t \in T$. Hence (1) is verified.

Basu's Theorem is useful in that it allows us to deduce the independence of two statistics without ever finding the joint distribution of the two statistics.

Complete Statistics in the Exponential Family: Let $X_1, \ldots, X_n$ be iid observation from an exponential family with pdf or pmf of the form

$$f(x|\theta) = h(x)c(\theta)\exp\left(\sum_{j=1}^{k} w(\theta_j)t_j(x)\right), \qquad (2)$$

where $\theta = (\theta_1, \theta_2, \ldots, \theta_k)$. Then the statistic

$$T(\boldsymbol{X}) = \left(\sum_{i=1}^{n} t_1(X_i), \sum_{i=1}^{n} t_2(X_i), \ldots, \sum_{i=1}^{n} t_k(X_i)\right)$$

is complete as long as the parameter space Θ contains an open set $\Re^k$.

The condition that the parameter space contains an open set is needed to avoid a situation like the following. The $N(\theta, \theta^2)$ distribution can be written in the form (7); however, the parameter space (θ, θ^2) does not contain a two-dimensional open set, as it consists of only the points on a parabola. As a result, we can find a transformation of the statistics $T(\boldsymbol{X})$ that is an unbiased estimator of θ

(Recall that exponential families such as the $N(\theta, \theta^2)$, where the parameter space is a lower-dimensional curve, are called curved exponential families)

Example: Let $X_1, \ldots, X_n$ be iid exponential observations with parameter θ. Consider computing the expected value of

$g(\boldsymbol{X}) = \frac{X_n}{X_1 + \cdots + X_n}$.

We first note that the exponential distributions form a scale parameter family and thus, $g(\boldsymbol{X})$ is an ancillary statistic. The exponential distributions also form an exponential family with $t(x) = x$ and so,

$$T(\boldsymbol{X}) = \sum_{\boldsymbol{i}}^{\boldsymbol{n}} \boldsymbol{X_i}$$

is a complete sufficient statistic. (As noted below, we need not verify that $T(\boldsymbol{X})$is minimal, although it could easily be verified) Hence, by Basu's theorem, $T(\boldsymbol{X})$ and $g(\boldsymbol{X})$ are independent.

Thus, we have

$$\theta = E_\theta X_n = E_\theta T(\boldsymbol{X}) g(\boldsymbol{X}) = \big(E_\theta T(\boldsymbol{X})\big)\big(E_\theta g(\boldsymbol{X})\big) = n\theta E_\theta g(\boldsymbol{X}).$$

Hence, for any $\theta, E_\theta g(\boldsymbol{X}) = n^{-1}$.

Theorem (without proof): If a minimal sufficient statistic exists, then any complete statistic is also a minimal sufficient statistic.

1.10 Self-Assessment Exercise

1. Let $P_\theta[X = x] = \left(\frac{\theta}{2}\right)^{|x|} (1-\theta)^{1-|x|}, x = -1, 0, 1, 0 < \theta < 1$. Then which of the following are/is sufficient based on a random sample (X_1, X_2) from $P_\theta[X = x]$?

(i) $T = X_1 + X_2$,

(ii) $T = X_1 X_2$,

(iii) $T = |X_1| + |X_2|$

2. Let (X_1, X_2) be a random sample from Poisson distribution $P(\theta)$. Let $T = X_1 + K_1 X_2 + K_2 X_3;\ K_1, K_2 = 2, 3, \ldots$. Show that T is not sufficient statistic for all K_1, K_2.

3. Let $(X_1, X_2, \ldots, X_n)$ be a random sample from a pdf or pmf given below. Show that the pdf or pmf belongs to one-parameter exponential family and identify $t(x), c(\theta), w(\theta)\ and\ h(x)$. Also, show that, in each case, $\sum_{i=1}^n t(X_i)$ is a sufficient statistic for θ. Find the pdf or pmf of $T^* = \sum_{i=1}^n t(X_i)$ belongs to one − parameter exponential family.

(i) $f(x;\theta) = N(\theta, 1)$

(ii) $f(x;\theta) = N(0,\theta)$

(iii) $f(x;\theta) =$ Gamma distribution $= G(1,\theta)$

(iv) $f(x;\theta) = b(1,\theta) =$ Binomial distribution

(v) $f(x;\theta) = P(\theta) =$ Poisson distribution

(vi) $f(x;\theta) = \theta x^{\theta-1}, 0 < x < 1; \theta > 0$

$= 0$, otherwise

(vii) $f(x;\theta) = \left(\frac{\theta}{c}\right)\left(\frac{c}{x}\right)^{\theta+1}, x > 0;\ \theta > 0$

$= 0$, otherwise

(viii) $f(x;\theta) = N(\theta,\theta), \theta > 0$

(ix) $f(x;\theta) = \theta,$ if $x = 1$

$= 1-\theta,$ if $x = 2;\ 0 < \theta < 1$

4. Let $(X_1, X_2, \ldots, X_n)$ be a random sample from a pdf or pmf given below. Show that the pdf's or pmf's do not belong to one-parameter exponential family. Why?

(i) $f(x;N) = P_N[X = x] = \frac{1}{N}, x = 1,2, \ldots., N, N$ is + ve integer

(ii) $f(x;\theta) = U(0,\theta) =$ Uniform distribution over the interval $(0,\theta)$

(iii) $f(x;\theta) = \frac{1}{\pi[x^2+\theta^2]}, -\infty < x, \theta < \infty$

(iv) $f(x;\theta) = exp-(x-\theta),\ x > \theta$

$= 0$ otherwise

(v) $f(x;\theta) = exp-|x-\theta|, -\infty < x < \infty$

5. Let X be a random variable from the following pdf or pmf:

(i) $b(n,\theta), 0 < \theta < 1$

(ii)$P(\theta), \theta > 0$

(iii) $NB(r,\theta), 0 < \theta < 1$

(iv) $G(1,\theta), \theta > 0$

(v)$N(\theta, 1)$

(vi) $f_\theta(x) = \theta x^{\theta-1}, 0 < x < 1; \theta > 0$

$= 0$ otherwise.

(vii) $P_\theta[X = 1] = \theta, P_\theta[X = 2] = 1 - \theta;\ 0 < \theta < 1$

Show that the family of pdf's or pmf's of X is complete.

6. Let X be a random variable from the following pdfs:

(i) $N(0,\theta), \theta > 0$

(ii)$U(-\theta, \theta), \theta > 0$

(iii) $f_\theta(x) = \frac{\theta}{\pi[\theta^2+x^2]}, \theta > 0$

(iv) $P_\theta[X = x] = \left(\frac{\theta}{2}\right)^{|x|} (1-\theta)^{1-|x|}, x = -1, 0, 1;\ 0 < \theta < 1$

Show that the family of pdf's of X is not complete.

7. See whether the following family of pdf's are complete?

(i) $\{U(-\theta,\theta);\ \theta > 0\}$

(ii) U$(\theta, \theta + 1)$(Hint: Let $g(x) = sin2\pi x$)

(iii)U$(\theta - 1, \theta + 1)$(Let $g(x) = sin\pi x$)

(iv)$U\left(\theta - \frac{1}{2}, \theta + \frac{1}{2}\right)$ (Let $g(x) = sin2\pi x$)

8. Let $(X_1, X_2, \ldots, X_n)$ be a random sample from a pdf $U\left(\theta - \frac{1}{2}, \theta + \frac{1}{2}\right)$. Let $T = \left(\min_{1\le i\le n} X_i, \max_{1\le i\le n} X_i\right)$. Then, show that the family of pdf's of T is not complete.

9. Let X be a random variable with pmf

$$P_\theta[X = -1] = \theta,$$

$$P_\theta[X = x] = \theta^x(1-\theta)^2, x = 0,1,2, \ldots$$

Is $\{P_\theta : \theta\ \varepsilon\ (0,1)\}$ complete?

10. Let X be a random variable with pmf

$$P_\theta\ [X = 0] = \theta, P_\theta\ [X = 1] = 3\theta, P_\theta\ [X = 2] = 1 - 4\theta\ , \qquad 0 < \theta < 1$$

Is $\{P_\theta : \theta\ \varepsilon\ (0,1)\}$ complete?

11. Let $(X_1, X_2, \ldots, X_n)$ be a random sample from $N(\theta, \theta^2)$. Then, show that $T = \left(\sum_{i=1}^n X_i\ , \sum_{i=1}^n {X_i}^2\right)$ is sufficient but not complete.

12. Let $(X_1, X_2, \ldots, X_n)$ be a random sample from $N(\theta, 1)$. Let $T = \sum_{i=1}^n X_i$. Show that the family of pdf's of T is complete.

13. Let $(X_1, X_2, \ldots, X_n)$ be a random sample from a pdf

$f_\theta(x) = \theta x^{\theta-1}, 0 < x < 1; \theta > 0$

$= 0$ otherwise.

Let $= \sum_{i=1}^n -\log X_i$. Show that the family of pdf's of T is complete.

14. Let (X_1, X_2) be a random sample from pmf

$$P_\theta[X = x] = \left(\frac{\theta}{2}\right)^{|x|} (1-\theta)^{1-|x|}, x = -1, 0, 1, 0 < \theta < 1.$$

Show that $T = X_1 + X_2$ is not complete. What can you say about the completeness of $T = X_1X_2$?

15. Let $(X_1, X_2, \ldots, X_n)$ be a random sample from a pmf or pdf

(i) $b(1,\theta), 0 < \theta < 1$

(ii) $P(\theta), \theta > 0$

(iii) $NB(1,\theta), 0 < \theta < 1$

(iv) $G(1,\theta), \theta > 0$

Define $= \sum_{i=1}^{n} X_i$. Show that the family of pmf's or pdf's of T is complete

16. Let X_1 and X_2 be i.i.d. random variables having the Poisson distribution with parameter λ. Show, by considering the joint (conditional) distribution of X_1 and X_2, given $X_1 + X_2$, that $X_1 + X_2$ is sufficient for λ but not $X_1 + 2X_2$.

17. Let X_1 and X_2 be i.i.d. random variables having the discrete uniform distribution on $\{1, 2, \dots, N\}$, where N is unknown. Obtain the conditional distribution of X_1, X_2, given $T =$ max (X_1, X_2). Hence show that T is sufficient for N but $X_1 + X_2$ is not.

18 (i). If X_i $(i = 1, 2, \dots, n)$ be a random sample form a rectangular distribution with p.d.f.

$$f_\theta(x) = \begin{cases} \frac{1}{\theta} & if \quad 0 < x < \theta \\ 0 & \text{otherwise} \end{cases}$$

where $0 < \theta < \infty$, show that $X_{(n)}$ is a sufficient statistic for θ.

18(ii) Generalise this result by considering the p.d.f.

$$f_\theta(x) = \begin{cases} Q(\theta)M(x) & if \quad 0 < x < \theta \\ 0 & \text{otherwise} \end{cases}$$

where $0 < \theta < \infty, M(x) > 0$ and $[Q(\theta)]^{-1} = \int_0^\theta M(x)dx$.

19. For a random sample X_i $(i = 1, 2, \dots, n)$ from an exponential distribution with p.d.f.

$$f_\theta(x) = \begin{cases} \frac{1}{\theta}\exp[-\frac{x}{\theta}] & if \quad 0 < x < \infty \\ 0 & otherwise \end{cases}$$

where $0 < \theta < \infty$, show that $\sum_i X_i$ is a sufficient statistic for θ.

20. Let X_i $(i = 1, 2, \dots n)$ be a random sample from an exponential distribution with p.d.f.

$$f_\theta(x) = \begin{cases} \exp[-(x-\theta)] & if \quad 0 < x < \infty \\ 0 & otherwise \end{cases}$$

where $-\infty < \theta < \infty$, show that $X_{(1)}$ is a sufficient statistic for θ.

21. If a random sample of size $n \geq 2$ from a Cauchy distribution, whose p.d.f. is

$$f_\theta(x) = \frac{1}{\pi[1 + (x-\theta)^2]}$$

where $-\infty < \theta < \infty$, is considered, then can you have a single sufficient statistic for θ?

22. For a random sample size n from

$$f_\theta(x) = \begin{cases} \frac{1}{2} & if \quad \theta - 1 < x < \theta + 1 \\ 0 & otherwise \end{cases}$$

where $-\infty < \theta < \infty$, show that $X_{(1)}$ and $X_{(n)}$ are jointly sufficient for θ.

23. Let X_i $(i = 1, 2, \dots, n)$ be a random sample from a distribution with p.d.f.

$$f_\theta(x) = \begin{cases} \theta x^{\theta-1} & if \quad 0 < x < 1 \\ 0 & otherwise \end{cases}$$

where $0 < \theta < \infty$. Find a sufficient statistic for θ.

24. Show that for a random sample of size n form a distribution with p.d.f.

$$f_\theta(x) = \begin{cases} \frac{1}{\theta_2} \exp[-(x-\theta_1)/\theta_2] & if \quad \theta_1 < x < \infty \\ 0 & otherwise \end{cases}$$

where $\theta = (\theta_1, \theta_2)$ and $-\infty < \theta_1 < \infty,\ 0 < \theta_2 < \infty$, the statistics $X_{(1)}$ and $\sum_i X_i$ are jointly sufficient for θ_1 and θ_2.

25. Let $X_1, X_2, \dots, X_n$ be a random sample from a gamma distribution with p.d.f.

$$f_\theta(x) = \begin{cases} \frac{\alpha^p}{\Gamma(\mathrm{p})} \exp[-\alpha x] x^{p-1} & if \quad 0 < x < \infty \\ 0 & otherwise \end{cases}$$

where $\alpha > 0$ and $p > 0$.

Show that

(i) $\sum_i x_i$ and $\Pi_i x_i$ are jointly sufficient for $\theta = (\alpha, p)$;

(ii) if α is known, $\Pi_i x_i$ are sufficient for p;

(iii) if p is known, $\sum_i x_i$ is sufficient for α.

26. Show that if $X_1, X_2, \dots, X_n$ is a sample from any family of continuous distributions with distribution function $F_\theta(\theta \in \Theta)$, then the order statistics $X_{(1)}, X_{(2)}, \dots, X_{(n)}$ are jointly sufficient for θ.

27. The distribution of the discrete random variable X is known to be represented by either the p.m.f.

$$f_1(x) = \begin{cases} \frac{1}{5} & if\ x = 1, 2, \ldots, 5 \\ 0 & otherwise \end{cases}$$

or the p.m.f.

$$f_2(x) = \begin{cases} \frac{1}{5} & if\ x = 1, 2, \ldots, 5 \\ 0 & otherwise \end{cases}$$

Suggest a minimal sufficient statistic for the family (i.e. pair) of distributions.

28. Let X_1 be a sample of size 1 from $N(0, \sigma^2)$. Show that X_1^2 is a complete sufficient statistic for σ^2, while X_1 is sufficient but not complete.

29. Show that the following families of continuous probability, distributions are not complete :

(1) $$f_\theta(t) = \begin{cases} \frac{1}{2\theta} & if\ -\theta < t < \theta \\ 0 & otherwise \end{cases}$$

where $0 < \theta < \infty$;

(2) $$f_\theta(t) = \begin{cases} \frac{1}{\theta\sqrt{2\pi}} \exp(-t^2/2\theta^2) & if\ -\infty < t < \infty \\ 0 & otherwise \end{cases}$$

where $0 < \theta < \infty$.

(3) (Continuation) Let T have, for every $\theta \in \Theta$, a distribution which is symmetrical about 0. Show that if $\psi(T)$ be some odd function of T such that $E_\theta[\psi(T)]$ exists for every θ, then the family of distributions is incomplete.

UNIT 2: UNBIASED ESTIMATION, MINIMUM VARIANCE AND CRLB

Structure

2.1 Introduction

The unbiased estimation of the parameter of interest ensures that estimators are correct on average, without any systematic errors. The unbiasedness property is crucial for identifying the "best" estimator, which finding otherwise is a tricky affair. The unbiasedness provides the foundation for defining the concept of Uniformly Minimum Variance Unbiased Estimator (UMVUE). The UMVUE are vital for achieving the most accurate estimations with the least variability, enhancing the reliability of statistical inference. The UMVUE can be identified majorly in two ways: One by Cramér-Rao Inequality and Second by Lehmann-Scheffé Theorem. The Cramér-Rao Inequality, provides a benchmark to assess the efficiency of estimators, indicating whether an estimator is as precise as theoretically possible. Lehmann-Scheffé Theorem offers a systematic approach to finding the best possible unbiased estimator, streamlining the estimation process in practical applications.

Understanding and applying these principles allows statisticians to make precise and reliable inferences from data, which is essential in many fields such as economics, engineering, medicine, and the social sciences.

2.2 Objective

The objective of this unit is to provide us an basic understanding of the concepts related to unbiasedness, minimum variance unbiased estimation & Cramer Rao Lower Bound. The main emphasis of this unit is on the unbiased estimators, how to identify the best unbiased estimators, the concept of minimum variance unbiased estimator and methods of obtaining it using Cramer Rao Inequality as well as by using Lehmann Scheffe & Rao Blackwell Theorems. These concepts should be clear after reading this material.

2.3 Unbiasedness

Definition: A statistic $T(X)$ is called an unbiased estimator for a function of the parameter $g(\theta)$,provided that for every choice of θ.

$$ET(X) = g(\theta) \tag{1}$$

Any function $g(\theta)$ for which there exists a T satisfying (1) is usually referred to as an estimable function.

Any estimator that is not unbiased is called biased. The basis is denoted by $b(\theta)$.

$$b(\theta) = ET(X) - g(\theta) \tag{2}$$

Definition: The mean square error (MSE) of a statistic is defined as

$MSE[T(X)] = E[T(X) - g(\theta)]^2$

The MSE can be expressed as

$$MSE[T(X)] = E[T(X) - g(\theta)]^2 = E[T(X) - ET(X) + b(\theta)]^2$$

$$= E[T(X) - ET(X)]^2 + 2b(\theta)E[T(X) - ET(X)] + b^2(\theta)$$

$$= V[T(X)] + b^2(\theta)$$

$$= \text{Variance of } [T(X)] + [\text{bias of } T(X)]^2$$

Example: Let $(X_1, X_2, \ldots X_n)$ be Bernoulli rvs with parameter θ, where θ is unknown. $\bar{X}$ is an estimator for θ. Is it unbiased?

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

Thus, $\bar{X}$ is an unbiased estimator for θ.

We denote it as $\hat{\theta} = \bar{X}$.

$$Var(\bar{X}) = \frac{1}{n^2}\sum_{i=1}^{n} V(X_i) = \frac{n\theta(1-\theta)}{n^2} = \frac{\theta(1-\theta)}{n}$$

Example: Let $X_i (i = 1, 2, \ldots n)$ be iid rvs from $N(\mu, \sigma^2)$, where μ and σ^2 are unknown.

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

Consider

$$\sum_{i=1}^{n}(X_i - \mu)^2 = \sum_{i=1}^{n}(X_i - \bar{X} + \bar{X} - \mu)^2$$

$$= \sum_{i=1}^{n}(X_i - \bar{X})^2 + 2\sum_{i=1}^{n}(X_i - \mu)(\bar{X} - \mu) + n(\bar{X} - \mu)^2$$

$$= \sum_{i=1}^{n}(X_i - \bar{X})^2 + n(\bar{X} - \mu)^2$$

Therefore, $\sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n}(X_i - \mu)^2 - n(\bar{X} - \mu)^2$

$$E\left[\sum_{i=1}^{n}(X_i - \bar{X})^2\right] = E\left[\sum_{i=1}^{n}(X_i - \mu)^2\right] - nE[(\bar{X} - \mu)^2]$$

$$= n\sigma^2 - \frac{n\sigma^2}{n} = n\sigma^2 - \sigma^2$$

Hence, $E(S^2) = \sigma^2 - \frac{\sigma^2}{n} = \sigma^2\left(\frac{n-1}{n}\right)$

Thus, S^2 is a biased estimator of σ^2.

Hence, $b(\sigma^2) = \sigma^2 - \frac{\sigma^2}{n} - \sigma^2 = -\frac{\sigma^2}{n}$

Further, $\frac{nS^2}{n-1}$ is an unbiased estimator of σ^2.

2.4 Best Unbiased Estimators

A comparison of estimators based on MSE considerations may not yield a clear favourite. Indeed, there is no one "best MSE" estimator. Many find this troublesome or annoying, and rather than doing MSE comparisons of candidate estimators, they would rather have a "recommended" one.

If W_1 and W_2 are both unbiased estimators of a parameter θ, that is, $E_\theta W_1 = E_\theta W_2 = \theta$, then their mean squared errors are equal to their variances, so we should choose the estimator with the smaller variance. If we can find an unbiased estimator with uniformly smallest variance i.e. a best unbiased estimator then our task is done.

Suppose W^* is an estimator of θ with $E_\theta W^* = \tau(\theta) \neq \theta$, and we are interested in investigating the worth of W^*. consider the class of estimators

$$C_\tau = \{W: E_\theta W = \tau(\theta)\}$$

For any $W_1, W_2 \in C_\tau$, $Bias_\theta W_1 = Bias_\theta W_2$,

So $E_\theta(W_1 - \theta)^2 - E_\theta(W_2 - \theta)^2 = Var_\theta W_1 - Var_\theta W_2$,

and MSE comparisons, within the class C_τ can be based on variance alone. Thus, although we speak in terms of unbiased estimators, we really are comparing estimators that have the same expected value,$\tau(\theta)$.

Remark: The definition of unbiasedness, in particular, requires that T be integrable (summable), that is, $E_\theta(T)$ exists for all $\theta\ \varepsilon\ \Omega$. This definition can be extended to the case where θ is a parameter vector and T is a random vector.

Let θ $(g(\theta) = \theta\ )$ be estimable, and T be an unbiased estimator of θ. Let T_1 be another unbiased estimator of θ, different from T . This means that there exists at least one θ such that $P_\theta[T \neq T_1] > 0$. In this case there exists infinitely many unbiased estimators of θ of the form $\alpha T + (1-\alpha)T_1, \quad 0 < \alpha < 1$. It is therefore, desirable to find a procedure to differentiate between these estimators.

Definition: An estimator W^* is a best unbiased estimator of $\tau(\theta)$ if it satisfies $E_\theta W^* = \tau(\theta)$ for all θ and, for any other estimator W with $E_\theta W = \tau(\theta)$, we have $var_\theta W^* \leq var_\theta W$ for all θ. W^* is also called a uniform minimum variance unbiased estimator (UMVUE) of $\tau(\theta)$.

Remark: Let $\theta_0\ \varepsilon\ \Omega$ and $U(\theta_0)$ be class of all unbiased estimators T of θ_0 such that $E_{\theta_0}(T^2) < \infty$. Then $T_0\ \varepsilon\ U(\theta_0)$ is called a locally minimum variance unbiased estimator (LMVUE) at θ_0 if

$$E_{\theta_0}(T_0 - \theta_0)^2 \leq E_{\theta_0}(T - \theta_0)^2$$

holds for all $T\ \varepsilon\ U(\theta_0)$.

Theorem: Let U be the non-empty class of all unbiased estimators T of a parameter $\theta\ \varepsilon\ \Omega$ with $E_\theta(T^2) < \infty$ for all $\theta\ \varepsilon\ \Omega$. Let U_0 be the class of all unbiased estimators v of zero, that is

$$U_0 = \{v: E_\theta(v) = 0, E_\theta(v^2) < \infty, for\ all\ \theta\ \varepsilon\ \Omega\ \}.$$

Then, $T_0\ \varepsilon\ U$ is UMVUE if and only if.

$$E_\theta(v\ T_0) = 0, \text{for all } \theta\ \varepsilon\ \Omega \text{ and all } v\ \varepsilon\ U_0 \qquad (3)$$

Proof: The condition of the Theorem guarantee the existence of $E_\theta(v\ T_0)$ for all $\theta\ \varepsilon\ \Omega$ and all $v\ \varepsilon\ U_0$.

Suppose $T_0 \, \varepsilon \, U$ is a UMVUE and $E_{\theta_0}(\nu_0 \, T_0) \neq 0$ for some $\theta_0 \, \varepsilon \, \Omega$ and some $\nu_0 \, \varepsilon \, U_0$.

Then $T_0 + \lambda \, \nu_0 \, \varepsilon \, U$ for all real λ.

If $E_{\theta_0}(\, {\nu_0}^2) = 0$, then $E_{\theta_0}(\nu_0 \, T_0) = 0$ must hold since $P_{\theta_0}[\, \nu_0 = 0] = 1$.

Let $E_{\theta_0}(\, {\nu_0}^2) > 0$, Choose $\lambda_0 = \frac{-E_{\theta_0}(\nu_0 \, T_0)}{E_{\theta_0}(\, {\nu_0}^2)}$. Then

$$E_{\theta_0}(\, T_0 + \lambda \, \nu_0)^2 = \, E_{\theta_0}\left({T_0}^2\right) - \frac{\left(E_{\theta_0}(\nu_0 \, T_0)\right)^2}{E_{\theta_0}(\, {\nu_0}^2)} < E_{\theta_0}\left({T_0}^2\right) \qquad (4)$$

Since $T_0 + \lambda \, \nu_0 \, \varepsilon \, U$, and $T_0 \, \varepsilon \, U$, it follows from (4) that

$Var_{\theta_0}(T_0 + \lambda \, \nu_0) < Var_{\theta_0}(T_0)$,

which is a contradiction. Hence, (3) holds.

Conversely, let (3) holds for some $T_0 \, \varepsilon \, U$, and for all $\theta \, \varepsilon \, \Omega$ and all $\nu \, \varepsilon \, U_0$.

Let $T \, \varepsilon \, U$. Then $T_0 - T \, \varepsilon \, U_0$ and for every $\theta \, \varepsilon \, \Omega$

$E_\theta[\, T_0(\, T_0 - T)] = E_\theta\left({T_0}^2\right) - E_\theta(\, T_0 T) = 0$ (by (3)).

Hence, by Cauchy –Schwartz Inequality, we have

$$E_\theta\left({T_0}^2\right) = E_\theta(\, T_0 T) \leq \left[E_\theta\left({T_0}^2\right)\right]^{\frac{1}{2}} [E_\theta(T^2)]^{\frac{1}{2}}$$

If $E_\theta\left({T_0}^2\right) = 0$

$\Rightarrow P_\theta[\, T_0 = 0] = 1$, there is nothing to prove.

Otherwise

$$\left[E_\theta\left({T_0}^2\right)\right]^{\frac{1}{2}} \leq [E_\theta(T^2)]^{\frac{1}{2}}$$

$$\Rightarrow \left[E_\theta\left({T_0}^2\right)\right] \leq [E_\theta(T^2)]$$

or

$V_\theta(T_0) \leq V_\theta(T)$.

Since T is arbitrary, Theorem is proved.

Theorem: Let U be the non-empty class of all unbiased estimators T of a parameter $\theta\ \varepsilon\ \Omega$ with $E_\theta(T^2) < \infty$ for all $\theta\ \varepsilon\ \Omega$. Then there exists at most one UMVUE for θ.

Proof: If $T\ and\ T_0\ \varepsilon\ U$ are both UMVUE, then $T - T_0 \varepsilon\ U_0$ and

$E_\theta[\ T_0(T - \ T_0)] = 0$ for all $\theta\ \varepsilon\ \Omega$

$\Rightarrow E_\theta({T_0}^2) = E_\theta(\ T_0 T)$

$\Rightarrow V_\theta(\ T_0) = Cov_\theta(T,\ T_0)$ for all $\theta\ \varepsilon\ \Omega$.

Since $T\ and\ T_0$ are both UMVUE therefore $V_\theta(\ T_0) = V_\theta(T)$.

It follows that the correlation coefficient between T and T_0 is one.

This implies that

$P_\theta[\ aT + bT_0 = 0] = 1$ for some a, b and all $\theta\ \varepsilon\ \Omega$.

Since $T\ and\ T_0$ are both unbiased for $\varepsilon\ \Omega$, we must have

$P_\theta[\ T = T_0] = 1$ for all $\theta\ \varepsilon\ \Omega$.

Theorem: Let $\{T_n\}$ be a sequence of UMVUE's and T be a statistic with $E_\theta(T^2) < \infty$ for all $\theta\ \varepsilon\ \Omega$ and such that $E_\theta(T_n - T)^2 \to 0\ as\ n \to \infty$ for all $\theta\ \varepsilon\ \Omega$. Then, T is also the UMVUE.

Proof: T is unbiased follows from

$|E_\theta(T) - \theta| \leq |E_\theta(T - T_n)| \leq E_\theta|T - T_n| \leq [E_\theta(T - T_n)^2]^{\frac{1}{2}}$

For all $\nu\ \varepsilon\ U_0$, all $\theta\ \varepsilon\ \Omega$, and every $n = 1, 2, \ldots$

$E_\theta(T_n \nu) = 0$ (using (3)). Therefore,

$E_\theta(\nu T) = E_\theta(\nu T) - E_\theta(T_n \nu) = E_\theta[\nu(T - T_n)]$

$|E_\theta(\nu T)| \leq [E_\theta(\nu^2)]^{\frac{1}{2}}[E_\theta(T - T_n)^2]^{\frac{1}{2}} \to 0$ as $n \to \infty$

$\Rightarrow E_\theta(\nu T) = 0$ for all $\theta\ \varepsilon\ \Omega$ and all $\nu\ \varepsilon\ U_0$.

Hence, T is UMVUE.

Example: Let $(X_1, X_2, \ldots, X_n)$ be a random sample from $N(\theta, 1)$. Show that $\bar{X}$ is UMVUE for θ.

Let T be any unbiased estimator of θ.

For $\bar{X}$ to be UMVUE for θ we must have $Cov(\bar{X}, \bar{X} - T) = 0$.

Therefore, if

$Cov(\bar{X}, \bar{X} - T) = \sigma_{\bar{X}}^2 - Cov(\bar{X}, T) = \sigma_{\bar{X}}^2 - \rho_{(\bar{X},T)}\sigma_{\bar{X}}\sigma_T \neq 0 \Rightarrow \bar{X}$ is UMVUE for θ

CaseI: $\sigma_{\bar{X}}^2 - \rho_{(\bar{X},T)}\sigma_{\bar{X}}\sigma_T > 0$

$\Rightarrow \sigma_{\bar{X}}^2 > \rho_{(\bar{X},T)}\sigma_{\bar{X}}\sigma_T$

$\Rightarrow \sigma_{\bar{X}} > \rho_{(\bar{X},T)}\sigma_T, for\ all\ T$

$\Rightarrow$ for $T = \bar{X}$ that $\frac{1}{\sqrt{n}} > \frac{1}{\sqrt{n}}$ which is impossible.

Hence, $\sigma_{\bar{X}}^2 - \rho_{(\bar{X},T)}\sigma_{\bar{X}}\sigma_T \not> 0$.

Case II: $\sigma_{\bar{X}}^2 - \rho_{(\bar{X},T)}\sigma_{\bar{X}}\sigma_T < 0$

$\Rightarrow \sigma_{\bar{X}}^2 < \rho_{(\bar{X},T)}\sigma_{\bar{X}}\sigma_T$

$\Rightarrow \sigma_{\bar{X}} < \rho_{(\bar{X},T)}\sigma_T, for\ all\ T$

$\Rightarrow$for $T = X_1 \Rightarrow \frac{1}{\sqrt{n}} < \frac{1}{\sqrt{n}}$, which is not possible.

Hence, $\sigma_{\bar{X}}^2 - \rho_{(\bar{X},T)}\sigma_{\bar{X}}\sigma_T \not< 0$.

Therefore, in any case $\sigma_{\bar{X}}^2 - Cov(\bar{X}, T) = 0$. Hence, $\bar{X}$ is UMVUE.

Example (Poisson Unbiased Estimation): Let $X_1, \ldots, X_n$ be iid Poisson (λ), and let $\hat{X}$ and S^2 be the sample mean and variance, respectively. Recall that for the Poisson pmf both the mean and variance are equal to λ. Therefore, we have

$E_\lambda \bar{X} = \lambda$, for all λ,

and $E_\lambda S^2 = \lambda$, for all λ

so both $\bar{X}$ and S^2 are unbiased estimators of λ.

Even if we can establish that X is better than S^2, consider the class of estimators

$W_a(\bar{X}, S^2) = a\bar{X} + (1 - a)S^2$.

for every constant a, $E_\lambda W_a(\bar{X}, S^2) = \lambda$, so we now have infinitely many unbiased estimators of λ. Even if $\bar{X}$ is better than S^2, is it better than every $W_a(\bar{X}, S^2)$? Furthermore, how can we be sure that there are not other, better estimators lurking about?

This example shows some of the problems that might be encountered in trying to find a best unbiased estimator, and perhaps that a more comprehensive approach is desirable. Suppose that, for estimating a parameter $\tau(\theta)$ of a distribution $f(x|\theta)$, we Can specify a lower bound, say $B(\theta)$, on the Variance of any unbiased estimator of $\tau(\theta)$. If we can then find an unbiased estimator W^* satisfying $Var_\theta W^* = B(\theta)$, we have found a best unbiased estimator. This is the approach taken with the use of the Cramér-Rao Lower Bound.

2.5 Cramer Rao Lower Bound

According to this method, we first establish a lower bound for the variance of all unbiased estimators and then we attempt to identify an unbiased estimator with variance equal to the lower bound found. If that is possible, the problem is again solved.

The following regularity conditions will be employed in proving the main result in this section. We assume that $\Omega \subset \boldsymbol{R}$ and that $g(.)$ is real-valued function and differentiable for all $\theta\ \varepsilon\ \Omega \subset \boldsymbol{R}$.

Regularity Conditions:

Let X be a random variable with pdf or pmf $f(x|\theta), \theta\ \varepsilon\ \Omega \subset \boldsymbol{R}$. Then it is assumed that

(i) $f(x;\theta)$ is positive on a set x independent of $\theta\ \varepsilon\ \Omega$.

(ii) Ω is an open interval in $\boldsymbol{R}$ (finite or infinite).

(iii) $\frac{\partial f(x;\theta)}{\partial \theta}$ exists for all $\theta\ \varepsilon\ \Omega$ and all $x\ \varepsilon\ x$.

(iv) $\int_x \int_x \ldots \int_x f(x_1;\theta)f(x_2;\theta)\ldots f(x_n;\theta)\, dx_1 dx_2 \ldots dx_n$ or $\sum_x \sum_x \ldots \sum_x f(x_1;\theta)f(x_2;\theta)\ldots f(x_n;\theta)$ may be differentiated under the integral or summation sign, respectively.

(v) $E_\theta\left[\frac{\partial \log f(x;\theta)}{\partial \theta}\right]^2$ to be denoted by $I(\theta)$ is positive for all $\theta\ \varepsilon\ \Omega$.

(vi) $\int_x \ldots \int_x W(x_1, x_2, \ldots, x_n)\, f(x_1;\theta)\ldots f(x_n;\theta)dx_1 \ldots dx_n$ or $\sum_x \ldots \sum_x W(x_1, x_2, \ldots, x_n)\, f(x_1;\theta)\ldots f(x_n;\theta)$ may be differentiated under the integral or summation sign, respectively, $W(X_1, X_2, \ldots, X_n)$ is any unbiased estimator of $g(\theta)$.

Cramer-Rao Inequality: Let $X_1, \ldots, X_n$ be a sample with pdf $f(x|\theta)$, and let $W(\boldsymbol{X}) = W(X_1, \ldots, X_n)$ be any estimator satisfying

$$\frac{d}{d\theta} E_\theta W(\boldsymbol{X}) = \int_{\boldsymbol{x}} \frac{\partial}{\partial \theta}[W(x)f(x|\theta)]d\boldsymbol{x} \tag{4}$$

and $Var_\theta W(\boldsymbol{X}) < \infty$.

Then

$$Var_\theta\big(W(\boldsymbol{X})\big) \geq \frac{\left(\frac{d}{d\theta}E_\theta W(\boldsymbol{X})\right)^2}{E_\theta\left(\left(\frac{\partial}{\partial\theta}\log f(\boldsymbol{X}|\theta)\right)^2\right)} \tag{5}$$

Proof: The proof of this theorem is elegantly simple and is a clever application of the Cauchy-Schwarz Inequality or, stated statistically, the fact that for any two random variables X and Y,

$$[Cov(X,Y)]^2 \leq (VarX)(Var\ Y). \tag{6}$$

If we rearrange (6) we can get a lower bound on the variance of X,

$$Var\ X \geq [Cov(X,Y)]^2/Var\ Y.$$

The cleverness in this theorem follows from choosing X to be the estimator $W(\boldsymbol{X})$ and Y to be the quantity $\frac{\partial}{\partial\theta}\log f(\boldsymbol{X}|\theta)$ and applying the Cauchy-Schwarz Inequality.

First note that

$$\frac{d}{d\theta}E_\theta W(\boldsymbol{X}) = \int_x W(\boldsymbol{x})\left[\frac{\partial}{\partial\theta}f(\boldsymbol{x}|\theta)\right]dx$$

$$= E_\theta\left[W(\boldsymbol{X})\left(\frac{\frac{\partial}{\partial\theta}f(\boldsymbol{X}|\theta)}{f(\boldsymbol{X}|\theta)}\right)\right] \text{ (Multiply \& divide by } f(\boldsymbol{X}|\theta)) \quad (7)$$

$$= E_\theta\left[W(\boldsymbol{X})\frac{\partial}{\partial\theta}\log f(\boldsymbol{X}|\theta)\right], \qquad \text{(properly of logs)}$$

Which suggests a covariance between $W(\boldsymbol{X})$ and $\frac{\partial}{\partial\theta}\log f(\boldsymbol{X}|\theta)$. For it to be a covariance, we need to subtract the product of the expected values, so we calculate $E_\theta\left(\frac{\partial}{\partial\theta}\log f(\boldsymbol{X}|\theta)\right)$. But if we apply (7) with $W(\boldsymbol{x}) = 1$, we have

$$E_\theta\left(\frac{\partial}{\partial\theta}\log f(\boldsymbol{X}|\theta)\right) = \frac{d}{d\theta}E_\theta[1] = 0. \tag{8}$$

Therefore $Cov_\theta(W(\boldsymbol{X}), \frac{\partial}{\partial\theta}\log f(\boldsymbol{X}|\theta)$ is equal to the expectation of the product, and it follows from (7) and (8) that

$$Cov_\theta\left((W(\boldsymbol{X}), \frac{\partial}{\partial\theta}\log f(\boldsymbol{X}|\theta\right) = E_\theta\left(W(\boldsymbol{X})\frac{\partial}{\partial\theta}\log f(\boldsymbol{X}|\theta\right)$$

$$= \frac{d}{d\theta}E_\theta W(\boldsymbol{X}) \qquad (9)$$

Also, since $E_\theta\left(\frac{\partial}{\partial\theta}\log f(\boldsymbol{X}|\theta) = 0\right)$ we have

$$Var_\theta\left(\frac{\partial}{\partial\theta}\log f(\boldsymbol{X}|\theta)\right) = E_\theta\left(\left(\frac{\partial}{\partial\theta}\log f(\boldsymbol{X}|\theta)\right)^2\right) \qquad (10)$$

Using the Cauchy-Schwarz Inequality together with (9) and (10), we obtain

$$Var_\theta(W(\boldsymbol{X})) \geq \frac{\left(\frac{d}{d\theta}E_\theta W(\boldsymbol{X})\right)^2}{E_\theta\left(\left(\frac{\partial}{\partial\theta}\log f(\boldsymbol{X}|\theta\right)^2\right)},$$

Hence, proving the theorem.

Corollary (Cramer-Rao Inequality, iid case): If the assumptions of Cramer Rao Inequality are satisfied and, additionally, if $X_1, \ldots, X_n$ are iid with pdf $f(x|\theta)$, then

$$Var_\theta(W(\boldsymbol{X}) \geq \frac{\left(\frac{d}{d\theta}E_\theta W(\boldsymbol{X})\right)^2}{nE_\theta\left(\left(\frac{\partial}{\partial\theta}\log f(X|\theta\right)^2\right)}$$

Proof: We only need to show that

$$E_\theta\left(\left(\frac{\partial}{\partial\theta}\log f(\boldsymbol{X}|\theta)\right)^2\right) = nE_\theta\left(\left(\frac{\partial}{\partial\theta}\log f(X|\theta)\right)^2\right)$$

Since $X_1, \ldots, X_n$ are independent.

$$E_\theta\left(\frac{\partial}{\partial\theta}\log f(\boldsymbol{X}|\theta)\right)^2 = E_\theta\left(\left(\frac{\partial}{\partial\theta}\log\Pi_{i=1}^n f(X_i|\theta)\right)^2\right)$$

$$= E_\theta\left(\left(\sum_{i=1}^n \frac{\partial}{\partial\theta}\log f(X_i|\theta)\right)^2\right) \quad (property\ of\ logs)$$

$$= \sum_{i=1}^n E_\theta\left(\left(\frac{\partial}{\partial\theta}\log f(X_i|\theta)\right)^2\right)$$

$$+ \sum_{i\neq j} E_\theta\left(\frac{\partial}{\partial\theta}\log f(X_i|\theta)\frac{\partial}{\partial\theta}\log f(X_j|\theta)\right) \quad (11)$$

For $i \neq j$ we have

$$E_\theta\left(\frac{\partial}{\partial\theta}\log f(X_i|\theta)\frac{\partial}{\partial\theta}\log f(X_j|\theta)\right)$$

$$= E_\theta\left(\frac{\partial}{\partial\theta}\log f(X_i|\theta)\right)E_\theta\left(\frac{\partial}{\partial\theta}\log f(X_j|\theta)\right) \quad \text{(independence)}$$

$$= 0$$

Therefore, the second sum in (11) is 0, and the first term is

$$\sum_{i=1}^n E_\theta\left(\left(\frac{\partial}{\partial\theta}\log f(X_i|\theta)\right)^2\right) = nE_\theta\left(\left(\frac{\partial}{\partial\theta}\log f(X|\theta)\right)^2\right) \quad \text{(identicality)}$$

which establishes the corollary.

The quantity $E_\theta\left(\left(\frac{\partial}{\partial\theta}\log f(\boldsymbol{X}|\theta)\right)^2\right)$ is called the information number, or Fisher information of the sample. This terminology reflects the fact that the information number gives a bound on the variance of the best unbiased estimator of θ. As the information number gets bigger and we have more information about θ, we have a smaller bound on the variance of the best unbiased estimator.

In fact, the term Information Inequality is an alternative to Cramér-Rao Inequality. Before looking at some examples, we present a computational result that aids in the application of this theorem. Its proof is left as an Exercise.

Lemma: If $f(x|\theta)$ satisfies

$$\frac{d}{d\theta}E_\theta\left(\frac{\partial}{\partial\theta}\log f(X|\theta)\right)=\int\frac{\partial}{\partial\theta}\left[\left(\frac{\partial}{\partial\theta}\log f(x|\theta)\right)f(x|\theta)\right]dx$$

(true for an exponential family), then

$$E_\theta\left(\left(\frac{\partial}{\partial\theta}\log f(X|\theta)\right)^2\right)=-E_\theta\left(\frac{\partial^2}{\partial\theta^2}\log f(X|\theta)\right).$$

Example (Continuation): Here $\tau(\lambda)=\lambda$, so $\tau'(\lambda)=1$. Also, since we have an exponential family, using Lemma (11) gives us

$$E_\lambda\left(\left(\frac{\partial}{\partial\lambda}\log\Pi_{i=1}^n f(X_i|\lambda)\right)^2\right)=-nE_\lambda\left(\frac{\partial^2}{\partial\lambda^2}\log f(X|\lambda)\right)$$

$$=nE_\lambda\left(\frac{\partial^2}{\partial\lambda^2}\log\left(\frac{e^{-\lambda}\lambda^X}{X!}\right)\right)$$

$$=-nE_\lambda\left(\frac{\partial^2}{\partial\lambda^2}(-\lambda+X\log\lambda-\log X!)\right)$$

$$=-nE_\lambda\left(-\frac{X}{\lambda^2}\right)=\frac{n}{\lambda}.$$

Hence for any unbiased estimator, W, of λ, we must have

$$Var_\lambda W\geq\frac{\lambda}{n}.$$

Since $Var_\lambda\bar{X}=\frac{\lambda}{n}$, $\bar{X}$ is a best unbiased of λ.

It is important to remember that a key assumption in the Cramér-Rao Theorem is the ability to differentiate under the integral sign, which, of course, is somewhat restrictive. As we have seen, densities in the exponential class will satisfy the assumptions but, in general, such assumptions need to be checked, or contradictions such as the following will arise.

Example (Unbiased estimator for the scale uniform): Let $X_1,\ldots,X_n$ be iid with pdf $f(x|\theta)=\frac{1}{\theta},0<x<\theta$. Since $\frac{\partial}{\partial\theta}\log f(x|\theta)=-1/\theta$, we have

$$E_\theta\left(\left(\frac{\partial}{\partial\theta}\log f(X|\theta)\right)^2\right)=\frac{1}{\theta^2}.$$

The Cramer-Rao Theorem would seem to indicate that if W is any unbiased estimator of θ,

$$Var_\theta W \geq \frac{\theta^2}{n}.$$

We would now like to find an unbiased estimator with small variance. As a first guess, consider the sufficient statistic $Y=\max(X_1,\ldots,X_n)$, the largest order-statistic. The pdf of Y is $f_Y(y|\theta)=\frac{ny^{n-1}}{\theta^n}, 0<y<\theta$, so

$$E_\theta Y=\int_0^\theta \frac{ny^n}{\theta^n}dy=\frac{n}{n+1}\theta,$$

Showing that $\frac{n+1}{n}Y$ is an unbiased estimator of θ. We nest calculate

$$Var_\theta\left(\frac{n+1}{n}Y\right)=\left(\frac{n+1}{n}\right)^2 Var_\theta Y$$

$$=\left(\frac{n+1}{n}\right)^2\left[E_\theta Y^2-\left(\frac{n}{n+1}\theta\right)^2\right]$$

$$=\left(\frac{n+1}{n}\right)^2\left[\frac{n}{n+2}\theta^2-\left(\frac{n}{n+1}\theta\right)^2\right]$$

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

which is uniformly smaller than θ^2/n. This indicates that the Cramer Rao Theorem is not applicable to this pdf. To see that this is so, we can use Leibnitz's Rule to calculate

$$\frac{d}{d\theta}\int_0^\theta h(x)f(x|\theta)dx=\frac{d}{d\theta}\int_0^\theta h(x)\frac{1}{\theta}dx$$

$$=\frac{h(\theta)}{\theta}+\int_0^\theta h(x)\frac{\partial}{\partial\theta}\left(\frac{1}{\theta}\right)dx$$

$$\neq \int_0^{\theta} h(x)\ \frac{\partial}{\partial\theta} f(x|\theta)dx,$$

unless $\frac{h(\theta)}{\theta} = 0$ for all θ. Hence, the Cramér-Rao Theorem does not apply. In general, if the range of the pdf depends on the parameter, the theorem will not be applicable.

Example (Normal Variance Bound): Let $X_1, \ldots, X_n$ be iid $n(\mu, \sigma^2)$, and consider estimation of σ^2, where μ is unknown. The normal pdf satisfies the assumptions of the Cramer-Rao Theorem and Lemma 11, so we have

$$\frac{\partial^2}{\partial(\sigma^2)^2} \log\left(\frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} e^{-\frac{\left(\frac{1}{2}\right)(x-\mu)^2}{\sigma^2}}\right) = \frac{1}{2\sigma^4} - \frac{(x-\mu)^2}{\sigma^6}$$

and

$$-E\left(\frac{\partial^2}{\partial(\sigma^2)^2} \log f(X|\mu,\sigma^2)|\mu,\sigma^2\right) = -E\left(\frac{1}{2\sigma^4} - \frac{(x-\mu)^2}{\sigma^6}|\mu,\sigma^2\right)$$

$$= \frac{1}{2\sigma^4}.$$

Thus, any unbiased estimator,W, of σ^2 must satisfy

$$Var(W|\mu,\sigma^2) \geq \frac{2\sigma^2}{n}.$$

$$Var(S^2|\mu,\sigma^2) = \frac{2\sigma^4}{n-1},$$

so S^2 does not attain the Cramer-Rao Lower Bound.

In the above example we are left with an incomplete answer; that is, is there a better unbiased estimator of σ^2 than S^2, or is the Cramer Rao Lower Bound unattainable? The conditions for attainment of the Cramer Rao Lower Bound are actually quite simple. Recall that the bound follows from an application of the Cauchy-Schwarz Inequality, so conditions for attainment of the bound are the conditions for equality in the Cauchy-Schwarz Inequality.

Corollary (Attainment): Let $X_1, \dots, X_n$ be iid $f(x|\theta)$, where $f(x_i|\theta)$ satisfies the conditions of the Cramer Rao Thoerem. Let $L(\theta|\boldsymbol{x}) = \Pi_{i=1}^{n} f(x_i|\theta)$ denote the likelihood function. If $W(\boldsymbol{X}) = \boldsymbol{W}(X_1, \dots, Xn)$ is any unbiased estimator of $\tau(\theta)$, then $W(\boldsymbol{X})$ attains the Cramer Rao Lower Bound if and only if

$$a(\theta)[W(\boldsymbol{x}) - \tau(\theta)] = \frac{\partial}{\partial\theta} \log L(\theta|\boldsymbol{x}) \tag{12}$$

for some function $a(\theta)$.

Proof: The Cramer Rao Inequality, as given in (6), can be written as

$$\left[Cov_\theta\left(W(\boldsymbol{X}), \frac{\partial}{\partial\theta} \log \Pi_{i=1}^{n} f(\boldsymbol{X_i}|\theta)\right)\right]^2 \leq Var_\theta W(\boldsymbol{X}) Var_\theta\left(\frac{\partial}{\partial\theta} \log \Pi_{i=1}^{n} f(\boldsymbol{X_i}|\theta)\right)$$

and, recalling that $E_\theta W = \tau(\theta), E_\theta\left(\frac{\partial}{\partial\theta} \log \Pi_{i=1}^{n} f(\boldsymbol{X_i}|\theta)\right) = 0$, and using the results of Theorem 7, we can have equality if and only if $W(\boldsymbol{x}) - \tau(\theta)$ is proportional to $\frac{\partial}{\partial\theta} \log \Pi_{i=1}^{n} f(x_i|\theta)$. That is exactly what is expressed in (12). Hence proved.

Theorem: For any random variables X and Y,

$-1 \leq \rho_{XY} \leq 1$.

$|\rho_{XY}| = 1$ if and only if there exist numbers $a \neq 0$ and b such that $P(Y = aX + b) = 1$. If $\rho_{XY} = 1$, then $a > 0$, and if $\rho_{XY} = -1$, then $a < 0$.

Example (Continuation): Here we have

$$L(\mu, \sigma^2|\boldsymbol{x}) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} e^{-\left(\frac{1}{2}\right)\Sigma_{i=1}^{n}(x_i-\mu)^2/\sigma^2},$$

and hence

$$\frac{\partial}{\partial\sigma^2} \log L(\mu, \sigma^2|\boldsymbol{x}) = \frac{n}{2\sigma^4}\left(\sum_{i=1}^{n} \frac{(x_i - \mu)^2}{n} - \sigma^2\right).$$

Thus, taking $a(\sigma^2) = n/(2\sigma^4)$ shows that the best unbiased estimator of σ^2 is $\sum_{i=1}^{n}(x_i - \mu)^2/n$, which is calculable only if μ is known. If μ is unknown, the bound cannot be attained.

Remark: The expression $E_\theta\left[\frac{\partial}{\partial\theta}\log f(X_i;\theta)\right]^2$, denoted by $I(\theta)$, is called Fisher's Information (about θ) number; $nI(\theta)$ is called the Information (about θ) contained in the sample $(X_1, X_2, \dots, X_n)$.

Theorem: If equality occurs for some unbiased estimator $U = U(X_1, X_2, \dots, X_n)$ of $g(\theta)$ and if the indefinite integrals $\int k(\theta)\,d\theta$, $\int g(\theta)k(\theta)\,d\theta$ exist, where $k(\theta) = \pm\frac{\sigma_\theta(V_\theta)}{\sigma_\theta(U)}$, $V_\theta = \left[\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\log f(x_i;\theta)\right]$

then, for all $\theta\ \varepsilon\ \Omega$,

$$\prod_{i=1}^{n} f(x_i;\theta) = C(\theta)exp[Q(\theta)U(x_1, x_2, \dots, x_n)]\ h(x_1, x_2, \dots, x_n)$$

where $(\theta) = exp - [\int g(\theta)k(\theta)\,d\theta]\ and\ Q(\theta) = \int k(\theta)\,d\theta$.

Proof: The equality in CRI holds if and only if

$Cov_\theta^2(U, V_\theta) = \sigma_\theta^2(U)\sigma_\theta^2(V_\theta)$.

Hence, by Schwartz inequality this is equivalent to

$V_\theta = E_\theta(V_\theta) + k(\theta)[U - E_\theta(U)]$ with probability one,

where $k(\theta) = \pm\frac{\sigma_\theta(V_\theta)}{\sigma_\theta(U)}$

But $E_\theta(V_\theta) = 0$.

Therefore, $V_\theta = k(\theta)[U - E_\theta(U)] = k(\theta)[U - g(\theta)]$

$\Rightarrow \frac{\partial}{\partial\theta}\log\prod_{i=1}^{n} f(x_i;\theta) = k(\theta)[U - g(\theta)] = k(\theta)U - k(\theta)g(\theta)$

$\Rightarrow \quad \log \prod_{i=1}^{n} f(x_i;\theta) = U \int k(\theta)d\theta - \int g(\theta)k(\theta)\, d\theta + h^*(x_1, x_2, \dots, x_n)$

$$\Rightarrow \prod_{i=1}^{n} f(x_i;\theta) = exp\left[U \int k(\theta)d\theta - \int g(\theta)k(\theta)\, d\theta + h^*(x_1, x_2, \dots, x_n)\right]$$

$\Rightarrow \prod_{i=1}^{n} f(x_i;\theta) = C(\theta)exp[Q(\theta)U(x_1, x_2, \dots, x_n)]h(x_1, x_2, \dots, x_n)$

for all $\theta\ \varepsilon\ \Omega$, where, $(\theta) = exp[-\int g(\theta)k(\theta)\, d\theta]$, $Q(\theta) = exp[\int k(\theta)d\theta]$, and $h(x_1, x_2, \dots, x_n) = exp[h^*(x_1, x_2, \dots, x_n)]$.

Thus, if equality occurs in CR bound for some unbiased estimator, then the joint pdf or pmf of random sample $(X_1, X_2, \dots, X_n)$ is of one parameter exponential form, provided certain conditions are satisfied.

Theorem: Let $(X_1, X_2, \dots, X_n)$ be a random sample from a distribution with pdf or pmf $f(x|\theta)$ and assume that the regularity conditions (i)-(vi) are fulfilled. Then, for any unbiased estimator $U(X_1, X_2, \dots, X_n)$ of $g(\theta)$, $\sigma_\theta^2(U)$ is equal to CR bound if and only if there exists a real-valued function $k(\theta)$ of θ such that

$U(X_1, X_2, \dots, X_n) = \ g(\theta) + k(\theta)V_\theta(X_1, X_2, \dots, X_n) = g(\theta) + k(\theta)V_\theta$.

Proof: Under the regularity conditions (i)-(vi) we have

$\sigma_\theta^2(U) \geq \frac{[g'(\theta)]^2}{nI(\theta)} = \frac{[g'(\theta)]^2}{\sigma_\theta^2(V_\theta)}$, for all $\theta\ \varepsilon\ \Omega$.

The equality holds if and only if

$\sigma_\theta^2(U)\sigma_\theta^2(V_\theta) = [g'(\theta)]^2 = Cov_\theta^2(U, V_\theta)$

$\Rightarrow E_\theta[U - E_\theta(U)]^2 E_\theta[V_\theta - E_\theta(V_\theta)]^2 = {E_\theta}^2\left[\left(U - E_\theta(U)\left(V_\theta - E_\theta(V_\theta)\right)\right)\right]$

$$\Rightarrow E_\theta\left[\frac{U - E_\theta(U)}{\sigma_\theta(U)} \mp \frac{V_\theta - E_\theta(V_\theta)}{\sigma_\theta(V_\theta)}\right]^2 = 0 \text{ with probability one}$$

$\Leftrightarrow \frac{U - E_\theta(U)}{\sigma_\theta(U)} \mp \frac{V_\theta - E_\theta(V_\theta)}{\sigma_\theta(V_\theta)} = 0$ with probability one

$\Leftrightarrow \frac{U-E_\theta(U)}{\sigma_\theta(U)} = \pm \frac{V_\theta - E_\theta(V_\theta)}{\sigma_\theta(V_\theta)}$ with probability one

$\Leftrightarrow U - E_\theta(U) = \pm \frac{\sigma_\theta(U)}{\sigma_\theta(V_\theta)} V_\theta - E_\theta(V_\theta)$

$= \pm \frac{\sigma_\theta(U)}{\sigma_\theta(V_\theta)} V_\theta \; as \; E_\theta(V_\theta) = 0$

$\Leftrightarrow U = E_\theta(U) \pm \frac{\sigma_\theta(U)}{\sigma_\theta(V_\theta)} V_\theta$

$\Rightarrow U = g(\theta) \pm \frac{\sigma_\theta(U)}{\sigma_\theta(V_\theta)} V_\theta$

$= g(\theta) \pm k(\theta) V_\theta$

Example: Let $(X_1, X_2, \dots, X_n)$ be a random sample from a Poisson distribution $P(\theta)$. Show that UMVUE of $g(\theta) = P_\theta[X_1 = 0]$ does not attain the CR lower bound. Why?

We know that UMVUE of $g(\theta) = P_\theta[X_1 = 0]$ is $\left(1 - \frac{1}{n}\right)^T$ and its variance is $e^{-2\theta}\left(e^{\frac{\theta}{n}} - 1\right)$.

Now, we will find the CR bound.

$f(x;\theta) = \frac{e^{-\theta}\theta^x}{x!}$

$\Rightarrow \log f(x;\theta) = -\theta + x \log \theta - \log x!$

$\Rightarrow \frac{\partial \log f(x;\theta)}{\partial \theta} = -1 + \frac{x}{\theta} = \frac{x - \theta}{\theta}$

$\Rightarrow E_\theta\left[\frac{\partial \log f(x;\theta)}{\partial \theta}\right] = E_\theta\left(\frac{X-\theta}{\theta}\right) = 0$, and

$\Rightarrow E_\theta\left[\frac{\partial \log f(X;\theta)}{\partial \theta}\right]^2 = I(\theta) = E_\theta\left(\frac{X-\theta}{\theta}\right)^2 = \frac{1}{\theta}$.

Also, $\frac{dg(\theta)}{d\theta} = -e^{-\theta} = g'(\theta)$

$\Rightarrow$ CR Lower bound $= \frac{[g'(\theta)]^2}{nI(\theta)} = \frac{\theta e^{-2\theta}}{n}$.

Clearly,

Lower bound $\frac{\theta e^{-2\theta}}{n} < Variance\ of\ UMVUE\ \left(1-\frac{1}{n}\right)^T = e^{-2\theta}\left(e^{\frac{\theta}{n}}-1\right)$. This is because we can not write

$U = E_\theta(U) \pm \frac{\sigma_\theta(U)}{\sigma_\theta(V_\theta)} V_\theta$, where $U = \left(1-\frac{1}{n}\right)^T$.

Empirical Distribution Function

Let $X_1, X_2, \ldots X_n$ be a random sample from a continuous population with df F and pdf f. Then the order statistics $X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$ is a sufficient statistic.

Define $\hat{F}(x) = \frac{Number\ of\ X_i's \le x}{n}$, same thing can be written in terms of order statistics as,

$$\hat{F}(x) = \begin{cases} 0; & X_{(1)} > x \\ \frac{k}{n};\ X_{(k)} \le x \le X_{(k+1)} = \frac{1}{n}\sum_{j=1}^{n} \boldsymbol{I}(x - X_{(j)} \\ 1; & x \ge X_{(n)} \end{cases}$$

where $I(y) = \begin{cases} 1; & y \ge 0 \\ 0; & otherwise \end{cases}$

Example: Show that empirical distribution function is an unbiased estimator of $F(x)$

Consider

$$E\,\hat{F}(x) = \frac{1}{n}\sum_{j=1}^{n} P\left[X_{(j)} \le x\right]$$

$$= \tfrac{1}{n}\Sigma_{j=1}^{n}\Sigma_{k=j}^{n}\binom{n}{k}[F(x)]^k[1-F(x)]^{n-k}$$

$$= \tfrac{1}{n}\Sigma_{j=1}^{k}\Sigma_{k=1}^{n}\binom{n}{k}[F(x)]^k[1-F(x)]^{n-k}$$

$$= \tfrac{1}{n}\Sigma_{j=1}^{k}\Sigma_{k=1}^{n}\binom{n}{k}[F(x)]^k[1-F(x)]^{n-k}\sum_{\mathrm{j}=1}^{\mathrm{k}}(1)$$

$$= \tfrac{1}{n}\sum_{k=1}^{n} k\binom{n}{k}[F(x)]^k[1-F(x)]^{n-k}$$

$$= \tfrac{1}{n}[nF(x)]$$

$$= F(x)$$

Note: One can see that $\boldsymbol{I}(x - X_{(j)})$ is a Bernoulli random variable. Then $E\boldsymbol{I}\left(\boldsymbol{x} - X_{(j)}\right) = F(x)$, so that $E\,\hat{F}(x) = F(x)$, we observe that $\hat{F}(x)$ has a Binomial distribution with mean $F(x)$ and variance $\frac{F(x)[1-F(x)]}{n}$. Using central limit theorem, for iid rvs, we can show that as $n \to \infty$

$$\sqrt{n}\left[\frac{\hat{F}(x) - Fx)}{\sqrt{F(x)[1-F(x)]}}\right] \to N(0,1).$$

Recall that if X and Y are any two random variables, then, provided the expectations exist, we have

$EX = E[E(X|Y)]$ and

$Var\,X = Var[E(X|Y)] + E[Var(X|Y)]$.

2.6 Rao Blackwell & Lehman Scheffe's Theorems

Rao Blackwell Theorem: Let $h(X)$ be an unbiased estimator of $g(\theta)$. Let $T(X)$ be a sufficient statistic for θ. Define $\psi(T) = E(h|T)$. Then $E[\psi(T)] = g(\theta)$ and $V[\psi(T)] \leq V(h)] \ \forall \ \theta$.

The process of minimizing the MSE using sufficient statistic is known as Blackwellization.

Proof:

$$E[h(X)] = E[Eh(X)|T = t]$$

$$= E[\psi(T)]$$

$$= g(\theta) \qquad (6)$$

Hence $\psi(T)$ is unbiased estimator of $g(\theta)$

$$V[h(X)] = V\big[E\big(h(X)\big|T(X)\big)\big] + E\big[V\big(h(X)\big|T(X)\big)\big]$$

$$= V[\psi(T)] + E\big[V\big(h(X)\big|T(X)\big)\big]$$

Since $V[h(X)|T(X)] \geq 0$ and $E\big[V\big(h(X)\big|T(X)\big)\big] > 0$

Therefore,

$$V[\psi(T)] \leq V[h(X)] \qquad (7)$$

From the definition of sufficiency, we can conclude that the distribution that the distribution of $h(X)$ given $T(X)$ is independent of θ. Hence $\psi(T)$ is an estimator.

Lehmann-Scheffe Theorem: If T is a complete sufficient statistic and there exists and unbiased estimate h of $g(\theta)$, there exists a unique UMVUE of θ, which is given by $E(h|T)$.

Proof: Let h_1 and h_2 be two unbiased estimators of $g(\theta)$ Rao-Blackwell theorem, $E(h_1|T)$ and $E(h_2|T)$ are both UMVUE of $g(\theta)$.

Hence $E[E(h_1|T) - E(h_2|T)] = 0$

But T is complete therefore

$$[E(h_1|T) - E(h_2|T)] = 0$$

This implies $[E(h_1|T) = E(h_2|T)]$.

Hence, UMVUE is unique.

Example: Let $(X_1, X_2, \dots, X_n)$ be a random sample from discrete Uniform distributionP_N with pmf

$$P_N[X = x] = \frac{1}{N}, x = 1,2, \dots, N$$

Find UMVUE of N.

Consider,

$P[X_1 = x_1, T = t] = P[X_1 = x_1, T \leq t] - P[X_1 = x_1, T \leq t-1]$

$= P[X_1 = x_1, \text{all } X_i \leq t] - P[X_1 = x_1, \text{all } X_i \leq t-1]$

$= P[X_1 = x_1, X_2 \leq t, \dots, X_n \leq t] - P[X_1 = x_1, X_2 \leq t, \dots, X_n \leq t-1]$, if $x_1 = 1,2, \dots, t-1$

$= \frac{t^{n-1} - (t-1)^{n-1}}{N^n}$, if $x_1 = 1,2, \dots, t-1$

$= \frac{t^{n-1}}{N^n}$, if $x_1 = t$

Therefore,

$$P[X_1 = x_1 \,|T = t] = \frac{P_N[X_1 = x_1, T = t]}{P_N[T = t]}$$

$$= \frac{t^{n-1} - (t-1)^{n-1}}{t^n - (t-1)^n}, \text{if } x_1 = 1,2, \dots, t-1$$

$$= \frac{t^{n-1}}{t^n - (t-1)^n}, \text{if } x_1 = t$$

$E[V \,|T = t] = E[2X_1 - 1 \,|T = t]$

$$= \sum_{x_1=1}^{t-1}(2x_1 - 1)\frac{t^{n-1}-(t-1)^{n-1}}{t^n-(t-1)^n} + (2t-1)\frac{t^{n-1}}{N^n}$$

$$= \frac{t^{n-1}-(t-1)^{n-1}}{t^n-(t-1)^n}\sum_{x_1=1}^{t-1}(2x_1 - 1) + (2t-1)\frac{t^{n-1}}{N^n}$$

$$= \frac{t^{n+1}-(t-1)^{n+1}}{t^n-(t-1)^n}.$$

This is the UMVUE of N.

Example: Let $(X_1, X_2, \dots, X_n)$ be a random sample from a normal distribution $N(\mu, \sigma^2)$ with both μ *and* σ^2 unknown, and let $\xi_{\frac{1}{2}}$ for the 50% quantile of the distribution. Find UMVUE of $\xi_{\frac{1}{2}}$.

From the definition of $\xi_{\frac{1}{2}}$, we know that

$$P_\theta\left(X_1 \geq \xi_{\frac{1}{2}}\right) = \frac{1}{2}, \text{where } \theta = (\mu, \sigma^2)$$

$$P_\theta\left(X_1 \geq \xi_{\frac{1}{2}}\right) = P_\theta\left(\frac{X_1 - \mu}{\sigma} \geq \frac{\xi_{\frac{1}{2}} - \mu}{\sigma}\right)$$

$$= 1 - \Phi\left(\frac{\xi_{\frac{1}{2}}-\mu}{\sigma}\right)$$

$$\Rightarrow \Phi\left(\frac{\xi_{\frac{1}{2}} - \mu}{\sigma}\right) = \frac{1}{2}$$

Hence,

$$\Rightarrow \frac{\xi_{\frac{1}{2}}-\mu}{\sigma} = \Phi^{-1}\left(\frac{1}{2}\right)$$

and $\Rightarrow \xi_{\frac{1}{2}} = \mu + \sigma\Phi^{-1}\left(\frac{1}{2}\right)$.

Therefore, UMVUE of $\xi_{\frac{1}{2}}$ is

$\bar{X} + \frac{S}{c_n}\Phi^{-1}\left(\frac{1}{2}\right)$, where $S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$, $E_\theta\left(\frac{S}{c_n}\right) = \sigma$.

Example: Let $(X_1, X_2, \ldots, X_n)$ be a random sample from a Gamma distribution $G(1, \theta)$ with pdf

$f(x, \theta) = \frac{1}{\theta}\, e^{-\frac{x}{\theta}}$, $x, \theta > 0$. Find UMVUE of $g(\theta) = e^{-\frac{k_0}{\theta}}$, $k_0 > 0$.

Let $U(X_1) = 1$ if $X_1 > k_0$

$= 1$ otherwise

Then $E[U(X_1)] = e^{-\frac{k_0}{\theta}} = g(\theta)$.

Now we have to find $f(x_1 \mid t)$, where $t = \sum_{i=1}^{n} x_i$. Let us define transformation $v = t - x_1, w = x_1$, T and W are independent. Now, $t = v + w$, $x_1 = w$. The Jacobian of transformation is 1.

Hence joint pdf of (x_1, t) is

$f(x_1, t) = f_{V,W}[v(x_1, t), w(x_1, t)]$

$= \frac{1}{\Gamma(n-1)\theta^n}\,(t - x_1)^{n-2}\, e^{-\frac{t}{\theta}}, \quad 0 < x_1 < t < \infty$

$f(x_1 \mid t) = \frac{f(x_1, t)}{f_T(t)}$

$= \frac{\Gamma(n)\,(t-x_1)^{n-2}}{\Gamma(n-1)t^{n-1}}$, $\quad 0 < x_1 < t < \infty$

$\Phi(t) = E[U \mid T = t]$

$= P\left[X_1 > {k_0}/{t}\right]$

$= \int_{k_0}^{t} \frac{\Gamma(n)\,(t-x_1)^{n-2}}{\Gamma(n-1)t^{n-1}}\, dx_1$

$= \frac{(t-k_0)^{n-1}}{t^{n-1}}$

$= \left(\frac{t-k_0}{t}\right)^{n-1}$

Thus $\Phi(T) = \left(1 - \frac{k_0}{T}\right)^{n-1}$ is UMVUE because T is complete and sufficient statistic.

Example: Let $X_1, X_2, \ldots, X_n$ are iid rvs with $B(n,p), 0 < p < 1$. In this case, we have to find the UMVUE of $p^r q^s, q = 1-p, r, s \neq 0$ and $P[X \leq c]$. Assume n is known.

Binomial distribution belongs to exponential family. So that $\sum_{i=1}^{n} X_i$ is sufficient and complete for p.

The distribution of T is $B(mn,p)$. Let $U(t)$ be unbiased estimator for $p^r q^s$

$$\sum_{t=0}^{mn} u(t) \binom{nm}{t} p^t q^{nm-t} = p^r q^s \qquad (8)$$

$$\sum_{t=0}^{nm} u(t) \binom{nm}{t} p^{t-r} q^{nm-t-s} = 1$$

$$\sum_{t=r}^{nm-s} u(t) \frac{\binom{nm}{t}}{\binom{nm-s-r}{t-r}} \binom{nm-s-r}{t-r} p^{t-r} q^{nm-t-s} = 1$$

Then

$$u(t) = \frac{\binom{nm}{t}}{\binom{nm-s-r}{t-r}} = 1$$

Hence

$$u(t) = \begin{cases} \frac{\binom{nm-s-r}{t-r}}{\binom{nm}{t}}; & t = r, r+1, r+2, \ldots, nm-s \\ 0 \quad ; & otherwise \end{cases} \qquad (9)$$

To find UMVUE of $P[X \leq c]$, now

$$P[X \leq c] = \sum_{x=0}^{c} \binom{n}{x} p^x q^{n-x}$$

Then UMVUE of

$$p^x q^{n-x} = \frac{\binom{nm-n}{t-x}}{\binom{nm}{t}}$$

Hence UMVUE of $P[X \leq c]$

$$= \begin{cases} \sum_{x=0}^{c} \binom{n}{x} \frac{\binom{nm-n}{t-x}}{\binom{nm}{t}}; & t = x, x+1, x+2, \dots, nm-n+x, c \leq \min(t,n) \\ 0 \quad ; & otherwise \end{cases} \quad (9)$$

Note: UMVUE of $P[X = x] = \binom{n}{x} p^x q^{n-x}$ is $\frac{\binom{n}{x}\binom{nm-n}{t-x}}{\binom{nm}{t}}$; $x = 0, 1, 2, \dots, t$

Particular cases:

$r = 1, s = 0$. From (8), we will get UMVUE of p,

$$u(t) = \frac{\binom{nm-1}{t-1}}{\binom{nm}{t}} = \frac{t}{nm} \quad (10)$$

$r = 0, s = 1$. From (8), we will get UMVUE of q,

$$u(t) = \frac{\binom{nm-1}{t}}{\binom{nm}{t}} = \frac{nm-t}{nm} \quad 1 - \frac{t}{nm} \quad (11)$$

$r = 1, s = 1$. From (8), we will get UMVUE of pq,

$$u(t) = \left(\frac{t}{nm}\right) \left(\frac{nm-t}{nm-1}\right) \quad (12)$$

Example: Let $X_1, X_2, \dots, X_m$ are iid rvs with $P(\lambda)$. In this case we have to find UMVUE of

(i) $\lambda^r e^{-(s\lambda)}$ (ii) $P[X \leq c]$

Poisson distribution belongs to exponential family. So that $T = \sum_{i=1}^{n} X_i$ is sufficient and complete for λ.

The distribution of T is $P(m\lambda)$. Let $U(t)$ be unbiased estimator for $\lambda^r e^{-(s\lambda)}$

$$\sum_{t=}^{\infty} u(t) \frac{e^{-m\lambda}(m\lambda)^t}{t!} = e^{-s\lambda}\lambda^r \qquad (13)$$

$$\sum_{t=}^{\infty} u(t) \frac{e^{-(m-s)\lambda} m^t \lambda^{t-r}}{t!} = 1$$

$$\sum_{t=r}^{\infty} u(t) \frac{m^t}{(m-s)^{t-r}} \frac{(t-r)!}{t!} \frac{e^{-(m-s)\lambda}[(m-s)\lambda]^{t-r}}{(t-r)!} = 1$$

Then

$$u(t) \frac{m^t}{(m-s)^{t-r}} \frac{(t-r)!}{t!} = 1$$

$$u(t) = \begin{cases} \frac{(m-s)^{t-r}}{m^t} \frac{t!}{(t-r)!} ; t = r, r+1, \dots, s \le m \\ 0 \quad ; \quad otherwise \end{cases} \qquad (14)$$

To find UMVUE of $P[X \le c]$

$$P[X \le c] = \sum_{x=0}^{c} \frac{e^{-\lambda}\lambda^x}{x!}$$

Now, UMVUE of $e^{-\lambda}\lambda^x$ is $\frac{(m-1)^{t-x}}{m^t} \frac{t!}{(t-x)!}$ so

UMVUE of $P[X \le c] = \sum_{x=0} \frac{t!}{(t-x)!x!} \left(\frac{m-1}{m}\right)^t \left(\frac{1}{m-1}\right)^x$

$$= \begin{cases} \sum_{x=0}^{c} \binom{t}{x} \left(\frac{1}{m}\right)^x \left(\frac{m-1}{m}\right)^{t-x} ; \quad c \le t \\ 1 \quad ; \quad otherwise \end{cases} \qquad (15)$$

Remark:

UMVUE of $P|X = x| = \frac{e^{-\lambda}\lambda^x}{x!}$ Is $\binom{t}{x} \left(\frac{1}{m}\right)^x \left(\frac{m-1}{m}\right)^{t-x} ; x = 0, 1, \dots, t$

Particular cases:

$s = 0, r = 1$; from (14), we will get the UMVUE of λ.

$$u(t) = \frac{m^{t-1}t!}{m^t(t-1)!} = \frac{t}{m} \tag{16}$$

$s = 1, r = 0$; from (14), we will get the UMVUE of $e^{-\lambda}$,

$$u(t) = \left(\frac{m-1}{m}\right)^t \tag{17}$$

$s = 1, r = 1$; from (14), we will get the UMVUE of $\lambda e^{-\lambda}$,

$$u(t) = \frac{(m-1)^{t-1}t!}{m^t(t-1)!} = \left(\frac{m-1}{m}\right)^t \frac{t}{m-1} \tag{18}$$

Remark: Note that UMVUE of $\lambda e^{-\lambda} \neq (UMVUE\ of\ \lambda)\ (UMVUE\ of\ e^{-\lambda}$

Theorem: Let $\{f_\theta : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ be a $k-$ parameter exponential family given by

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \boldsymbol{exp}\left[\sum_{\boldsymbol{j=1}}^{\boldsymbol{k}} Q_j(\boldsymbol{\theta})\boldsymbol{T_j}(\boldsymbol{x}) + D(\boldsymbol{\theta}) + S(\boldsymbol{x})\right], \tag{1}$$

where $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_k) \in \Theta$, an interval in $\mathcal{R}_k, T_1, T_2, \dots, T_k$, and S are defined on $\mathcal{R}_n, \boldsymbol{T} = (T_1, T_2, \dots, T_k)$, and $\boldsymbol{x} = (x_1, x_2, \dots, x_n), k \leq \boldsymbol{n}$. Let $\boldsymbol{Q} = (Q_1, Q_2, \dots, Q_k$, and suppose that the range of $\boldsymbol{Q}$ contains an open set in $\boldsymbol{\mathcal{R}_n}$. Then

$$\boldsymbol{T} = (T_1(\boldsymbol{X}), T_2(\boldsymbol{X}), \dots, T_k(\boldsymbol{X}))$$

is a complete sufficient statistic.

Proof. For a complete proof in a general setting, we reader to Lehmann. Essentially, the unicity of the Laplace transform is used on the probability distribution induced by $\boldsymbol{T}$. We will content ourselves here by proving the result for the $k = 1$ case when f_θ is a PMF

Let us write $Q(\theta) = \theta$ in (2), and let $(\alpha, \beta) \subseteq \Theta$. We wish to show that

$$E_\theta g\big(T(\boldsymbol{X})\big) = \sum_t g\,(t) P_\theta\{T(\boldsymbol{X}) = t\}$$

$$= \sum_t g(t) \exp[\theta t + D(\theta) + S^*(t)] = 0 \ \forall \ \theta \qquad (2)$$

Implies that $g(t) = 0$.

Let us write $x^+ = x$ if $x \geq 0$, if $x < 0$, and $x^- = -x$ if $x < 0$,=0 if $x \geq 0$.

Then $g(t) = g^+(t) - g^-(t)$, and both g^+ and g^- are nonnegative functions. In terms of g^+ and g^-, (3) is the same as

$$\sum_t g^+(t) e^{\theta t + S^*(t)} = \sum_t g^-(t) e^{\theta t + S^*(t)} \qquad (3)$$

For all θ, let $\theta_0 \in (\alpha, \beta)$ be fixed, and write

$p^+(t) = g^+(t) e^{\theta_0 t + S^*(t)} / \sum_t g^+(t) e^{\theta_0 t + S^*(t)}$ and

$p^-(t) = g^-(t) e^{\theta_0 t + S^*(t)} / \sum_t g^-(t) e^{\theta_0 t + S^*(t)}$ (4)

Then both p^+ and p^- are PMFs, and it follows from (3) that

$$\sum_t e^{\delta t} p^+(t) = \sum_t e^{\delta t} p^-(t) \qquad (5)$$

For all $\delta \in (\alpha - \theta_0, \beta - \theta_0)$. By the uniqueness of MGFs (5) implies that

$$p^+(t) = p^-(t) \quad for\ all\ t$$

and hence that

$$g^+(t) = g^-(t) \quad for\ all\ t,$$

which is equivalent to

$$g(t) = 0 \quad for\ all\ t.$$

Since T is clearly sufficient (by factorization criterion), it is proved that T is a complete sufficient statistic.

Example: Let $X_1, X_2, \dots, X_n$ be iid $N(\mu, \sigma^2)$ RVs where both μ and σ^2 are unknown .We know that family of distributions of $\boldsymbol{X} = (X_1, X_2, \dots, X_n)$ is a two parameter exponential family with $T(X_1, X_2, \dots, X_n) = (\sum_1^n X_i, \sum_1^n X_i^2)$. Hence, it follows that T is a complete sufficient statistic.

Example: Let $X_1, X_2, \dots, X_n$ be iid $b(1, p)$ RVs. Then $T = \sum_1^n X_i$ is a sufficient statistic. We show that T is also complete; that is, the family of distributions of $T, \{b(n, p), 0 < p < 1\}$, is complete.

Example: Let X be $N(0, \theta)$. Then the family of PDFs $\{N0, \theta), \theta > 0\}$ is not complete since $EX = 0$ and $g(x) = x$ is not identically zero. Note that $T(X) = X^2$ is complete, for the PDF of $X^2 \sim \theta\chi^2(1)$ is given by

$$f(t) = \begin{cases} \dfrac{e^{-\left(\frac{t}{2\theta}\right)}}{\sqrt{2\pi\theta}}, & t > 0 \\ 0, & otherwise \end{cases}$$

$$E_\theta g(T) = \frac{1}{\sqrt{2\pi\theta}} \int_0^\infty g(t) t^{-\left(\frac{1}{2}\right)} e^{-\left(\frac{t}{2\theta}\right)} dt = 0 \qquad \forall \ \ \theta > 0,$$

which holds if and only of

$$\int_0^\infty g(t) t^{-\left(\frac{1}{2}\right)} e^{-\left(\frac{t}{2\theta}\right)} dt = 0,$$

and using the uniqueness property of Laplace transforms, it follows that

$$g(t) t^{-\frac{1}{2}} = 0 \qquad for\ all\ \ t > 0$$

that is, $g(t) = 0$. Hence the proof.

2.7 Self-Assessment Exercise

1. Let $(X_1, X_2, \ldots, X_n)$ be a random sample from pdf

$$f(x,\theta) = \theta x^{\theta-1}, \quad 0 < x < 1;\ \theta > 0$$

$$= 0 \qquad \text{otherwise.}$$

Show that $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n}(-\log X_i)$ is UMVUE for θ.

2. Let $(X_1, X_2, \ldots, X_n)$ be a random sample from gamma distribution $G\left(1, \frac{1}{\beta}\right)$. Show that $\bar{X}$ is UMVUE for $\theta = \frac{1}{\beta}$. Let $T_{(1)} = \min_{1\le i\le n}(X_1, X_2, \ldots, X_n)$.

Show that $E_\beta\left(T_{(1)}\right) = \frac{1}{\beta}$, and

$$Var\left(T_{(1)}\right) < Var(\bar{X}).$$

Note that, sometimes an estimator with larger variance may be preferred.

3. Let $(X_1, X_2, \ldots, X_n)$ be a random sample from pdf or pmf given below. Find UMVUE of $g(\theta)$ indicated against each problem:

(i) $f(x;\theta) = N(\theta, 1);\ g(\theta) = \theta^2$, by finding $f(x_1 \mid t)$, where $t = \bar{x}$

(ii) $f(x;\theta) = N(0, \theta);\ g(\theta) = \sqrt{\theta}$,

by finding $f(x_1 \mid t)$, where $t = \sum_{i=1}^{n} x_i^2$

(iii) $f(x;\theta_1,\theta_2) = N(\theta_1,\theta_2);\ g(\theta_1,\theta_2) = \frac{\theta_1}{\sqrt{\theta_2}}$, where θ_1 and θ_2 are unkown

(iv) $f(x;\theta_1,\theta_2) = N(\theta_1,\theta_2);\ g(\theta_1,\theta_2) = \theta_1\sqrt{\theta_2}$, where θ_1 and θ_2 are unkown

(v) $f(x;\theta_1,\theta_2) = N(\theta_1,\theta_2);\ g(\theta_1,\theta_2) = {\theta_1}^2$, where θ_1 and θ_2 are unkown

(vi) $P_\theta[X = x] = P(\theta);\ \ g(\theta) = e^{-2\theta}$

(vii) $P_\theta[X = x] = b(1,\theta);\ \ g(\theta) = Var_\theta(X)$

(viii) $P_\theta[X = x] = b(1, \theta);\ \ g(\theta) = \theta^2$

(ix) $P_\theta[X = x] = NB(1, \theta);\ \ g(\theta) = E_\theta(X)$

(x) $P_\theta[X = x] = NB(1, \theta);\ \ g(\theta) = P_\theta[X = 1]$

(xi) $f(x; \theta) = U(0, \theta);\ g(\theta) = \theta$

by finding $f(x_1 \mid t)$, where $t = \max_{1 \le i \le n} X_i$

(xii) $f(x; \theta) = \ C(\theta)\ exp\{Q(\theta)T(x)\}\ \ h(x).\ g(\theta) = E_\theta(X)$

(xiii) $f(x; \mu, \sigma^2) = N(\mu, \sigma^2), g(\Theta) = \xi_p,$

the pth quantile of the distribution, $\Theta = (\mu, \sigma^2)$

4. Let $(X_1, X_2, \dots, X_n)$ be a random sample from pdf or pmf given below. Find CR lower bound for an unbiased estimator of $g(\theta)$ as mentioned against each problem:

(i) $P_\theta[X = x] = P(\theta);\ \ g(\theta) = \theta, and\ e^{-2\theta}$

(ii) $P_\theta[X = x] = b(1, \theta);\ \ g(\theta) = Var_\theta(X)$

(iii) $P_\theta[X = x] = b(1, \theta);\ \ g(\theta) = \theta, and\ \theta^2$

(iv) $P_\theta[X = x] = NB(1, \theta);\ \ g(\theta) = E_\theta(X)$

(v) $P_\theta[X = x] = NB(1, \theta);\ \ g(\theta) = P_\theta[X = 1]$

(vi) $f(x; \theta) = N(\theta, 1);\ g(\theta) = \theta, and\ \theta^2$

(vii) $f(x; \theta) = N(0, \theta);\ g(\theta) = \sqrt{\theta}$

Also identify the function $g(\theta)$ and pdf or pmf for which the UMVUE does not attain the lower bound.

5. Let T_1 and T_2 be two unbiased estimators of $\gamma(\theta)$ having the same variance. Show that their correlation coefficient ρ_θ cannot be smaller than $2e_\theta - 1$, where e_θ is the efficiency of each estimator.

6. Is the sample mean necessarily an efficient estimator of the population mean for every distribution?

7. If $X_1, X_2, \dots, X_n$ be i.i.d. random variables with mean μ and unknown variance. Then show that $\bar{X}$ is the minimum- variance unbiased linear estimator of μ.

8. Show that $\bar{X}$ in random sampling from

$$f_\theta(x) = \begin{cases} \frac{1}{\theta} \exp[-x/\theta] & if \quad \theta < x < \infty \\ 0 & otherwise \end{cases}$$

where $0 < \theta < \infty$, is an MVB estimator of θ and has variance θ^2/n.

9. Let $X_{(1)}, X_{(2)}$ and $X_{(3)}$ be the order statistics of a random sample of size 3 from

$$f_\theta(x) = \begin{cases} \frac{1}{\theta} & if \quad \theta < x < \theta \\ 0 & otherwise \end{cases}$$

where $0 < \theta < \infty$. Show that $4X_{(1)}, 2X_{(2)}$ and $\frac{4}{3}X_{(3)}$ are all unbiased estimators for θ. Find the variance and state your conclusion.

10. If Y has the binomial distribution $b(m, n$, where $0 < p < 1$ amd m is known, of what quantity is Y^2 an unbiased estimator? Hence obtain unbiased estimators of p^2 and $var_p(Y)$.

11. For a random sample $X_1, X_2, \dots, X_n$ from the binomial population $b(m, p)$, where $0 < p < 1$ and m is known, obtain the MVU estimator of

$$q^m + \binom{m}{1} q^{m-1}p + \binom{m}{2} q^{m-2}p^2.$$

12. Suppose $X_1, X_2, \dots, X_n$ are a random sample from the distribution with p.d.f.

$$f_\theta(t) = \begin{cases} \frac{1}{\theta} \exp(-\frac{x}{\theta}) & if \quad 0 < x < \infty \\ 0 & otherwise, \end{cases}$$

Where $0 < \theta < \infty$. Show that the sample mean $\bar{X}$ is sufficient for θ and that $\frac{n-1}{n\bar{X}}$ is the only unbiased estimator of $1/\theta$ based on $\bar{X}$.

UNIT 3: MAXIMUM LIKELIHOOD ESTIMATION, CONSISTENCY & OTHER LOWER BOUNDS

Structure

3.1 Introduction

Maximum Likelihood Estimation (MLE) is a method for estimating the parameters of a statistical model. It identifies the values that maximize the likelihood function. The MLE have many desirable properties like consistency, efficiency, invariance and applicability as it can be applied to a wide variety of distributions and models, making them versatile and widely used. Sometimes, the closed form solutions by using conventional derivative may not yield proper solution. Then, in such situations, numerical methods like the Newton-Raphson (NR) method are used. The NR method is an iterative numerical technique that is applied to maximize the likelihood function in MLE by solving the equation where the derivative of the log-likelihood function equals zero.

3.2 Objective

The objective of this unit is to provide us a basic understanding of the concepts related to maximum likelihood estimation, consistency & other lower bounds. we have focussed mainly on maximum likelihood estimation, its importance and using numerical analysis-based solution if closed form cannot be easily obtained like Newton Raphson method. The concept of consistency and other large sample properties. The Bhattacharya and Chapman – Robbin – Kiefer Bounds have been discussed in this unit. These concepts should be clear after reading this material.

3.3 Maximum Likelihood Estimation

Consider the joint pdf $f(x_1,..,x_n,\theta)$ of the random sample $(X_1,..,X_n)$ of size n from the distribution of X having the þdf $f(x,\theta)$ whose parameters θ is unknown and needs to be estimated. When the values $(x_1,..,x_n)$ of the random sample $(X_1,..,X_n)$ are given, $f(x_1,..,x_n,..\theta)$ may be looked upon as a function of θ which is called the likelihood function of θ and is denoted by $L(\theta) = L(\theta, x_1,..,x_n)$. The likelihood function gives likelihood of the possible values of θ given that the random variables $(X_1,..,X_n)$ assumes the value $(x_1,..,x_n)$.

We want to know for what value of θ, the likelihood is largest for this set of observations. In other words, we want to find the value of θ, denoted by $\hat{\theta}$ which maximizes $L(x_1,..,x_n,\theta)$. The value $\hat{\theta}$ maximizes the likelihood function is in general, a function of $x_1,..,x_n$

$$\hat{\theta} = \hat{\theta}(x_1,..,x_n)$$

such that

$$L(\hat{\theta}) = \max \;\; L(\theta,,x_1,..,x_n) \;\; \theta \in \Omega$$

Then $\hat{\theta}$ is called the maximum likelihood estimator or MLE.

In many cases it would be more convenient to deal with log $L(\theta)$, rather then $L(\theta)$, since log $L(\theta)$ is maximized for some value of θ as $L(\theta)$. For obtaining MLE we find the value of θ for which

$$\frac{\partial}{\partial\theta}\log L(\theta) = 0$$

We must however, check that this provides the absolute maximum. It the derivative does not exists at $\theta = \hat{\theta}$ or equation (1) is not solvable this method of solving (1) will fail. In such situation, we shall obtain the MLE by inspection.

Theorem: Let T be a sufficient statistic for the family of pdf (pmf) $f(x|\theta, \theta \in \Theta)$. If an MLE of θ exists and it is unique then it is a function of T.

Proof: It is given that T is sufficient, from the factorization theorem,

$$f(x|\theta) = h(x)g(T|\theta)$$

Maximization of the likelihood function with respect to θ is therefore equivalent to the maximization of $g(T|\theta)$, which is a function of T alone.

Remark: This theorem does not say that a MLE is itself a sufficient statistic.

In Example 1, we have shown that MLE need not be a function of sufficient statistics (see Remark 1).

Example: Let $X_1, X_2, \ldots, X_n$ be iid rvs with the following uniform pdf

$$\cup\,(0, \theta)$$
$$\cup\,(\theta, 2\theta)$$
$$\cup\,(\theta - 1, \theta + 1)$$
$$\cup\,(\theta, \theta + 1)$$

The pdf of X is given by

$$f(x|\theta) = \begin{cases} \frac{1}{\theta} & ; 0 < x < \theta, \\ 0 & ; otherwise \end{cases} \qquad (1)$$

and the corresponding likelihood function is given by

$$L(\theta|x) = \theta^{-n}; 0 < x_i < \theta, i = 1,2,\ldots,n$$

Consider the order statistics $X_{(1)} < X_{(2)} < X_{(n)}$. Hence $0 < X_1 < X_2 < \ldots, < X_n < \theta < \infty$. Note that the support of θ is $X_{(n)} < \theta < \infty$

We have to minimize $L(\theta|x)$ which is equivalent to finding the minimum value of θ, and it is given by

$$\hat{\theta} = X_{(n)}. \tag{2}$$

Thus, MLE of θ is $X_{(n)}$

The pdf is given

$$f(x|\theta) = \begin{cases} \frac{1}{\theta} & ; 0 < x < \theta, \\ 0 & ; otherwise \end{cases}$$

and the corresponding likelihood function is given by

$$L(x|\theta) = \begin{cases} \theta^{-n} & ; \ \theta < X_{(1)} < X_{(n)} < 2\theta, i = 1,2,\ldots,n \\ 0 & ; \ otherwise \end{cases}$$

so that $\theta < X_{(1)}$ and $\frac{X_{(n)}}{2} < \theta$

$\Rightarrow$ $$\frac{X_{(n)}}{2} < \theta < X_{(1)}$$

Maximizing $L(\theta|x)$ occurs at minimum value of θ

That is, $$\hat{\theta} = \frac{X_{(n)}}{2} \tag{3}$$

The pdf its corresponding likelihood functions are given by

$$f(x|\theta) = \begin{cases} \frac{1}{\theta} & , \theta - 1 < x < \theta + 1, \\ 0 & , \quad \text{otherwise} \end{cases}$$

$$L(\theta|x) = \begin{cases} \frac{1}{2n} & ; \ \theta - 1 < X_{(1)} < X_{(n)} < \theta + 1, \\ 0 & ; \quad otherwise \end{cases}$$

The support of θ and $\left(X_{(n)} - 1\right) \leq \theta \leq (X_{(1)} + 1)$. Here any value of θ is MLE. Therefore,

$$\hat{\theta} = \alpha\left(X_{(n)} - 1\right) + (1 - \alpha)(X_{(1)} + 1) \tag{4}$$

where $\alpha \in [0,1]$

In this case

$$L(\theta|x) = \begin{cases} 1\,; & \theta < x < \theta + 1, \\ 0\,; & otherwise \end{cases}$$

and

$$L(\theta|x) = \begin{cases} 1\,; & \theta < X_{(1)} < X_{(n)} < \theta + 1, \\ 0\,; & otherwise \end{cases}$$

The support of θ and $\left(X_{(n)} - 1\right) \leq \theta \leq (X_{(1)} + 1)$. Here any value of θ is MLE.

$$\hat{\theta} = \alpha\left(X_{(n)} - 1\right) + (1 - \alpha)(X_{(1)} + 1) \qquad (5)$$

Remark:

In (iii) and (iv), from (4) and (5), we can conclude that MLE is not a function of sufficient statistics, if $\alpha = 0$ or 1. Thus, from (4) and (5), we can say that MLE is not unique.

Example: Find the MLE of parameter p or σ of the following pdf

$$f(x|p,\sigma) = \frac{1}{\Gamma p}\left(\frac{p}{\sigma}\right)^{p} e^{-px/\sigma} x^{p-1}; x > 0, p, \sigma > 0$$

For large value of p one should use $\Psi(p)$,

$$\Psi(p) = \log p - \frac{1}{2p} \; and \; \Psi'(p) = \frac{1}{p} + \frac{1}{2p^2},$$

where $\Psi(p)$ and $\Psi'(p)$ are known as trigamma functions,

$$\frac{d \log \Gamma p}{dp} = \Psi(p) \; and \; d\frac{\Psi(p)}{dp} = \Psi'(p) \qquad (6)$$

The corresponding likelihood function given by,

$$L(p,\sigma|x) = \frac{1}{\Gamma p}\,\frac{p}{n}\, p e^{-p\sigma \sum_{i=1}^{n} x_i} \prod_{i=1}^{n} x_i^{p-1}$$

$$\log L = -n \log \Gamma p + \mathrm{np}[\log p - \log \sigma] - \frac{\mathrm{p}}{\sigma}\sum_{i=1}^{n} x_i + (p-1)\sum_{i=1}^{n} \log x_i$$

Let G be the geometric mean of $x_1, x_2, \ldots, x_n$, then

$$\log G = \frac{1}{n}\sum_{i=1}^{n} \log x_i \Rightarrow n \log G = \sum_{i}^{n} \log x_i$$

$$\log L = -n \log \Gamma p + np\,[\log p - \log \sigma] - \frac{pn\bar{x}}{\sigma} + (p-1)n \log G$$

$$\frac{\partial \log L}{\partial \sigma} = -\frac{np}{\sigma} + \frac{np\bar{x}}{\sigma^2} = 0 \Rightarrow \hat{\sigma} = \bar{x}$$

$$\frac{\partial \log L}{\partial \sigma} = -n\Psi(p) + n\,[\log p - \log \sigma] + \frac{np}{\sigma} - \frac{n\bar{x}}{p} + n \log G$$

$$\Rightarrow \left[-n \log p + \frac{n}{2p}\right] + n\,[\log p - \log \sigma + 1] - n + n \log G = 0$$

$$\Rightarrow \frac{1}{2p} - \log \bar{x} + 1 - 1 + \log G = 0$$

$$\frac{1}{2p} - \log \frac{\bar{x}}{\mathrm{G}} = 0$$

$$\hat{p} = \frac{1}{2\log \frac{\bar{x}}{G}} \qquad (7)$$

Hence, MLE of p and σ are

$\hat{p} = \frac{1}{2\log \frac{\bar{x}}{G}}$ and $\hat{\sigma} = \bar{x}$ (66)

Further, we state the following theorems on MLE without proof.

Theorem: (Invariance Property of MLE): If $\hat{\theta}$ is the MLE of θ, then $\mathrm{h}(\hat{\theta})$is the MLE of $h(\theta)$, where $h(\theta)$is any continuous function of θ.

Theorem: Let $X_1, X_2, \ldots, X_n$ be an iid rvs having common pdf $f(x|\theta), \theta \in \Theta$.

Assumption: The derivative $\frac{\partial^i \log f(x|\theta)}{\partial \theta^i}, i = 1, 2, 3$ exists for almost all x and for every θ belonging to a non-degenerate interval in Θ

There exists function $H_1(x), H_2(x)$ and $H_3(x)$ such that $\left|\frac{\partial f}{\partial \theta}\right| < H_1(x), \left|\frac{\partial^2 f}{\partial \theta^2}\right| < H_2(x), \frac{\partial^3 f}{\partial \theta^3} < H_3(x), \forall \theta \in \Theta, \int H_1(x)dx < \infty, \int H_2(x)dx < \infty, \int H_3(x)dx < \infty,$

$$\int \left[\partial \log \frac{f(x|\theta)}{\partial \theta}\right]^2 f(x|\theta)dx$$

is finite and positive for every $\theta \in \Theta$.

If assumption (a)-(c) are satisfied and true parameter point θ_0 is an inner point then for sufficiently large n,

$$\sum_{j=1}^{n} \partial \log \frac{f(x_j|\theta)}{\partial \theta} = 0$$

has at least one root $\hat{\theta}_n$ which converges in probability to θ_0.

$\sqrt{n}(\hat{\theta}_n - \theta_0)$ converges in distribution to $N\left(0, I^1(\theta)\right)$, where

$$I(\theta) = \int \left(\frac{\partial \log f(x|\theta)}{\partial \theta}\right)^2 f(x|\theta)dx,$$

which is the Fisher information contained in the sample size n.

Huzurbazar Theorem: The consistent root is unique.

Wald Theorem: The estimate which maximizes the likelihood absolutely is a consistent estimate.

3.4 Newton-Raphson Method

The Newton-Raphson method is a powerful technique for solving equations numerically. Likes so much of the differential calculus, it is based on the simple idea of linear approximation.

Let $f(x)$ be a well-behaved function. Let x^* be a root of the equation $f(x) = 0$ which we want to find. To find let us start with an initial estimate x_0. From x_0, we produced to an improved estimate x_1 (if possible) then from x_1 and x_2 and so on. Continue the procedure until two consecutive values x_i and X_{i+1} in ith and $(i + 1)$th steps are very close or it is clear that two consecutive values are away from each other. This style of proceeding is called 'iterative procedure'.

Consider the equation $f(x) = 0$ with root x^*. Let x_0 be a initial estimate. Let $x^* = x_0 + h$ then $h = x^* - x_0$, the number h measures how far the estimate x_0 is from the truth. Since h is small, we can use linear approximation to conclude that

$$0 = f(x^*) = f(x_0 + h) \simeq f(x_0) + hf'(x_0)$$

and therefore, unless $f'(x_0)$ is close to 0,

$$h \simeq \frac{f(x_0)}{f'(x_0)}$$

This implies,

$$x^* = x_0 + h \simeq x_0 - \frac{f(x_0)}{f'(x_0)}$$

Our new improved estimate x_1 of x^* is given by

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}$$

Next estimate x_2 is obtained from x_1 in exactly the same way as x_1 was obtained from x_0

$$x_2 = x_1 - \frac{f(x_1)}{f'(x_1)}$$

Continuing in this way, if x_n is the current estimate, then next estimate x_{n+1} is given by

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)},$$

Example (Continuation): Consider the previous problem

$$\log L = -n \log \Gamma p + np\,[\log p - \log \sigma] - \frac{pn\bar{x}}{\sigma} + (p-1)n \log G$$

$$\frac{\partial \log L}{\partial \sigma} = -\frac{np}{\sigma} + \frac{np\bar{x}}{\sigma^2} = 0 \Rightarrow \hat{\sigma} = \bar{x}$$

$$\frac{\partial \log L}{\partial \sigma} = -n\Psi(p) + n\,[\log p - \log \sigma] + \frac{np}{\sigma} - \frac{n\bar{x}}{p} + n \log G$$

$$\Rightarrow \log p - \Psi(p) = \log \frac{\bar{x}}{G}$$

Let $\log \frac{\bar{x}}{G} = C$

Hence $\log p - \Psi(p) = C$. By Newton-Raphson iteration method gives

$$\hat{p}_k = \hat{p}_{k-1} - \frac{\log(\hat{p}_{k-1}) - \Psi(\hat{p}_{k-1}) - C}{(\hat{p}_{k-1})^{-1} - \Psi'(\hat{p}_{k-1})}$$

$\hat{p}_k$ denotes the kth iterate starting with initial trial value $\hat{p}_0$ and $\Psi'(p) = \frac{d\Psi}{d}$

The function $\Psi(p)$ and $\Psi'(p)$ are tabulated in Abramowitz and Stegun.

3.5 Consistent Estimators

Definition: A sequence of rvs $\{X_n\}$ is said to converge to X in probability,

denoted as $X_n \xrightarrow{P} X$, if for every $\epsilon > 0$, as $n \to \infty$

$$P[|X_n - X| \geq \epsilon] \to 0. \qquad (1)$$

Equivalently, $X_n \xrightarrow{P} X$, if for every $\epsilon > 0$, as $n \to \infty$

$$P[|X_n - X| \geq \epsilon] \to 1. \qquad (2)$$

Definition: The sequence of rvs $\{X_n\}$ is said to converge to X almost surely (a.s.) or almost certainly, denoted as $X_n \xrightarrow{a.s.} X \; iff \; X_n(w) \to X(w)$, for all w, except those belonging to a null set N.

Thus,

$X_n \xrightarrow{a.s.} X \; iff \; X_n(w) \to X(w) < \infty$, for $w \epsilon N^c$,

where $P(N) = 0$. Hence we can write as

$$P\left[\lim_n X_n = X\right] \to 1 \qquad (3)$$

Definition: Let $F_n(x)$ be the df of a rv X_n and $F(x)$, the df of X. Let $c(F)$ be the set of points of continuity of F. Then $\{X_n\}$ is said to converge to X in distribution or in law or weakly, denoted as $X_n \xrightarrow{L} X$ and/or $F_n \xrightarrow{W} F$, for every $x \in C(F)$.

It may be written as $X_n \xrightarrow{d} X$ or $F_n \xrightarrow{d} F$.

Theorem: Let k be a constant, $X_n \xrightarrow{L} k \Leftrightarrow X_n \xrightarrow{P} k$.

Definition: A sequence of rvs $\{X_n\}$ is said to converge to X in the rth mean if $E|X_n - X|^r \to 0$ as $n \to \infty$. It is denoted as $X_n \xrightarrow{r} X$.

For $r = 2$, it is called converges in quadratic mean or mean square.

Definition: Let $X_1, X_2, \ldots, X_n$ be a sequence of iid rvs with pdf(pmf) $f(x|\theta)$.

A sequence of point estimates T is called consistent estimator of θ, where $T = T(X_1, X_2, \ldots, X_n)$ if for a given $\epsilon, \delta > 0$, these exists $(n_0(\epsilon, \delta, \theta)$ such that $\forall \theta \in \Theta$

$$P[|T - \theta| < \epsilon] \geq 1 - \delta, \forall\, n > n_0 \qquad (4)$$

or, using the Definition 1.1, we can say that $T\theta \xrightarrow{P}$as $n \to \infty$.

Moreover, we can say that

$$P[|T - \theta| < \epsilon] \to 1 \qquad (5)$$

Note: Some authors (5) define as a weak consistency and if we use a Definition 2 then they define it as a strong consistency.

Chebyshev's Inequality:

Let X be a rv with $E(X) = \mu$ and $Var\ X = \sigma^2 < \infty$, for any $k > 0$

$$P[|X - \mu| \geq k\sigma] \leq \frac{1}{k^2} \qquad (6)$$

Theorem: $X_n \xrightarrow{r} X \Rightarrow X_n \xrightarrow{P} X$.

$$\lim_{n\to\infty} MSE\big(T(X)\big) = 0 \qquad (7)$$

which means that as the number of observations increase, the mse decreases to zero. For example, if $X_1, X_2, \ldots, X_n \sim N(\theta, 1)$, then $MSE(\bar{X}) = \frac{1}{n}$. Hence $\lim_{n\to\infty} MSE(\bar{X}) = 0, \bar{X}$ is consistent estimator of θ.

Example: Let $\{X_i\}_1^m$ be iid $B(n, p)$, then $\frac{\bar{X}}{n}$ is consistent estimator for $p\ \bar{X} = \frac{\sum_{i=1}^m X_i}{m}$.

Now, $MSE\left(\frac{\bar{X}}{n}\right) = \frac{pq}{mn}, q = 1 - p$. As $m \to \infty \Rightarrow MSE\left(\frac{\bar{X}}{n}\right) \to 0$

Example: Let $\{X_i\}_1^n$ be iid rvs with $P(\lambda)\lambda > 0$ then $\bar{X}$ is a consistent estimator of $\lambda, \bar{X} = n^{-1}\sum_{i=1}^n X_i$.

Now, $E(\bar{X} = \lambda)$ and $MSE(\bar{X}) = \frac{\lambda}{n} \to 0$ as $n \to \infty$.

Example: Let $\{X_i\}_1^n$ be iid rvs with $\cup\ (0, \theta), \theta > 0$.

$\bar{X}$ is not a consistent estimator of θ.

$MSEE(\bar{X}) = \frac{(3n+1)\theta^2}{12n}$ and $\lim_{n\to\infty} \frac{(3n+1)\theta^2}{12n} = \frac{\theta^2}{4} \neq 0$

But $X_{(n)}$ is a consistent estimator.

$E\left(X_{(n)}\right) = \frac{n\theta}{n+1}$ and $MSE(X_n) = \frac{2\theta}{(n+1)(n+2)} \to 0 \; as \; n \to \infty$

Use the Definition and assume $\epsilon \leq \theta$, so that

$$P\left[\left|X_{(n)} - \theta\right| < \epsilon\right] = P[\theta - \epsilon < X_{(n)} < \theta + \epsilon]$$

$$= \int_{\theta-\epsilon}^{\theta} \frac{nx^{n-1}}{\theta} dx$$

$$= 1 - \left(\frac{\theta-\epsilon}{\theta}\right)^n \to 1 \text{ as } n \to \infty$$

Consider a df of $X_{(n)}$, let it be $H_n(x, \theta)$

$$H_n = P\left[X_{(n)} \leq x\right]$$

$$= \begin{cases} 0 & ; \; x < 0 \\ \left(\frac{x}{\theta}\right)^n & ; 0 \leq x < \theta \\ 1 & ; \; x \geq \theta \end{cases}$$

$$\lim_{n\to\infty} H_n(x, \theta) = H(x, \theta),$$

where

$$H(x, \theta) = \begin{cases} 0 \; ; x < 0 \\ 1 \; ; x \geq 0 \end{cases}$$

we have

$$X_{(n)} \xrightarrow{P} \theta \Leftrightarrow H_n \xrightarrow{d} H$$

In this case $H(x, \theta)$ is a df of a singular random variable, i.e., $P[X = \theta] = 1$, then

$$X_n \xrightarrow{d} X \Rightarrow X_{(n)} \xrightarrow{P} \theta.$$

Example: Consider $\{X_i\}_i^n$ are iid rvs as Cauchy distribution with location parameter θ.

$$f(x|\theta) = \frac{1}{\pi}\left[\frac{1}{1 + (x-\theta)^2}\right]; x \in R, \theta \in R$$

then $\bar{X}$ is not consistent estimator for θ.

The distribution of $\bar{X}$ is Cauchy with parameter θ. Using the Definition

$$P[|\bar{X}-\theta|<\epsilon] = P[\theta-\epsilon<\bar{X}<\theta+\epsilon] \qquad (8)$$

$$= \int_{\theta-\epsilon}^{\theta+\epsilon} \frac{1}{\pi}\left[\frac{dx}{1+(x-\theta)^2}\right]$$

$$= \frac{2}{\pi}\tan^{-1}\epsilon$$

This does not tend to 1. Hence $\bar{X}$ is not a consistent estimator.

Example: Let $X_1, X_2, \ldots, X_n$ be iid $N(\mu, \sigma^2)$ rvs. We have to find the consistent estimator for σ^2.

We know that $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{(n-1)}, where\ S^2 = \frac{\sum(x_1-\bar{x})^2}{n-1}$.

From Chebyshev's Inequality, we set

$$k\sigma = \epsilon \Rightarrow k = \frac{\epsilon}{\sigma}$$

so that

$$P[|S^2-\sigma^2|>\epsilon] \leq \frac{Var(S^2)}{\epsilon^2}$$

$$= \frac{2\sigma^4}{(n-1)\epsilon^2} \to 0 \text{ as } n \to \infty$$

Hence S^2 is consistent estimator for σ^2.

Theorem: Let T be a consistent estimator for θ and let g be a continuous function then $g(t)$ is consistent for $g(\theta)$

Proof: Given any $\epsilon > 0$, there exist a $\delta > 0$, such that, $|g(t) - g(\theta)| < \epsilon$ whenever $|T - \theta| < \delta$

Therefore, $\{x||T - \theta| < \delta\} \subseteq \{x||g(t) - g(\theta)| < \epsilon\}$

Then $P\{x||g(t) - g(\theta)| < \epsilon\} \geq P\{x||T - \theta| < \delta\}$,

Hence,

$$P\{x|\ |g(t) - g(\theta)| < \epsilon\} \to 1$$

Because

$$P\{x|\ \ |T - \theta| < \delta\} \to 1$$

$g(t)$ is consistent for $g(\theta)$.

Example: Let $X_1, X_2, \ldots, X_n$ be iid $p(\lambda)$ rvs. To find the consistent estimator for $g(\lambda) = e^{-s\lambda}\lambda^r$. We know that $\bar{X}$ is consistent for λ.

Using the Theorem 2, $g(\bar{X}) = e^{-s\bar{X}}(\bar{X})^r$ is consistent for $g(\lambda) = e^{-s\lambda}\lambda^r$.

Example: Let $X_1, X_2, \ldots, X_m$ be iid $B(n, p)$ rvs. We know that $\frac{\bar{X}}{n} = (mn)^{-1}\sum_{i=1}^{m} X_i$ is consistent for p.

Now, using Theorem 2.1 $\binom{n}{x}\bar{X}^x(1 - \bar{X})^{n-x}$ is consistent for $\binom{n}{x}p^x q^{n-x}$, when $m \to \infty$.

Example: Let $X_1, X_2, \ldots, X_m$ be iid $B(n, p)$ rvs, where p is a function of θ, in Bioassay problem, $p(\theta) = \frac{\exp(\theta y)}{1+\exp(\theta y)}$, where $y > 0$ is a given dose level.

Now $\frac{\bar{X}}{n}$ is consistent for p.

$$\frac{\bar{X}}{n} = \frac{\exp(\theta y)}{1+\exp(\theta y)}$$

$$\Rightarrow \hat{\theta} = \frac{1}{y}\log\frac{\frac{\bar{X}}{n}}{1-\frac{\bar{X}}{n}}$$

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

Example: Let $X_1, X_2, \ldots, X_m$ be iid with $f(x|\theta)$,

$$f(x|\theta) = \theta x^{\theta-1} \;;\; 0 < x < 1, \qquad \theta > 0$$

Let $y = -\log x$

$$g(y|\theta) = \theta e^{-\theta y} \;;\; y > 0, \qquad \theta > 0$$

One can easily see that $-\frac{n}{\log\sum x_i}$ is consistent for θ.

Definition: Let X be a rv with its df $Fx|\theta, \theta \in \Theta$ then population quantile q_p is defined as

$$P[X \leq q_p] = p, \qquad 0 < p < 1$$

If $p = \frac{1}{2}$ then $q_{\frac{1}{2}}$ is median. If $p = \frac{i}{4}$ $(i = 1, 2, 3)$, then $q_{\frac{i}{4}}$ is called as i[th] Quartile. They are generally denoted as Q_1, Q_2, and Q_3.

If $p = \frac{i}{10}$ $(i = 1, 2, \ldots, 9)$, then $q_{\frac{i}{10}}$ is called as ith Decile. They are generally denoted by $D_1, D_2, \ldots, D_9$.

Let the r. v. X have exponential distribution with mean θ, then to find $Q_1, Q_2, Q_3, D_1, D_2, D_3$ and D_8:

$$f(x|\theta) = \frac{1}{\theta} e^{-\left(\frac{x}{\theta}\right)} \;; x > 0, \theta > 0$$

$$P[X \leq Q_1] = \frac{1}{4}$$

$$1 - e^{\frac{Q_1}{\theta}} = \frac{1}{4} \Rightarrow Q_1 = -\theta \log \frac{3}{4}$$

Similarly,

$Q_2 = -\theta \log \frac{1}{2}$ and $Q_3 = -\theta \log \frac{1}{4}$

$D_1 = -\theta \log \frac{9}{10}, D_3 = -\theta \log \frac{7}{10}$ and $D_8 = -\theta \log \frac{2}{10}$.

Lemma: Let X be a random variable with its $df\ F\ (x)$.The distribution of $F\ (x)$ is U(0,1)

Proof: Consider

$$P[F(X) \leq z] = P[X \leq F^{-1}(z)] = F[F^{-1}(z)] = z$$

Hence $F(x)$ is U (0,1).

Example: Let X_1, X_2, ...be iid $b(1, p)$RVs. Then $EX_1 = p$ and it follows by the WLLN that

$$\frac{\sum_1^n X_i}{n} \xrightarrow{P} p.$$

Thus $\bar{X}$ is consistent for p.Also, $(\sum_1^n X_i + 1)/(n+2) \xrightarrow{P} p$, so that a consistent estimator need not be unique. Indeed, if $T_n \xrightarrow{P} p\ and\ c_n \to 0$ as $n \to \infty$, then $T_n + c_n \xrightarrow{P} p$.

Theorem: If X_1, X_2 ...are iid RVs with common law $\mathcal{L}(X)$, and $E|X|^p < \infty$ for some positive integer p, then

$$\frac{\sum_1^n X_i^k}{n} \to EX^k \quad \text{for } 1 \leq k \leq p,$$

and $n^{-1}\ \sum_1^n X_i^k$ is consistent for $EX^k, 1 \leq k \leq p$.

Moreover, if c_n is any sequence of constant such that $c_n \to 0$ as $n \to \infty$, then $(n^{-1} \sum X_i^k + c_n)$ is also consistent for $EX^k, 1 \leq k \leq p$.

Also, if $c_n \to 1$ as $n \to \infty$, then $(c_n n^{-1} \sum X_i^k)$ is consistent for EX^k. This is simply a restatement of the WLLN for iid RVs.

Theorem: if T_n is a sequence of estimator such that $ET_n \to \psi(\boldsymbol{\theta})$ and $var(T_n) \to 0$ as $n \to \infty$, then T_n is consistent for $\psi(\boldsymbol{\theta})$.

Proof: Consider

$$P\{|T_n - \psi(\boldsymbol{\theta})| > \varepsilon\} \leq \varepsilon^{-2} E[T_n - ET_n + ET_n - \psi(\boldsymbol{\theta})]^2$$

$$= \varepsilon^{-2}\{var(T_n) + [ET_n - \psi(\boldsymbol{\theta})]^2\} \to 0 \text{ as } n \to \infty.$$

Other large-sample properties of estimators are asymptotic unbiasedness, asymptotic normality, and asymptotic efficiency.

Definition: Asymptotically Unbiased

A sequence of estimators $\{T_n\}$ is asymptotically unbiased for $\psi(\boldsymbol{\theta})$ if

$$\lim_{n\to\infty} E_\theta T_n(\boldsymbol{X}) = \psi(\boldsymbol{\theta})$$

for all $\boldsymbol{\theta}$.

Definition: Consistent Asymptotically Normal

A consistent sequence of estimators $\{T_n\}$ is said to be consistent asymptotically normal (CAN) for $\psi(\boldsymbol{\theta})$ if $T_n \sim AN(\psi(\boldsymbol{\theta}), v(\boldsymbol{\theta})/n)$ for all $\boldsymbol{\theta} \in \Theta$.

Definition: Best Asymptotically Normal

If $v(\theta) = 1/(I(\theta))$, where $I(\theta)$ is the Fisher information, then $\{T_n\}$ is known as a best asymptotically normal (BAN) estimator.

3.6 Bhattacharya Bounds

Hodge's Lemma: Let $S_1, S_2, \ldots, S_k$ and $T_1, T_2, \ldots, T_k$ be the two sets of random variables such that with probability one $S_i's$ are linearly independent, i.e.,

$$P[a_1S_1 + a_2S_2 + \ldots + a_kS_k = 0] = 1 \quad (1)$$

Further, we denote by

$$\Lambda = \text{Covariance matrix of } S_i, i = 1, 2, \ldots, k$$

$$M = \text{Covariance of matrix of } T_j, j = 1, 2, \ldots, k$$

$$N = \text{Covariance of matrix of } S_i \text{ and } T_j, i \neq j$$

Then the matrix $(M - N'\Lambda^{-1}N) \geq 0$ is positive semi-definite, i.e.,

$$v'(M - N'\Lambda N)v \geq 0$$

Proof: Without loss of generality, assume that $ES_i = 0, ET_j = 0$ and if ES_i and $ET_j \neq 0$, then let $S_i^* = S_i - ES_i$ and $T_j^* = T_j - ET_j$, then $Var(S_i) = Var(S_i^*)$ and $Var(T_j) = Var(T_j^*)$.

Using Cauchy-Schwarz inequality, we have

$$Cov^2(u'S, v'T) \leq Var(u's)Var(v'T) \quad (2)$$

$$[u'Cov(S, T)v]^2 \leq [u'Var(S)uVar(T)v]$$

$$(u'Nv)^2 \leq (u'\Lambda u)(v'Mv)$$

Suppose $\Lambda u = Nv \Rightarrow u = \Lambda Nv$

$$[(\Lambda^{-1}Nv)'Nv]^2 \leq [(\Lambda^{-1}Nv)'\Lambda\Lambda^{-1}Nv][v'Mv]$$

$$[v'N'\Lambda^{-1}Nv]^2 \leq [v'\mathrm{N}'\Lambda^{-1}\Lambda\Lambda^{-1}Nv][v'Mv]$$

$$[v'N'\Lambda^{-1}Nv]^2 \leq [v'\mathrm{N}'\Lambda^{-1}Nv][v'Mv]$$

$$v'N'\Lambda^{-1}Nv \leq [v'Mv]$$

Therefore,

$$v'(M - N'\Lambda^{-1}N)v \geq 0 \qquad (3)$$

Hence the proof.

Theorem: Let $X_1, X_2, \ldots\ldots, X_n$ be iid rvs with joint pdf $f(x_1, x_2, \ldots. x_n|\theta)$ satisfying the regularity conditions.

Let

$$S_i = \frac{1}{f(x|\theta)} \frac{\partial^i f(x|\theta)}{\partial \theta^i}$$

Then $ES_i = 0, i = 1,2, \ldots\ldots, k$

$$\Lambda = Covariance\ matrix\ of\ S_i, i = 1,2, \ldots\ldots k$$

$$N' = \left[g^{(1)}(\theta), g^{(2)}(\theta), \ldots, g^{(k)}(\theta)\right],$$

where $g^{(i)}(\theta) = \frac{\partial^i g(\theta)}{\partial \theta^i}; i = 1,2, \ldots\ldots, k$

$u(x_1, x_2, \ldots x_n)$ is an unbiased estimator of $g(\theta)$, then

$$V\big(u(x)\big) \geq L_k \qquad (4)$$

where $L_k = N'\,\Lambda^{-1}N$, (4) is called k-th Bhattacharya bound.

Proof: Let $u(x)$ be an unbiased estimator of $g(\theta)$

Hence,

$$\iint \ldots \int u(x) f(x|\theta) dx = g(\theta)$$

$$\frac{\partial}{\partial \theta} \iint \ldots \int u(x) f(x|\theta) dx = \frac{\partial g(\theta)}{\partial (\theta)}$$

$$\iint \dots \int \frac{u(x)}{f(x|\theta)} \frac{\partial f(x|\theta)}{\partial \theta} f(x|\theta) dx = g^{(1)}(\theta)$$

$$\iint \dots \int u(x) S_1 f(x|\theta) dx = g^{(1)}(\theta)$$

$$E[u(x)S_1] = g^{(1)}(\theta)$$

In general,

$$\iint \dots \int \frac{u(x)}{f(x|\theta)} \frac{\partial^i f(x|\theta)}{\partial \theta^i} f(x|\theta) dx = g^{(i)}(\theta)$$

$$E[u(x)S_1] = g^{(i)}(\theta)$$

We know that $ES_i = 0 \Rightarrow cov[u(x), S_1] = g^{(i)}(\theta)$

By using Hodge's Lemma (Theorem 1),

$$M - N'\Lambda^{-1}N \geq 0$$

In this case $M = Var[u(x)]$

$$Var[U(x)] \geq L_k$$

Hence , it can be easily seen that

$$L_k \geq L_{k-1} \geq \dots \geq L_1 \qquad (5)$$

Note:

For $k = 1$, $Var[u(x)] \geq L_1$

$$L_1 = \frac{\left[g^{(1)}(\theta)\right]^2}{Var(S_1)},$$

$$S_1 = \frac{1}{f(x|\theta)} \frac{\partial f(x|\theta)}{\partial \theta} = \frac{\partial \log f(x|\theta)}{\partial \theta}$$

$$Var(S_1) = Var\left[\frac{\partial \log f(x|\theta)}{\partial \theta}\right]$$

CR bound because a particular case of Bhattacharya bound for $k = 1$.

Steps to find Bhattacharya Bounds:

1. To get N', differentiate the given parametric function $g(\theta)$.

i.e. $N' = \left[\frac{\partial g(\theta)}{\partial \theta}, \frac{\partial^2 g(\theta)}{\partial \theta^2}, \ldots, \frac{\partial^k g(\theta)}{\partial \theta^k}\right]$

2. Find $S_i = \frac{1}{f(x|\theta)} \frac{\partial f(x|\theta)}{\partial \theta}; i = 1, 2, \ldots, k$ and verify $ES_i = 0$

3. Find $Var(S_i) = E(S_i)^2$ and $Cov(S_i, S_j) = E(S_i S_j)(i \neq j)$. Then obtain the covariance matrix of $(S_i, S_j), (i \neq j)$, i.e., Λ.

4. Calculate $N'\Lambda^{-1}N$.

Example: Let $X_1, X_2, \ldots, X_n$ be iid rvs with $N(\theta, 1)$. We will obtain the Bhattacharya bound for $g(\theta) = \theta^2$

$$N' = \left[g^{(1)}(\theta), g^{(2)}(\theta), \ldots, g^{(k)}(\theta)\right] = [2\theta, 2, \ldots, 0] \qquad (6)$$

Here, we can take $N' = [2\theta, 2]$.

$$f(x|\theta) = (2\pi)^{-\frac{n}{2}} \exp\left[-\frac{1}{2}\sum_{i=1}^{n}(x_i - \theta)^2\right]$$

$$\frac{\partial f(x|\theta)}{\partial \theta} = (2\pi)^{-\frac{n}{2}} n(\bar{x} - \theta) \exp\left[-\frac{1}{2}\sum_{i=1}^{n}(x_i - \theta)^2\right]$$

$$S_1 = \frac{1}{f(x|\theta)} \frac{\partial f(x|\theta)}{\partial \theta} = n(\bar{x} - \theta) \qquad (7)$$

Then $ES_1 = 0$

$$\frac{\partial^2 f(x|\theta)}{\partial\theta^2} = (2\pi)^{-\frac{n}{2}}\left[\left\{-n\exp\left[-\frac{1}{2}\sum_{i=1}^{n}(x_i-\theta)^2\right]\right\} + n^2(\bar{x}-\theta)^2\exp\left[-\frac{1}{2}\sum_{i=1}^{n}(x_i-\theta)^2\right]\right]$$

$$= (2\pi)^{-\frac{n}{2}}\exp\left[-\frac{1}{2}\sum_{i=1}^{n}(x_i-\theta)^2\right][-n+n^2(\bar{x}-\theta)^2]$$

$$S_2 = \frac{1}{f(x|\theta)}\frac{\partial^2 f(x|\theta)}{\partial\theta^2} = -n + n^2(\bar{x}-\theta)^2 \qquad (8)$$

Similarly, one can find $S_3, S_4, \ldots, S_k$.

$$ES_2 = -n + n = 0$$

$$Var(S_1) = ES_1^2 = n^2E(\bar{x}-\theta)^2 = n \qquad (9)$$

$$Var(S_2) = E[n^2(\bar{x}-\theta)^2 - n]^2$$

$$= n^2E[n(\bar{x}-\theta)^2 - 1]^2 \qquad (10)$$

$$= n^2E[n^2(\bar{x}-\theta)^4 - 2n(\bar{x}-\theta)^2 + 1]$$

Now, since $E(\bar{x}-\theta^4) = \frac{3}{n^2}$, we have

$$Var(S_2) = n^2\left[n^2\frac{3}{n^2} - 2n\frac{1}{n} + 1\right]$$

$$= n^2[3-2+1]$$

$$= 2n^2$$

$$Cov(S_1, S_2,) = ES_1S_2$$

$$= E[\{n(\bar{x}-\theta)\}\{n^2(\bar{x}-\theta)^2 - n\}]$$

$$= E[n^3(\bar{x}-\theta)^3] - E[n^2(\bar{x}-\theta)] = 0 \qquad (11)$$

Hence, we have

$$\Lambda = \begin{pmatrix} n & 0 \\ 0 & 2n^2 \end{pmatrix}$$

so that

$$\Lambda^{-1} = \begin{pmatrix} \frac{1}{n} & 0 \\ 0 & \frac{1}{2n^2} \end{pmatrix}$$

Now,

$$L_2 = N'\Lambda^{-1}N$$

$$= (2\theta \;\; 2)\begin{pmatrix} \frac{1}{n} & 0 \\ 0 & \frac{1}{2n^2} \end{pmatrix}\binom{2\theta}{2}$$

$$= \frac{4\theta^2}{n} + \frac{2}{n^2} \qquad (12)$$

and $L_1 = \frac{4\theta^2}{n}$ = CR lower bound

therefore, it can be easily seen that

$$L_1 < L_2.$$

3.7 Chapman – Robbin – Kiefer Bound

Theorem: Let the random vector X have a pdf (pmf) $f(x|\theta)$. Let $T(X)$ be an unbiased estimator $g(\theta)$, where $g(\theta)$ defined on Θ. Further, assume that $ET^2 < \infty$ for all $\theta \epsilon \Theta$. If $\theta \neq \theta \neq \alpha$, then assume that $f(x|\theta)$ and $f(x|\alpha)$ are different. Assume that $S(\theta) = \{f(x|\theta) > 0\}$, $S(\alpha) = \{f(x|\alpha) > 0\}$ and $S(\alpha) \subset S(\theta)$.

Then,

$$Var[T(X)] \geq \sup_{S(\alpha)\subset S(\theta),\alpha\neq 0} \frac{[g(\alpha)-g(\theta)]^2}{Var\left\{\frac{f(X|\alpha)}{f(X|\theta)}\right\}} \; \forall\theta \in \Theta \qquad (1)$$

Proof: Consider under $f(x|\theta)$ and $f(x|\alpha)$

$$ET(X) = \int T(X) f(x|\theta) dx = g(\theta)$$

$$ET(X) = \int T(X) f(x|\alpha) dx = g(\alpha)$$

$$g(\alpha) - g(\theta) = \int T(X) \frac{\left(f(x|\alpha) - f(x|\theta)\right)}{f(x|\theta)} f(x|\theta) dx$$

$$= \int T(X) \left[\frac{f(x|\theta)}{f(x|\theta)} - 1\right] f(x|\theta) dx$$

$$Cov\left[T(X), \frac{f(x|\theta)}{f(x|\theta)} - 1\right] = g(\alpha) - g(\theta)$$

Using Cauchy-Schwarz inequality,

$$Cov^2\left[T(X), \frac{f(x|\theta)}{f(x|\theta)} - 1\right] \leq Var[T(X)] Var\left[\frac{f(x|\theta)}{f(x|\theta)} - 1\right]$$

$$VarT(X) Var\left[\frac{f(x|\theta)}{f(x|\theta)}\right]$$

Therefore,

$$[g(\alpha) - g(\theta)]^2 \leq Var\, T(X) Var\left[\frac{f(x|\theta)}{f(x|\theta)}\right]$$

$$Var[T(X)] \geq \frac{[g(\alpha) - g(\theta)]^2}{Var\left\{\frac{f(x|\alpha)}{f(x|\theta)}\right\}} \;\forall\; \theta \in \Theta$$

Then, (1) follows immediately.

Chapman and Robbins (1951) had given the same above-mentioned theorem in different form.

3.8 Self-Assessment Exercise

1. Consider a random sample $X_1, X_2, \dots, X_n$ from the Cauchy distribution

$$f_\theta(x) = \frac{1}{\pi[1+(x-\theta)^2]}, -\infty < x < \infty, 0 < \theta < \infty.$$

(a) Show that the sample mean $\bar{X}$ is not a consistent estimator of θ.

(b) Show that the sample median $X^{\sim}$ is a consistent estimator of θ and that it has asymptotic efficiency

2. Let T and U be consistent estimators of $\gamma(\theta)$. Show that $aT + bU$, such that $a + b = 1$, is consistent for $\gamma(\theta)$.

3. Let T and U be consistent estimators of $\gamma(\theta)$ and $\delta(\theta)$, respectively. Show that

(a) $kT + lU$ is consistent for $k\gamma(\theta) + l\delta(\theta)$;

(b) TU is consistent for $\gamma(\theta)\delta(\theta)$; and

(c) T/U is consistent for $\gamma(\theta)/\delta(\theta)$, provided $\delta(\theta) \neq 0$ for every $\theta \in \Theta$.

4. Let T_i $(i = 1, 2, \dots, k)$ be independent unbiased estimators of $\gamma(\theta)$ and $var_\theta(T_i) = \sigma_i^2 < \sigma^2$, for each i.

If

$$\bar{T}_k = \frac{T_1 + T_2 + \cdots + T_k}{k},$$

Show that $\overline{T_k}$ is consistent for $\gamma(\theta)$, What will happen if T_i, for each i, has a bias β?

5. Find the ML estimators of the parameter θ for random samples from rectangular distributions of the following types:

(a) $f_\theta(x) = \begin{cases} \frac{1}{\theta} \text{ if } -\theta \leq x \leq 0 \\ 0 \quad otherwise \end{cases}$

(b) $f_\theta(x) = \begin{cases} \frac{1}{2\theta} \text{ if } -\theta \leq x \leq \theta \\ 0 \quad otherwise \end{cases}$

(c) $f_\theta(x) = \begin{cases} 1 \text{ if } \theta - \frac{1}{2} \leq x \leq \theta + \frac{1}{2} \\ 0 \quad otherwise \end{cases}$

6. Find the ML estimators of θ for random samples from the following continuous distributions:

(a) $f_\theta(x) = \begin{cases} \frac{1}{\theta}\exp[-\frac{x}{\theta}] \text{ if } 0 \le x < \infty \\ 0 \quad otherwise \end{cases}$

(b) $f_\theta(x) = \begin{cases} \exp[-(x-\theta)] \text{ if } \theta \le x < \infty \\ 0 \quad otherwise \end{cases}$

(c) $f_\theta(x) = \begin{cases} \frac{1}{\beta}\exp[-(x-\alpha)/\beta] \text{ if } \alpha \le x < \infty \\ 0 \quad otherwise \end{cases}$

where $\theta = (\alpha, \beta)$ and $\beta > 0$.

(d) $f_\theta(x) = \frac{1}{2}\exp[-|x-\theta|], -\infty \le x < \infty.$

(e) $f_\theta(x) = \theta x^{\theta-1}, 0 < x < 1, 0 < \theta < \infty.$

(f) $f_\theta(x) = \theta^x \exp(-\theta)\Gamma(x), x = 0, 1, 2, \ldots\ldots, 0 < \theta < \infty.$

(g) $f_\theta(x) = x^{l-1}\exp(-x/\theta)/\theta^l\Gamma(l)\ 0 < x < \infty, 0 < \theta < \infty,$ l is a known constant.

7. If $X_1, X_2, \ldots\ldots, X_m$ are a random sample from the distribution $N(\mu, \sigma^2)$ and $X_{m+1}, X_{m+2}, \ldots\ldots, X_n$ from the distribution $N(\mu, \gamma\sigma^2)$, γ being known, find the maximum likelihood estimator of μ.

8. In a set of Bernoulli trials if X denotes the number of trials necessary to produce n successes, show that the ML estimator of p, the probability of success, is n/X.

9. Let X be binomially distributed with p.m.f

$$f_\theta(x) = \begin{cases} \theta^x(1-\theta)^{1-x} \; if \; x = 0, 1 \\ 0 \qquad otherwise, \end{cases}$$

where $\frac{1}{4} \le \theta \le \frac{3}{4}$ Show that the ML estimator of θ is $(2X+1)/4$.

U.P. Rajarshi Tandon Open University, Prayagraj

MScSTAT – 102N / MASTAT – 102N
Statistical Inference

Course Design Committee

Prof. Ashutosh Gupta **Chairman**
Director, School of Sciences
U. P. Rajarshi Tandon Open University, Prayagraj

Prof. Anup Chaturvedi (Retd.) **Member**
Ex. Head, Department of Statistics
University of Allahabad, Prayagraj

Prof. S. Lalitha (Retd.) **Member**
Ex. Head, Department of Statistics
University of Allahabad, Prayagraj

Prof. Himanshu Pandey **Member**
Department of Statistics, D. D. U. Gorakhpur University, Gorakhpur

Prof. Shruti **Member-Secretary**
Professor, School of Sciences
U. P. Rajarshi Tandon Open University, Prayagraj

Course Preparation Committee

Prof. Shashi Bhushan **Writer**
Department of Statistics
University of Lucknow, Lucknow

Prof. S. K. Pandey **Editor**
Department of Statistics
University of Lucknow, Lucknow

Prof. Shruti **Course Coordinator**
School of Sciences
U. P. Rajarshi Tandon Open University, Prayagraj

MScSTAT – 102N/ MASTAT – 102N STATISTICAL INFERENCE
©UPRTOU
First Edition: ***July 2024***
ISBN : 978-93-48987-21-1

Printed and Published by Mr. Vinay Kumar, Registrar, Uttar Pradesh Rajarshi Tandon Open University, 2025.

Block & Units Introduction

The ***Block – 2, Testing of Hypotheses & Interval Estimation,*** is the second block, which is divided into four units.

In ***Unit – 4 – Basic Concepts*** discusses the fundamental concepts and definitions, concept of randomized test using test functions. The concept of UMP test using a BCR has been introduced. Neyman Pearson Lemma has been derived and some examples have been given for illustration.

In ***Unit – 5 – Generalized Neyman Pearson Lemma & UMPU*** has been discussed, Generalized NP Lemma has been derived and its utility in finding UMPU where UMP test does not exist has been extensively discussed. The general result for UMPU Test for exponential family has been derived & discussed extensively.

Unit – 6 – Likelihood Ratio Tests dealt with derivation of some important LR test.

Unit – 7 – Interval Estimation dealt with Confidence estimation and the relationship between UMP test and UMA confidence intervals.

At the end of every block/unit the summary, self-assessment questions and further readings are given.

UNIT 4: BASIC CONCEPTS

Structure

4.1 Introduction

The utility of hypothesis testing lies in its ability to make informed decisions about population parameters based on sample data. It provides a systematic framework to evaluate the plausibility of a claim (null hypothesis) by comparing it against an alternative hypothesis. This process helps in determining the statistical significance of results, guiding decisions in scientific research, medicine, business, and other fields. Hypothesis testing minimizes errors through controlled significance levels and *p*-values, ensuring robust and reliable conclusions, thus facilitating evidence-based practices and advancing knowledge across various domains.

4.2 Objective

The objective of this unit is to provide us a basic understanding of the concepts related to basic concepts of testing of hypotheses. In this unit, the fundamental concepts and definitions, concept of randomized test using test functions are defined and discussed with examples. The concept of UMP test using a BCR has been introduced. Neyman Pearson Lemma has been derived and some examples have been given for illustration. The concept

of unbiased test has been introduced as well. These concepts should be clear after reading this material.

4.3 Some Definitions

Before moving any further, we should discuss some basic concepts and definitions.

Let $(X_1, X_2, \ldots\ldots, X_n)$ be iid from $f(x|\theta), \theta \, \epsilon \Theta$ and $\Theta \subseteq \mathcal{R}$. Further, we assume that the functional form of $f(x|\theta)$ is known except the parameter θ. Also, we assume that θ contains at least two points.

Definition (Hypothesis): A parametric hypothesis is an assertion or statement about the unknown parameter θ of the population or the corresponding probability distribution.

Definition (Null Hypothesis): Any statement about the population or its parameter from which a given random sample $(x_1, x_2, \ldots\ldots, x_n)$ is purportedly have been drawn is called a null hypothesis. It is denoted by H_0.

Definition (Alternative Hypothesis) : The alternative hypothesis is denoted by H_1. It is an alternative statement that is tested against the null hypothesis.

It is a statement that something is happening. In most situations, this hypothesis is what the statistician hopes to prove. It may be a statement that the assumed status quo is false, or that there is a relationship, or that there is a difference.

Consider the following example of null hypothesis:

- Men and women have same I.Q.
- There is no difference between the mean pulse rate of men and women.
- The accused is not guilty.

Some examples of alternative hypothesis

- Men and women do not have the same I.Q.
- There is a difference between the mean pulse rate of men and women.
- The accused is guilty.

In notation, we can write:

$H_0: \theta \, \epsilon \Theta,$ where $\Theta_0 \subset \Theta$

$H_0: \theta \, \epsilon \Theta_1$, where $\Theta_1 \subset \Theta$

Definition (Simple and Composite Hypotheses) : If Θ_0 *or* $\Theta_1$contains only one point, we say that Θ_0 *or* Θ_1 is simple hypothesis, otherwise it is composite hypothesis.

If the hypothesis is simple, the probability distribution of X is completely specified.

For example, if the r.v. X is $N(\mu, 1)$ and $H_0: \mu = 4$ and $H_1: \mu = 5$. In this case under H_0, X is $N(4,1)$ and under H_1, X is $N(5,1)$. Hence it is simple hypothesis.

In another case let the rv X is $N(\mu, \sigma^2)$ and both μ and σ^2 are unkown.

$\Theta = \{(\mu, \sigma^2): -\infty < \mu < \infty, \sigma^2 > 0\}$

Let $H_0: \mu \leq \mu_0,\ \sigma^2 > 0$, where μ_0 is known constant against $H_1: \mu > \mu_0,\ \sigma^2 > 0$.

In this case, both null and alternative hypothesis are composite.

Type-1 and Type-2 Errors: Given the sample point $X = (x_1, x_2, \ldots., x_n)$, we have to find a decision rule which will lead to a decision to accept or reject the null hypothesis. Further partition the n-dimensional Euclidean space $\Re_n$ into two disjoint sets $A$and A^c.

If $x \epsilon A$, reject H_0 and $x \epsilon A^c$, accept H_0.

Definition (Critical Region): Let $X_1, X_2, \ldots, X_n$ are iid rvs from $f(x|\theta), \theta \epsilon$. The subset $'A'$ of $\Re_n$ such that if $x \epsilon A$ then H_0 is rejected with probability 1 is called the critical region or rejection region.

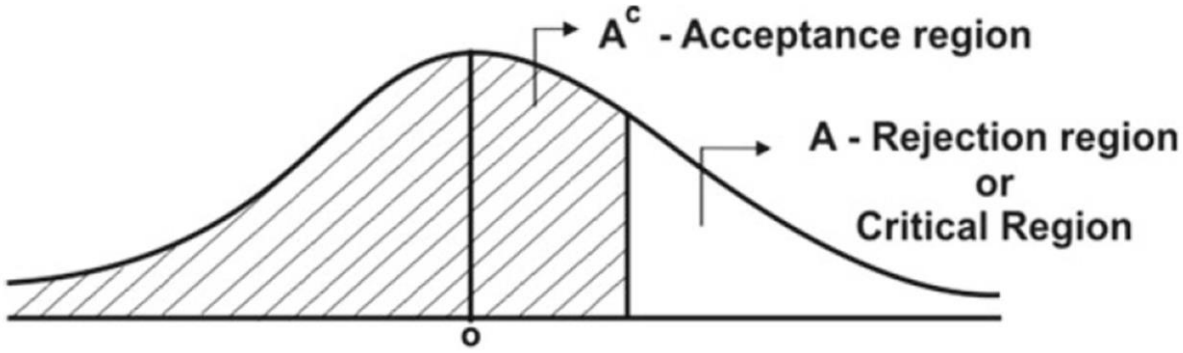


Fig.1. Acceptance and Rejection (Critical) region

In notation, we can write, $A = \{x \epsilon \Re_n, H_0 \text{ is rejected } if\ x \epsilon A\}$; see Fig.1.

Also, before proceeding we must remember the following two possibilities

(i) We may reject the hypothesis when we ought to accept it: i.e., when it is true

(ii) We may accept the hypothesis when we ought to reject it: i.e., when it is false

	Decision	
Test	Reject H_0	Accept H_0
$H_0\ True$	Type -1 error	Correct
$H_0 False$	correct	Type-2 error

Let $\alpha =$ Probability of Type-1 error and

$\beta =$ Probability of Type-2 error.

Choose the critical region so as to minimize both types of errors simultaneously, but this is, in general, not possible for a sample of fixed size. In fact, decreasing one type of error may very likely increase the other type. Thus, by deciding to always accept the hypothesis, we can reduce type-1 error to zero, but in that case βwould have its largest value, i.e., 1. In practice, we keep type-1 error fixed at a specified value and then, out of these critical regions all of which give this type-1 error, we choose that region which minimizes the type-2 error. The type-2 error, which is the same for all these regions, is sometimes called the size of the critical regions.

Definition (Test Function): The test function ϕ is defined as $\phi: \Re \to [0,1]$

Examples of test functions are

(i) $\phi(x) = 1 \forall\ x \epsilon \Re_n$

(ii) $\phi(x) = 0\ \forall\ x \epsilon\ \Re_n$

(iii) $\phi(x) = \delta, 0 \leq \delta \leq 1, \forall\ x \epsilon\ \Re_n$

Definition (Level of Significance & Size of Test): The test function ϕ is said to be a test of hypothesis $H_0: \theta\ \epsilon\ \Theta_0$ against the alternative $H_1: \theta\ \epsilon\ \Theta_1$ with the error probability α (it is also called level of significance) if

$$E_\theta \phi(x) \leq \alpha\ \forall\ \theta\ \epsilon \Theta_0 \qquad (1)$$

Further, our objective will be to seek a test ϕ for a given $\alpha, 0 \leq \alpha \leq 1$, let $\beta_\phi(\theta) = E_\theta \phi(x)$, such that

$$\sup_{\theta \epsilon \Theta_0} \beta_\phi(\theta) \leq \alpha \qquad (2)$$

LHS of (2) is also known as the size of the test ϕ. From (1) and (2), we can conclude that it restricts attention to those tests whose size does not exceed a given level of significance α.

In the hypothesis testing problems involving discrete distribution, it is usually not possible to choose a critical region consisting of realizable values of the statistics of size exactly α, where α is some prescribed value. In simple hypothesis, we can write

α=Probability of type 1 error,

$= P$[Reject $H_0 | H_0$ is true], and

β=Probability of type 2 error,

$=P$[Accept $H_0 | H_0$ is false]

For a non-randomized test with rejection region A, ϕ for a region A is just an indicator function. That is,

$$\phi = \begin{cases} 1; & x \epsilon A \\ 0; & x \epsilon A^c \end{cases}$$

We will extend this to allow or some different action (other than reject and accept) if the outcome X is on the boundary of the critical region. The other action effectively is performing an auxiliary experiment such as tossing a coin with P[heads] $= p$; if head results, reject H_0: if tail results,H_0 is accepted. The value of p is chose to make P[Reject $H_0 | H_0$ is true], the desired value.

More formally, for a test with critical region A and a value of $X = x_0$ on the boundary, we may define

$$\phi(x) = \begin{cases} 1\,; & x < x_0 \\ \gamma\,; & x = x_0 \\ 0;\ otherwise \end{cases}$$

where $0 < \gamma < 1$.

Such a test is known as "Randomized Test".

Example: Suppose X has a Poission distribution with mean λ. A sample of size $n = 10$ is used to test $H_0: \lambda = 0.1$ against $H_1: \lambda > 0.1$.

Note that $Y = \sum_{i=1}^{10} X_i$ has a Poission distribution with mean 10λ.

The test is to reject H_0 for large values of Y. suppose we wish to have a significance level of $\alpha = 0.05$.

Now, $P[Y \geq 3] = 0.08$ and $P[Y \geq 4] = 0.019$. The desired level of significance can be achieved by the test

$$\phi(Y) = \begin{cases} 1;\ Y \geq 4 \\ \gamma;\ Y = 3 \\ 0;\ Y < 3 \end{cases}$$

$P[\text{Reject } H_0 | H_0 \text{ is true}] = 0.05$

$$E\phi(y) = 1 \times P[Y \geq 4] + \gamma \times P[Y = 3] + 0$$

$$= 0.019 + \gamma[P(Y \geq 3) - P(Y \geq 4)]$$

$$= 0.019 + \gamma[0.08 - 0.019]$$

$$= 0.019 + \gamma(0.016)$$

Therefore,

$$= 0.019 + \gamma(0.016) = 0.05$$

$$\Rightarrow \gamma = \frac{31}{61}$$

Definition (Power Function): Let ϕ be any tests function for $H_0: \theta \in \Theta_0$ and $H_1: \theta \in \Theta_1$. For every $\theta \epsilon \Theta_1$, define

$$\beta_\phi(\theta) = E_\theta \phi(X); \quad \theta \in \Theta_1 \qquad (3)$$

As a function of $\theta, \beta_\phi(\theta)$ is called the power function of the test ϕ, for $\theta \in \Theta_1$. Further in simple hypothesis if β if type II error then power of the test is $1-\beta$. Moreover, in composite hypothesis $\beta_\phi(\theta)$ for $\theta \epsilon \Theta_1$ is a power function.

Problem In Testing of Hypothesis: Let $X_1, X_2, \ldots, X_n$ be iid rvs from $f(x|\theta), \theta \in \Theta$. assume $0 < \alpha < 1$ be given. Given a sample point X, find a test $\phi(x)$ such that

$$\sup_{\theta \epsilon \Theta_0} \beta_\phi(\theta) \leq \alpha$$

and $\beta_\phi(\theta)$ for $\theta \epsilon \Theta_1$ is maximum.

Here, we seek a critical region A such that its power defined in (3) is as large as possible. Then, in addition to having controlled to probability of type-1 error defined in (1) or (2), we have to minimize the probability of type-2 error.

Example: Let the rv X is distributed as $U(0, \theta)$. We are testing the hypothesis $H_0: \theta = 1$ against $H_1: \theta = 2$ based on a single observation. Calculate the probabilities of type-1 and type-2 errors based on the following critical regions. Also obtain the power of the test.

(i) $A_1 = \{x | 0.9 < x\}$

(ii) $A_2 = \{x | 1 < x < 1.5\}$

(i) $\mathrm{P[Type-1\ error]} = \mathrm{P[Reject\ H_0 | H_0 is\ true]}$

$$= \mathrm{P[X > 0.9 | \theta = 1]}$$

$$= \int_{0.9}^{1} dx$$

$$= 0.10$$

$\mathrm{P[Type-2\ error]} = \mathrm{P[Accept\ H_0 | H_1\ is\ true]}$

$$= \mathrm{P[0 \leq X < 0.9 | \theta = 2]}$$

$$= \int_{0}^{0.9} dx$$

$$= 0.45$$

$\mathrm{Power} = \mathrm{P[Reject\ H_0 | H_1\ is\ true]}$

$= \text{P}[X > 0.9|\theta = 2]$

$= P[0.9 < X < 2|\theta = 2]$

$= \int_{0.9}^{2} \frac{dx}{2}$

$= 0.55$

(ii) $A_2 = \{x: 1 \leq x \leq 1.5\}$

$\text{P}[\text{Type} - 1 \text{ error}] = \text{P}[\text{Reject } \text{H}_0|\text{H}_0 \text{is true}]$

$= P\{1 \leq x \leq 1.5|\theta = 1]$

$= \int_{-1}^{1} dx = 0$

$\text{P}[\text{Type} - 2 \text{ error}] = \text{P}[\text{Accept } \text{H}_0|\text{H}_1 \text{ is true}]$

$= P[x\epsilon A_2^c|\theta = 2]$

$= 1 - P[x\epsilon A_2|\theta = 2]$

$= 1 - \int_{1}^{1.5} \frac{dx}{2}$

$= 0.75$

$\text{Power} = \text{P}[\text{Reject } \text{H}_0|\text{H}_1 \text{ is true}]$

$= P[x\epsilon A_2|\theta = 2]$

$= \int_{1}^{1.5} \frac{dx}{2}$

$= 0.25$

4.4 Most Powerful Test and Best Critical Region

A Critical region whose power is no smaller than that of any other region of the same size for testing a hypothesis H_0 against the alternative H_1 is called a best critical

region (BCR) and a test based on BCR is called a most powerful test (MP) test. We will illustrate this BCR in notation.

Let A denote a subset of the sample space S. Then A is called BCR of size α for testing H_0 against H_1 if for every subset C of the sample space S for which $P[(X_1, X_2, \ldots\ldots, X_n) \epsilon C] \leq \alpha$.

Let $X = (X_1, X_2, \ldots\ldots, X_n)$

(i) $P[X \epsilon A | H_0 \; is \; true] \leq \alpha$ (3)

(ii) $P[X \epsilon A | H_1 \; is \; true] \geq P[X \epsilon C | H_1 \; is \; true]$

A test based on A is called the MP test.

Definition (Most Powerful Test): Let ϕ_α be the class of all test whose size is α. A test $\phi_0 \; \epsilon \; \phi_\alpha$ is said to be most powerful test (MP) test against an alternative $\theta \; \epsilon \; \Theta_1$ if

$$\beta_{\phi_0}(\theta) \geq \beta_\phi(\theta), \;\; \forall \; \phi \; \epsilon \phi_\alpha \qquad (4)$$

Note: If Θ_1 contains only one point, this definition suffices. If on the other hand Θ_1 contains at least two points as will usually be the ease, we will have a MP test corresponding to each Θ_1.

Definition (Uniformly Most Powerful Test): A test $\phi_0 \epsilon \phi_\alpha$ for testing $H_0 : \theta \; \epsilon \; \Theta_0$ against $H_1 : \theta \; \epsilon \; \Theta_1$ of size α is said to be uniformly most powerful test (UMP) test if

$$\beta_{\phi_0}(\theta) \geq \beta_\phi(\theta), \;\; \forall \; \phi \; \epsilon \phi_\alpha \;\; \text{uniformly in } \theta \epsilon \Theta_1 \qquad (5)$$

There will be many critical regions C with size α but suppose one of the critical regions, say A is such that its power is greater than or equal to the power of any such critical region C.

Neyman Pearson Lemma: Let $(X_1, X_2, \ldots, X_n)$ be a random sample from a distribution function $F_\theta(.)$,$\theta \varepsilon \Omega = \{\theta_0, \theta_1\}$ whose pdf or pmf is $f_\theta(.)$. We are interested in testing the hypothesis $H_0 : \theta = \theta_0$ against the alternative $H_1 : \theta = \theta_1$ at level $\alpha \; (0 < \alpha < 1)$. Let ϕ be the test defined as follows

$$\phi(x_1, x_2, \ldots, x_n) = 1 \quad if \; f(x_1, x_2, \ldots, x_n; \theta_1) > K \, f(x_1, x_2, \ldots, x_n; \theta_0)$$

$$= \gamma \quad if \; f(x_1, x_2, \ldots, x_n; \theta_1) = K \, f(x_1, x_2, \ldots, x_n; \theta_0) \qquad (6)$$

$= 0 \qquad otherwise$

where the constants $\gamma\ (0 \leq \gamma \leq 1)$ and $K\ (> 0)$ are determined so that

$E_{\theta_0}\phi(X_{1,}X_{2,}\dots,X_n) = \alpha.$ (7)

Then, for testing H_0 against H_1 at levelα, the test defined by (6) and (7) is MP within the class of all tests whose level is $\leq \alpha$.

Proof: For convenient writing, we set

$z = (x_1, x_2, \dots, x_n),\ dz = dx_1 dx_2 \dots dx_n,\ Z = (X_{1,}X_{2,}\dots,X_n)$, and

$f(z;\ \theta) = f(x_1;\theta)f(x_2;\theta)\dots f(x_n;\theta), f(Z;\theta) = f(X_{1,}\theta)f(X_2;\theta)\dots f(X_n;\theta)$, respectively

Let B be the set of points z in $\boldsymbol{R^n}$ such that $f_0(z) > 0$, where $f_0(z) = f(z;\ \theta_0)$ and $f_1(z) = f(z;\ \theta_1)$ and let B^c be the set of points z in R^n such that $f_0(z) \leq 0$. Let $D = Z^{-1}(B)\ and\ D^c = Z^{-1}(B^c)$,then

$P_{\theta_0}(D^c) = P_{\theta_0}(Z\ \varepsilon\ B^c) = \int_{B^c} f_0(z)dz = 0$,

and therefore, in calculating P_{θ_0}- probabilities we may redefine and modify random variables on the set D^c. Thus, we have, in particular

$$E_{\theta_0}[\phi(Z)] = P_{\theta_0}[f_1(Z) > Kf_0(Z)] + \gamma P_{\theta_0}[f_1(Z) = Kf_0(Z)]$$

$$= P_{\theta_0}[\{f_1(Z) > Kf_0(Z)\} \cap D] + \gamma P_{\theta_0}[\{f_1(Z) = Kf_0(Z)\} \cap D]$$

$$= P_{\theta_0}\left[\left\{\frac{f_1(Z)}{f_0(Z)} > K\right\} \cap D\right] + \gamma P_{\theta_0}\left[\left\{\frac{f_1(Z)}{f_0(Z)} = K\right\} \cap D\right]$$

$$= P_{\theta_0}[\{Y > K\} \cap D] + \gamma P_{\theta_0}[\{Y = K\} \cap D] \qquad (8)$$

$$= P_{\theta_0}[\{Y > K\}] + \gamma P_{\theta_0}[\{Y = K\}]$$

where $Y = \frac{f_1(Z)}{f_0(Z)}$ on D and let Y be arbitrary on D^c. Now let

$a(K) = P_{\theta_0}[\{Y > K\}]$, so that

$G(K) = 1 - a(K) = P_{\theta_0}[\{Y \leq K\}]$

is the distribution function of the random variable Y . Since G is a d.f., we have $G(-\infty) = 0, G(\infty) = 1$, G is non-decreasing and continuous from the right. These properties of G imply that the function $a(.)$ is such that $a(-\infty) = 1, a(\infty) = 0$, $a(.)$ is non-increasing and continuous from the right. Furthermore,

$P_{\theta_0}[\{Y = K\}] = G(K) - G(K - 0)$

$= [1 - a(K)] - [1 - a(K - 0)]$

$= a(K - 0) - a(K)$,

and $a(K) = 1$ for $K < 0$, since $P_{\theta_0}[\{Y \geq 0\}] = 1$.

Now for any α $(0 < \alpha < 1)$ there exists $K_0 (\geq 0)$ such that

$a(K_0) \leq \alpha \leq a(K_0 - 0)$.

At this point there are two cases to consider.

First $a(K_0) = a(K_0 - 0)$, that is K_0 is continuity point of the function $a(.)$. Then $\alpha = a(K_0)$ and if in (6) K is replaced by K_0 and $\gamma = 0$ the resulting test is of level α. In fact, in this case (8) becomes

$E_{\theta_0}[\phi(Z)] = P_{\theta_0}[\{Y > K_0\}] = a(K_0) = \alpha$ as was to be seen.

Next, we assume that K_0 is a discontinuity point of $a(.)$. In this case, take again $K = K_0$ in (6) and also set

$$\gamma = \frac{\alpha - a(K_0)}{a(K_0 - 0) - a(K_0)} \text{ (so that } 0 \leq \gamma \leq 1) .$$

Again we assert that the resulting test is of level α. In the present case, (8) becomes as follows

$E_{\theta_0}[\phi(Z)] = P_{\theta_0}[\{Y > K_0\}] + \gamma P_{\theta_0}[\{Y = K_0\}]$

$$= a(K_0) + \frac{\alpha - a(K_0)}{a(K_0 - 0) - a(K_0)} [a(K_0 - 0) - a(K_0)] = \alpha$$

$\gamma = \frac{\alpha - a(K_0)}{a(K_0 - 0) - a(K_0)}$ is to be interpreted as 0 whenever it is of the form $\frac{0}{0}$. The test (6) is of level. That is, (7) is satisfied.

Now it remains for us to show that the test so defined is MP, as described in the Lemma. To see this, let ϕ^* be any test of level $\leq \alpha$ and set

$B^+ = \{z \varepsilon R^n : \phi(z) - \phi^*(z) > 0\} = (\phi - \phi^* > 0)$

$B^- = \{z \varepsilon R^n : \phi(z) - \phi^*(z) < 0\} = (\phi - \phi^* < 0)$

Then $B^+ \cap B^- = \emptyset$ and, clearly,

$B^+ = (\phi > \phi^*) \subseteq (\phi = 1) \cup (\phi = \gamma) = (f_1 > K f_0)$

$B^- = (\phi < \phi^*) \subseteq (\phi = 0) \cup (\phi = \gamma) = (f_1 \leq K f_0)$ (9)

Therefore

$\int_{R^n} [\phi(z) - \phi^*(z)] [f_1(z) - K f_0(z)] dz$

$= \int_{B^+} [\phi(z) - \phi^*(z)][f_1(z) - K f_0(z)] dz + \int_{B^-} [\phi(z) - \phi^*(z)][f_1(z) - K f_0(z)] dz \geq 0$ by [10.5]

That is

$\int_{R^n} [\phi(z) - \phi^*(z)] [f_1(z) - K f_0(z)] dz \geq 0$ which is equivalent to

$\int_{R^n} [\phi(z) - \phi^*(z)] f_1(z) dz \geq \int_{R^n} [\phi(z) - \phi^*(z)] f_0(z) dz$ (10)

But

$$\int_{R^n} [\phi(z) - \phi^*(z)] f_0(z) dz = \int_{R^n} \phi(z) f_0(z) dz - \int_{R^n} \phi^*(z) f_0(z) dz$$

$$= E_{\theta_0}[\phi(Z)] - E_{\theta_0}[\phi^*(Z)] \geq 0$$

$$= \alpha - E_{\theta_0}[\phi^*(Z)] \geq 0 \quad (11)$$

and similarly,

$$\int_{R^n}[\phi(z)-\phi^*(z)]\,f_1(z)dz = \int_{R^n}\phi(z)\,f_1(z)dz - \int_{R^n}\phi^*(z)\,f_1(z)dz$$

$$= E_{\theta_1}[\phi(Z)] - E_{\theta_1}[\phi^*(Z)]$$

$$= \beta_\phi(\theta_1) - \beta_{\phi^*}(\theta_1) \qquad (12)$$

Relations (10), (11) and (12) yield $\beta_\phi(\theta_1) - \beta_{\phi^*}(\theta_1) \geq 0$, or

$\beta_\phi(\theta_1) \geq \beta_{\phi^*}(\theta_1)$.

This completes the proof of the NP Lemma. NP Lemma also guarantees that the power $\beta_\phi(\theta_1)$ is at least α.

Corollary: Let ϕ be defined by (6) and (7). Then $\beta_\phi(\theta_1) \geq \alpha$.

Proof: Let us consider a test $\phi^*(z) = \alpha$ is of the level α,

and since ϕ is MP test, we have

$\beta_\phi(\theta_1) \geq \beta_{\phi^*}(\theta_1) = \alpha$.

Remark: The NP Lemma shows that there is always a test of the structure (6) and (7) which is MP. The converse is also true, namely, if ϕ is a MP level α test, then ϕ necessarily has the form (6) unless there is a test of size $< \alpha$ and power 1.

Example: Let (X_1, X_2, X_3) be a random sample from $b(1,\theta)$. Find MP test of size $\frac{1}{9}$ for testing $H_0: \theta = \frac{1}{3}$ against $H_1: \theta = \frac{2}{3}$. Also, find the power of the test.

We know that if ϕ_1 and ϕ_2 are two test functions then $\phi = \alpha\,\phi_1 + (1-\alpha)\phi_2, 0 \leq \alpha \leq 1$ is also a test function. We use this fact in the solution.

Define

$$\lambda(z;\theta_0,\theta_1) = \frac{f(x_1, x_2, x_3;\theta_1)}{f(x_1, x_2, x_3;\theta_0)}$$

$$= \frac{\theta_1^{\sum_{i=1}^3 x_i}(1-\theta_1)^{3-\sum_{i=1}^3 x_i}}{\theta_0^{\sum_{i=1}^3 x_i}(1-\theta_0)^{3-\sum_{i=1}^3 x_i}}$$

$$= \left(\frac{4}{3}\right)^{\sum_{i=1}^3 x_i} \frac{1}{8} \text{ as } \theta_0 = \frac{1}{3} \text{ and } \theta_1 = \frac{2}{3}.$$

Thus, $\lambda(z;\theta_0,\theta_1) \uparrow \sum_{i=1}^3 x_i$.

By N.P. Lemma

$\lambda(z;\theta_0,\theta_1) > K_1 \Leftrightarrow \sum_{i=1}^3 x_i > K_0$.

Now we calculate the values of $\alpha, \beta \; and \; 1-\beta$ for different values of K_0 and then find the value of K_0 corresponding to $\alpha = \frac{1}{9}$ which will correspond to a randomized test as below:

K_0	α	β	$1-\beta$
< 0	1	0	1
=0	$\frac{19}{27}$	$\frac{1}{27}$	$\frac{26}{27}$
=1	$\frac{7}{27}$	$\frac{7}{27}$	$\frac{20}{27}$
=2	$\frac{1}{27}$	$\frac{19}{27}$	$\frac{8}{27}$
=3	0	1	0

Let $T = \sum_{i=1}^3 X_i \sim b(3,\theta)$

$\Rightarrow$ under $H_0, T \sim b\left(3,\frac{1}{3}\right)$ and under, H_1 $T \sim b\left(3,\frac{2}{3}\right)$

$\Rightarrow \alpha = P_{\frac{1}{3}}[T > K_0] = \sum_{t=K_0+1}^3 \binom{n}{t} \left(\frac{1}{3}\right)^t \left(\frac{2}{3}\right)^{3-t}$

$1-\beta = P_{\frac{2}{3}}[T > K_0] = \sum_{t=K_0+1}^3 \binom{n}{t} \left(\frac{2}{3}\right)^t \left(\frac{1}{3}\right)^{3-t}$

Define

$\phi_1(T) = \phi_1(X_1, X_2, X_3) = 1 \;\; if \; T > 1$

$= 0 \;\; if \; T \leq 1$

And

$\phi_2(T) = \phi_2(X_1, X_2, X_3) = 1 \;\; if \; T > 2$

$$= 0 \ \ if\ T \leq 2$$

Now, let

$\phi(\mathrm{T}) = \gamma\, \phi_1(T) + (1-\gamma)\, \phi_2(T)$,

then

$E_{\frac{1}{3}}[\phi(\mathrm{T})] = \ \gamma P_{\frac{1}{3}}[T > 1] + (1-\gamma) P_{\frac{1}{3}}[T > 2] = \ \alpha = \frac{1}{9}$

$\Rightarrow \gamma = \frac{1}{3}$, and

$$\phi(\mathrm{T}) = 1 \ \ \text{if T} > 2$$

$$= \tfrac{1}{3} \quad if\ T = 2$$

$$= 0 \quad if\ T < 2$$

$E_{\frac{2}{3}}[\phi(\mathrm{T})] = 1 - \beta = \ Power = \ P_{\frac{2}{3}}[T > 2] + \frac{1}{3} P_{\frac{2}{3}}[T = 2]$

Thus, Power = $1 - \beta = \ \frac{8}{27} + \frac{1}{3}.3.\frac{4}{27} = \ \frac{12}{27} = \ \frac{4}{9}$

Example: Let (X_1, X_2, X_3) be a random sample from $P(\lambda)$. Find MP test of size $.02$ for testing $H_0: \lambda = 1$ against $H_1: \lambda = 2$. Also find the power of the test.

Let $= \sum_{i=1}^{3} X_i$, $T \sim P(3\lambda)$. We know from Poisson distribution table that

$$P_{\lambda=1}[T > 5] = \ .08$$

$$P_{\lambda=1}[T > 6] = \ .03$$

$$P_{\lambda=1}[T > 7] = \ .01$$

$$P_{\lambda=2}[T > 7] = 0.256$$

$P_{\lambda=2}[T \leq 6] = \ 0.6063.$

Define

$\mathrm{R}(z; \lambda_0, \lambda_1) = \frac{f(x_1, x_2, x_3; \lambda_1)}{f(x_1, x_2, x_3; \lambda_0)}$

$$= \frac{\frac{e^{-3\lambda_1}\lambda_1^{\sum_{i=1}^{3} x_i}}{\prod_{i=1}^{3} x_i!}}{\frac{e^{-3\lambda_0}\lambda_0^{\sum_{i=1}^{3} x_i}}{\prod_{i=1}^{3} x_i!}}$$

$= 2^{\sum_{i=1}^{3} x_i} e^{-3}$ as $\lambda_0 = 1 \; and \; \lambda_1 = 2.$

Thus, $\mathrm{R}(z; \lambda_0, \lambda_1) \uparrow \sum_{i=1}^{3} x_i$.

By N.P. Lemma

$\mathrm{R}(z; \lambda_0, \lambda_1) > K_1$

$\Leftrightarrow \sum_{i=1}^{3} X_i > K_0 \; or \; T > K_0$.

Let

$$\phi_1(T) = \phi_1(X_1, X_2, X_3) = 1 \quad if \; T > 6$$

$$= 0 \quad if \; T \leq 6$$

$$\phi_2(T) = \phi_2(X_1, X_2, X_3) = 1 \quad if \; T > 7$$

$$= 0 \quad if \; T \leq 7$$

$\phi(\mathrm{T}) = \gamma \, \phi_1(T) + (1 - \gamma) \, \phi_2(T)$,

$$\phi(\mathrm{T}) = 1 \quad \text{if } \mathrm{T} > 7$$

$$= \gamma \quad \text{if } \mathrm{T} = 7$$

$$= 0 \quad if \; T < 7$$

Then

$E_{\lambda=1}[\phi(\mathrm{T})] = \gamma P_{\lambda=1}[T > 6] + (1 - \gamma) P_{\lambda=1}[T > 7] = \alpha = .02$

$\Rightarrow \gamma = \frac{1}{4}$, and

$$\phi(\mathrm{T}) = 1 \quad \text{if } \mathrm{T} > 7$$

$= \frac{1}{4} \quad if\ T = 7$

$= 0 \quad if\ T < 7$

$E_{\lambda=2}[\phi(\mathrm{T})] = 1 - \beta = \ Power = \ P_{\lambda=2}[T > 7] + \frac{1}{4}P_{\lambda=2}[T = 7]$

Thus, Power = $1 - \beta = \ 0.256 + \frac{1}{4}.[0.744 - 0.6063] = \ 0.2899$

Example: Let $(X_{1,}X_{2,} \ldots, X_{16})$ be a random sample from$N(\theta, 1)$. Find MP test of size $.05$ for testing $H_0: \theta = 1$ against $H_1: \theta = 2$. Also find the power of the test.

Let $T = \sum_{i=1}^{16} X_i\ , T \sim N(16\theta, 16)\ or\ \bar{X} \sim N\left(\theta, \frac{1}{16}\right)$.

Let = $(x_{1,}x_{2,} \ldots x_{16})$.

Define

$\mathrm{R}(z; 1{,}2) = \frac{f(x_{1,}x_{2,}\ldots x_{16};2)}{f(x_{1,}x_{2,}\ldots,x_{16};1)}$

$= \frac{\frac{1}{(2\pi)^8}\, exp - \sum_{i=1}^{16}(x_i - 2)^2}{\frac{1}{(2\pi)^8}\, exp - \sum_{i=1}^{16}(x_i - 1)^2}$

$= exp(-24)\ exp(32\ \bar{x})$

Clearly, $\mathrm{R}(z; 1{,}2)$ ↥ $\bar{x}$

$\Rightarrow \mathrm{R}(z; 1{,}2)\ > K_1$

$\Rightarrow \bar{x} > K_0$.

Since the distribution is continuous, therefore, the test defined by N.P. Lemma is

$\phi(\overline{\mathrm{X}}) = 1$ if $\overline{\mathrm{X}} > K_0$

$= 0 \quad if\ T \leq K_0$

where K_0 is obtained by the size condition $E_{\theta=1}[\phi(\overline{X})] = .05$.

$$E_{\theta=1}[\phi(\overline{X})] = P_{\theta=1}[\overline{X} > K_0] = 0.05$$

$$P[(\overline{X} - 1)4 > 4(K_0 - 1)] = 0.05$$

$$P[Z > 4(K_0 - 1)] = 0.05, \text{where } Z \sim N(0,1)$$

From Standard Normal Table

$4(K_0 - 1) = 1.65$

$\Rightarrow K_0 = 1.4125$

Power $= 1 - \beta$

$= E_{\theta=2}[\phi(\overline{X})] = P_{\theta=2}[\overline{X} > K_0] = P[4(\overline{X} - 2) > 4(1.4125 - 2)]$

$= P[Z > -2.35] = 1 - P[Z \leq 2.35] = 0.9906$

So far, an MP rest was constructed for the problem of testing a simple hypothesis against a simple alternative. However, in most problems of practical interest at least one of the hypotheses H_0 *or* H_1 is composite. In cases like this it so happens that for certain families of distributions and certain H_0 *and* H_1, UMP tests may not exist. We will restricted to our self only the cases of one parameter exponential families of distributions for which UMP tests may exist.

Theorem: Let $(X_1, X_2, \ldots, X_n)$ be a random sample from a distribution function $F_\theta(.)$,$\theta \varepsilon \Omega \subseteq R$ whose pdf or pmf is $f_\theta(.)$ belongs to one parameter exponential family

$$f(x;\theta) = C(\theta)e^{Q(\theta)T(x)}\, h(x) \qquad (13)$$

where $C(\theta) > 0$ for all $\theta \varepsilon \Omega \subseteq R$ and set of positivity of h is independent of θ. Suppose $Q(\theta)$ is monotone function of θ . Let $\theta_0 \varepsilon \Omega$ and set

$\omega = \{\theta \varepsilon \Omega; \theta \leq \theta_0\}$. Then for testing (composite) $H_0: \theta\, \varepsilon\, \omega$ against the (composite) alternative $H_1: \theta\, \varepsilon\, \omega^c$ of at level of significance α, there exists a test ϕ which is UMP with in the class of all tests of level $\leq \alpha$. In the case that the $Q(\theta)$ is increasing, the test is given by

$\phi(x_1, x_2, \ldots, x_n) = \phi(z) = 1$ if $V(Z) > K_0$

$$= \gamma \ if \ \ \mathrm{V(Z)} = \mathrm{K_0} \qquad (14)$$

$$= \ 0 \quad if \ \mathrm{V(Z)} < \mathrm{K_0}$$

where $\mathrm{V(Z)} = \sum_{i=1}^{n} \mathrm{T(X_i)}$, and $\mathrm{K_0}$ and γ are determined by

$$E_{\theta_0}[\phi(\mathrm{Z})] = P_{\theta_0}[\mathrm{V(Z)} > \mathrm{K_0}] + \gamma P_{\theta_0}[\mathrm{V(Z)} = \mathrm{K_0}] = \alpha \qquad (15)$$

If $Q(\theta)$ is decreasing, the test is taken from (14) and (15) with the inequality signs reversed.

Proof: Consider first that $Q(\theta)$ is increasing function of θ. Let $\theta' \varepsilon \, \omega^c$ be arbitrary and consider the problem of testing the hypothesis $H_0: \theta = \theta_0$ against the (simple) alternative $H_1: \theta = \theta'$ at level α. Then, by N. P. Lemma, the M.P. test ϕ^* is given by

$$\phi^*(z) = 1 \ if \ g(z;\theta') > K^* g(z;\theta_0)$$

$$= \gamma^* \ if \ g(z;\theta') = K^* g(z;\theta_0) \qquad (16)$$

$$= 0 \ if \ g(z;\theta') < K^* g(z;\theta_0)$$

where $g(z;\theta) = f(x_1, x_2, \dots, x_n;\theta) = \prod_{i=1}^{n} f(x_i;\theta)$, and $K^* \, and \, \gamma^*$ are defined by

$$E_{\theta_0}[\phi^*(\mathrm{Z})] = \alpha \qquad (17)$$

Let

$$\frac{g(z;\theta')}{g(z;\theta_0)} = \frac{\left(C(\theta')\right)^n exp[Q(\theta') \sum_{i=1}^{n} T(x_i)] \ \prod_{i=1}^{n} h(x_i)}{\left(C(\theta_0)\right)^n exp[Q(\theta_0) \sum_{i=1}^{n} T(x_i)] \ \prod_{i=1}^{n} h(x_i)}$$

$$= \frac{\left(c(\theta')\right)^n}{\left(c(\theta_0)\right)^n} \ exp\left[\sum_{i=1}^{n} T(x_i) \left(Q(\theta') - Q(\theta_0)\right)\right]$$

Thus, as

$$Q(\theta) \uparrow \theta \Rightarrow \left(Q(\theta') - Q(\theta_0)\right) > 0$$

$$\Rightarrow \frac{g(z;\theta')}{g(z;\theta_0)} > \ K^* \text{ iff} \sum_{i=1}^{n} T(x_i) > \ K_0$$

$$E_{\theta_0}[\phi^*(\mathrm{Z})] = P_{\theta_0}\left[\frac{g(z;\theta')}{g(z;\theta_0)} > K^*\right] + \gamma^* P_{\theta_0}\left[\frac{g(z;\theta')}{g(z;\theta_0)} = K^*\right] = \alpha$$

$$= P_{\theta_0}[\mathrm{V(Z)} > K_0] + \gamma^* P_{\theta_0}[\mathrm{V(Z)} = K_0] = \alpha$$

where $\mathrm{V(Z)} = \sum_{i=1}^{n} T(X_i)$. Therefore, (16) and (17) becomes as

$$\phi^*(z) = 1 \ if \ \mathrm{V(Z)} > K_0$$

$$= \gamma^* \ if \ \mathrm{V(Z)} = K_0 \qquad (18)$$

$$= 0 \ if \ \mathrm{V(Z)} < K_0$$

and $E_{\theta_0}[\phi^*(\mathrm{Z})] = P_{\theta_0}[\mathrm{V(Z)} > K_0] + \gamma^* P_{\theta_0}[\mathrm{V(Z)} = K_0] = \alpha \qquad (19)$

The value $\gamma^* = \gamma$ by means of (14) and (15). It follows from (18) and (19) that the test ϕ^* is independent of $\theta' \varepsilon\, \omega^c$. In other words, we have that $\gamma^* = \gamma$ and test given by (14) and (15) is UMP for testing $H_0: \theta = \theta_0$ against $H_1: \theta\, \varepsilon\, \omega^c$ at level α. Therefore, all that remains to show is that $E_{\theta'}[\phi(\mathrm{Z})] \leq \alpha$ for all $\theta' \, \varepsilon \, \omega$. That this is indeed so is seen as follows:

Let $E_{\theta'}[\phi(\mathrm{Z})] = \alpha(\theta')$ and consider the problem of testing the (simple) hypothesis $H_0': \theta = \theta'$ against the (simple) alternative $H_0: \theta = \theta_0$ at level $\alpha(\theta')$. Once more, by N.P. Lemma, the MP test ϕ' is given by

$$\phi'(z) = 1 \ if \ g(z;\theta_0) > K' g(z;\theta')$$

$$= \gamma' \ if \ g(z;\theta_0) = K' g(z;\theta') \qquad (20)$$

$$= 0 \ if \ g(z;\theta_0) < K' g(z;\theta')$$

or

$$\phi'(z) = 1 \ if \ V(z) > K_0'$$

$$= \gamma' \ if \ V(z) = K_0' \qquad (21)$$

$$= 0 \ if \ V(z) < K_0'$$

where $K_0' \ and \ \gamma'$ are determined by

$$E_{\theta'}[\phi'(\mathrm{Z})] = P_{\theta'}[\mathrm{V(Z)} > K_0'] + \gamma' P_{\theta'}[\mathrm{V(Z)} = K_0'] = \alpha(\theta') \qquad (22)$$

Replacing θ_0 by θ' in the expression on the left side of (15) and comparing the resulting expression with (22), one has $K_0' = K_0$ and $\gamma' = \gamma$. Therefore, the tests ϕ and ϕ' are identical. Furthermore, by the corollary of NP Lemma one has that $\alpha(\theta') \leq \alpha$, since α is the power of the test ϕ'. To complete the proof of the Theorem, let Φ be the class of all level α

tests for testing $H_0: \theta \leq \theta_0$ and let Φ_0 be the class of all level α tests for testing $H_0: \theta = \theta_0$. Then, clearly, $\Phi \subset \Phi_0$. The test ϕ defined by (14) and (15) belongs to Φ and maximizes the power among all tests in the class Φ_0. Hence it maximizes the power among all tests in the class Φ.

Definition: Let $(X_1, X_2, \ldots, X_n)$ be a random sample from a pdf or pmf $(.;\ \theta), \theta \varepsilon \Omega \subseteq R^r$. Then for testing the hypothesis $H_0: \theta\ \varepsilon\ \omega$ against the alternative $H_1: \theta\ \varepsilon\ \omega^c$, where $\omega \subset \Omega \subseteq R^r$ at level of significance α, a test ϕ based on $(X_1, X_2, \ldots, X_n)$ is said to be unbiased if $E_\theta[\phi(X_1, X_2, \ldots, X_n)] \leq \alpha$ for all $\theta\ \varepsilon\ \omega$ and $E_\theta[\phi(X_1, X_2, \ldots, X_n)] \geq \alpha$ for all $\theta\ \varepsilon\ \omega^c$.

Definition: In the notation of Definition 10.5, a test is said to be uniformly most powerful unbiased (UMPU) if it is UMP with in the class of all unbiased tests.

Remark: A UMP test is always UMPU. In fact, in the first place it is unbiased because it is at least as powerful as the test which is identically equal to α. Next, it is UMPU because it is UMP within a class including the class of unbiased tests.

Now, we will present a Theorem (without proof) by which certain class of distributions and certain hypotheses, UMPU tests do exist.

Theorem: Let $(X_1, X_2, \ldots, X_n)$ be a random sample from a pdf or pmf $f(.;\ \theta)$ given by

$$f(x;\theta) = C(\theta)e^{\theta T(x)}h(x),\ \theta\ \varepsilon\ \Omega \subseteq R \qquad (23)$$

Let $\omega = \{\theta_0\}$. Then for testing the hypothesis $H_0: \theta\ \varepsilon\ \omega$ against $H_1: \theta\ \varepsilon\ \omega^c$ at level of significance α, there exists UMPU test which is given by

$$\phi(\mathrm{z}) = 1 \text{ if } \mathrm{V(z)} < \mathrm{K_1} \text{ or } \mathrm{V(z)} > \mathrm{K_2}$$

$$= \gamma_i\ if\ \ \mathrm{V(z)} = \mathrm{K_i}\ (\mathrm{i} = 1, 2)(\mathrm{K_1} < \mathrm{K_2})$$

$$= 0\ \ if\ \ \mathrm{K_1} < V(\mathrm{z}) < \mathrm{K_2}$$

where with $z = (x_1, x_2, \ldots, x_n), \mathrm{V(z)} = \sum_{i=1}^{n} T(x_i)$ the constants $\mathrm{K_i}, \gamma_i, i == 1,2$ are given by

$$E_{\theta_0}[\phi(\mathrm{Z})] = \alpha, E_{\theta_0}[\mathrm{V(Z)}\phi(\mathrm{Z})] = \alpha E_{\theta_0}[V(\mathrm{Z})]$$

Remark: We would expect that the cases like Binomial, Poisson and Normal fall under Theorem 10.3, while they seemingly do not. However, a simple reparameterization of the families brings them under (22).

1. ***For Binomial* case,** $Q(\theta) = \log\left[\frac{\theta}{1-\theta}\right]$. Then setting $\log\left[\frac{\theta}{1-\theta}\right] = \tau$, the family is brought under the form (22). From this transformation we get $\theta = \frac{e^{\tau}}{1+e^{\tau}}$ and the hypothesis $H_0: \theta = \theta_0$ becomes equivalently $H_0': \tau = \tau_0$
2. ***For the Poisson case,*** $Q(\theta) = \log\theta$ and the transformation $\log\theta = \tau$ brings the family under the form (22). The transformation implies $\theta = e^{\tau}$ and the hypothesis $H_0: \theta = \theta_0$ becomes equivalently $H_0': \tau = \tau_0$
3. ***For the Normal case*** *with σ known and $\mu = \theta, Q(\theta) = \frac{\theta}{\sigma^2}$* and the factor $\frac{1}{\sigma^2}$ may be absorbed into $T(x)$.
4. ***For the Normal case*** *with μ known and $\sigma^2 = \theta, Q(\theta) = -\frac{1}{2\theta}$* and the transformation $-\frac{1}{2\theta} = \tau$ brings the family under the form (22) again, and the hypothesis $H_0: \theta = \theta_0$ becomes equivalently $H_0': \tau = \tau_0$.

The condition $E_{\theta_0}[\mathrm{V(Z)}\phi(\mathrm{Z})] = \alpha E_{\theta_0}[V(\mathrm{Z})]$ is the unbiasedness condition as $\beta_\phi(\theta) = E_\theta[\phi(\mathrm{Z})] = C_1(\theta)\int \phi(\mathrm{z})e^{\theta V(z)}dV$, $V(\mathrm{Z})$ is sufficient statistic. Differentiating w.r.t θ, we have $\beta_\phi'(\theta) = C_1'(\theta)\int \phi(\mathrm{z})e^{\theta V(z)}dV + C_1(\theta)\int \phi(\mathrm{z})V(z)\,e^{\theta V(z)}dV$

$$= \frac{C_1'(\theta)}{C_1(\theta)}C_1(\theta)\int \phi(\mathrm{z})e^{\theta V(z)}dV + C_1(\theta)\int \phi(\mathrm{z})V(z)\,e^{\theta V(z)}dV$$

$$= \frac{C_1'(\theta)}{C_1(\theta)}E_\theta[\phi(\mathrm{Z})] + E_\theta[\phi(\mathrm{Z})V(\mathrm{Z})]\,.$$

At $\theta = \theta_0$ $\beta_\phi(\theta)$ has minimum value and $\phi(\mathrm{Z}) \equiv \alpha \Rightarrow \beta_\phi'(\theta_0) = 0 \Rightarrow E_{\theta_0}[V(\mathrm{Z})] =$

$$-\frac{C_1'(\theta)}{C_1(\theta)} \Rightarrow E_{\theta_0}[\mathrm{V(Z)}\phi(\mathrm{Z})] = \alpha E_{\theta_0}[V(\mathrm{Z})].$$

Example: Let $(X_1, X_2, \dots, X_n)$ be a random sample from $N(\theta, 1)$. Find UMP test of size α for testing $H_0: \theta \leq \theta_0$ against $H_1: \theta > \theta_0$. Also show that the power function is an increasing function of θ .

We know that

$$f(x;\theta) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\theta)^2}{2}} = \frac{1}{\sqrt{2\pi}} e^{-\frac{\theta^2}{2}} e^{\theta x} e^{-\frac{x^2}{2}} = C(\theta) e^{Q(\theta)T(x)} h(x),$$

Here $C(\theta) = \frac{1}{\sqrt{2\pi}} e^{-\frac{\theta^2}{2}}, T(x) = \quad x, h(x) = e^{-\frac{x^2}{2}}, and\ Q(\theta) = \theta.\ Q(\theta)\uparrow\theta$ and

$$g(z;\theta) = \frac{1}{(2\pi)^{\frac{n}{2}}} e^{-\frac{n\theta^2}{2}} e^{\theta \sum_{i=1}^{n} x_i} \prod_{i=1}^{n} e^{-\frac{x_i^2}{2}}$$

$$\Rightarrow V(z) = \sum_{i=1}^{n} x_i,\ z = (x_1, x_2, \dots, x_n) .$$

Since the distribution $N(\theta, 1)$ is continuous, the test function is given by previous theorem is

$$\phi(z) = 1 \quad \text{if } V(z) > K_0'$$
$$= 0\ if\ V(z) \leq K_0'$$

or

$$\phi(z) = 1 \quad \text{if } \bar{x} > K_0$$
$$= 0\ if\ \bar{x} \leq K_0$$

where K_0 is obtained by the size condition

$$E_{\theta_0}[\phi(Z)] = P_{\theta_0}[\bar{X} > K_0] = \alpha$$

But

$$P_{\theta_0}[\bar{X} > K_0] = P\left[\sqrt{n}\,(\bar{X} - \theta_0) > \sqrt{n}\,(K_0 - \theta_0)\right] = P\left[Y > \sqrt{n}\,(K_0 - \theta_0)\right],$$

$Y \sim N(0, 1)$. Thus K_0 is obtained by the condition

$$\Phi\left(\sqrt{n}\,(K_0 - \theta_0)\right) = 1 - \alpha.$$

From standard normal table, corresponding to the value $1 - \alpha$ we can find the value $\sqrt{n}\,(K_0 - \theta_0)$ and hence the value of K_0. The power function

$$\beta(\theta) = P_\theta[\text{Rejecting } H_0]$$
$$= P_\theta[\bar{X} > K_0]$$
$$= P\left[\sqrt{n}\,(\bar{X} - \theta) > \sqrt{n}\,(K_0 - \theta)\right]$$
$$= P\left[Y > \sqrt{n}\,(K_0 - \theta)\right]$$
$$= 1 - \Phi\left[\sqrt{n}\,(K_0 - \theta)\right]$$

Clearly $\beta(\theta)\uparrow\theta$ under H_1 and $\beta(\theta)\downarrow\theta$ under H_0. Hence, by previous theorem, the test given above is UMP.

Example: Let $(X_1, X_2, \ldots, X_n)$ be a random sample from $N(\theta, 1)$. Find UMPU test of size α for testing $H_0: \theta = \theta_0$ against $H_1: \theta \neq \theta_0$. Also draw the curve of power function $\beta(\theta)$.

Since

$$f(x;\theta) = \frac{1}{\sqrt{2\pi}}\, exp[-(x-\theta)^2] = C(\theta)e^{\theta x}\, h(x)$$

where $C(\theta) = \frac{1}{\sqrt{2\pi}} e^{-\frac{\theta^2}{2}}, h(x) = e^{-\frac{x^2}{2}}, T(x) = x$, hence by previous theorem UMPU test is given by

$$\phi(z) = 1 \;\; if \;\; n\bar{x} \;<\; K_1 \; or \; n\bar{x} \;>\; K_2$$
$$= 0 \;\; if \; K_1 \;\leq\; n\bar{x} \;\leq\; K_2$$

where K_1 and K_2 are determined by the conditions

$$E_{\theta_0}[\phi(Z)] = \alpha, E_{\theta_0}[V(Z)\phi(Z)] = \alpha\, E_{\theta_0}[V(Z)]$$

Now ϕ can be expressed equivalently

$$\phi(z) = 1 \text{ if } \quad \sqrt{n}(\bar{x} - \theta_0) \;<\; \frac{K_1}{\sqrt{n}} - \sqrt{n}\,\theta_0 \; or$$
$$\sqrt{n}(\bar{x} - \theta_0) > \tfrac{K_2}{\sqrt{n}} - \sqrt{n}\,\theta_0$$
$$= 0 \text{ if } \tfrac{K_1}{\sqrt{n}} - \sqrt{n}\,\theta_0 \;\leq\; \sqrt{n}(\bar{x} - \theta_0) \;\leq\; \tfrac{K_2}{\sqrt{n}} - \sqrt{n}\,\theta_0$$

We know that under H_0 , $\sqrt{n}(\bar{X} - \theta_0)$ is $N(0,1)$. Since $N(0,1)$ is symmetric about zero, this implies that

$\left(\frac{K_1}{\sqrt{n}} - \sqrt{n}\,\theta_0\right) = K_1' = -K_2' = \; -\left(\frac{K_2}{\sqrt{n}} - \sqrt{n}\,\theta_0\right) = -K_0$, say $(K_0 > 0)$.

Thus

$$K_1 = \tfrac{-K_0}{\sqrt{n}} + \theta_0, K_2 = \tfrac{K_0}{\sqrt{n}} + \theta_0.$$

$$\phi(y) = 1 \;\; if \;\; |y| > K_0$$
$$= 0 \;\; if \; |y| \leq K_0$$

where $y = \sqrt{n}(\bar{x} - \theta_0)$ and K_0 is obtained by the condition

$$E_{\theta_0}[\phi(Y)] = \alpha \Rightarrow K_0 = z_{\frac{\alpha}{2}}, \int_{z_{\frac{\alpha}{2}}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy = \frac{\alpha}{2}$$

Unbiasedness condition becomes $\int_{-\frac{z_\alpha}{2}}^{\frac{z_\alpha}{2}} \frac{y}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy = 0.$

The power function

$$\beta(\theta) = 2\left(1 - \Phi\left[\mathrm{K}_0 + \sqrt{\mathrm{n}}\ (\theta_0 - \theta)\right]\right)\ if\ \theta \geq \theta_0$$

$$= 2\ \Phi\left[-\mathrm{K}_0 + \sqrt{\mathrm{n}}\ (\theta_0 - \theta)\right]\ if\ \theta \leq \theta_0$$

$\beta(\theta)\uparrow\theta$ as $\theta\to\infty\ (\theta > \theta_0)$, and $\beta(\theta)\uparrow$ as $\theta\to -\infty(\theta < \theta_0)$.

Example: Let $\left(X_{1,}X_{2,} \ldots, X_n\right)$ be a random sample from $N(0, \theta)$. Find UMPU test of size α for testing $H_0: \theta = \theta_0 > 0$ against $H_1: \theta \neq \theta_0$. Also draw the curve of power function $\beta(\theta)$.

Since

$$f(x;\theta) = \frac{1}{\sqrt{2\pi\theta}}\ exp\left[-\frac{x^2}{2\theta}\right] = C(\tau)e^{\tau T(x)}\ h(x)$$

where $\tau = \frac{-1}{2\theta}, C(\tau) = \sqrt{\frac{|\tau|}{2\pi}}, h(x) \equiv 1, T(x) = x^2$, hence by previous theorem UMPU test is given by

$$\phi(z) = 1\ \text{if}\ V(z) < K_1\ or\ V(z) > K_2$$

$$= 0\ \text{if}\ K_1 \leq V(z) \leq K_2$$

where $V(z) = \sum_{i=1}^{n} x_i^2$, and $K_1\ and\ K_2$ are obtained by the size condition

$$E_{\theta_0}[\phi(Z)] = \alpha$$

This is equivalent to

$$\int_{\frac{K_1}{\theta_0}}^{\frac{K_2}{\theta_0}} f_n(y)dy = (1 - \alpha)$$

where $f_n(y), y = \frac{\sum_{i=1}^{n} x_i^2}{\theta_0}$, is the pdf of $\mathcal{X}^2 -$ distribution with n degrees of freedom. The unbiasedness condition becomes

$$\int_{\frac{K_1}{\theta_0}}^{\frac{K_2}{\theta_0}} y f_n(y)dy = n(1 - \alpha)$$

$$\Rightarrow \int_{\frac{K_1}{\theta_0}}^{\frac{K_2}{\theta_0}} f_{n+2}(y)dy = (1 - \alpha), \text{as}$$

$$y f_n(y) = n\ f_{n+2}(y)$$

or $$\int_{\frac{K_1}{\theta_0}}^{\frac{K_2}{\theta_0}} f_n(y)dy = \frac{1}{n}\int_{\frac{K_1}{\theta_0}}^{\frac{K_2}{\theta_0}} y f_n(y)dy = (1-\alpha).$$

In practice we take equal tail test given by

$$\frac{K_1}{\theta_0} = \chi_n^2\left(\frac{\alpha}{2}\right), \frac{K_2}{\theta_0} = \chi_n^2\left(1-\frac{\alpha}{2}\right)$$

where $\int_0^{\chi_n^2\left(\frac{\alpha}{2}\right)} f_n(y)dy = \frac{\alpha}{2}$ and $\int_0^{\chi_n^2\left(1-\frac{\alpha}{2}\right)} f_n(y)dy = \left(1-\frac{\alpha}{2}\right)$.

The popular equal tail test is not UMPU; it is close to the UMPU test when *n* is large.

4.5 Self-Assessment Exercise

1. State and Prove Neyman Pearson Lemma for simple versus simple hypotheses.

2 (a) Let $X_{(n)}$ denote the largest order statistic in a random sample of size n from the rectangular distribution

$$f_\theta(x) = \begin{cases} 1/\theta & if\ \ 0 < x < \theta \\ 0 & elsewhere, \end{cases}$$

where $0 < \theta < \infty$. Suppose a test for $H_0: \theta = 1$ against $H_0: \theta \neq 1$ involves rejection of H_0 if $x_{(n)} < k_1$ or $x_{(n)} > k$ where $x_{(n)}$ is the observed value of $X_{(n)}$ and $0 < k_1 < k_2$. Find the power function of the test and hence determine the level of significance.

(b) Show that for testing $H_0: \theta = \theta_0$ in random sampling from the rectangular distribution

$$f_\theta(x) = \begin{cases} 1 & if\ \theta \leq x \leq \theta + 1 \\ 0 & otherwise, \end{cases}$$

there is a pair of UMP tests for one-sided alternatives and hence no UMP test for all alternatives.

3. A sample of size one is taken from the distribution with p.d.f.

$$f_\theta(x) = \frac{2}{\theta^2}(\theta - x)\ if\ \ 0 < x < \theta$$

$$= 0 \quad otherwise.$$

Find an MP test of $H_0: \theta = \theta_0$ against $H_1: \theta = \theta_1 (\theta_1 < \theta_0)$

4. Let $X_1, X_2, \ldots\ldots, X_n$ be a random sample of size n from an exponential distribution with p.d.f.

$$f_\theta(x) = \begin{cases} \theta \exp[-\theta x] \; if \; x > 0 \\ 0 \quad otherwise, \end{cases}$$

where $0 < \theta < \infty$. Derive the MP critical regions of size α for testing $H_0: \theta = \theta_0$ against (i) $H_1: \theta = \theta_1 \;\; (\theta_1 > \theta_0)$ and (i) $H_1: \theta = \theta_1 (\theta_1 < \theta_0)$. Show that the tests are actually UMP against one-sided alternatives. Also show that there exists no UMP test against two-sided alternatives.

5. Let $X_1, X_2, \ldots\ldots, X_n$ be independent random variables such that X_i is distributed as $N(\mu, \sigma_i^2)$, where σ_i^2 is known but μ is unknown. Derive UMP tests of level α for $H_0: \mu = \mu_0$ against two-sided alternatives. Is there a UMP test of level a for H_0 against two-sided alternatives?

6. Let $X_1, X_2, \ldots\ldots, X_n$ be a random sample from a distribution which has p.d.f.

$$f_\theta(x) = \begin{cases} \theta x^{\theta - 1} \quad if \; 0 < x < 1 \\ 0 \quad otherwise, \end{cases}$$

where $0 < \theta < \infty$. Show that the MP test of level α for testing $H_0: \theta = \theta_0$ against the alternative $H_1: \theta = \theta_1 (\theta > \theta_0)$ is given by the critical region

$$\left\{ x \middle| \prod_i x_i > \exp\left[\frac{\chi^2_{1-\alpha, 2n}}{2\theta_0} \right] \right\},$$

where $\chi^2_{1-\alpha, 2n}$ is the lower α-point of the χ^2 distribution with $df = 2n$.

7. Let (X_1, X_2, X_3) be a random sample from pdf or pmf given below. Find MP test of size $.05$ for testing $H_0: \theta = 1$ against $H_1: \theta = 2$ in each case. Also find the power of the test.

(i) $f(x; \theta) = N(0, \theta)$

(ii) $f(x; \theta) = N(2, \theta)$

(iii) $f(x; \theta) = G(1, \theta)$

(iv)$f(x;\theta) = NB(1,\theta) =$ Geometric distribution,

$$H_0: \theta = \frac{1}{4}, H_1: \theta = \frac{1}{2}$$

(v) $f(x;\theta) = \theta x^{\theta-1} \;\; 0 < x < 1, \theta > 0$

$= 0$ Otherwise

UNIT 5: GENERALIZED NEYMAN PEARSON LEMMA & UMPU

Structure

5.1 Introduction

The Most Powerful Test (MPT) maximizes the probability of detecting true effects for a given significance level, enhancing reliability and decision-making and the Neyman-Pearson Lemma provides a method to derive the most powerful test for simple hypotheses using a likelihood ratio. But in many situations Uniformly Most Powerful Tests do not exist. In such situations, we resort to some alternative course. One such alternative is Most Powerful Unbiased Test that ensures maximal power among all unbiased tests, maintaining fairness and accuracy. The Generalized Neyman-Pearson Lemma extends the lemma to find the most powerful unbiased tests, ensuring both power and unbiasedness in hypothesis testing. These principles are essential for developing robust statistical tests that provide reliable and fair conclusions in various applications.

5.2 Objective

The objective of this unit is to provide us a basic understanding of the concepts related to Generalized Neyman Pearson Lemma & construction of uniformly most powerful unbiased test. In unit**,** Generalized NP Lemma has been discussed and derived along with its utility in finding UMPU where UMP test does not exist has been extensively discussed. The general result for UMPU Test for exponential family has been derived & discussed extensively. These concepts should be clear after reading this material.

5.3 Generalized Neyman Pearson Lemma

Theorem: Suppose we have $(m+1)$ function $g_0(x), g_1(x), \ldots, g_m(x)$ which are integrable and let $0 \leq \phi(x) \leq 1$ such that

$$\int \phi(x) g_i(x) dx = c_i \;\; i = 1,2,\ldots,m \tag{1}$$

where c_i's are known constants.

Let $\phi_0(x)$ be a function such that

$$\phi_0(x) = \begin{cases} 1 \; ; g_0(x) > \sum_{i=1}^{m} k_i g_i(x) \\ 0 \; ; \; otherwise \end{cases}$$

then $\int \phi_0(x) g_0(x) dx \geq \int \phi(x) g_0(x) dx$

where $\phi_0(x)$ and $\phi(x)$ satisfy (1).

Proof: Consider

$$[\phi_0(x) - \phi(x)][g_0(x) - \textstyle\sum_{i=1}^{m} k_i g_i(x)] \geq 0 \tag{2}$$

One can easily see that if $\phi(x) = 1$ then (1) is nonnegative.

Similarly, if $\phi_0(x) = 1$ then also (1) is nonnegative.

From (2)

$$[\phi_0(x) - \phi(x)] \left[g_0(x) - \sum_{i=1}^{m} k_i g_i(x) \right] \geq 0$$

$$\Rightarrow \int [\phi_0(x) - \phi(x)] g_0(x) dx \;\; \geq \sum_{i=1}^{m} k_i \left[\int [\phi_0(x) g_i(x) dx] - \int \phi(x) g_i(x) dx \right]$$

$$\geq \textstyle\sum_{i=1}^{m} k_i [c_i - c_i] = 0 \qquad \text{(using (1))}$$

$$\Rightarrow \int [\phi_0(x) - \phi(x)] g_0(x) dx \geq 0$$

$$\Rightarrow \int \phi_0(x) g_0(x) dx \geq \int \phi(x) g_0(x) dx$$

Remark: We can see the difference between Neyman Pearson lemma and its extension

(i) There is equality in (1)

(ii) The function $g_0(x), g_1(x), \ldots, g_m(x)$ need not be pdf.

(iii) k_i's need not be nonnegative.

Definition (Unbiased Test): A test $\phi(x)$ is called an unbiased test of size α if

$E_{H_0}\phi(x) \leq \alpha$ or $\beta_\phi(\theta) \leq \alpha, \theta \in \Theta_0$ or $\sup_{\theta \in \Theta_0} \beta_\phi(\theta) = \alpha$, and

$E_{H_1}\phi(x) \geq \alpha$ or $\beta_\phi(\theta) \geq \alpha, \theta \in \Theta_1$,

Construction of UMP Unbiased (UMPU) Test: Assume that $f(x|\theta)$ involves single parameter. In this case, we will find an UMPU test for testing $H_0: \theta = \theta_0$ against $H_1: \theta \neq \theta_0$ for a size of α.

From the above definition,

$$\beta_\phi(\theta) \leq \alpha, \theta \in \Theta_0 \quad \text{(i)}$$

and

$$\beta_\phi(\theta) \geq \alpha, \theta \neq \theta_0 \quad \text{(ii)}$$

Suppose that the power function $E_{H_1}\phi(x)$ is a continuous function of θ.

From (i) and (ii), $E_{H_0}\phi(x)$ has minimum at $\theta = \theta_0$.

Therefore, we want to find ϕ such that

$$\beta_\phi(\theta_0) = \alpha \quad \text{(iii)}$$

and

$$\beta_\phi(\theta) > \alpha \text{ for } \theta \neq \theta_0 \quad \text{(iv)}$$

If $\beta_\phi(\theta)$ is differentiable, then

$$\frac{d\beta_\phi(\theta)}{d\theta}|_{\theta=\theta_0} = 0 \Rightarrow \beta'_\phi(\theta_0) = 0 \qquad \text{(v)}$$

Class of test satisfying (i) and (ii) is a subset of a class of tests satisfying (i) and (v). Now we have to find a MP test satisfying (i) and (v). If the test is independent of θ, then it is UMPU.

Now our problem reduces to fins a test ϕ such that

$$E_{H_0}\phi(x) = \int \phi(x) f(x|\theta_0 dx) = \alpha \qquad \text{(vi)}$$

and if regularity conditions are satisfied then

$$\frac{dE_{H_0}\phi(x)}{d\theta}|_{\theta=\theta_0} = \frac{d\beta_\phi(\theta)}{d\theta}|_{\theta=\theta_0}$$

$$= \int \frac{d}{d\theta}\phi(x) f(x|\theta_0) dx|_{\theta=\theta_0} = 0 \qquad \text{(vii)}$$

Now $\phi(x)$ maximizes the power $\int \phi(x) f(x|\theta_1) dx$ such that to find $\phi_0(x)$ that satisfies (vi) and (vii).

$$\text{i.e., } \int \phi_0(x) f(x|\theta_1) dx \geq \int \phi(x) f(x|\theta_1) dx \; \forall \; \theta$$

using theorem (1.1), i.e., GNP lemma,

$$g_0(x) = f(x|\theta_1), g_1(x) = f(x|\theta_0) \text{ and}$$

$$g_2(x) = \frac{df(x|\theta)}{d\theta}|_{\theta=\theta_0}$$

$$\phi_0(x) = \begin{cases} 1; f(x|\theta_1) > k_1 f(x|\theta_0) + k_2 \left[\frac{df(x|\theta)}{d\theta}\right]_{\theta=\theta_0} \\ 0; \quad otherwise \end{cases}$$

$$\phi_0(x) = \begin{cases} 1; & \frac{f(x|\theta_1)}{f(x|\theta_0)} > k_1 + k_2 \left[\frac{d \log f(x|\theta)}{d\theta}\right]_{\theta=\theta_0} \\ 0; & otherwise \end{cases} \quad (3)$$

where k_1 and k_2 are such that $E_{H_0}\phi_0(x) = \alpha$ and $\frac{dE_{\phi_0}(x)}{d\theta}|_{\theta=\theta_0} = 0$

5.4 UMPU Test for Exponential Family

Theorem: Let the rv X has pdf (pmf) $f(x|\theta), \theta \in \Theta$. Assume that $f(x|\theta)$ belongs to a one parameter exponential family.

$$f(x|\theta) = A(x) \exp[\theta T(x) + D(\theta)], x \in R, \theta \in \Theta$$

Then prove the test if

(i) U is continuous $\phi_0(x) = \begin{cases} 1 ; u < c_1 \text{ or } u > c_2 \\ 0 ; otherwise \end{cases}$ (4)

(ii) U is discrete $\phi_0(x) = \begin{cases} 1 ; u < c_1 \text{ or } u > c_2 \\ \gamma_i ; u = c_i, i = 1,2 \\ 0 ; otherwise \end{cases}$ (5)

Is UMPU of size α for testing $H_0: \theta = \theta_0$ against $H_1: \theta \neq \theta_0$, where $u = \sum_{i=1}^{n} T(x_i)$

Proof: Consider

$f(x_1, x_2, \ldots, x_n|\theta) = \prod_{i=1}^{n} A(x_i) \exp[\theta \sum_{i=1}^{n} T(x_i) + nD(\theta)]$

Using GNP and (3),

$$\phi_0(x) = \begin{cases} 1 ; \frac{f(x|\theta_1)}{f(x|\theta_0)} > k_1 + k_2 \left[\frac{d \log f(x|\theta)}{d\theta}\right]_{\theta=\theta_0} \\ 0 ; \quad otherwise \end{cases} \quad (6)$$

$$\frac{f_1(x_1, x_2, \ldots, x\_n|\theta_1}{f_0(x_1, x_2, \ldots, x\_n|\theta_0} = \frac{\exp[\theta_1 \sum_{i=1}^{n} T(x_i) + nD(\theta_1)]}{\exp[\theta_0 \sum_{i=1}^{n} T(x_i) + nD(\theta_0)]}$$

$$= \exp[(\theta_1 - \theta_0) \sum_{i=1}^{n} T(x_i) + n\{D(\theta_1) - D(\theta_0)\}] \quad (7)$$

Next,

$$\log f(x|\theta) = \sum_{i=1}^{n} \log A(x_i) + \theta \sum_{i=1}^{n} T(x_i) + nD(\theta)$$

$$\frac{d \log f(x|\theta)}{d\theta} = \sum_{i=1}^{n} T(x_i) + D'(\theta) \qquad (8)$$

From (6), (7), and (8)

$$\phi_0(x)$$

$$= \begin{cases} 1 \,;\, \exp\left[(\theta_1 - \theta_0)\sum_{i=1}^{n} T(x_i) + n\{D(\theta_1) - D(\theta_0)\}\right] > k_1 + k_2\left[\sum_{i=1}^{n} T(x_i) + nD'(\theta_0)\right] \\ 0 \,;\, \exp\left[(\theta_1 - \theta_0)\sum_{i=1}^{n} T(x_i) + n\{D(\theta_1) - D(\theta_0)\}\right] \leq k_1 + k_2\left[\sum_{i=1}^{n} T(x_i) + nD'(\theta_0)\right] \end{cases}$$

$$= \begin{cases} 1 \,;\, \exp\left[(\theta_1 - \theta_0)\sum_{i=1}^{n} T(x_i)\right] > k_1^* + k_2^* \sum_{i=1}^{n} T(x_i) \\ 0 \,;\, \exp\left[(\theta_1 - \theta_0)\sum_{i=1}^{n} T(x_i)\right] \leq k_1^* + k_2^* \sum_{i=1}^{n} T(x_i) \end{cases}$$

where

$$k_1^* = \frac{k_1}{n[D(\theta_1) - D(\theta_0)]} + k_2 \frac{nD'(\theta_0)}{n[D(\theta_1) - D(\theta_0)]} \,, k_2^* = \frac{k_2}{n[D(\theta_1) - (\theta_0)]}$$

Let $h[\sum_{i=1}^{n} T(x_i)] = \exp[(\theta_1 - \theta_0)\sum T(x)] - k_2^* \sum_{i=1}^{n} T(x_i)$

Then $$\phi_0(x) = \begin{cases} 1 \,;\, h[\sum_{i=1}^{n} T(x_i)] > k_1^* \\ 0 \,;\, h[\sum_{i=1}^{n} T(x_i)] \leq k_1^* \end{cases}$$

Let's observe the nature of $h[\sum_{i=1}^{n} T(x_i)]$. Let $u(x) = \sum_{i=1}^{n} T(x_i)$

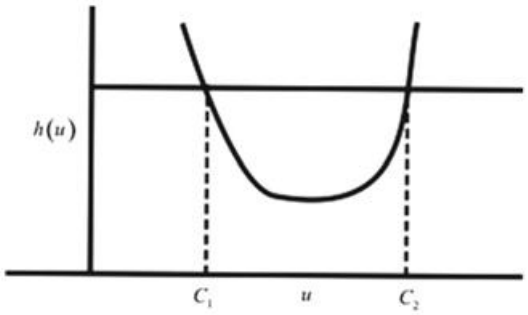
$$h(u) = \exp[(\theta_1 - \theta_0)u(x)] - k_2^* u(x)$$

$$h'(u) = (\theta_1 - \theta_0)\exp[(\theta_1 - \theta_0)u(x)] - k_2^*$$

$$h''(u) = (\theta_1 - \theta_0)^2 \exp\ [(\theta_1 - \theta_0)u(x)] > 0;$$

This implies that $h(u)$ is convex in U. If $h(u) > k_1^* \Rightarrow u < c_1$ or $u > c_2$

Graph of $h(u)$



The UMPU test is given as

$$\phi_0(x) = \begin{cases} 1\ ;\ u < c_1\ or\ u > c_2 \\ 0\ ;\ otherwise \end{cases}$$

If U is descrete

$$\phi_0(x) = \begin{cases} 1\ ;u < c_1\ or\ u > c_2 \\ \gamma_i\ ;u = c_i,\ \ i = 1,2 \\ 0\ ;otherwise \end{cases}$$

$\gamma_i (i = 1, 2)$ can be determined such that $E\phi_0(x) = \alpha$ and

$$\frac{d}{d\theta}[P(u < c_1) + P(u > c_2)]_{\theta=\theta_0} = 0.$$

Conclusion: For one parameter exponential family, we have seen how to obtain UMPU test for testing $H_0: \theta = \theta_0$ against $H_1: \theta \neq \theta_0$.

These conditions are as follows:

(i) $E\phi(x) = \alpha$

(ii) $\frac{d}{d\theta}E\phi(x)|_{\theta=\theta_0} = 0$

These conditions can be put in different form. Hence, we consider the following theorem:

Theorem: Let $f(x|\theta) = c(\theta)e^{\theta T(x)}h(x); x \in R, \theta \in \Theta$.

Prove that

$$E\phi(x)T(x) = \alpha ET(x), \tag{9}$$

where $\phi(x)$ is defined in (4) or (5)

Proof The test $\phi(x)$ defined in (4) or (5) satisfies (i) and (ii) defined in the conclusion.

$$\frac{d}{d\theta}E\phi(x) = \frac{d}{d\theta}\int \phi(x)c(\theta)e^{\theta T(x)}h(x)|_{\theta=\theta_0} = 0$$

$$\int \phi(x)c(\theta)e^{\theta T(x)}h(x)d(x) + \int \phi(x)c'^{(\theta)e^{\theta T(x)}}h(x)|_{\theta=\theta_0} = 0$$

$$\Rightarrow E[\phi(x)T(x)] + c'(\theta)\int \phi(x)e^{\theta T(x)}h(x)|_{\theta=\theta_0} = 0$$

$$\Rightarrow E[\phi(x)T(x)] + \frac{c'(\theta)}{c(\theta)}\int \phi(x)c(\theta)e^{\theta T(x)}h(x)|_{\theta=\theta_0} = 0$$

$$\Rightarrow E[\phi(x)T(x)] + \frac{c'(\theta)}{c(\theta)}E\phi(x) = 0 \tag{10}$$

This is true for all ϕ, which are unbiased.

Let $\phi(x) = \alpha$

$$\alpha E[T(x)] + \frac{c'(\theta)}{c(\theta)}\alpha = 0$$

Then $$E[T(x)] = -\frac{c'(\theta)}{c(\theta)} \tag{11}$$

From (10) and (11),

$$E[\phi(x)T(x)] = E[\phi(x)]E[T(x)]$$

Then we get the result as in (9),

$$E[\phi(x)T(x)] = \alpha E[T(x)]$$

Remark: We can find the constants c_1 and c_2 from

(i) $E[\phi(x)] = \alpha$ and

(ii) $E[\phi(x)T(x)] = \alpha E[T(x)]$.

Remark: A simplification of the test is possible if for $\theta = \theta_0$, the distribution of T is symmetric about some point a.

i.e., $$P_{\theta_0}[T < a - u] = P_{\theta_0}[T > a + u] \ \forall\, u$$

Then any test which is symmetric about a and satisfying $E_{H_0}\phi(x) = \alpha$, i.e., it satisfies (9).

Let $\psi(t)$ be symmetric about a and

$$E\psi(t) = \alpha,$$

then we have to show that $\psi(t)$ is unbiased, i.e., it satisfies (9).

$$E_{H_0}\psi(t)T(x) = E_{H_0}[(T-a)\psi(t) + a\psi(t)]$$

$$= E_{H_0}(T-a)\psi(t) + aE_{H_0}\psi(t)$$

$$= 0 + aE_{H_0}\psi(t)$$

(As T is symmetric about a then $ET = a$.)

$$= a\alpha = \alpha ET(x)$$

Hence it satisfies (9). Therefore $\psi(t)$ is unbiased.

Remark 3 c_i's and γ_i's can be found out such that $E_{H_0}\psi(t) = \alpha$

$$P_{H_0}[T < c_1] + \gamma_1 P_{H_0}[T = c_1] = \frac{\alpha}{2}$$

and

$$P_{H_0}[T < c_2] + \gamma_2 P_{H_0}[T = c_2] = \frac{\alpha}{2}$$

Further $\gamma_1 = \gamma_2$ and $a = \frac{(c_1+c_2)}{2}$

Example: Let the rvs $X_1, X_2, \ldots, X_n$ are iid rvs $N(\theta, 1)$. Find UMPU test for testing $H_0: \theta = \theta_0$ against $H_1: \theta \neq \theta_0$

$$f(x_1, x_2, \ldots, x_n|\theta) = (2\pi)^{-\frac{n}{2}} \exp\left[-\frac{1}{2}\sum_{i=1}^{n}(x_i - \theta)^2\right]$$

This belongs to a one parameter exponential family. Hence, from Theorem (2), $U = \sum T(x_i) = \sum x_i$

The UMPU test is

$$\phi(x) = \begin{cases} 1\,; \sum x_i < c_1 \; or \; \sum x_i > c_2 \\ 0\,; otherwise \end{cases}$$

or

$$\phi(x) = \begin{cases} 1\,; \bar{x} < c_1 \; or \; \bar{x} > c_2 \\ 0\,; otherwise \end{cases}$$

$\bar{X}$ is distributed as $N\left(\theta, \frac{1}{n}\right)$, thus

$$E_{H_0}\phi(x) = \alpha$$

$$P_{H_0}(\bar{X} < c_1) + P(\bar{X} > c_2) = \alpha$$

$$\Rightarrow \int_{-\infty}^{c_1} \frac{\sqrt{n}}{\sqrt{2\pi}} \exp\left[-\frac{n(\bar{x}-\theta_0)^2}{2}\right] d\bar{x} + \int_{c_2}^{-\infty} \frac{\sqrt{n}}{\sqrt{2\pi}} \exp\left[-\frac{n(\bar{x}-\theta_0)^2}{2}\right] d\bar{x} = \alpha$$

Next, $\frac{dE\phi(x)}{d\theta}|_{\theta=\theta_0} = 0$

$$\Rightarrow \frac{d}{d\theta}\left[\int_{-\infty}^{c_1} \frac{\sqrt{n}}{\sqrt{2\pi}} \exp\left[-\frac{n(\bar{x}-\theta_0)^2}{2}\right]\right] d\bar{x} + \int_{c_2}^{-\infty} \frac{\sqrt{n}}{\sqrt{2\pi}} \exp\left[-\frac{n(\bar{x}-\theta_0)^2}{2}\right] d\bar{x}|_{\theta=\theta_0} = 0$$

$$\Rightarrow \frac{d}{d\theta}\left[\int_{-\infty}^{\sqrt{n}(c_1-\theta)} \frac{e^{-\frac{t^2}{2}}}{\sqrt{2\pi}} \int_{\sqrt{n}(c_2-\theta)}^{\infty} \frac{e^{-\frac{t^2}{2}}}{\sqrt{2\pi}}\right]|_{\theta=\theta_0} = 0$$

$$\Rightarrow -\frac{\sqrt{n}}{\sqrt{2\pi}} \exp\left[-\frac{n(c_1-\theta_0)^2}{2}\right] + \frac{\sqrt{n}}{\sqrt{2\pi}} \exp\left[-\frac{n(c_2-\theta_0)^2}{2}\right] = 0$$

$$\Rightarrow \exp\left[-\frac{n(c_1-\theta_0)^2}{2}\right] = \exp\left[-\frac{n(c_2-\theta_0)^2}{2}\right]$$

$$\Rightarrow (c_1-\theta_0)^2 = (c_2-\theta_0)^2$$

$$\Rightarrow (c_1-\theta_0)^2 - (c_2-\theta_0)^2 = 0$$

$$\Rightarrow (c_1-c_2)(c_1+c_2-2\theta_0) = 0$$

$$\Rightarrow c_1 = c_2 \text{ and } c_1+c_2-2\theta_0 = 0$$

$$\Rightarrow c_1 = 2\theta_0 - c_2 \text{ or } c_2 = 2\theta_0 - c_1$$

Now $E\phi(x) = \alpha$ and $\sqrt{n}(c_2-\theta_0) = \sqrt{n(2\theta_0-c_1-\theta_0)} = \sqrt{n}(\theta_0-c_1)$

Hence,

$$\int_{-\infty}^{\sqrt{n}(c_1-\theta_0)} \frac{e^{-\frac{t^2}{2}}}{\sqrt{2\pi}} dt + \int_{\sqrt{n}(\theta_0-c_1)}^{\infty} \frac{e^{-\frac{t^2}{2}}}{\sqrt{2\pi}} dt = \alpha$$

$$\Rightarrow 2\int_{-\infty}^{\sqrt{n}(c_1-\theta_0)} \frac{e^{-\frac{t^2}{2}}}{\sqrt{2\pi}}dt = \alpha$$

$$\Rightarrow \int_{-\infty}^{\sqrt{n}(c_1-\theta_0)} \frac{e^{-\frac{t^2}{2}}}{\sqrt{2\pi}}dt = \frac{\alpha}{2}$$

$$\Rightarrow \int_{-\sqrt{n}(c_1-\theta_0)}^{\infty} \frac{e^{-\frac{t^2}{2}}}{\sqrt{2\pi}}dt = \frac{\alpha}{2}$$

$$\Rightarrow -\sqrt{n}(c_1-\theta_0) = Z_{\frac{\alpha}{2}}$$

$$\Rightarrow c_1 = \theta_0 - \frac{Z_{\frac{\alpha}{2}}}{\sqrt{n}} \text{ and } c_2 = \theta_0 + \frac{Z_{\frac{\alpha}{2}}}{\sqrt{n}}$$

Hence, the UMPU test is

$$\phi(x) = \begin{cases} 1\,; \bar{x} < \theta_0 - \dfrac{Z_{\frac{\alpha}{2}}}{\sqrt{n}} \ or\ \bar{x} > \theta_0 + \dfrac{Z_{\frac{\alpha}{2}}}{\sqrt{n}} \\ 0\,; otherwise \end{cases}$$

5.5 Self-Assessment Exercise

1. Let $(X_1, X_2, \ldots, X_n)$ be a random sample from pdf given below. Find UMPU test of size α for testing H_0 against H_1 mentioned against each problem at level .

(i) $f(x;\theta) = \theta x^{\theta-1},\ 0 < x < 1, \theta > 0;$

$= 0$ otherwise

$H_0: \theta = \theta_0$ against $H_1: \theta \neq \theta_0$

(ii) $f(x;\theta) = G(1,\theta)$

$H_0: \theta = \theta_0\ against\ H_1: \theta \neq \theta_0$

(iii) $f(x;\theta) = N(\theta, \sigma^2)$, both θ and σ^2 are unknown

$H_0: \sigma^2 = \sigma_0^2\ against\ H_1: \sigma^2 \neq \sigma_0^2$

UNIT 6: LIKELIHOOD RATIO TESTS

Structure

6.1 Introduction

The likelihood ratio test (LRT) is a powerful statistical tool used to compare the fit of two models: a null model and an alternative model. Its utility lies in assessing whether the more complex model significantly improves the fit to the data compared to the simpler model. LRT is widely applicable across different statistical models, including regression and ANOVA. It provides a systematic approach to hypothesis testing by calculating the ratio of the maximum likelihoods under both models and using this ratio to determine significance. This helps in making informed decisions about model adequacy and selecting the best-fitting model.

6.2 Objective

The objective of this unit is to provide us a basic understanding of the concepts related to likelihood ratio test has been put forth & its various properties have been discussed. Further, many important likelihood ratio tests have been derived & discussed extensively. These concepts should be clear after reading this material.

6.3 Introduction to Likelihood Ratio Tests

In this unit, we consider a classical procedure for constructing tests that has some intuitive appeal and that frequently, though not necessarily, leads to optimal tests. The procedure also leads to tests that have some desirable large-sample properties. Neyman and Pearson suggested a simple method for testing a general testing problem.

Consider a random sample $X_1, X_2, \ldots, X_n$ from $f(x|\theta), \theta \in \Theta$ and we have to tests

$$H_0: \theta \in \Theta_0 \text{ against } H_1: \theta \in \Theta_1 \tag{1}$$

The likelihood ratio test for testing (1) is defined as

$$\lambda(x) = \frac{\sup_{\theta \in \Theta_0} L(\theta|x)}{\sup_{\theta \in \Theta} L(\theta|x)} \tag{2}$$

where $L(\theta|x)$ is the likelihood function of x.

Definition: Let $L(\theta|x)$ be a likelihood for a random sample having the joint pdf(pmf) $f(x|\theta)$ for $\theta \in \Theta$. The likelihood ratio is defined to be

$$\lambda(x) = \frac{\sup_{\theta \in \Theta_0} L(\theta|x)}{\sup_{\theta \in \Theta} L(\theta|x)}$$

The numerator of the likelihood ratio $\lambda(x)$ is the best explanation of X that the null hypothesis H_0 can provide and the denominator is the best possible explanation of X. H_0 is rejected if there is a much better explanation of X than the best one provided by H_0 Further, it implies that smaller values of λ leads to the rejection of H_0 and larger values of λ leads to acceptance of H_0. Therefore, the critical region would be of the type $\lambda(x) < k$. Note that $0 \leq \lambda \leq 1$. The constant k is determined by

$$\sup_{\theta \in \Theta_0} P[\lambda(x) < k] = \alpha.$$

If the distribution of $\lambda(x)$ is continuous, then the size α is exactly attained and no randomization on the boundary is required. Similarly, if the distribution of $\lambda(x)$ is discrete, the size may not attain α and then we require randomization. We will see the following theorems without proof.

Theorem: For $0 \leq \alpha \leq 1$ nonrandomized Neyman-Pearson and likelihood ratio test of a simple hypothesis against a simple alternative exist, they are equivalent.

Proof: Left as an exercise.

Theorem: For testing $\theta \in \Theta_0$ against $\theta \in \Theta_1$, the likelihood ratio test is a function of every sufficient statistic for θ.

Proof: Left as an exercise.

The notation $\lambda = \frac{L(\hat{\omega})}{L(\hat{\Omega})}$ has been in wide use. Notice that $0 < \lambda \leq 1$. Also the statistic $-2\log\lambda$ rather than λ itself is employed, the reason being that, under certain regularity conditions, the asymptotic distribution of $-2\log\lambda$ under H_0 , is known. Then in terms of this statistic, one rejects H_0 whenever $-2\log\lambda > K,$ where K is determined by the desired level of the test. Of course, this test is equivalent to LR test. In carrying out the likelihood ratio test in actuality, one is apt to encounter two sorts of difficulties. First is the problem of determining the cut-off point K and second is problem of actually determining $L(\hat{\omega})$ and $L(\hat{\Omega})$. The first difficulty is removed at the asymptotic level, in the sense that we may use as an approximation (for large n) the limiting distribution of $-2\log\lambda$ for specifying. The problem of finding $L(\hat{\Omega})$ is essentially that of finding the MLE of θ. Calculating $L(\hat{\omega})$ is much harder a problem. In many cases, however, H_0 is simple and then no problem exists. In many cases of practical interest and for a fixed sample size, the LR test happens to coincide with or to be close to other tests which are known to have some well-defined optimal properties such as being UMP or being UMPU.

Theorem: (Without Proof) Let $(X_1, X_2, \ldots, X_n)$ be a random sample from a pdf or pmf $(.;\ \theta)$; $\theta \varepsilon \Omega \subseteq R^r$, and let ω be an m-dimensional subset of Ω. Suppose also that the set of positivity

of the pdf or pmf does not depend on θ. The under some additional regularity conditions, the asymptotic distribution of $-2\log\lambda$ is $\mathcal{X}^2_{r-m}$, provided $\theta\ \varepsilon\ \omega$; that is as $n \mapsto \infty$

$$P_\theta[-2\log\lambda \leq x] \mapsto G(x), \qquad x \geq 0\ for\ all\ \theta\ \varepsilon\ \omega$$

where G is the distribution function of a $\mathcal{X}^2_{r-m}$ distribution.

6.4 Some Important Likelihood Ratio Tests

Example: Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from a normal distribution with mean μ with variance σ^2. Obtain the likelihood ratio test for testing $H_0: \mu = \mu_0$ against $H_1: \mu \neq \mu_0$

Case (i) σ^2 is known

Note that there is no UMP test for this problem.

The likelihood function is

$$L(\mu|X_1, X_2, \ldots, X_n) = \Pi_{i=1}^n \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma^2}(x_i-\mu)^2\right]$$

$$= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu)^2\right]$$

Consider

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

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

$$L(\mu|X) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left[-\frac{n}{2\sigma^2}(\bar{x}-\mu)^2\right] \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\bar{x})^2\right]$$

$$\sup_{\theta\in\Theta} L(\mu|x) = \sup_{\mu=\mu_0} L(\mu|x) = L(\mu_0)$$

$$= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left[-\frac{n}{2\sigma^2}(\bar{x}-\mu_0)^2\right] \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\bar{x})^2\right]$$

$\sup_{\theta\in\Theta} L(\mu|x) = L(\hat{\mu}|x)$, $\hat{\mu}$ is the mle of μ. Also, mle of $\mu = \hat{\mu} = \bar{X}$

$$L(\hat{\mu}|x) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right]$$

From (2)

$$\lambda(x) = \exp\left[-\frac{n}{2\sigma^2}(\bar{x}-\mu_0)\right]$$

The likelihood ratio is just a function of $\frac{n(\bar{x}-\mu_0)^2}{\sigma^2}$ and will be small when the quantity is large

The LR test is given as

$$\phi(x) = \begin{cases} 1\ ; \left|\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}\right| > k \\ 0\ ; otherwise \end{cases}$$

Since $\left(\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}\right)$ is $N(0,1)$. Then k can be obtained as $P\left[\left(\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}\right) > k\right] = \frac{\alpha}{2}$.

Therefore, $k = Z_{\frac{\alpha}{2}} = \frac{\alpha}{2}$th quantile of $N(0,1)$

Then the test is given as

$$\phi(x) = \begin{cases} 1\ ; \left|\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}\right| > Z_{\frac{\alpha}{2}} \\ 0\ ; otherwise \end{cases}$$

Case (ii) σ^2 unknown

We have to test $H_0: \mu = \mu_0$ against $H_1: \mu \neq \mu_0$

From (1),

$$\sup_{\mu=\mu_0,\sigma^2>0} L(\mu|x) = \sup_{\sigma^2>0} \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu_0)^2\right]$$

MLE of $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu_0)^2$

Let $s_0^2 = \sum_{i=1}^{n}(x_i-\mu_0)^2 \Rightarrow \hat{\sigma}^2 = \frac{s_0^2}{n}$

Under H_0,

$$\sup_{\sigma^2>0} L(\mu|x) = \sup_{\sigma^2>0} \left(\frac{\sqrt{n}}{s_0\sqrt{2\pi}}\right)^n \exp\left(-\frac{n}{2}\right) \qquad (3)$$

Further, MLE of μ and σ^2 is $\hat{\mu}$ and $\widehat{\sigma^2}$ given by

$\hat{\mu} = \bar{X}$ and $\hat{\sigma}^2 = \frac{s^2}{n}, s^2 = \sum(x_i-\bar{x})^2$

$$\sup_{\mu,\sigma^2} \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^{\frac{n}{2}} \exp\left[-\frac{2}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right] = \left(\frac{\sqrt{n}}{s\sqrt{2\pi}}\right)^n \exp\left(-\frac{n}{2}\right) \qquad (4)$$

From (2), (3) and (4), we have

$$\lambda(x) = \left(\frac{s}{s_0}\right)^n = \left(\frac{s^2}{s_0^2}\right)^{\frac{n}{2}}$$

Since

$$s_0^2 = \sum_{i=1}^{n}(x_i-\mu_0)^2 = \sum_{i=1}^{n}(x_i-\bar{x}+\bar{x}+\mu_0)^2$$

$$= \sum_{i=1}^{n} (x_i - \bar{x})^2 + n(\bar{x} - \mu_0)^2 \; = s^2 + n(\bar{x} - \mu_0)^2$$

so that

$$\lambda(x) = \left[\frac{s^2}{s^2 + n(\bar{x} - \mu_0)^2}\right]^{\frac{n}{2}} = \left[\frac{1}{1 + \frac{n(\bar{x} - \mu_0)^2}{s^2}}\right]^{\frac{n}{2}}$$

Further, the critical region is $\lambda(x) < k$

$$\Rightarrow \frac{n(\bar{x} - \mu_0)^2}{s^2} > k$$

$$\Rightarrow \left|\frac{(\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}\right| > k$$

Thus, the likelihood ratio test is

$$\phi(x) = \begin{cases} 1 \; ; \left|\frac{\sqrt{n}(\bar{x} - \mu_0)}{s}\right| > k \\ 0 \; ; otherwise \end{cases}$$

Now, since $\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$ and $\frac{s^2}{\sigma^2} \sim \chi^2_{n-1}$

Hence, $$\frac{\left(\frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}\right)}{\sqrt{\frac{(n-1)s^2}{\frac{\sigma^2}{n-1}}}} \sim t_{n-1}$$

The distribution t_{n-1} is symmetric about 0,

$$P_{H_0}\left\{\frac{\sqrt{n(\bar{x} - \mu_0)}}{s} > k\right\} = \frac{\alpha}{2}$$

The likelihood ratio test is given as

$$\phi(x) = \begin{cases} 1\ ; & \left|\frac{\sqrt{n}(\bar{x}-\mu_0)}{s}\right| > t_{n-1,\frac{\alpha}{2}} \\ 0\ ; & \text{otherwise} \end{cases}$$

Example: Let $X_1, X_2, \ldots, X_n$ be a random sample from $N(\mu, \sigma^2)$. Obtain a LR test to test $H_0: \sigma^2 = \sigma_0^2$ against $H_1: \sigma^2 \neq \sigma_0^2$ with population mean μ is unknown.

$$L(\mu, \sigma^2|X) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right]$$

so that

$$\sup_{(\mu,\sigma_0^2)\in\Theta_0} L(\mu, \sigma^2|X) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left[-\frac{s^2}{2\sigma^2}\right]$$

Also, the maximum likelihood estimate of μ and σ^2 are

$\hat{\mu} = \bar{x}$ and $\widehat{\sigma^2} = \frac{\sum(x_i-\bar{x})^2}{n} = \frac{s^2}{n}$

$$\sup_{\mu,\sigma^2} L(\mu, \sigma^2|X) = \left(\frac{\sqrt{n}}{s\sqrt{2\pi}}\right)^n \exp\left(-\frac{n}{2}\right)$$

The LR statistic is given by

$$\lambda(x) = \left(\frac{s}{\sqrt{n}\sigma_0}\right)^n \exp\left[-\frac{1}{2}\left\{\frac{s^2}{\sigma_0^2} - n\right\}\right]$$

The CR is $\lambda(x) < k$

$$\Rightarrow \left(\frac{s}{\sqrt{n}\sigma_0}\right)^n \exp\left[-\frac{1}{2}\left\{\frac{s^2}{\sigma_0^2} - n\right\}\right] < k$$

Let $w = \frac{s^2}{\sigma_0^2} \sim \chi^2_{n-1}$

$$\Rightarrow \left(\frac{w}{n}\right)^{\frac{n}{2}} \exp\left(-\frac{w}{2}\right) < k$$

$$\Rightarrow w < k_1 \; or \; w > k_2$$

Under H_0

$$P_{H_0}[w < k_1] + P_{H_0}[w > k_2] = \alpha$$

Distributing the error probability, i.e., α equally in tails, we get $k_1 = \chi^2_{n-1,1-\frac{\alpha}{2}}$ and $k_2 = \chi^2_{n-1,\frac{\alpha}{2}}$

The LR test is

$$\phi(x) = \begin{cases} 1 \; ; \frac{s^2}{\sigma^2} \leq \chi^2_{n-1,1-\frac{\alpha}{2}} \; or \; \frac{s^2}{\sigma^2} \geq \chi^2_{n-1,\frac{\alpha}{2}} \\ 0 \; ; \qquad\qquad \text{otherwise} \end{cases}$$

Example: Let $X_1, X_2, \ldots, X_m$ and $Y_1, Y_2, \ldots, Y_n$ be independent rvs from $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$ respectively. Obtain the likelihood ratio test for $H_0: \mu_1 = \mu_2$ against $H_1: \mu_1 \neq \mu_2$ under two conditions.

i. σ_1^2 and σ_2^2 known

ii. $\sigma_1^2 = \sigma_2^2 = \sigma^2$ unknown

i. The likelihood function for $(\mu_1, \mu_2) \in \Theta$ is given as

$$L(\mu_1, \mu_2 | X, Y) = \left(\frac{1}{\sigma_1\sqrt{2\pi}}\right)^m \exp\left[-\frac{1}{2\sigma_1^2}\sum_{i=1}^{m}(x_i - \mu_1)^2\right]$$

$$\times \left(\frac{1}{\sigma_2\sqrt{2\pi}}\right)^n \exp\left[-\frac{1}{2\sigma_2^2}\sum_{i=1}^{n}(y_i - \mu_2)^2\right] \qquad (5)$$

The MLE of $\mu_1 = \widehat{\mu_1} = \bar{x}$ and MLE of $\mu_2 = \widehat{\mu_2} = \bar{y}$.

$$\sup_{\mu_1,\mu_2} L(\mu_1, \mu_2 | x, y) = \left(\frac{1}{\sigma_1\sqrt{2\pi}}\right)^m \exp\left[-\frac{1}{2\sigma_1^2}\sum_{i=1}^{m}(x_i - \bar{x})^2\right]$$

$$\times \left(\frac{1}{\sigma_2\sqrt{2\pi}}\right)^n \exp\left[-\frac{1}{2\sigma_2^2}\sum_{i=1}^{n}(y_i-\bar{y})^2\right] \qquad (6)$$

The likelihood function for $\mu \in \Theta_0$ is given as

$$L(\mu|X,Y) = \left(\frac{1}{\sigma_1\sqrt{2\pi}}\right)^m \left(\frac{1}{\sigma_1\sqrt{2\pi}}\right)^n \exp\left[-\frac{1}{2\sigma_1^2}\sum_{i=1}^{m}(x_i-\mu_1)^2\right]$$
$$\times \exp\left[-\frac{1}{2\sigma_2^2}\sum_{i=1}^{n}(y_i-\mu_2)^2\right]$$

$$\frac{\partial \log L(\mu|x,y)}{\partial \mu} = 0$$

$$\Rightarrow \hat{\mu} = MLE\ of\ \mu = \frac{\left(\frac{m\,\bar{x}}{\sigma^2}+\frac{n\,\bar{y}}{\sigma_2^2}\right)}{\frac{m}{\sigma_1^2}+\frac{n}{\sigma_2^2}} \qquad (7)$$

$$\sup_{\mu\in\Theta_0} L(\mu|X,Y) = \left(\frac{1}{\sigma_1\sqrt{2\pi}}\right)^m \left(\frac{1}{\sigma_1\sqrt{2\pi}}\right)^n$$
$$\times \exp\left[-\frac{1}{2\sigma_1^2}\sum_{i=1}^{m}(x_i-\hat{\mu})^2 - \frac{1}{2\sigma_2^2}\sum_{i=1}^{n}(y_i-\hat{\mu})^2\right]$$
$$= \left(\frac{1}{\sigma_1\sqrt{2\pi}}\right)^m \left(\frac{1}{\sigma_1\sqrt{2\pi}}\right)^n \exp\left[-\frac{1}{2\sigma_1^2}\left\{\sum_{i=1}^{m}(x_i-\bar{x})^2 + m(\bar{x}-\hat{\mu})^2\right\}\right]$$
$$\times \exp\left[-\frac{1}{2\sigma_2^2}\left\{\sum_{i=1}^{m}(y_i-\bar{y})^2 + n(\bar{y}-\mu)^2\right\}\right] \qquad (8)$$

Now, consider the LR statistic given by

$$\lambda(x,y) = \sup_{\mu\in\Theta_0} L(\mu|X,Y) / \sup_{\mu_1,\mu_2} L(\mu_1,\mu_2|x,y) \qquad (9)$$

From (6) and (8), we have

$$\lambda(x,y) = \exp\left[-\frac{m(\bar{x}-\hat{\mu})^2}{2\sigma_1^2} - \frac{n(\bar{y}-\hat{\mu})^2}{2\sigma_2^2}\right] \tag{10}$$

From (7), we have

$$\mu = \frac{\left(\frac{\bar{x}}{\frac{\sigma^2}{m}} + \frac{n\bar{y}}{\frac{\sigma_2^2}{n}}\right)}{\frac{m}{\sigma_1^2} + \frac{n}{\sigma_2^2}}$$

Let $a_1 = \frac{\sigma_1^2}{m},\ a_2 = \frac{\sigma_2^2}{n}$

$$\mu = \frac{\left(\frac{\bar{x}}{a_1} + \frac{\bar{y}}{a_2}\right)}{\frac{1}{a_1} + \frac{1}{a_2}} = \frac{a_2\bar{x} + a_1\bar{y}}{a_2 + a_1}$$

$$\hat{\mu} = \frac{\frac{\sigma_2^2}{n}\bar{x} + \frac{\sigma_1^2}{m}\bar{y}}{\frac{\sigma_2^2}{n} + \frac{\sigma_1^2}{m}}$$

Consider $\bar{x} - \mu = \bar{x} - \frac{\frac{\sigma_2^2}{n}\bar{x} + \frac{\sigma_1^2}{m}\bar{y}}{\frac{\sigma_2^2}{n} + \frac{\sigma_1^2}{m}}$

$$= \frac{\bar{x}\left[\frac{\sigma_2^2}{n} + \frac{\sigma_1^2}{m}\right] - \frac{\sigma_2^2}{n}\bar{x} - \frac{\sigma_1^2\bar{y}}{m}}{\frac{\sigma_2^2}{n} + \frac{\sigma_1^2}{m}}$$

$$= \frac{\frac{\sigma_1^2}{m}(\bar{x}-\bar{y})}{\frac{\sigma_2^2}{n} + \frac{\sigma_1^2}{m}} \tag{11}$$

Proceeding similarly, we have

$$\bar{y} - \mu = \frac{\frac{\sigma_2^2}{m}(\bar{y} - \bar{x})}{\frac{\sigma_2^2}{m} + \frac{\sigma_1^2}{n}} \qquad (12)$$

From (10), we have

$$\lambda(x, y) = \exp\left[-\frac{m}{2\sigma_1^2}\left\{\frac{\frac{\sigma_1^2}{m}(\bar{x} - \bar{y})}{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}\right\}^2 - \frac{n}{2\sigma_2^2}\left\{\frac{\frac{\sigma_2^2}{m}(\bar{y} - \bar{x})}{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}\right\}^2 \right]$$

$$= \exp\left[-\frac{1}{2}\frac{(\bar{x} - \bar{y})^2}{\left(\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}\right)} \right]$$

$$\lambda(x, y) < k \Leftrightarrow \left(\frac{(\bar{x} - \bar{y})^2}{\left(\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}\right)} \right) > k$$

$$\Leftrightarrow \left| (\bar{x} - \bar{y}) / \sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}} \right| > k \qquad (13)$$

Since, $\bar{x} \sim N\left(\mu_1, \frac{\sigma_1^2}{m}\right)$, $\bar{y} \sim N\left(\mu_2, \frac{\sigma_2^2}{n}\right)$ and $\bar{X}$ and $\bar{Y}$ are independent , then $\bar{x} - \bar{y} \sim N(\mu_1 - \mu_2, \frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n})$

$$\frac{(\bar{x} - \bar{y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}} \sim N(0, 1) \qquad (14)$$

Under H_0, $\frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}} \sim N(0, 1)$

LR test is given as $\phi(x, y) = \begin{cases} 1 \ ; \frac{|(\bar{x} - \bar{y})|}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}} > k \\ 0 \ ; otherwise \end{cases}$

$$E_{H_0}\phi(x,y) = \alpha \Rightarrow P\left[\frac{(\bar{x}-\bar{y})}{\sqrt{\frac{\sigma_1^2}{m}+\frac{\sigma_2^2}{n}}} > k\right] = \frac{\alpha}{2} \Rightarrow k = Z_{\frac{\alpha}{2}}$$

The LR test is given by

$$\phi(x,y) = \begin{cases} 1\ ; \dfrac{|(\bar{x}-\bar{y})|}{\sqrt{\frac{\sigma_1^2}{m}+\frac{\sigma_2^2}{n}}} > Z_{\frac{\alpha}{2}} \\ 0\ ; \quad \text{otherwise} \end{cases}$$

(ii) The maximum likelihood estimate of μ_1, μ_2 and σ^2 will be $\widehat{\mu_1} = \bar{x}, \widehat{\mu_2} = \bar{y}$

$$\hat{\sigma}^2 = \frac{1}{m+n}[\sum_{i=1}^{m}(x_i-\bar{x})^2 + \sum_{i=1}^{n}(y_i-\bar{y})^2] = \frac{s_1^2+s_2^2}{m+n}$$

From (5) and substituting these ML estimates, we get

$$\sup_{\mu_1,\mu_2\in\Theta} L(\mu_1,\mu_2,\sigma^2|x,y) = \left[\frac{m+n}{\sqrt{2\pi(s_1^2+s_2^2)}}\right]^{m+n} \exp\left[-\frac{1}{2}(m+n)\right]$$

Under H_0,

$$L(\mu,\sigma^2|x,y) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^{m+n} \exp\left[-\frac{1}{2\sigma^2}\left\{\sum_{i=1}^{m}(x_i-\mu)^2 + \sum_{i=1}^{n}(y_i-\mu)^2\right\}\right]$$

The ML estimates of μ and σ^2 are $\hat{\mu} = \frac{m\bar{x}+n\bar{y}}{m+n}$ and

$$\widehat{\sigma^2} = \frac{1}{m+n}\left[\sum_{i=1}^{m}(x_i-\mu)^2 + \sum_{i=1}^{n}(y_i-\mu)^2\right]$$

$$= \frac{1}{m+n}\left[\sum_{i=1}^{m}(x_i-\bar{x})^2 + m(\bar{x}-\hat{\mu})^2 + \sum_{i=1}^{n}(y_i-\bar{y})^2 + n(\bar{y}-\hat{\mu})^2\right]$$

Now,

$$(\bar{x}-\hat{\mu})^2=\left[\bar{x}-\frac{m\bar{x}+n\bar{y}}{m+n}\right]^2=\left[\frac{n(\bar{x}-\bar{y})}{m+n}\right]^2$$

$$(\bar{y}-\hat{\mu})^2=\left[\bar{y}-\frac{m\bar{x}+n\bar{y}}{m+n}\right]^2=\left[\frac{n(\bar{y}-\bar{x})}{m+n}\right]^2$$

$$\widehat{\sigma^2}=\frac{1}{m+n}\left[s_1^2+s_2^2+\frac{mn^2(\bar{x}-\bar{y})^2}{(m+n)^2}+\frac{nm^2(\bar{y}-\bar{x})^2}{(m+n)^2}\right] \quad (15)$$

$$=\frac{1}{m+n}\left[s_1^2+s_2^2+\frac{mn(\bar{x}-\bar{y})^2}{m+n}\right] \quad (16)$$

$$\sup_{\mu,\sigma^2\in\Theta_0} L(\mu,\sigma^2|x,y)=\left[\frac{m+n}{\sqrt{2\pi\left[s_1^2+s_2^2+\frac{mn}{m+n}(\bar{x}-\bar{y})^2\right]}}\right]^{m+n} e^{-\left(\frac{m+n}{2}\right)} \quad (17)$$

The LR statistic is given by

$$\lambda(x,y)=\frac{\sup_{\mu,\sigma^2\in\Theta_0} L(\mu,\sigma^2|x,y)}{\sup_{\mu_1,\mu_2,\sigma^2\in\Theta_0} L(\mu_1,\mu_2,\sigma^2|x,y)}$$

$$=\left[\frac{s_1^2+s_2^2}{s_1^2+s_2^2+\frac{mn}{m+n}(\bar{x}-\bar{y})^2}\right]^{\frac{m+n}{2}}$$

$$=\left[1+\frac{mn(\bar{x}-\bar{y})^2}{(m+n)(s_1^2+s_2^2)}\right]^{-\left(\frac{m+n}{2}\right)} \quad (18)$$

$\bar{x}\sim N\left(\mu_1,\frac{\sigma^2}{m}\right)$ and $\bar{y}\sim N\left(\mu_2,\frac{\sigma^2}{n}\right)$, $\bar{x}-\bar{y}\sim N(\mu_1-\mu_2,\sigma^2\left(\frac{1}{m}+\frac{1}{n}\right)$

$$\left[\frac{(\bar{x}-\bar{y})-(\mu_1-\mu_2)}{\sigma\sqrt{\frac{1}{m}+\frac{1}{n}}}\right]\sim N(0,1)$$

$$\frac{s_1^2}{\sigma^2} = \sum \frac{(x_i - \bar{x})^2}{\sigma^2} \chi^2_{m-1}, \qquad \frac{s_2^2}{\sigma^2} = \sum \frac{(y_i - \bar{y})^2}{\sigma^2} \sim \chi^2_{n-1}$$

$$\frac{s_1^2 + s_2^2}{\sigma^2} = \chi^2_{m+n-2}$$

Thus, under $H_0: \mu_1 = \mu_2$

$$t = \frac{(\bar{x} - \bar{y})}{\left[\sigma\sqrt{\frac{1}{m} + \frac{1}{n}}\right]} \Big/ \sqrt{\frac{s_1^2 + s_2^2}{\sigma^2(m+n-2)}} \sim t_{m+n-2}$$

$$= \frac{\sqrt{\frac{mn}{m+n}}(\bar{x} - \bar{y})}{\sqrt{\frac{s_1^2 + s_2^2}{\sigma^2(m+n-2)}}} \sim t_{m+n-2}$$

$$t^2 = \frac{mn(\bar{x} - \bar{y})^2}{(m+n)(s_1^2 + s_2^2)}(m+n-2)$$

$$\frac{t^2}{m+n-2} = \frac{mn(\bar{x} - \bar{y})^2}{(m+n)(s_1^2 + s_2^2)} \tag{19}$$

From (18), we have

$$\lambda^{\frac{2}{m+n}}(x, y) = \frac{1}{1 + \frac{mn(\bar{x} - \bar{y})^2}{(m+n)(s_1^2 + s_2^2)}}$$

$$= \left[1 + \frac{mn(\bar{x} - \bar{y})^2}{(m+n)(s_1^2 + s_2^2)}\right]^{-1}$$

Using (19), we get

$$\lambda^{\frac{2}{m+n}}(x, y) = \left[1 + \frac{t^2}{m+n-2}\right]^{-1}$$

$$= \frac{m+n-2}{m+n-2+t^2}$$

The critical region is given by

$$\lambda^{\frac{2}{m+n}}(x,y) < k$$

$$\Rightarrow \frac{m+n-2}{m+n-2+t^2} < k$$

$$\Rightarrow t^2 > k \qquad \text{or} \qquad \Rightarrow |t| > k$$

k is obtained such that

$$P_{H_0}[|T| > k] = \alpha$$

so that k is an upper $\frac{\alpha}{2}$th quantile of t distribution with df $m+n-2$.

Then LR test is given as

$$\phi(x,y) = \begin{cases} 1 \ ; \dfrac{|(\bar{x}-\bar{y})|}{s\sqrt{\dfrac{1}{m}+\dfrac{1}{n}}} > t_{\frac{\alpha}{2},m+n-2} \\ 0 \quad ; \qquad \text{otherwise} \end{cases}$$

where $s^2 = \frac{s_1^2+s_2^2}{m+n-2}$.

6.5 Self-Assessment Exercise

1. Let $(X_1, X_2, \ldots, X_n)$ be a random sample from pdf given below. Find UMP test of size α for testing H_0 against H_1 mentioned against each problem at level α. Also draw the curve of power function $\beta(\theta)$.

(i) $f(x;\theta) = N(0,\theta)$, $H_0: \theta \leq \theta_0$ against $H_1: \theta > \theta_0$

(ii) $f(x;\theta) = N(0,\theta)$, $H_0: \theta \geq \theta_0$ against $H_1: \theta < \theta_0$

(iii) $f(x;\theta) = N(\theta,1)$, $H_0: \theta \geq \theta_0$ against $H_1: \theta < \theta_0$

(iv) $f(x;\theta) = \theta x^{\theta-1},\ 0 < x < 1, \theta > 0;$

$= 0$ otherwise

(a) $H_0: \theta \leq \theta_0$ against $H_1: \theta > \theta_0$

(b) $H_0: \theta \leq \theta_0$ against $H_1: \theta > \theta_0$

(v) $f(x;\theta) = G(1,\theta),$

(a) $H_0: \theta \leq \theta_0$ against $H_1: \theta > \theta_0$

(b) $H_0: \theta \geq \theta_0$ against $H_1: \theta < \theta_0$

2. Let (X_1, X_2, X_3) be a random sample from $b(1,\theta)$. Find UMP test of size $\frac{1}{9}$ for testing $H_0: \theta \leq \frac{1}{3}$ against $H_1: \theta > \frac{1}{3}$. Also find the power function of the test.

UNIT 7: INTERVAL ESTIMATION

Structure

7.1 Introduction

Confidence estimation provides a range (confidence interval) within which a parameter is likely to lie, reflecting the uncertainty of the estimate. It offers precision and reliability by quantifying the estimation's accuracy and conveying variability. Confidence intervals aid in informed decision-making across fields like medicine, economics, and engineering, by assessing the certainty of estimates. They also facilitate comparative analysis, enabling the evaluation of different parameter estimates and supporting hypothesis testing. Overall, confidence estimation enhances the robustness of statistical inferences and ensures more reliable conclusions from data analysis.

7.2 Objective

The objective of this unit is to provide us a basic understanding of the concepts related to interval estimation has been put forth & its various results have been discussed. Further, the concepts of uniformly most accurate, shortest length confidence intervals have been derived & discussed extensively. The relationship between UMP test and UMA confidence intervals has been discussed. These concepts should be clear after reading this material.

7.3 Introduction to Confidence Estimation

Let X be an RV, and a, b be two given positive real numbers.

Then, consider the following probability statement

$$P\{a < X < b\} = P\{a < X \text{ and } X < b\}$$

$$= P\left\{\frac{bX}{a} > b \text{ and } X < b\right\}$$

$$= P\left\{X < b < \frac{bX}{a}\right\},$$

and if we know the distribution of X and a, b, 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 endpoints that are functions of the RV X, and hence it takes the value $\left(x, \frac{bx}{a}\right)$ when X takes the value x. In other words, $I(X)$ assumes the value $I(x)$ whenever X assumes the value x. Thus $I(X)$ is a random quantity and is a random *interval*. Note that I(X) includes the value b with a certain fixed probability. For example, if $b = 1, a = \frac{1}{2}$ and X is $U(0,1)$, the interval $(X, 2X)$ includes point 1 with probability $\frac{1}{2}$. We note that $I(X)$ is a family of intervals with associated coverage probability $P(I(X) \ni 1) = \frac{1}{2}$. It has (random) length of the interval, the larger the coverage probability. Let us formalize these notations.

Definition: Let $P_{\boldsymbol{\theta}}$, $\boldsymbol{\theta} \in \Theta \subseteq \boldsymbol{\mathcal{R}_k}$, be the set of probability distributions of an *RV* $\boldsymbol{X}$. A family of subsets $S(\boldsymbol{x})$ of Θ, where $S(\boldsymbol{x})$ depends on the observation $\boldsymbol{x}$ but not on $\boldsymbol{\theta}$, is called a *family of random sets*.

If, in particular, $\Theta \subseteq \mathcal{R}$ and $S(\boldsymbol{x})$ is an interval $\left(\underline{\theta}(\boldsymbol{x}), \bar{\theta}(\boldsymbol{x})\right)$, where $\underline{\theta}(\boldsymbol{x})$ and $\bar{\theta}(\boldsymbol{x})$ are functions of $\boldsymbol{x}$ alone(and not θ), we call $S(\boldsymbol{X})$ a random interval with $\underline{\theta}(\boldsymbol{X})$ *and* $\bar{\theta}(\boldsymbol{X})$ as lower and upper bounds, respectively.$\bar{\theta}(\boldsymbol{X})$ may be $-\infty$, and $\bar{\theta}(\boldsymbol{X})$ may be $+\infty$.

We are interested in the problem of *confidence estimation*, namely, that of finding a family of random sets $S(\boldsymbol{x})$ for a parameter $\boldsymbol{\theta}$ such that for a given $\alpha, 0 < \alpha < 1$(usually small),

$P_{\boldsymbol{\theta}}\{S(\boldsymbol{X}) \ni \boldsymbol{\theta}\} \geq 1 - \alpha$ for all $\boldsymbol{\theta} \in \Theta$. (1)

We restrict our attention mainly to the case where $\theta \in \Theta \subseteq \mathcal{R}$.

Definition: A family of subsets $S(\boldsymbol{x})$ of $\Theta \subseteq \mathcal{R}_k$ is said to constitute a family of confidence sets at confidence level $1-\alpha$ if

$$P_{\boldsymbol{\theta}}\{S(\boldsymbol{X}) \ni \boldsymbol{\theta}\} \geq 1-\alpha \text{ for all } \boldsymbol{\theta} \in \Theta, \qquad (2)$$

That is, the random set $S(\boldsymbol{X})$ covers the true parameter value $\boldsymbol{\theta}$ with probability $\geq 1-\alpha$.

If $S(\boldsymbol{x})$ is of the form

$$S(\boldsymbol{x}) = \left(\underline{\theta}\,(\boldsymbol{x}), \bar{\theta}(\boldsymbol{x})\right) \qquad (3)$$

We will call it a confidence interval at confidence level $1-\alpha$, provided that

$$P_{\theta}\left\{\underline{\theta}(\boldsymbol{X}) < \theta < \bar{\theta}(\boldsymbol{X})\right\} \geq 1-\alpha \text{ for all } \theta, \qquad (4)$$

and the quantity

$$\inf_{\theta} P_{\theta}\{\theta(\boldsymbol{X}) < \theta < \bar{\theta}(\boldsymbol{X})\} \qquad (5)$$

will be referred to as the confidence coefficient associated with the random interval.

Remark: We write $S(\boldsymbol{X}) \ni \boldsymbol{\theta}$ to indicate that $\boldsymbol{X}$, and hence $S(\boldsymbol{X})$, is random here and not $\boldsymbol{\theta}$, so the probability distribution referred to is that of $\boldsymbol{X}$.

Remark: When $\boldsymbol{X} = \boldsymbol{x}$ is the realization, the confidence interval (set) $S(\boldsymbol{x})$ is a fixed subset of $\mathcal{R}_k$.No probability is attached to $S(\boldsymbol{x})$ itself since neither $\boldsymbol{\theta}$ nor $S(\boldsymbol{x})$ has a probability distribution. In fact, either $S(\boldsymbol{x})$ covers $\boldsymbol{\theta}$ or it does not, and we will never know which since $\boldsymbol{\theta}$ is unknown. One can give a relative frequency interpretation. If $(1-\alpha)$-level confidence sets for $\boldsymbol{\theta}$ were computed a large number of times, a fraction (approximately) $1-\alpha$ of those would contain the true (but unknown) parameter value.

Definition: A family of $(1-\alpha)$-level confidence sets $\{S(\boldsymbol{x})\}$ is said to be a **UMA** family of confidence sets at level $1-\alpha$ if

$$P_{\boldsymbol{\theta}}\{S(\boldsymbol{X}) \ \ contains\ \boldsymbol{\theta}'\} \leq P_{\theta}\{S'(\boldsymbol{X}).\ contains\ \boldsymbol{\theta}'\}$$

for all $\boldsymbol{\theta} \neq \boldsymbol{\theta}' and\ any\ (1-\alpha)$ level family of confidence sets $S'(\boldsymbol{X})$.

Theorem: Let $(X_1, X_2, \ldots, X_n)$ be a random sample from a distribution function $F_\theta(.)$, $\theta \varepsilon \Theta$, where Θ is an interval of real line $\boldsymbol{R}$. Let $T(X,\theta) = T(X_1, X_2, \ldots, X_n; \theta)$ be a real valued function defined on $\boldsymbol{R_n}$ X Θ, such that , for each θ, $T(X,\theta)$ is a statistic, and as a function of θ, T is strictly increasing or decreasing at every $x = (x_1, x_2, \ldots, x_n) \varepsilon \boldsymbol{R_n}$. Let $\Lambda \subseteq \boldsymbol{R}$ be the range of T, and for every $\lambda\ \varepsilon\ \Lambda$ and $x\ \varepsilon\ \boldsymbol{R_n}$, let the equation $\lambda = T(x,\theta)$ be solvable. If the distribution of $T(X,\theta)$ is independent of θ, one can construct a confidence interval for θ at any level.

Proof: Let $0 < \alpha < 1$. Then we can choose a pair of numbers $\lambda_1(\alpha)$ and $\lambda_2(\alpha)$ in Λ not necessarily unique, such that

$$P_\theta[\lambda_1(\alpha) < T(X,\theta) < \lambda_2(\alpha)] \geq (1-\alpha) \quad \text{for all } \theta \varepsilon \Theta \qquad (6)$$

Since the distribution of $T(X,\theta)$ is independent of , it is clear that $\lambda_1\ and\ \lambda_2$ are independent of θ. Since., moreover, T is monotone in θ, we can solve the equations

$$T(x,\theta) = \lambda_1(\alpha) \text{ and}$$

$$T(x,\theta) = \lambda_2(\alpha)$$

for every x uniquely for θ so that we have

$$P_\theta\left[\underline{\theta}(X) < \theta < \overline{\theta}(X)\right] \geq (1-\alpha) \text{ for all } \theta \varepsilon \Theta \text{ ,}$$

where $\underline{\theta}(X)$ and $\overline{\theta}(X)$ $\left(\underline{\theta}(X) < \overline{\theta}(X)\right)$ are random variables.

Remark: The condition that $\lambda = T(x,\theta)$ be solvable will be satisfied if, for example, T is continuous and strictly increasing or decreasing function of θ in Θ .

Remark: It is usually possible to find a function T that is monotone and has a distribution independent of θ. For example, if F_θ is continuous and monotone in Θ , we can take

$$T(X,\theta) = \prod_{i=1}^{n} F_\theta(X_i)$$

Since F_θ is continuous, $F_\theta(X_i)$ are i.i.d. with $(0,1)$. Then

$$-\log T(X,\theta) = -\sum_{i=1}^{n} \log F_\theta(X_i)$$

where $-\log F_\theta(X_i)$ are i.i.d. with $G(1,1)$ distribution. It follows $-\log T(X,\theta)$ is distributed as $G(n,1)$, and we can find λ_1 and λ_2 such that

$$P_\theta[-\log\lambda_2 < -\log T(X,\theta) < -\log\lambda_1] = \frac{1}{\Gamma(n)}\int_{-\log\lambda_2}^{-\log\lambda_1} x^{n-1}\ e^{-x}\ dx$$

$$= 1-\alpha$$

Thus,

$$P_\theta[\lambda_1 < T(X,\theta) < \lambda_2] = 1-\alpha$$

or

$$P_\theta\left[\underline{\theta}(X) < \theta < \overline{\theta}(X)\right] = 1-\alpha$$

Note that, in the continuous case, we can find a confidence interval with equality on the right side of (6). In the discrete case, however, this is usually not possible.

Remark: Relation (6) is valid even if the assumption of monotonicity condition of T in the Theorem is dropped. In that case, the inequalities may yield a set of intervals (or random set) $S(X)$ in Θ instead of a confidence interval.

Example: Let $X_1, X_2, \ldots, X_n$ be iid Rvs, $X_i \sim N(\mu,\sigma^2)$. Consider the interval $(\bar{X}-c_1, \bar{X}+c_2)$. In order for this to be a $(1-\alpha)$ level confidence interval, we must have

$$P\{\bar{X}-c_1 < \mu < \bar{X}+c_2\} \geq 1-\alpha,$$

which is same as

$$P\{\mu-c_2 < \bar{X} < \mu+c_1\} \geq 1-\alpha.$$

Thus

$$P\left\{-\frac{c_2}{\sigma}\sqrt{n} < \frac{\bar{X}-\mu}{\sigma}\sqrt{n} < \frac{c_1}{\sigma}\sqrt{n}\right\} \geq 1-\alpha.$$

Since $\sqrt{n}\,(\bar{X}-\mu)/\sigma \sim N(0,1)$. We can choose c_1 and c_2 to have equality,namely,

$$P\left\{-\frac{c_2}{\sigma}\sqrt{n} < \frac{\bar{X}-\mu}{\sigma}\sqrt{n} < \frac{c_1}{\sigma}\sqrt{n}\right\} \geq 1-\alpha.$$

provided that σ is known. There are infinitely many such pairs of values (c_1, c_2). In particular, an intuitively reasonable choice is $c_1 = -c_2 = c$, say. In that case

$$\frac{c\sqrt{n}}{\sigma} = z_{\alpha/2,}$$

and the confidence interval is $(\bar{X} - \frac{\sigma}{\sqrt{n}} z_{\frac{\alpha}{2}}, \bar{X} + \frac{\sigma}{\sqrt{n}} z_{\frac{\alpha}{2}})$. The length of this interval is $2\frac{\sigma}{\sqrt{n}} z_{\alpha/2.}$ Given σ and α, we can choose n to get a confidence interval of a fixed length.

If σ is not known, we have from

$$P\{-c_2 < \bar{X}-\mu < c_1\} \geq 1-\alpha$$

that

$$P\left\{-\frac{c_2}{S}\sqrt{n} < \frac{\bar{X}-\mu}{S/\sqrt{n}}\sqrt{n} < \frac{c_1}{S}\sqrt{n}\right\} \geq 1-\alpha.$$

And once again we can choose pairs of values(c_1, c_2) using a t-distribution with $n-1$ d.f. such that

$$P\left\{-\frac{c_2\sqrt{n}}{S} < \frac{\bar{X}-\mu}{S}\sqrt{n} < \frac{c_1\sqrt{n}}{S}\right\} = 1-\alpha.$$

In particular, if we take $c_1 = -c_2 = c, say, then$

$$c\frac{\sqrt{n}}{S} = t_{n-1,\alpha/2},$$

and $(\bar{X} - \left(\frac{S}{\sqrt{n}}\right) t_{n-1,\frac{\alpha}{2}}, \bar{X} + \left(\frac{S}{\sqrt{n}}\right) t_{n-1,\frac{\alpha}{2}})$, is a $1-\alpha$-level confidence interval for μ. The length of this interval is $(2S/\sqrt{n})t_{n-1,\alpha/2,}$ which is no longer constant.

Therefore, we cannot choose n to get a fixed-width confidence interval of level $1-\alpha$. Indeed, the length of this interval can be quite large if σ is large. Its expected length is

$$\frac{2}{\sqrt{n}} t_{n-1,\alpha/2}\, E_\sigma S = \frac{2}{\sqrt{n}} t_{n-1,\alpha/2} \sqrt{\frac{2}{n-1}}\, \frac{\Gamma(n/2)}{\Gamma[(n-1)/2]} \sigma,$$

which can be made as small as small as we please by choosing n large enough.

Example: In previous example, suppose that we wish to find a confidence interval for σ^2 instead when μ is unknown. Consider the interval $(c_1 S^2, c_2 S^2), c_1, c_2 > 0$. We have

$$P\{c_1 S^2 < \sigma^2 < c_2 S^2\} \geq 1-\alpha,$$

so that

$$\left\{c_2^{-1} < \frac{S^2}{\sigma^2} < c_1^{-1}\right\} \geq 1-\alpha.$$

Since $(n-1)S^2/\sigma^2$ is $\chi^2(n-1)$, we can choose pairs of values (c_1, c_2) from the tables of the chi-square distribution. In particular, we can choose c_1, c_2 so that

$$P\left\{\frac{S^2}{\sigma^2} \geq \frac{1}{c_1}\right\} = \frac{\alpha}{2} = P\left\{\frac{S^2}{\sigma^2} \leq \frac{1}{c_2}\right\}.$$

Then

$$\frac{n-1}{c_1} = \chi^2_{n-1,\alpha/2}\ and \frac{n-1}{c_2} = \chi^2_{n-1,1-\alpha/2.}$$

Thus

$$\left(\frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}}, \frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}}\right)$$

is a $(1-\alpha)$-level confidence interval for σ^2 whenever μ is unknown. If μ is known, then

$$\sum_{1}^{n} \frac{(X_i-\mu)^2}{\sigma^2} \sim \chi^2(n).$$

Thus we can base the confidence interval on $\sum_1^n (X_i-\mu)^2$. Proceeding similarly, we get a $(1-\alpha)$ level confidence interval as

$$\left(\frac{\sum_1^n (X_i-\mu)^2}{\chi^2_{n,\frac{\alpha}{2}}}, \qquad \frac{\sum_1^n (X_i-\mu)^2}{\chi^2_{n,1-\frac{\alpha}{2}}} \right).$$

Remark: If a sufficiently large sample size is available and the condition of Central Limit Theorem holds, then it is possible to construct approximate confidence intervals of any desired level.

Example: Let $(X_1, X_2, \dots, X_n)$ be a random sample from a distribution with finite variance σ^2. Let $E(X) = \theta, E(X^2) = \theta^2 + \sigma^2$. Find the approximate confidence interval for θ of level $1-\alpha$.

Since from C.L.T. it follows that

$$Y_n = \frac{\bar{X}-\theta}{\sigma}\sqrt{n} \xrightarrow{L} Z$$

where Z is distributed with $(0,1)$.

Suppose we want $1-\alpha$ level confidence interval for θ when σ is not known. We know that $\frac{\sqrt{n}(\bar{X}-\theta)}{S}$ is approximately normal $N(0,1)$.

Hence for large n, we can find constants C_1 and C_2 such that

$$P\left[C_1 < \frac{\sqrt{n}(\bar{X}-\theta)}{S} < C_2\right] = 1-\alpha$$

In particular $-C_1 = C_2 = z_{\alpha/2}$ to give

$$\left(\bar{X} - \frac{S}{\sqrt{n}} z\alpha_{/2}, \bar{X} + \frac{S}{\sqrt{n}} z\alpha_{/2}\right)$$

as an approximate $1-\alpha$ level confidence interval for θ.

Remark: Yet another possible procedure has usual applicability and hence can be based for large or small samples. Unfortunately, however, this procedure usually yields confidence intervals that are much too large. The method employs well known Chebyshev's Inequality

$$P\left[|X - E(X)| < \epsilon \sqrt{Var(X)}\right] > 1 - \frac{1}{\epsilon^2}.$$

Example: Let $(X_1, X_2, \ldots, X_n)$ be a random sample from $U(0, \theta)$. Find $1-\alpha$ confidence interval of .

Let $M_n = \max_i X_i = max(X_1, X_2, \ldots, X_n)$

The pdf of $Y = M_n$ is given by

$$f(y) = \frac{ny^{n-1}}{\theta^n} \quad ; 0 < y < \theta$$

$$= 0 \quad ; \text{otherwise}$$

Let $T(M_n, \theta) = \frac{M_n}{\theta}$. Then pdf of T is given by

$$f(t) = nt^{n-1} \; ; \; 0 < t < 1$$

$$= 0 \quad ; \text{otherwise}$$

which is independent of θ. Also T is monotone in θ. By previous theorem , we can find $\lambda_1(\alpha)$ and $\lambda_2(\alpha)$ such that

$$P[\lambda_1(\alpha) < T < \lambda_2(\alpha)] = 1 - \alpha \text{ for all } \theta$$

or

$$n \int_{\lambda_1(\alpha)}^{\lambda_2(\alpha)} t^{n-1} dt = 1 - \alpha$$

or

$$\lambda_2^n(\alpha) - \lambda_1^n(\alpha) = 1 - \alpha$$

This equation has infinitely many solutions. If we choose $\lambda_2(\alpha) = 1, \text{then } \lambda_1(\alpha) = \alpha^{\frac{1}{n}}$ and $1 - \alpha$ confidence interval is obtained by

$$T(M_n, \theta) = \frac{M_n}{\theta} = \lambda_1(\alpha) = \alpha^{\frac{1}{n}}$$

$$\Rightarrow \breve{\theta} = \frac{M_n}{\alpha^{\frac{1}{n}}}$$

and

$$T(M_n, \theta) = \frac{M_n}{\theta} = \lambda_2(\alpha) = 1 \Rightarrow \breve{\theta} = M_n$$

Thus,

$$P\left[\lambda_1(\alpha) < \frac{M_n}{\theta} < \lambda_2(\alpha)\right] = 1 - \alpha \ \ for\ all\ \theta$$

$$P\left[\frac{M_n}{\lambda_2(\alpha)} < \theta < \frac{M_n}{\lambda_1(\alpha)}\right]$$

the interval is

$$\left(\frac{M_n}{\lambda_2(\alpha)}, \frac{M_n}{\lambda_1(\alpha)}\right) = \left(M_n, M_n(\alpha)^{-\frac{1}{n}}\right)$$

Now, we use the Chebyshev's Inequality to find approximate confidence interval. We know that

$$E(M_n) = \frac{n}{n+1}\theta,$$

$$E(M_n - \theta)^2 < Var(M_n) \text{ and}$$

$$E(M_n - \theta)^2 = \frac{2\theta^2}{(n+1)(n+2)}$$

so that by using Chebychev's Inequality, we have

$$P\left[|M_n - E(M_n)| < \epsilon\sqrt{E(M_n-\theta)^2}\right] \geq P\left[|M_n - E(M_n)| < \epsilon\sqrt{Var(M_n)}\right]$$

$$> 1 - \frac{1}{\epsilon^2}$$

or

$$P\left[|M_n - E(M_n)| < \epsilon\sqrt{E(M_n-\theta)^2}\right] > 1 - \frac{1}{\epsilon^2}$$

$$\Rightarrow P\left[\left|M_n - \frac{n\theta}{n+1}\right| < \epsilon\sqrt{\frac{2\theta^2}{(n+1)(n+2)}}\right] > 1 - \frac{1}{\epsilon^2} \Rightarrow$$

$$P\left[\left(\frac{n+1}{n}\right)M_n - \epsilon\sqrt{E(M_n-\theta)^2} < \theta < \left(\frac{n+1}{n}\right)M_n + \epsilon\sqrt{E(M_n-\theta)^2}\right]$$

$$> 1 - \frac{1}{\epsilon^2}$$

$\Rightarrow$ The interval is $\left(\left(\frac{n+1}{n}\right)M_n - \epsilon\theta\sqrt{\frac{2}{(n+1)(n+2)}}, \left(\frac{n+1}{n}\right)M_n + \epsilon\theta\sqrt{\frac{2}{(n+1)(n+2)}}\right)$ For large n, $\left(\frac{n+1}{n}\right)M_n \cong M_n$ and $\sqrt{\frac{1}{(n+1)(n+2)}} \cong \frac{1}{n}$. We have $M_n \xrightarrow{P} \theta$ and replacing θ by M_n, we have the interval

$$\left(M_n - \epsilon\, M_n\frac{\sqrt{2}}{n}, M_n + \epsilon\, M_n\frac{\sqrt{2}}{n}\right)$$

Letting $1 - \frac{1}{\epsilon^2} = 1 - \alpha \Rightarrow \epsilon = \frac{1}{\sqrt{\alpha}}$, we have the interval

$$\left(M_n - \frac{M_n}{n}\sqrt{\frac{2}{\alpha}}, M_n + \frac{M_n}{n}\sqrt{\frac{2}{\alpha}}\right)$$

Also, $M_n \leq \theta$, the lower limit may be taken as M_n . Hence the interval is

$$\left(M_n, M_n + \frac{M_n}{n}\sqrt{\frac{2}{\alpha}}\right) = \left(M_n, M_n\left(1+\frac{1}{n}\right)\sqrt{\frac{2}{\alpha}}\right)$$

an approximate confidence interval of level $1-\alpha$.

Shortest- Length Confidence Interval

For a given a given confidence level $1-\alpha$ a wide choice of confidence intervals is available. Clearly, the larger the interval, the better is the chance of trapping a true parameter value. Thus, the interval $(-\infty, \infty)$ which ignores the data completely will include the real-valued parameter θ with confidence level 1. However, the larger the confidence interval, the less meaningful it is. Therefore, for a given confidence level $1-\alpha$, it is desirable to choose the shortest possible confidence interval.

For $T(X, \theta)$, we use the simplest random variable that is a function of a sufficient statistic and parameter θ.

Definition: A random variable (X, θ) , which is a function of $\boldsymbol{X} = \left(X_{1,} X_{2,} \ldots, X_n\right)$ and θ, whose distribution is independent of θ, is called a Pivot.

Remark: An alternative to minimize the length of the confidence interval is to minimize the expected length $E_\theta\left[\overline{\theta}(X) - \underline{\theta}(X)\right]$. Unfortunately, this also is quite unsatisfactorily since, in general, there does not exist a member in the class of all $1-\alpha$ level confidence intervals that minimizes $E_\theta\left[\overline{\theta}(X) - \underline{\theta}(X)\right]$ for all θ .The procedures applied in finding the shortest-length confidence interval based on a pivot are also applicable in finding an interval that minimizes the expected length. We remark here that the restriction to unbiased confidence intervals is natural if we wish to minimize $E_\theta\left[\overline{\theta}(X) - \underline{\theta}(X)\right]$.

Definition: A family $\{S(X)\}$ of confidence sets for parameter θ is said to be unbiased at confidence level $1-\alpha$ if

$P_\theta[S(X) \text{ contains } \theta] \geq 1-\alpha$ and

$P_{\theta'}[S(X) \text{ contains } \theta] \leq 1-\alpha$ for all θ and $\theta' \varepsilon\, \Theta$.

Example: Let $(X_1, X_2, \ldots, X_n)$ be a random sample from $U(\theta, \theta+1)$. Find $1-\alpha$ confidence intervals of θ.

Let us first consider

$$Y = M_n = \max_i X_i = max(X_1, X_2, \ldots, X_n)$$

so that

$$F_Y(y) = [F_X(y)]^n = (y-\theta)^n$$

Let $Z = (Y-\theta)^n \sim U(0,1)$ as Z is a distribution function.

Therefore,

$$P[a < (Y-\theta)^n < b] = P[a < Z < b] = b - a = 1-\alpha$$

$$\frac{db}{da} = 1 \qquad (1)$$

$$P\left[a^{\frac{1}{n}} < (Y-\theta) < b^{\frac{1}{n}}\right] = b-a$$

Thus, the interval is $\left(Y - b^{\frac{1}{n}}, Y - a^{\frac{1}{n}}\right)$ and the length of the interval is $L = b^{\frac{1}{n}} - a^{\frac{1}{n}}$

$$\frac{dL}{da} = \frac{1}{n}\left[b^{\frac{1-n}{n}}\frac{db}{da} - a^{\frac{1-n}{n}}\right] \qquad (2)$$

From (1) and (2) we get that $\frac{dL}{da} > 0 \Rightarrow L \uparrow a$

Thus, L is minimum when $a = 0$. Therefore, $b = 1-\alpha$.

Hence, the required confidence interval of level $1-\alpha$ is

$$\left(Y - (1-\alpha)^{\frac{1}{n}}, Y\right) = \left(M_n - (1-\alpha)^{\frac{1}{n}},\ M_n\right)$$

Let us now consider

$W = T_n = \min_i X_i = min(X_1, X_2, \dots, X_n)$

so that

$F_W(w) = 1 - (\theta + 1 - w)^n$

Let $Z = 1 - (\theta + 1 - W)^n \sim U(0,1)$ as Z is a distribution function.

Therefore,

$$P[a < 1 - (\theta + 1 - W)^n < b] = P[a < Z < b] = b - a = 1 - \alpha$$

$$\frac{db}{da} = 1$$

$$P\left[(1-b)^{\frac{1}{n}} < (\theta + 1 - W) < (1-a)^{\frac{1}{n}}\right] = b - a$$

$$P\left[W - 1 + (1-b)^{\frac{1}{n}} < \theta < W - 1 + (1-a)^{\frac{1}{n}}\right] = b - a = 1 - \alpha$$

Thus the interval is $\left(W - 1 + (1-b)^{\frac{1}{n}}, W - 1 + (1-a)^{\frac{1}{n}}\right)$ and the length of the interval is $L = (1-a)^{\frac{1}{n}} - (1-b)^{\frac{1}{n}}$

$\frac{dL}{db} = \frac{1}{n}\left[(1-b)^{\frac{1-n}{n}} - (1-a)^{\frac{1-n}{n}} \frac{da}{db}\right]$

From (1) and (2) we get that $\frac{dL}{db} < 0 \Rightarrow L \downarrow b$

Thus, L is minimum when b= 1. Therefore a= α. Hence, the required confidence interval of level $1 - \alpha$ is

$\left(W - 1, (1-\alpha)^{\frac{1}{n}} - 1 + W\right) = \left(T_n - 1, (1-\alpha)^{\frac{1}{n}} - 1 + T_n\right)$

Example: Let $(X_1, X_2, \dots, X_n)$ be a random sample from $U(0, \theta)$. Find $1 - \alpha$ confidence intervals of θ.

Let us consider

$Y = M_n = \max_i X_i = max(X_1, X_2, \ldots, X_n)$ which is sufficient for θ and its pdf is given by

$$f(y) = \frac{ny^{n-1}}{\theta^n} \quad ; 0 < y < \theta$$

$$= 0 \quad ; otherwise$$

Let $T(M_n, \theta) = \frac{M_n}{\theta} = \frac{Y}{\theta} = T$

$\Rightarrow f_T(t) = nt^{n-1}$; $0 < t < 1$

$= 0$; otherwise

Using T as Pivot, we see that

$$P[a < T < b] = b^n - a^n = 1 - \alpha$$

Or $P\left[\frac{Y}{b} < \theta < \frac{Y}{a}\right] = 1 - \alpha$ (3)

Thus, $\left(\frac{Y}{b}, \frac{Y}{a}\right) = \left(\frac{M_n}{b}, \frac{M_n}{a}\right)$ is confidence interval of length

$L = Y\left[\frac{1}{a} - \frac{1}{b}\right]$ (4)

We minimize L given by (4) subject to condition (3).

$$\frac{dL}{db} = Y\left[\frac{1}{b^2} - \frac{1}{a^2}\frac{da}{db}\right]$$

Again, from (3), we have

$$\frac{d}{db}[b^n - a^n] = 0$$

$$\Rightarrow \frac{da}{db} = \frac{b^{n-1}}{a^{n-1}}$$

$$\Rightarrow \frac{dL}{db} = Y\left[\frac{a^{n+1} - b^{n+1}}{b^2 a^{n+1}}\right] < 0 \text{ as } a < b$$

Thus $L \downarrow b$ and the minimum occurs if $b = 1$. Thus, the shortest confidence interval is

$$\left(M_n, \frac{M_n}{a}\right)$$

Now $b = 1 \Rightarrow a = (\alpha)^{\frac{1}{n}}$. Therefore, the shortest length of interval is $\left(M_n, M_n(\alpha)^{-\frac{1}{n}}\right)$. We note that $L \mapsto 0 \; as \; n \mapsto \infty$ and

$E_\theta(L) = \left[(\alpha)^{-\frac{1}{n}} - 1\right]\frac{n\theta}{n+1}$.

Example: Let $\left(X_1, X_2, \ldots, X_n\right)$ be a random sample from $N(\theta, \sigma^2)$, where σ^2 is known. Find shortest length confidence interval of θ at level $1 - \alpha$.

The Pivot in this case is

$$T(\bar{X}, \theta) = \frac{(\bar{X} - \theta)\sqrt{n}}{\sigma}$$

Then

$$P[a < T(\bar{X}, \theta) < b] = 1 - \alpha$$

or

$$\Phi(b) - \Phi(a) = 1 - \alpha \qquad (4)$$

where $\Phi(.)$ is the distribution function of $N(0, 1)$. Therefore,

$$P\left[\bar{X} - \frac{b\sigma}{\sqrt{n}} < \theta < \bar{X} - \frac{a\sigma}{\sqrt{n}}\right] = 1 - \alpha$$

The interval is $\left(\bar{X} - \frac{b\sigma}{\sqrt{n}}, \bar{X} - \frac{a\sigma}{\sqrt{n}}\right)$.

$$L = \frac{\sigma}{\sqrt{n}}[b - a]$$

We minimize L subject the condition (4)

$$\frac{dL}{da} = \frac{\sigma}{\sqrt{n}}\left[\frac{db}{da} - 1\right] = 0$$

From (4)

$$f(b)\frac{db}{da} - f(a) = 0 \Rightarrow \frac{db}{da} = \frac{f(a)}{f(b)}$$

$$\frac{dL}{da} = 0 \Rightarrow f(b) = f(a), \text{where } f(.) \text{ is pdf of } N(0,1).$$

$$\Rightarrow a = -b \text{ as } f(.) \text{ is symmetric about zero.}$$

Therefore, if $b = z_{\alpha/2} = -a$, where $\int_{z_{\alpha/2}}^{\infty} f(u)du = {}^{\alpha}/_{2}$, the confidence interval is

$\left(\bar{X} - \frac{\sigma z_{\alpha/2}}{\sqrt{n}}, \bar{X} + \frac{z_{\alpha/2}\sigma}{\sqrt{n}}\right)$ $\qquad L = \frac{2\sigma z_{\alpha/2}}{\sqrt{n}}$

In this case we can plan our experiment to give a prescribed confidence level and prescribed length for the interval. To have level $1-\alpha$ and length $\leq 2\alpha$, we choose sample size such that

$\frac{2\sigma z_{\alpha/2}}{\sqrt{n}} \leq$2d

$\Rightarrow n \geq \frac{z^2_{\alpha/2}\sigma^2}{d^2}$.

We can interpret this as follows: If estimate θ by $\bar{X}$, taking a sample size $n \geq \frac{z^2_{\alpha/2}\sigma^2}{d^2}$, we have $(1-\alpha)$ percent confidence that the error in our experiment is at the most d

Example: Let $(X_1, X_2, \ldots, X_n)$ be a random sample from $N(\theta, \sigma^2)$, where σ^2 is unknown. Find shortest length confidence interval of θ at level $1-\alpha$.

The Pivot in this case is

$$T(\bar{X}, \theta) = \frac{(\bar{X} - \theta)\sqrt{n}}{S}, S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

$$T(\bar{X}, \theta) = Y \sim t_{n-1}$$

Thus

$$P[a < Y < b] = 1 - \alpha$$

$$\int_a^b f_{n-1}(t)\, dt = 1 - \alpha,$$

where$f_{n-1}(t)$ is pdf of $t-$ distribution with $n-1$ degrees of freedom.

$$P\left[\bar{X} - b\frac{S}{\sqrt{n}} < \theta < \bar{X} - a\frac{S}{\sqrt{n}}\right] = 1 - \alpha$$

Thus, the interval is $\left(\bar{X} - b\frac{S}{\sqrt{n}}, \bar{X} - a\frac{S}{\sqrt{n}}\right)$.

$$L = \frac{S}{\sqrt{n}}(b - a)$$

We have $\frac{dL}{da} = 0 \Rightarrow \frac{db}{da} = 1$

Again

$$\frac{d}{da}\int_a^b f_{n-1}(t)\, dt = 0$$

$$\Rightarrow \frac{db}{da} f_{n-1}(b) - f_{n-1}(a) = 0$$

$$\Rightarrow f_{n-1}(b) = f_{n-1}(a)$$

$$\Rightarrow a = -b$$

as $f_{n-1}(t)$ is symmetric about zero. Therefore, $b = t_{n-1}(\alpha/2)$, where

$\int_{t_{n-1}(\alpha/2)}^{\infty} f_{n-1}(t)dt = \alpha/2$.

Therefore, we have the confidence interval of level $1 - \alpha$ as

$$\left(\bar{X} - t_{n-1}(\alpha/2)\frac{S}{\sqrt{n}}, \bar{X} + t_{n-1}(\alpha/2)\frac{S}{\sqrt{n}}\right)$$

$L = 2t_{n-1}(\alpha/2)\frac{S}{\sqrt{n}}$, which is a random variable may be arbitrary large.

Example: Let $(X_1, X_2, \ldots, X_n)$ be a random sample from $N(\theta, \sigma^2)$, where θ is known. Find shortest length confidence interval of σ^2 at level $1-\alpha$.

The Pivot in this case is

$$T(X, \sigma^2) = \frac{\sum_{i=1}^{n}(X_i - \theta)^2}{\sigma^2} \sim \mathcal{X}^2_{(n)}$$

$$P[a < T(X, \sigma^2) < b] = 1 - \alpha \tag{5}$$

$$P\left[\frac{\sum_{i=1}^{n}(X_i - \theta)^2}{b} < \sigma^2 < \frac{\sum_{i=1}^{n}(X_i - \theta)^2}{a}\right] = 1 - \alpha$$

Confidence interval is

$$\left(\frac{\sum_{i=1}^{n}(X_i - \theta)^2}{b}, \frac{\sum_{i=1}^{n}(X_i - \theta)^2}{a}\right)$$

Length of interval is $L = (\sum_{i=1}^{n}(X_i - \theta)^2)\left[\frac{1}{a} - \frac{1}{b}\right]$.

From (5), we have

$$\int_a^b f_n(t)dt = 1 - \alpha \tag{6}$$

where $f_n(t)$ is the pdf of $\mathcal{X}^2_{(n)}$ -distribution.

$$\frac{dL}{da} = \left(\sum_{i=1}^{n}(X_i - \theta)^2\right)\left[\frac{1}{b^2}\frac{db}{da} - \frac{1}{a^2}\right]$$

From (6), we have

$$\frac{db}{da} f_n(b) - f_n(a) = 0$$

$$\Rightarrow \frac{db}{da} = \frac{f_n(a)}{f_n(b)}$$

and

$$\frac{dL}{da} = 0$$

$$\Rightarrow \frac{f_n(a)}{f_n(b)} = \frac{b^2}{a^2}$$

$$\Rightarrow a^2 f_n(a) = b^2 f_n(b)\ .$$

Thus we have a number of shortest confidence intervals satisfying the condition $a^2 f_n(a) = b^2 f_n(b)$.

In practice, we take equal tails interval i.e.

$$b = \mathcal{X}^2_{(n)}\left(\alpha/2\right), a = \mathcal{X}^2_{(n)}\left(1 - \left(\alpha/2\right)\right)$$

where $\int_{\mathcal{X}^2_{(n)}(\alpha/2)}^{\infty} f_n(t)dt = \alpha/2$ and $\int_{\mathcal{X}^2_{(n)}\left(1-(\alpha/2)\right)}^{\infty} f_n(t)dt = 1 - \left(\alpha/2\right)$.

Remark: When θ is unknown, then the Pivot is

$T(X, \sigma^2) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{\sigma^2} = \frac{(n-1)S^2}{\sigma^2} \sim \mathcal{X}^2_{(n-1)}$

In this case the confidence intervals of shortest length will be

$\left(\frac{(n-1)S^2}{\mathcal{X}^2_{(n-1)}(\alpha/2)}, \frac{(n-1)S^2}{\mathcal{X}^2_{(n-1)}\left(1-(\alpha/2)\right)}\right)$ of equal tails.

7.4 Relationship Between UMP Tests and UMA Confidence Intervals

We next consider the method of test inversion and explore the relationship between a test of hypothesis for a parameter θ and confidence interval for θ.Consider the following example.

Example: Let $X_1, X_2, \ldots, X_n$ be a sample from $N(\mu, \sigma_o^2)$ where σ_0 is known.

It can be easily shown that

$$\left(\bar{X} - \frac{1}{\sqrt{n}} z_{\frac{\alpha}{2}}\sigma_0, \bar{X} + \frac{1}{\sqrt{n}}\ z_{\frac{\alpha}{2}}\sigma_0\right)$$

is a $(1-\alpha)$ level confidence interval for μ. If we define a test ϕ that rejects a value of $\mu = \mu_0$ if and only if μ_0 lies outside this interval; that is, if and only if

$$\frac{\sqrt{n}|\bar{X} - \mu_0|}{\sigma_0} \geq z_{\alpha/2,}$$

then

$$P_{\mu_0}\left\{\frac{\sqrt{n}|\bar{X} - \mu_0|}{\sigma_0} \geq z_{\alpha/2}\right\} = \alpha,$$

and the test ϕ is a size α test of $\mu = \mu_0$ against the alternatives $\mu \neq \mu_0$.

Conversely, a family of α-level tests for the hypothesis $\mu = \mu_0$ generates a family of confidence intervals for μ by simply taking, as the confidence interval for μ_0, the set of those μ for which one cannot reject $\mu = \mu_0$.

Similarly, we can generate a family of α-level tests from a $(1-\alpha)$-level lower (or upper) confidence bound. Suppose that we start with the $(1-\alpha)$-level lower confidence bound $\bar{X} - z_{\alpha(\sigma_0/\sqrt{n})}$ for μ. Then, by defining a test $\phi(\boldsymbol{X})$ that rejects $\mu \leq \mu_0$ if and only if $\mu_0 < \bar{X} - z_{\alpha(\sigma_0/\sqrt{n}),}$ we get an α-level test for a hypothesis of the form $\mu \leq \mu_0$.

Theorem: Let $A(\theta_0), \theta_0 \in \Theta$, denote the region of acceptance of an α-level test of $H_0(\theta_0)$. For each observation $x = (x_1, x_2, \ldots, x_n)$, Let $S(\boldsymbol{x})$ denote the set

$$S(\boldsymbol{x}) = \{\theta : \boldsymbol{x} \in A(\theta), \theta \in \Theta\}. \qquad (6)$$

Then $S(\boldsymbol{x})$ is a family of confidence sets for θ at confidence level $1-\alpha$. If, moreever, $A(\theta_0)$ is UMP for the problem $\left(\alpha, H_0(\theta_0), H_1(\theta_0)\right)$, then $S(\boldsymbol{X})$ minimizes

$$P_\theta\{S(\boldsymbol{X}) \ni \theta'\} \quad for\ all\ \theta \in H_1(\theta') \qquad (7)$$

among all $(1-\alpha)$-level families of confidence sets. That is, $S(\boldsymbol{X})$ is UMA.

Proof: Since, we have

$$S(\boldsymbol{x}) \ni \theta \quad if\ and\ only\ if\ \boldsymbol{x} \in A(\theta), \qquad (8)$$

so that

$P_\theta\{S(\boldsymbol{X}) \ni \theta\} = P_\theta\{\boldsymbol{X} \in A(\theta)\} \geq 1 - \alpha,$

as asserted. If $S^*(\boldsymbol{X})$ is any other family of $(1-\alpha)$ level confidence sets, let $A^*(\theta) = \{x: S^*(x) \ni \theta\}$. Then

$$P_\theta\{\boldsymbol{X} \in A^*(\theta)\} = P_\theta\{S^*(\boldsymbol{X}) \ni \theta\} \geq 1 - \alpha;$$

and since $A(\theta_0)$ is UMP for $\big(\alpha, H_0(\theta_0), H_1(\theta_0)\big)$, it follows that

$$P_\theta\{\boldsymbol{X} \in A^*(\theta_0)\} \geq P_\theta\{\boldsymbol{X} \in A(\theta_0)\} \ \ for\ any\ \theta \in H_1(\theta_0).$$

Hence

$$P_\theta\{S^*(\boldsymbol{X}) \ni \theta_0\} \geq P_\theta\{\boldsymbol{X} \in A(\theta_0)\} = P_\theta\{S(\boldsymbol{X}) \ni \theta_0\}$$

For all $\theta \in H_1(\theta_0)$. This completes the proof.

7.5 Self-Assessment Exercise

1. Let $\big(X_1, X_2, \ldots, X_n\big)$ be a random sample from the following pdf's. Find a shortest length confidence interval of level $(1-\alpha)$ for the parameter θ involved in the pdf.

(i) $f_\theta(x) = U(\theta, 1)$

(ii) $f_\theta(x) = e^{-(x-\theta)}$ if $x > \theta$

$= 0$ otherwise

(iii) $f_\theta(x) = G(1, \theta)$

2. Let $\big(X_1, X_2, \ldots, X_n\big)$ be a random sample from $N(\theta, \sigma^2)$, where θ and σ^2 are unknown. Find shortest length confidence interval of (θ, σ^2) at level $1-\alpha$.

U.P. Rajarshi Tandon Open University, Prayagraj

MScSTAT – 102N / MASTAT – 102N
Statistical Inference

Course Design Committee

Prof. Ashutosh Gupta — **Chairman**
Director, School of Sciences,
U. P. Rajarshi Tandon Open University, Prayagraj

Prof. Anup Chaturvedi (Retd.) — **Member**
Ex. Head, Department of Statistics,
University of Allahabad, Prayagraj

Prof. S. Lalitha (Retd.) — **Member**
Ex. Head, Department of Statistics,
University of Allahabad, Prayagraj

Prof. Himanshu Pandey — **Member**
Department of Statistics,
D. D. U. Gorakhpur University, Gorakhpur

Prof. Shruti — **Member-Secretary**
Professor, School of Sciences,
U.P. Rajarshi Tandon Open University, Prayagraj

Course Preparation Committee

Dr. Pratyasha Tripathi — **Writer**
Department of Statistics (Unit 8, 9, 10 & 11)
Tilka Manjhi Bhagalpur University, Bhagalpur, Bihar

Dr. Prabhat Kr. Singh — **Writer**
Department of Statistics (Unit 12)
Jamshedpur Co-Operative College, Jamshedpur

Dr. Sunit Kumar — **Writer**
Department of Statistics (Unit 13)
Central University of South Bihar, Gaya, Bihar

Prof. Shruti — **Editor**
School of Sciences **(**Unit 8, 9, 10, 11, 12 & 13)
U. P. Rajarshi Tandon Open University, Prayagraj

Prof. Shruti — **Course Coordinator**
School of Sciences
U. P. Rajarshi Tandon Open University, Prayagraj

MScSTAT – 102N/ MASTAT – 102N STATISTICAL INFERENCE

First Edition: ***July 2024***
ISBN : 978-93-48987-21-1

Printed and Published by Mr. Vinay Kumar, Registrar, Uttar Pradesh Rajarshi Tandon Open University, 2025.

Block & Units Introduction

The ***Block - 3 – Testing of Hypotheses & Confidence Estimation*** is the second block with four units.

In ***Unit – 8 –Methods of Estimation – I*** & ***Unit – 9 –Methods of Estimation – II,*** deals with the Maximum likelihood estimation, method of moments, MVUE, necessary and sufficient conditions for MVUE, etc., Zehna theorem for invariance, Cramer theorem for weak consistency. Cramer-Huzurbazar theorem

In ***Unit – 10 – Criterion for Good Estimators – I*** & ***Unit – 11 – Criterion for Good Estimators - II*** discussed about the Criterion for Good Estimators, Bhattacharya bound, Chapman Robbins and Kiefer (CRK) bound, asymptotic normality, BAN and CAN estimators, asymptotic efficiency, equivariant consistency.

In ***Unit – 12 – Confidence Estimation*** discussed about the Confidence interval and confidence coefficient, shortest length confidence interval, relation between confidence estimation and hypotheses testing.

In ***Unit – 13 – Hypothesis Testing*** defined about the Generalized Neyman Pearson lemma, MP and UMP tests for distributions with MLR, LR tests and their properties, UMPU tests, similar regions, Neyman structure, Invariant tests.

At the end of every block/unit the summary, self-assessment questions and further readings are given.

UNIT 8: METHODS OF ESTIMATION-I

Structure

8.1 Introduction

Estimation theory is a fundamental pillar of statistical analysis, dealing with the process of deducing population parameters from sample data. In real-world scenarios, it is often impractical to have complete population data due to constraints like time, cost, or accessibility. Instead, we rely on samples, which offer only partial information. Estimation methods are used to infer unknown population parameters, such as the mean or variance, based on the available data.

The historical development of estimation theory can be traced back to the late 19th and early 20th centuries with foundational contributions by statisticians like Karl Pearson and Sir Ronald A. Fisher. Karl Pearson pioneered the **Method of Moments (MoM)**, one of the earliest formal estimation techniques, where parameters are estimated by equating sample moments to theoretical moments of a distribution. This method provided a simple yet powerful approach for parameter estimation in various scientific applications, especially in biological and social sciences.

However, it was Sir Ronald A. Fisher who revolutionized the field by introducing the **Maximum Likelihood Estimation (MLE)** method in the early 20th century. MLE marked a significant advancement, as it systematically identified parameter values that maximized the likelihood of observing the given sample data, making it a robust and versatile tool in

statistical inference. Fisher's contributions also laid the groundwork for modern inferential techniques, as he demonstrated that under regularity conditions, MLE estimators are consistent, asymptotically normal, and efficient.

Further refinements in estimation theory came with contributions from statisticians like C.R. Rao, who developed the **Cramér-Rao Lower Bound (CRLB)**, which establishes a lower bound for the variance of unbiased estimators. This theoretical result gave statisticians a way to judge the efficiency of an estimator, ensuring that no unbiased estimator could have a variance smaller than the CRLB. Over time, the **Lehmann-Scheffé theorem** and the **Rao-Blackwell theorem** were introduced to improve the precision of unbiased estimators and identify those with minimum variance.

These historical developments culminated in the formation of modern estimation theory, which is now used extensively in fields such as economics, medicine, engineering, and environmental sciences. Whether estimating the average income of a population or determining the failure rate of a machine component, the choice of estimation method can significantly affect the accuracy and reliability of inferences.

This unit begins by introducing two major estimation techniques: **Maximum Likelihood Estimation (MLE)** and the **Method of Moments (MoM)**. These methods form the foundation for understanding more advanced estimation concepts, such as the **Minimum Variance Unbiased Estimator (MVUE)** and various theorems that provide the necessary conditions for optimal estimation. Throughout this unit, we explore the theoretical underpinnings of these methods, their historical context, and their practical applications, equipping learners with the tools needed to make reliable statistical inferences.

8.2 Objectives

By the end of this unit, learners will be able to:

1. **Understand the Role of Estimation in Statistical Analysis**:
    - Comprehend the importance of estimation in making inferences about population parameters based on sample data, especially when complete data is unavailable.
2. **Apply Maximum Likelihood Estimation (MLE)**:

- Understand the principles behind Maximum Likelihood Estimation and its application in parameter estimation.
- Be able to formulate and solve likelihood functions to determine MLEs for different types of data.

3. **Apply the Method of Moments (MoM)**:
    - Learn how the Method of Moments works by equating sample moments with theoretical moments.
    - Apply MoM to estimate population parameters in cases where MLE may be difficult to use or intractable.
4. **Compare and Contrast MLE and MoM**:
    - Evaluate the strengths and limitations of both Maximum Likelihood Estimation and Method of Moments in terms of bias, consistency, and efficiency.
5. **Understand and Identify Minimum Variance Unbiased Estimators (MVUE)**:
    - Learn the importance of unbiasedness and minimum variance in selecting an optimal estimator.
    - Recognize the conditions under which an estimator is considered MVUE and its relevance in ensuring precise and reliable estimations.
6. **Analyse the Cramér-Rao Lower Bound (CRLB)**:
    - Understand how the Cramér-Rao Lower Bound provides a benchmark for the variance of unbiased estimators.
    - Explore how CRLB is used to judge the efficiency of different estimators.
7. **Explore the Historical Development of Estimation Methods**:
    - Gain an understanding of how estimation methods evolved through the contributions of Karl Pearson, Ronald Fisher, C.R. Rao, and others.
8. **Solve Practical Problems Using Estimation Techniques**:
    - Apply estimation methods to real-world data to solve problems and make inferences about population parameters.
    - Work through practical examples involving MLE and MoM to reinforce understanding of theoretical concepts.

By mastering these objectives, learners will develop a comprehensive understanding of key estimation methods, allowing them to apply these concepts in statistical analysis and real-world problem solving.

8.3 Maximum Likelihood Estimation (MLE)

The Maximum Likelihood Estimation (MLE) method is one of the most widely used techniques in statistical inference for estimating the parameters of a model. The central idea behind MLE is to find the parameter values that maximize the likelihood of observing the given sample data. In other words, MLE selects the parameter values that make the observed data most probable under the assumed statistical model.

MLE has become a cornerstone of modern statistics due to its flexibility, general applicability, and asymptotic properties. This section delves into the principles of MLE, its mathematical formulation, and the conditions that guarantee its optimal performance in large samples.

The estimation process begins with defining a likelihood function, which represents the probability of observing the sample data given a set of parameters. By maximizing this likelihood function, we obtain the parameter estimates that best explain the observed data. MLE is particularly valuable in practical applications, as it is adaptable to a wide range of statistical models, including complex and high-dimensional settings

The roots of Maximum Likelihood Estimation trace back to the early 20th century, with **Sir Ronald A. Fisher** as its principal architect. Fisher introduced the MLE method in a series of pioneering papers starting in 1912 and further developed it in his seminal works during the 1920s. Before Fisher's innovations, statisticians primarily relied on methods like the Method of Moments (introduced by Karl Pearson) to estimate population parameters. However, these earlier methods lacked the general applicability and theoretical foundation that MLE provided.

Fisher’s breakthrough came with the realization that likelihood could be maximized to provide estimates that were both intuitive and theoretically sound. He demonstrated that, under certain regularity conditions, MLE estimators possess desirable asymptotic properties: they are **consistent** (converging to the true parameter value as the sample size increases),

efficient (having the smallest possible variance among all unbiased estimators in large samples), and **asymptotically normal**.

Fisher's MLE theory marked a shift in statistical thinking, bridging the gap between practical data analysis and rigorous mathematical theory. It revolutionized how statisticians approach parameter estimation, providing a foundation for modern statistical methods. Over the years, MLE has been extended and refined, influencing fields as diverse as economics, biology, and machine learning.

Mathematically, let us say we have a random sample $X_1, X_2, \ldots, X_n$ from a distribution with probability density function (pdf) or probability mass function (pmf) $f(x;\theta)\ where\ \theta$ is an unknown parameter. The likelihood function $L(x;\theta)$ is given by

$$L(x;\theta) = \prod_{i=1}^{n} f(x_i;\theta)$$

The MLE $\hat{\theta}$ is the value of θ that maximises $L(x;\theta)$;

$$\hat{\theta} = \arg max_{\theta} L(x;\theta)$$

Maximum Likelihood Estimation (MLE) is a widely used method for estimating the parameters of statistical models. The likelihood function captures the probability of observing the given sample, given a set of parameters. The MLE is the parameter value that maximizes this likelihood function.

There are several regularity conditions (often termed necessary and sufficient conditions) that are generally assumed to ensure that the MLE has good properties, such as consistency and asymptotic normality. However, it is important to note that these conditions do not guarantee the existence or uniqueness of an MLE, but they are conditions under which the MLE behaves well asymptotically.

Necessary and Sufficient Conditions for MLE to be Asymptotically Normal and Consistent:

The Maximum Likelihood Estimator (MLE) is one of the most widely used methods for parameter estimation due to its desirable properties, especially in large samples. Two critical

properties of MLE are **consistency** and **asymptotic normality**. These properties ensure that as the sample size increases, the MLE converges to the true parameter value (consistency), and the distribution of the MLE approaches a normal distribution (asymptotic normality).

1. ***Consistency of MLE:*** An estimator $\widehat{\theta_n}$ of parameter θ is said to be consistent if $\widehat{\theta_n}$ converges in probability to the true value θ as the sample size *n* tends to infinity.

$$\widehat{\theta_n} \xrightarrow{P} \theta \; as \; n \; \to \infty.$$

Necessary and Sufficient Condition for Consistency:

The MLE $\widehat{\theta_n}$ is consistent if the following conditions are satisfied:

1. **Identifiability:** The parameter θ must be identifiable from the probability distribution $f(x;\theta)$. This means that for every distinct θ_1 and θ_2 in the parameter space, the probability distributions they generate must also be distinct. Formally, for all $\theta_1 \neq \theta_2$ we must have

$$P(X \in A;\, \theta_1) \neq P(X \in A;\, \theta_2), for\ some\ event\ A.$$

2. Existence of MLE: For each sample, the MLE must exist, *i.e.* the likelihood function must have a maximiser for every sample size *n*.

3. Log-Likelihood Function Smoothness: The log-likelihood function $l(\theta) = \log L(\theta)$, where $L(\theta)\ is\ the\ likelihood\ function,$ must be differentiable with respect to θ and the first and second derivatives of $l(\theta)$ must exist and be continuous in θ.

4. Fisher Information: The Fisher information matrix $I(\theta)$must be finite and positive definite at the true parameter value θ. Formally,

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

Must be finite and positive definite.

5. Law of Large Numbers (LLN) for the Score Function: The score function must satisfy the law of large numbers. This means that as *n* tends to infinity.

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\log f(X_i;\theta)\xrightarrow{P}0$$

Where $X_i's$ are independently and identically distributed random variables.

Proof: (Consistency)

To prove the consistency of MLE, we start by showing that the likelihood function is maximized at the true parameter value θ_0.

1. Log-Likelihood Maximization: By the Law of Large Numbers, as $n \to \infty$ the average log-likelihood converges to its expectation:

$$\frac{1}{n}l(\theta)\xrightarrow{P}E[\log(X;\theta)]$$

the true parameter value θ_0 maximises the expected log-likelihood because:

$$E[\log\ \mathrm{f}(X;\ \theta_0)] \geq E[\log(X;\theta)]\ for\ all\ \theta \neq \theta_0$$

2. Score Function Convergence: The score function $S(\theta) = \frac{\partial}{\partial\theta}\log L(\theta)$, must satisfy $S(\theta_0) = 0$. Asymptotically, ensuring that the likelihood is maximised at θ_0.

3. Identifiability Ensures Uniqueness: Because θ_0 is identifiable, the likelihood function peaks uniquely at θ_0, ensuring that the MLE converges in probability to the true value θ_0 as n tends to infinity.

Thus, the MLE is consistent.

Asymptotic Normality of MLE

The MLE $\hat{\theta}_n$ is said to be asymptotically normal if the distribution of $\hat{\theta}_n$ approaches a normal distribution as the sample size increases. Specifically:

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \xrightarrow{D} N(0, I(\theta_0)^{-1})$$

Where $I(\theta_0)$ is the Fisher information at the true parameter θ_0

Necessary and Sufficient Conditions for Asymptotic Normality

The MLE $\widehat{\theta_n}$ is asymptotically normal if the following conditions are satisfied:

1. Second-Order Differentiability: the log likelihood function $l(\theta)$ must be twice continuously differentiable with respect to θ in the neighbourhood of the true parameter value θ_0.

2. Fisher Information Regularity: the Fisher Information matrix I(θ) must be positive definite and finite at the true parameter value θ_0. This ensures that the likelihood function has a quadratic shape around θ_0 which leads to normality.

3. Central Limit Theorem (CLT) for the Score Function: The score function (the first derivative of the log-likelihood) must satisfy the central limit theorem. That is, as *n* tends to infinity.

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\log f(X_i;\theta_0) \xrightarrow{D} N(0, I(\theta_0))$$

4. Law of Large Numbers for The Second Derivative: The second derivative of the log-likelihood (Hessian matrix) must satisfy the law of large numbers. Specifically, the negative Hessian evaluated at the MLE must converge in probability to the Fisher information matrix. Formally,

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial\theta^2}\log f(X_i;\theta) \xrightarrow{P} -I(\theta_0)$$

Proof of Asymptotic Normality:

The proof of asymptotic normality relies on applying a Taylor series expansion of the log-likelihood function around the true parameter θ_0.

Taylor Expansion: begin by expanding the log=likelihood around θ_0 up to the second order:

$$l(\theta) \approx l(\theta_0) + \left(\hat{\theta}_n - \theta_0\right)\frac{\partial}{\partial\theta} l(\theta_0) + \frac{1}{2}\left(\hat{\theta}_n - \theta_0\right)^2 \frac{\partial^2}{\partial\theta^2} l(\theta_0) \ .$$

Score Function: the first derivative (score function) at θ_0 is asymptotically normally distributed by the central limit theorem:

$$S(\theta_0) = \frac{\partial}{\partial\theta} l(\theta_0) \sim N\ (0, nI(\theta_0))$$

Normality Derivation: by solving the score equation $S\left(\hat{\theta}_n\right) = 0$, using the Taylor expansion, we get the approximation:

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \approx \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial}{\partial\theta} \log f(X_i; \theta_0) \xrightarrow{D} N(0, I(\theta_0))$$

Which shows that $\hat{\theta}_n$ is asymptotically normal.

The uniqueness of the Maximum Likelihood Estimator (MLE) is not guaranteed in general, but under certain conditions, it can be. The conditions for the uniqueness of the MLE, along with a simplified proof, are as follows:

Theorem (Uniqueness of MLE):

Suppose we have a random sample $X_1, X_2, \ldots, X_n$ from a distribution with probability density function (pdf) or probability mass function (pmf) $f(x; \theta)\ where\ \theta$ is an unknown parameter in the parameter space Θ. If the log-likelihood function $l(\theta)$ is strictly concave over Θ, then the MLE $\hat{\theta}$ is unique.

Proof:

Given the log-likelihood function $l(\theta)$, is the natural logarithm of the likelihood function. Our goal is to find θ that maximises $l(\theta)$.

1. A necessary condition for $l(\theta)$ to have a maximum at $\hat{\theta}$ is that it's first derivative with respect to θ is zero i.e.

$$\frac{\partial}{\partial\theta} l(\theta) = 0.$$

2. The log-likelihood function $l(\theta)$ is strictly concave over Θ, if it's second derivative is strictly negative everywhere in Θ that is

$$\frac{\partial^2}{\partial\theta^2} l(\theta) < 0 \text{ for all } \theta \; in \; \Theta.$$

Now, assuming the log-likelihood function is twice differentiable and the second derivative is strictly negative, then $l(\theta)$ is strictly concave.

For a strictly concave function, any critical point (where the first derivative is zero) is a unique global maximum. Thus, under the given conditions, the MLE $\hat{\theta}$ which maximizes $l(\theta)$ is unique.

It should be noted that the conditions provided are sufficient for the uniqueness of MLE but are not necessary. There exist situations where the MLE is unique, even if the conditions are not met. However, this theorem offers a clear and generalizable guideline for when the MLE is guaranteed to be unique.

Properties of MLE

1. ***Invariance Property of MLE:*** This property states that if $\hat{\theta}$ is MLE of $\theta \; and \; g\,(\theta)$ is a function, then $g(\widehat{\theta})$ is MLE of $g\,(\theta)$, provided that $g(.)$ is a continuous function.

2. ***Consistency of MLE:*** under certain regulatory conditions, as the sample size n approaches infinity, the MLE converges in probability to the true parameter value.

3. ***Asymptotic Normality of MLE:*** under certain conditions, the MLE is approximately normally distributed around the true parameter value with a variance that decreases as the sample-size increases.

Example: Suppose $X_1, X_2, \ldots, X_n$ are independently and identically distributed random variables from a normal distribution with known variance σ^2 and unknown mean μ. Find MLE of μ.

Solution: Given the pdf of the normal distribution:

$$f(x;\mu) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

the likelihood function is:

$$L(\mu;x) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

Taking log likelihood:

$$l(\mu;x) = -\frac{n}{2}\log(2\pi) - nlog\ (\sigma) - \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2}$$

Differentiating $l(\mu;x)$ with respect to μ and setting it to zero, we find:

$$\hat{\mu} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Thus, the sample mean is the MLE of the population mean. This is just a brief overview of the MLE with some foundational theorems and an example.

Example: Fish in a Pond

Scenario: Imagine you are trying to estimate the percentage of goldfish in a pond that also contains several other types of fish. You catch 10 fish at random, and 7 out of them are goldfish.

Application of MLE: Based on this sample, which estimate of the percentage of goldfish in the pond would make your observed sample (7 out of 10 are goldfish) most probable or "likely"? Intuitively, the maximum likelihood estimate would be 70% goldfish.

Example: Guessing the Contents of a Cookie Jar

You have a cookie jar filled with chocolate chip and oatmeal cookies. Without looking, you draw 5 cookies, and 4 of them are chocolate chip.

Application of MLE: If you had to guess the proportion of chocolate chip cookies in the jar based on the idea of making your observed sample most "likely", you would probably estimate that the jar is 80% chocolate chip cookies.

Example: Detecting a Loaded Die

You are trying to determine if a six-sided die is fair or loaded towards a specific number. You roll the die 60 times, and it shows the number "6" on 20 of those rolls.

Application of MLE: Based on your observations, if you had to guess the probability of the die landing on "6", which estimate would make your observed results the most "likely"? An intuitive maximum likelihood estimate would be that the die has a 1/3 chance of landing on "6", given your sample.

In these non-numerical examples, the underlying idea of MLE is to choose parameter values that make the observed data most probable. In formal statistical settings, this involves setting up a likelihood function based on the data and the model, and then finding the parameter values that maximize this function.

8.4 Method of Moments

The **Method of Moments (MoM)** is one of the earliest techniques developed for estimating population parameters from sample data. The method involves equating sample moments, such as the mean or variance, with their corresponding theoretical moments derived from a probability distribution, and then solving for the unknown parameters. This approach is simple, intuitive, and widely applicable, making it a useful tool in both classical and modern statistics.

The central idea of the Method of Moments is that the moments of a distribution—such as the mean, variance, skewness, and kurtosis—are unique to that distribution. Therefore, by matching the sample moments to the theoretical moments of a given distribution, we can obtain parameter estimates for the population. For example, if we want

to estimate the parameters of a normal distribution, we equate the sample mean and variance to the theoretical mean and variance of the normal distribution, and solve for the parameters. This section explores the principles behind MoM, its application, and its strengths and limitations in comparison to other methods like Maximum Likelihood Estimation (MLE).

The Method of Moments was introduced by the British statistician **Karl Pearson** in the late 19th century. Pearson, who is often regarded as one of the founders of modern statistics, developed MoM to fit distributions to data. His work laid the foundation for statistical inference, and the Method of Moments became widely adopted due to its simplicity and ease of application in various fields of science.

Before the advent of more complex and computationally intensive methods like Maximum Likelihood Estimation, MoM was the dominant technique for parameter estimation. Pearson's development of MoM was closely tied to his work in curve fitting and distribution theory, particularly in biology and the social sciences, where he applied statistical methods to real-world data.

Despite its early success, the Method of Moments was gradually supplemented (and in many cases replaced) by more efficient methods, particularly MLE, which was developed by **Ronald A. Fisher** in the early 20th century. Fisher's work showed that MLE had better statistical properties, such as consistency and efficiency, under certain regularity conditions. However, MoM remains an important and accessible tool for parameter estimation, especially in cases where MLE is difficult to compute or where sample moments provide meaningful estimates.

The Method of Moments is based on the principle that the moments of a distribution can be expressed as functions of the unknown parameters. By equating the empirical moments (computed from the sample) with the theoretical moments (based on the distribution), we can derive equations that can be solved for the unknown parameters.

Definition: For the random variable X, the k-th moment about the origin is defined as :

$$\mu_k' = E[X^k]$$

The kth central-moment is defined as:

$$\mu_k = E\left[\left(X - E(X)\right)^k\right]$$

For a sample $x_1, x_2, \ldots, x_n$ the k-th central moment about the origin is:

$$m'_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k$$

the $k-th$ sample central moment is :

$$m_k = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^k$$

where $\bar{x}$ is sample mean.

The Method:

1. For a given probability distribution with parameters $\theta_1, \theta_2, \ldots, \theta_p$ express the first p moments (either about the origin or central moments) in terms of these parameters.
2. Equate these theoretical moments to the corresponding sample moments.
3. Solve the resulting system of equations for the parameters.

Consistency and Asymptotic Results of the Method of Moments (MoM)

The **Method of Moments (MoM)** is a classical technique for parameter estimation, and it shares many of the desirable properties of other estimation methods, such as **consistency** and **asymptotic normality**. Below, we will explore these properties in detail, providing formal statements and proofs.

Theorem: (Consistency of MoM) Let $X_1, X_2, \ldots, X_n$be independent and identically distributed (i.i.d.) random variables with a probability distribution $f(X;\theta)$ that depends on an unknown parameter θ. Let $\hat{\theta}_n$ be the estimator of θ derived using the Method of Moments by equating the sample moments m'_k to the theoretical moments μ'_k. If the sample moments converge in probability to the theoretical moments, i.e.,

$$m'_k \xrightarrow{P} \mu'_k \; for \; k = 1,2, \ldots, p$$

Then the method of moments estimator $\hat{\theta}_n$ is consistent for n tending to infinity.

Proof: we know that the sample moments m'_k converges in probability to the corresponding theoretical moments μ'_k for each k. By the definition

$$m'_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k$$

By the law of large numbers, as n tends to infinity,

$$m'_k \xrightarrow{P} \mu'_k = E[X^k]$$

Therefore, the sample moments provide consistent estimates of the theoretical moments.

The MoM constructs estimators by equating the sample moments to the theoretical moments, which are functions of the unknown parameter θ. The system of equations is given as

$$m'_k = \mu'_k(\theta); k = 1,2, \dots, p$$

As the sample moments m'_k converge in probability to $\mu'_k(\theta)$, the solution of these, moment equations that is the MoM estimator $\hat{\theta}_n$, must converge to the true parameter value θ, provided that the moment equations are invertible.

For the estimator $\hat{\theta}_n$ to be consistent, the system of equations derived from the moments must be uniquely solvable for θ. This requires the model to be identifiable, that means different values of θ should yield different values for the theoretical moments. Under this condition the solution $\hat{\theta}_n$ to the system of moment equations will converge to the true parameter θ.

Thus, as sample size increases, $\hat{\theta}_n \xrightarrow{P} \theta$, providing the consistency of the MoM estimators.

Asymptotic Normality: Under further regularity conditions, the method of moments estimators are asymptotically normal, meaning they can be approximated by a normal distribution for large sample sizes.

Theorem: Let $X_1, X_2, \dots, X_n$be independent and identically distributed (i.i.d.) random variables with a probability distribution $f(X;\theta)$ that depends on an unknown parameter θ. Let $\hat{\theta}_n$ be the MoM estimator of θ, and let the true value of the parameter be θ_0. Under regularity condition, as n tends to infinity,

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \xrightarrow{D} N(0, \Sigma)$$

Where $\Sigma = J^{-1}VJ^{-1}$; V is asymptotic variance covariance matrix of the sample moments and J is the Jacobian matrix of the partial derivatives of the theoretical moments with respect to the parameters.

Proof:

1. By the Central Limit Theorem (CLT), the sample moments are asymptotically normally distributed. Specifically, for each $k = 1, 2, \dots, p$ we have

$$\sqrt{n}(m'_k - \mu'_k) \xrightarrow{D} N(0, \sigma_k^2)$$

Where σ_k^2 *is the variance of kth sample moment.*

Since the sample moments are functions of i.i.d. random variables $X_1, X_2, \dots, X_n$ the vector of sample moments $\left(m'_1, m'_2, \dots, m'_p\right)$ follows a multivariate normal distribution in large samples

$$\sqrt{n}(m'_k - \mu'_k) \xrightarrow{D} N(0, V)$$

Where V is asymptotic variance-covariance matrix of the sample moments.

2. **Delta Method for Parameter Estimators:** The MoM estimators $\hat{\theta}_n$ *are* obtained by solving the moment equations

$$m'_k = \mu'_k(\theta); k = 1,2, \dots, p$$

We can apply delta method to obtain the asymptotic distribution of $\hat{\theta}_n$. This method is used to approximate the distribution of a function of a random variable, and in this case, the function is the inverse of the moment equations.

Let $g(\theta)$ be the vector of the theoretical moments $\mu_k'(\theta), and\ let\ J = \frac{\partial g}{\partial \theta}$ be the Jacobian matrix of partial derivatives of $g(\theta)$ w.r.t θ, by delta method we have,

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \xrightarrow{D} N(0, J^{-1}VJ^{-1})$$

Where $J\ is\ evaluated\ at\ \theta_0$.

3. **Asymptotic Variance:** the asymptotic variance of the MoM estimator is given by $\Sigma = J^{-1}VJ^{-1}$, which combines the variability of the sample moments with the sensitivity of the moments to changes in the parameters.

Thus, the Method of Moments estimator is asymptotically normally distributed, with the variance covariance matrix $\Sigma = J^{-1}VJ^{-1}$.

Let us look at some theoretical examples to understand how the Method of Moments (MoM) works in practice.

Example: suppose $X_1, X_2, \ldots, X_n$ is a random sample from an exponential distribution with probability density function;

$$f(x|\lambda) = \lambda e^{-\lambda x}; for\ x \geq 0\ and\ \lambda > 0$$

The expected value or first moment of the exponential distribution is;

$$E[X] = \frac{1}{\lambda}$$

Using MoM, we equate the sample mean to the theoretical mean to estimate λ

$$\overline{x} = \frac{1}{\lambda_{MoM}}$$

So,

$$\lambda_{MoM} = \frac{1}{\overline{x}}$$

Example: Weighing Mystery Boxes

Scenario: Imagine you are at a warehouse where boxes of the same size are filled with items of varying weights. You know the boxes are either filled with stones or feathers, but they are all sealed. Your goal is to estimate the average weight of the stones in a box without opening them.

Application of MoM:

Randomly pick a few boxes and weigh each one.

Assume the weight of a box with feathers is negligible.

Calculate the average weight of the boxes you picked. This average weight is your MoM estimate for the average weight of a box filled with stones.

Example: Guessing the Height of Students

Scenario: Imagine you are a teacher trying to guess the average height of the students in your school, but you do not want to measure each student's height.

Application of MoM:

You randomly select a few students from different grades and measure their heights.

You then calculate the average height of this sample.

This average height becomes your MoM estimate for the average height of all students in the school.

Example: Estimating Average Reading Time

Scenario: You are a librarian, and you want to know how long, on average, a visitor spends reading in the library.

Application of MoM:

- Over a few days, you discreetly observe and note down the reading time of several visitors.
- Calculate the average reading time from your observations.
- This average time is your MoM estimate for the average reading time of a visitor to the library.

In each of these non-numerical examples, the core idea remains the same: obtain a sample, calculate its moments (like the mean), and use these moments as estimates for the population parameters. This approach is analogous to the way MoM is applied in a statistical context.

Strengths and Limitations of Method of Moments:

The Method of Moments is valued for its simplicity and computational ease, especially when theoretical moments are easily derivable. It is particularly useful in cases where MLE is intractable or computationally expensive. However, MoM does have certain limitations:

- **Efficiency**: MoM estimators are not always efficient, meaning they may not have the smallest possible variance. In contrast, MLE often provides more efficient estimators, particularly in large samples.
- **Bias**: MoM estimators can sometimes be biased, particularly in small samples, as they do not inherently minimize bias or variance.
- **Robustness**: MoM may be less robust than MLE in certain cases, particularly when the data exhibits extreme values or non-standard behaviour.

Historical Significance and Contemporary Use:

Although the Method of Moments was historically a pioneering technique, it has since been supplemented by more sophisticated methods, such as MLE. However, MoM continues to play an important role in parameter estimation in certain applications. For example, MoM is commonly used in econometrics, actuarial science, and areas where the computation of MLE may be difficult.

In recent years, MoM has also seen renewed interest in the context of **Generalized Method of Moments (GMM)**, an extension of MoM developed by **Lars Peter Hansen** in the 1980s. GMM is widely used in econometrics and finance for estimating parameters in models with more complex structures, such as those involving instrumental variables or moment conditions.

8.5 Self-Assessment Questions

1. What is the difference between point estimation and interval estimation?
2. Why is estimation crucial when dealing with incomplete data?
3. How does sampling affect the accuracy of estimation?
4. Define Maximum Likelihood Estimation (MLE) and explain the steps involved in the MLE process.
5. Under what conditions is MLE considered consistent, asymptotically normal, and efficient?
6. What is the likelihood function, and how is it used to derive parameter estimates?
7. Provide a real-world example of how MLE can be applied.
8. Explain how the Method of Moments (MoM) is used to estimate parameters.
9. Compare the computational complexity of MoM versus MLE.
10. In which situations would MoM be preferable to MLE, and why?
11. Provide an example where MoM can be applied to estimate the parameters of a distribution.
12. What are the trade-offs between using MLE and MoM in terms of bias, variance, and computational complexity?
13. Why is MLE often considered more efficient than MoM?
14. Can both methods lead to the same estimator? Provide an example where they do.
15. What makes an estimator "unbiased," and why is it important in estimation theory?
16. Explain the significance of the "minimum variance" property in the context of MVUE.
17. How does the Rao-Blackwell theorem lead to the construction of an MVUE?
18. What is the Lehmann-Scheffé theorem, and how does it relate to identifying an MVUE?
19. What is the Cramér-Rao Lower Bound (CRLB), and why is it used in estimation?
20. How does the Fisher Information relate to the CRLB?
21. Why is it impossible for any unbiased estimator to have a variance lower than the CRLB?

22. Provide a scenario where the CRLB can be applied to evaluate the efficiency of an estimator.
23. Define and explain the significance of consistency in an estimator.
24. What is asymptotic normality, and how does it help in making statistical inferences?
25. Explain the concept of efficiency and how it is assessed for different estimators.
26. What are the necessary conditions for an estimator to be considered efficient?
27. Who were the key figures in the development of modern estimation theory, and what were their contributions?
28. How did Fisher's development of MLE revolutionize statistical inference?
29. What role did the Method of Moments play in the early stages of statistical estimation theory?
30. How has estimation theory evolved to meet the needs of modern applications such as economics and machine learning?

8.6 Summary

This unit introduced the foundational principles of estimation theory, a critical area of statistical inference that allows us to make informed guesses about unknown population parameters based on sample data. Estimation plays a crucial role in fields as diverse as economics, biology, engineering, and social sciences, where complete data is often inaccessible, making it necessary to infer population characteristics from a subset of observations.

We began by exploring the two primary methods of estimation: **Maximum Likelihood Estimation (MLE)** and the **Method of Moments (MoM)**. MLE was shown to be a powerful technique for parameter estimation, where the objective is to find the parameter values that maximize the likelihood of observing the given sample data. MLE's key properties—consistency, asymptotic normality, and efficiency—make it a preferred choice in many practical applications. It ensures that, with increasing sample size, the estimated parameter converges to the true population parameter and that the estimator is as

precise as possible. MLE has its roots in the groundbreaking work of Sir Ronald A. Fisher, who revolutionized statistical theory by providing a systematic framework for estimation.

On the other hand, the Method of Moments (MoM), developed by Karl Pearson, offers a simpler alternative by equating sample moments to theoretical moments of a distribution to derive estimators. While less efficient than MLE in certain scenarios, MoM can be more straightforward to apply, particularly when the likelihood function is difficult to compute or maximize. MoM is also historically significant, having laid the groundwork for early developments in estimation theory.

The unit also delved into the concept of **Minimum Variance Unbiased Estimators (MVUE)**, which combines two desirable properties of estimators—unbiasedness and minimum variance. An MVUE provides an estimate that is both accurate (unbiased) and precise (minimum variance). We explored key theorems such as the **Rao-Blackwell** and **Lehmann-Scheffé theorems**, which offer systematic approaches for improving estimators to achieve minimum variance. The Rao-Blackwell theorem demonstrates how an unbiased estimator can be refined by conditioning on a sufficient statistic, while the Lehmann-Scheffé theorem provides the necessary and sufficient conditions for identifying an MVUE.

A critical component of this unit was the discussion of the **Cramér-Rao Lower Bound (CRLB)**, a theoretical benchmark that provides the smallest possible variance for an unbiased estimator. The CRLB is essential for assessing the efficiency of an estimator, as it sets a limit below which no unbiased estimator can go in terms of variance. Understanding the CRLB allows statisticians to evaluate whether an estimator is optimally efficient and helps guide the selection of appropriate estimation methods in practical applications.

Throughout the unit, we emphasized the historical development of these concepts, tracing how statistical estimation theory evolved through the contributions of Pearson, Fisher, Rao, and others. The progression from basic methods such as MoM to more sophisticated techniques like MLE highlights the ongoing refinement of statistical tools designed to improve the accuracy and reliability of inferences drawn from data.

In conclusion, this unit equipped learners with a deep understanding of the theoretical foundations of estimation, the practical applications of MLE and MoM, and the importance of unbiasedness, efficiency, and consistency in choosing estimators. With the knowledge of key theorems and benchmarks such as the CRLB, learners are now better positioned to apply

estimation techniques to real-world problems, making reliable and efficient statistical inferences from sample data.

8.7 References

- Casella, G., & Berger, R. L. (2002). *Statistical Inference* (2nd ed.). Duxbury Press.
- Lehmann, E. L., & Casella, G. (1998). *Theory of Point Estimation* (2nd ed.). Springer.
- Hogg, R. V., McKean, J. W., & Craig, A. T. (2018). *Introduction to Mathematical Statistics* (8th ed.). Pearson.
- Rao, C. R. (1973). *Linear Statistical Inference and Its Applications* (2nd ed.). Wiley.
- Cramér, H. (1946). *Mathematical Methods of Statistics*. Princeton University Press.

8.8 Further Readings

Here are the recommendations for further reading, along with their respective publishers:

- "Statistical Inference" by George Casella and Roger L. Berger, Duxbury Press
- "Theory of Point Estimation" by E.L. Lehmann and George Casella, Springer
- "A Course in Large Sample Theory" by Thomas S. Ferguson, Chapman & Hall/CRC.
- "Linear Statistical Inference and Its Applications" by C.R. Rao, Wiley
- "Elements of Large-Sample Theory" by E.L. Lehmann, Springer
- "Introduction to the Theory of Statistics" by Alexander M. Mood, Franklin A. Graybill, and Duane C. Boes, McGraw-Hill
- "Asymptotic Statistics" by A.W. van der Vaart, Cambridge University Press
- "All of Statistics: A Concise Course in Statistical Inference" by Larry Wasserman, Springer.

UNIT - 9: METHODS OF ESTIMATION - II

Structure

9.1 Introduction

Estimation is a central concept in statistical inference, where the objective is to deduce the value of an unknown population parameter from sample data. Two fundamental types of estimation are point estimation and interval estimation. Point estimation aims to provide a single "best guess" for the parameter, while interval estimation presents a range of plausible values for it. In this unit, we focus on advanced point estimation methods that form the bedrock of modern statistical theory.

The development of estimation theory, particularly in the 20th century, was driven by the need for rigorous methods that ensure accuracy and efficiency in decision-making across various fields, including economics, biology, and engineering. Early contributions by statisticians like Karl Pearson and Francis Galton laid the foundation for the method of moments and least squares, respectively. However, the most significant advancements came from Sir Ronald A. Fisher, who introduced the concept of Maximum Likelihood Estimation (MLE) in the 1920s. Fisher's work revolutionized estimation by offering a method that was

not only intuitive but also mathematically grounded, with optimal properties under certain conditions.

In the following decades, several other statisticians contributed to refining the theory of estimation. C.R. Rao's introduction of the Cramér-Rao Lower Bound (CRLB) in the 1940s marked a pivotal moment, establishing the minimum variance an unbiased estimator can achieve, thereby defining the concept of efficiency in estimation. Further developments such as the Rao-Blackwell and Lehmann-Scheffé theorems offered practical tools for constructing Minimum Variance Unbiased Estimators (MVUE).

Thus, modern estimation methods are built upon these historical advancements, with MVUE playing a crucial role in ensuring that estimators not only provide unbiased estimates but also have the least possible variance among all unbiased estimators. This section delves into the theoretical underpinnings, historical evolution, and significance of these estimation methods, with a particular focus on their applications in statistical inference.

9.2 Objectives

The primary aim of this unit is to provide a comprehensive understanding of advanced methods of estimation, particularly focusing on unbiasedness and efficiency, two critical properties of estimators in statistical inference. By the end of this unit, learners will be able to:

1. **Understand the Concept of Minimum Variance Unbiased Estimators (MVUE):**
    - Gain a clear understanding of what constitutes an MVUE.
    - Explore the significance of unbiasedness and minimum variance in the context of estimation.
    - Learn how MVUE offers an optimal solution by combining the two essential qualities of unbiasedness and minimum variance.

2. **Apply the Rao-Blackwell and Lehmann-Scheffé Theorems:**
    - Study and comprehend the Rao-Blackwell theorem, which provides a method to improve an unbiased estimator by conditioning on a sufficient statistic.

- Understand the Lehmann-Scheffé theorem, which offers a necessary and sufficient condition for an estimator to be MVUE.
- Utilize these theorems to derive and recognize MVUEs in real-world statistical applications.

3. **Explore Historical Developments in Estimation Theory:**
   - Learn about the historical contributions of key figures like Sir Ronald A. Fisher and C.R. Rao in developing the theoretical framework for modern estimation methods.
   - Understand how the Cramér-Rao Lower Bound (CRLB) sets a benchmark for the efficiency of unbiased estimators.
   - Appreciate the evolution of these ideas and their impact on the development of robust estimation techniques.

4. **Examine the Invariance Property of Maximum Likelihood Estimators (MLE):**
   - Study the invariance property of MLE, which states that the MLE of a function of a parameter is the same function of the MLE of the parameter.
   - Analyse the practical implications of this property for simplifying estimation in applied settings.

5. **Understand Weak Consistency and Its Role in Estimation:**
   - Learn the concept of weak consistency, which ensures that an estimator converges in probability to the true value of the parameter as the sample size increases.
   - Examine Cramér's theorem and its conditions for the weak consistency of MLEs.
   - Understand the importance of weak consistency in the reliability of estimators when dealing with large datasets.

6. **Explore the Cramér-Huzurbazar Theorem for Multivariate Estimation:**
   - Understand the Cramér-Huzurbazar theorem and its application in multivariate estimation problems.

- Learn how the theorem extends principles of consistency and asymptotic normality to multidimensional contexts, a critical aspect of modern statistical models.

7. **Apply Theoretical Concepts to Practical Problems:**
   - Use examples and exercises to apply the theoretical concepts of MVUE, MLE, and the various theorems to real-world estimation problems.
   - Develop skills to select and justify the use of appropriate estimators in different statistical scenarios, ensuring optimal and efficient inferences.

9.3 Minimum Variance Unbiased Estimators (MVUE)

In statistics, an **estimator** is a rule or formula used to estimate unknown population parameters based on sample data. Among the many criteria used to judge the quality of an estimator, the concept of a **Minimum Variance Unbiased Estimator (MVUE)** is significant. The MVUE is an estimator that satisfies two key properties: **unbiasedness** and **minimum variance** among all unbiased estimators.

Unbiasedness ensures that, on average, the estimator will give the true value of the parameter being estimated, without systematic overestimation or underestimation. Minimum variance ensures that the estimator is as precise as possible, meaning that the spread or variability of the estimator's sampling distribution is the smallest among all unbiased estimators.

In this section, we explore the theoretical underpinnings of MVUE, its historical development, and its implications for statistical inference.

The concept of an MVUE emerged from the development of **point estimation theory** in the early 20th century, primarily through the works of **Sir Ronald A. Fisher** and **C.R. Rao**. Fisher, known for his revolutionary contributions to statistical theory, laid the groundwork with the concept of **sufficiency** and introduced **Fisher's Information** as a measure of the precision of an estimator.

Later, in the 1940s, **C.R. Rao** expanded on Fisher's ideas and developed what is now known as the **Cramér-Rao Lower Bound (CRLB)**. This fundamental result provides a lower bound for the variance of any unbiased estimator, establishing a benchmark for

identifying the MVUE. An unbiased estimator that attains this bound is considered to have the smallest possible variance and is thus optimal.

E.L. Lehmann and **Joseph Scheffé** also contributed to this theory through the **Lehmann-Scheffé Theorem**, which provided a necessary and sufficient condition for an unbiased estimator to be the MVUE. These contributions collectively formed the foundation of modern estimation theory and shaped how statisticians think about optimal estimators.

An estimator $\hat{\theta}$ is called a **Minimum Variance Unbiased Estimator (MVUE)** if it satisfies the following two properties:

i. ***Unbiasedness:*** On average, the estimator gives you the correct value of the parameter you are trying to estimate. In other words, it doesn't systematically overestimate or underestimate the true value.

ii. ***Minimum Variance*****:** Among all the unbiased estimators that you could possibly use, it spreads out or varies the least from the parameter it's estimating. This means that it is the most precise or reliable unbiased estimator you can have.

To put it simply, an MVUE is like an archer who, on average, hits the bullseye (unbiasedness) and whose arrows cluster very tightly around the bullseye (minimum variance). It is essentially the "best" unbiased estimator you can get in terms of precision.

Example 1: Estimating Average Apples in a Basket

Scenario:
Imagine an orchard with many baskets of apples. Each basket has a different number of apples. You want to estimate the average number of apples in a basket.

Application of MVUE:

Unbiasedness: Suppose you have three methods:
Method A: You always pick the basket near the entrance.
Method B: You close your eyes and pick a basket at random.
Method C: You pick baskets only from the sunny side.

Of the three, Method B is unbiased because it doesn't favor any particular basket. Methods A and C could be biased because the location might influence the number of apples.

Minimum Variance: Now, among all unbiased methods, you want the one that gives the most consistent results. Suppose:

Method B1: Picking a single basket at random.
Method B2: Picking five baskets at random and averaging the count.
Method B2 will likely have a smaller spread or variance in estimates because averaging over multiple baskets reduces the chance variability. Hence, B2 would be closer to an MVUE compared to B1.

Example 2: Measuring the Height of Trees in a Forest
You want to estimate the average height of trees in a vast forest.

Application of MVUE:

Unbiasedness: Methods to measure include:
Method A: Measuring trees near the pathway.
Method B: Randomly selecting trees throughout the forest.
Method C: Only measuring trees in open clearings.

Among these, Method B is unbiased because it does not favour any specific area of the forest.

Minimum Variance: For consistency:
Method B1: Measuring a single randomly chosen tree.
Method B2: Measuring ten randomly chosen trees and taking the average.
By averaging heights from ten trees, Method B2 will give more consistent estimates, making it closer to an MVUE compared to B1.

Example 3: Estimating Average Time Spent in a Bookstore
You are curious about the average time people spend in a bookstore.

Application of MVUE:

Unbiasedness:

Method A: Timing only people who visit during morning hours.

Method B: Randomly timing visitors throughout the day.

Method C: Timing only people who look like students.

Method B is unbiased because it doesn't have a preference for a particular group of visitors.

Minimum Variance:

Method B1: Timing a single random visitor.

Method B2: Timing ten random visitors and taking the average.

Method B2 provides a more consistent estimate by averaging multiple timings, making it a better MVUE candidate than B1.

Real-Life Example:

- **Estimating the Average Lifespan of Light Bulbs:**
  In a manufacturing plant, the goal is to estimate the average lifespan of light bulbs produced. A random sample of bulbs is tested to failure, and the lifespans are recorded.
  - **Unbiasedness:** A method is unbiased if it estimates the true average lifespan without systematic error. For instance, randomly selecting bulbs from various production batches provides an unbiased estimate, compared to selecting only from specific production runs (which might favor newer machines).
  - **Minimum Variance:** The method that minimizes variability between sample estimates is the one that averages results from multiple random batches rather than just relying on one.

In each example, the concept is to find a method that doesn’t favor any particular outcome (unbiasedness) and gives the most consistent results (minimum variance).

Importance of MUVE: The Minimum Variance Unbiased Estimator (MVUE) holds a significant place in statistical estimation theory due to its desirable properties. Here's why MVUE is considered important:

Optimality: MVUE is optimal in the sense that among all unbiased estimators, it has the smallest variance. This means that if we use an MVUE to estimate a parameter repeatedly, our estimates will tend to be tightly clustered around the true parameter value, making the estimator more reliable.

Efficiency: Because MVUE has the smallest variance among the unbiased estimators, it is the most efficient unbiased estimator. Efficiency means we're getting the most precision possible from our data.

Consistency: Many MVUEs are consistent, which means that as the sample size grows, the estimator converges to the true parameter value.

Unbiasedness: By definition, an MVUE is unbiased. This means that on average, it correctly estimates the parameter of interest. The estimates do not systematically overestimate or underestimate the true value.

Asymptotic Normality: Many MVUEs, under certain regularity conditions, are asymptotically normal. This property is crucial for making statistical inferences using the normal approximation.

Facilitates Inference: Because of its properties, MVUE provides a solid foundation for hypothesis testing and constructing confidence intervals. When an MVUE is used, we can be more confident in the validity of our statistical conclusions.

Simplifies Decision Making: In situations where multiple unbiased estimators are available, the MVUE offers a criterion for selecting the best one: the one with the minimum variance. This simplifies decision-making for statisticians and researchers.

Transparency and Credibility: Using an estimator known to have minimum variance and unbiased properties can lend credibility to research findings. It's easier for peers to accept and validate conclusions based on recognized optimal estimators.

In summary, the MVUE is valued because it combines two essential qualities: being unbiased (giving correct average estimates) and having the minimum variance (being the most precise among unbiased estimators). Utilizing an MVUE ensures that we are making the best possible estimations from our data, which is fundamental in statistical analysis and research.

9.4 Necessary and Sufficient conditions for MVUE

The concept of Minimum Variance Unbiased Estimators (MVUE) is closely tied with the properties of unbiasedness and efficiency of an estimator.

MVUE stands for Minimum Variance Unbiased Estimator. It is an estimator that has the smallest variance among all unbiased estimators of a parameter.

Necessary and Sufficient conditions for MVUE: For an estimator to be MVUE, it should be unbiased and efficient. The conditions for an estimator $\hat{\theta}$ of a parameter $\theta\ to\ be\ MVUE$ are:

1. **Unbiasedness:** the estimator $\hat{\theta}$ is unbiased for θ i.e.

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

2. **Completeness:** The distribution of the estimator belongs to a complete class of statistics. A class of statistics is said to be complete if the expectation of any function of the statistic is zero implies that the function itself is zero almost everywhere.

3. **Sufficiency:** The estimator $\hat{\theta}$ is based on a sufficient statistic for θ. A statistic $T(x)$ is said to be sufficient for θ if the distribution of the sample x given $T(x)$ does not depend on θ.

If an estimator is unbiased and based on a complete and sufficient statistic, then it is the MVUE i.e.

For an estimator to be MVUE, it should:

Be unbiased: E(θ̂) = θ.

Have minimum variance among all unbiased estimators.

The Rao-Blackwell and Lehmann-Scheffé theorems provide necessary and sufficient conditions for an estimator to be an MVUE.

Rao-Blackwell Theorem:

Statement: Given an unbiased estimator T(Y) of a parameter θ, where Y is a random sample, and given a sufficient statistic S(Y) for θ, the conditional expectation E[T(Y)|S(Y)] is also an unbiased estimator of θ. Furthermore, this improved estimator has variance less than or equal to that of T(Y), provided the variances exist.

Proof:

Unbiasedness: To show that

$T^*(Y)=E[T(Y)|S(Y)]$ is unbiased, we compute its expected value:

$E[T^*(Y)] =E[E[T(Y)|S(Y)]]$

Using the law of iterated expectations (also known as the tower property), we have:

$E[E[T(Y)|S(Y)]] =E[T(Y)]$

Since T(Y) is an unbiased estimator of θ,

$E[T(Y)] =\theta$. Thus,

$E[T^*(Y)] =\theta$, proving that $T^*(Y)$ is also unbiased.

Variance Reduction:

To prove the variance reduction, we start with a fundamental identity related to conditional variances:

$Var(T(Y)) =E[Var(T(Y)|S(Y))] +Var(E[T(Y)|S(Y)])$

Here, E[Var(T(Y)|S(Y))] is the average of the conditional variances of T(Y) given S(Y), and Var(E[T(Y)|S(Y)]) is the variance of the conditional expectations.

Now, note that:

$\text{Var}(T^*(Y)) = \text{Var}(E[T(Y)|S(Y)])$

From the identity, it is clear that:

$\text{Var}(T(Y)) = E[\text{Var}(T(Y)|S(Y))] + \text{Var}(T^*(Y))$

Given that

$\text{Var}(T(Y)|S(Y)) \geq 0$ for all values of S(Y), the term $E[\text{Var}(T(Y)|S(Y))]$ is non-negative.

Hence: $\text{Var}(T(Y)) \geq \text{Var}(T^*(Y))$

This proves that the conditional expectation $T^*(Y)$ has a variance that is less than or equal to the variance of T(Y).

The Rao-Blackwell theorem tells us that, if we have an unbiased estimator, we can potentially improve (reduce the variance of) this estimator by conditioning on a sufficient statistic. This result is foundational in the construction and understanding of optimal statistical estimators.

Lehmann-Scheffé Theorem:

Statement: If T(Y) is an unbiased estimator of θ and is a function of a complete and sufficient statistic S(Y), then T(Y) is the MVUE of θ.

Proof: Assuming S(Y) is a complete and sufficient statistic, and T(Y) is an unbiased estimator based on S(Y). We need to show that for any other unbiased estimator $T_1(Y)$ based on S(Y), $\text{Var}(T(Y)) \leq \text{Var}(T_1(Y))$.

Let $g(S(Y)) = E[T1(Y)|S(Y)] - T(Y)$. Due to the completeness of S(Y), $E[g(S(Y))] = 0$ implies $g(S(Y)) = 0$ almost surely.

Now, $E[T_1(Y)|S(Y)] = T(Y) + g(S(Y))$. But since $g(S(Y)) = 0$ almost surely, $E[T_1(Y)|S(Y)] = T(Y)$.

Using the Rao-Blackwell theorem, it follows that

$\text{Var}(T(Y)) \leq \text{Var}(T_1(Y))$.

Thus, the unbiased estimator T(Y) based on a complete and sufficient statistic S(Y) is indeed the MVUE.

Both theorems together provide a powerful framework for deriving MVUEs. First, find a sufficient statistic using factorization criteria. Then, if the statistic is also complete, any unbiased estimator based on this statistic is an MVUE. If the estimator is not based on the statistic, Rao-Blackwell can be used to improve it.

Some examples to illustrate the principles behind the Rao-Blackwell and Lehmann-Scheffé theorems.

Real-Life Example:

- **Surveying Average Income in a City:**
  Suppose a city government wants to estimate the average income of its citizens. An unbiased estimator can be achieved by randomly sampling from various income brackets without favouring any specific group.
  - **Sufficiency:** By basing the estimator on sufficient statistics, such as total income from a broad, representative sample, the government ensures that the estimate reflects the true population more accurately.
  - **Completeness:** The survey method ensures that there is no other way to achieve a better or more comprehensive estimate for the given sample.

Rao-Blackwell Theorem: You are trying to estimate the average number of books a person reads in a year in a town. You have the data of ten randomly chosen individuals from last year.

Initial Estimator: You decide to randomly pick one person's reading count from last year as your estimate.

Sufficient Statistic: The total number of books read by the ten individuals. This is a sufficient statistic because it contains all the information you need about the average.

Rao-Blackwell Improvement: Instead of just picking one person's count, you use the sufficient statistic (the total count) and estimate the average by dividing the total count by ten. This is essentially the conditional expectation given the total.

Intuitively: Your initial random method might give you vastly different results each time. But once you condition on the total number of books read (sufficient statistic), the variability in your estimate reduces.

Lehmann-Scheffé Theorem: Continuing from the previous example, you want to find the best way to estimate the average number of books a person reads in a year.

Sufficient and Complete Statistic: Assume (for the sake of this example) that the total number of books read by the ten individuals is both sufficient and complete. This means it captures all the information about the average, and there is no other unbiased estimator based on this statistic that has a smaller variance.

Unbiased Estimator: The average number of books read by the ten individuals (total count divided by ten).

Lehmann-Scheffé's MVUE: The theorem tells us that the above estimator (average) is not only unbiased but also has the minimum variance among all unbiased estimators based on the sufficient statistic. Thus, it is the MVUE.

Intuitive Takeaway: If you are using the total number of books read (a complete and sufficient statistic) to estimate the average, you cannot find a better unbiased estimator than simply taking the average of the ten individuals.

In both examples, the core idea is about refining and improving the estimation process, first by reducing variability using available data (Rao-Blackwell), and then ensuring the estimator is the best possible one given the data (Lehmann-Scheffé).

9.5 Zehna Theorem

Zehna or Invariance Property of MLE:

If you have an MLE for a parameter, then the MLE for any function of that parameter is simply the same function of the MLE.

In other words, if you have found the MLE of a parameter and you want the MLE of some transformation of that parameter, you can just apply the transformation directly to the MLE.

***Proof*:** Imagine we have a random sample of data and a likelihood function based on that data. This likelihood function tells us, for any given parameter value, how "compatible" that parameter value is with the observed data. The parameter value that makes the observed data most "likely" (or most "compatible") is the MLE of the parameter.

Now, suppose you have found the MLE of a parameter, say θ. Let us call this MLE$\hat{\theta}$. If you now want the MLE of a transformed parameter, say g(θ), then what you are really asking is: "What is the value of g(θ) that makes the observed data most likely, given that θ is most 'likely' at $\hat{\theta}$?

The answer is straightforward: if θ is "most likely" at $\hat{\theta}$, then g(θ) will be "most likely" at g($\hat{\theta}$).

This is the essence of the invariance property. The likelihood for the transformed parameter is maximized at the same relative location as the likelihood for the original parameter.

In other words, if you know the MLE for a parameter, you do not have to do a whole new set of calculations to find the MLE for a function of that parameter. You can simply apply the function to the known MLE.

To summarize, the invariance property is a time-saving feature of MLEs that simplifies the process of finding estimates for transformed parameters.

Real-Life Example:

- **Estimating the Rate of Infection in a Population:**

  Suppose public health officials estimate the infection rate of a disease using MLE based on a sample of test results. If they want to estimate the square of the infection

rate (to calculate certain risk measures), they can simply square the MLE of the infection rate, without needing to conduct an additional estimation process.

- **Invariance Property:** This simplifies calculations and ensures that transformations of the MLE are straightforward and consistent.

9.6 Weak Consistency: Cramer's Theorem

In statistical inference, **consistency** is a desirable property of an estimator, ensuring that as the sample size increases, the estimator converges to the true value of the parameter being estimated. Specifically, an estimator is said to be **weakly consistent** if, as the sample size grows, the probability that the estimator is arbitrarily close to the true parameter approaches 1. This concept is fundamental for validating that estimators perform well with sufficiently large datasets, a scenario commonly encountered in real-world applications.

One of the key theoretical results that supports weak consistency in Maximum Likelihood Estimation (MLE) is **Cramér's Theorem**. This theorem provides conditions under which the Maximum Likelihood Estimator is weakly consistent, ensuring that it converges to the true parameter as the sample size increases. Cramér's Theorem, named after the Swedish mathematician **Harald Cramér**, is foundational in modern estimation theory.

In this section, we explore the concept of weak consistency, the intuition behind Cramér's Theorem, and its application to MLE. We will also examine the regularity conditions necessary for weak consistency to hold and the practical implications for statistical modelling.

Weak consistency, also called convergence in probability, means that the probability of the estimator being close to the true parameter value increases as the sample size grows. Cramer's theorem provides conditions under which an estimator will be weakly consistent.
Cramér's theorem for the Maximum Likelihood Estimator (MLE) is a foundational result establishing the consistency of the MLE under certain conditions.

Statement: let $\{X_1, X_2, \dots, X_n\}$ be a sequence of independent and identically distributed random variable with a common probability density function$f(x;\theta)$ that depends on a parameter θ lying a parameter space Θ. Let $\widehat{\theta_n}$ be the MLE based on the first n observations. If certain regularity conditions are met (conditions related to the properties of the likelihood

function, such as differentiability and boundedness of certain derivatives), then the MLE is weakly consistent, i.e.,

$$\widehat{\theta_n} \xrightarrow{P} \theta_0$$

Where θ_0 is the true value of the parameter.

Proof: the MLE $\widehat{\theta_n}$ is found by solving the likelihood equations:

$$\frac{\partial}{\partial\theta}\log L(\theta; x_1, x_2, \dots, x_n) = 0$$

Where $L(\theta; x_1, x_2, \dots, x_n)$ is likelihood function.

Under the given regularity conditions, we can show that the expected value of the score function, which is the derivative of the log-likelihood with respect to θ, is zero:

$$\frac{\partial}{\partial\theta}\log f(\theta_0; x_1, x_2, \dots, x_n) = 0$$

Where $f(.)$ is the probability density function and θ_0 is the true value of the parameter.

Again, using the regularity conditions, the central limit theorem (CLT) can be applied to the score function to show that:

$$\sqrt{n}\left[\frac{\partial}{\partial\theta}\log L(\theta; x_1, x_2, \dots, x_n)\right] \xrightarrow{D} N(0, I(\theta_0))$$

Where $I(\theta_0)$ is the Fisher information.

The above results, along with some additional regularity and technical conditions, can be combined to show that $\widehat{\theta_n}$ converges in probability to θ_0.

To fully appreciate and rigorously prove Cramér's theorem, one typically needs to delve deep into the mathematics and conditions underlying the theorem. The above provides an outline and intuitive idea behind the proof rather than a rigorous mathematical demonstration. If you need a detailed proof, advanced texts in mathematical statistics would be the right place to consult.

Cramér's theorem for the consistency of the Maximum Likelihood Estimator (MLE) lays out foundational conditions under which the MLE is a consistent estimator. The theorem essentially says that as you gather more data (increase your sample size), the MLE tends to get closer to the true parameter value, assuming certain conditions are met.

Examples: Let us illustrate with non-numerical examples:

Coin Toss: Suppose you are trying to estimate the probability p of getting a head when tossing a biased coin. The MLE for p based on observed data is just the sample proportion of heads.

Weak Consistency: If you keep tossing the coin an infinite number of times, the sample proportion of heads (the MLE) will converge to the true probability p of getting a head, assuming the conditions of Cramér's theorem are met (like the tosses being independent).

Height of Students: Imagine you're trying to estimate the mean height of all students in a large university based on a sample. The MLE for the mean is simply the sample average height.

Weak Consistency: As you increase your sample size to include more students, the sample average height (the MLE) will get closer and closer to the true average height of all students in the university, again assuming the conditions of Cramér's theorem hold (e.g., the sampled heights are independent and identically distributed).

Factory Production: A factory produces light bulbs, and you are trying to estimate the failure rate λ of these bulbs. The MLE for λ could be based on observing a sample of bulbs and calculating the sample failure rate.

Weak Consistency: If you keep testing more and more bulbs, the sample failure rate (the MLE) will converge to the true failure rate λ of the bulbs produced by the factory, provided the conditions of Cramér's theorem are satisfied.

In each of these examples, the essence is the same: as you gather more data, the MLE becomes a better and better estimate of the true parameter, assuming the conditions of the theorem are met. The MLE "converges in probability" to the true parameter value.

Real-Life Example:

- **Estimating the Proportion of Defective Products in a Factory:** A quality control team monitors the proportion of defective items in a large production facility. Using MLE, they estimate this proportion based on a sample of produced items.
- **Weak Consistency:** As they increase the sample size over time, their estimate for the defect rate becomes more accurate, converging on the true proportion of defective products. This is especially important for long-term quality control.

9.7 Cramer-Huzurbazar Theorem

The **Cramér-Huzurbazar Theorem** is a powerful result in statistical theory that extends the principles of consistency and asymptotic normality to multivariate estimators. This theorem plays a critical role in multivariate analysis and characterizes the behaviour of estimators when dealing with more than one parameter. It builds upon the insights of **Cramér's Theorem** and the work of **V.V. Huzurbazar**, a notable statistician who contributed to the understanding of convergence in distribution and sufficiency in multivariate contexts.

In its essence, the Cramér-Huzurbazar Theorem provides conditions under which multivariate estimators converge to a normal distribution as the sample size increases. This result is particularly important for multivariate models, where understanding the joint behaviour of multiple parameters is essential for accurate inference.

The theorem relies on concepts such as the **Cramér-Wold device**, which is a method for reducing a multivariate problem to a univariate one by taking linear combinations of the parameters. By establishing conditions for convergence in distribution, the Cramér-Huzurbazar Theorem enables statisticians to apply normal theory to complex, multidimensional problems.

In this section, we will explore the Cramér-Huzurbazar Theorem in detail, outlining its assumptions, theoretical foundations, and practical applications in multivariate statistics. This theorem offers valuable insights into the behaviour of estimators in multivariate models, and it serves as a cornerstone in the study of asymptotic theory.

This is sometimes called the Cramér-Huzurbazar theorem. Here is a statement and a brief proof:

Statement: let $X_n \text{ and } X \text{ be random vectors in } \mathbb{R}^k$. Then $X_n \xrightarrow{D} X$ i.e. X_n converges in distribution to X if and only if $t'X_n \xrightarrow{D} t'X$ for all $t \in \mathbb{R}^k$, where t' Denotes the transpose of t.

Proof: Suppose $X_n \xrightarrow{D} X$. Then for any fixed $t \in \mathbb{R}^k$, $t'X$ is a real vslued random variable, and its distribution function converges to that of $t'X$ by the distribution of convergence in distribution.

Conversely, suppose $t'X_n \xrightarrow{D} t'X$ for all $t \in \mathbb{R}^k$. Using the Cramer-Wold device, we want to show that $X_n \xrightarrow{D} X$. Without loss of generality, let $X = 0$, the zero vector.

Given any sequence $t_n \; converging \; to \; t \; with \; \|t\| = 1$ we know thast $t'X_n \xrightarrow{D} t'X = 0$, by the Skorohod Representation Theorem. Essentially, we can find a probability spacve on which $t'X_n \; and \; t'X$ are defined and $t'X_n \rightarrow t'X$ almost surely.

Using this result and the tightness of the family of probability measures (a technical condition ensuring that the probability measures don't escape to infinity), one can then show that the convergence in distribution of the one-dimensional marginals (the $t'X_n \xrightarrow{D} t'X \; for \; all \; t \; part$) implies convergence in distribution of the multivariate i.e. $X_n \xrightarrow{D} X$.

This is a simplified explanation, and the complete proof requires a more thorough examination of the properties of convergence in distribution and the characteristics of the underlying probability measures. Advanced texts on multivariate statistical theory or convergence of random vectors would provide a detailed discussion.

Real-Life Example:

- **Multivariate Analysis in Marketing:** A marketing company gathers data on multiple factors affecting sales (e.g., advertising budget, customer demographics, and product prices). The goal is to estimate the joint impact of these factors on sales performance.
- **Multivariate Estimation:** The Cramér-Huzurbazar theorem ensures that, under certain conditions, the estimates for multiple variables (such as the effect of price and

advertising on sales) will behave like normally distributed estimates as the sample size grows. This is crucial for making reliable predictions and decisions in complex multivariate systems.

9.8 Self-Assessment Exercises

1. What are the two key properties of an MVUE?
2. Explain the significance of minimum variance in an estimator.
3. How does the Lehmann-Scheffé theorem help in identifying an MVUE?
4. What are the necessary and sufficient conditions for an estimator to be MVUE?
5. Define sufficiency and completeness in the context of statistical estimation.
6. Describe how the Rao-Blackwell theorem improves an estimator.
7. Explain the invariance property of MLE using a real-life example.
8. How does the Zehna theorem simplify the process of finding MLE for transformed parameters?
9. Define weak consistency in estimation.
10. Under what conditions does Cramér's theorem ensure the weak consistency of MLE?
11. What is the role of the Cramér-Huzurbazar theorem in multivariate estimation?
12. How does the Cramér-Wold device assist in reducing a multivariate problem to a univariate one?

9.9 Summary

This unit provided an in-depth exploration of advanced estimation methods, focusing on the critical concepts of unbiasedness and efficiency in statistical estimation. It began with a detailed examination of Minimum Variance Unbiased Estimators (MVUE), explaining that MVUEs are estimators that, while being unbiased, have the smallest possible variance among all unbiased estimators. These properties make MVUEs optimal for statistical inference, ensuring that estimates are both accurate and precise.

We then explored the necessary and sufficient conditions for an estimator to be classified as an MVUE, emphasizing the importance of sufficiency and completeness in

deriving such estimators. The Rao-Blackwell theorem was introduced as a method to improve an existing unbiased estimator by conditioning it on a sufficient statistic, thereby reducing its variance. This was followed by a discussion of the Lehmann-Scheffé theorem, which provides a necessary and sufficient condition for an unbiased estimator to be an MVUE, helping to formalize the process of selecting the best estimator.

The Zehna theorem was presented to highlight the invariance property of Maximum Likelihood Estimators (MLE). This property simplifies the estimation process for functions of parameters, as it allows us to apply transformations directly to the MLE of a parameter without requiring further calculations. This practical result streamlines complex estimation procedures in real-world applications.

Next, we discussed the concept of weak consistency, a desirable property that ensures estimators become more accurate as the sample size increases. Cramér's theorem was introduced to formalize this idea, showing the conditions under which the MLE is weakly consistent, ensuring that the estimator converges to the true value of the parameter as the data set grows. This is especially important in large-sample scenarios, where consistency is critical for reliable inference.

Finally, we examined the Cramér-Huzurbazar theorem, which extends the principles of estimation to multivariate cases. This theorem provides conditions under which multivariate estimators converge to a normal distribution, a vital result for complex models involving multiple parameters. The Cramér-Wold device, which reduces multivariate problems to simpler univariate ones, was also discussed as a useful tool for analysing the joint behaviour of several parameters in large-sample estimation problems.

Throughout this unit, the interplay between theory and application was emphasized, ensuring that students not only understand the theoretical underpinnings of these estimation methods but also how they apply to real-world scenarios. The tools and theorems discussed in this unit—such as the Rao-Blackwell theorem, Lehmann-Scheffé theorem, and Cramér's theorem—are foundational to modern statistical inference, equipping students with the knowledge and skills to make more accurate, efficient, and reliable estimations in diverse fields of study.

9.10 References

- Casella, G., & Berger, R. L. (2002). *Statistical Inference* (2nd ed.). Duxbury Press.
- Lehmann, E. L., & Casella, G. (1998). *Theory of Point Estimation* (2nd ed.). Springer.
- Hogg, R. V., McKean, J. W., & Craig, A. T. (2018). *Introduction to Mathematical Statistics* (8th ed.). Pearson.
- Rao, C. R. (1973). *Linear Statistical Inference and Its Applications* (2nd ed.). Wiley.
- Cramér, H. (1946). *Mathematical Methods of Statistics*. Princeton University Press.
- Fisher, R. A. (1925). *Statistical Methods for Research Workers*. Oliver & Boyd.
- Huzurbazar, V. S. (1976). *Sufficiency and Statistical Decision Functions*. Academic Press.

9.11 Further Readings

- Ferguson, T. S. (1996). *A Course in Large Sample Theory*. Chapman & Hall.
- Schervish, M. J. (1995). *Theory of Statistics*. Springer.
- Spanos, A. (1999). *Probability Theory and Statistical Inference*. Cambridge University Press.
- van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press.
- Amemiya, T. (1985). *Advanced Econometrics*. Harvard University Press.
- Bickel, P. J., & Doksum, K. A. (2015). *Mathematical Statistics: Basic Ideas and Selected Topics* (Vol. I, 2nd ed.). CRC Press.
- Tanner, M. A. (1996). *Tools for Statistical Inference*. Springer.

UNIT - 10: CRITERION FOR GOOD ESTIMATORS-I

Structure

10.1 Introduction

Estimation theory is a central area in statistics that deals with determining the most accurate ways to estimate unknown parameters based on sample data. The focus of this unit is to explore the **criteria for good estimators**, providing the foundation for understanding the quality of estimators used in statistical inference. A good estimator should not only provide an accurate estimate but should also be reliable, efficient, and consistent, especially as the sample size increases.

The historical development of estimation theory introduced various methods to evaluate the quality of estimators. Early work by Karl Pearson, particularly with the **Method of Moments (MoM)**, emphasized simple techniques to estimate population parameters by matching sample moments to theoretical moments. However, the more advanced framework for judging estimators came with Sir Ronald A. Fisher's introduction of **Maximum Likelihood Estimation (MLE)** in the early 20th century. MLE revolutionized statistical

estimation by providing consistent and efficient parameter estimates under specific conditions, making it a widely used method in various fields.

This unit is focuses on key properties of good estimators, such as **unbiasedness**, **consistency**, and **efficiency**. An estimator is said to be unbiased if its expected value equals the true parameter, ensuring that, on average, the estimator provides the correct value. **Consistency** guarantees that as the sample size grows, the estimator converges to the true parameter, making it reliable in large samples. **Efficiency** relates to minimizing the variance of an estimator, and the **Cramér-Rao Lower Bound (CRLB)** provides a benchmark for evaluating this property by setting a lower limit on the variance of unbiased estimators.

Additionally, this unit introduces important bounds that are used to evaluate the performance of estimators in more complex scenarios. The **Bhattacharya bound** offers a lower bound on the mean squared error (MSE) of biased estimators, expanding the criteria for evaluation beyond unbiasedness. The **Chapman-Robbins-Kiefer (CRK) bound** further refines these ideas by offering tighter bounds on the variance of estimators, particularly when regularity conditions are not met.

The concept of **asymptotic normality** is also covered in this unit. As sample sizes increase, the distribution of an estimator can often be approximated by a normal distribution, a property known as asymptotic normality. This is a key principle in large-sample inference, allowing the use of normal distribution-based methods for hypothesis testing and constructing confidence intervals.

10.2 Objectives

By the end of this unit, learners will be able to:

1. **Understand the Criterion for Good Estimators**:
    - Comprehend the essential properties that define a good estimator, including unbiasedness, consistency, and efficiency.
    - Learn to apply these criteria when evaluating different estimators in practical statistical problems.
2. **Explain the Bhattacharya Bound**:

- Understand the Bhattacharya bound as a measure that provides a lower bound on the mean squared error for biased estimators.
- Apply the Bhattacharya bound to assess estimator performance when dealing with biased estimators, extending beyond the limitations of the Cramér-Rao Lower Bound (CRLB).

3. **Describe the Chapman-Robbins-Kiefer (CRK) Bound**:
   - Gain a deep understanding of the Chapman-Robbins-Kiefer (CRK) bound, which refines the evaluation of estimator efficiency beyond the conditions required for the Cramér-Rao Lower Bound.
   - Learn how to apply the CRK bound in scenarios where the CRLB does not hold, particularly in complex statistical models.
4. **Comprehend Asymptotic Normality**:
   - Grasp the concept of asymptotic normality and its importance in large-sample estimation.
   - Understand how estimators, when properly normalized, converge to a normal distribution as the sample size increases, facilitating hypothesis testing and the construction of confidence intervals.
5. **Differentiate Between BAN and CAN Estimators**:
   - Learn the difference between Best Asymptotic Normal (BAN) and Consistent Asymptotically Normal (CAN) estimators.
   - Recognize the significance of BAN estimators in achieving the lowest possible variance and the role of CAN estimators in ensuring consistency in large samples.
6. **Understand the Principle of Asymptotic Efficiency**:
   - Comprehend the concept of asymptotic efficiency, where an estimator achieves the smallest variance possible in large samples.
   - Explore how asymptotic efficiency is connected to the Cramér-Rao Lower Bound and how it affects estimator selection in large-sample scenarios.
7. **Explore Equivariant Consistency**:
   - Learn about equivariant consistency, which ensures that estimators maintain their consistency under transformations of the data, such as scaling or shifts.

- Apply the concept of equivariant consistency to statistical problems where the data undergo transformations, ensuring the robustness and reliability of the estimators.

10.3 Criterion for good estimators

In statistical estimation, the primary goal is to find the best possible method to estimate unknown population parameters from sample data. However, not all estimators are created equal, and various criteria have been developed to judge the quality of an estimator. This section focuses on the **criterion for a good estimator**, which involves properties such as unbiasedness, consistency, efficiency, and sufficiency. Understanding these criteria is fundamental to selecting an estimator that provides accurate, reliable, and efficient results.

The historical development of these concepts can be traced back to the early 20th century, with the groundbreaking work of Sir Ronald A. Fisher. Fisher introduced the **Maximum Likelihood Estimation (MLE)** method, which became a cornerstone of estimation theory due to its desirable properties like consistency and asymptotic normality. Fisher also formalized the idea of **efficiency** through the **Cramér-Rao Lower Bound (CRLB)**, which establishes a theoretical lower limit on the variance of unbiased estimators, serving as a benchmark for evaluating the efficiency of an estimator.

At the same time, the concept of **unbiasedness** was recognized as a critical property. An estimator is unbiased if, on average, it provides the true value of the parameter being estimated. However, as estimation theory progressed, it became evident that unbiasedness alone was insufficient. **Consistency** was introduced as a vital property, ensuring that as the sample size increases, the estimator converges to the true value of the parameter. This was essential in fields like econometrics and biology, where large datasets are common.

Fisher's work, along with contributions from other statisticians such as C.R. Rao and Jerzy Neyman, led to the development of more refined criteria, including **sufficiency** and **efficiency**. Sufficiency ensures that an estimator fully utilizes the available data, while efficiency focuses on minimizing the variance of the estimator among all unbiased estimators. These properties form the foundation of modern estimation theory and remain key tools for statisticians in evaluating the performance of estimators.

In this section, we will explore each of the key criteria—unbiasedness, consistency, efficiency, and sufficiency—that define a good estimator. Understanding these properties allows statisticians to choose the most appropriate estimator for a given situation, ensuring that their estimates are both accurate and reliable.

1. Unbiasedness

Perhaps the most intuitive criterion, unbiasedness, refers to an estimator's expected value being equal to the true parameter it aims to estimate. Unbiasedness: On average, the estimator should be equal to the true parameter.

Definition: An estimator $\hat{\theta}$ of a parameter θ is said to be unbiased if:

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

$for\ instance, the\ sample\ mean\ is\ an\ unbiased\ estimator\ of\ the\ population\ mean.$

Unbiasedness is a fundamental concept in statistics, ensuring that, on average, an estimator gives the correct value of a parameter. To understand this property better, let us delve into some supporting results:

1. Linearity of Expectation

One of the foundational results supporting unbiasedness is the linearity of expectation. For any random variables X and Y and constants a and b:

E(aX+bY) = aE(X)+bE(Y)

This property is instrumental in proving the unbiasedness of many estimators, especially when they are linear combinations of random variables.

2. Sample mean of iid random variables:

If $X_1, X_2, \ldots, X_n$ are independent and identically distributed random variables with mean $\mu\ then\ the\ sample\ mean\ \bar{X}$ is an unbiased estimator of μ.

Proof:

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

Using linearity of expectation:

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

3. Sample variance of iid random variables

For iid random variables $X_1, X_2, \ldots, X_n$ with mean and variance $\mu \; and \; \sigma^2$ respectively, the sample variance defined as:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Here, S^2 is an unbiased estimator of σ^2.

This result might seem non-intuitive because of the n−1 in the denominator instead of n. This correction is called Bessel's correction and is necessary to account for the fact that the sample mean $\bar{X}$ is used in the place of μ.

4. Transformation of unbiased estimators

If an estimator $\hat{\theta}$ is unbiased estimator of θ and $g(\theta)$ is a differentiable function, then $g(\hat{\theta})$ is generally not an unbiased estimator of $g(\theta)$, except in specific cases. This result highlights the need to be caution when transforming unbiased estimators.

5. Bias of the difference

If $\widehat{\theta_1}$ and $\widehat{\theta_2}$ are unbiased estimators of the same parameter θ, then the difference between them, $\widehat{\theta_1} - \widehat{\theta_2}$, also has an expected value of zero, but it does not mean this difference is an unbiased estimator of θ. It is an unbiased estimator of 0.

Unbiasedness is a central concept, but it's essential to approach it with a nuanced perspective. While an unbiased estimator gives the correct value on average, it doesn't guarantee that it's the best or most efficient estimator. Moreover, understanding the properties and results associated with unbiasedness helps in better designing and evaluating statistical procedures.

Advantages:

- Offers a straightforward measure of the estimator's accuracy on average.
- Useful in many practical applications where bias can lead to systematic errors.

Limitations:

- Being unbiased doesn't guarantee that the estimator will be close to the true parameter for any specific sample.
- Sometimes, a slightly biased estimator might perform better in terms of other criteria.

Understanding unbiasedness through non-numerical examples can help illustrate the concept without getting caught up in calculations. Here are a few illustrative examples:

Example 1: Weighing a Fruit

Situation: Imagine you have a weighing scale that you use to measure the weight of fruits. Every time you weigh an apple, the scale consistently adds 0.5 grams to its true weight.

Estimator: The reading of the weighing scale.

Discussion: Since the scale consistently overestimates the weight by 0.5 grams, the expected value of the weight (as given by the scale) will not equal the true weight of the apple. Therefore, the scale is a biased estimator of the apple's weight.

Example 2: Tossing a Fair Coin

Situation: You are tossing a fair coin, and you're interested in estimating the probability p that it lands heads.

Estimator: The proportion of times the coin lands heads after n tosses.

Discussion: If you were to repeatedly toss the coin many times and calculate the proportion of heads each time, the average of those proportions would converge to the true probability p of getting a head (which is 0.5 for a fair coin). Hence, the proportion of heads in a sample of coin tosses is an unbiased estimator of p.

Example 3: Measuring Student Heights

Situation: You are a school teacher, and you want to estimate the average height of students in your school. You decide to measure the heights of students in your class and use the average height of your class as an estimator for the average height of the entire school.

Estimator: The average height of students in your class.

Discussion: If your class is a representative sample of the entire school population (i.e., it's a random sample), then the expected value of the average height of your class would equal the true average height of the school. In this case, your estimator would be unbiased. However,

if, say, your class was specifically for tall students or short students, then the estimator would be biased.

Example 4: Movie Ratings

Situation: You are trying to estimate the average rating of a movie based on online reviews. However, you notice that only those who either really loved it or really hated it are motivated to leave a review.

Estimator: The average rating from the online reviews.

Discussion: Given that the reviews are skewed towards extreme opinions, the expected value of the average rating from online reviews might not match the true average rating of all viewers. Hence, in this situation, the average online rating would likely be a biased estimator for the true average rating of the movie.

1. **Consistency:**

Consistency is a property that concerns large samples. An estimator is consistent if, as the sample size grows to infinity, it converges in probability to the true parameter value. As the sample size increases, the estimator should converge to the true parameter.

Definition: An estimator $\widehat{\theta_n}$ $is\ consistent\ for\ \theta$ if:

$\widehat{\theta_n} \xrightarrow{P} \theta\ as\ n \to \infty$ where $n\ is\ the\ sample\ size.$

Consistency, as mentioned, is a crucial property of an estimator. Let's discuss some foundational results related to consistency and provide proofs for them:

Weak Law of Large Numbers

Statement: If $X_1, X_2, \dots, X_n$ are independent and identically distributed random variables with a finite mean $\mu, then\ the\ sample\ mean\ \overline{X_n}$ converges in probability to $\mu\ as\ n \to \infty$.

Proof: to prove convergence in probability, we need to show:

$$\log_{n\to\infty} P(|\overline{X_n} - \mu| > \epsilon) = 0; for\ all\ \epsilon > 0$$

Using Chebyshev's inequality:

$$P(|\overline{X_n} - \mu| > \epsilon) \leq \frac{Var(\overline{X_n})}{\epsilon^2}$$

Given that $Var(\overline{X_n}) = \frac{\sigma^2}{n}$,

$P(|\overline{X_n} - \mu| > \epsilon) \leq \frac{\sigma^2}{n\epsilon^2}$, as $n \to \infty$, the right-side approaches to 0, which implies $\overline{X_n}$ is consistent for μ.

Convergence of Sample Variance to Population Variance

Statement: For $X_1, X_2, \ldots, X_n$ are independent and identically distributed random variables with a finite mean μ *and variance* σ^2, the sample variance S_n^2 defined by:

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X_n})^2$$

Is consistent for σ^2.

Proof: This proof is a bit more involved. Essentially, one can expand S_n^2 and take expectations, showing that the bias decreases to zero as n→∞. Further, one can use Chebyshev's inequality (similarly to the WLLN) to show convergence in probability to σ^2.

3. Continuous Mapping Theorem

Statement: If X_n converges in probability to X and g is a continuous function, then $g(X_n)$ converges in probability to $g(X)$.

Proof:

Given $\epsilon > 0$, since X_n converges in probability to X, for every $\delta > 0$, when n is large

$$P(|X_n - X| > \delta) < \frac{\epsilon}{2}$$

Due to the continuity of g, there exists a δ such that for all $y\ with\ |y - X| < \delta$, $|g(y) - g(X)| < \epsilon$.

Combining the two we get,

$$P(|g(X_n) - g(X)| > \epsilon) < \frac{\epsilon}{2}$$

Hence, $g(X_n)$ converges in probability to $g(X)$.

These are foundational results that highlight the behaviour of certain statistics or sequences of random variables as the sample size grows. They provide the backbone for understanding consistency in various statistical contexts.

Advantages:

- Ensures that with larger samples, our estimates become more accurate.
- Particularly useful in the era of big data, where large datasets are common.

Limitations:

- Does not provide any assurance about the estimator's performance for small sample sizes.

Consistency is a property of an estimator where, as the sample size grows, the estimator converges in probability to the true parameter value. Let us explore this concept with some intuitive, non-numerical examples.

Example 1: Tasting Soup in a Large Pot

Situation: Imagine you have a massive pot of soup, and you wish to judge its saltiness. You cannot taste the entire pot, so you take a spoonful to gauge the flavour.

Estimator: The saltiness based on the spoonful you tasted.

Discussion: If you taste just one spoonful, you might hit a particularly salty or bland pocket. However, as you taste more and more spoonful's (increasing your "sample size"), you will get a better sense of the overall saltiness. If by tasting more, you can more accurately gauge the pot's saltiness, then your tasting method is consistent.

Example 2: Audience Opinion in a Theatre

Situation: You are a director previewing a new film in a theatre, and you want to estimate how well the entire audience likes it.

Estimator: The opinions of a few members of the audience you interview after the show.

Discussion: If you ask only one or two people, their opinions might not represent the entire audience. However, if you interview more and more people (increasing your "sample size"), your estimate of the overall audience opinion will become more accurate. If the general sentiment you gather from more interviews aligns closer to the true sentiment of the entire audience, then your method is consistent.

Example 3: Estimating Forest Health

Situation: You are an ecologist trying to gauge the health of a vast forest based on the health of its trees.

E**stimator**: The health assessment based on a few trees you examine.

Discussion: By examining only, a small patch, you might end up in a particularly healthy or diseased area, not representative of the entire forest. However, as you examine more areas and more trees (increasing your "sample size"), you get a better representation of the forest's overall health. If the larger the area you examine, the closer your assessment is to the actual forest health, your method is consistent.

Example 4: Measuring Fabric Quality

Situation: You are a tailor receiving a new roll of fabric. You want to judge its quality, but you cannot inspect every inch.

Estimator: The fabric's quality based on small patches you inspect.

Discussion: By examining just, a tiny patch, you might judge based on an anomaly - maybe a particularly rough or smooth spot. However, as you inspect larger and larger portions (increasing your "sample size"), you get a better understanding of the overall fabric quality. If your judgments become more accurate as you inspect more of the fabric, your evaluation method is consistent.

In essence, consistency captures the idea that with more data (or a larger sample size), our estimator provides a better approximation of the true parameter.

Efficiency: Efficiency, in the context of statistical estimation, refers to the precision of an estimator in terms of its variance. Among unbiased estimators, an efficient estimator is one with the smallest variance. Efficiency can be expressed as a ratio, known as the "efficiency," which compares the variance of an estimator with that of another estimator.

Definition: Suppose $\widehat{\theta_1}$ and $\widehat{\theta_2}$ are two unbiased estimators of a parameter θ. The efficiency e of $\widehat{\theta_1}$ relative to $\widehat{\theta_2}$ is defined as:

$$e = \frac{Var\ (\widehat{\theta_1})}{Var(\widehat{\theta_2})}$$

If $e > 1$, then $\widehat{\theta_1}$ is more efficient than $\widehat{\theta_2}$ since it has a smaller variance.

An estimator is said to be "efficient" if it achieves the Cramér-Rao Lower Bound (CRLB).

Cramér-Rao Lower Bound (CRLB):

The Cramér-Rao Lower Bound (CRLB) is a fundamental theorem in the field of statistical estimation that provides a lower bound on the variance of unbiased estimators of a parameter. It is a measure of the best possible accuracy (in terms of lowest variance) that any unbiased estimator can achieve. When an estimator reaches this bound, it is considered to be an efficient estimator.

Definition: Let $\hat{\theta}$ be an unbiased estimator of a parameter θ based on a sample $X_1, X_2, \dots, X_n$ from a probability distribution with probability density function or probability mass function $f(x;\theta)$, and certain regularity conditions are met, then the Cramer -rao lower bound states that the variance of $\hat{\theta}$ is bounded from below as follows:

$$Var(\hat{\theta}) \geq \frac{1}{nI(\theta)}$$

Where n is the sample size and $I(\theta)$ is the expected fisher information in a single observation, which is given by:

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

or equivalently,

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

if the expectation of the second derivative exists.

Intuition:

The CRLB tells us that there is a limit to how precise an estimator can be. This limit is determined by the information that the random sample carries about the parameter θ. The more information available, the lower the bound on the variance, and the more precise the estimation can potentially be.

Implications:

If an unbiased estimator achieves this lower bound, it is considered to be the minimum variance unbiased estimator (MVUE) for the parameter θ.

The bound is a function of the Fisher Information, which measures the amount of information that an observable random variable X carries about an unknown parameter θ upon which the probability of X depends.

The bound underscores the importance of the sample size n; as the sample size increases, the bound gets tighter, implying that the potential variance of the estimator decreases, allowing for more precise estimates.

The Cramér-Rao Lower Bound (CRLB) provides a lower bound on the variance of unbiased estimators of a parameter, but for the bound to be valid, certain regularity conditions must be met. These conditions ensure that the mathematical operations used in deriving the CRLB are legitimate.

Here are the common regularity conditions that are often stated in the context of the CRLB:

Differentiability: The probability density function (pdf) or probability mass function (pmf), f(x;θ), must be differentiable with respect to the parameter θ, and the derivative must be an integrable function of x.

Identifiability: The parameter θ must be identifiable from the pdf or pmf; that is, the mapping from θ to the distribution must be one-to-one. In other words, no two different values of θ produce the same distribution.

Expectation of the Score is Zero: The score, which is the derivative of the log-likelihood with respect to θ, $U(\theta) = \frac{\partial}{\partial\theta}\log f(X;\theta)$

Mathematically,

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

Exchange of Integration and Differentiation: It should be possible to interchange the order of integration (or summation in the discrete case) and differentiation with respect to θ. This requires that the integral (or sum) of the derivative of the likelihood function converges uniformly.

Finiteness of the Fisher Information: The Fisher Information I(θ) must be finite, which implies that the score has finite second moments. Mathematically,

$$I(\theta) = -E\left[\frac{\partial^2}{\partial^2\theta}\log f(X;\theta)\right] < \infty$$

Completeness: This is not always stated as a necessary condition for the CRLB, but it is related to the sufficiency of the data. A family of *pdfs* or *pmfs* {f(x;θ)} is said to be complete if, for any function g(x), if E[g(X)]=0 for all θ, then g(x) must be zero almost everywhere. Completeness ensures that the bound is the best possible for unbiased estimators.

Support of the Distribution: The support of the *pdf* or *pmf* (the set of all x where f(x;θ)>0) does not depend on the parameter θ.

These regularity conditions are essential to ensure that the CRLB is applicable, providing a theoretical benchmark for the lowest possible variance that an unbiased estimator of θ can achieve. However, in practice, many real-world problems may not satisfy all these conditions, and statisticians must be careful when applying the CRLB.

Proof: let $U(\theta) = \frac{\partial}{\partial\theta}\log f(X;\theta)$, be the score function. We begin with two key properties of the score function:

1. The expected value of the score function is zero.

$$E[U(\theta)] = \int U(\theta)\log f(X;\theta)dx$$

$$= \int \frac{\partial}{\partial\theta}\log f(X;\theta)dx$$

$$= \frac{\partial}{\partial\theta}\int \log f(X;\theta)dx = 0$$

2. The Fisher information can be represented as the variance of the score function

$$I(\theta) = Var[U(\theta)] = E[U(\theta)^2] - \{E[U(\theta)]\}^2 = E[U(\theta)^2]$$

Now consider the covariance between the estimator $\hat{\theta}$ and the score function:

$$Cov\left(\hat{\theta}\,, \mathrm{U}(\theta)\right) = E\left[\left(\hat{\theta} - E\left(\hat{\theta}\right)\right)(U(\theta) - E(U(\theta)))\right]$$

$$since, \qquad E\left(\hat{\theta}\right) = \theta \; and \; E\left(U(\theta)\right) = 0$$

This simplifies to:

$$Cov\left(\hat{\theta}\,, \mathrm{U}(\theta)\right) = E\left(\hat{\theta}\;\mathrm{U}(\theta)\right)$$

$by\ the\ Cauchy - Schwartz\ inequality, we\ have$

$$\left(Cov\left(\hat{\theta}\,, \mathrm{U}(\theta)\right)\right)^2 \leq Var\widehat{(\theta)}\mathrm{Var}\,(\mathrm{U}(\theta))$$

$substituting$ $\mathrm{Var}\left(\mathrm{U}(\theta)\right)$ $with\ I(\theta)$ and expanding the left side:

$$E\left(\hat{\theta}\,, \mathrm{U}(\theta)\right)^2 \leq Var\widehat{(\theta)}\;\mathrm{I}(\theta)$$

Since $\hat{\theta}$ is unbiased, we can differentiate inside the expectation:

$$E\left(\hat{\theta}\,\mathrm{U}(\theta)\right) = E\left(\hat{\theta}\,\frac{\partial}{\partial\theta}\log f(X;\theta)\right) = \frac{\partial}{\partial\theta}E\left(\hat{\theta}\right) = 1$$

Therefore,

$$1 \leq Var\widehat{(\theta)}I(\theta)$$

$$thus\ we\ get$$

$$\frac{1}{I(\theta)} \leq Var\widehat{(\theta)}$$

If we consider *n i.i.d.* random variables, the Fisher Information in the samples in $nI(\theta)$, hence:

This concludes the proof of the Cramér-Rao Lower Bound. The bound indicates that no unbiased estimator of θ can have a variance smaller than the reciprocal of the Fisher Information, and it provides a standard by which to measure the efficiency of an estimator.

10.4 Bhattacharya bound

The **Bhattacharya Bound** is an important concept in estimation theory that provides a lower bound on the mean squared error (MSE) of biased estimators. Unlike the **Cramér-Rao Lower Bound (CRLB)**, which applies strictly to unbiased estimators, the Bhattacharya Bound extends the evaluation of estimator quality by offering a way to assess biased estimators. This bound is particularly useful when unbiased estimators either do not exist or are not the most efficient option, and biased estimators are considered for practical reasons.

Developed as part of the advancements in estimation theory during the mid-20th century, the Bhattacharya Bound is named after the Indian statistician Anil Kumar Bhattacharya, who made significant contributions to statistical distance measures and estimation theory. The bound plays a critical role in situations where biased estimators are preferred for their lower variance, despite their potential to deviate systematically from the true parameter value.

The Bhattacharya Bound provides a more general framework for understanding the trade-off between bias and variance in estimators. It demonstrates that even when some bias

is introduced, the estimator's total error (measured by MSE) can still be minimized under certain conditions, making biased estimators more attractive in certain applied settings, especially when unbiased estimators have high variance.

In this section, we will examine the derivation of the Bhattacharya Bound, explore its relationship with the CRLB, and discuss how it can be applied to improve estimation in cases where biased estimators are preferred. Real-life examples will illustrate the practical importance of this bound in fields like economics and engineering, where the use of biased but efficient estimators is often necessary to achieve the best possible results.

Definition: Given an estimator $\hat{\theta}$ of a parameter θ, the mean square of $\hat{\theta}$ is defined as:

$$MSE(\hat{\theta}) \geq \frac{\left[1 + \left(\frac{d}{d\theta}\right) b(\theta)\right]^2}{nI(\theta)} + b(\theta)^2$$

Where,

$b(\theta) = \mathrm{E}(\hat{\theta}) - \theta$, is the bias of the estimator,

n is the sample size,

$I(\theta)$ is the Fisher Information for a single observation, and

$\left(\frac{d}{d\theta}\right) b(\theta)$ is the derivative of the bias with respect to θ.

Intuition:

The Bhattacharya bound indicates that the mean square error of an estimator must account for the information provided by the data (as captured by the Fisher Information), as well as any bias that the estimator has. Unlike the CRLB, which holds only for unbiased estimators, the Bhattacharya bound recognizes that bias can be a significant component of an estimator's error, and it quantifies how bias and variance contribute to the overall MSE.

The term $\left[\left(\frac{d}{d\theta}\right) b(\theta)\right]^2$ adjusts the Fisher Information part of the bound to account for the presence of bias in the estimator. If the estimator is unbiased, b(θ)=0 and its derivative is also zero, which reduces the Bhattacharya bound to the CRLB.

Application:

The Bhattacharya bound is particularly useful when we are dealing with biased estimators or when it is not desirable or possible to have an unbiased estimator. In such cases, the bound provides a more applicable measure of the estimator's performance than the CRLB.

Limitations:

As with the CRLB, the Bhattacharya bound is based on certain regularity conditions that must hold for the bound to be applicable. It assumes the existence of the first and second derivatives of certain functions and requires that the estimator and the underlying probability density functions behave well enough for these mathematical operations to be valid.

10.5 Chapman, Robbins, and Kiefer (CRK) Bound

The Chapman-Robbins-Kiefer (CRK) bound is a lower bound on the variance of estimators that is a generalization of the Cramér-Rao Lower Bound (CRLB). It can be used even when the regularity conditions necessary for the CRLB do not hold. This bound is particularly useful for complex problems where the probability density function may not be differentiable with respect to the parameter, or where the support of the distribution depends on the parameter.

Statement of the Chapman-Robbins-Kiefer Bound:

Suppose X is a random variable with probability density function $f(x;\theta), where\ \theta$ is a scalar parameter. $\hat{\theta}(X) be\ an\ estimator\ of\ \theta$. The CRK bound state that for any $\theta' \neq \theta$ and any unbiased estimator $\hat{\theta}$, the following inequality holds;

$$Var(\hat{\theta}) \geq \frac{(\theta' - \theta)^2}{nE_\theta\left[\left(\frac{f(x;\theta')}{f(x;\theta)} - 1\right)^2\right]}$$

where n is sample size E_θ denotes the expectation with respect to the distribution $f(x;\theta)$

Proof:

The proof of the CRK bound builds on an application of the Cauchy-Schwarz inequality to a cleverly chosen set of random variables.

Let $g(X)$ be an arbitrary measurable function with $E_\theta[g(X)] = 0$ and $E_\theta[g(X)^2] < \infty$

Consider two parameter values $\theta' and\ \theta$ with $\theta' \neq \theta$. Define the random variable Y as follows:

$$Y = \frac{f(x;\theta')}{f(x;\theta)} - 1$$

Now consider the covariance between $g(X)\ and\ Y$ under distribution parameterized by θ.

$$Cov_\theta(g(X), Y) = E_\theta[g(X)Y] - E_\theta[g(X)]E_\theta(Y)$$

Since $E_\theta[g(X)] = 0$, by design, the covariance simplifies to;

$$Cov_\theta(g(X), Y) = E_\theta[g(X)Y]$$

Now by Cauchy-Schwartz inequality, we have

$$E_\theta[g(X)Y]^2 \leq E_\theta[g(X)]^2 E_\theta[Y]^2$$

Choose $g(X) = \hat{\theta}(X) - \theta$, note that $E_\theta[g(X)] = E_\theta\left[\hat{\theta}(X)\right] - \theta = 0$

Since $\hat{\theta}$ is unbiased, we then have:

$$E_\theta\left[\left(\hat{\theta} - \theta\right)Y\right]^2 \leq Var_\theta(\hat{\theta}) E_\theta[Y]^2$$

Now, $E_\theta\left[\left(\hat{\theta} - \theta\right)Y\right]$ simplifies as follows

$$E_\theta\left[\left(\hat{\theta} - \theta\right)Y\right] = \theta' - \theta$$

Thus we obtain,

$$(\theta' - \theta)^2 \leq Var_\theta(\hat{\theta}) E_\theta[Y]^2$$

Hence,

$$Var(\hat{\theta}) \geq \frac{(\theta' - \theta)^2}{nE_\theta\left[\left(\frac{f(x;\theta')}{f(x;\theta)} - 1\right)^2\right]}$$

The Chapman-Robbins-Kiefer (CRK) bound is a theoretical construct in the field of statistics and probability, and while it can be challenging to apply directly, especially in a non-numerical context, we can discuss its implications through conceptual examples.

Example 1: Estimation with Censored Data

Suppose we are dealing with survival data where the lifetime of certain components is being estimated, but some of the components are removed from the study before failure (right-censored data). The CRLB might not be applicable here because the lifetime distribution's support depends on the parameter (the life expectancy), which violates one of the regularity conditions of CRLB.

However, the CRK bound still provides a way to understand the lower bound on the variance of an estimator in this context. Even though we do not have a specific numerical bound, the CRK bound informs us that any estimator we use for the parameter (like the life expectancy) will have a variance that is at least as large as the CRK bound, which takes into account the censored nature of the data.

Example 2: Estimation in Discontinuous Distributions Imagine

we have a probability distribution that is not smooth and has discontinuities, perhaps representing a scenario where a sudden policy change affects the data generation process at a certain point. The CRLB might not be applicable because the probability density function is not differentiable everywhere with respect to the parameter of interest.

The CRK bound can still be applied to give us insight into the variance of estimators in this situation. Conceptually, it tells us there is a fundamental limit to how precise our estimates can be, given the "jumps" in the distribution, and this limit is encapsulated by the CRK bound, even though we might not explicitly calculate it.

Example 3: Estimation of a Parameter with Discrete and Continuous Components

Consider a complex system where the parameter of interest has both discrete and continuous components, such as a queueing system with a finite number of servers (discrete) and service times that are continuously distributed.

The CRLB is not immediately applicable due to the discrete nature of part of the parameter (number of servers). The CRK bound, on the other hand, provides a conceptual understanding that for any estimation procedure we adopt to estimate the system's efficiency or capacity, there is an inherent variance that cannot be surpassed as dictated by the CRK bound.

In all these examples, the CRK bound serves more as a theoretical guideline than a tool for direct application. It tells researchers and practitioners that there are limits to the precision of their estimates, and these limits are not just a factor of the sample size or the measurement precision, but also a fundamental property of the statistical models and the estimation problems they are working with. The specific value of the CRK bound in each case would depend on the details of the distribution and the estimator used, and calculating it would require numerical methods.

10.6 Asymptotic Normality

Asymptotic normality is a fundamental concept in the field of statistics, particularly in the study of estimator sequences. An estimator is said to be asymptotically normal if, as the sample size increases indefinitely, the distribution of the estimator, after proper standardization, converges to a normal distribution. This property is crucial for making statistical inferences because it allows for the application of normal distribution theory to construct confidence intervals and hypothesis tests.

Definition: An estimator $\hat{\theta}_n$ for a parameter θ, based on a sample of size n, is said to be asymptotically normal if there exists a sequence of positive numbers {an} such that the standardized estimator:

$$\frac{\hat{\theta}_n - \theta}{a_n}$$

converges in distribution to a standard normal distribution as n→∞. This is often written as:

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) \xrightarrow{d} N\ (0, \sigma^2)$$

where $\xrightarrow{d}$ denotes convergence in distribution, $N\ (0, \sigma^2)$ represents the normal distribution with mean 0 and variance σ^2, and σ^2 is the asymptotic variance of the estimator. The scale factor $\sqrt{n}$ is often used because many estimators, under certain regularity conditions, have variances that decrease at a rate of 1/n, making the product $\sqrt{n}\left(\hat{\theta}_n - \theta\right)$ have a stable variance as n increases.

For many estimators, particularly Maximum Likelihood Estimators (MLEs), under certain regularity conditions, the asymptotic normality can be established. This is a powerful result as it allows the usage of the normal distribution for inferential purposes even when the original data or the sampling distribution of the estimator is not normal, provided the sample size is sufficiently large.

Asymptotic normality is a fundamental concept in statistics, particularly in the realm of inferential techniques, and it plays a central role in the application of many statistical methods.

Applications of Asymptotic Normality:

Simplification of Inference: Asymptotic normality simplifies the process of making statistical inferences because the normal distribution is well-understood and tables of its distribution are widely available.

Central Limit Theorem: It forms the basis for the Central Limit Theorem (CLT), which allows for the normal approximation of the sampling distribution of sums (and means) of independent random variables, regardless of their original distribution, given a sufficiently large sample size.

Econometrics: In econometrics, it is often assumed that estimators are asymptotically normal, which allows for straightforward hypothesis testing and the construction of confidence intervals for economic parameters.

Method of Moments and Maximum Likelihood Estimation: Asymptotic normality is used to derive the properties of estimators obtained through methods like the Method of Moments and Maximum Likelihood Estimation (MLE). These estimators are often asymptotically normal, allowing for standard inference procedures.

Large Sample Theory: It is a key concept in large sample theory, which is a cornerstone of theoretical statistics and underpins the asymptotic analysis of estimators and test statistics.

Limitations of Asymptotic Normality

Sample Size: The main limitation is that it pertains to the asymptotic, i.e., large-sample, behaviour of estimators. In small samples, the distribution of the estimator may not be well approximated by a normal distribution.

Speed of Convergence: The speed at which the distribution of an estimator converges to the normal can vary widely between different estimators. For some estimators and distributions, this convergence may be slow, and normal approximation may not be appropriate for practical sample sizes.

Dependence on Conditions: The assurance of asymptotic normality often depends on specific conditions being met (e.g., i.i.d. samples, finite variance), which may not hold in all practical situations, particularly in the presence of dependent or heavy-tailed data.

Robustness: Asymptotically normal estimators might not be robust to outliers or deviations from model assumptions. If the underlying assumptions are violated, the estimator's distribution may significantly deviate from normality.

Misapplication: There may be a tendency to apply normal-based methods inappropriately or interpret results too optimistically, especially when sample sizes are not sufficiently large for the central limit theorem to kick in.

Tail Behaviour: Asymptotic normality does not provide information about the tail behavior of an estimator's distribution, which can be important in risk assessment and extreme value theory.

In practical applications, researchers must carefully consider these limitations when using asymptotically normal estimators, especially in terms of sample size and the nature of the data. In some cases, bootstrap methods or other resampling techniques may provide a more accurate inference in finite samples.

10.7 Self-Assessment Exercises

1. What are the key properties that define a good estimator, such as unbiasedness, consistency, and efficiency?
2. Provide examples to illustrate each property in practical situations.
3. Define the Bhattacharya Bound and explain its significance in assessing the quality of biased estimators.
4. How does it compare to the Cramér-Rao Lower Bound (CRLB)?
5. Provide a real-world example where the Bhattacharya Bound is applied.
6. Describe the CRK Bound and explain how it improves upon the CRLB in complex or non-standard scenarios.
7. Illustrate with an example where the CRK Bound is more appropriate than the CRLB.
8. What is asymptotic normality, and why is it important for large-sample inference?
9. Provide an example of an estimator that becomes asymptotically normal and explain how this property aids in practical applications like hypothesis testing or constructing confidence intervals.
10. Explain the differences between BAN and CAN estimators in terms of their efficiency and consistency properties.
11. Provide examples of situations where BAN and CAN estimators would be preferred.
12. Define asymptotic efficiency and describe how it relates to the CRLB.
13. Why is asymptotic efficiency a critical property in large-sample estimation?

Short Answer Questions:

1. What is the importance of unbiasedness in an estimator? Give an example of an unbiased estimator.
2. Define sufficiency in the context of statistical estimation. Why is it important?
3. Explain the Chapman-Robbins-Kiefer Bound in simple terms.

4. Differentiate between biased and unbiased estimators. Can a biased estimator sometimes be more efficient than an unbiased estimator?
5. What is the role of asymptotic normality in statistical estimation?

10.8 Summary

In this unit, we explored the essential criteria that define a good estimator in statistical estimation. The primary focus was on understanding the properties that ensure an estimator provides reliable, efficient, and accurate results when used to estimate unknown parameters from sample data. The unit covered the key concepts of unbiasedness, consistency, and efficiency, which are fundamental to selecting and evaluating estimators.

Unbiasedness was discussed as a crucial property, where an estimator is considered unbiased if its expected value equals the true parameter being estimated. This ensures that, on average, the estimator produces the correct value without systematic error. However, unbiasedness alone does not guarantee the optimal performance of an estimator, as it must also minimize variance to be considered efficient. Efficiency was defined in terms of the Cramér-Rao Lower Bound (CRLB), which provides a theoretical benchmark for the minimum variance that an unbiased estimator can achieve. An estimator that meets this bound is considered efficient, ensuring it extracts the maximum possible information from the sample data.

The unit also introduced the Bhattacharya Bound, which expands the evaluation of estimator quality to biased estimators. While the CRLB is restricted to unbiased estimators, the Bhattacharya Bound allows for the assessment of biased estimators by setting a lower bound on their mean squared error (MSE). This is particularly important in practical applications where biased estimators may offer advantages, such as lower variance, over unbiased alternatives. Similarly, the Chapman-Robbins-Kiefer (CRK) Bound was explored as a refinement of the CRLB, offering tighter bounds on estimator variance, particularly in cases where regularity conditions do not hold.

A significant portion of the unit focused on asymptotic properties of estimators, particularly asymptotic normality and asymptotic efficiency. Asymptotic normality ensures that as the sample size grows, the distribution of an estimator approaches a normal

distribution, facilitating the use of normal distribution-based inference methods. This property is essential for constructing confidence intervals and conducting hypothesis tests in large-sample scenarios. The distinction between Best Asymptotic Normal (BAN) and Consistent Asymptotically Normal (CAN) estimators was also emphasized. BAN estimators are both asymptotically normal and efficient, achieving the lowest possible variance, while CAN estimators ensure consistency, converging to the true parameter as the sample size increases.

The concept of sufficiency was also introduced, highlighting the importance of using all relevant information in the sample to make the most accurate estimation. A sufficient statistic captures all the information about the parameter of interest, ensuring that no information is lost in the estimation process.

In summary, this unit provided a comprehensive framework for evaluating estimators based on their properties of unbiasedness, consistency, efficiency, and sufficiency. The exploration of the Bhattacharya and CRK bounds extended the analysis to biased estimators, offering a broader perspective on estimator quality. By understanding these criteria, students are better equipped to select appropriate estimators for various statistical problems, ensuring reliable and accurate parameter estimation in both theoretical and applied contexts.

10.9 References

- Lehmann, E. L., & Casella, G. (2006). *Theory of Point Estimation* (2nd ed.). Springer.
- Kay, S. M. (1993). *Fundamentals of Statistical Signal Processing: Estimation Theory* (Vol. 1). Prentice Hall.
- Casella, G., & Berger, R. L. (2002). *Statistical Inference* (2nd ed.). Duxbury Press.
- Bickel, P. J., & Doksum, K. A. (2015). *Mathematical Statistics: Basic Ideas and Selected Topics* (2nd ed.). CRC Press.
- Rao, C. R. (1973). *Linear Statistical Inference and Its Applications* (2nd ed.). Wiley.
- Ferguson, T. S. (1996). *A Course in Large Sample Theory*. CRC Press.

- Stuart, A., & Ord, K. (1994). *Kendall's Advanced Theory of Statistics* (6th ed., Vol. 1). Wiley.
- Van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press.
- Zehna, P. W. (1966). Invariance of Maximum Likelihood Estimators. *The Annals of Mathematical Statistics*, 37(3), 744–748.
- Huzurbazar, V. V. (1976). *Sufficiency and Statistical Inference*. Academic Press.

10.10 Further Readings

- Cramér, H. (1946). *Mathematical Methods of Statistics*. Princeton University Press.
- Rao, C. R. (1973). *Linear Statistical Inference and Its Applications*. John Wiley & Sons.
- Ferguson, T. S. (1996). *A Course in Large Sample Theory*. Chapman & Hall/CRC.
- Lehmann, E. L., & Casella, G. (1998). *Theory of Point Estimation* (2nd ed.). Springer.
- Mood, A. M., Graybill, F. A., & Boes, D. C. (1974). *Introduction to the Theory of Statistics* (3rd ed.). McGraw-Hill.
- Van der Vaart, A. W. (2000). *Asymptotic Statistics*. Cambridge University Press.
- Wasserman, L. (2004). *All of Statistics: A Concise Course in Statistical Inference*. Springer.
- Casella, G., & Berger, R. L. (2002). *Statistical Inference* (2nd ed.). Duxbury Press.
- Zehna, P. W. (1966). "Necessary and Sufficient Conditions for MVUE." *The Annals of Mathematical Statistics*, 37(3), 701-708.
- Huzurbazar, V. S. (1950). "Sufficiency and Statistical Decision Functions." *The Annals of Mathematical Statistics*, 21(4), 556-569.
- Stuart, A., & Ord, K. (1994). *Kendall's Advanced Theory of Statistics* (6th ed., Vol. 1). Wiley.

- Bickel, P. J., & Doksum, K. A. (2015). *Mathematical Statistics: Basic Ideas and Selected Topics* (2nd ed.). CRC Press.
- Fisher, R. A. (1925). *Statistical Methods for Research Workers*. Oliver & Boyd.

UNIT - 11: CRITERION FOR GOOD ESTIMATORS-II

Structure

11.1 Introduction

Estimation theory forms the backbone of statistical inference, where the objective is to deduce unknown population parameters using sample data. The evolution of this field has been driven by the need for estimators that not only provide accurate predictions but also ensure efficiency and reliability. In modern statistics, the emphasis is not just on obtaining estimates, but on ensuring that these estimators exhibit properties like consistency, asymptotic normality, and efficiency.

Historically, much of estimation theory was developed during the early 20th century, led by the pioneering work of statisticians like Sir Ronald A. Fisher. Fisher's **Maximum Likelihood Estimation (MLE)** method revolutionized estimation by providing a systematic framework to estimate parameters that maximize the probability of observed data. As statistical problems became more complex, there was a growing need for estimators that performed well with large datasets and could be applied in diverse fields such as economics, biology, and engineering.

Asymptotic properties of estimators became increasingly important as large samples became more common in real-world data. The focus shifted towards ensuring that estimators

remain consistent and efficient as sample size grows. **Bounded Asymptotic Normal (BAN)** and **Consistent Asymptotically Normal (CAN)** estimators emerged as key concepts, addressing the need for estimators that not only converge to the true parameter values but also exhibit minimal variance in large samples. These concepts are integral to modern estimation, where the performance of estimators in large samples is a critical criterion for their application.

BAN estimators are a class of estimators that are not only asymptotically normal but also asymptotically efficient. That is, they have the smallest possible variance among unbiased estimators as sample size increases, ensuring that they are the best possible estimators in large samples. Meanwhile, **CAN estimators** are those that are consistent, meaning they converge to the true parameter value as the sample size grows, and also exhibit asymptotic normality. While all BAN estimators are CAN, not all CAN estimators achieve the same level of efficiency as BAN estimators, making the distinction between them crucial for choosing the right estimator in practice.

Another important concept covered in this unit is **asymptotic efficiency**, which characterizes the performance of an estimator in large samples. An estimator is asymptotically efficient if it reaches the **Cramér-Rao Lower Bound (CRLB)**, the theoretical minimum variance for an unbiased estimator. This property ensures that the estimator extracts the maximum possible information from the data as the sample size increases, which is particularly important in fields like econometrics and biostatistics.

In addition, the concept of **equivariant consistency** plays a key role in ensuring that estimators behave predictably under transformations. An estimator is said to be equivariantly consistent if it remains consistent under transformations of the data, such as changes in units or scales. This is important in practical scenarios where data transformations are common, and the estimator needs to adapt accordingly while maintaining its consistency.

This unit provides a deep dive into these advanced estimation concepts, offering learners the tools to assess and select appropriate estimators based on their asymptotic properties. The emphasis is on understanding how these properties ensure the reliability and efficiency of estimators in real-world applications, especially when working with large datasets. By the end of this unit, learners will have a comprehensive understanding of the theoretical framework underpinning BAN and CAN estimators, asymptotic efficiency, and

equivariant consistency, all of which are essential for making accurate and efficient inferences from data.

11.2 Objectives

By the end of this unit, learners will be able to:

1. **Understand the Concept of Estimators in Large Samples:**
    - Grasp the fundamental role of estimators in statistical inference, specifically focusing on their behaviour in large samples.
    - Identify the importance of asymptotic properties like consistency, normality, and efficiency in choosing an appropriate estimator.
2. **Comprehend the Properties of Bounded Asymptotic Normal (BAN) Estimators:**
    - Define a Bounded Asymptotic Normal (BAN) estimator and explain its significance in large-sample estimation.
    - Understand the conditions under which an estimator is considered BAN and how this property ensures asymptotic efficiency.
    - Recognize the importance of achieving minimal variance in unbiased estimators and how BAN estimators fulfill this criterion in large datasets.
3. **Understand the Characteristics of Consistent Asymptotically Normal (CAN) Estimators:**
    - Define Consistent Asymptotically Normal (CAN) estimators and understand their role in ensuring the reliability of estimation as sample size increases.
    - Explain the difference between BAN and CAN estimators, particularly in terms of variance and efficiency, and understand why all BAN estimators are CAN, but not all CAN estimators are BAN.
    - Apply the concept of asymptotic normality in constructing confidence intervals and performing hypothesis tests in large samples.
4. **Apply the Concept of Asymptotic Efficiency:**
    - Define asymptotic efficiency and understand its role in comparing the performance of different estimators in large samples.

- Explain how asymptotic efficiency is related to the Cramér-Rao Lower Bound (CRLB), and how an efficient estimator achieves the lowest possible variance.
- Explore real-world examples where asymptotically efficient estimators are crucial, such as in econometrics, biostatistics, and machine learning, where large datasets are the norm.

5. **Explore the Concept of Equivariant Consistency:**
   - Understand what it means for an estimator to be equivariantly consistent, ensuring that the estimator's performance remains stable under data transformations.
   - Learn how equivariant consistency is important in practical applications where data might undergo transformations such as scaling, shifts, or changes in units.
   - Apply the concept of equivariant consistency in real-world statistical problems to ensure robust and adaptable estimation processes.
6. **Analyse the Relationship Between BAN, CAN, and Other Estimation Methods:**
   - Compare BAN and CAN estimators with other common estimation methods such as Maximum Likelihood Estimation (MLE) and the Method of Moments (MoM), particularly in terms of their asymptotic properties.
   - Understand the role of regularity conditions and assumptions in ensuring the performance of BAN and CAN estimators.
   - Explore the theoretical foundations that guarantee the asymptotic properties of estimators and apply these principles to evaluate the performance of estimators in different contexts.
7. **Solve Practical Problems Using Asymptotic Estimation Techniques:**
   - Apply BAN and CAN estimators to practical problems involving large datasets, ensuring accurate and efficient parameter estimation.
   - Use the properties of asymptotic normality and efficiency to construct confidence intervals and conduct hypothesis testing in real-world scenarios.
   - Evaluate the performance of different estimators in terms of their asymptotic behavior, selecting the most appropriate estimator based on efficiency and consistency.

By mastering these objectives, learners will develop a thorough understanding of advanced estimation techniques, with a focus on their application in large sample scenarios. They will be equipped to make informed decisions on selecting appropriate estimators based on properties like bounded normality, consistency, and efficiency, and apply these concepts to real-world statistical problems.

11.3 Bounded Asymptotic Normal (BAN) Estimator

In the field of estimation theory, **Bounded Asymptotic Normal (BAN)** estimators play a crucial role, particularly in the context of large samples. BAN estimators are a class of estimators that exhibit desirable asymptotic properties, making them highly effective for statistical inference when the sample size is large. The central concept behind BAN estimators is that they possess asymptotic unbiasedness, minimum variance, and asymptotic normality, which means they approach the properties of an ideal estimator as the sample size grows indefinitely.

The "bounded" property in BAN estimators refers to the fact that the variance of the estimator is bounded asymptotically, ensuring that the estimator's precision does not deteriorate even as the sample size increases. BAN estimators are considered to be **best** among asymptotically normal estimators because they not only provide accurate parameter estimates but also achieve the lowest possible variance in large samples. This property makes BAN estimators optimal for many large-sample problems encountered in real-world applications, such as econometrics, biological research, and engineering.

Key Properties of BAN Estimators:

1. **Asymptotic Unbiasedness:** As the sample size increases, the expected value of the BAN estimator converges to the true parameter being estimated. This means that any systematic bias in the estimation process diminishes with increasing sample size, ensuring that the estimator is accurate in the long run.

2. **Asymptotically Minimum Variance:** Among the class of asymptotically unbiased estimators, a BAN estimator achieves the lowest possible variance as the sample size approaches infinity. This property is critical for large-sample efficiency, as it

guarantees that the estimator is the most precise in the asymptotic sense, minimizing the uncertainty around the estimated parameter.

3. **Asymptotic Normality:** A BAN estimator's distribution, when properly normalized, converges to a normal distribution as the sample size becomes large. This property is essential because it allows the use of normal distribution-based inference methods, such as constructing confidence intervals and conducting hypothesis tests.

4. **Bounded Variance:** The variance of a BAN estimator does not grow unbounded as the sample size increases. This boundedness ensures the estimator's stability and reliability in large samples, as it prevents excessive variability in the estimates even in very large datasets.

Important Results of BAN Estimators:

- **Asymptotic Efficiency:** BAN estimators achieve asymptotic efficiency, meaning they reach the **Cramér-Rao Lower Bound (CRLB)** for the variance of unbiased estimators in large samples. This makes them the best possible estimators in terms of precision for large datasets, as no other unbiased estimator can have a smaller variance asymptotically.

- **Maximum Likelihood Estimators (MLE) as BAN Estimators:** In many cases, Maximum Likelihood Estimators (MLE) are BAN estimators under regularity conditions. The MLE has been shown to be asymptotically normal and efficient, making it one of the most widely used estimation methods in large-sample inference. Under these conditions, the MLE achieves the smallest asymptotic variance, ensuring optimal estimation performance.

- **BAN Estimators and Sufficiency:** BAN estimators are often linked to the concept of **sufficient statistics**, where the estimator fully captures all the relevant information in the sample data. This sufficiency allows the BAN estimator to provide the most efficient and accurate estimates by using all the information available in the data.

Real-Life Examples of BAN Estimators:

1. **Econometric Analysis in Large Datasets:** In econometrics, researchers often deal with large datasets to estimate parameters such as the impact of economic policies on employment or inflation rates. BAN estimators are frequently used in these situations to ensure that the parameter estimates are both unbiased and efficient as the sample size increases. For example, when estimating the effect of interest rates on consumer spending using large national datasets, BAN estimators provide precise and reliable estimates with minimal variance.

2. **Estimating the Mean in Biological Studies:** In large-scale biological experiments, such as estimating the average height of a population of plants, BAN estimators ensure that the mean estimate is both accurate and precise as the number of samples grows. As the experiment involves hundreds or thousands of plants, BAN estimators offer stability by achieving the lowest possible variance, making them ideal for large-sample scenarios in biological research.

3. **Estimating the Failure Rate in Engineering Systems:** Engineers often use BAN estimators when analysing the reliability of systems, such as estimating the failure rate of a machine component over time. In cases where data is collected over long periods or across numerous systems, BAN estimators ensure that the estimated failure rate is unbiased and has minimal variability. For example, in the aerospace industry, where failure rates of critical components are monitored across multiple flights, BAN estimators help ensure that the failure probability is estimated with the highest possible precision.

4. **Survey Research in Social Sciences:** In large-scale survey research, such as estimating the proportion of people who support a particular policy based on national surveys, BAN estimators provide unbiased and efficient estimates. As sample sizes grow into the thousands or millions, BAN estimators ensure that the estimated proportion is both accurate and stable, making them essential for reliable policy analysis and decision-making.

Bounded Asymptotic Normal (BAN) estimators are fundamental tools in modern statistical estimation, particularly in large-sample inference. Their properties of asymptotic unbiasedness, minimal variance, and normality make them optimal for situations where large datasets are common, such as in econometrics, biology, engineering, and social sciences. BAN estimators are efficient, achieving the lowest possible variance asymptotically, which makes them a preferred choice for precise parameter estimation in large samples. Their application across various fields highlights their versatility and importance in providing reliable statistical inferences in real-world scenarios.

11.4 CAN Estimators (Consistent Asymptotically Normal)

In statistical estimation, an estimator's performance often hinges on its behaviour as the sample size increases. This is where **Consistent Asymptotically Normal (CAN)** estimators become critical. A CAN estimator is defined by two key properties: consistency and asymptotic normality. These properties ensure that, as the sample size grows indefinitely, the estimator converges to the true parameter value (consistency) and its distribution approaches a normal distribution (asymptotic normality).

CAN estimators are particularly valuable in the context of large samples, where their asymptotic properties facilitate reliable and efficient inference. While they may not always have the smallest variance among consistent estimators (as opposed to BAN estimators), their asymptotic normality allows for the application of familiar statistical techniques, such as constructing confidence intervals and performing hypothesis tests using the normal distribution. CAN estimators play a central role in various fields, including econometrics, biostatistics, and engineering, where large datasets are common and precision in estimation is critical.

Key Properties of CAN Estimators:

1. **Consistency:**
    - A CAN estimator is consistent if, as the sample size increases, it converges in probability to the true value of the parameter it is estimating. This means that with enough data, the CAN estimator will produce estimates that are arbitrarily

close to the true parameter. Consistency is a fundamental requirement for any good estimator, as it ensures that the estimator is reliable in the long run.

2. **Asymptotic Normality:**
   - Asymptotic normality means that the distribution of the CAN estimator, when appropriately normalized (scaled and centred), approaches a normal distribution as the sample size increases. This property is crucial because it enables the use of normal-based inferential methods, such as confidence intervals and hypothesis testing, even when the sample data may not initially follow a normal distribution.

3. **Practical Utility in Large Samples:**
   - While CAN estimators may not always achieve the minimum possible variance (unlike BAN estimators), their asymptotic normality ensures that they perform well in large samples. In real-world applications, where data is abundant, CAN estimators provide reliable estimates that converge to the true values with increasing sample size, while also simplifying the inference process through their normal approximation.

Important Results for CAN Estimators:

1. **CAN Estimators and Maximum Likelihood Estimation (MLE):**
   - Under regularity conditions, **Maximum Likelihood Estimators (MLE)** are often CAN estimators. This means that MLEs not only converge to the true parameter value as the sample size grows but also exhibit asymptotic normality, allowing for normal-based inference. MLE's asymptotic normality simplifies statistical procedures, making it easier to derive confidence intervals and perform hypothesis tests in large-sample contexts.

2. **CAN Estimators and Efficiency:**
   - CAN estimators do not necessarily have the smallest possible variance among consistent estimators (unlike BAN estimators). However, they still offer practical utility, particularly in cases where finding the most efficient estimator may be difficult or computationally intensive. The ease of computation and the

robustness of CAN estimators often make them preferable in applied settings, where computational simplicity and consistency are key.

3. **Relationship Between CAN and BAN Estimators:**
    - All BAN estimators are CAN, but not all CAN estimators are BAN. The distinction lies in the efficiency of the estimator: while BAN estimators achieve the lowest possible asymptotic variance, CAN estimators may be less efficient but still maintain consistency and asymptotic normality. This distinction is important when selecting estimators in practice, especially when dealing with large datasets where computational efficiency is a consideration.

Real-Life Examples of CAN Estimators:

1. **Econometric Models:**
    - In econometrics, where researchers often estimate parameters such as the relationship between variables like income and consumption, CAN estimators are used extensively. For example, when estimating the parameters of a regression model involving a large number of observations, the Ordinary Least Squares (OLS) estimator, under certain regularity conditions, becomes a CAN estimator. While it may not always be the most efficient estimator, its consistency and asymptotic normality make it a reliable choice in large samples.

2. **Survey Sampling:**
    - In national surveys that aim to estimate population parameters like the average household income, CAN estimators are commonly employed. For instance, a simple random sample from a large population can yield a CAN estimator for the average income, ensuring that the estimate converges to the true population mean as the sample size increases. Despite not being the most efficient estimator, its asymptotic normality allows for easy inference using normal distribution techniques, such as confidence intervals.

3. **Clinical Trials in Biostatistics:**

- In large clinical trials, where researchers estimate the effect of a new drug or treatment, CAN estimators are often used to ensure that the estimated treatment effect converges to the true effect as the sample size grows. For example, when estimating the mean difference in treatment outcomes between two groups, a CAN estimator ensures that the estimate will be accurate and reliable with sufficient data. Moreover, its asymptotic normality enables the construction of confidence intervals around the estimated treatment effect, facilitating hypothesis testing.

4. **Quality Control in Manufacturing:**
    - In quality control processes in manufacturing industries, where companies monitor the rate of defective products over time, CAN estimators are used to estimate the true defect rate. As more data is collected, the CAN estimator for the defect rate becomes increasingly accurate, converging to the true rate as the sample size grows. This is particularly useful for long-term monitoring, where the normality property allows for easy hypothesis testing to determine whether the defect rate has changed significantly.

Consistent Asymptotically Normal (CAN) estimators are fundamental tools in modern statistical estimation, especially when dealing with large datasets. Their properties of consistency and asymptotic normality ensure that the estimator becomes increasingly accurate with more data, while also enabling straightforward inference using normal distribution techniques. While CAN estimators may not always achieve the smallest variance, their ease of use, computational simplicity, and robustness make them invaluable in real-world applications ranging from econometrics and biostatistics to survey research and quality control. By understanding the properties and applications of CAN estimators, statisticians and researchers can make informed decisions in selecting appropriate estimation methods for large-sample problems.

The Relationship Between BAN and CAN Estimators

All BAN estimators are CAN, but not all CAN estimators are BAN. This is because BAN estimators not only have to be consistent and asymptotically normal, but they also have to have the smallest possible asymptotic variance.

BAN estimators are often obtained through methods like Maximum Likelihood Estimation (MLE) under regularity conditions, which guarantee that the MLE is efficient (in the sense of having minimum variance) and asymptotically normal.

CAN estimators, on the other hand, might be obtained through other methods, like Method of Moments or less efficient versions of MLE, and while they are consistent and become normally distributed as the sample size increases, they may not achieve the lowest possible variance.

Applications and Importance:

Statistical Inference: Both BAN and CAN estimators are widely used for statistical inference because their asymptotic properties facilitate the creation of confidence intervals and the performance of hypothesis tests.

Econometrics: In econometric analyses, where large datasets are common, these estimators are essential for consistent and efficient estimation of economic parameters.

Practical Preference: In practice, BAN estimators are preferred for large sample inference due to their efficiency, but CAN estimators may still be useful, especially if they are easier to compute or if the efficiency loss is not substantial.

Limitations:

Finite Samples: The properties of BAN and CAN estimators are asymptotic, meaning they may not hold in small samples. Practitioners must be cautious when applying results based on these properties to finite samples.

Assumptions: The asymptotic properties often rely on certain assumptions (e.g., i.i.d. samples, regularity conditions) that may not hold in practice, which could lead to invalid inferences.

In summary, BAN and CAN estimators are foundational concepts in the theory of statistical estimation, with BAN estimators being a subset of CAN estimators that also possess the property of having the smallest asymptotic variance. Understanding these concepts is crucial for statisticians and researchers who work with large datasets and rely on asymptotic theory to guide their inference.

11.5 Asymptotic Efficiency

Asymptotic efficiency is a concept in statistical estimation theory that characterizes the performance of an estimator as the sample size becomes large. It is a property of point estimators in the context of large-sample, or asymptotic, properties.

Definition of Asymptotic Efficiency:

An estimator is said to be asymptotically efficient if it meets two key criteria:

Consistency: The estimator must be consistent; that is, it converges in probability to the true parameter value as the sample size grows without bound. This means that the estimator's probability distribution collapses onto the true parameter value as the number of observations increases.

Asymptotic Normality with Minimum Variance: The distribution of the estimator, after being cantered and scaled (normalized), converges in distribution to a standard normal distribution as the sample size increases. Additionally, an asymptotically efficient estimator achieves the lowest possible variance among all consistent estimators, which is given by the Cramér-Rao Lower Bound (CRLB) in regular cases.

Explanation and Importance:

Benchmark for Comparing Estimators: Asymptotic efficiency serves as a benchmark to compare the performance of different estimators. An asymptotically efficient estimator is considered the best (in terms of variance) among the class of consistent estimators for large samples.

Role in Maximum Likelihood Estimation: In many cases, maximum likelihood estimators (MLEs) are asymptotically efficient under certain regularity conditions. This means that, for large samples, MLEs have desirable properties such as unbiasedness or nearly unbiased, and the variance of the estimator reaches the lower bound set by the Cramér-Rao inequality.

Practical Significance: Asymptotic efficiency is a desirable property because it means that the estimates become increasingly precise as more data is gathered. This efficiency is particularly important in fields like econometrics and biostatistics, where estimators are often applied to large datasets.

Influence on Statistical Procedures: The concept of asymptotic efficiency influences the development and choice of statistical procedures, especially in hypothesis testing and the construction of confidence intervals.

Asymptotic efficiency is an idealized property that assumes an infinite sample size. In practice, statisticians must also consider the finite-sample properties of an estimator and not rely solely on asymptotic efficiency, especially when the sample size is moderate or small.

Here are some non-numerical, conceptual examples that illustrate the principle of asymptotic efficiency:

Example 1: Estimating the Mean

Consider the problem of estimating the mean of a normal distribution. The sample mean is an asymptotically efficient estimator of the population mean because as the sample size increases, its variance approaches the inverse of the Fisher information in the sample, which meets the CRLB. This holds true regardless of whether we're measuring the average height of a population or the average temperature in a given city.

Example 2: Estimating Population Proportion

In a political survey, estimating the proportion of the population that supports a particular policy can be done using the sample proportion. As the sample size becomes very large, the sample proportion becomes asymptotically efficient for the true population

proportion, assuming a binomial distribution. This is because its variance approaches the CRLB for large sample sizes.

Example 3: Estimating the Rate of a Poisson Process

If we are looking at the rate of a certain type of event, like system failures per month, the number of events in a given time frame follows a Poisson distribution. The sample mean (the average number of events per time period) is an asymptotically efficient estimator of the event rate, as it achieves the CRLB as the number of observed time periods (or sample size) becomes large.

Example 4: Estimating Regression Coefficients

In linear regression, the ordinary least squares (OLS) estimators of the coefficients are asymptotically efficient under the Gauss-Markov assumptions, which include linearity, independence, and homoscedasticity of errors. This means that as the sample size increases, these estimators achieve the lowest possible variance among all linear unbiased estimators.

Example 5: Estimating Variance

When estimating the variance of a normal distribution, the unbiased sample variance (using $n-1$ in the denominator) is asymptotically efficient. It meets the CRLB asymptotically, which implies that no other unbiased estimator of variance will have a smaller variance for large sample sizes.

In each of these cases, the key point is not the numerical value of the variance or efficiency but the fact that as the sample size becomes very large, the estimator in question achieves the lowest possible variance among all unbiased estimators, which can be quantified through asymptotic analysis.

The concept of asymptotic efficiency is highly regarded in statistical theory, particularly in the context of large samples or when approaching the limit as the sample size goes to infinity. Let us delve into the applications and limitations of this concept.

Applications of Asymptotic Efficiency

Statistical Inference: Asymptotic efficiency is crucial for constructing confidence intervals and conducting hypothesis tests in large samples, as it ensures minimal width of confidence intervals and maximal power of hypothesis tests for a given sample size.

Model Selection: In econometrics and other fields that use regression models, the choice between different estimators can be guided by their asymptotic properties, including efficiency. This guides practitioners in choosing estimators that yield the most reliable inference in the limit.

Benchmarking: Asymptotic efficiency provides a benchmark for evaluating and comparing estimators. An estimator that is asymptotically efficient is often considered the gold standard against which other estimators are judged.

Econometric Theory: In econometrics, where models are often applied to large datasets, asymptotically efficient estimators are preferred because they are expected to exploit the information in the data most effectively as the sample size increases.

Machine Learning: Asymptotic properties can be important in machine learning, especially in understanding the long-run behaviour of algorithms as the amount of data approaches infinity, guiding algorithm design and selection.

Limitations of Asymptotic Efficiency

Finite Sample Relevance: Asymptotic efficiency does not necessarily reflect the properties of an estimator in small samples. An estimator that is efficient in the limit may not be the best choice for small or moderate sample sizes.

Assumptions May Not Hold: The theoretical results that guarantee asymptotic efficiency often rely on assumptions that may not hold in practice, such as the independence of observations or certain distributional properties.

Slow Convergence: For some estimators, the convergence to their asymptotic distribution can be slow, which means that even moderately large samples may not be "large enough" for the asymptotic properties to be relevant.

Practical Implementation: In practice, computing an asymptotically efficient estimator may be complex or computationally intensive compared to other more straightforward methods.

Robustness: Asymptotically efficient estimators may be sensitive to deviations from model assumptions, such as heteroscedasticity or non-normality of errors, and hence may not be robust in practice.

Misinterpretation of Asymptotic Results: There is a risk that asymptotic results are misinterpreted as applying to the finite sample scenario, which can lead to overconfidence in the results obtained using such estimators.

Model Dependence: Asymptotic efficiency is model-dependent, which means that an estimator's efficiency is only valid within the framework of a given model. If the model is mis specified, the estimator may not perform well.

In conclusion, while asymptotic efficiency is a desirable property and drives much of the theoretical development in statistics, its practical implications should be considered carefully. Researchers must assess whether the conditions for asymptotic efficiency are met and whether alternative estimators might be more suitable for finite samples or under model uncertainty.

11.6 Equivariant Consistency

Equivariant consistency is a concept that relates to the behaviour of estimators under transformations of the data. In statistics, an estimator is said to be equivariant if its form remains the same under certain transformations. Equivariant consistency refers to a property of such estimators when they are also consistent. Let us break this down further:

Equivariant Estimators: An estimator is called equivariant under a group of transformations if the estimates transform in the same way as the data under these transformations. For example, consider an estimator $\hat{\theta}(X)$ for a parameter θ based on data X. If g is a transformation function, then the estimator is equivariant if for every g, we have:

$$\hat{\theta}\ \mathrm{g(X)) = g(}\hat{\theta}(X)).$$

This means that applying the transformation to the data and then estimating the parameter should give the same result as estimating the parameter first and then applying the transformation to the estimate.

Consistent Estimators: A consistent estimator is one that converges in probability to the true parameter value as the sample size increases indefinitely. That is, for a parameter θ, an estimator $\hat{\theta}_n$ based on a sample size of n is consistent if:

$$\hat{\theta}_n \xrightarrow{P} \theta$$

as $n \to \infty$, where $\xrightarrow{P}$ denotes convergence in probability.

Equivariant consistency combines these two properties. An estimator is equivariantly consistent if it is both equivariant and consistent. In other words, it respects the symmetries of the parameter space under transformations even as it converges to the true parameter value with increasing sample size.

Example: Consider estimating the location parameter of a distribution. If we have an estimator $\hat{\mu}$ for the mean (a location parameter), and we add a constant c to every observation in our data, an equivariant estimator would satisfy:

$$\hat{\mu}\,(X+c) = \hat{\mu}\,(X) + c.$$

If this estimator $\hat{\mu}$ also converges in probability to the true mean μ as the sample size increases, then it is equivariantly consistent.

Importance: Equivariant consistency is important because it ensures that the estimator behaves in a predictable and sensible way under transformations, which is a desirable property in many practical situations. For instance, if one is estimating physical quantities that could be measured in different units (like pounds versus kilograms), an equivariantly consistent estimator would automatically adjust the estimates in accordance with the change in units.

11.7 Self-Assessment Exercises

1. What are the key properties that define a Bounded Asymptotic Normal (BAN) estimator?
2. How does the asymptotic normality of a BAN estimator assist in large-sample statistical inference?
3. Explain the difference between Bounded Asymptotic Normal (BAN) and Consistent Asymptotically Normal (CAN) estimators.
4. Under what conditions does the Maximum Likelihood Estimator (MLE) serve as a BAN estimator?
5. Define consistency and explain why it is crucial for an estimator to be considered consistent in large samples.
6. How does the asymptotic efficiency of BAN estimators relate to the Cramér-Rao Lower Bound (CRLB)?
7. Give a real-life example where a CAN estimator would be preferable to a BAN estimator and explain why.
8. Why are CAN estimators important in situations involving large datasets, even if they do not always minimize variance?
9. Explain the role of sufficiency in BAN estimators and how it relates to achieving minimum variance.
10. How can the normality property of CAN estimators be applied to hypothesis testing and constructing confidence intervals?

11.8 Summary

This unit delves into the key concepts of Bounded Asymptotic Normal (BAN) and Consistent Asymptotically Normal (CAN) estimators, emphasizing their importance in large-sample estimation. BAN estimators, known for their asymptotic unbiasedness, minimum variance, and normality, are recognized as efficient estimators when the sample size increases indefinitely. BAN estimators, particularly under conditions like those required by Maximum Likelihood Estimation (MLE), achieve optimal performance by reaching the smallest possible variance. On the other hand, CAN estimators, while ensuring consistency and

normality, may not always have the smallest variance but offer robust performance in large datasets, making them essential in practical applications.

The unit also explores the relationship between BAN and CAN estimators, highlighting that while all BAN estimators are CAN, not all CAN estimators are BAN due to the variance efficiency differences. The concept of asymptotic efficiency was discussed in relation to the Cramér-Rao Lower Bound (CRLB), providing a benchmark for estimator performance. Furthermore, the practical application of these estimators across various fields such as econometrics, biostatistics, and engineering was illustrated, emphasizing the usefulness of these estimation methods in real-world problems. The focus on large-sample properties like normality and variance underscores their critical role in statistical inference, particularly for constructing confidence intervals and performing hypothesis tests. Understanding these concepts equips researchers with the tools needed to make reliable and efficient inferences from large datasets.

11.9 References

- Casella, G., & Berger, R. L. (2002). *Statistical Inference* (2nd ed.). Duxbury Press.
- Lehmann, E. L., & Casella, G. (1998). *Theory of Point Estimation* (2nd ed.). Springer.
- Hogg, R. V., McKean, J. W., & Craig, A. T. (2018). *Introduction to Mathematical Statistics* (8th ed.). Pearson.
- Rao, C. R. (1973). *Linear Statistical Inference and Its Applications* (2nd ed.). Wiley.
- Cramér, H. (1946). *Mathematical Methods of Statistics*. Princeton University Press.
- Fisher, R. A. (1925). *Statistical Methods for Research Workers*. Oliver & Boyd.
- Huzurbazar, V. S. (1976). *Sufficiency and Statistical Decision Functions*. Academic Press.

11.10 Further Readings

- Ferguson, T. S. (1996). *A Course in Large Sample Theory*. Chapman & Hall.
- Schervish, M. J. (1995). *Theory of Statistics*. Springer.
- Spanos, A. (1999). *Probability Theory and Statistical Inference*. Cambridge University Press.
- van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press.
- Lehmann, E. L. (2004). *Elements of Large-Sample Theory*. Springer.

UNIT-12 CONFIDENCE ESTIMATION

Structure

12.1 Introduction

In the unit of Estimation, we have studied the methods which provide the estimate of the value of one or many unknown parameters. These methods provide functions of sample value which is called the estimator. These estimators provide unique estimate for a given sample. As it is obvious, that these estimates may differ from the parameters in many cases. Therefore, a margin of uncertainty exists. The sampling variance of the estimator expresses the extent of the uncertainty. We can say that it is probable that θ (parameter) lies in the range $t \pm \sqrt{\text{var}(t)}$. Very probable that it lies in the range $t \pm 2\sqrt{\text{var}(t)}$ and so on. We now abandon the attempts to estimate θ by a function which, for a specified sample, gives a unique estimate. Instead, we shall consider the specification of a range in which θ lies. This method is known as method of **confidence interval.**

12.2 Objectives

After studying this a learner will able to understand about confidence interval and confidence coefficient. You will able to calculate confidence interval for mean, variance, etc of different distribution. You will also understand the meaning of shortest confidence interval and method of calculating it. You will know about the relationship between Confidence Estimation and Hypothesis Testing.

12.3 Confidence Interval and Confidence Coefficient

In this unit we will discuss the basic ideas and method of confidence interval estimation, which is due to Neyman (1937b). In short, we can say that problem of estimating the actual values of population parameters is said to belong to "Point Estimation". It is equally important to know within what limits we can confidently expect the parameter values to lie with any specified degree of confidence and such problems belong to "***Interval Estimation.***

Let T_1 and T_2 are two statistics, which is two measurable functions of the random variable $X_1,\ldots,X_n$ also, let T_1 and T_2 $(T_1 < T_2)$ be such that

$$P(T_1 > \theta) = \alpha_1 \qquad (12.1)$$

$$P(T_2 < \theta) = \alpha_2 \qquad (12.2)$$

Where α_1 and α_2 are independent of θ. Putting $\alpha_1 + \alpha_2 = \alpha$, we have the result that

$$P[T_1 \le \theta \le T_2] = 1-\alpha \qquad (12.3)$$

and make the assertion that θ lies in the interval T_1 to T_2, which is called a confidence interval for θ. T_1 and T_2 are known as the lower and upper confidence limits respectively. They depend only on $1-\alpha$ and the sample values. For any fixed $1-\alpha$, the totality of confidence intervals for different samples determines a field within which θ is supposed to

lie. This field is called the ***confidence belt***. We shall give a graphical representation of the idea below. The fraction $1-\alpha$ is called the ***confidence coefficient.***

Example-1 : Suppose we have a sample of size n from the normal population with mean μ and variance unity.

$$f(x,\mu;1)=\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}(x-\mu)^2\right\}dx, \qquad -\infty \le x \le \infty \qquad (12.4)$$

The distribution of the sample mean $\bar{x}$ is

$$f\left(\bar{x},\mu;1\right)=\sqrt{\frac{n}{2\pi}}\exp\left\{-\frac{n}{2}(x-\mu)^2\right\}d\bar{x} \qquad -\infty \le \bar{x} \le \infty \qquad (12.5)$$

From the tables of the normal integral we know that the probability of a positive deviation from the mean not greater than twice the standard deviation is 0.97725. We have then

$$P\left(\bar{x}-\mu \le 2/\sqrt{n}\right)=0.97725$$

Which is equivalent to $$P\left(\bar{x}-2/\sqrt{n} \le \mu\right)=0.97725$$

Thus, if we assert that μ is greater than or equal to $\bar{x}-2/\sqrt{n}$ we shall be accurate in about 97.725 percent of cases in the long run.

Similarly, we have $$P\left(\bar{x}-\mu \ge -2/\sqrt{n}\right)=P\left(\mu \le \bar{x}+\ 2/\sqrt{n}\right)=0.97725$$

Thus, combining the two results,

$$P\left(\bar{x}-2/\sqrt{n}\le\mu\le\bar{x}+2/\sqrt{n}\right)=1-2(1-0.97725)=0.9545 \tag{12.6}$$

Hence, if we assert that μ lies in the range $\bar{x}\pm 2/\sqrt{n},$ we shall be accurate in about 95.45 percent of cases in the long run.

Conversely, given the confidence coefficient, we can easily find from the tables of the normal integral the deviation d such that

$$P\left(\bar{x}-d/\sqrt{n}\le\mu\le\bar{x}+d/\sqrt{n}\right)=1-\alpha \tag{12.7}$$

For instance, if $1-\alpha=0.8, d=1.28,$ so that if we assert that μ lies in the range $\bar{x}\pm 1.28/\sqrt{n}$, the odds are 4 to 1 that we shall be right.

Another point of interest in this example is that the upper and lower confidence limits which is derived above are equidistant from the mean $\bar{x}$. This is not by any means necessary, and it is easy to see that we can obtain any number of alternative limits for the same confidence coefficient $1-\alpha$. Assume, for instance, we take $1-\alpha=0.9545$, and select two numbers α_1 and α_2 which obey the condition $\alpha_1+\alpha_2=\alpha=0.0455$, say

$\alpha_1=0.01$ and α_2 =0.355. From the tables of the normal integral we have

$$P\left(\bar{x}-\mu\le 2.326/\sqrt{n}\right)=0.99$$

$$P\left(\bar{x}-\mu\le 1.806/\sqrt{n}\right)=0.9645$$

and hence

$$P\left(\bar{x}-\frac{2.326}{\sqrt{n}}\le\mu\le\frac{1.806}{\sqrt{n}}\right)=0.9545 \tag{12.8}$$

Thus, with the same confidence coefficient we can assert that μ lies in the interval $\bar{x}-2/\sqrt{n}$ to $\bar{x}+2/\sqrt{n}$, or in the interval $\bar{x}-2.326/\sqrt{n}$ to $\bar{x}+1.806/\sqrt{n}$. In either case we shall be right in about 95.45 percent of cases in the long run.

We note that in the first case the interval has length $4/\sqrt{n}$, while in the second case its length is $4.123/\sqrt{n}$. Other things being equal, we should choose the first set of limits since they locate the parameter in a narrower range. We shall consider this point in more detail below. It does not always happen that there is infinity of possible confidence limits or that the choice between them can be made on such clear-cut grounds as in this example.

12.4 Graphical Representation

In a number of simple cases, including that of above example 1, the confidence limits can be represented in a useful graphical form. We take two orthogonal axes, *OX* relating to the observed $\bar{x}$ and *OY* to μ (see Fig. 1).The two straight lines shown have as their equations $\mu=\bar{x}+2$ and $\mu=\bar{x}-2.$

Consequently, for any point between the lines, $\bar{x}-2 \le \mu \le \bar{x}+2$.

Hence, if for any observed $\bar{x}$ we read off the two ordinates on the lines corresponding to that value, we obtain the two confidence limits. The vertical interval between the limits is the confidence interval (shown in the above diagram for $\bar{x}=1$), and the total zone between the lines is the confidence belt. We may refer to the two lines as upper and lower confidence lines respectively.

This example relates to the case $n=1$ in above example. For different values of *n*, there will be different confidence lines, all parallel to $\mu=\bar{x},$ and getting closer to each other as *n* increases. They may be shown on a single diagram for selected values of *n*, and a figure so constructed provides a useful method of reading off confidence limits in practical work.

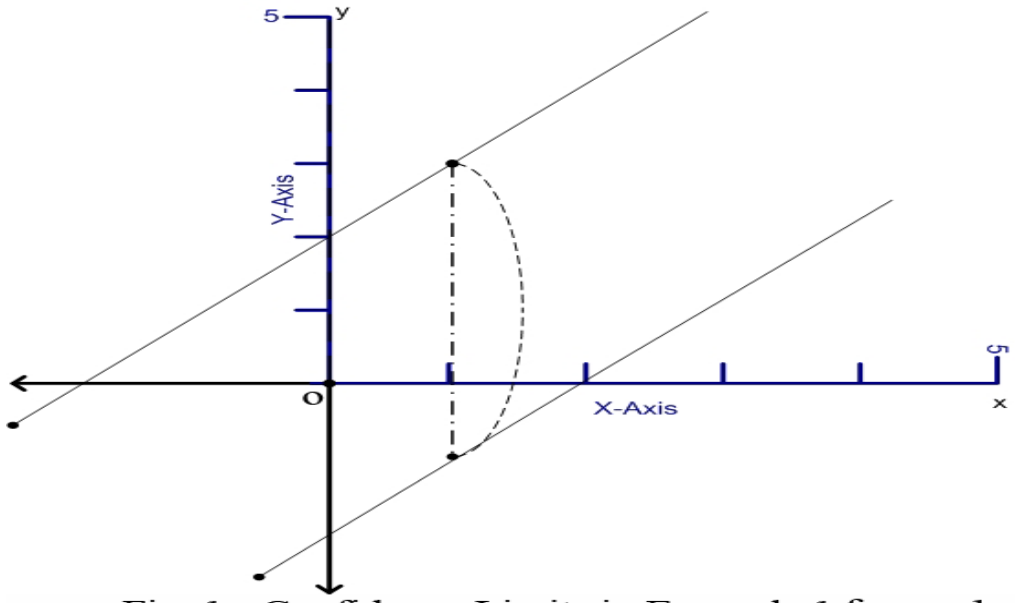


Fig. 1.- Confidence Limits in Example 1 for n =1

Alternatively, we may wish to vary the confidence coefficient $1-\alpha$, which is our example is 0.9545. Again, we may show a series of pairs of confidence lines, each pair corresponding to a selected value $1-\alpha$, on a single diagram relating to some fixed value n. In this case, of course, the lines become farther apart with increasing $1-\alpha$. In fact, in many practical situations, we are interested in the variation of the confidence interval with $1-\alpha$, and we may validly make assertions of the structure of eq(1.3) $P[T_1 \le \theta \le T_2] = 1-\alpha$, simultaneously for a number of values of α: each will be true in the corresponding proportion of cases in the long run. Indeed, this procedure may be taken to its extreme form, when we consider all values of $1-\alpha$ in (0,1) simultaneously, and thus generate a ***confidence distribution*** of the parameter-the term is due to D.R Cox (e.g.1958b): We then have an infinite sequence of simultaneous confidence statement, each contained within the preceding one, with increasing value $1-\alpha$.

12.5 Central and Non-Central Intervals

In example 1 the sampling distribution on which the confidence intervals were based was symmetrical, and hence, by taking equal deviations from the mean, we obtained equal value of

$$1-\alpha_1 = P\left(T_1 \le \theta\right)$$
$$1-\alpha_2 = P\left(\theta \le T_2\right)$$

In general, we cannot achieve this result with equal deviations, but subject always to the condition $\alpha_1 + \alpha_2 = \alpha,\ \alpha_1$ and α_2 may be chosen arbitrarily.

If α_1 and α_2 are taken to be equal, we shall say that the intervals are ***central***. In such a case we have

$$P\left(T_1 > \theta\right) = P\left(\theta > T_2\right) = \alpha/2$$

In the contrary case the intervals will be called ***non-central***. It should be observed that centrality in this sense does not mean that the confidence limits are equidistant from the sample statistic, unless the sampling distribution is symmetrical.

In the absence of other considerations, it is usually convenient to employ central intervals, but circumstances sometimes arise in which non-central intervals are more useful. Suppose, for instance, we are estimating the proportion of some drug in a medicinal preparation and the drug is toxic in large doses. We must then clearly err on the safe side, an excess of the true value over our estimate being more serious than a deficiency. In such a case we might like to take α_2 equal to zero, so that

$$P\left(\theta \le T_2\right) = 1$$
$$P\left(T_1 \le \theta\right) = 1-\alpha$$

In order to be certain that θ is not greater than T_2. But if our statistic has a sampling distribution with infinite range, this is only possible with T_2 infinite, so we must content ourselves with making α_2 very close to zero.

Again, if we are estimating the proportion of viable seed in a sample of material that is to be placed on the market, we are more concerned with the accuracy of the lower limit than that of the upper limit, for a deficiency of germination is more serious than an excess from the grower's point of view. In such circumstances we should probably take α_1 as small as conveniently possible so as to be near to certainty about the minimum value of viability. This kind of situation often arises in the specification of the quality of a manufactured product, the seller wishing to guarantee a minimum standard but being much less concerned with whether his product exceed expectation.

On a somewhat similar point, it may be remarked that in certain circumstances it is enough to know that $P[T_1 \le \theta \le T_2] \ge 1-\alpha$. We then knew that in asserting θ to lie in the range T_1 to T_2 we shall be right in at least a proportion $1-\alpha$ of the cases. Mathematical difficulties in ascertaining confidence limits exactly for given $1-\alpha$, or theoretical difficulties when the distribution is discontinuous may, for example, lead us to be content with this inequality rather than the equality of (1.3).

Example-2

Find out the $100(1-\alpha)\%$ Confidence interval for the parameters of a normal distribution when **(a)** σ^2 is known **(b)** σ^2 is unknown **(c)** μ is known **(d)** μ is unknown.

Let $X_1,\ldots,X_n$ be i.i.d. variates from $N(\mu,\sigma^2)$

(a) Taking the condition σ^2 is known. The statistics $z = \dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$ with n degrees of freedom. Hence $100(1-\alpha)\%$ confidence limits for μ are given by:

$$P\left(|z| \le z_{\alpha/2}\right) = 1-\alpha \quad \Rightarrow P\left(|\bar{x}-\mu| \le \frac{\sigma}{\sqrt{n}} z_{\alpha/2}\right) = 1-\alpha$$

$$P\left(\bar{X} - z_{\alpha/2}.\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X} + z_{\alpha/2}.\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha \qquad (12.9)$$

Where $z_{\alpha/2}$ is the tabulated value of z for n d.f. at significant level α. Hence the required confidence interval for μ is $\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$.

(b) The statistics $t = \frac{\bar{X}-\mu}{S/\sqrt{n}}$ follows student's t-distribution with n-1 degrees of freedom. Hence $100(1-\alpha)\%$ confidence limits for μ are given by:

$$P\left(|t| \le t_{\alpha/2}\right) = 1-\alpha \quad \Rightarrow P\left(|\bar{x}-\mu| \le \frac{S}{\sqrt{n}} t_{\alpha/2}\right) = 1-\alpha$$

$$P\left(\bar{X} - t_{\alpha/2}.\frac{S}{\sqrt{n}} \le \mu \le \bar{X} + t_{\alpha/2}.\frac{S}{\sqrt{n}}\right) = 1-\alpha \qquad (12.10)$$

Where $t_{\alpha/2}$ is the tabulated value of't' for n-1 d.f. at significant level α. Hence the required confidence interval for μ is $\left(\bar{X} - t_{\alpha/2}\frac{S}{\sqrt{n}}, \bar{X} + t_{\alpha/2}\frac{S}{\sqrt{n}}\right)$.

(c) We suppose that μ is known, so that σ^2 is a parameter. We determine confidence interval for σ^2.

The statistics $\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\sigma^2} = \frac{ns^2}{\sigma^2} \sim \chi^2_{(n)}$, if we define $\chi_\alpha{}^2$ as the value χ^2 such that

$P\left(\chi^2 > \chi_\alpha{}^2\right) = \int_{\chi_\alpha{}^2}^{\infty} p\left(\chi^2\right) d\chi^2 = \alpha$, where $p\left(\chi^2\right)$ is the p.d.f. of χ^2-distribution with n degree of freedom, then the required confidence interval is given by

$$P\left(\chi^2{}_{1-(\alpha/2)} \le \chi^2 \le \chi^2{}_{\alpha/2}\right) = 1-\alpha \quad \Rightarrow P\left\{\chi^2{}_{1-(\alpha/2)} \le \frac{ns^2}{\sigma^2} \le \chi^2{}_{\alpha/2}\right\} = 1-\alpha$$

$$P\left\{\frac{ns^2}{\chi^2{}_{\alpha/2}} \le \sigma^2 \le \frac{ns^2}{\chi^2{}_{1-(\alpha/2)}}\right\} = 1-\alpha \qquad (12.11)$$

Hence the required confidence interval for σ^2 is $\left(\frac{ns^2}{\chi^2{}_{\alpha/2}}, \frac{ns^2}{\chi^2{}_{1-(\alpha/2)}}\right)$.

(d) We suppose that μ is unknown so that σ^2 is a parameter. We determine confidence interval for σ^2.

The statistics $\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\sigma^2} = \frac{ns^2}{\sigma^2} \sim \chi^2{}_{(n-1)}$, here also confidence interval for σ^2 is given in equation 1.11, where now $\chi_\alpha{}^2$ is the significant value of χ^2 with α level of significance and (n-1) degree of freedom.

Note: Hence above we had seen that in normal distribution when mean (μ) is known, the statistics $\frac{ns^2}{\sigma^2}$ follows χ^2 with "n" degree of freedom and otherwise when μ unknown the statistics $\frac{ns^2}{\sigma^2} \sim \chi^2{}_{(n-1)}$.

12.6 Confidence Interval for Large Sample

We have studied that the first derivative of the logarithm of the Likelihood Function is, under regularity conditions, asymptotically normally distributed with zero mean and

$$\text{var}\left(\frac{\partial \log L}{\partial \theta}\right) = E\left\{\left(\frac{\partial \log L}{\partial \theta}\right)^2\right\} = -E\left\{\frac{\partial^2 \log L}{\partial \theta^2}\right\} \tag{12.12}$$

We may use this fact to set confidence intervals for θ in large samples. Writing

$$\psi = \frac{\partial \log L}{\partial \theta} \Big/ \left[E\left\{\left(\frac{\partial \log L}{\partial \theta}\right)^2\right\}\right]^{\frac{1}{2}}, \tag{12.13}$$

So that ψ is a standardized normal variate in large samples, we may from the normal integral determine confidence limits for θ in large samples if ψ is a monotonic function of θ, so that inequalities in one may be transformed to inequalities in the other. The following example illustrates the procedure.

Example-3

Consider the Poisson distribution whose general term is

$$f(x,\lambda) = \frac{e^{-\lambda}\lambda^x}{\lfloor \underline{x}}, \qquad x = 0,1,..... \tag{12.14}$$

We have $$\frac{\partial \log L}{\partial \lambda} = \frac{n}{\lambda}\left(\bar{x} - \lambda\right)$$

Hence $$-\frac{\partial^2 \log L}{\partial \lambda^2} = \frac{n\bar{x}}{\lambda^2}$$

and $$E\left(-\frac{\partial^2 \log L}{\partial \lambda^2}\right) = \frac{n}{\lambda}$$

Hence, we use the above equation in equation 12.12 and 12.13, we have

$$\psi = \left(\bar{x} - \lambda \right) \sqrt{n / \lambda} \tag{12.15}$$

For example, with $1-\alpha$=0.95, corresponding to a normal deviate ± 1.96, we have, for the central confidence limits,

$$\left(\bar{x} - \lambda \right) \sqrt{\left(n / \lambda \right)} = \pm 1.96$$

Giving, on solution for λ, $$\lambda^2 - \left(2\bar{x} + \frac{3.84}{n} \right) \lambda + \bar{x}^2 = 0$$

or $$\lambda = \bar{x} + \frac{1.92}{n} + \sqrt{\left(\frac{3.84\,\bar{x}}{n} + \frac{3.69}{n^2} \right)} \tag{12.16}$$

the ambiguity in the square root giving upper and lower limits respectively. To order $n^{-1/2}$ this is equivalent to

$$\lambda = \bar{x} + 1.96\sqrt{\bar{x}/n}$$

From which the upper and lower limits are seen to be equidistant from the mean $\bar{x}$, as we should expect.

12.7 Shortest Confidence Interval

As we had discussed example in section 12.3 there are some circumstances in which more than one set of confidence interval exists. Therefore, it is necessary to find whether any particular set can be considered as better than the others in some useful sense. The said example problem presented itself in specialized form. We have seen that for the intervals based on mean $\bar{x}$. There were infinitely many sets of interval exists, according to the way in which we selected α_1 and α_2 (subject to the condition that $\alpha_1 + \alpha_2 = \alpha$). Among these the

central intervals are obviously the shortest, for a given range will include the greatest area of the normal distribution if it is centred at the mean.

Example-4

Let $X_1,\ldots,X_n$ be a random sample from $N\left(\mu,\sigma^2\right)$, where the variance σ^2 is supposed to be known. If $\bar{X}$ be the sample mean, then, for every $\mu\left(-\infty<\mu<\infty\right)$, then the confidence interval for μ is

$$P\left(\bar{X}-z_{\alpha/2}.\frac{\sigma}{\sqrt{n}}\le\mu\le\bar{X}+z_{\alpha/2}.\frac{\sigma}{\sqrt{n}}\right)=1-\alpha \qquad (12.17)$$

In this case we have taken $\alpha_1=\alpha_2=\alpha/2$. However, by taking α_1 and α_2 differently but subject to the conditions $\alpha_1\ge 0$, $\alpha_2\ge 0$ and $\alpha_1+\alpha_2=\alpha$. We might obtain a different pair of confidence limits with the same confidence coefficient $1-\alpha$. We have

$$p\left[z_{1-\alpha_1}\le\sqrt{n}\left(\bar{X}-\mu\right)/\sigma\le z_{\alpha_2}\right]=1-\alpha \qquad (12.18)$$

So that

$$P\left[\bar{X}-z_{\alpha_2}\frac{\sigma}{\sqrt{n}}\le\mu\le\bar{X}+z_{1-\alpha_1}\frac{\sigma}{\sqrt{n}}\right]=1-\alpha \qquad (12.19)$$

The length of corresponding confidence interval is $\left(z_{\alpha_2}-z_{1-\alpha_1}\right)\sigma/\sqrt{n}$. Our problem is to minimize $\left(z_{\alpha_2}-z_{1-\alpha_1}\right)$ under given above condition. By taking consideration of the symmetry of the distribution of $\sqrt{n}\left(\bar{X}-\mu\right)/\sigma$ about 0, the difference $\left(z_{\alpha_2}-z_{1-\alpha_1}\right)$ will be minimum when $z_{1-\alpha_1}=-z_{\alpha_2}$ i.e. $\alpha_1=\alpha_2=\alpha/2$.

Hence the interval present in above equation 1.17, in fact the shortest confidence interval based on the distribution of $\bar{X}$.

12.8 Relation between Confidence Estimation and Hypothesis Testing

Suppose that $I(X)$ is a random interval having the property $P\left(\theta \in I\left(X\right);\theta\right)=\gamma$ for each $\theta \in \Omega$. $I(X)$ is a $100\gamma\%$ confidence interval on θ. Now suppose that we wish to test $H_0:\theta=\theta_0$ against $H_1:\theta\neq\theta_0$, where θ_0 is known constant. In testing the null hypothesis that a coin is fair, we might set p (head) $=\theta$ and $\theta_0=1/2$. In a paired sample 't' test, we might choose $\theta_0=\mu_0=0$. Consider the test of H_0 against H_1 which rejects H_0 whenever $\theta_0 \notin I\left(X\right)$. This test has the critical region $w\left(\theta_0\right)=\left\{x \mid \theta_0 \notin I\left(x\right)\right\}$,acceptance region $\bar{w}\left(\theta_0\right)=\left\{x \mid \theta_0 \in I\left(x\right)\right\}$.

Then $P\left(X \in w\left(\theta_0\right);\theta_0\right)=P\left(\theta_0 \notin I\left(X\right);\theta_0\right)=1-P\left(\theta_0 \in I\left(X\right);\theta_0\right)=1-\gamma$.

Thus, 95% confidence interval [0.32, 0.41] on a binomial parameter p indicates that for any $\alpha=0.05$level test of $H_0:p=p_0$ against $H_1:p\neq p_0$ the test would reject $H_0$for p_0 not in the interval, accept (fail to reject) for p_0 in the interval. Vice versa, if $w\left(\theta_0\right)$is an α-level critical region for a test of $H_0:\theta=\theta_0$ against$H_1:\theta\neq\theta_0$ define$I\left(x\right)=\left\{\theta_0 \mid x \notin w\left(\theta_0\right)\right\}$. Then $P\left(\theta_0 \in I\left(X\right);\theta_0\right)=1-P\left(\theta_0 \notin I\left(X\right);\theta_0\right)=1-P\left(X \in w\left(\theta_0\right);\theta_0\right)=1-\alpha$, so that $I\left(X\right)$ is a $100\left(1-\alpha\right)\%$ confidence interval on θ.

The power function $\gamma\left(\theta\right)$ for an α-level test of $H_0:\theta=\theta_0$ against$H_1:\theta\neq\theta_0$ can be useful in studying the performance of the corresponding confidence interval $I\left(X\right)$, since $\gamma\left(\theta\right)=P\left(X \in w\left(\theta_0\right);\theta\right)=P\left(\theta_0 \notin I\left(X\right);\theta\right)$. If $|\theta-\theta_0|$ is large, we want this probability to be large.

Now consider $H_0 : \theta \le \theta_0$ against $H_1 : \theta > \theta_0$ and let $L(X)$ be a $100\gamma\%$ lower confidence limit on θ. This mean that for each $\theta \in \Omega$, $P(L(X) \le \theta; \theta) = \gamma$. Consider the test that rejects H_0 whenever $L(X) > \theta_0$. This test has level $P(L(X) > \theta_0; \theta_0) = 1 - P(L(X) \le \theta_0; \theta_0) = 1 - \gamma$. Similarly, the upper $100\gamma\%$ confidence limit $U(X)$ corresponds to the test of $H_0 : \theta \ge \theta_0$ against $H_1 : \theta < \theta_0$ which rejects at level $(1-\gamma)$ for $U(X) < \theta_0$.

12.9 Self-Assessment Exercises

1. Given one observation from a population with $p.d.f.$ $f(x,\theta) = \frac{2}{\theta^2}(\theta - x), 0 \le x \le \theta$ obtain $100(1-\alpha)\%$ confidence interval for θ.

2. A random sample of size n is drawn from $U(0,\theta)$. Find the shortest confidence interval for θ for a confidence coefficient $1-\alpha$.

3. Obtain $100(1-\alpha)\%$ confidence limits (for large samples) for the parameter λ of the Poisson distribution.

4. Show that for the distribution $f(x,\theta) = \theta e^{-\theta x} dx, 0 \le x \le \infty$ central confidence limits for large samples with $\alpha = 0.95$ are given by $\theta = \left(1 \pm \frac{1.96}{\sqrt{n}}\right) / \bar{x}$

5. Develop a general method for obtaining confidence intervals. Obtain a $100(1-\alpha)\%$ confidence interval for large sample size for the parameter θ of the Poisson distribution $f(x,\theta) = \frac{e^{-\theta}\theta^x}{\lfloor x}, \; x = 0,1,2.....$

12.10 Summary

In this unit we had discussed the basic ideas and method of confidence interval estimation, which is due to Neyman (1937b). Problem of estimating the actual values of population parameters are said to belong to "Point Estimation". On other hand within what limits we can confidently expect the parameter values to lie with any specified degree of confidence and such problems belong to "***Interval Estimation.***

If T_1 and T_2 are two statistics, that is two measurable functions of the random variable $X_1,\ldots,X_n$ and θ lies in the interval T_1 to T_2, which is called a confidence interval for θ. T_1 and T_2 are known as the **lower and upper confidence limits** respectively. They depend only on $1-\alpha$ and the sample values. For any fixed $1-\alpha$, the totality of confidence intervals for different samples determines a field within which θ is supposed to lie. This field is called the ***confidence belt.*** The fraction $1-\alpha$ is called the ***confidence coefficient.***

The sampling distribution on which the confidence intervals were based was symmetrical, and hence, by taking equal deviations from the mean, we obtained equal value of

$$1-\alpha_1 = P\left(T_1 \le \theta\right)$$
$$1-\alpha_2 = P\left(\theta \le T_2\right)$$

And subject to the condition $\alpha_1+\alpha_2=\alpha,\ \alpha_1$ and α_2 may be chosen arbitrarily.

If α_1 and α_2 are taken to be equal, we shall say that the intervals are ***central***. In such a case we have

$$P\left(T_1 > \theta\right) = P\left(\theta > T_2\right) = \alpha/2$$

In the contrary case the intervals will be called ***non-central***.

The central intervals are obviously the **shortest**, for a given range will include the greatest area of the normal distribution if it is centred at the mean.

12.11 References and Further Readings

- Devore, Jay L. Berk, Kenneth N. Modern Mathematical Statistics with Applications Second Edition, Springer-Verlag 2012

- Gun A.M., Gupta M.K., Dasgupta B., An Outline of Statistical Theory, The World Press Private limited, Kolkata.

- Gupta S. C. and Kapoor V. K. Fundamentals of mathematical statistics, Sultan Chand & Sons. New Delhi.

- Hogg, Robert V., McKean, Joseph W., and Craig, Allen T. Introduction to Mathematical Statistics, 6th ed., Pearson Prentice Hall, 2005.

- Kendall M.G. and Stuart A., The Advance Theory of Statistics Vol-2, Charles Griffin & Co. Limited, London.

- Larsen J Richard. and Marx L Morris, An Introduction to Mathematical Statistics and Its Applications, 5th Edition, Prentice Hall.

- Rohatgi, V. K., An Introduction to Probability Theory and Mathematical Statistics. John Wiley & Sons, 1976, New York.

- Shao, Jun. Mathematical Statistics: Springer-Verlag ,1999, New York.

- Stapleton, James H. Models for Probability and Statistical Inference Theory and Applications, John Wiley & Sons, 2008.

- Wilks, Samuel S. Mathematical Statistics, John Wiley & Sons, 1962, New York.

We acknowledge all above authors as we had referred them extensively while creating this Self-Study material.

UNIT-13 HYPOTHESIS TESTING

Structure

13.1 Introduction

We have previously studied that when the tested hypothesis is simple (specifying the distribution completely) we always have best critical region (BCR), which provides a most powerful test against a simple alternative hypothesis. Also, this is possibly a Uniformly Most Powerful (UMP) test against the class of simple hypothesis which constitute a composite parametric alternative hypothesis, and that there will not, in general, be a UMP test if the parameter whose variation generates the alternative hypothesis is free to vary in both direction from the value tested.

In many practical problems samples are examined in order to test some hypothesis about the parent population, for example, that the mean is not greater than some specified value, or that the proportion of individuals with some definite characteristic lies within assigned limits. In general problem of testing of hypothesis, we have a set of random variables $X_1,\ldots,X_N$, with a joint probability distribution $F\left(X_1,\ldots,X_N\right)$. The statistical

hypothesis H_ω is that $F(X_1,\ldots,X_N)$ belongs to a certain subclass ω of distribution function. If the class ω consists of a single element, the hypothesis is said to simple. Otherwise, it is called composite.

13.2 Objective

After studying this unit, the learner will understand about Generalized Neyman Pearson lemma. You will be acquainted with UMP, UMPU, and Invariant test for different distribution. You will also understand the meaning of LR test and their Properties, Similar Region, Neyman Structure, and method to obtain these.

13.3 Generalized Neyman Pearson Lemma

Generalized Neyman Pearson lemma is an important result for establishing most powerful statistical tests. While testing a simple hypothesis against a simple alternative hypothesis i.e. choosing between two completely specified distributions, the problem of finding a Best Critical Region (BCR) of a given size is straightforward. Its solution is given by the following Lemma.

Lemma 1- Let $f_0, f_1\ldots\ldots$ are integrable function over space S with respect to a measure v and let w be any region such that

$$\int_w f_i \phi dv = c_i \quad \text{(given)}, \; i = 1,2,\ldots\ldots \tag{13.1}$$

Further, let there exist constants $a_1, a_2, \ldots\ldots$ such that for the region w_0 within which

$f_0 \geq a_1 f_1 + a_2 f_2 + \ldots\ldots\ldots$ and outside which $f_0 \leq a_1 f_1 + a_2 f_2 + \ldots\ldots\ldots$, the conditions (13.1) are satisfied. Then

$$\int_{w_0} f_0 dv \geq \int_w f_0 dv \tag{13.2}$$

Let the common part of the region w and w_0 be denoted by ww_0. From (8.1), subtracting the common part, we have

$$\int_{w-ww_0} f_i dv = \int_{w_0-ww_0} f_i dv, \qquad i = 1, 2, ... \tag{13.3}$$

Consider the difference

$$\int_{w_0} f_0 dv - \int_{w} f_0 dv = \int_{w_0-ww_0} f_0 dv - \int_{w-ww_0} f_0 dv \geq \int_{w_0-ww_0} \sum` a_i f_i dv - \int_{w-ww_0} \sum` a_i f_i dv \tag{13.4}$$

The equation (13.4) is zero, using (13.3). Hence the result (13.2).

Lemma 2 (Generalized Lemma 1) - Let $f_0, f_1 .....$ be as in Lemma1, and φ be a point function over S such that $0 \leq \phi \leq 1$ and

$$\int f_i \phi dv = c_i \quad \text{(given)}, \quad i = 1, 2, ... \tag{8.5.}$$

Further, let there exist a ϕ^* such that

(a) the condition (8.5) is satisfied,

(b) $\phi^* = 0,$ if $f_0 < a_1 f_1 + a_2 f_2 + ...$

$=1,$ if $f_0 > a_1 f_1 + a_2 f_2 + ...$

=arbitrary, if $f_0 = a_1 f_1 + a_2 f_2 + ...$

then $$\int f_0 \phi^* dv \geq \int f_0 \varphi dv \tag{13.6}$$

Let S_1, S_2 and S_3 be the regions $f_o < \sum` a_i f_i$ , $f_0 > \sum` a_i f_i$ and $f_0 = \sum` a_i f_i$. In S_2, $\phi^* - \phi \geq 0$ and in S_1, $\phi^* - \phi \leq 0$ since $\phi^* = 0$. Therefore,

$$f_0\left(\phi^* - \phi\right) \geq \left(\sum a_i f_i\right)\left(\varphi^* - \phi\right)$$

in each of S_1, S_2 and trivally S_3 and therefore, at all points in S thus

$$\int_S f_0\left(\phi^* - \phi\right) dv \geq \int_S \left(\sum a_i f_i\right)\left(\varphi^* - \phi\right) dv = 0$$

using the condition (13.5), and hence the result (13.6).

13.4 Uniformly Most Powerful (UMP) test for Distribution with MLR

In this section we compare one value or one set of value of a parameter against a series of its values. A critical region, whose power is no smaller than that of any other region of the same size for testing of hypothesis, H_0 against a simple alternative H_1, is called a best critical region (BCR) and a test based on a BCR is called a most powerful test. Such a critical region which offers most powerful test for a series of alternative hypothesis where the same Null hypothesis is compared against each of them is called *Uniformly Most Powerful critical region* with regards to this series of alternatives and the test based on this, is known as ***Uniformly Most Powerful*** **(UMP)** test**.**

Lemma- The power of a BCR for testing a simple hypothesis against a simple alternative is never less than its size.

Proof- Let w_0 be a BCR of size α for testing a simple hypothesis H_0 against the simple alternative H_1 .

Then $$\int_{w_0} L\left(x, H_0\right) dx = \alpha \qquad (13.7)$$

By Neyman-Pearson lemma we have, within the BCR w_0

$$a\,L(x,H_1) \geq L(x,H_0)$$

i.e. $$a\int_{w_0} L(x,H_1)dx \geq \int_{w_0} L(x,H_0)dx$$

i.e. $$a\,(1-\beta) \geq \alpha \qquad (13.8)$$

for $\int_{w_0} L(x,H_1)\,dx = (1-\beta)$ is the ***power of the test***.

again, inside $S-w_0$ i.e. outside w_0, we have the relation

$$a\,L(x,H_1) \leq L(x,H_0)$$

i.e. $$a\int_{S-w_0} L(x,H_1)dx \leq \int_{S-w_0} L(x,H_0)dx$$

i.e. $$(1-\alpha) \geq a\beta$$

(13.9) From (13.8) and (13.9) we get

$$a(1-\beta)(1-\alpha) \geq a(\alpha\beta)$$

i.e. $(1-\beta)(1-\alpha) \geq \alpha\beta$; as a is positive

This is the same as $1-\beta-\alpha \geq 0$ or $(1-\beta) \geq \alpha$ and this proves the lemma.

Unlike the MP tests, UMP tests do not always exist. But there are certain cases in which they exist. One such case of existence of UMP test is considered in the following theorem. A real parameter family of distributions $f(x,\theta)$ is said to have ***Monotone Likelihood Ratio*** **(MLR)** if there exists a real valued function $D(x)$ such that for any $\theta < \theta'$,

the likelihood functions $L(x,\theta)$ and $L(x,\theta')$ are distinct and the ratio $L(x,\theta')/L(x,\theta)$ is a non-decreasing function of $D(x)$.

Theorem- If θ is a real parameter and $f(x,\theta)$, the distribution of a random variable x, has a monotone likelihood ratio in $D(x)$, then there exists a UMP test for testing $H:\theta \le \theta_0$ against the alternative $H':\theta > \theta_0$.

Proof- We know that, for testing a simple hypothesis $H_0:\theta=\theta_0$ against a simple alternative $H_1:\theta=\theta_1(>\theta_0)$ there exists a BCR w_0 defined by

$$\frac{L(x,H_1)}{L(x,H_0)} \geq \text{ a constant}$$

(8.10)

Since the ratio of the likelihood function, that is $L(x,\theta')/L(x,\theta)$ is a non-decreasing function of $D(x)$ for $\theta' > \theta$, the BCR determined by (8.10) is also given by

$$D(x) \geq a_1 \quad \text{inside } w_0$$

(8.11)

Let the size and power function of this test be respectively α and $P(\theta)$ so that $P(\theta_0)=\alpha$

The BCR for testing $\theta=\theta'$ against $\theta=\theta''(>\theta')$ is given by $\frac{L(x,\theta'')}{L(x,\theta')} \geq$ a constant inside the critical region.

As the likelihood ratio is monotone in $D(x)$, this BCR is also determined in terms of $D(x)$, say by the relation

$$D(x) \geq a_2 \tag{13.12}$$

If we now take $a_2 = a_1$ in (13.12), the critical region obtained is identical with w_0 defined in (13.11) and is still most powerful for testing $\theta = \theta'$ against $\theta = \theta''(>\theta')$ with size of the region equal to $\alpha' = P(\theta')$. As the test is most powerful, $P(\theta'') \geq P(\theta')$ and this is true for any θ' and $\theta''(>\theta')$. Thus, the power function $P(\theta)$ is strictly increasing for $P(\theta) < 1$.

Therefore, for testing the hypothesis $H : \theta \leq \theta_0$, the critical region defined by (13.11) can be used with size less than or equal to α. The power of this test for any alternative $\theta_1(>\theta_0)$ is maximum and this is so for all alternatives greater than θ_0. Hence the critical region (13.11) is a UMP for testing $H : \theta \leq \theta_0$ against $H' : \theta > \theta_0$.

Note: In the same way we can show that there exists a UMP test for testing $H : \theta \geq \theta_0$ against $H' : \theta < \theta_0$ under the same conditions in which above theorem exists.

Corollary: If $B(\theta)$ is a strictly monotone function of a real parameter θ and $R(\theta)e^{B(\theta)N(x)}l(x)$ is the joint probability density of the n observations $x_1, x_2 \ldots, x_n$; then there exists a UMP test for testing, (i) $\theta \leq \theta_0$ against $\theta > \theta_0$ (ii) $\theta \geq \theta_0$ against $\theta < \theta_0$.

Proof-(i) If $B(\theta)$ increasing, let $\theta_1(>\theta_0)$ denote a simple alternative. Then

$$\frac{L(x,\theta_1)}{L(x,\theta_0)} = \frac{R(\theta_1)}{R(\theta_0)} e^{\{B(\theta_1)-B(\theta_0)\}N(x)}$$

This is a non-decreasing function of $N(x)$ and hence by above theorem and the note to it the result follows.

(ii) If $B(\theta)$ decreasing, for $\theta_1 > \theta_0$,

$$\frac{L(x,\theta_1)}{L(x,\theta_0)} = \frac{R(\theta_1)}{R(\theta_0)} e^{\{B(\theta_1)-B(\theta_0)\}N(x)}$$

This is a non-decreasing function of $T(x) = \frac{1}{N(x)}$ and the result is immediate.

When a UMP test does not exist, we may use the same approach used in estimation problems, i.e., imposing a reasonable restriction on the tests to be considered and finding optimal tests within the class of tests under the restriction. Two such types of restriction in estimation problems are unbiasedness and invariance. We discussed unbiased test in section 13.4.1 and class of invariant test in section 13.4.2.

13.4.1 Uniformly Most Powerful Unbiased (UMPU) Test

The power of any statistical test is probability of a correct decision and the size of the corresponding critical region is always the probability of wrong decision. In any test it is desirable that the probability of a correct decision is as large as possible as; at least larger than the probability of a wrong decision. Therefore, we formulate another criterion for the choice of a critical region namely that the power of the test based on it is never less than its size. Such a critical region is said to be an unbiased critical region and test based on it is called an unbiased test. If $H_0 : \theta \in \Omega_{H0}$ and $H_1 : \theta \in \Omega_{H1}$ where Ω_{H0} and Ω_{H1} are subspace of Ω and w any critical region for testing H_0 against H_1, unbiasedness requires that

$$P_r\{x \in w \mid H_0\} \leq \alpha$$

and

$$P_t\{x \in w \mid H_0\} \geq \alpha$$

If H_1 is composite, the test is said to be unbiased if for each simple alternative the test has power at least equal to α. A test is said to be uniformly unbiased or completely unbiased if the power of the test is always greater than α. A Uniformly Most Powerful Unbiased

(UMPU) test is one which is at least as powerful as or more powerful than any other unbiased critical region of equal size, with respect to all alternative hypothesis. If Uniformly Most Powerful Unbiased test exists and we accept the principle of unbiasedness for the choice of tests. Therefore, it is obvious that this is the most advantageous test to use. Neyman and Pearson called a critical region corresponding to a UMPU test, a critical region of Type A_1. The Type A_1 region exists in an important but restricted class of cases. The advantage with the UMPU test is that even when the UMP tests do not exist, the former may exist.

13.4.2 Invariant Test

In the above the unbiasedness principle is considered to derive an optimal test within the class of unbiased tests when UMP test does not exist. We study same problem with unbiasedness replaced by invariance under a given group of transformation. The principle of unbiasedness and invariance often complement each other in that each is successful in case where the other is not.

Definition- Let X be a random sample from Ᵽ $\in$ p

1. A class G of one-to-one transformation of X is called a group if and only if $g_i \in G$ implies $g_1 o g_2 \in G$ and $g_i^{-1} \in G$

2. We say that p is invariant under G if and only if $\bar{g}$ ($Ᵽ_x$) = $Ᵽ_{g(X)}$ is a one-to-one transformation from p on to p for each $g \in G$.

3. We say that the problem of testing H_0: Ᵽ$\in$ p_0 against H_1: Ᵽ$\in$ p_1 is invariant under G if and only if both p_0 and p_1 are invariant under G.

4. In an invariant testing problem, a test $T(X)$ is said to be invariant under G if and only if

$$T\{g(x)\} = T(x) \qquad \text{for all } x \text{ and } g \tag{13.13}$$

5. A test size α is said to be uniformly most power invariant (UMPI) test if and only if it is UMP within the class level α test that are invariant under G.

6. A statistic $M(X)$ is said to be maximal invariant under G if and only if (13.13) hold with T replaced by M and

$$M(x_1) = M(x_2) \qquad \text{implies } x_1 = g(x_2) \text{ for some } g \in G(B)$$

Example-1 Examine whether a UMP test exists for testing $H_0 : \mu = \mu_0, \sigma = \sigma_0$ in $N(\mu, \sigma^2)$.

If $H_1 : \mu = \mu_1, \sigma = \sigma_1$ is any simple alternative, then

$$\frac{L(x, H_1)}{L(x, H_0)} = \left(\frac{\sigma_0}{\sigma_1}\right)^n \exp\left[-\frac{1}{2}\left\{\frac{\sum(x_i - \mu_1)^2}{\sigma_1^2} - \frac{\sum(x_i - \mu_2)^2}{\sigma_0^2}\right\}\right] \geq a$$

determine the BCR. This may put in the form

$$s^2\left(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_0^2}\right) + \frac{(\bar{x} - \mu_1)^2}{\sigma_1^2} - \frac{(\bar{x} - \mu_0)^2}{\sigma_0^2} \leq \frac{2}{n}\log\left[\frac{1}{k}\left(\frac{\sigma_1}{\sigma_0}\right)^n\right]$$

i.e. $$s^2\left(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_0^2}\right) + \bar{x}^2\left(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_0^2}\right) + 2\bar{x}\left(\frac{\mu_0}{\sigma_0^2} - \frac{\mu_1}{\sigma_1^2}\right) \leq \text{constant}$$

i.e. $$\left(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_0^2}\right)\left\{s^2 + (\bar{x} - \delta)^2\right\} \leq \text{constant, where } \delta = \frac{\mu_0\sigma_1^2 - \mu_1\sigma_0^2}{\sigma_1^2\sigma_0^2}$$

i.e. $$(\sigma_0^2 - \sigma_1^2)\sum(x_i - \delta)^2 \leq \text{constant}$$

This means that if $\sigma_0 > \sigma_1$, the BCR is bounded by a hyper sphere centered at $(\delta, \ldots\ldots\delta)$ where δ itself is dependent on H_1. When $\sigma_1 > \sigma_0$, the BCR lies outside this sphere. In both

cases the BCR changes with the alternative. Therefore, there does not exist any UMP test for any set of alternatives.

8.5 Likelihood Ratio (L.R.) Test

A method of test construction closely allied to Maximum Likelihood Estimated (MLE) is the Likelihood Ratio (LR) method, proposed by Neyman and Pearson (1928). It has played a similar role in the theory of tests to that the maximum Likelihood method in the theory of estimation.

We have a Likelihood function (LF)

$$L(x \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta) \tag{13.14}$$

Where $\theta = (\theta_r, \theta_s)$ is a vector of $r + s = k$ parameters $(r \geq 1, s \geq o)$ and x may also be vector. We wish to test the hypothesis

$$H_0 : \theta_r = \theta_{r0}$$

Which is composite unless $s = 0$, against

$$H_1 : \theta_r \neq \theta_{r0}$$

We know that there is generally no UMP test in this situation but that there may be a UMPU test.

The LR method first requires us to find ML estimators of (θ_r, θ_s), giving the unconditional maximum of the LF

$$L\left(x \mid \hat{\theta}_r, \hat{\theta}_s\right) \tag{13.15}$$

and also to find the ML estimators of θ_s, where H_0 holds *(when* $s=0$, H_0 *being simple, no maximization process is needed, for* L *is uniquely determined)* giving the conditional maximum of LF

$$L\left(x \mid \theta_{r0}, \hat{\hat{\theta}}_s \right) \tag{13.16}$$

$\hat{\hat{\theta}}_s$ in (8.16) has been given a double circumflex to emphasize that it does not in general coincide with $\hat{\theta}_s$ in (8.15). Now consider the likelihood ratio

$$\lambda = L_{\max}\left(x, \theta_{r0}\, \hat{\hat{\theta}}_s \right) / L_{\max}\left(x, \theta_r\, \hat{\theta}_s \right) \tag{13.17}$$

Since (13.17) is the ratio of a conditional maximum of the LF to its unconditional maximum, we clearly have

$$0 < \lambda < 1$$

(13.18) Intuitively, λ is a reasonable test statistics for H_0, it is maximum likelihood under H_0 as a fraction of its largest possible value, and large values of λ signify that H_0 is reasonably acceptable. The critical region for the test statistics is therefore,

$$\lambda \le c_\alpha$$

(13.19) Where c_α is determined from the distribution $g(\lambda)$ of λ to give a size-α test, i.e.

$$\int_0^{c_\alpha} g(\lambda)\, d\lambda = \alpha \tag{13.20}$$

Example-3 Let $X_1, \ldots, X_n$ be random sample from $N\left(\mu, \sigma^2\right)$ population. We wish to test

$H_0 : \mu = \mu_0$ (Specified) against $H_1 : \mu \neq \mu_0$, σ^2 is unknown.

Here
$$\Theta = \left\{ \left(\mu, \sigma^2 \right); -\infty < \mu < \infty, \sigma^2 > 0 \right\}$$

$$\Theta_0 = \left\{ \left(\mu_0, \sigma^2 \right); -\infty < \mu_0 < \infty, \sigma^2 > 0 \right\}$$

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

$$L\left(x, \theta \right) = \left(2\pi\sigma^2 \right)^{-n/2} \exp\left\{ -\sum \left(x_i - \mu \right)^2 / 2\sigma^2 \right\} \tag{13.21}$$

$$L\left(x, \theta_0 \right) = \left(2\pi\sigma^2 \right)^{-n/2} \exp\left\{ -\sum \left(x_i - \mu_0 \right)^2 / 2\sigma^2 \right\} \tag{13.22}$$

We know that the MLE for $\left(\mu, \sigma^2 \right)$ is $\left(\bar{X}, S^2 \right)$ and that for $\sigma^2 \left(\mu = \mu_0 \right)$ known is ${S_0}^2$, where

$$nS^2 = \sum (X_i - \bar{X})^2, n{S_0}^2 = \sum (X_i - \mu_0)^2$$

$$\sum (X_i - \mu_0)^2 = \sum \left[\left(X_i - \bar{X} \right) + \left(\bar{X} - \mu_0 \right) \right]^2 \tag{13.23}$$

Since

$$= \sum \left(X_i - \bar{X} \right)^2 + n \left(\bar{X} - \mu_0 \right)^2$$

$$ {S_0}^2 = S^2 + \left(\bar{X} - \mu_0 \right)^2 \Rightarrow \left({S_0}^2 / S^2 \right) = 1 + \left[\left(\bar{X} - \mu_0 \right)^2 / S^2 \right] \tag{13.24}$$

As $T = \left(\bar{X} - \mu_0 \right) \sqrt{n-1} / S \sim t_{(n-1)}$, the preceding result becomes

$${S_0}^2 / S^2 = 1 + \left[T^2 / (n-1) \right]$$

$$\max L(x,\theta) = \left(2\pi S^2\right)^{-n/2} \exp\left\{-\sum (X_i - \bar{X})^2 / 2S^2\right\} \tag{13.25}$$

using the result equation (13.23), we find max $L(x,\theta) = \left(2\pi S^2\right)^{-n/2} \exp(-n/2)$

$$\max L(x,\theta_0) = \left(2\pi S_0^{\,2}\right)^{-n/2} \exp\left\{-\sum (X_i - \mu_0)^2 / 2S_0^{\,2}\right\} \tag{13.26}$$

again using the result equation (13.23), we find max $L(x,\theta_0) = \left(2\pi S_0^{\,2}\right)^{-n/2} \exp(-n/2)$

$$\lambda = \frac{\max L(x,\theta_0)}{\max L(x,\theta)} = \left(\frac{S_0^{\,2}}{S^2}\right) = \left(1 + \frac{T^2}{n-1}\right)^{-n/2}$$

We reject H_0 if $\lambda < c$ where c is some constant. Here this gives

$$\left(1 + \frac{T^2}{n-1}\right)^{-n/2} < c \;\Rightarrow T^2 > b^2 \text{ or } |T| > b$$

This gives the critical region. The constant b can determine by the size of C.R., i.e. α and so

$$\alpha = P\left[|T| > b \mid H_0\right] \Rightarrow b = t_{(n-1),\alpha/2}$$

Thus, b is the upper $(\alpha/2)^{th}$ quartile of $t_{(n-1)}$ distribution. Obviously, LRT is equivalent to the test function ϕ given by

$$\phi(x) = 1, \quad \text{if } |T| > t_{(n-1),\alpha/2}$$

= 0, otherwise

Example-4 Let $X_1, \ldots, X_n$ be random sample from $N(\mu, \sigma^2)$ population. We wish to test

$H_0 : \sigma^2 = \sigma_0^{\,2}$ (Specified) against $H_1 : \sigma^2 \neq \sigma_0^{\,2}$

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

$$L\left(x,\theta\right) = \left(2\pi\sigma^2\right)^{-n/2} \exp\left\{-\sum\left(x_i-\mu\right)^2/2\sigma^2\right\} \tag{13.27}$$

$$L\left(x,\theta_0\right) = \left(2\pi\sigma_0^{\,2}\right)^{-n/2} \exp\left\{-\sum\left(x_i-\mu\right)^2/2\sigma_0^{\,2}\right\} \tag{13.28}$$

We know that the MLE for $\left(\mu,\sigma^2\right)$ is $\left(\bar{X},S^2\right)$ and that for μ is $\bar{X}$ ($\sigma = \sigma_0^{\,2}$ known), where

$nS^2 = \sum (X_i - \bar{X})^2, n\bar{X} = \sum X_i$

Hence max $L\left(x,\theta_0\right) = \left(2\pi\sigma_0^{\,2}\right)^{-n/2} \exp\left(-nS^2/2\sigma_0^{\,2}\right)$ and max $L\left(x,\theta\right) = \left(2\pi S^2\right)^{-n/2} \exp\left(-n/2\right)$

$$\lambda = \frac{\max L\left(x,\theta_0\right)}{\max L\left(x,\theta\right)} = \left(\frac{S^2}{\sigma_0^{\,2}}\right)^{n/2} \exp\left\{\frac{1}{2}n - \frac{1}{2}n\left(\frac{S^2}{\sigma_0^{\,2}}\right)\right\} \tag{13.29}$$

Under H_0, we know that $W = \left(ns^2/\sigma_0^{\,2}\right) \sim \chi^2_{(n-1)}$ and equation (8.29) is equivalent to

$$\lambda = \left(W/n\right)^{n/2} \exp\left\{-\frac{1}{2}\left(W-n\right)\right\} = \frac{W^{n/2}}{\left(n\right)^{n/2}} e^{n/2} e^{-W/2}$$

Hence λ increase if $W < n$ and λ decrease if $W > n$. Consequently $\lambda < c$ provides $W < b_1$ or $W > b_2$ which provides the critical region (rejection of H_0). Since W is $\chi^2_{(n-1)}$, the constants b_1 and b_2 can be uniquely determined by

$$P\left(W \le b_1\right) + P\left(W \ge b_2\right) = \alpha, \qquad \lambda\left(b_1\right) = \lambda\left(b_2\right)$$

However, in practice, b_1 and b_2 are usually considered by

$$P(W \le b_1) = P(W \ge b_2) = \alpha / 2$$

So that b_1 is the upper $\alpha / 2$ quartile of $\chi^2_{(n-1)}$ distribution.

The Properties of LR Test:

Likelihood ratio (L.R.) test's theory is an intuitive one. If we like to test a simple hypothesis H_0 against a simple alternative hypothesis H_1 then the LR principle leads to the similar test as suggested by the Neyman-Pearson lemma. This tells that LR test has some wanted properties, especially large sample properties.

In LR test, the probability of type *I* error is controlled by properly choosing the cut-off point λ_0. LR test is generally UMP if there is an UMP test exists. The two asymptotic properties of LR tests is given below.

1. Under certain conditions, $-2\log_e \lambda$ has an asymptotic chi-square distribution.
2. Under certain assumptions, LR test is consistent.

13.6 Composite Hypothesis and Similar Region

As we know, the null hypothesis H_0 is said to be composite if it does not specify the joint probability distribution of $X_1, \ldots, X_n$ completely. Since here we are apprehensive with parametric hypotheses, we can write the null hypothesis in the following form

$$H_0 : \theta \in \Theta_0 \tag{13.30}$$

Where Θ_0 consists of more than a single point.

In the process of finding to suitable test for H_0, we shall as before; we look into two type of error that one may commit by rejecting or accepting H_0 on the basis of a set of observations on the random variables $X_1,\ldots,X_n$. We would, first of all, like that the test be of the desired level $\alpha(0<\alpha<1)$. In other words, if W be the critical region of the test then it is required that

$$P_\theta(W)=\alpha \text{ for all } \theta\in\Theta_0 \tag{13.31}$$

This means, in the first place, that W should be such that its size is the same for all parameter points within Θ_0 and, secondly, that this common size should be equal to α. Now, not all regions of $\Omega=R^n$ has this property of having a size independent of $\theta\in\Theta_0$. Those that have are called regions similar to the sample space or in other word similar regions, because the sample space Ω has the property that

$$P_\theta(\Omega)=1$$

Identically in θ.

In this case of a composite null hypothesis, then, the selection of a suitable test involves three essential stages, viz. finding all similar regions, finding those similar regions which are of size α and lastly, finding a similar region of size α that is best from the point of view of power.

Taking this discussion further, we know that a composite hypothesis leaves some parameter values unspecified, a new problem immediately arises, for the size of the test of H_0 will obviously be a function, in general, of these unspecified parameter value θ_s.

If we to keep Type I errors down to some preassigned level, we must seek critical regions whose size can be kept down to that level for all possible value of θ_s.

Thus, we require

$$\alpha(\theta_s) \leq \alpha$$

If a critical region has

$$\alpha(\theta_s) = \alpha$$

as a strict equality for all θ_s, it is called (critical) region *similar* to the sample space with respect θ_s, or, more briefly, a *similar* (critical) *region.* The test based on a similar critical region is called ***a similar size-*** α ***test***.

It is not obvious that similar region exist at all generally, but, in one sense, as Feller (1938) pointed out, they exist whenever we are dealing with a set of n independent identically distributed observations on a continuous variate x. For no matter what the distribution or its parameters, we have

$$P(x_1 < x_2 < x_3 < ..... < x_n) = 1/n! \qquad (13.32)$$

Since any of $n!$ permutations of the x_i is equally likely. Thus, for α an integral multiple of $1/n!$, there are similar regions based on $n!$ hyper simplices in the sample space obtained by permuting the n suffixes in (13.32).

If we confine ourselves to regions defined by symmetric functions of the observations it is easy to see that, where similar region do exist, they need not exist for all sample size. For example, for a sample n observation from the Normal distribution with mean θ and variance 1,

$f(x;\theta,1) = (2\pi)^{-1/2} \exp\left\{-\frac{1}{2}(x-\theta)^2\right\} dx$, there is no similar region with respect to θ for n=1, but for n$\geq$2 the fact that $ns^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2$ has a chi-square distribution with (n-1)

degree of freedom, what-ever the value of θ, ensures that similar regions of any size can be found from the distribution of s^2.

13.6.1 Similar Region and Complete Sufficient Statistics

Let T be the sufficient statistics for $\theta \in \Theta_0$. Also, let I_w be the indicator function of the critical region W, so that

$$I_w = \begin{cases} 1 & if\ x \in W \\ 0 & otherwise \end{cases} \tag{13.33}$$

We then have the following theorem:

Theorem- If W is such that the conditional expectation (The conditional expectation must be independent of θ since T is sufficient)

$$E(I_w \mid t) = \alpha \quad \text{a.e.} \ (\theta \in \Theta_0) \tag{13.34}$$

then W is a similar region of size α.

Proof- we have

$$\begin{aligned} E_\theta(I_w) &= E_\theta\{E(I_w \mid T)\} \\ &= E_\theta(\alpha) \\ &= \alpha, \ \text{for all } \theta \in \Theta_0 \end{aligned}$$

This means

$$\begin{aligned} P_\theta(W) &= E_\theta(I_w) \\ &= \alpha, \ \text{for all } \theta \in \Theta_0 \end{aligned}$$

So that W is a similar region of size α.

13.6.2 Neyman Structure

Definition: A test with critical region W is said to be ***Neyman Structure*** with respect to T if $E(I_w \mid t)$ is the same almost everywhere for all $\theta \in \Theta_0$.

Thus, a test with Neyman structure and size α is characterized by the fact that the conditional probability of rejection H_0 when it is true is α on each of the surfaces $T = t$.

Since the distribution of $X_1, \ldots, X_n$ on each surface is independent of θ for $\theta \in \Theta_0$, condition (13.34) in essence reduces the problem of testing H_0 to that of testing a simple hypothesis for each value of t. Quite frequently, it would be easy to obtain a most powerful test for H_0 among those having *Neyman structure*, by obtaining a most powerful test for the simple hypothesis on each surface separately. The resulting test will then be most powerful among all similar tests, provided every similar test is of Neyman *structure*. A condition under which this will hold is given by following theorem.

Theorem: Every test based on a similar region will have Neyman structure with respect to a sufficient statistic T if the family P_0^T of distributions of T for $\theta \in \Theta_0$, i.e.

$$P_0^T = \left\{ p_\theta{}^T \mid \theta \in \Theta_0 \right\} \qquad (13.35)$$

is boundedly complete.

Proof: **S**uppose that P_0^T is boundedly complete and let W be any similar region of size α for $\theta \in \Theta_0$. Then

$$\begin{aligned} E_\theta(I_w) &= P_\theta(W) \\ &= \alpha \text{ for all } \theta \in \Theta_0 \\ E_\theta(I_w - \alpha) &= 0 \text{ for all } \theta \in \Theta_0 \end{aligned}$$

Hence, if $\psi(t)$ denotes the conditional expectation of $(I_w - \alpha)$, given $T = t$, then

$$E_\theta\left[\psi(T)\right] = 0 \text{ for all } p_\theta{}^T \in P_\theta{}^T \tag{13.36}$$

Now, $$-\alpha \le I_w(X) - \alpha \le 1 - \alpha$$

implying that $$-\alpha \le \psi(t) \le 1 - \alpha$$

Thus $\psi(t)$ is bounded, and it follows from (13.36) and the bounded completeness of P_0^T that

$$\psi(t) = 0 \qquad \text{a.e.}\left(P_0^T\right)$$

i.e. $$E\left(I_w \mid t\right) = \alpha \qquad \text{a.e.}\left(P_0^T\right)$$

Hence the test corresponding to W has ***Neyman Structure***.

13.7 Self -Assessment Exercises

1. State and prove Neyman-Pearson lemma and use it to obtain the most powerful test of a simple null hypothesis against a simple alternative hypothesis.

2. For the Normal distribution with unknown mean μ and unknown standard deviation σ, show that a UMP test exists for testing $H_0 : \sigma = \sigma_0$ against $H_1 : \sigma > \sigma_0$ and obtain the power function of the UMP test. Give an example when a UMP does not exist.

3. Discuss the general method of construction of likelihood ratio test. Consider "*n*" Bernoullian trials with probability of success *p* for each trial. Derive the likelihood-ratio test for testing $H_0 : p = p_0$ against $H_1 : p > p_0$.

4. Let $X_1, \ldots, X_n$ be a random sample from a Poission distribution with parameter θ. Derive the likelihood ratio test for $H_0 : \theta = \theta_0$ against $H_1 : \theta > \theta_0$. Show that this is identical with the corresponding UMP test.

5. Obtain the LR test for testing $\sigma_1 = \sigma_2$ in two normal populations with known means. Show that the test is uniformly unbiased.

13.8 Summery

A critical region, whose power is no smaller than that of any other region of the same size for testing of hypothesis, H_0 against a simple alternative H_1, is called a ***Best critical Region*** (BCR) and a test based on a BCR is called a most powerful test.

Generalized Neyman Pearson lemma is an important result for establishing most powerful statistical tests. While testing a simple Hypothesis against a simple alternative hypothesis i.e. choosing between two completely specified distributions, the problem of finding a Best critical Region (BCR) of a given size is straightforward. Its Solution is given by Neyman Pearson lemma.

Such a critical region which offers most powerful test for a series of alternative hypothesis where the same Null hypothesis is compared against each of them is called Uniformly Most Powerful critical region with regards to this series of alternatives and test based on it is known as ***Uniformly Most Powerful*** (UMP) test.

The power of a BCR for testing a simple hypothesis against a simple alternative is never less than its size.

A ***Uniformly Most Powerful Unbiased*** (UMPU) test is one which is at least as powerful as or more powerful than any other unbiased critical region of equal size, with respect to all alternative hypothesis.

13.9 Further Readings

- Devore, Jay L. Berk, Kenneth N. Modern Mathematical Statistics with Applications Second Edition, Springer-Verlag 2012

- Gun A. M., Gupta M. K., Dasgupta B., An Outline of Statistical Theory, The World Press Private limited, Kolkata.
- Gupta S. C. and Kapoor V. K. Fundamentals of mathematical statistics, Sultan Chand & Sons. New Delhi.
- Hogg, Robert V., McKean, Joseph W., and Craig, Allen T. Introduction to Mathematical Statistics, 6th ed., Pearson Prentice Hall, 2005.
- Kendall M.G. and Stuart A., The Advance Theory of Statistics Vol-2, Charles Griffin & Co. Limited, London.
- Larsen J Richard. and Marx L Morris, An Introduction to Mathematical Statistics and Its Applications, 5th Edition, Prentice Hall.
- Rohatgi, V. K., An Introduction to Probability Theory and Mathematical Statistics. John Wiley & Sons, 1976, New York.
- Saxena, H. C. and Surendran, P. U. Statistical Inference: S. Chand & Company Limited,1990, New Delhi.
- Shao, Jun. Mathematical Statistics: Springer-Verlag ,1999, New York.
- Stapleton, James H. Models for Probability and Statistical Inference Theory and Applications, John Wiley & Sons, 2008.
- Wilks, Samuel S. Mathematical Statistics, John Wiley & Sons, 1962, New York.

We acknowledge all above authors as we had referred them extensively while creating this self-Study material.

Notes

Notes