2. Estimation

1. Sample Analogue Estimation

Let $x = (x_1, \ldots, x_n)'$ be an observation from $X = (X_1, \ldots, X_n)'$. Assume that the random variables $X_1, \ldots, X_n$ are independent and has common underlying distribution, which we parameterize as

$$\mathcal{P} = \{\mathrm{P}_\theta | \, \theta \in \Theta\}$$

The underlying distribution is often called the *population*.

We denote by E_θ the integration with respect to the probability distribution P_θ on $\mathbf{R}$, i.e.,

$$\mathrm{E}_\theta(f) = \int f \, d\mathrm{P}_\theta$$

for a measurable function $f : \mathbf{R} \to \mathbf{R}$. Furthermore, we define P_n be a probability distribution, called the *empirical distribution*, which assigns probability mass $1/n$ on each of points x_i, $i = 1, \ldots, n$. Then let E_n be the integration with respect to P_n on $\mathbf{R}$, i.e.,

$$\mathrm{E}_n(f) = \int f \, d\mathrm{P}_n = \frac{1}{n} \sum_{i=1}^{n} f(x_i)$$

again for any measurable function $f : \mathbf{R} \to \mathbf{R}$.

Estimator Suppose

$$\pi = \mathrm{E}_\theta(f)$$

is the parameter of interest. The sample analogue principle suggests that we estimate π by

$$\hat{\pi} = \mathrm{E}_n(f)$$

Loosely speaking, the principle proposes to estimate a population characteristic by its sample analogue.

Remark If $f(x) = x^k$ and π is a moment of the underlying distribution, then the method reduces to what is known as the method of moments.

Examples

(a) Let X_i be i.i.d. Poisson(λ). To get the sample analogue estimator $\hat{\lambda}$ of λ, we note that

$$\mathrm{E}_\lambda(f) = \lambda$$

for $f(x) = x$. It therefore follows that

$$\hat{\lambda} = \mathrm{E}_n(f) = \bar{x}$$

(b) Let X_i be i.i.d. $\mathbf{N}(\mu, \sigma^2)$. Suppose we want to estimate two parameters μ and σ^2. Notice that

$$\mathrm{E}_{\mu,\sigma^2}(f) = \mu \quad \text{and} \quad \mathrm{E}_{\mu,\sigma^2}(g) = \sigma^2$$

for $f(x) = x$ and $g(x) = (x - \mu)^2$. However,

$$\mathrm{E}_n(f) = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{and} \quad \mathrm{E}_n(g) \;=\; \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

The sample analogue estimators for μ and σ^2 are therefore given by

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

2. Maximum Likelihood Estimation

As before, let $x = (x_1, \ldots, x_n)'$ be an observation from $X = (X_1, \ldots, X_n)'$, whose density is assumed to belong to the family

$$\mathcal{P} = \{p(\cdot, \theta) | \, \theta \in \Theta\}$$

Note that here we use the notation $p(\cdot, \theta)$, instead of $p_\theta(\cdot)$, to denote a member in the class. This is because we now wish to regard a density p also as a function of θ. If a density is thought of as a function of the unknown parameter θ, then it is called the *likelihood function.*

Estimator The *maximum likelihood estimate* (MLE) of θ is defined by

$$\hat{\theta}_{\mathrm{ML}} = \operatorname*{argmax}_{\theta \in \Theta} p(x, \theta)$$

Remarks

(a) Computationally, it is often much easier to maximize the *log-likelihood function*

$$\ell(x, \theta) = \log\, p(x, \theta)$$

which is legitimate since log function is monotone increasing.

(b) Usually, the function $\ell(x, \cdot)$ is differentiable and globally concave for every x. The maximizer can therefore be found simply by solving the first-order condition (FOC)

$$\frac{\partial}{\partial \theta}\, \ell(x, \theta) = 0$$

for θ in terms of x.

(c) The MLE of a function of θ, say $\pi = f(\theta)$, is given by

$$\hat{\pi}_{\mathrm{ML}} = f(\hat{\theta}_{\mathrm{ML}})$$

since any other value of π results in values of θ different from $\hat{\theta}_{\mathrm{ML}}$ which yield smaller likelihood. It can, in particular, be said that the ML estimation is invariant with respect to reparametrization.

Examples

(a) Let X_i, $i = 1, \ldots, n$, be i.i.d. Bernoulli(θ). Then the log-likelihood function is given by

$$\ell(x, \theta) = \left(\sum_{i=1}^{n} x_i\right) \log \theta + \left(n - \sum_{i=1}^{n} x_i\right) \log(1 - \theta)$$

with the FOC

$$\frac{\partial \ell}{\partial \theta} = \frac{\sum_{i=1}^{n} x_i}{\theta} - \frac{n - \sum_{i=1}^{n} x_i}{1 - \theta} = 0$$

and this yields $\hat{\theta}_{\mathrm{ML}} = \bar{x}$.

(b) Let X_i, $i = 1, \ldots, n$, be i.i.d. $\mathbf{N}(\mu, \sigma^2)$. Then the log-likelihood function is given by

$$\ell(x, \mu, \sigma^2) = -\frac{n}{2} \log 2\pi - \frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2$$

and solving the FOC's yields

$$\begin{aligned} \hat{\mu}_{\mathrm{ML}} &= \bar{x} \\ \hat{\sigma}^2_{\mathrm{ML}} &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \end{aligned}$$

simultaneously.

(c) Let X_i, $i = 1, \ldots, n$, be i.i.d. $U[0, \theta]$. Then

$$\begin{aligned} p(x, \theta) &= \frac{1}{\theta^n}\prod_{i=1}^{n} \mathrm{I}\{0 \le x_i \le \theta\} \\ &= \frac{1}{\theta^n}\mathrm{I}\left\{\min_{1\le i\le n} x_i \ge 0\right\}\mathrm{I}\left\{\max_{1\le i\le n} x_i \le \theta\right\} \end{aligned}$$

it follows that $\hat{\theta}_{\mathrm{ML}} = \max\{x_1, \ldots, x_n\}$.

3. Uniformly Minimum Variance Unbiased Estimators

As desirable properties for a good estimator, we introduce the concepts of *unbiasedness* and *minimum mean squared error*. We first define unbiasedness. Here and elsewhere, we denote by $\mathbf{P}_\theta$ the probability in Ω that corresponds to P_θ in $\mathcal{X}$. Expectation with respect to $\mathbf{P}_\theta$ is denoted by $\mathbf{E}_\theta$.

Definition 1 *An estimator $T = \tau(X)$ is called unbiased if*

$$\mathbf{E}_\theta(T) = \theta$$

for all $\theta \in \Theta$.

The *mean squared error* (MSE) of an estimator $T = \tau(X)$ can be decomposed as the sum of the variance and the squared bias, as shown below

$$\mathbf{E}_\theta(T - \theta)^2 = \mathbf{E}_\theta(T - \mathbf{E}_\theta(T))^2 + (\mathbf{E}_\theta(T) - \theta)^2$$

Remarks

(a) Needless to say, it is preferable for an estimator to have small MSE. For an unbiased estimator, mean squared error reduces to variance.

(b) MSE is, in general, a function of the unknown parameter θ. An estimator that has the smallest MSE for all $\theta \in \Theta$ does not exist. This is because the trivial estimator $\hat{\theta} = \theta_0$ for some fixed value θ_0 has zero MSE at $\theta = \theta_0$. The MSE of any other estimator is strictly positive at $\theta = \theta_0$, since it takes other values with positive probabilities.

It is sometimes, if not always, possible to find an estimator with minimum variance for all $\theta \in \Theta$, if we restrict ourselves to the class of unbiased estimators. We are thus led to define

Definition 2 *An estimator $T_* = \tau_*(X)$ is called a* uniformly minimum variance unbiased (UMVU) *estimator if it satisfies*

(a) *T_* is unbiased, and*

(b) $\mathbf{E}_\theta(T_* - \theta)^2 \leq \mathbf{E}_\theta(T - \theta)^2$ *for any unbiased estimator $T = \tau(X)$.*

4. Information Inequality

We let $\ell(x, \theta)$ be the log-likelihood function, as defined before. Define

Definition 3

(a) (score function) $s(x, \theta) = \dfrac{\partial}{\partial \theta} \ell(x, \theta)$

(b) (Hessian) $h(x, \theta) = \dfrac{\partial^2}{\partial \theta \partial \theta'} \ell(x, \theta)$

(c) ((Fisher) information) $I(\theta) = \mathbf{E}_\theta(s(X, \theta) s(X, \theta)')$

(d) (expected Hessian) $H(\theta) = \mathbf{E}_\theta \, h(X, \theta)$

Remark Let X_i, $i = 1, 2$, be independent and define $X = (X_1', X_2')'$. If we denote scores and Hessians of X_i's by $s(x_i, \theta)$ and $h(x_i, \theta)$ respectively, then the score $s(x, \theta)$ and Hessian $h(x, \theta)$ of X are given by

$$s(x, \theta) = s(x_1, \theta) + s(x_2, \theta) \quad \text{and} \quad h(x, \theta) = h(x_1, \theta) + h(x_2, \theta)$$

Similarly, if we let $I_i(\theta)$ and $H_i(\theta)$ respectively be the information and expected Hessian of X_i, then the information $I(\theta)$ and expected Hessian $H(\theta)$ of X are

$$I(\theta) = I_1(\theta) + I_2(\theta) \quad \text{and} \quad H(\theta) = H_1(\theta) + H_2(\theta)$$

If, in particular, X_i, $i = 1, \ldots, n$, are independent random samples with the information and expected Hessian denoted respectively by $\iota_i(\theta)$ and $h_i(\theta)$, then the information $I(\theta)$ and expected Hessian $H(\theta)$ of $X = (X_1, \ldots, X_n)'$ are given by

$$I(\theta) = \sum_{i=1}^{n} \iota_i(\theta) \quad \text{and} \quad H(\theta) = \sum_{i=1}^{n} h_i(\theta)$$

If X_i's have identical distribution with $\iota(\theta)$ and $h(\theta)$, then $I(\theta) = n\,\iota(\theta)$ and $H(\theta) = n\,h(\theta)$.

In what follows, we let X have density $p(x, \theta)$ with respect to Lebesque measure μ, and let τ (or $\tau(X)$) be an estimator for θ. The information and expected Hessian of X are denoted respectively by $I(\theta)$ and $H(\theta)$. The following assumptions will be needed for our subsequent development.

Assumption 1 *(Regularity Conditions)*

(a) $\dfrac{\partial}{\partial\theta}\displaystyle\int p(x,\theta)d\mu(x) = \int \frac{\partial}{\partial\theta}p(x,\theta)d\mu(x)$

(b) $\dfrac{\partial^2}{\partial\theta\partial\theta'}\displaystyle\int p(x,\theta)d\mu(x) = \int \frac{\partial^2}{\partial\theta\partial\theta'}p(x,\theta)d\mu(x)$

(c) $\displaystyle\int \tau(x)\frac{\partial}{\partial\theta'}p(x,\theta)d\mu(x) = \frac{\partial}{\partial\theta'}\int \tau(x)p(x,\theta)d\mu(x)$

Proposition 1 *Suppose that* Assumption 1(a) *holds. Then*

$$\mathbf{E}_\theta\, s(X, \theta) = 0$$

Proof Notice that

$$\int s(x,\theta)p(x,\theta)d\mu(x) \;=\; \int \frac{\partial}{\partial\theta}\ell(x,\theta)p(x,\theta)d\mu(x)$$

$$\begin{aligned} &= \int \frac{\frac{\partial}{\partial\theta}p(x,\theta)}{p(x,\theta)}p(x,\theta)d\mu(x) \\ &= \frac{\partial}{\partial\theta}\int p(x,\theta)d\mu(x) \\ &= 0 \end{aligned}$$

as we wanted to show. ■

Remark The information $I(\theta)$ is therefore the variance of random score $s(X,\theta)$, which has expectation zero.

Proposition 2 *Suppose that* Assumption 1(b) *holds. Then*

$$I(\theta) = -H(\theta)$$

Proof Notice that

$$\frac{\partial^2}{\partial\theta\partial\theta'}\ell(x,\theta) = \frac{\frac{\partial^2}{\partial\theta\partial\theta'}p(x,\theta)}{p(x,\theta)} - \frac{\partial}{\partial\theta}\log p(x,\theta)\frac{\partial}{\partial\theta'}\log p(x,\theta)$$

we get

$$\begin{aligned} H(\theta) &= \int \left(\frac{\partial^2}{\partial\theta\partial\theta'}\ell(x,\theta)\right)p(x,\theta)d\mu(x) \\ &= \frac{\partial^2}{\partial\theta\partial\theta'}\int p(x,\theta)d\mu(x) - I(\theta) \\ &= -I(\theta) \end{aligned}$$

as was to be shown. ■

Lemma 3 *Let $T = \tau(X)$ be an unbiased estimator for θ for which* Assumption 1(c) *holds. Then*

$$\mathbf{E}_\theta(\tau(X)s(X,\theta)') = I$$

Proof Note that

$$\begin{aligned}\mathbf{E}_\theta(\tau(X)s(X,\theta)') &= \int \tau(x)\frac{\dfrac{\partial}{\partial\theta'}p(x,\theta)}{p(x,\theta)}p(x,\theta)d\mu(x) \\ &= \frac{\partial}{\partial\theta'}\int \tau(x)p(x,\theta)d\mu(x) \\ &= I\end{aligned}$$

from which the stated result is immediate. ■

Remark Since the expectation of random score is zero, the lemma implies that covariance between an unbiased estimator and random score is identity for all θ.

Theorem 4 (Cramer-Rao Bound) *Let $T = \tau(X)$ be an unbiased estimator of θ, and suppose that* Assumption 1(c) *holds. Then*

$$\operatorname{var}_\theta \tau(X) \geq I(\theta)^{-1}$$

Proof Note that

$$\operatorname{var}_\theta \begin{pmatrix} \tau(X) \\ s(X,\theta) \end{pmatrix} = \begin{pmatrix} \operatorname{var}_\theta \tau(X) & I \\ I & I(\theta) \end{pmatrix}$$

To get the stated result, observe that for any matrix

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \geq 0$$

we have $A_{11} \geq A_{12}A_{22}^{-1}A_{21}$. This is because we may write

$$A_{11} - A_{12}A_{22}^{-1}A_{21} = B'AB \geq 0$$

with $B' = (I, \ -A_{12}A_{22}^{-1})$. ■

Examples

(a) Let $X_1, \ldots, X_n$ be i.i.d. Poisson(λ). Then the log-likelihood ℓ, score s and information ι of each X_i, $i = 1, \ldots, n$, is given by

$$\begin{aligned} \ell(x_i, \lambda) &= -\lambda + x_i \log \lambda - \log x_i! \\ s(x_i, \lambda) &= -1 + \frac{x_i}{\lambda} \\ \iota(\lambda) &= \frac{1}{\lambda^2} \mathbf{E}_\lambda (X_i - \lambda)^2 = \frac{1}{\lambda} \end{aligned}$$

and therefore, information I of $X = (X_1, \ldots, X_n)'$ becomes

$$I(\lambda) = n\, \iota(\lambda) = \frac{n}{\lambda}$$

The unbiased estimator $\tau(X) = \bar{X}$ achieves the Cramer-Rao bound, since

$$\mathrm{var}_\lambda(\overline{X}) = \frac{\lambda}{n}$$

It is thus an UMVU estimator.

(b) Let $X_1, \ldots, X_n$ be i.i.d. $\mathbf{N}(\mu, \sigma^2)$. Then the log-likelihood ℓ, score s and hessian h of each X_i, $i = 1, \ldots, n$, is given by

$$\begin{aligned} \ell(x_i, \mu, \sigma^2) &= -\frac{1}{2} \log 2\pi - \frac{1}{2} \log \sigma^2 - \frac{(x_i - \mu)^2}{2\sigma^2} \\ s(x_i, \mu, \sigma^2) &= \begin{pmatrix} \dfrac{x_i - \mu}{\sigma^2} \\ -\dfrac{1}{2\sigma^2} + \dfrac{(x_i - \mu)^2}{2\sigma^4} \end{pmatrix} \\ h(x_i, \mu, \sigma^2) &= \begin{pmatrix} -\dfrac{1}{\sigma^2} & -\dfrac{x_i - \mu}{\sigma^4} \\ -\dfrac{x_i - \mu}{\sigma^4} & \dfrac{1}{2\sigma^4} - \dfrac{(x_i - \mu)^2}{\sigma^6} \end{pmatrix} \end{aligned}$$

We may now easily deduce that

$$I(\mu, \sigma^2) = -H(\mu, \sigma^2) = \begin{pmatrix} \dfrac{n}{\sigma^2} & 0 \\ 0 & \dfrac{n}{2\sigma^4} \end{pmatrix}$$

where I and H are, respectively, the information and expected Hessian of $X = (X_1, \ldots, X_n)'$. It can now be easily seen that the estimator $\overline{X}$ for μ achieves the Cramer-Rao bound, and therfore is an UMVU estimator. However, the variance of the estimator S^2 for σ^2 is strictly greater than the Cramer-Rao bound.

5. Exercises

1. Let $X_1, \ldots, X_n$ be a random sample from the underlying distribution given by the density

$$p(x, \theta) = \frac{2x}{\theta^2} \mathrm{I}\{0 \le x \le \theta\}$$

(a) Find the MLE of θ.

(b) Show that $T = \max\{X_1, \ldots, X_n\}$ is sufficient.

(c) Let

$$\begin{aligned} S_1 &= (\max\{X_1, \ldots, X_m\}, \max\{X_{m+1}, \ldots, X_n\}) \\ S_2 &= (\max\{X_1, \ldots, X_m\}, \min\{X_{m+1}, \ldots, X_n\}) \end{aligned}$$

where $1 < m < n$. Discuss the sufficiency of S_1 and S_2.

2. Let $X_1, \ldots, X_n$ be iid Uniform$[\alpha - \beta, \alpha + \beta]$, where $\beta > 0$, and let $\theta = (\alpha, \beta)$.

(a) Find a minimal sufficient statistic τ for θ.

(b) Find the ML estimator $\hat{\theta}_{\mathrm{ML}}$ of θ. (Hint: Graph the region for θ such that the joint density $p(x, \theta) > 0$.)

(c) Given the fact that τ in (a) is complete, find the UMVU estimator of α. (Hint: Note that $\mathbf{E}_\theta(X_1) = \alpha$.)

3. Let $X_1, \ldots, X_n$ be a random sample from a normal distribution with mean μ and variance σ^2. Define

$$\overline{X} = \frac{\sum_{i=1}^n X_i}{n} \quad \text{and} \quad S^2 = \frac{\sum_{i=1}^n (X_i - \overline{X})^2}{n-1}$$

(a) Obtain the Cramer-Rao lower bound.

(b) See whether $\overline{X}$ and S^2 attain the lower bound.

(c) Show that $\overline{X}$ and S^2 are jointly sufficient for μ and σ^2.

(d) Are $\overline{X}$ and S^2 the UMVU estimators?