3 Minimum Variance Unbiased Estimation

3.1 The Cramér-Rao inequality

Recall the following result from Lecture 3:

Lemma 1 (The Cauchy-Schwarz Inequality) For any random variables Y and Z with finite second moments,

$$[\mathrm{cov}(Y, Z)]^2 \leq \mathrm{var}(Y)\mathrm{var}(Z).$$

Equality is attained in the above inequality if and only if Y and Z are linearly related, i.e., there exist constants a and b such that (with probability 1) $Z = aY + b$.

The Cauchy-Schwarz Inequality is equivalent to the statement that the correlation coefficient ρ satisfies $|\rho(Y, Z)| \leq 1$.

We shall further require to assume certain regularity conditions about the family of p.d.f.'s,

$$\mathcal{F} = \{f(\mathbf{x}; \theta),\ \mathbf{x} \in \mathcal{X} : \theta \in \Theta\}.$$

We shall restrict attention to the case when θ is a single parameter, so that $\Theta \subseteq \mathbb{R}$. We shall use integral notation, referring to continuous distributions, but the results will be equally valid for families of discrete distributions

Because $f(\cdot)$ is a p.d.f., it follows that

$$\int_{\mathcal{X}} f(\mathbf{x}; \theta) d\mathbf{x} = 1 \qquad \theta \in \Theta.$$

Hence

$$\frac{\partial}{\partial\theta} \int_{\mathcal{X}} f(\mathbf{x}; \theta) d\mathbf{x} = 0 \qquad \theta \in \Theta.$$

Assuming that f is a differentiable function of θ for (almost) all $\mathbf{x} \in \mathcal{X}$, and further assuming regularity conditions that allow us to differentiate through the integral sign, it follows that

$$\int_{\mathcal{X}} \frac{\partial f(\mathbf{x}; \theta)}{\partial\theta} d\mathbf{x} = 0 \qquad \theta \in \Theta. \tag{1}$$

Among the regularity conditions required for Equation (1) to be valid is that the range of values of $\mathbf{x}$ for which $f(\mathbf{x}; \theta)$ is strictly positive should not depend upon θ, thus ruling out, for example, the family of uniform Uniform$(0, \theta)$ distributions.

Theorem 1 (The Cramér-Rao Inequality) *Let $\mathbf{X}$ be a vector of observations from the family $\mathcal{F}$ of probability distributions and $\hat{\tau}$ an unbiased estimator of $\tau(\theta)$, where τ is a differentiable function of θ. Under suitable regularity conditions,*

$$\mathrm{var}_\theta(\hat{\tau}) \geq \frac{\left(\frac{d\tau}{d\theta}\right)^2}{I(\theta)} \qquad \theta \in \Theta,$$

where $I(\theta)$ is the Fisher information*,*

$$I(\theta) = E_\theta\left[\left(\frac{\partial \ln f(\mathbf{X};\theta)}{\partial\theta}\right)^2\right] \qquad \theta \in \Theta. \tag{2}$$

Proof. In the Cauchy-Schwarz inequality (Lemma 1), for any given $\theta \in \Theta$, we shall take

$$Y = \hat{\tau}(\mathbf{X})$$

and

$$Z = \frac{\partial \ln f(\mathbf{X};\theta)}{\partial\theta}.$$

From Equation (1),

$$E_\theta[Z] = E_\theta\left[\frac{\partial \ln f(\mathbf{X};\theta)}{\partial\theta}\right] = \int_{\mathcal{X}} \frac{\partial f(\mathbf{x};\theta)}{\partial\theta} d\mathbf{x} = 0.$$

Hence

$$\mathrm{var}_\theta(Z) = E_\theta\left[\left(\frac{\partial \ln f(\mathbf{X};\theta)}{\partial\theta}\right)^2\right] = I(\theta) \tag{3}$$

and

$$\begin{aligned} \mathrm{cov}_\theta(Y,Z) &= E_\theta[YZ] \\ &= \int_{\mathcal{X}} \hat{\tau}(\mathbf{x}) \frac{\partial f(\mathbf{x};\theta)}{\partial\theta} d\mathbf{x} \\ &= \frac{\partial}{\partial\theta} \int_{\mathcal{X}} \hat{\tau}(\mathbf{x}) f(\mathbf{x};\theta) d\mathbf{x} \\ &= \frac{\partial E_\theta[\hat{\tau}(\mathbf{X})]}{\partial\theta} \\ &= \frac{d\tau}{d\theta}, \end{aligned} \tag{4}$$

where we have again assumed a regularity condition that allows to differentiate through the integral sign and have used the fact that $\hat{\tau}$ is an unbiased estimator of $\tau(\theta)$.

Rearranging the Cauchy-Schwarz inequality, we have

$$\mathrm{var}(Y) \geq \frac{[\mathrm{cov}(Y,Z)]^2}{\mathrm{var}(Z)}.$$

Substituting the expressions (3) and (4) we obtain the result of the theorem.

An important special case of Theorem 1 is the following.

Theorem 2 *If $\hat{\theta}$ is an unbiased estimator of θ then, provided that appropriate regularity conditions are satisfied,*

$$\mathrm{var}_\theta(\hat{\theta}) \geq \frac{1}{I(\theta)} \qquad \theta \in \Theta.$$

Proof: immediate from Theorem 1 with $\tau(\theta) \equiv \theta$.

A sometimes useful alternative expression for the Fisher information $I(\theta)$ is provided by Theorem 3.

Theorem 3 *Under appropriate regularity conditions,*

$$I(\theta) = E_\theta\left[-\frac{\partial^2 \ln f(\mathbf{X};\theta)}{\partial\theta^2}\right] \qquad \theta \in \Theta.$$

Proof.

$$\begin{aligned} \frac{\partial^2 \ln f(\mathbf{x};\theta)}{\partial\theta^2} &= \frac{\partial}{\partial\theta}\left(\frac{1}{f}\frac{\partial f}{\partial\theta}\right) \\ &= -\frac{1}{f^2}\left(\frac{\partial f}{\partial\theta}\right)^2 + \frac{1}{f}\frac{\partial^2 f}{\partial\theta^2} \\ &= -\left(\frac{\partial \ln f(\mathbf{x};\theta)}{\partial\theta}\right)^2 + \frac{1}{f}\frac{\partial^2 f}{\partial\theta^2}. \end{aligned}$$

Hence

$$\begin{aligned} E_\theta\left[-\frac{\partial^2 \ln f(\mathbf{X};\theta)}{\partial\theta^2}\right] &= E_\theta\left[\left(\frac{\partial \ln f(\mathbf{X};\theta)}{\partial\theta}\right)^2\right] - \int_{\mathcal{X}} \frac{\partial^2 f}{\partial\theta^2} d\mathbf{x} \\ &= I(\theta), \end{aligned}$$

where we have assumed the regularity condition that we can differentiate through the integral in Equation (1) to obtain

$$\int_{\mathcal{X}} \frac{\partial^2 f}{\partial\theta^2} d\mathbf{x} = 0.$$

It may sometimes be straightforward to evaluate $I(\theta)$ directly from Equation (2) or from the result of Theorem 3, using known properties of a given family $\mathcal{F}$ of distributions. On some occasions we may need to evaluate $I(\theta)$ as an integral or a summation, for continuous or discrete distributions, respectively. For example, in the continuous case, Equation (2) may be written as

$$I(\theta) = \int_{\mathcal{X}} \left(\frac{\partial \ln f(\mathbf{x};\theta)}{\partial\theta}\right)^2 f(\mathbf{x};\theta)\ d\mathbf{x}. \tag{5}$$

Usually, the components $X_1, X_2, \ldots, X_n$ of X will represent a random sample, so that, for any given θ, the X_i are i.i.d. r.v.s and their joint p.d.f. may be factorized as

$$f(\mathbf{x};\theta) = \prod_{i=1}^{n} f(x_i;\theta),$$

where on the right hand side of the above expression $f(x_i;\theta)$ now represents the marginal p.d.f. of X_i. It follows that

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

and hence that

$$\frac{\partial \ln f(\mathbf{X};\theta)}{\partial\theta} = \sum_{i=1}^{n} \frac{\partial \ln f(X_i;\theta)}{\partial\theta}, \tag{6}$$

the terms on the right hand side of Equation (6) being i.i.d. r.v.s for any given θ. Hence

$$\mathrm{var}_\theta\left(\frac{\partial \ln f(\mathbf{X};\theta)}{\partial\theta}\right) = \sum_{i=1}^{n} \mathrm{var}_\theta\left(\frac{\partial \ln f(X_i;\theta)}{\partial\theta}\right),$$

i.e., using the arguments of the proof of Theorem 1 and Equation (3) in particular,

$$I(\theta) = ni(\theta), \tag{7}$$

where $i(\theta)$ is the Fisher information corresponding to a single observation X_i. Any of the previously given expressions for evaluating the Fisher information may be used to determine $i(\theta)$ but with the p.d.f. f taken to be the marginal p.d.f. of a single observation.

According to Equation (7), the information from a random sample of size n is the sum of the information from each of the individual observations, i.e, n times the information from a single observation.

Definition

An estimator $\hat{\tau}^*$ is said to be a *minimum variance unbiased estimator* (MVUE) of $\tau(\theta)$ if it is unbiased and for every other unbiased estimator $\hat{\tau}$ of $\tau(\theta)$

$$\mathrm{var}_\theta(\hat{\tau}^*) \le \mathrm{var}_\theta(\hat{\tau}) \qquad \theta \in \Theta.$$

The importance of the Cramér-Rao inequality of Theorems 1 and 2 in conjunction with the definition of the Fisher information is that it provides a lower bound for the variance of unbiased estimators. Hence, if we find an unbiased estimator that attains the Cramér-Rao lower bound for all $\theta \in \Theta$ then we know that it is a MVUE.

Example 1

Consider a random sample $X_1, X_2, \ldots, X_n$ of size n from a Poisson distribution with parameter θ, so that the marginal p.m.f. for each observation is given by

$$f(x;\theta) = \frac{e^{-\theta}\theta^x}{x!}, \qquad x = 0, 1, 2, \ldots.$$

The mean and variance of the Poisson distribution are both equal to θ.

Perhaps a natural estimator of θ, obtained, for example, by the method of moments or the method of maximum likelihood, is $\hat{\theta}(\mathbf{X}) \equiv \bar{X}$. Its mean and variance are θ and θ/n, respectively. Thus $\hat{\theta}$ is unbiased as an estimator of θ. **For a single observation**, using the result of Theorem 3,

$$\begin{aligned} i(\theta) &= E_\theta\left[-\frac{\partial^2 \ln f(X;\theta)}{\partial\theta^2}\right] \\ &= E_\theta\left[-\frac{\partial^2}{\partial\theta^2}[-\theta + X\ln\theta - \ln(X!)]\right] \\ &= E_\theta\left[\frac{X}{\theta^2}\right] = \frac{1}{\theta}. \end{aligned}$$

Hence, from Equation (7), $I(\theta) = n/\theta$. Applying Theorem 2, it follows that the Cramér-Rao lower bound is θ/n. Thus $\mathrm{var}(\bar{X})$ attains the Cramér-Rao lower bound and so $\bar{X}$ must be a MVUE of θ.

Definition

Let $\hat{\tau}$ be an unbiased estimator of $\tau(\theta)$, where τ is a differentiable function of θ. The *efficiency*, $\mathrm{eff}_\theta(\hat{\tau})$, of $\hat{\tau}$ is defined by

$$\mathrm{eff}_\theta(\hat{\tau}) = \frac{\left(\frac{d\tau}{d\theta}\right)^2}{\mathrm{var}_\theta(\hat{\tau})I(\theta)} \qquad \theta \in \Theta.$$

The efficiency is the ratio of the Cramér-Rao lower bound to the variance of the estimator. From Theorem 1 it follows that

$$0 \le \mathrm{eff}_\theta(\hat{\tau}) \le 1 \qquad \theta \in \Theta.$$

If $\mathrm{eff}_\theta(\hat{\tau}) = 1,\ \theta \in \Theta$ then $\hat{\tau}$ is a MVUE. More generally, if the efficiency of an estimator is near 1, uniformly for all $\theta \in \Theta$, then this may be regarded as a recommendation for $\hat{\tau}$ as an estimator of $\tau(\theta)$.

However, a MVUE, if one exists, does not always attain the Cramér-Rao lower bound, so a small efficiency is not necessarily an indication that the corresponding estimator is a poor one.

In the case $\tau(\theta) = \theta$, if the Cramér-Rao bound is attained, $\mathrm{var}(\hat{\theta}) = 1/I(\theta)$. Thus $I(\theta)$ is indeed a measure of information, a measure of the precision of the estimator $\hat{\theta}$, the inverse of its mean square error.

3.2 The Rao-Blackwell theorem

The result of Theorem 1 by itself is not always enough to settle the question of whether a given unbiased estimator is a MVUE. Further results have to be established before we can develop a more complete treatment of MVUEs. An important result, which we state without proof, is the Rao-Blackwell theorem.

Theorem 4 (The Rao-Blackwell Theorem) *Let $\hat{\tau}$ be any unbiased estimator of $\tau(\theta)$, and T a sufficient statistic for θ. Consider the function $\hat{\phi}$ defined by $\hat{\phi} = E[\hat{\tau}|T]$. Then $\hat{\phi}$ is an unbiased estimator of $\tau(\theta)$ and*

$$\mathrm{var}_\theta(\hat{\phi}) \leq \mathrm{var}_\theta(\hat{\tau}) \qquad \theta \in \Theta.$$

The Rao-Blackwell Theorem shows that, given any unbiased estimator and any sufficient statistic T, we can find an unbiased estimator which is at least as good, according to the minimum variance criterion, by conditioning on T. If we have been able to find any unbiased estimator and any sufficient statistic T, the Rao-Blackwell theorem provides us with a method for finding, where possible, an improved estimator that is a function of T.

It also follows that in our search for a MVUE we can restrict attention to estimators that are functions of any given sufficient statistic. Thus the Rao-Blackwell Theorem demonstrates that use of the minimum variance criterion for comparing unbiased estimators is consistent with the sufficiency principle.

In fact, the theory of minimum variance unbiased estimation can be taken further. It turns out that if we are dealing with an exponential family of distributions and we are using the canonical sufficient statistic as described at the end of Section 1.4 then any unbiased estimator that is a function of this statistic is the unique MVUE.

If $\tau(\theta)$ has an unbiased estimator then it is said to be *estimable.* If $\tau(\theta)$ is estimable and we are dealing with an exponential family of distributions then we have the following method for constructing a MVUE of $\tau(\theta)$.

1. Find <u>any</u> unbiased estimator $\hat{\tau}$ of $\tau(\theta)$.

2. Determine the estimator $\hat{\phi} = E[\hat{\tau}|T]$, where T is the canonical sufficient statistic.

It follows that $\hat{\phi}$ is the unique MVUE of $\tau(\theta)$.

Example 1 (continued)

For a random sample $X_1, X_2, \ldots, X_n$ of size n from a Poisson distribution with parameter θ, consider estimation of the function $e^{-\theta}$. (In this case $\tau(\theta) \equiv e^{-\theta}$, but its MLE $e^{-\bar{X}}$ is not unbiased.)

It is easy to check that the Poisson distributions constitute an exponential family:

$$f(x;\theta) = \frac{e^{-\theta}\theta^x}{x!} = \frac{1}{x!}\exp(x\ln\theta - \theta), \qquad x = 0, 1, 2, \ldots,$$

so that $\ln\theta$ is the natural parameter and $T \equiv \sum_{i=1}^n X_i$ is the canonical sufficient statistic.

Note that $e^{-\theta} = f(0;\theta)$, the probability that an observation from the Poisson distribution takes the value 0. This enables us to construct a crude, unbiased estimator of $e^{-\theta}$, based upon the first observation X_1 alone. Define the estimator $\hat{\tau}$ by

$$\hat{\tau}(\mathbf{X}) = I_{\{X_1=0\}},$$

where I is the indicator function, so that

$$\hat{\tau}(\mathbf{X}) = \begin{cases} 1 & \text{if } X_1 = 0 \\ & \\ 0 & \text{otherwise} \end{cases}$$

First we check that $\hat{\tau}$ is indeed an unbiased estimator of $e^{-\theta}$:

$$E_\theta[\hat{\tau}(\mathbf{X})] = \mathbb{P}(X_1 = 0; \theta) = e^{-\theta}.$$

Define the estimator $\hat{\phi}(T)$ by conditioning on the statistic T:

$$\begin{aligned}
\hat{\phi}(t) &= E[\hat{\tau}(\mathbf{X})|T = t] \\
&= \mathbb{P}(X_1 = 0|T = t) \\
&= \frac{\mathbb{P}(X_1 = 0, T = t)}{\mathbb{P}(T = t)} \\
&= \frac{\mathbb{P}(X_1 = 0, S = t)}{\mathbb{P}(T = t)},
\end{aligned}$$

where

$$S = \sum_{i=2}^{n} X_i.$$

Note that X_1 and S are independently distributed. Furthermore S and T have Poisson distributions with parameters $(n-1)\theta$ and $n\theta$, respectively. Hence

$$\begin{aligned}
\hat{\phi}(t) &= \frac{\mathbb{P}(X_1 = 0)\mathbb{P}(S = t)}{\mathbb{P}(T = t)} \\
&= \frac{e^{-\theta} \times e^{-(n-1)\theta}[(n-1)\theta]^t/t!}{e^{-n\theta}[n\theta]^t/t!} \\
&= \left(1 - \frac{1}{n}\right)^t.
\end{aligned}$$

Thus the MVUE of $e^{-\theta}$ is given by $\hat{\phi}(T) \equiv (1 - 1/n)^T$.

3.3 Consistency and asymptotic efficiency of MLEs

Let $X_1, X_2, \ldots, X_n, \ldots$ be a sequence of i.i.d. r.v.s, each having p.d.f. $f(x; \theta)$ with parameter $\theta \in \Theta$.

Definition

A sequence of estimators $\hat{\tau}_n \equiv \hat{\tau}_n(X_1, X_2, \ldots, X_n)$ is said to be a *(weakly) consistent* sequence of estimators of the real-valued function $\tau(\theta)$ if for every $\epsilon > 0$ and every $\theta \in \Theta$

$$\mathbb{P}(|\hat{\tau}_n(X_1, X_2, \ldots, X_n) - \tau(\theta)| \geq \epsilon; \theta) \to 0 \quad \text{as} \quad n \to \infty.$$

Put another way, $\hat{\tau}_n \to \tau(\theta)$ as $n \to \infty$, in probability.

Although MLEs are not in general unbiased, under certain regularity conditions it can be shown that they are consistent.

Recall that the Fisher information for a sample of size n is given by $I(\theta) = ni(\theta)$, where $i(\theta)$ is the information corresponding to a single observation.

It can be shown, using the Central Limit theorem, that MLEs are asymptotically normally distributed and *asymptotically efficient*, i.e., in the case of $\Theta \subseteq \mathbb{R}$, for all $\theta \in \Theta$

$$\sqrt{n}(\hat{\tau}_n - \tau(\theta)) \to N\left(0, \frac{\left(\frac{d\tau}{d\theta}\right)^2}{i(\theta)}\right).$$

Put less formally, for large n,

$$\hat{\tau}_n \sim N\left(\tau(\theta), \frac{\left(\frac{d\tau}{d\theta}\right)^2}{I(\theta)}\right)$$

approximately. Thus, in the limit as the sample size tends to infinity, the MLEs $\hat{\tau}_n$ are asymptotically normally distributed about $\tau(\theta)$ with variance that attains the Cramér-Rao lower bound $\left(\frac{d\tau}{d\theta}\right)^2 / I(\theta)$. In this sense, as the sample size increases, MLEs are asymptotically the best possible estimators according to the criterion of minimum mean squared error.

The asymptotic distribution may be used to construct approximate confidence intervals. For example, when θ itself is being estimated, an approximate $100(1-\alpha)\%$ confidence interval for θ is given by

$$\hat{\theta} \pm \frac{z_{\alpha/2}}{\sqrt{I(\hat{\theta})}},$$

where $z_{\alpha/2}$ is the upper $100\frac{\alpha}{2}\%$ point of the standard normal distribution.