Module 33

Statistical Inference Problems: Point Estimation

Introduction to Statistical Inference

The basic situation in **statistical inference problems** is as follows:

- We seek information about characteristics of a collection of elements, called **population**;
- Due to various considerations (such as time, cost etc.) we may not wish or be able to study each individual element of the population;
- Our object is to draw conclusions about the unknown population characteristics on the basis of information on characteristics of a suitably selected **sample** from the population;
- Formally, let the r.v. $\underline{X}$ (which may be vector valued) describes the characteristics of the population under investigation and let $F(\cdot)$ be the d.f. of $\underline{X}$;

Parametric Statistical Inference:

- Here the r.v.$\underline{X}$ has a d.f. $F \equiv F_{\underline{\theta}}(\cdot)$ with a known functional form (except perhaps for the parameter $\underline{\theta}$, which may be a vector valued);

- Let Θ be the set of possible values of the unknown parameter $\underline{\theta}$. In problems of **parametric statistical inference**, the statistician's job is to decide, on the basis of a suitably selected sample (generally a random sample) from $F_{\underline{\theta}}(\cdot)$, which member or members of the family $\{F_{\underline{\theta}}(\cdot) : \underline{\theta} \in \Theta\}$ can represent the d.f. of $\underline{X}$;

Nonparametric Statistical Inference:

- Here we know nothing about the functional form of the d.f. $F(\cdot)$ (except perhaps that $F(\cdot)$ is, say, continuous or discrete);
- Our goal is to make inferences about the unknown d.f. $F(\cdot)$;

- In statistical inference problems, the statistician can observe n independent observations on $\underline{X}$, the r.v. describing the population under investigation, i.e., the statistician observes n values $\underline{x}_1, \dots, \underline{x}_n$ assumed by the r.v. $\underline{X}$;

- Each $\underline{x}_i$ can be regarded as the value assumed by a r.v. $\underline{X}_i$, $i = 1, \dots, n$, having the d.f. $F(\cdot)$;

- The observed values $(\underline{x}_1, \dots, \underline{x}_n)$ are then the values assumed by $(\underline{X}_1, \dots, \underline{X}_n)$;

- The set $\{\underline{X}_1, \dots, \underline{X}_n\}$ is then a **random sample** of size n taken from the population having d.f. $F(\cdot)$;

- The observed value $(\underline{x}_1, \dots, \underline{x}_n)$ is called a **realization** of the random sample.

Definition 1

(a) The space of the possible values of the random sample $(\underline{X}_1, \ldots, \underline{X}_n)$ is called the **sample space**. We will denote the sample space by χ. Generally the sample space χ is the same as the support $S_{\underline{X}}$ of the distribution of random sample or its interior.

(b) In the parametric statistical inference problems, the set Θ of possible values of the unknown parameter $\underline{\theta}$ is called the **parameter space**.

Some Parametric Statistical Inference Problems

Consider the following example.

Example 1.

- A manager wants to make inferences about the mean lifetime of a brand of an electric bulb manufactured by a certain company;
- Here the population under investigation consists of lifetimes of all the electric bulbs produced by that company;
- Suppose that the r.v. X represents the lifetime of a typical electric bulb manufactured by the company, i.e., the r.v. X describes the given population;

- Probability modelling on the past experience with testing of similar electric bulbs indicates that X has an exponential distribution with mean θ, i.e., X has the d.f.

$$F_X(x|\theta) = \begin{cases} 0, & \text{if } x < 0 \\ 1 - e^{-\frac{x}{\theta}}, & \text{if } x \geq 0 \end{cases};$$

- But the value of $\theta \in \Theta = (0, \infty)$ is not evident from the past experience and the manager wants to make inferences about the unknown parameter $\theta \in \Theta$;

- Here $\Theta = (0, \infty)$ is the parameter space. Due to various considerations (e.g., time, cost etc.), the statistician can not obtain the lifetimes of all the bulbs produced by the company;

- One way to obtain information about the unknown θ is to do testing, under identical conditions, on a number, say n, of electric bulbs produced by the company;

- This leads to observing a realization $\underline{x} = (x_1, \ldots, x_n)$ of a random sample $\underline{X} = (X_1, \ldots, X_n)$ from the population;

- Here, $\underline{X} = (X_1, \ldots, X_n) \in \chi = \mathbb{R}_+ = \{(t_1, \ldots, t_n) : 0 \leq t_i < \infty, i = 1, \ldots, n\}$ and χ is the sample space;

- On the basis of the realization $\underline{x}$ of the random sample $\underline{X}$, the manager may want answer several questions concerning unknown θ. Some of these may be:

(a) How to obtain a point estimate of θ? This is an example of a **point estimation problem**;

(b) How to obtain an appropriate interval in which the unknown θ lies with certain confidence? This is an example of a **confidence interval estimation problem** of finding an appropriate random interval (depending on $\underline{X}$) for the unknown θ such that the given random interval contains the true θ with given confidence (probability);

(c) To verify the claim (hypothesis) that $\theta \in \Theta_0$, where $\Theta_0 \subset \Theta$. This is an example of **hypothesis testing problem**.

Point Estimation Problems

- $\underline{X}$: a r.v. defined on a probability space $(\Omega, \mathcal{F}, P)$;
- $\underline{\mathrm{X}}$ has a d.f. $F(\cdot|\underline{\theta})$, the functional form of which is known and $\underline{\theta} \in \Theta$ is unknown; here Θ is the parameter space;
- The basic situation in point estimation problems is as follows:
  - We observe r.v.s $\underline{X}_1, \ldots, \underline{X}_n$ (say, a random sample) from the population described by the d.f. $F(\cdot|\underline{\theta})$;
  - based on random sample $\underline{X}_1, \ldots, \underline{X}_n$ we seek an approximation (or an estimate) of $\underline{\theta}$ (or some function of $\underline{\theta}$).

Definition 2.

(a) Let $\underline{g}(\underline{\theta})$ (possibly a vector valued) be a function of $\underline{\theta}$ which we want to estimate. Then $g(\underline{\theta})$ is called the **estimand**.

(b) Let $\Lambda = \{\underline{g}(\underline{\theta}) : \underline{\theta} \in \Theta\} \subseteq \mathbb{R}^q$ be the range of possible values of the estimand $\underline{g}(\underline{\theta})$. A statistic $\underline{\delta} \equiv \underline{\delta}(\underline{X})$ is said to be an **estimator** of $\underline{g}(\underline{\theta})$ if $\underline{\delta}$ maps the sample space χ into $\mathbb{R}^q$; here $\underline{X} = (\underline{X}_1, \ldots, \underline{X}_n)$.

Definition 2 (continued)

(c) Let $\underline{x} = (\underline{x}_1, \ldots, \underline{x}_n)$ be a sample realization of $\underline{X} = (\underline{X}_1, \ldots, \underline{X}_n)$ and let $\underline{\delta} \equiv \underline{\delta}(\underline{X})$ be an estimator of $\underline{g}(\underline{\theta})$. Then $\underline{\delta}(\underline{x})$ is called an **estimate** of $\underline{g}(\underline{\theta})$ (i.e., an estimate is a realization of an estimator);

Note: An estimator is a r.v.. To cover more general situations, in the definition of an estimator we allow it to assume values outside Λ, the set of possible values of the estimand $\underline{g}(\underline{\theta})$ (although it may look absurd).

Example 2.

- Let $X_1, \ldots, X_n$ be a random sample from a Poisson(θ) distribution, where $\theta \in \Theta = (0, \infty)$;

- Let the estimand be $g(\theta) = \theta$;

- Then $\delta_1(\underline{X}) = \overline{X}$ is an estimator of $g(\theta)$, so also is $\delta_2(\underline{X}) = S^2$.

- By the definition, $\delta_3(\underline{X}) = (-1)^{X_1} X_1$ is also an estimator of $g(\theta)$, but it is absurd since it can assume negative values whereas the estimand $g(\theta)$ is positive.

We will now discuss two commonly used methods of parametric point estimation, namely the **Method of Moments** and the **Method of Maximum Likelihood**.

The Method of Moments

- $X_1, \ldots, X_n$: a random sample of size n from a population having distribution function $F_{\underline{\theta}}(x)$, $x \in \mathbb{R}$ (i.e., the r.v. describing the population has the d.f. $F_{\underline{\theta}}(\cdot)$), where $\underline{\theta} = (\theta_1, \ldots, \theta_p) \in \Theta$ is an unknown parameter;

- Suppose that, for $k = 1, \ldots, p$, $m_k = E_{\underline{\theta}}(X_1^k)$ exists and is finite. Here and elsewhere $E_{\underline{\theta}_0}(\cdot)$ $(P_{\theta_0}(\cdot))$ represents that the expectation (probability) is calculated under the d.f. $F_{\underline{\theta}_0}(\cdot)$, $\underline{\theta}_0 \in \Theta$. Let $m_k = h_k(\underline{\theta})$, $k = 1, \ldots, p$;

- Define

$$A_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k, \quad k = 1, \ldots, p.$$

Definition 3.

(a) m_k $(k = 1, \ldots, p)$ is called the k^{th} population moment (about origin) of the d.f. $F_{\underline{\theta}}(\cdot)$, $\underline{\theta} \in \Theta$;

(b) A_k $(k = 1, \ldots, p)$ is called the k^{th} sample moment (about origin) based on the random sample $X_1, \ldots, X_n$;

(c) The **method of moments** consists of equating A_k with $h_k(\theta_1, \ldots, \theta_p)$, for $k = 1, \ldots, p$, and solving for $\theta_1, \ldots, \theta_p$. The value $(\theta_1, \ldots, \theta_p) = (\hat{\theta}_1, \ldots, \hat{\theta}_p)$, say, so obtained is called the **method of moment estimator** (**M.M.E.**) of $\underline{\theta} = (\theta_1, \ldots, \theta_p)$;

Definition 3 (continued)

(d) Let $g : \Theta \to \Lambda$ be a mapping of Θ onto Λ. If $\hat{\underline{\theta}}$ is the M.M.E. of $\underline{\theta}$, then $g(\hat{\underline{\theta}})$ is called the M.M.E. of $g(\underline{\theta})$.

Remark 1. (a) The method of moments is not applicable when m_k $(k = 1, \ldots, p)$ do not exist (e.g., for the Cauchy distribution with median θ).

(b) M.M.E. may not exist when the underlying equations do not have a solution. Also the M.M.E. may not be unique as the underlying equations may have more than one solution.

Example 3.

Let $X_1, \ldots, X_n$ be a random sample from a Poisson(θ) distribution, where $\theta \in \Theta = (0, \infty)$ is unknown. Then $\hat{\theta} = \overline{X}$ is the M.M.E. of θ.

Solution: We have $m_1 = E(X_1) = \theta$. Thus M.M.E. $\hat{\theta}$ is the solution of equation

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

Example 4.

Let $X_1, \ldots, X_n$ $(n \geq 2)$ be a random sample from $N(\mu, \sigma^2)$ distribution, where $\underline{\theta} = (\mu, \sigma^2) \in \Theta = \{(z_1, z_2) : -\infty < z_1 < \infty, z_2 > 0\}$ is unknown. Then $\hat{\underline{\theta}} = (\overline{X}, \frac{n-1}{n} S^2)$ is the M.M.E. of $\underline{\theta}$.

Solution: We have

$$m_1 = E(X_1) = \mu \qquad \text{and} \qquad m_2 = E(X_1^2) = \sigma^2 + \mu^2.$$

Thus M.M.E. $\hat{\underline{\theta}} = (\hat{\mu}, \hat{\sigma}^2)$ is the solution of

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} X_i, \qquad \hat{\sigma}^2 + \hat{\mu}^2 = \frac{1}{n} \sum_{i=1}^{n} X_i^2$$

$$\Rightarrow \quad \hat{\mu} = \overline{X} \qquad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} X_i^2 - \overline{X}^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X})^2 = \frac{n-1}{n} S^2.$$

The Method of Maximum Likelihood

- $X_1, \ldots, X_n$: a random sample of size n from a population having p.d.f. (or p.m.f.) $f_{\underline{\theta}}(x)$, $x \in \mathbb{R}$ (i.e., the r.v. describing the population has the p.d.f. (or p.m.f.) $f_{\underline{\theta}}(\cdot)$), where $\underline{\theta} = (\theta_1, \ldots, \theta_p) \in \Theta$ is an unknown parameter;

- Then the joint p.d.f. of $\underline{X} = (X_1, \ldots, X_n)$ is

$$f_{\underline{X}}(\underline{x}|\underline{\theta}) = \prod_{i=1}^{n} f_{\underline{\theta}}(x_i), \quad \underline{\theta} \in \Theta.$$

Definition 4. For a given sample realization $\underline{x} = (x_1, \ldots, x_n)$ of the observation on $\underline{X} = (X_1, \ldots, X_n)$, the function

$$L_{\underline{x}}(\underline{\theta}) = f_{\underline{X}}(\underline{x}|\underline{\theta}),$$

considered as a function of $\underline{\theta} \in \Theta$, is called the **likelihood function**.

Remark 2.

In the discrete case, given the sample realization $\underline{x} = (x_1, \ldots, x_n)$,

$$L_{\underline{x}}(\underline{\theta}_0) = f_{\underline{X}}(\underline{x}|\underline{\theta}_0)$$

is the probability of obtaining the observed sample $\underline{x}$, when $\underline{\theta}_0 \in \Theta$ is the true value of $\underline{\theta}$. Therefore, intuitively, it is appealing to find $\underline{\hat{\theta}} \equiv \underline{\hat{\theta}}(\underline{x})$ (provided it exists) such that $L_{\underline{x}}(\underline{\hat{\theta}}) = \sup_{\underline{\theta} \in \Theta} L_{\underline{x}}(\underline{\theta})$, since if such a $\underline{\hat{\theta}}$ exists then it is more probable that $\underline{x}$ came from the distribution with p.d.f. (or p.m.f.) $f_{\underline{X}}(\cdot|\underline{\hat{\theta}})$ than from any of the other distribution $f_{\underline{X}}(\cdot|\underline{\theta})$, $\underline{\theta} \in \Theta - \{\underline{\hat{\theta}}\}$. A similar argument can also be given for absolutely continuous distributions.

Definition 5.

(a) For a given sample realization $\underline{x}$, the **maximum likelihood estimate (m.l.e.)** of the unknown parameter $\underline{\theta}$ is the value $\hat{\underline{\theta}} \equiv \hat{\underline{\theta}}(\underline{x})$ (provided it exists) such that

$$L_{\underline{x}}(\hat{\underline{\theta}}) = \sup_{\underline{\theta} \in \Theta} L_{\underline{x}}(\underline{\theta}).$$

(b) Let $g : \Theta \to \Lambda$ be a mapping of Θ into Λ. Define, for $\underline{\lambda} \in \Lambda$, $\Theta_{\underline{\lambda}} = \{\underline{\theta} \in \Theta : g(\underline{\theta}) = \underline{\lambda}\}$. Then, for a given sample realization $\underline{x}$, the function

$$M_{\underline{x}}(\underline{\lambda}) = \sup_{\underline{\theta} \in \Theta_{\underline{\lambda}}} L_{\underline{x}}(\underline{\theta}),$$

considered as a function of $\underline{\lambda} \in \Lambda$, is called the **likelihood function induced by** $g(\underline{\theta})$.

(c) For a given sample realization $\underline{x}$, the **maximum likelihood estimate** (**m.l.e.**) of the estimand $g(\underline{\theta})$ is the value $\hat{\underline{\lambda}} \equiv \hat{\underline{\lambda}}(\underline{x})$ (provided it exists) such that

$$M_{\underline{x}}(\hat{\underline{\lambda}}) = \sup_{\underline{\lambda} \in \Lambda} M_{\underline{x}}(\underline{\lambda}),$$

where $M_{\underline{x}}(\underline{\lambda})$ is as defined in (b) above.

- (d) The estimator (a r.v.) corresponding to the m.l.e. is called the **maximum likelihood estimator** (M.L.E.).

Remark 3.

(a) (**Maximum likelihood estimate may not be unique**). Let $\underline{x} = (x_1, \ldots, x_n)$ be a sample realization based on a random sample from $U(\theta - \frac{1}{2}, \theta + \frac{1}{2})$ distribution, where $\theta \in \Theta = (-\infty, \infty)$ is an unknown parameter. Then, for $x_{(1)} = \min\{x_1, \ldots, x_n\}$ and $x_{(n)} = \max\{x_1, \ldots, x_n\}$,

$$L_{\underline{x}}(\theta) = \begin{cases} 1, & \text{if } x_{(n)} - \frac{1}{2} \le \theta \le x_{(1)} + \frac{1}{2} \\ 0, & \text{otherwise} \end{cases}.$$

Clearly any estimate $\delta(\underline{x})$ such that $x_{(n)} - \frac{1}{2} \le \delta(\underline{x}) \le x_{(1)} + \frac{1}{2}$ is a m.l.e. In particular $\delta^*(\underline{x}) = \frac{x_{(1)} + x_{(n)}}{2}$ is a m.l.e. of θ.

(b) **Maximum likelihood estimate may be absurd**. Let $\underline{x} = (0, 0, \ldots, 0)$ be a sample realization based on a random sample of size n from a $\text{Bin}(1, \theta)$ distribution, where $\theta \in \Theta = (0, 1)$ is unknown. In this case

$$L_{\underline{x}}(\theta) = (1 - \theta)^n, \quad 0 < \theta < 1.$$

and $\hat{\theta} = \overline{x} = 0$ is the m.l.e., while $\hat{\theta}$ does not belong to Θ.

(c) Since $L_{\underline{x}}(\underline{\theta})$ and $\ln L_{\underline{x}}(\underline{\theta})$ attain their maximum for same values of $\underline{\theta}$, sometimes it is more convenient to work with $\ln L_{\underline{x}}(\underline{\theta})$.

(d) If Θ is an open rectangle in $\mathbb{R}^p$ and $L_{\underline{x}}(\underline{\theta})$ is a positive and differentiable function of $\underline{\theta}$ (i.e., the first order partial derivatives exist in the components of $\underline{\theta}$), then if a m.l.e. $\hat{\underline{\theta}}$ exists, it must satisfy

$$\begin{aligned} \frac{\partial}{\partial\theta_j} \ln L_{\underline{x}}(\underline{\theta}) \Big|_{\underline{\theta}=\hat{\underline{\theta}}} &= 0, \quad j=1,\ldots,p; \quad \underline{\theta}=(\theta_1,\ldots,\theta_p) \\ \Leftrightarrow \quad \frac{\partial}{\partial\theta_j} L_{\underline{x}}(\underline{\theta}) \Big|_{\underline{\theta}=\hat{\underline{\theta}}} &= 0, \quad j=1,\ldots,p. \end{aligned}$$

Result 1.

(**Invariance of the m.l.e.**) Suppose that $\Theta \subseteq \mathbb{R}^p$. Let $g : \Theta \to \Lambda$ be a mapping of Θ into Λ, where Λ is a region in $\mathbb{R}^q$ $(1 \leq q \leq p)$. If $\underline{\hat{\theta}} \equiv \underline{\hat{\theta}}(\underline{x})$ is the m.l.e. of $\underline{\theta}$ and $\underline{\hat{\theta}}(\underline{X}) \in \Theta$ with probability one, then $g(\underline{\hat{\theta}})$ is the m.l.e. of $g(\underline{\theta})$.

Proof: We have $\Theta_{\underline{\lambda}} = \{\underline{\theta} \in \Theta : h(\underline{\theta}) = \underline{\lambda}\}$, $\underline{\lambda} \in \Lambda$ and

$$M_{\underline{x}}(\underline{\lambda}) = \sup_{\underline{\theta} \in \Theta_{\underline{\lambda}}} L_{\underline{x}}(\underline{\theta}), \quad \underline{\lambda} \in \Lambda.$$

Clearly $\{\Theta_{\underline{\lambda}} : \underline{\lambda} \in \Lambda\}$ forms a partition of Θ. Now

$$\underline{\hat{\theta}} \text{ is m.l.e. of } \underline{\theta} \in \Theta \;\Rightarrow\; \underline{\hat{\theta}} \in \Theta \;\Rightarrow\; \underline{\hat{\theta}} \in \Theta_{\underline{\lambda}}, \text{ for some } \underline{\lambda} \in \underline{\Lambda}.$$

Let $\underline{\hat{\theta}} \in \Theta_{\underline{\hat{\lambda}}}$, where $\underline{\hat{\lambda}} \in \Lambda$. Then $h(\underline{\hat{\theta}}) = \underline{\hat{\lambda}}$ (by definition of $\Theta_{\underline{\hat{\lambda}}}$). Also, since $\underline{\hat{\theta}} \in \Theta_{\underline{\hat{\lambda}}}$,

$$L_{\underline{x}}(\underline{\hat{\theta}}) \leq \sup_{\underline{\theta}\in\Theta_{\underline{\hat{\lambda}}}} L_{\underline{x}}(\underline{\theta}) = M_{\underline{x}}(\underline{\hat{\lambda}}) \leq \sup_{\underline{\lambda}\in\Lambda} M_{\underline{x}}(\lambda) = \sup_{\underline{\theta}\in\Theta} L_{\underline{x}}(\underline{\theta}) = L_{\underline{x}}(\underline{\hat{\theta}})$$

$$\Rightarrow \quad M_{\underline{x}}(\underline{\hat{\lambda}}) = \sup_{\underline{\lambda}\in\Lambda} M_{\underline{x}}(\lambda)$$

$$\Rightarrow \quad \underline{\hat{\lambda}} = h(\underline{\hat{\theta}}) \text{ is an m.l.e. of } h(\underline{\theta}).$$

Regularity Conditions R_1:

(a) The parameter space Θ is an open interval in $\mathbb{R}$ (finite, infinite or semi-finite);

(b) The support $S_{\underline{X}} = \{\underline{x} : f_{\underline{X}}(\underline{x}|\theta) > 0\}$ does not depend on θ.

(c) For any $\underline{x} \in S_{\underline{X}}$ and any $\theta \in \Theta$, the derivative $\frac{\partial}{\partial\theta} f_{\underline{X}}(\underline{x}|\theta)$, $\theta \in \Theta$, exists and is finite and

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{\underline{X}}(\underline{x}|\theta) d\underline{x} = 1,\ \theta \in \Theta,$$

can be differentiated under the integral (or summation) sign, so that

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \frac{\partial}{\partial\theta} f_{\underline{X}}(\underline{x}|\theta) d\underline{x} = \frac{d}{d\theta} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{\underline{X}}(\underline{x}|\theta) d\underline{x} = 0, \quad \forall\, \theta \in \Theta,$$

with integrals replaced by the summation sign in the discrete case.

(d) For any $\underline{x} \in S_{\underline{X}}$ and any $\theta \in \Theta$, the second partial derivative $\frac{\partial^2}{\partial\theta^2} f_{\underline{X}}(\underline{x}|\theta)$, $\theta \in \Theta$, exists and is finite and

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \frac{\partial}{\partial\theta} f_{\underline{X}}(\underline{x}|\theta) d\underline{x} = 0, \quad \theta \in \Theta,$$

can be differentiated under the integral (summation) sign, so that

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \frac{\partial^2}{\partial\theta^2} f_{\underline{X}}(\underline{x}|\theta) d\underline{x} = \frac{d^2}{d\theta^2} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{\underline{X}}(\underline{x}|\theta) d\underline{x} = 0, \forall\theta \in \Theta,$$

with integrals replaced by the summation sign in the discrete case.

Remark 4.

(a) Using advanced mathematical arguments, it can be shown that the regularity conditions R_1 are satisfied for a large family of distributions, including the exponential family of distributions having associated p.d.f.s (or p.m.f.s) of the form

$$f_{\underline{X}}(\underline{x}|\theta) = c(\theta)h(\underline{x})e^{r(\theta)T(\underline{x})}, \quad \underline{x} \in \chi, \quad \theta \in \Theta,$$

for some functions $h(\cdot)$, $c(\cdot)$, $r(\cdot)$ and $T(\cdot)$ and an open interval $\Theta \subseteq \mathbb{R}$.

(b) For $\underline{x} \in S_{\underline{X}}$,

$$\Psi(\underline{x}, \underline{\theta}) = \left(\frac{\partial}{\partial\theta} \ln f_{\underline{X}}(\underline{x}|\theta)\right)^2 = \left(\frac{1}{f_{\underline{X}}(\underline{x}|\theta)} \frac{\partial}{\partial\theta} f_{\underline{X}}(\underline{x}|\theta)\right)^2$$

represents the relative rate at which the p.d.f. (or p.m.f.) $f_{\underline{X}}(\underline{x}|\theta)$ changes at $\underline{x}$. The average of this rate is denoted by

$$I(\theta) = E_{\underline{\theta}}\left(\left(\frac{\partial}{\partial\theta} \ln f_{\underline{X}}(\underline{X}|\theta)\right)^2\right), \quad \theta \in \Theta \subseteq \mathbb{R}.$$

The large value of $I(\theta_0)$ indicates that it is easier to distinguish θ_0 from the neighboring values of θ_0 and therefore more accurately θ can be estimated if true $\theta = \theta_0$. The quantity $I(\theta)$, $\theta \in \Theta$, is called the **Fisher's information** that $\underline{X}$ contains about the parameter θ. Note that $I(\theta)$ is a function of $\theta \in \Theta$.

(c) Let $X_1, \ldots, X_n$ be a random sample with common p.d.f./p.m.f. $f(\cdot|\theta)$, $\theta \in \Theta \subseteq \mathbb{R}$, and let $\underline{X} = (X_1, \ldots, X_n)$. Then

$$f_{\underline{X}}(\underline{x}|\theta) = \prod_{i=1}^{n} f(x_i|\theta),\ \theta \in \Theta.$$

Let $i(\theta)$ and $I(\theta)$, respectively, denote the Fisher's information contained in the single observation, say X_1, and the whole sample $\underline{X} = (X_1, \ldots, X_n)$. Then, for $\theta \in \Theta$,

$$\begin{aligned} I(\theta) &= E_{\underline{\theta}}\left(\left(\frac{\partial}{\partial\theta} \ln f_{\underline{X}}(\underline{X}|\theta)\right)^2\right) \\ &= E_{\underline{\theta}}\left(\left(\sum_{i=1}^{n} \frac{\partial}{\partial\theta} \ln f(X_i|\theta)\right)^2\right) \end{aligned}$$

$$
\begin{aligned}
&= E_{\underline{\theta}}\left(\sum_{i=1}^{n}\left(\frac{\partial}{\partial\theta}\ln f(X_i|\theta)\right)^2\right) \\
&\qquad + E_{\underline{\theta}}\left(\sum_{\substack{i=1\\ i\neq j}}^{n}\sum_{j=1}^{n}\frac{\partial}{\partial\theta}\ln f(X_i|\theta)\frac{\partial}{\partial\theta}\ln f(X_j|\theta)\right) \\
&= nE_{\underline{\theta}}\left(\left(\frac{\partial}{\partial\theta}\ln f(X_1|\theta)\right)^2\right) \\
&= ni(\theta),
\end{aligned}
$$

since $X_1, \ldots, X_n$ are i.i.d., and

$$
\begin{aligned}
E_{\underline{\theta}}\left(\frac{\partial}{\partial\theta}\ln f(X_1|\theta)\right) &= \int_{-\infty}^{\infty}\frac{\partial}{\partial\theta}f(x|\theta)dx \\
&= \frac{d}{d\theta}\int_{-\infty}^{\infty}f(x|\theta)dx \\
&= 0, \quad \forall\ \theta\in\Theta.
\end{aligned}
$$

Result 2.

Let $X_1, X_2, \ldots$ be a sequence of i.i.d. one-dimensional r.v.s with common Fisher's information $i(\theta) = E_\theta((\frac{\partial}{\partial\theta} \ln f(X_1|\theta))^2)$, $\theta \in \Theta \subseteq \mathbf{R}$, where $f(\cdot|\theta)$, $\theta \in \Theta$, is the common p.d.f. (or p.m.f.) of the sequence $X_1, X_2, \ldots$, and Θ is an open interval in $\mathbb{R}$. Let $\hat{\theta}_n$ be the unique M.L.E. of θ based on $X_1, \ldots, X_n$. Then, under regularity conditions R_1, as $n \to \infty$,

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} Y \sim N\left(0, \frac{1}{i(\theta)}\right) \quad \text{and} \quad \hat{\theta}_n \xrightarrow{p} \theta.$$

Corollary 1.

Under the regularity conditions R_1, let $\hat{g}_n \equiv \hat{g}_n(\underline{X})$ be the M.L.E. of one-dimensional estimand $g(\theta)$, where $g(\cdot)$ is a differentiable function. Then, under regularity conditions R_1, as $n \to \infty$,

$$\sqrt{n}(\hat{g}_n - g(\theta)) \xrightarrow{d} W \sim N\left(0, \frac{(g'(\theta))^2}{i(\theta)}\right) \text{ and } \hat{g}_n \xrightarrow{p} g(\theta), \quad \theta \in \Theta.$$

Example 5.

Let $X_1, \ldots, X_n$ $(n \geq 2)$ be a random sample from $N(\mu, \sigma^2)$ distribution, where $\underline{\theta} = (\mu, \sigma^2) \in \Theta = (-\infty, \infty) \times (0, \infty)$ is unknown. Show that the maximum likelihood estimator of $\underline{\theta}$ is $(\hat{\mu}, \hat{\sigma}^2) = (\overline{X}, \frac{n-1}{n} S^2)$.

Proof: For a given sample realization $\underline{x} = (x_1, \ldots, x_n)$

$$\begin{aligned} L_{\underline{x}}(\underline{\theta}) &= \prod_{i=1}^{n} \left\{ \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x_i - \mu)^2} \right\} \\ &= (2\pi\sigma^2)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}, \underline{\theta} \in \Theta. \end{aligned}$$

Then

$$\begin{aligned} \ln L_{\underline{x}}(\underline{\theta}) &= -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2, \\ \frac{\partial}{\partial\mu}\ln L_{\underline{x}}(\underline{\theta}) &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu) \end{aligned}$$

$$
\begin{aligned}
\frac{\partial}{\partial \sigma^2} \ln L_{\underline{x}}(\underline{\theta}) &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (x_i - \mu)^2 \\
\frac{\partial}{\partial \mu^2} \ln L_{\underline{x}}(\underline{\theta}) &= -\frac{n}{\sigma^2} \\
\frac{\partial}{\partial (\sigma^2)^2} \ln L_{\underline{x}}(\underline{\theta}) &= \frac{n}{2\sigma^4} - \frac{1}{\sigma^6} \sum_{i=1}^{n} (x_i - \mu)^2 \\
\frac{\partial}{\partial \mu \partial \sigma^2} \ln L_{\underline{x}}(\underline{\theta}) &= -\frac{1}{\sigma^4} \sum_{i=1}^{n} (x_i - \mu).
\end{aligned}
$$

Clearly $\underline{\hat{\theta}} = (\hat{\mu}, \hat{\sigma}^2) = (\overline{X}, \frac{n-1}{n} S^2)$ is the unique critical point. Also

$$
\begin{aligned}
\left[\frac{\partial}{\partial \mu^2} \ln L_{\underline{x}}(\underline{\theta})\right]_{\underline{\theta} = \underline{\hat{\theta}}} &= -\frac{n}{\hat{\sigma}^2} \\
\left[\frac{\partial}{\partial (\sigma^2)^2} \ln L_{\underline{x}}(\underline{\theta})\right]_{\underline{\theta} = \underline{\hat{\theta}}} &= \frac{n}{2\hat{\sigma}^4} - \frac{1}{\hat{\sigma}^6} \sum_{i=1}^{n} (x_i - \overline{x})^2
\end{aligned}
$$

$$\begin{aligned} &= -\frac{n}{2\hat{\sigma}^4} \\ \left[\frac{\partial}{\partial\mu\partial\sigma^2}\ln L_{\underline{x}}(\underline{\theta})\right]_{\underline{\theta}=\underline{\hat{\theta}}} &= 0. \end{aligned}$$

It follows that $\underline{\hat{\theta}} = (\hat{\mu}, \hat{\sigma}^2) = (\overline{X}, \frac{n-1}{n}S^2)$ is the m.l.e. of $\underline{\theta}$.

Example 6.

Let $X_1, \ldots, X_n$ be a random sample from $\text{Bin}(m, \theta)$ distribution, where $\theta \in \Theta = (0, 1)$ is unknown and m is a known positive integer. Show that $\delta_M(\underline{X}) = \overline{X}$ is the M.L.E. of θ.

Solution. For a sample realization $\underline{x} \in \chi = \{0, 1, \ldots, m\}^n$

$$\begin{aligned} L_{\underline{x}}(\theta) &= \prod_{i=1}^{n} \left\{ \binom{m}{x_i} \theta^{x_i} (1-\theta)^{m-x_i} \right\} \\ &= \left(\prod_{i=1}^{n} \binom{m}{x_i} \right) \theta^{\sum_{i=1}^{n} x_i} (1-\theta)^{mn - \sum_{i=1}^{n} x_i}. \end{aligned}$$

First, let $\theta \in (0, 1)$.

$$\ln L_{\underline{x}}(\theta) = \sum_{i=1}^{n} \ln \binom{m}{x_i} + \left(\sum_{i=1}^{n} x_i \right) \ln \theta + \left(mn - \sum_{i=1}^{n} x_i \right) \ln(1-\theta)$$

$$
\begin{aligned}
\frac{\partial}{\partial\theta}\ln L_{\underline{x}}(\theta) &= \frac{\sum_{i=1}^n x_i}{\theta} - \frac{mn - \sum_{i=1}^n x_i}{1-\theta} \\
\frac{\partial}{\partial\theta}\ln L_{\underline{x}}(\theta) > 0 &\Leftrightarrow \theta < \frac{\bar{x}}{m} \\
&\Leftrightarrow \frac{\bar{x}}{m} \text{ is the M.L.E. of } \theta \\
&\Rightarrow \delta_M(\underline{X}) = \frac{\bar{X}}{m} \text{ is the M.L.E. of } \theta.
\end{aligned}
$$

Example 7.

Let $X_1, \ldots, X_n$ be a random sample from $\mathrm{U}(\theta_1, \theta_2)$ distribution, where $\underline{\theta} = (\theta_1, \theta_2) \in \Theta = \{(z_1, z_2) : -\infty < z_1 < z_2 < \infty\}$ is unknown. Show that $\underline{\delta}_M(\underline{X}) = (X_{(1)}, X_{(n)})$ is the M.L.E. of $\underline{\theta}$.

Solution.

$$f_{X_i}(x) = \begin{cases} \frac{1}{\theta_2 - \theta_1}, & \text{if } \theta_1 < x < \theta_2 \\ 0, & \text{otherwise} \end{cases}, \quad i = 1, \ldots, n.$$

Let $\underline{x}$ be the fixed realization. Then

$$\begin{aligned} L_{\underline{x}}(\underline{\theta}) = L_{\underline{x}}(\theta_1, \theta_2) = f_{\underline{X}}(\underline{x}|\underline{\theta}) &= \prod_{i=1}^{n} f_{X_i}(x_i|\underline{\theta}) \\ &= \begin{cases} \frac{1}{(\theta_2 - \theta_1)^n}, & x_{(1)} > \theta_1, x_{(n)} < \theta_2 \\ 0, & \text{otherwise} \end{cases}. \end{aligned}$$

Here $L_{\underline{x}}(\theta)$ is an increasing function of θ_1 and decreasing function of θ_2. Thus $\underline{\delta}_M(\underline{X}) = (X_{(1)}, X_{(n)})$ is the M.L.E. of $\underline{\theta}$.

Example 8. (M.L.E. and M.M.E. may be different)

Let $X \sim \mathrm{U}(0, \theta)$, where $\theta \in \Theta = (0, \infty)$ is unknown. Show that the M.M.E. of θ is $\delta_{\text{MME}}(X) = 2X$, whereas the M.L.E. of θ is $\delta_{\text{MLE}}(X) = X$.

Solution. Since $E_\theta(X) = \frac{\theta}{2}$, it follows that $\delta_{\text{MME}}(X) = 2X$ is the M.M.E. of θ. Also, for a fixed realization $x > 0$,

$$L_x(\theta) = f_X(x|\theta) == \begin{cases} \frac{1}{\theta}, & \text{if } \theta > x \\ 0, & \text{if } 0 < \theta \leq x \end{cases}.$$

Clearly $L_x(\theta)$ is maximized at $\theta = x$. Thus the M.L.E. of θ is $\delta_{\text{MLE}}(X) = X$.

Properties of Estimators

Unbiased Estimators

Suppose that the estimand $g(\underline{\theta})$ is real-valued.

Definition 6.

(a) An estimator $\delta(\underline{X})$ is said to be an **unbiased estimator** of $g(\underline{\theta})$ if $E_{\underline{\theta}}(\delta(\underline{X})) = g(\underline{\theta}), \; \forall \; \underline{\theta} \in \Theta$.

(b) An estimator which is not unbiased for estimating $g(\underline{\theta})$ is called a **biased estimator** of $g(\underline{\theta})$.

(c) The quantity $B_{\underline{\theta}}(\delta) = E_{\underline{\theta}}(\delta(\underline{X})) - g(\underline{\theta})$, $\underline{\theta} \in \Theta$, is called the **bias** of the estimator $\delta(\underline{X})$.

Remark 5.

(a) Note that $B_{\underline{\theta}}(\delta)$ is a function of $\underline{\theta} \in \Theta$.

(b) Note that for an unbiased estimator $\delta(\underline{X})$, $B_{\underline{\theta}}(\delta) = 0$, $\forall\ \underline{\theta} \in \Theta$.

(c) An unbiased estimator, if evaluated a large number of times, on the average equals the true value of the estimand. Thus, the property of unbiasedness is a reasonable property for an estimator to have.

Example 8. (Unbiased estimators may not exist)

. Let $X \sim \text{Bin}(n, \theta)$, where $\theta \in \Theta = (0, 1)$ is unknown and n is a known positive integer. Show that the unbiased estimators for the estimand $g(\theta) = \frac{1}{\theta}$ do not exist.

Solution. On contrary suppose there exists an estimator $\delta(X)$ such that

$$\begin{aligned} E_\theta(\delta(X)) &= \frac{1}{\theta},\ \forall\ \theta \in \Theta \\ \text{i.e., } \sum_{j=0}^{n} \delta(j) \binom{n}{j} \theta^j (1-\theta)^{n-j} &= \frac{1}{\theta},\ \forall\ 0 < \theta < 1 \\ \Rightarrow \theta \sum_{j=0}^{n} \delta(j) \binom{n}{j} \theta^j (1-\theta)^{n-j} &= 1,\ \forall\ 0 < \theta < 1, \end{aligned}$$

which is not possible since, as $\theta \to 0$, L.H.S. $\to 0$, whereas R.H.S. $\to 1$.

Example 9. (Unbiased estimator may be absurd)

Let $X \sim \text{Poisson}(\theta)$, where $\theta \in \Theta = (0, \infty)$ is unknown, and let the estimand be $g(\theta) = e^{-3\theta}$. Show that $\delta(X) = (-2)^X$ is the unique unbiased estimator of $g(\theta)$ (here $\delta(X) = (-2)^X$ takes both positive and negative values, whereas the estimand $g(\theta)$ is always positive).

Solution. An estimator $\delta(X)$ is unbiased for estimating $g(\theta) = e^{-3\theta}$ iff

$$\begin{aligned} E_\theta[\delta(X)] &= g(\theta),\ \forall\ \theta \in \Theta \\ \Leftrightarrow \sum_{j=0}^{\infty} \delta(j) \frac{e^{-\theta}\theta^j}{j!} &= e^{-3\theta},\ \forall\ \theta > 0 \\ \Leftrightarrow \sum_{j=0}^{\infty} \frac{\delta(j)\theta^j}{j!} &= e^{-2\theta},\ \forall\ \theta > 0 \\ \Leftrightarrow \sum_{j=0}^{\infty} \frac{\delta(j)\theta^j}{j!} &= \sum_{j=0}^{\infty} \frac{(-2)^j\theta^j}{j!},\ \forall\ \theta > 0 \end{aligned}$$

The L.H.S. and R.H.S. are power series in θ and they match in an open interval. Thus,

$$\begin{aligned} \frac{\delta(j)}{j!} &= \frac{(-2)^j}{j!},\ j = 0, 1, 2, \ldots \\ \Rightarrow \delta(j) &= (-2)^j,\ j = 0, 1, 2, \ldots \end{aligned}$$

Thus $\delta(X) = (-2)^X$ is the unique unbiased estimator of $g(\theta) = e^{-3\theta}$.

Example 10. (M.M.E. and M.L.E. may not be unbiased)

Let $X \sim \mathrm{U}(0, \theta)$, where $\theta \in \Theta = (0, \infty)$ is unknown, and let the estimand be $g(\theta) = \sqrt{\theta}$. Show that M.M.E. and the M.L.E. of $g(\theta)$ are $\delta_{\text{MME}}(X) = \sqrt{2X}$ and $\delta_{\text{MLE}}(X) = \sqrt{X}$, respectively, and $E_\theta(\delta_{\text{MME}}(X)) = \frac{2\sqrt{2}}{3} g(\theta)$, $\theta \in \Theta$, $E_\theta(\delta_{\text{MLE}}(X)) = \frac{2}{3} g(\theta)$, $\theta \in \Theta$.

Solution. For the sample realization $x > 0$, the likelihood function

$$L_X(\theta) = f_X(x|\theta) = \begin{cases} \frac{1}{\theta}, & \text{if } \theta > x \\ 0, & \text{otherwise,} \end{cases}$$

is minimized at $\theta = \hat{\theta} = x$. Thus the MLE of θ is X and by the invariance property of MLEs, the MLE of $g(\theta) = \sqrt{\theta}$ is

$$\delta_{\text{MLE}}(X) = \sqrt{X}.$$

$$E_\theta[\delta_{\text{MLE}}(X)] = \int_0^\theta \frac{\sqrt{x}}{\theta} dx = \frac{2}{3}\sqrt{\theta} \neq \theta.$$

Also MME of θ is given by (since $E(X) = \frac{\theta}{2}$)

$$\frac{\hat{\theta}_{\text{MME}}}{2} = X \Rightarrow \hat{\theta}_{\text{MME}} = 2X.$$

Thus the MME of $g(\theta) = \sqrt{\theta}$ is

$$\delta_{\text{MME}}(X) = \sqrt{2X}$$

$$\Rightarrow E_\theta[\delta_{\text{MME}}(X)] = \sqrt{2}E_\theta(X) = \frac{2\sqrt{2}}{3}\sqrt{\theta} \neq \theta.$$

Example 11 (Typically, there are many unbiased estimators for a given estimand)

Let $X_1, \ldots, X_n$ be a random sample from a $N(\theta, 1)$ distribution, where $\theta \in \Theta = (-\infty, \infty)$ is unknown, and let the estimand be $g(\theta) = \theta$. Then $\delta_M(\underline{X}) = \overline{X}$, $\delta_i(\underline{X}) = X_i$, $\delta_{i,j} = \frac{X_i + X_j}{2}$, $i, j \in \{1, 2, \ldots, n\}$, $i \neq j$ $\delta_{i,j,k}(\underline{X}) = X_i + X_j - X_k$, $i, j, k \in \{1, \ldots, n\}$, etc., are all unbiased for estimating $g(\theta)$.

As seen in the above example, typically, there are many unbiased estimators for a given estimand. Therefore, it is useful to have some criterion for comparing unbiased estimators. One criterion which is often used is the variance of the unbiased estimator $\delta(\cdot)$ (denoted by $V_{\underline{\theta}}(\delta)$ to emphasize the dependence on $\underline{\theta} \in \Theta$). If $\delta_1 \equiv \delta_1(\underline{X})$ and $\delta_2 \equiv \delta_2(\underline{X})$ are two unbiased estimators of $g(\underline{\theta})$ and if

$$V_{\underline{\theta}}(\delta_1) = E_{\underline{\theta}}((\delta_1(\underline{X}) - g(\underline{\theta}))^2) < V_{\underline{\theta}}(\delta_2) = E_{\underline{\theta}}((\delta_2(\underline{X}) - g(\underline{\theta}))^2), \ \forall \ \underline{\theta} \in \Theta,$$

then $(\delta_1(\underline{X}) - g(\underline{\theta}))^2$ is, on the average, less than $(\delta_2(\underline{X}) - g(\underline{\theta}))^2$, which indicates that δ_1 is nearer to $g(\underline{\theta})$ than δ_2. For this reason we define:

Definition 7.

An unbiased estimator δ_1 is said to be better than the unbiased estimator δ_2 if $V_{\underline{\theta}}(\delta_1) \leq V_{\underline{\theta}}(\delta_2)$, $\forall\ \underline{\theta} \in \Theta$, with strict inequality for at least one $\underline{\theta} \in \Theta$.

Definition 8.

In an estimation problem where the M.L.E. exists, an estimator (not necessarily unbiased) which depends on observation $\underline{X} = (X_1, \ldots, X_n)$ only through the M.L.E. (i.e., an estimator which is a function of the M.L.E. alone) is called an **estimator based on the M.L.E.**.

Under fairly general conditions, it can be shown that the estimators which are not based on the M.L.E. are not desirable, i.e., given any unbiased estimator δ, which is not based on the M.L.E., there exists an unbiased estimator based on the M.L.E., say δ_M, such that δ_M is better than that δ. Thus, to find the best unbiased estimators one should consider only those estimators which are based on the M.L.E. Under fairly general conditions, it can also be shown that there is only one unbiased estimator based on the M.L.E., and that estimator is the best unbiased estimator. Therefore, in finding a sensible unbiased estimator for an estimand $g(\underline{\theta})$, we typically start with the M.L.E. of $g(\underline{\theta})$. If it is unbiased, then we have found the estimator we want. If it is not unbiased, we modify it to make it unbiased.

Example 12. Let $X_1, \ldots, X_n$ be a random sample from a Poisson(θ) distribution, where $\theta \in \Theta = (0, \infty)$ is unknown, and let the estimand be $g(\theta) = P_\theta(X = 0) = e^{-\theta}$. Then the M.L.E. of $g(\theta)$ is $\delta_M(\underline{X}) = e^{-\overline{X}}$ and the unbiased estimator based on the M.L.E. is $\delta_U(\overline{X}) = (1 - \frac{1}{n})^{n\overline{X}}$.

Solution. Let $T = \sum_{i=1}^n X_i$ so that $T \sim$ Poisson$(n\theta)$ and $\overline{X} = \frac{T}{n}$. We want the estimator $\delta(\overline{X}) = \delta(\frac{T}{n})$ such that

$$\begin{aligned} E_\theta(\delta(\overline{X})) &= e^{-\theta},\ \forall\ \theta > 0 \\ \Leftrightarrow\ \sum_{j=0}^{\infty} \delta(\frac{j}{n}) \frac{e^{-n\theta}(n\theta)^j}{j!} &= e^{-\theta},\ \forall\ \theta > 0 \\ \Leftrightarrow\ \sum_{j=0}^{\infty} \delta(\frac{j}{n}) \frac{n^j}{j!} \theta^j &= e^{(n-1)\theta},\ \forall\ \theta > 0 \\ \Leftrightarrow\ \sum_{j=0}^{\infty} \delta(\frac{j}{n}) \frac{n^j}{j!} \theta^j &= \sum_{j=0}^{\infty} \frac{(n-1)^j}{j!} \theta^j,\ \forall\ \theta > 0 \end{aligned}$$

$$\Leftrightarrow \quad \delta(\frac{j}{n}) = (1 - \frac{1}{n})^j, \; j = 0, 1, \ldots .$$

It follows that the unbiased estimator based on the M.L.E. is $\delta_U(\overline{X}) = (1 - \frac{1}{n})^{n\overline{X}}$.

Example 13.

Let $X_1, \ldots, X_n$ be a random sample from a $N(\mu, \sigma^2)$ distribution, where $\underline{\theta} = (\mu, \sigma^2) \in \Theta = \{(z_1, z_2) : -\infty < z_1 < \infty, z_2 > 0\}$ is unknown, and let the estimand be $g(\underline{\theta}) = \sigma^2$. Show that the M.L.E. of (μ, σ^2) is $(\overline{X}, \frac{(n-1)S^2}{n})$ and the unbiased estimator of $g(\underline{\theta})$ based on the M.L.E. is S^2.

Consistent Estimators

Let $g(\underline{\theta})$ be a real valued estimand and let $X_1, X_2, \ldots, X_n$ be a random sample based on which $g(\underline{\theta})$ is to be estimated. We consider the problem of estimating $g(\underline{\theta})$ as n, the number of observations, goes to infinity. Suppose that for each n, we have an estimator $\delta_n \equiv \delta_n(X_1, \ldots, X_n)$ of $g(\underline{\theta})$. For any sensible estimator δ_n we would expect that, as $n \to \infty$, the estimator δ_n would get close to $g(\underline{\theta})$ in some sense. Estimators defined below possess such property.

Definition 9. An estimator $\delta_n(X_1, \ldots, X_n)$, based on sample $X_1, \ldots, X_n$, is said to be a **consistent estimator** of (or consistent for estimating) $g(\underline{\theta})$ if , for each $\underline{\theta} \in \Theta$, $\delta_n(\underline{X}) \xrightarrow{p} g(\underline{\theta})$, as $n \to \infty$.

Remark 5.

(a) An estimator $\delta_n(\underline{X})$ is consistent for estimating $g(\underline{\theta})$ if and only if, for every $\underline{\theta} \in \Theta$,

$$\lim_{n\to\infty} P_{\underline{\theta}}(|\delta_n(\underline{X}) - g(\underline{\theta})| > \epsilon) = 0, \ \forall \ \epsilon > 0,$$

i.e., as n goes to infinity, the estimator $\delta_n(\underline{X})$ would get close to the estimand $g(\underline{\theta})$.

(b) Let $\Theta \subseteq \mathbb{R}$ and suppose that the regularity conditions R_1 are satisfied. Then, by Corollary 1, the M.L.E of any real-valued estimand $g(\underline{\theta})$ is consistent.

Remark 5 continued

(c) Consider the method of moments for estimating the estimand $\underline{\theta} = (\theta_1, \ldots, \theta_p)$. Let $A_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$, $k = 1, \ldots, p$, and let $m_k = E_{\underline{\theta}}(X_1^k) = h_k(\underline{\theta})$, $k = 1, \ldots, p$, say. By WLLN, $A_k \xrightarrow{P} m_k = h_k(\underline{\theta})$, $k = 1, \ldots, p$. If $(m_1, \ldots, m_p) = (h_1(\underline{\theta}), \ldots, h_p(\underline{\theta}))$ is one-to-one function of $\underline{\theta}$ and if the inverse functions $\theta_i = g_i(m_1, \ldots, m_p)$, $i = 1, \ldots, p$, are continuous in $m_1, \ldots, m_p$, then, as $n \to \infty$, $\hat{\theta}_i = g_i(A_1, \ldots, A_p) \xrightarrow{P} g_i(m_1, \ldots, m_p) = \theta_i$, $i = 1, \ldots, p$, so that $\hat{\theta}_i = g_i(A_1, \ldots, A_p)$ $(i = 1, \ldots, p)$ are consistent estimators of θ_i.

(d) If $\delta_n(\underline{X})$ is consistent for estimating $g(\underline{\theta})$ and if $\{a_n\}_{n\geq 1}$ and $\{b_n\}_{n\geq 1}$ are sequences of real numbers such that $a_n \to 1$ and $b_n \to 0$, as $n \to \infty$, then the estimator $T_n(\underline{X}) = a_n\delta_n(\underline{X}) + b_n$ is also consistent for estimating $g(\underline{\theta})$. Thus for an estimand, typically, many consistent estimators exist. Also it follows that a consistent estimator may not be unbiased.

Theorem 1.

(a) If, for each $\underline{\theta} \in \Theta$, $B_{\underline{\theta}}(\delta_n)$ and $V_{\underline{\theta}}(\delta_n)$ go to zero, as $n \to \infty$, then δ_n is consistent for estimating $g(\underline{\theta})$.

(b) If δ_n is consistent for estimating $g(\underline{\theta})$ and $h(t)$ is a real-valued continuous function, then $h(\delta_n)$ is consistent for estimating $h(g(\underline{\theta}))$.

Proof. (a) We have

$$\begin{aligned} E_\theta(\delta_n(\underline{X}) - g(\underline{\theta}))^2 &= E_\theta[(\delta_n(\underline{X}) - E_\theta(\delta_n(\underline{X})) + E_\theta(\delta_n(\underline{X})) - g(\underline{\theta}))^2] \\ &= V_\theta(\delta_n) + (B_\theta(\delta_n))^2 \to 0, \text{as } n \to \infty. \end{aligned}$$

Thus,

$$0 \leq P_{\underline{\theta}}(|\delta_n(\underline{X}) - g(\underline{\theta})| > \epsilon) \leq \frac{E_\theta(\delta_n(\underline{X}) - g(\underline{\theta}))^2}{\epsilon^2} \to 0, \text{as } n \to \infty, \forall \epsilon > 0.$$

$$\Rightarrow \lim_{n\to\infty} P_{\underline{\theta}}(|\delta_n(\underline{X}) - g(\underline{\theta})| > \epsilon) = 0, \forall \epsilon > 0$$

$$\Rightarrow \delta_n(\underline{X}) \xrightarrow{p} g(\underline{\theta}).$$

(b) Follows using the result done before.

Example 14. Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution having p.d.f. $f(x|\theta) = e^{-(x-\theta)}$, if $x \geq \theta$ and $= 0$, otherwise, where $\theta \in \Theta = (-\infty, \infty)$ is unknown. The M.L.E. of θ is $\delta_M(\underline{X}) = X_{(1)}$ and the unbiased estimator based on the M.L.E. is $\delta_U(\underline{X}) = X_{(1)} - \frac{1}{n}$. Both of these estimators are consistent for estimating θ.

Example 15. Let $X_1, X_2, \ldots, X_n$ be a random sample from the Cauchy distribution having p.d.f. $f(x|\theta) = \frac{1}{\pi} \cdot \frac{1}{1+(x-\theta)^2}$, $-\infty < x < \infty$, where $\theta \in \Theta = (-\infty, \infty)$ is unknown. Then $\delta_n(\underline{X}) = \overline{X}$ is neither unbiased nor consistent for estimating $g(\underline{\theta})$.

Criteria for Comparing Estimators

We discussed how to find the best unbiased estimator. Often estimators with some bias may be preferred over the unbiased estimators provided these estimators have some desirable properties which are not possessed by the unbiased estimators. Thus, it is useful to have a criterion for comparing estimators that are not necessarily unbiased. One such criterion is the mean squared error (m.s.e.), defined below.

Definition 10. (a) The mean squared error (m.s.e.) of an estimator $\delta(\underline{X})$ (possibly biased) of $g(\underline{\theta})$ is defined by

$$M_{\underline{\theta}}(\delta) = E_{\underline{\theta}}\left((\delta(\underline{X}) - g(\underline{\theta}))^2\right) = V_{\underline{\theta}}(\delta) + (B_{\underline{\theta}}(\delta))^2, \quad \underline{\theta} \in \Theta,$$

where $V_{\underline{\theta}}(\delta)$ is the variance of $\delta(\underline{X})$ and $B_{\underline{\theta}}(\delta)$ is the bias of $\delta(\underline{X})$.

(b) For estimating $g(\underline{\theta})$, we say that the estimator $\delta_1(\underline{X})$ is better than the estimator $\delta_2(\underline{X})$, under the m.s.e. criterion, if $M_{\underline{\theta}}(\delta_1) \leq M_{\underline{\theta}}(\delta_2)$, $\forall\ \underline{\theta} \in \Theta$, with strict inequality for at least one $\underline{\theta} \in \Theta$.

Under fairly general conditions, it can be shown that the estimators (possibly biased) which are not based on the M.L.E. are not sensible, i.e., given any estimator δ, which is not based on the M.L.E., there exists an unbiased estimator based on the M.L.E., say δ^*, such that δ^* has smaller m.s.e. than δ, for each parametric configuration. Thus, for finding a sensible estimator (not necessarily unbiased) of a real-valued estimand $g(\underline{\theta})$, we typically start with the M.L.E. of $g(\underline{\theta})$ and then consider an appropriate class, say $\mathcal{D}$, of estimators based on the M.L.E., of which M.L.E. is a particular member. This choice of class $\mathcal{D}$ is generally based on intuitive considerations. We then try to find the estimator having the smallest m.s.e. (if such an estimator exist) in this class $\mathcal{D}$ of estimators.

Example 16. Let $X_1, \ldots, X_n$ $(n \geq 2)$ be a random sample from $N(\mu, \sigma^2)$ distribution, where $\underline{\theta} = (\mu, \sigma^2) = \{(z_1, z_2) : -\infty < z_1 < \infty, z_2 > 0\}$ is unknown. Let $(\hat{\mu}, \hat{\sigma}^2)$ be the M.L.E. of (μ, σ^2). Then

(a) M.L.E. $\hat{\sigma}^2$ is not unbiased for estimating σ^2;

(b) The unbiased estimator of σ^2 based on the M.L.E. is
$\delta_U(\underline{X}) = \frac{n}{n-1}\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$;

(c) Among the estimators in the class $\mathcal{D} = \{\delta_c(\underline{X}) : \delta_c(\underline{X}) = c\hat{\sigma}^2\}$, the estimator $\delta_{c_0}(\underline{X}) = c_0\hat{\sigma}^2 = \frac{1}{n+1}\sum_{i=1}^{n}(X_i - \overline{X})^2$, where $c_0 = \frac{n}{n+1}$, has the smallest m.s.e., for each parametric configuration.

Thank you for your patience