UNIT 16 POINT ESTIMATION

Structure

16.1 INTRODUCTION

In Unit 15, you have been introduced to the problem of point estimation and also to some basic concepts of the theory of point estimation. There we have also discussed two desirable properties of an estimator, viz., unbiasedness and consistency. In this unit, the problem of point estimation will be discussed in greater detail. To begin with, we shall introduce some more concepts. Next, some methods of point estimation are discussed. In particular, we shall concentrate on two methods of estimation that are used widely in practice, viz., the method of moments and the method of maximum likelihood. The first one is easy to implement in practice and the latter leads to estimators with "good" properties.

Objectives

After reading this unit, you should be able to;

- list the criteria for the choice of a good estimator
- derive estimators by one of the methods discussed
- decide which one in a given class of estimators is best according to a given criterion
- assess the goodness or otherwise of any given estimator.

16.2 PROPERTIES OF AN ESTIMATOR

We have already discussed in Unit 15 two properties of an estimator, namely, unbiasedness and consistency. Let us recall the definitions of unbiasedness and consistency.

Definition 1: An estimator $T(X_1, X_2, \ldots, X_n)$, which is a function of the sample values $X_1, X_2, \ldots, X_n$ is unbiased for $g(\theta)$, a known function of the parameter θ, if

$$E_\theta \left[T(X_1, X_2, \ldots, X_n) \right] = g(\theta) \text{ for all } \theta \in \Omega$$

where E_θ denotes the expectation taken when θ is the parameter and Ω is the parameter space

Definition 2: An estimator $T_n = T(X_1, X_2, \ldots, X_n)$ is said to be a consistent estimator of θ if

$$P_\theta \left[|T_n - \theta| > \varepsilon \right] \to 0 \text{ as } n \to \infty \text{ for every } \varepsilon > 0$$

In a given problem there might exist more than one unbiased estimator for the same parameter θ or the same parametric function $g(\theta)$. How do we choose among these unbiased estimators? As mentioned in Unit 15, one way to choose among various unbiased estimators for the same parameter is to compare their variances. That is, if $T_1(X_1, X_2, \ldots X_n)$ and $T_2(X_1, X_2, \ldots X_n)$ are two unbiased estimators of $g(\theta)$, then T_1 will be preferred over T_2 if

$\text{Var}_\theta(T_1) \leq \text{Var}_\theta(T_2)$ for all $\theta \in \Omega$ and with strict inequality for at least one $\theta \in \Omega$.

This brings us to the concept of **uniformly minimum variance unbiased estimators** (UMVUE). We have the following definition.

Definition 3: For a fixed sample size, n, $T = T(X_1, X_2, \ldots X_n)$ is called a minimum variance unbiased estimator of $g(\theta)$ if (i) $E_\theta(T) = g(\theta)$ for all $\theta \in \Omega$, i.e., T is unbiased for $g(\theta)$, and (ii) $\text{Var}_\theta(T) \leq \text{Var}_\theta(T')$ for all $\theta \in \Omega$ with strict inequality for at least one $\theta \in \Omega$, where T' is any other estimator based on $X_1, X_2, \ldots X_n$ satisfying (i).

How do we locate a minimum variance unbiased estimator in a given problem ? From definition 3 alone, it may be a very difficult task, if not impossible, to find a minimum variance unbiased estimator. The following example illustrates this fact.

Example 1: Suppose a random variable X follows a normal distribution with mean θ and variance unity, and let $X_1, X_2, \ldots X_{10}$ be a random sample of size 10 from the population . We know that $\overline{X}$, the sample mean, is unbiased for θ and so is X_1. Now, $\text{Var}_\theta(\overline{X}) = 1/10$, $\text{Var}_\theta(X_1) = 1$. Therefore, $\overline{X}$ is superior to X_1 for estimating θ unbiasedly. However, this does not necessarily mean that $\overline{X}$ is the minimum variance unbiased estimator of θ. To check whether $\overline{X}$ indeed is the minimum variance unbiased estimator of θ, it will be necessary to compare the variance of $\overline{X}$ with the variances of all other unbiased estimators of θ, which is clearly an impossible task. One has therefore take recourse to other methods for locating an unbiased estimator with the smallest variance in the class of all unbiased estimators.

To formalize the concepts, we now consider a population with probability density function (if the random variable in question is continuous) or probability mass function (in the discrete case) $f(x ; \theta)$ where the parameter $\theta \in \Omega \subset R$ is a scalar. The set of all x where $f(x ; \theta) \neq 0$ is called the **support** of $f(x ; \theta)$. We shall **assume that the support of** $f(x ; \theta)$ **is independent of** θ. For example, our discussion will not be applicable to a uniform distribution over the interval $(0, \theta)$, since the support $(0, \theta)$ is dependent on the parameter θ.

The problem is to estimate a parameter θ on the basis of the data $X_1, X_2, \ldots X_n$, which is a random sample of size n from $f(x ; \theta)$. At this stage, it is important to bring in the notion of a likelihood function. Let $X_1, X_2, \ldots X_n$ be a random sample from $f(x ; \theta)$ where $f(x ; \theta)$ is the probability density (or mass) function of a random variable X. The joint probability density or mass function of $X_1, X_2, \ldots X_n$ for given θ, is

$$\prod_{i=1}^{n} f(x_i ; \theta) = L_n(\theta), \text{ say,}$$

where $x_1, x_2, \ldots x_n$ are a realization of $X_1, X_2, \ldots X_n$ for the given sample. If θ is unknown and varies over Ω, $L_n(\theta)$ may be regarded as a function of the variable θ, and is called the likelihood function of θ.

We shall henceforth assume that X is continuous and hence f (x ; θ) is a probability density function. The likelihood function based on the sample $X_1, X_2, \ldots X_n$ is

$$L_n(\theta) = f(x_1 ; \theta) f(x_2 ; \theta) \ldots . f(x_n ; \theta).$$

$\underline{X} = (X_1, X_2, \ldots, X_n)$
$d\underline{x} = (dx_1, dx_2, \ldots, dx_n)$

Suppose $g(\underline{x})$ is an estimator of θ such that

$$E_\theta[g(\underline{X})] < \infty. \text{ Let}$$

$$B(\theta) = E_\theta[g(\underline{X})] - \theta$$

B (θ) is called the bias of the estimator $g(\underline{X})$ in estimating θ. Clearly, if $g(\underline{X})$ is unbiased for θ, then B (θ) = 0. Now,

$$\ln L_n(\theta) = \sum_{i=1}^{n} \ln f(x_i ; \theta)$$

and assuming f (x ; θ) to be differentiable w.r.t. θ.

$$\frac{d}{d\theta} \ln L_n(\theta) = \sum_{i=1}^{n} \frac{d}{d\theta} \ln f(x_i ; \theta)$$

The function $\frac{d}{d\theta} \ln L_n(\theta)$ is called the score function based on the observations $X_1, X_2, \ldots X_n$. Now, since f (x ; θ) is a density function, we have

$$\underbrace{\int_{-\infty}^{\infty} \int_{-\infty}^{\infty}}_{\text{n times}} f(x_1 ; \theta) f(x_2 ; \theta) \ldots . f(x_n ; \theta)\, dx_1, dx_2, \ldots, dx_n = 1$$

for all θ.

For brevity, we write the above equation as

$$\int_{R^n} \prod_{i=1}^{n} f(x_i ; \theta)\, d\underline{x} = 1 \qquad (1)$$

Since $E_\theta[g(\underline{X})] = \theta + B(\theta)$, we have

$$\int_A g(\underline{x}) \prod_{i=1}^{n} f(x_i ; \theta)\, d\underline{x} = \theta + B(\theta) \qquad (2)$$

where A is that part of R^n where $L_n(\theta)$ is positive.

We now assume the (1) and (2) can be differentiated w.r.t. θ under the integral sign. Then.

$$\frac{d}{d\theta}\left[\int_A L_n(\theta)\, d\underline{x}\right] = \int_A \frac{d}{d\theta} L_n(\theta)\, d\underline{x} = 0 \qquad (3)$$

and

$$\frac{d}{d\theta}\left[\int_A g(\underline{x}) L_n(\theta)\, d\underline{x}\right] = \int_A g(\underline{x}) \frac{d}{d\theta} L_n(\theta)\, d\underline{x}$$

$$= 1 + B'(\theta) \qquad (4)$$

where $B'(\theta) = \frac{d}{d\theta} B(\theta)$

Making use of the relation

$\frac{d}{d\theta} L_n(\theta) = \left(\frac{d}{d\theta} \ln L_n(\theta) \right) L_n(\theta)$, we can write (3) and (4) alternatively as

$$\int_A \left[\frac{d}{d\theta} \ln L_n(\theta) \right] L_n(\theta)\, d\underline{x} = 0 \tag{5}$$

and $$\int_A g(\underline{x}) \frac{d}{d\theta} \left[\ln L_n(\theta) \right] L_n(\theta)\, d\underline{x} = 1 + B'(\theta) \tag{6}$$

respectively.

Since $L_n(\theta)$ is the joint density of $X_1, X_2, \ldots, X_n$ when θ is the parameter, the relations (5) and (6) may be written in terms of expectations, as

$$E_\theta \left[\frac{d}{d\theta} \ln L_n(\theta) \right] = 0 \tag{7}$$

and $$E_\theta \left[g(\underline{X}) \frac{d}{d\theta} \ln L_n(\theta) \right] = 1 + B'(\theta) \tag{8}$$

Combining (7) and (8) we have

$$E_\theta \left[(g(\underline{X}) - \theta) \frac{d}{d\theta} \ln L_n(\theta) \right] = 1 + B'(\theta). \tag{9}$$

The Cauchy-Schwarz inequality states that for any two random variables U and V with $E(U^2) < \infty, E(V^2) < \infty$,

$$\left[E_\theta(UV) \right]^2 \le E_\theta(U^2) E_\theta(V^2) \tag{10}$$

with equality if and only if U and V are linearly related.

Let $$U = g(\underline{X}) - \theta, V = \frac{d}{d\theta} \ln L_n(\theta)$$

Then from (10) we have

$$\left[1 + B'(\theta) \right]^2 = \left[E_\theta \left(g(\underline{X}) - \theta \right) \frac{d}{d\theta} \ln L_n(\theta) \right]^2$$

$$\le E_\theta \left[g(\underline{X}) - \theta \right]^2 E_\theta \left[\frac{d}{d\theta} \ln L_n(\theta) \right]^2$$

or $$E_\theta \left[g(\underline{X}) - \theta \right]^2 \ge \frac{\left[1 + B'(\theta) \right]^2}{I_n(\theta)} \tag{11}$$

where $I_n(\theta) = E_\theta \left[\frac{d}{d\theta} \ln L_n(\theta) \right]^2$. $I_n(\theta)$ is called the **Fisher information** in the sample $(X_1, X_2, \ldots, X_n)$. The equality (11) is known as the **Cramer-Rao inequality.**

It can be shown that

$$I_n(\theta) = nI_1(\theta)$$

where $I_n(\theta)$ is the Fisher information contained in one observation. The inequality (11) can then be written alternatively as

$$E[g(\underline{X}) - \theta]^2 \geq \frac{[1 + B'(\theta)]^2}{nI(\theta)} \qquad (12)$$

where, we write $I(\theta)$ in place of $I_1(\theta)$ for simplicity.

If $g(\underline{X})$ is unbiased for θ, that is, if $E_\theta(g(\underline{X})) = \theta$, then $E_\theta[g(\underline{X}) - \theta]^2 = Var_\theta[g(\underline{X})]$ and $B(\theta) = 0$ and hence $B'(\theta) = 0$. Thus, **for an unbiased estimator** $g(\underline{X})$ of θ, we have

$$Var_\theta[g(\underline{X})] \geq 1/\{nI(\theta)\} \qquad (13)$$

The lower bound $1/(nI(\theta))$ to the variance of an unbiased estimator $g(\underline{X})$ of θ, is called the **Cramer - Rao lower bound.** Thus if the regularity conditions assumed earlier hold, the variance of an unbiased estimator $g(\underline{X})$ of θ cannot be smaller than $1/(nI(\theta))$ and hence if an unbiased estimator of a θ has variance equal to $1/(nI(\theta))$. It is the minimum variance unbiased estimator of θ.

If $g(\underline{X})$ is an unbiased estimator of $\delta(\theta)$, a known function of θ, the Cramer-Rao inequality takes the form

$$Var_\theta[g(\underline{X})] \geq [\delta'(\theta)]^2/\{nI(\theta)\} \qquad (14)$$

We can now define an **efficient estimator.**

Definition 4: An unbiased estimator $g(\underline{X})$ of $\delta(\theta)$ is said to be efficient in the Cramer-Rao sense if its variance is equal to the lower bound $[\delta'(\theta)]^2/[nI(\theta)]$ where n is the sample size and $I(\theta)$ is the Fisher information in a single observation.

It is also a uniformly minimum variance unbiased estimator (UMVUE) of $\delta(\theta)$ in the sense that it has the smallest variance uniformly for all $\theta \in \Omega$ in the class of all unbiased estimators.

Note that it is possible that there exists a uniformly minimum variance unbiased estimator for $\delta(\theta)$ but the variance of this estimator does not attain the Cramer - Rao lower bound.

The Fisher information $I(\theta)$ can be shown to be equal to

$$-E_\theta\left[\frac{d^2}{d\theta^2} \ln f(x;\theta)\right].$$

This is sometimes computationally simpler compared to the formula

$E_\theta\left[\frac{d}{d\theta} \ln L_n(\theta)\right]^2$, given earlier.

Example 2: Let $X_1, X_2, \ldots, X_n$ be a random sample from a normal population with unknown mean μ and variance unity. The density function of a normal random variable with mean μ and variance unity is

$$f(x;\mu) = (2\pi)^{-1/2} \exp\left[-\frac{1}{2}(x-\mu)^2\right]$$

and thus

$$\ln f(x;\mu) = -\frac{1}{2}\ln 2\pi - \frac{1}{2}(x-\mu)^2,$$

$$\frac{d}{d\mu}\ln f(x;\mu) = x-\mu$$

$$\frac{d^2}{d\mu^2}\ln f(x;\mu) = -1.$$

Hence $I(\mu) = -E\left[\frac{d^2}{d\mu^2}\ln f(x;\mu)\right] = 1$ and the Cramer - Rao lower bound is $1/[\,nI(\mu)\,] = 1/n$. Now, we know that $\overline{X} = n^{-1}\sum_{i=1}^{n} X_i$ is unbiased for μ and $\mathrm{Var}_\mu(\overline{X}) = 1/n$. Therefore, $\mathrm{Var}_\mu(\overline{X})$ attains the Cramer-Rao lower bound and $\overline{X}$ is the UMVUE of μ. It can be shown that there is only one such UMVUE, that is, $\overline{X}$ is the unique UMVUE of μ.

Example 3: Let for $n \geq 3$, $X_1, X_2, \ldots, X_n$ denote a random sample of size n from a Poisson population with parameter λ. Then, $\overline{X} = n^{-1}\sum_{i=1}^{n} X_i$ is unbiased for λ, and $\mathrm{Var}_\lambda(\overline{X}) = \lambda/n$. Now,

$$f(x;\lambda) = e^{-\lambda}\lambda^x/x!,$$

$$\ln f(x;\lambda) = -\lambda + x\ln\lambda - \ln(x!)$$

$$\frac{d}{d\lambda}\ln f(x;\lambda) = -1 + x/\lambda,$$

and

$$\frac{d^2}{d\lambda^2}\ln f(x;\lambda) = -x/\lambda^2$$

Therefore, $I(\lambda) = -E_\lambda\left[\frac{d^2}{d\lambda^2}\ln f(x;\lambda)\right] = E(x)/\lambda^2 = \lambda^{-1}$.

So that the Cramer-Rao lower bound is $1/[nI(\lambda)] = \lambda/n$. Since $\mathrm{Var}_\lambda(\overline{X}) = \lambda/n$, $\overline{X}$ is the UMVUE of λ.

E1) Let $X_1, X_2, \ldots, X_n$ be independent Bernoulli random variables, that is, $X_1, X_2, \ldots, X_n$ are independent random variables with $P(X_i = 1) = p$, $P(X_i = 0) = 1-p$ for $i = 1, 2, \ldots, n$. Show that if $S = X_1 + X_2 + \ldots + X_n$, S/n is the UMVUE of p.

The next example demonstrates that a uniformly minimum variance unbiased estimator for a parameter might exist but the Cramer-Rao lower bound is not attained.

Example 4: Let X be a Poisson random variable with parameter θ and suppose we wish to estimate $\delta(\theta) = e^{-\theta}$ on the basis of a sample of size one. Consider the estimator

$$T(X) = 1, \text{if } X = 0$$

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

Then, $E_\theta[T(X)] = 1 \cdot P_\theta[X = 0] = e^{-\theta}$, so that $T(X)$ is unbiased for $e^{-\theta}$ Also,

$$\text{Var}_\theta[T(X)] = E_\theta\left[\{T((X)\}^2\right] - \left[E_\theta\{T((X)\}\right]^2$$

$$= E_\theta[\{T(X)\}]^2 - e^{-2\theta}$$

But $E_\theta\left[\{T(X)\}^2\right] = E_\theta[T(X)] = e^{-\theta}$ and hence

$$\text{Var}_\theta(T(X)) = e^{-\theta} - e^{-2\theta} = e^{-\theta}(1 - e^{-\theta}).$$

Now, the probability mass function of X is

$$f(x;\theta) = e^{-\theta}\theta^x/x!$$

and thus, $\ln f(x;\theta) = -\theta + x \ln \theta - \ln(x!)$,

$$\frac{d}{d\theta}\ln f(x;\theta) = -1 + x/\theta$$

$$\frac{d^2}{d\theta^2}\ln f(x;\theta) = -x/\theta^2.$$

$$\text{Hence } I(\theta) = -E_\theta\left[\frac{d}{d\theta}\ln f(x;\theta)\right] = \theta^{-2}E_\theta(x) = \theta^{-1}.$$

Also, $\delta(\theta) = e^{-\theta}$, so that $\delta'(\theta) = \frac{d}{d\theta}\delta(\theta) = -e^{-\theta}$. Hence, the Cramer-Rao lower bound to the variance of $T(X)$, using (14), is

$$[\delta'(\theta)]^2/I(\theta) = \theta e^{2\theta}, \text{ as } n = 1.$$

But $\text{Var}_\theta[T(X)] = e^{-\theta}(1 - e^{-\theta}) > \theta e^{-2\theta}$ for $\theta > 0$. Thus, $T(X)$, though unbiased for $\delta(\theta) = e^{-\theta}$, has a variance larger than the Cramer-Rao lower bound. However, it can be shown that $T(X)$ is the only unbiased estimator of $\delta(\theta) = e^{-\theta}$ and hence is the UMVUE of $e^{-\theta}$.

We now bring in another important concept, namely, that of sufficient statistic and touch upon it briefly. Let X be a random variable having probability density (or, mass) function $f(x;\theta)$ and $X_1, X_2, \ldots, X_n$ be independent observations on X that is, let $X_1, X_2, \ldots, X_n$ be a random sample from a population with density (mass) function $f(x;\theta)$. The joint distribution of $(X_1, X_2, \ldots, X_n)$ clearly depends on θ. Is it possible to find a statistic (a function of $(X_1, X_2, \ldots, X_n)$ which contains all the "information" about θ? Such a question becomes relevant when we want to summarize the available data, because storing large bodies of data is expensive and might give rise to errors of recording etc. Moreover , it is unnecessary if we are able to summarize the data without losing any "information". A statistic containing all information about θ is called a sufficient statistic. We give below a precise definition.

A statistic $T = T(X_1, X_2, \ldots, X_n)$ is said to be a sufficient statistic for the parameter θ if the conditional distribution of $(X_1, X_2, \ldots, X_n)$ given T does not depend on θ.

From the above definition, it is clear that if there is a sufficient statistic for θ, then since the conditional distribution of $X_1, X_2, \ldots, X_n$ given the sufficient statistic is independent of θ, no other function of the observations can have any additional information about θ, given the sufficient statistic.

16.3 METHODS OF ESTIMATION

In this Section, we shall discuss some common methods of finding estimators. We concentrate on two useful and commonly used methods, namely, the method of moments and the method of maximum likelihood.

16.3.1. Method of Moments

The method of moments for estimation of parameters is often used mainly because of its simplicity. The method consists in equating sample moments to population moments and solving the resulting equations to obtain the estimators.

Let $X_1, X_2, \ldots, X_n$ be a random sample from a population with distribution function depending on a k-dimensional parameter $\underline{\theta} = (\theta_1, \theta_2, \ldots\ldots, \theta_k)$. Let

$$m_r' = n^{-1} \sum_{i=1}^{n} X_i^r \quad , \quad n = 1, 2, \ldots,$$

be the r-th sample moment. Suppose $\mu_r' = E(X^r)$ exists for $r = 1, 2, \ldots, k$. The method of moments involves solving the equation

$$m_r' = \mu_r'(\theta_1, \theta_2, \ldots\ldots, \theta_k)., \; 1 \le r \le k.$$

In order to estimate the k components of $\underline{\theta}$, one clearly needs to equate at least k sample moments to k population moments. However, which of the k moments are to be equated is not specified. In practice, one generally takes the first k moments . The method is now illustrated by some examples.

Example 5: Let $X_1, X_2, \ldots, X_n$ be a random sample from a normal population with mean μ and variance σ^2. Here, the parameter $\underline{\theta} = (\mu, \sigma^2)$ is 2-dimensional. In order to obtain the method of moments estimators of μ and σ^2, we equate the first two sample moments to the corresponding population moments, that is,

$$m_1' = n^{-1} \sum_{i=1}^{n} X_i = \overline{X} \text{ is equated to } E(X) = \mu$$

and $m_2' = n^{-1} \sum_{i=1}^{n} X_1^2$ is equated to $E(X^2) = \mu^2 + \sigma^2$.

The first of these two equations gives $\overline{X}$ as an estimator of μ ; $\hat{\mu} = \overline{X}$. From the second, using $\hat{\mu} = \overline{X}$, we have an estimator of σ^2 as

$$\hat{\sigma}^2 = m'_2 - \overline{X}^2$$

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

Observe that $\hat{\mu} = \bar{X}$ is an unbia estimator of μ but $\hat{\sigma}^2$ is not unbiased for σ^2. However, both $\hat{\mu}$ and $\hat{\sigma}^2$ are co...ent estimators of μ and σ^2 respectively.

Example 6: Let $X_1, X_2, \ldots, X_n$ be a random sample from a uniform distribution with density function

$$f(x; \alpha, \beta) = \frac{1}{\beta - \alpha}, \quad \alpha \le x \le \beta$$

$$= 0, \qquad \text{elsewhere.}$$

Then, $\mu'_1 = E(X) = (\alpha + \beta)/2$, $E(X^2) = \mu'_2 = (\alpha^2 + \alpha\beta + \beta^2)/3$.

Instead of equating m'_1 to μ'_1 and m'_2 to μ'_2, we may as well equate m'_1 to μ'_1 and $m_2 = n^{-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$ to $\mu_2 = \text{Var}(X)$. It is easy to see that

$$\text{Var}(X) = E(X^2) - (E(X))^2$$

$$= (\beta - \alpha)^2/12.$$

Thus, the equations to be solved are

$$\bar{X} = (\alpha + \beta)/2$$

and $$n^{-1} \sum_{i=1}^{n} (X_i - \bar{X})^2 = (\beta - \alpha)^2/12.$$

The solution of these equations give us the method of moments estimators of α and β as

$$\hat{\alpha} = \bar{X} - \left[3 \sum_{i=1}^{n} (X_i - \bar{X})^2/n \right]^{1/2}$$

$$\hat{\beta} = \bar{X} + \left[3 \sum_{i=1}^{n} (X_i - \bar{X})^2/n \right]^{1/2}$$

E2) Let $X_1, X_2, \ldots, X_n$ be a random sample from a Poisson distribution with parameter λ. Obtain two estimators of λ using the method of moments.

E3) Let $X_1, X_2, \ldots, X_N$ be a random sample of size N from a binomial population with parameters n and p, both unknown. Obtain the method of moments estimators of n and p.

As we have mentioned earlier, the method of moments is useful in practice because of its simplicity. The properties of such estimators are not established in general and have to be investigated separately for each estimator. Another method, which gives "efficient" estimators for large samples, under some reasonable conditions, is the method of maximum likelihood. We study this method in the following subsection.

16.3.2 Method of Maximum Likelihood

To appreciate this method of estimation, it is perhaps best to start with an example.

Example 7: Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from a Poisson population with parameter θ. The likelihood function, based on the observations $X_1, \ldots, X_n$ is

$$L_n(\theta) = \prod_{i=1}^{n} e^{-\theta} \theta^{x_i} / x_i \,! \; ; \theta > 0.$$

The method of maximum likelihood consists in choosing as an estimator of θ that value of θ (say θ_0) which maximizes the likelihood function $L_n(\theta)$. θ_0 is called the maximum likelihood estimator of θ. Obviously, θ_0 depends on the observed sample $X_1, X_2, \ldots, X_n$. In order to find a maximum likelihood estimator of θ, we have to find the value of θ_0 at which $L_n(\theta)$ is maximum over the interval $(0, \infty)$, as $\theta > 0$ here. Now,

$$\ln L_n(\theta) = -n\theta + \left(\sum_{i=1}^{n} X_i \right) \ln \theta - \sum_{i=1}^{n} \ln (X_i \,!).$$

It is known that $\ln L_n(\theta)$ attains its maximum at a point θ_0 if and only if $L_n(\theta)$ attains its maximum at θ_0 Now,

$$\frac{d}{d\theta} \ln L_n(\theta) = -n + \sum_{i=1}^{n} X_i / \theta.$$

Therefore, $\frac{d}{d\theta} \ln L_n(\theta) \,|_{\theta = \theta_0} = 0$ provided $\theta_0 = n^{-1} \sum_{i=1}^{n} X_i$.

In order to verify whether $L_n(\theta)$ is indeed maximum at $\theta = \theta_0$, we compute the second derivative of $\ln L_n(\theta)$ at $\theta = \theta_0$ and check whether it is negative . Here,

$$\frac{d^2}{d\theta^2} \ln L_n(\theta) = -\theta^2 \sum_{i=1}^{n} X_i$$

and clearly, $\frac{d^2}{d\theta^2} \ln L_n(\theta) \,|_{\theta = \theta_0} < 0$. This shows that $L_n(\theta)$ is maximized at $\theta = \theta_0 = \sum_{i=1}^{n} X_i / n$. Since there is a unique maximum for $L_n(\theta)$ and the maximum is attained at $\theta = \theta_0 = n^{-1} \sum_{i=1}^{n} X_i$, θ_0 is the maximum likelihood estimator of θ.

We next consider an example where the parameter θ is a vector instead of a scalar as in Example 7.

Example 8: Suppose $X_1, X_2, \ldots, X_n$ is a random sample from a normal population with mean μ and variance σ^2, both unknown. The likelihood function is

$$L_n(\mu, \sigma^2) = (2\pi)^{-n/2} \cdot \exp \left[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (X_i - \mu)^2 \right] (\sigma^2)^{-n/2}$$

Thus,

$$\ln L_n(\mu, \sigma^2) = c - \frac{n}{2}\ln \sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2$$

where c is a constant independent of μ and σ^2. The partial derivatives of $\ln L_n(\theta)$ w.r.t. μ and σ^2 are

$$\frac{d}{d\mu}\ln L_n(\mu, \sigma^2) = \sum_{i=1}^{n}(X_i - \mu)/\sigma^2$$

$$\frac{d}{d\sigma^2}\ln L_n(\mu, \sigma^2) = -\frac{n}{2\sigma^2} + \sum_{i=1}^{n}(X_i - \mu)^2/2\sigma^4.$$

Equating these two partial derivatives to zero, we get the likelihood equations. These equations have unique solutions

$$\hat{\mu} = \sum_{i=1}^{n} X_i/n = \bar{X},\ \hat{\sigma}^2 = n^{-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

The verification of the fact that these solutions actually maximize the likelihood function is left to the reader. Hence, $\hat{\mu}$ and $\hat{\sigma}^2$ are the maximum likelihood estimators of μ and σ^2 respectively.

E4) Let $X_1, X_2, \ldots, X_n$ be a random sample from a population with density function

$$f(x;\theta) = \theta^{-1} e^{-x/\theta}, \qquad x > 0$$

$$= 0, \qquad \text{elsewhere.}$$

Find the maximum likelihood estimator of θ.

In the case of a scalar parameter, the likelihood function is a function of one variable (as in the case of Example 7) and if this function is twice differentiable in the domain of its definition, then one can use the methods of Calculus to find the maximum. However, if the parameter θ is a vector parameter, the likelihood function is a function of several variable and finding the points of maxima of such functions might be difficult in general. In such cases, special methods, depending on the problem on hand are needed. Of course, it is possible that the likelihood function may not be differentiable at all and in that case also, we might have to resort to special techniques. The following example is an illustration of such a situation.

Example 9: Let $X_1, X_2, \ldots, X_n$ be a random sample from a uniform distribution with density function

$$f(x;\theta) = 1/\theta, 0 \le X \le \theta$$

$$= 0, \quad \text{elsewhere.}$$

The likelihood function is

$$L_n(\theta) = \theta^{-n} \text{ if } 0 \le X_i \le \theta \text{ for } i = 1, 2, \ldots, n$$

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

We can write the likelihood function alternatively as

$$L_n(\theta) = \theta^{-n}, \text{ if } 0 \le x_{(n)} \le \theta$$

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

Where $x_{(n)}$ is the largest observation in the sample. The derivative of $L_n(\theta)$ does not vanish and hence, we cannot use the methods of Calculus to get a maximum likelihood estimator. However, $L_n(\theta)$ attains its maximum at $\hat{\theta} = x_{(n)}$ and $x_{(n)}$ is the unique maximum likelihood estimator of θ.

There is another way to look at the same problem. Since $L_n(\theta) = \theta^{-n}, 0 \le X_i \le \theta$ is an ever-decreasing function of θ, the maximum can be found by selecting θ as small as possible. Now, $\theta \ge X_i$ for $i = 1, 2, \ldots, n$ and in particular, $\theta \ge x_{(n)}$. Thus, $L_n(\theta)$ can be made no larger than $1/x_{(n)}^n$ and the unique maximum likelihood estimator of θ is $x_{(n)}$.

Are maximum likelihood estimators unbiased and unique in every situation?

The answer to both the above questions is in the negative. That maximum likelihood estimators need not be unbiased is demonstrated by making an appeal to Examples 8 and 9. In Example 8, we had seen that $n^{-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ is the maximum likelihood estimator of σ^2, the variance of a normal population with unknown mean μ. Clearly, this estimator is not unbiased for σ^2. Again, in Example 9, it was demonstrated that $x_{(n)}$, the largest observation in the sample is the maximum likelihood estimator of θ. But, it can be shown that $E_\theta\left(x_{(n)}\right) = n\theta/(n+1)$, so that $x_{(n)}$ is not unbiased for θ.

To see that maximum likelihood estimator need not be unique, consider the following example.

Example 10: Let $X_1, X_2, \ldots, X_n$ be a random sample from a uniform distribution over $\left[\theta - \frac{1}{2}, \theta + \frac{1}{2}\right]$, where θ is unknown and $\theta \in \Omega = \{x : -\infty < x < \infty\}$. The likelihood is

$$L_n(\theta) = 1, \text{ if } \theta - 1/2 \le X_i \le \theta + 1/2 \text{ for } i = 1, 2, \ldots, n$$

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

or,

$$L_n(\theta) = 1, \text{ if } \theta - 1/2 \le \min(X_1, \ldots, X_n) \le \max(X_1, \ldots, X_n) \le \theta + 1/2$$

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

Thus, $L_n(\theta)$ attains its maximum provided

$$\theta - 1/2 \le \min(X_1, \ldots, X_n)$$

and

$$\theta + 1/2 \ge \max(X_1, \ldots, X_n),$$

or, when

$$\theta \le \min(X_1, \ldots, X_n) + 1/2$$

and $\theta \leq \max(X_1, \ldots, X_n) - 1/2$.

This means that any statistic $T(X_1, \ldots, X_n)$ satisfying

$$\max X_i - 1/2 \leq T(X_1, \ldots, X_n) \leq \min_i X_i + 1/2$$

is a maximum likelihood estimator of θ. In fact, for $0 < \alpha < 1$,

$$T(X_1, X_2, \ldots, X_n) = (\max_i X_i - 1/2) + \alpha(\min_i X_i - \max_i X_i + 1)$$

lies in the interval $\max_i X_i - 1/2 \leq T \leq \min_i X_i + 1/2$. Thus, for any α, $0 < \alpha < 1$, the above estimator is a maximum likelihood estimator of θ. In particular, for $\alpha = 1/2$, we get an estimator $T_1 = (\max_i X_i + \min_i X_i)/2$ and for $\alpha = 1/3$, we get the estimator $T_2 = (4 \max_i X_i + 2 \min_i X_i - 1)/6$.

Both T_1 and T_2 are maximum likelihood estimators of θ.

Are there any "good" properties of maximum likelihood estimators?

Before we attempt to answer this question, we introduce the concept of asymptotic efficiency. An estimator T_n based on a sample of size n for a parameter θ is said to be **asymptotically efficient** if $\lim_{n \to \infty} n \operatorname{Var}_\theta(T_n) = 1/I(\theta)$ where $I(\theta)$ is the per observation (Fisher) information. Recall that the Cramer-Rao lower bound to $\operatorname{Var}_\theta(T_n)$ is $1/\{nI(\theta)\}$, under some regularity conditions.

The important properties of maximum likelihood estimators are that under certain regularity conditions, these estimators are

(i) Consistent

(ii) Asymptotically efficient

(iii) Asymptotically normal with mean θ and variance $1/\{nI(\theta)\}$.

The third property says that for large samples, the distribution of the maximum likelihood estimator $\hat{\theta}$ of θ is approximately normal with mean θ and variance $1/\{nI(\theta)\}$.

The exact statements of the above results and their proofs are beyond the scope of this course and are therefore not given here.

15.4 SUMMARY

In this unit, we have

1. discussed some properties that an estimator should preferably possess, like unbiasedness, consistency and efficiency,
2. derived the Cramer-Rao lower bound to the variance of an estimator and demonstrated the use of this bound in finding minimum variance unbiased estimators,
3. discussed two commonly used methods of estimation, namely, the method of moments and the method of maximum likelihood.

16.5 SOLUTIONS AND ANSWERS

E1) Here, the probability mass function of the random variable, X is

$$f(x;p) = p^x (1-p)^{1-x}, X = 0, 1.$$

Therefore $\ln f(X;p) = X \ln p + (1 - X) \ln (1-p)$,

$$\frac{d}{dp} \ln f(X;p) = X/p - (1-X)/(1-p),$$

and $$\frac{d^2}{dp^2} \ln f(X;p) = -X/p^2 - (1-X)/(1-p)^2.$$

Hence $$I(p) = E_p\left[-\frac{d^2}{dp^2} \ln f(X;p) \right]$$

$$= E_p\left[X/p^2 + (1-X)/(1-p)^2 \right]$$

$$= 1/p + 1/(1/p) = 1/\{ p(1-p) \},$$

since $E_p(X) = p$. Therefore, the Cramer-Rao lower bound to the variance is $p(1-p)/n$. Let $S = X_1 + X_2 + \ldots + X_n$. Then S/n is unbiased for p. Also,

$Var_p(S/n) = n^{-2} \sum_{i=1}^{n} Var_p(X_i) = n^{-2}[np(1-p)] = p(1-p)/n$. Hence S/n is the UMVUE of p.

E2) Here $X_1, X_2, \ldots . X_n$ is a random sample from a Poisson distribution with parameter λ. Hence $E(X_i) = \lambda$ for $i = 1, 2, \ldots, n$. Equating the sample mean to the population mean leads to the following equation:

$$\overline{X} = n^{-1} \sum_{i=1}^{n} X_i = \lambda$$

which gives a moments estimator of λ as $\hat{\lambda} = \overline{X}$. Again, since $E(X^2) = \lambda^2 + \lambda$, equating the second sample moment about zero, viz., $n^{-1} \sum_{i=1}^{n} X_i^2$ to the corresponding population moment yields the equation

$$n^{-1} \sum_{i=1}^{n} X_i^2 = \lambda^2 + \lambda.$$

Since $\lambda > 0$, a unique positive solution of the above equation gives the second moments estimator of λ as

$$\hat{\lambda} = \left[-1 + \left[(4/n) \sum_{i=1}^{n} X_i^2 + 1 \right]^{1/2} \right]/2$$

E3) We are given that $X_1, X_2, \ldots . X_N$ is a random sample from a binomial population with parameters n and p, both unknown. We know that if X has a binomial distribution with parameters n and p, then

$$E_p(X) = np, Var_p(X) = np(1-p).$$

Therefore, $E_p(X^2) = Var_p(X) + \left(E_p(X) \right)^2 = np(1-p) + n^2 p^2$. If we

equate the first two sample moments $N^{-1}\sum_{i=1}^{N} X_i$ and $N^{-1}\sum_{i=1}^{N} X_i^2$ to the first two population moments, the following equations result:

$$\bar{X} = N^{-1}\sum_{i=1}^{N} X_i = np$$

$$s_0^2 = N^{-1}\sum_{i=1}^{N} X_i^2 = np(1-p) + n^2p^2.$$

The first of these gives $\hat{p} = \bar{X}/n$ as an estimator of p, where $\hat{n}$ is an estimator of n. Using this estimator in the second equation and solving for n gives

$$\hat{n} = \frac{\bar{X}^2}{\bar{X}^2 + \bar{X} - s_0^2} = \frac{\bar{X}^2}{\bar{X}^2 + \bar{X} - N^{-1}\sum_{i=1}^{n} X_i^2}$$

E4) Here $X_1, X_2, \ldots . X_n$ is a random sample from a population with density function

$$f(X;\theta) = \theta^{-1}\exp(-X/\theta),\ X > 0,\ \theta > 0$$

$$= 0, \text{ elsewhere}$$

Therefore, the likelihood function is

$$L_n(\theta) = \theta^{-n}\exp\left(-\sum_{i=1}^{n} X_i/\theta\right)$$

$$\ln L_n(\theta) = -n\ln\theta - \sum_{x=1}^{n} X_i/\theta$$

and $\frac{d}{d\theta}\ln L_n(\theta) = -n/\theta + \sum_{x=1}^{n} X_i/\theta^2.$

Equating $\frac{d}{d\theta}\ln L_n(\theta)$ to zero, gives on solving for θ,

$\hat{\theta} = \sum_{i=1}^{n} X_i/n = \bar{X}$ the sample mean.

Also,

$$\frac{d^2}{d\theta^2}\ln L_n(\theta) = -n/\theta^2 - 2\sum_{i=1}^{n} X_i/\theta^3$$

which is negative at $\theta = \hat{\theta} = \bar{X}$. Hence $\bar{X}$ is the maximum likelihood estimator of θ.

16.6 ADDITIONAL EXERCISES

1. Let $X_1, X_2, \ldots . X_n$ be a random sample from a distribution with density function

$$f(x ; \theta) = \theta^{-1} e^{-x/\theta}, \theta > 0, \text{ if } x > 0$$

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

Show that $\overline{X} = n^{-1} \sum_{i=1}^{n} X_i$ is unbiased for θ and $\text{Var}_\theta (\overline{X}) = \theta^2/n$.

Does $\text{Var}_\theta (\overline{X})$ attain the Cramer-Rao lower bound?

2. Let $X_1, X_2, \ldots . X_n$ be a random sample from a normal population with mean zero and variance σ^2. Construct an unbiased estimator of σ as a function of $\sum_{i=1}^{n} | X_i |$. You are given that if X is normal with mean zero and variance σ^2,

$$E (| X |) = \sigma \sqrt{\frac{2}{\pi}}.$$

3. Let $X_1, X_2, \ldots . X_n$ be a random sample from a distribution having finite mean μ and finite variance σ^2. Show that $T (X_1, X_2, \ldots . X_n) = \frac{2}{n (n + 1)} \sum_{i=1}^{n} i \, X_i$ is unbiased for μ.

4. Let $X_1, X_2, \ldots . X_n$ be a random sample of size n from a distribution with probability density function

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

$$= 0, \text{ elsewhere.}$$

Obtain a maximum likelihood estimator of θ.