$$a_n\left\{f\left(\hat{\theta}_n\right)-f\left(\theta\right)\right\}\xrightarrow{D} f'\left(\theta\right)X$$

2 ESTIMATION

2.1 STATISTICAL MODELS

Statistical inference starts by specifying the underlying statistical model, which consists of:

- A random vector $\mathbf{X}=(X_1,...,X_n)\in\chi$ which is observed;
- An unknown parameter vector $\boldsymbol{\theta}=(\theta_1,...,\theta_k)\in\Theta$;
- A function $f_{\mathbf{X}}(\mathbf{x};\boldsymbol{\theta})$ (or $p_{\mathbf{X}}(\mathbf{x};\boldsymbol{\theta})$) which represents the p.d.f. (or p.m.f.) of $\mathbf{X}$ for each $\boldsymbol{\theta}$.

χ is called the support (or sample space) and Θ is called the parameter space. Note that $\mathbf{X}$ is a sample measure and as such is a r.v. whereas $\boldsymbol{\theta}$ is a population measure and as such is a constant.

Any function $T=T(X_1,...,X_n)$ is called a statistic (note that *T* is also a r.v.). Note that *T* must not involve any unknown parameter. When used in the context of providing a numerical value for a parameter, a statistic is called an estimator.

One of the major aims of statistical inference is to use the observed values of suitable *T* to make conclusions about the unknown $\boldsymbol{\theta}$.

Example 2.1A

Consider the following statistical model: suppose an observation is made on each of $X_1,...,X_{10}$, where each $X_i\overset{iid}{\sim}N(\mu,\sigma^2)$, μ and σ being unknown parameters. Then $\mathbf{X}=(X_1,...,X_{10})$, $\boldsymbol{\theta}=(\mu,\sigma)$, and

$$f_{\mathbf{X}}(\mathbf{x};\boldsymbol{\theta}) = \frac{1}{\sigma^{10}(2\pi)^5}\exp\left\{-\sum_{i=1}^{10}\frac{(x_i-\mu)^2}{2\sigma^2}\right\}.$$

The support for this model is R^{10} and the parameter space is the half-plane

$$\Theta = \{(\mu,\sigma): -\infty < \mu < \infty, 0 < \sigma < \infty\}.$$

For a random sample $X_1,...,X_n$, examples of statistics include $T_1 = \bar{X}$ and $T_2 = \{X_{(1)} + X_{(n)}\}/2$ (where $X_{(i)}$ is the ith order statistic). Both T_1 and T_2 may be used as estimators of the population mean. On the other hand, although $\sqrt{n}(\bar{X}-\mu)/\sigma$ is a random variable, it is neither a statistic nor an estimator. When the values of μ and σ are known, then $\sqrt{n}(\bar{X}-\mu)/\sigma$ becomes a statistic.

2.2 METHOD OF MOMENTS ESTIMATION

Consider a random sample $X_1, X_2,..., X_n$, where each $X_i \sim f_{X_i}(x;\boldsymbol{\theta})$ [or p.m.f. $p_{X_i}(x;\boldsymbol{\theta})$]. Then, from the sample, the rth *sample* moment is defined by

$$m_r = \sum_{i=1}^{n}\frac{X_i^r}{n}, \qquad r = 1, 2, ...$$

On the other hand, the rth *population* (uncentered) moment is given by

$$\mu_r^{/} = \mathrm{E}X^r = \int_{-\infty}^{\infty} x^r f_X(x;\boldsymbol{\theta})dx$$

The method of moments (MoM) for estimating $\boldsymbol{\theta} = (\theta_1,...,\theta_k)$ proceeds by setting

$$m_r = \mu_r^{/}, \qquad r = 1, 2, ... \tag{2-1}$$

and by taking as many equations as is necessary to estimate $\boldsymbol{\theta}$. The justification of the method is that

$$\mathrm{E}m^r = \mu_r^{/}.$$

Example 2.2A

Consider a random sample $X_1, X_2,..., X_n$ where each X_i has density

$$f_X(x;\theta)=\frac{1}{\theta}e^{-x/\theta}I(x\geq 0)$$

Obtain the MOM estimator of θ.

Solution.

We have

$$\begin{aligned}
\mu_1' = \mathrm{E}X &= \frac{1}{\theta}\int_0^\infty xe^{-x/\theta}dx \\
&= \frac{1}{\theta}\left\{\left[x.-\theta e^{-x/\theta}\right]_0^\infty + \int_0^\infty \theta e^{-x/\theta}dx\right\} \\
&= \frac{1}{\theta}\left\{\left[x.-\theta e^{-x/\theta}\right]_0^\infty + \left[-\theta^2 e^{-x/\theta}\right]_0^\infty\right\} \\
&= \frac{1}{\theta}\left\{(0-0)-\theta^2(0-1)\right\} \\
&= \theta
\end{aligned}$$

By setting $\mu_1' = m_1$, we obtain

$$\theta_{MoM} = \frac{1}{n}\sum_{i=1}^{n}X_i = \overline{X}$$

Example 2.2B

Consider a random sample $X_1, X_2, ..., X_n$ where each X_i has density

$$f_X(x;\alpha,\beta)=\frac{x^{\alpha-1}e^{-x/\beta}}{\beta^\alpha\Gamma(\alpha)}I(x\geq 0)$$

Obtain the MOM estimator of α and β.

Solution.

We have

$$\mu_1' = \mathrm{E}X = \int_0^\infty x.\frac{x^{\alpha-1}e^{-x/\beta}}{\beta^\alpha\Gamma(\alpha)}dx = \alpha\beta\int_0^\infty \frac{x^\alpha e^{-x/\beta}}{\beta^{\alpha+1}\Gamma(\alpha+1)}dx = \alpha\beta$$

$$\mu_2' = \mathrm{E}X^2 = \int_0^\infty x^2.\frac{x^{\alpha-1}e^{-x/\beta}}{\beta^\alpha\Gamma(\alpha)}dx = \alpha(\alpha+1)\beta^2\int_0^\infty \frac{x^{\alpha+1}e^{-x/\beta}}{\beta^{\alpha+2}\Gamma(\alpha+2)}dx = \alpha(\alpha+1)\beta^2$$

We set

$$\hat{\alpha}\hat{\beta} = m_1 = \overline{X}$$

$$\hat{\alpha}(\hat{\alpha}+1)\hat{\beta}^2 = m_2 = \overline{X^2}$$

Therefore,

$$\hat{\alpha}^2\hat{\beta}^2 = \overline{X}^2$$

$$\hat{\alpha}^2\hat{\beta}^2 + \hat{\alpha}\hat{\beta}^2 = \overline{X^2}$$

Subtracting the first from the second equation above,

$$\hat{\alpha}\hat{\beta}^2 = \overline{X^2} - \overline{X}^2$$

Using the above and $\hat{\alpha}^2\hat{\beta}^2 = \overline{X}^2$, we have by division

$$\hat{\alpha} = \frac{\overline{X}^2}{\overline{X^2} - \overline{X}^2}.$$

Using the above and $\hat{\alpha}\hat{\beta} = \overline{X}$, we have

$$\hat{\beta} = \frac{\overline{X^2} - \overline{X}^2}{\overline{X}}.$$

2.3 LIKELIHOOD

The concept of likelihood leads to a powerful estimation method. Suppose $\mathbf{X} = (X_1, X_2, ..., X_n)$ is a vector r.v. with p.d.f. $f_{\mathbf{X}}(\mathbf{x};\boldsymbol{\theta})$ (or p.m.f. $p_{\mathbf{X}}(\mathbf{x};\boldsymbol{\theta})$) . Then the likelihood function is defined by

$$L_{\mathbf{X}}(\boldsymbol{\theta}) = \begin{cases} f_{\mathbf{X}}(\mathbf{X};\boldsymbol{\theta}) & \text{if } X \text{ is continuous} \\ p_{\mathbf{X}}(\mathbf{X};\boldsymbol{\theta}) & \text{if } X \text{ is discrete} \end{cases} \tag{2-2}$$

The likelihood is thus numerically equal to the joint density function (or joint mass function) and is a function of the parameters.

Suppose $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$ are two possible values of $\boldsymbol{\theta}$. If $L_{\mathbf{X}}(\boldsymbol{\theta}_0) > L_{\mathbf{X}}(\boldsymbol{\theta}_1)$, then $\boldsymbol{\theta}_0$ is said to be more likely than $\boldsymbol{\theta}_1$ (in the sense that the observed sample is more likely to have arisen under $\boldsymbol{\theta}_0$ than under $\boldsymbol{\theta}_1$).

Example 2.3A

(a) Given that $X \sim \text{binomial}(n, p)$, the p.m.f. of X is

$$p_X(x;p) = \binom{n}{x} p^x (1-p)^{n-x}\, I\left(x \in \{0, 1, ..., n\}\right)$$

The likelihood function is then

$$L_X(p) = \binom{n}{X} p^X (1-p)^{n-X}\, I\left(X \in \{0,1,...,n\}\right), \qquad 0 \le p \le 1$$

(b) Given that $\mathbf{X} = (X_1, ..., X_n)$, where the X_i's are i.i.d. $N(\mu, \sigma^2)$, the p.d.f. of $\mathbf{X}$ is

$$f_{\mathbf{X}}(x_1 ..., x_n; \mu, \sigma) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left\{-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}\right\}, \qquad -\infty < x_1, ..., x_n < \infty .$$

The likelihood function is then

$$L_{\mathbf{X}}(\mu, \sigma) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left\{-\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{2\sigma^2}\right\}, \qquad -\infty < \mu < \infty, 0 < \sigma < \infty .$$

An extremely useful method of finding estimators is thorough the method of maximum likelihood. $\hat{\boldsymbol{\theta}}$ is a maximum likelihood estimator (MLE) of $\boldsymbol{\theta}$ if

$$L_{\mathbf{X}}\left(\hat{\boldsymbol{\theta}}\right) \geq L_{\mathbf{X}}\left(\boldsymbol{\theta}\right) \quad \text{for all} \quad \boldsymbol{\theta} \in \Theta \tag{2-3}$$

An important result, known as the invariance principle, is as follows. Suppose the MLE of $\boldsymbol{\theta}$ is $\hat{\boldsymbol{\theta}}$. If we wish to estimate some function (not necessarily one-to-one) $\tau(\boldsymbol{\theta})$ of $\boldsymbol{\theta}$, then the MLE of $\tau(\boldsymbol{\theta})$ is $\tau(\hat{\boldsymbol{\theta}})$.

Example 2.3B

(a) Given that $X \sim \text{binomial}(n, p)$, find the MLE of p.

(b) Given that $\mathbf{X} = (X_1, ..., X_n)$, where the X_i's are i.i.d. $N(\mu, \sigma^2)$, find the MLE of μ and σ^2.

Solution.

(a)

$$L_X\left(p\right) = \binom{n}{X} p^X \left(1-p\right)^{n-X} I\left(X \in \{0,1,...,n\}\right) \quad \text{for} \quad 0 \leq p \leq 1.$$

Taking logarithms on both sides,

$$\log L_X\left(p\right) = \log\binom{n}{X} + X \log p + \left(n - X\right)\log\left(1-p\right) + \log I\left(X \in \{0,1,...,n\}\right)$$

$$\frac{\partial}{\partial p}\log L_X\left(p\right) = \frac{X}{p} - \frac{n-X}{1-p}$$

At a maximum,

$$\frac{\partial}{\partial p}\log L_X\left(p\right) = 0 \quad \Rightarrow \quad X - pX = np - pX \quad \Rightarrow \quad \hat{p} = \frac{X}{n}.$$

[It can further be shown that $L_X^{//}(\hat{p}) < 0$, so that $\hat{p} = X/n$ indeed *maximizes* $L_X(p)$].

(b)

$$L_{\mathrm{X}}(\mu,\sigma)=\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left\{-\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{2\sigma^2}\right\}, \qquad -\infty<\mu<\infty,\ 0<\sigma<\infty$$

Therefore,

$$\log L_{\mathrm{X}}(\mu,\sigma)=-n\log\sigma-\frac{n}{2}\log(2\pi)-\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{2\sigma^2}$$

$$\Rightarrow\begin{cases}\dfrac{\partial}{\partial\mu}\log L_{\mathrm{X}}(\mu,\sigma)=\dfrac{1}{\sigma^2}\displaystyle\sum_{i=1}^{n}(X_i-\mu)=\dfrac{1}{\sigma^2}\left(\displaystyle\sum_{i=1}^{n}X_i-n\mu\right)=0\\ \dfrac{\partial}{\partial\sigma}\log L_{\mathrm{X}}(\mu,\sigma)=-\dfrac{n}{\sigma}+\dfrac{1}{\sigma^3}\displaystyle\sum_{i=1}^{n}(X_i-\mu)^2=0\end{cases}$$

From the first equation, the MLE of μ

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

Substituting for μ in the second equation,

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

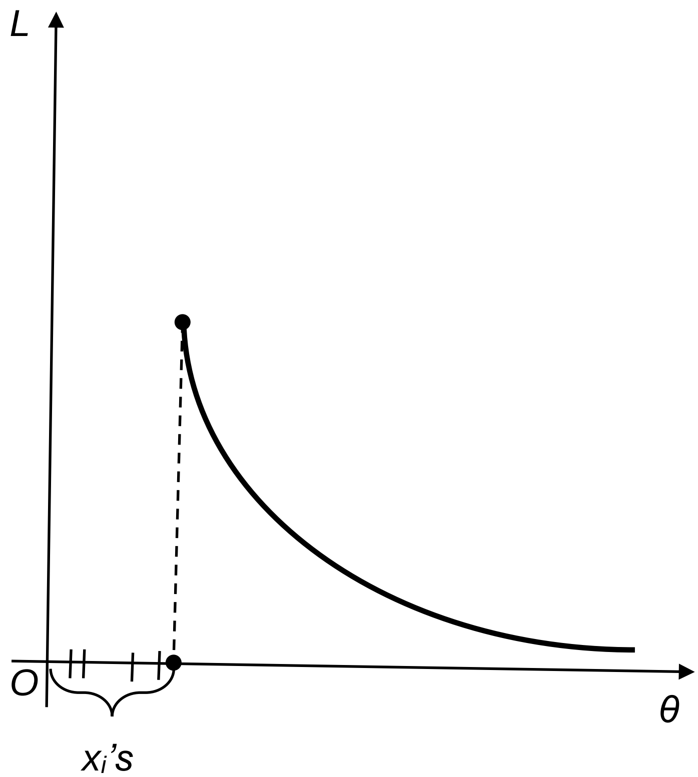
By the invariance principle, the MLE of σ^2 is

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

Remarks: (i) When finding the MLE, differentiating may not always be the best approach. For example, let $X_1,\dots,X_n$ be i.i.d. r.v.'s with each $X_i\sim$ uniform$(0,\theta)$, where $\theta>0$. Then $f(x_i;\theta)=1/\theta$ for $0<x_i<\theta$, i.e. $f(x_i;\theta)=I(0<x_i<\theta)/\theta$ and

$$L_{\mathrm{X}}(\theta)=\frac{1}{\theta^n}I\left(0<X_1,\dots,X_n<\theta\right).$$

Differentiating $L_{\mathbf{x}}(\theta)$ does not work work. It is also *wrong* to argue that $L_{\mathbf{x}}(\theta)$ is maximized when $\theta = 0$ and that the MLE should therefore be $\hat{\theta} = X_{(n)}$. This is because θ cannot take the value zero since $0 < X_1, ..., X_n < \theta$.



A better approach is to graph $L_{\mathbf{x}}(\theta)$ as a function of θ. It is seen that $L_{\mathbf{x}}(\theta)$ is maximum when θ is minimum. Since $\theta > X_{(n)}$, the minimum value of θ is $X_{(n)}$. Hence the MLE of θ is $\hat{\theta} = X_{(n)}$.

(ii) Although an MLE always exists, it may not be unique. For example, let $X_1,...,X_n$ be i.i.d. r.v.'s with each X_i ~ uniform$(\theta-1/2,\theta+1/2)$. Then $f(x_i;\theta)=1$ for $\theta-1/2<x<\theta+1/2$, i.e. $f(x_i;\theta)=I(\theta-1/2<x_i<\theta+1/2)$ and

$$L_{\mathbf{X}}(\theta)=I\left(\theta-\frac{1}{2}<X_1,...,X_n<\theta+\frac{1}{2}\right)=I\left(X_{(1)}>\theta-\frac{1}{2}\right)I\left(X_{(n)}<\theta+\frac{1}{2}\right).$$

It is seen that $L_{\mathbf{X}}(\theta)$ is maximized when $X_{(n)}-1/2<\hat{\theta}<X_{(1)}+1/2$. Thus *any* $\hat{\theta}$ satisfying this inequality is an MLE and there are infinitely many of them.

2.4 PROPERTIES OF ESTIMATIORS (1): UNBIASEDNESS

Suppose $\tau(\boldsymbol{\theta})$ is a function of some parameter θ and let $T=T(\mathbf{X})$ be an estimator of $\tau(\boldsymbol{\theta})$. Then the bias of *T* is defined by

$$\text{bias}(T)=\text{E}_{\boldsymbol{\theta}}T-\tau(\boldsymbol{\theta}). \tag{2-4}$$

In the above, the expectation E is written with a subscript $\boldsymbol{\theta}$ to indicate its dependence on $\boldsymbol{\theta}$. If $\text{bias}(T)=0$, then *T* is said to be unbiased. Otherwise, it is biased. The lower the bias the *more accurate* the estimator is.

Example 2.4A

Use the results in (1-12) to deduce the biases of

(a) $S^2=\dfrac{\sum_i\left(X_i-\bar{X}\right)^2}{n-1}$,

(b) $\hat{\sigma}^2=\dfrac{\sum_i\left(X_i-\bar{X}\right)^2}{n}$ (the MLE of σ^2)

as estimators of σ^2

Solution

(a) $\text{bias}(S^2)=\text{E}_{\boldsymbol{\theta}}S^2-\sigma^2=\sigma^2-\sigma^2=0$, so that S^2 is unbiased for σ^2

(b) $\quad \text{bias}(\hat{\sigma}^2) = \mathrm{E}_{\boldsymbol{\theta}}\hat{\sigma}^2 - \sigma^2 = \left(\frac{n-1}{n}\right)\sigma^2 - \sigma^2 = \frac{-\sigma^2}{n}$.

Remarks. (a) For all its utility, the method of ML does not always lead to unbiased estimators, as Example 2.4A shows. However, as the example also shows, it is sometimes possible to modify a biased MLE to obtain an unbiased estimator: thus if we multiple the MLE of σ^2 by $n/(n-1)$, we obtain S^2 which is unbiased.

(b) The unbiasedness property is not invariant under transformations. For example, in Sec. 1.4, we saw that S^2 is unbiased for σ^2 but S is still biased for σ.

In general, there several unbiased estimators of a given parameter. For example, if $X_1,...,X_{10}$ are i.i.d. with $\mathrm{E}X_i = \mu$, all of the following (among infinitely many) are unbiased estimators of μ:

$$U_1 = X_1, \quad U_2 = \frac{X_1 + X_2}{2}, \quad U_3 = \frac{X_1 + 2X_2}{3}, \quad U_4 = \frac{X_1 + ... + X_{10}}{10}.$$

The question is, which one to prefer? In general, if T_1 and T_2 are two unbiased estimators of $\tau(\boldsymbol{\theta})$, and

if $\text{var}_\theta T_1 < \text{var}_\theta T_2$, then T_1 is better than T_2.

The lower the variance the *more precise* the estimator is.

If T_1 is unbiased and $\text{var}_\theta T_1 \le \text{var}_\theta T$ for all θ, where T is any other unbiased estimator, then T_1 is the uniform best unbiased estimator or uniform minimum variance unbiased estimator (UMVUE).

Example 2.4B

If $X_1,...,X_{10}$ are i.i.d. with $\mathrm{E}X_i = \mu$ and $\text{var}\, X_i = \sigma^2$, which of the following unbiased estimators of μ is best:

$$U_1 = X_1, \quad U_2 = \frac{X_1 + X_2}{2}, \quad U_3 = \frac{X_1 + 2X_2}{3}, \quad U_4 = \frac{X_1 + \ldots + X_{10}}{10}.$$

Solution.

$$\begin{aligned}
\operatorname{var} U_1 &= \operatorname{var} X_1 = \sigma^2, \\
\operatorname{var} U_2 &= \operatorname{var} \frac{X_1 + X_2}{2} = \frac{1}{4}\left(\sigma^2 + \sigma^2\right) = \frac{1}{2}\sigma^2, \\
\operatorname{var} U_3 &= \operatorname{var} \frac{X_1 + 2X_2}{3} = \frac{1}{9}\left(\sigma^2 + 4\sigma^2\right) = \frac{5}{9}\sigma^2, \\
\operatorname{var} U_4 &= \operatorname{var} \frac{X_1 + \ldots X_{10}}{10} = \frac{1}{100}\left(\sigma^2 + \ldots. + \sigma^2\right) = \frac{\sigma^2}{10}.
\end{aligned}$$

Since $\operatorname{var} U_4$ is smallest, U_4 is the best estimator out of the other ones.

Theorem 2.4A

UMVUE's are unique in the sense that if T_1 and T_2 are two UMVUE's then $\Pr\{T_2 = T_1\} = 1$ (i.e. $T_2 = T_1$ almost surely (a.s.))

Proof

Let $\mathrm{E}T_1 = \mathrm{E}T_2 = \theta$ and $\operatorname{var} T_1 = \operatorname{var} T_2 = \sigma^2$. Consider a new unbiased estimator $T = (T_1 + T_2)/2$. We have

$$\begin{aligned}
\operatorname{var} T &= \frac{1}{4}\operatorname{var}\left(T_1 + T_2\right) \\
&= \frac{1}{4}\left\{\operatorname{var} T_1 + \operatorname{var} T_2 + 2\operatorname{cov}\left(T_1, T_2\right)\right\} \\
&= \frac{1}{4}\left(\sigma^2 + \sigma^2 + 2\rho\sqrt{\operatorname{var} T_1 \operatorname{var} T_2}\right) \qquad \text{[where } \rho = corr\left(T_1, T_2\right)\text{]} \\
&= \frac{\sigma^2}{2}\left(1 + \rho\right)
\end{aligned}$$

Since T_1 is an UMVUE,

$$\sigma^2 \le \frac{\sigma^2}{2}\left(1 + \rho\right) \quad \Rightarrow \quad \rho \ge 1$$

But $|\rho| \le 1$, therefore $\rho = 1$, so that $T_2 = c_1 T_1 + c_2$, where c_1 and c_2 are constants. Since $\mathrm{E}T_1 = \mathrm{E}T_2$ we have $c_1 = 1$ and $c_2 = 0$. Hence $T_2 = T_1$ a.s. ■

How can we verify if an estimator is an UMVUE? Later we will show how this can be done in some cases.

When estimators are biased, a criterion that can be used to choose between estimators is the mean squared error (MSE) of an estimator *T*, where

$$\begin{aligned} MSE(T) &= \mathrm{E}\{T - \tau(\boldsymbol{\theta})\}^2 \\ &= \mathrm{E}\{T - \mathrm{E}T + \mathrm{E}T - \tau(\boldsymbol{\theta})\}^2 \\ &= \mathrm{E}(T - \mathrm{E}T)^2 + \{\mathrm{E}T - \tau(\boldsymbol{\theta})\}^2 + 2\mathrm{E}(T - \mathrm{E}T)\{\mathrm{E}T - \tau(\boldsymbol{\theta})\} \\ &= \mathrm{var}\, T + \mathrm{bias}^2 T + 2\{\mathrm{E}T - \tau(\boldsymbol{\theta})\}\underbrace{\mathrm{E}(T - \mathrm{E}T)}_{0} \end{aligned}$$

$$\therefore \quad MSE_{\boldsymbol{\theta}}(T) = \mathrm{var}_{\boldsymbol{\theta}}\, T + \mathrm{bias}^2_{\boldsymbol{\theta}} T \tag{2-5}$$

If T_1 and T_2 are two estimators of $\tau(\theta)$, and

if $MSE_{\boldsymbol{\theta}}(T_1) < MSE_{\boldsymbol{\theta}}(T_2)$, then T_1 is better than T_2.

2.5 PROPERTIES OF ESTIMATIORS (2): CONSISTENCY

Unbiasedness is a finite-sample property. In contrast, consistency is a large-sample property. Consider a parameter $\tau(\theta)$ which is estimated by T_n. We would like T_n to get closer to $\tau(\theta)$ as *n* becomes larger. This property is called consistency and is defined as follows: T_n is a consistent estimator of $\tau(\theta)$ if T_n converges in probability to $\tau(\theta)$, i.e. for any $\varepsilon > 0$,

$$\Pr\{|T_n - \tau(\theta)| < \varepsilon\} \to 1 \quad \text{as} \quad n \to \infty \tag{2-6}$$

Theorem 2.5A

(a) If T_n is consistent for $\tau(\theta)$ and *h* is a continuous function of *T*, then $h(T_n)$ is consistent for $h(\tau(\theta))$.

(b) If $\text{bias}(T_n) \to 0$ and $\text{var}\, T_n \to 0$ then T_n is a consistent estimator. ■

The sufficient conditions in Theorem 2.4b are useful in establishing consistency.

Example 2.5A

Consider the two estimators of σ^2 for a random sample $X_1, X_2, ..., \mathrm{X}_n$ of i.i.d. $N(\mu,\, \sigma^2)$ r.v.s

(i) $S^2 = \dfrac{\sum_i \left(X_i - \bar{X}\right)^2}{n-1}$,

(ii) $\hat{\sigma}^2 = \dfrac{\sum_i \left(X_i - \bar{X}\right)^2}{n}$.

Show that both estimators are consistent.

Solution

(i) We have, from Eq. (1-10), $(n-1)S^2 / \sigma^2 \sim \chi^2_{n-1}$ so that

$$\mathrm{E}\frac{(n-1)S^2}{\sigma^2} = n-1 \quad \Rightarrow \quad \mathrm{E}S^2 = \sigma^2 \quad \Rightarrow \quad \text{bias}\left(S^2\right) = 0$$

$$\text{var}\frac{(n-1)S^2}{\sigma^2} = 2n-2 \quad \Rightarrow \quad \text{var}\, S^2 = \frac{2\sigma^4}{(n-1)} \to 0$$

Hence S^2 is consistent.

(ii)We have $\hat{\sigma}^2 = \left(\dfrac{n-1}{n}\right) S^2$. Therefore,

$$\mathrm{E}\hat{\sigma}^2 = \left(\frac{n-1}{n}\right)\sigma^2 \quad \Rightarrow \quad \text{bias}\left(\hat{\sigma}^2\right) = \left(\frac{n-1}{n}\right)\sigma^2 - \sigma^2 = \frac{-\sigma^2}{n} \to 0$$

$$\text{var}\,\hat{\sigma}^2 = \left(\frac{n-1}{n}\right)^2 . \frac{2\sigma^4}{n-1} \to 0$$

Hence $\hat{\sigma}^2$ is consistent.

<u>Proof Theorem 2.5A</u>

(a) We have, by the definition of continuity of $h(T_n)$

$$\left|T_n - \tau(\theta)\right| < \delta \quad \Rightarrow \quad \left|h(T_n) - h(\tau(\theta))\right| < \varepsilon$$

for arbitrarily small $\delta, \varepsilon > 0$.

Therefore,

$$\Pr\left\{\left|h(T_n) - h(\tau(\theta))\right| < \varepsilon\right\} \geq \Pr\left\{\left|T_n - \tau(\theta)\right| < \delta\right\}$$

Since T_n is consistent and probabilities are less or equal to unity,

$$1 \geq \Pr\left\{\left|h(T_n) - h(\tau(\theta))\right| < \varepsilon\right\} \geq \Pr\left\{\left|T_n - \tau(\theta)\right| < \delta\right\} \to 1$$

Hence $\Pr\left\{\left|h(T_n) - h(\tau(\theta))\right| < \varepsilon\right\} \to 1$ and $h(T_n)$ is consistent.

(b) We have

$$\left|T_n - \tau(\theta)\right| = \left|T_n - \mathrm{E}_\theta T_n + \mathrm{E}_\theta T_n - \tau(\theta)\right| \leq \left|T_n - \mathrm{E}_\theta T_n\right| + \left|\mathrm{E}_\theta T_n - \tau(\theta)\right| \qquad \text{[by triangle inequality]}$$

$$\left|T_n - \mathrm{E}_\theta T_n\right| < \varepsilon \quad \Rightarrow \quad \left|T_n - \tau(\theta)\right| < \varepsilon + \left|\mathrm{E}_\theta T_n - \tau(\theta)\right|$$

Therefore,

$$\Pr\left\{\left|T_n - \tau(\theta)\right| < \varepsilon + \left|\mathrm{E}_\theta T_n - \tau(\theta)\right|\right\} \geq \Pr\left\{\left|T_n - \mathrm{E}_\theta T_n\right| < \varepsilon\right\} > 1 - \frac{\operatorname{var} T_n}{\varepsilon^2}$$

i.e. $$\Pr\left\{\left|T_n - \tau(\theta)\right| < \varepsilon + \left|\mathrm{E}_\theta T_n - \tau(\theta)\right|\right\} > 1 - \frac{\operatorname{var} T_n}{\varepsilon^2}$$

As $n \to \infty$, $\left|\mathrm{E}_\theta T_n - \tau(\theta)\right| \to 0$ and $\operatorname{var} T_n \to 0$. Also all probabilities are ≤ 1, therefore

$$\Pr\left\{\left|T_n - \tau(\theta)\right| < \varepsilon\right\} \to 1$$

and T_n is consistent. ■

A final important result: although MLEs are sometimes biased, they are always consistent.

2.6 CRAMÉR-RAO INEQUALITY

Let $\mathbf{X} = (X_1, ..., X_n)$ be a vector r.v. with likelihood function $L_{\mathbf{X}}(\theta)$, where θ is a univariate parameter. Assume that: (i) the p.d.f. of each X_i has a range that does not depend on θ; (ii) $L_{\mathbf{X}}(\theta)$ is differentiable w.r.t. θ; (iii) derivatives w.r.t. θ can be moved inside and outside integrals involving $L_{\mathbf{X}}(\theta)$.

We now define two important quantities associated with families of distributions:

$$S = S(\mathbf{X};\theta) = \frac{\partial}{\partial\theta}\log L_{\mathbf{X}}(\theta),$$
$$I_{\mathbf{X}}(\theta) = \operatorname{var} S(\mathbf{X};\theta).$$

$S(\mathbf{X};\theta)$ is called the <u>score function</u>. $I_{\mathbf{X}}(\theta)$ is called the <u>Fisher information</u> in $\mathbf{X}$ (and is used to measure the amount of information about θ in the *n* observations).

<u>Theorem 2.6A</u>

(a) $\mathrm{E}S(\mathbf{X};\theta) = 0$;

(b) $I_{\mathbf{X}}(\theta) = \mathrm{E}S(\mathbf{X};\theta)^2 = -\mathrm{E}\frac{\partial}{\partial\theta}S(\mathbf{X};\theta) = -\mathrm{E}\frac{\partial^2}{\partial\theta^2}\log L_{\mathbf{X}}(\theta)$

<u>Proof:</u> (a) We have