5.5 Minimum Variance Estimators

We discussed in the previous section how to compare two unbiased estimators for the same parameter. The one with smaller variance should be considered "better".

QUESTION: Is there a "best" estimator, in the sense of possessing a minimum variance? How do we know if an estimator is "best"?

In this section we discuss an answer to this question. We shall see that the variance of an unbiased estimator cannot be smaller than certain bound, called the Cramér-Rao bound.

The Cramer-Rao Inequality and Lower Bound

The Cramer-Rao Lower Bound (CRLB) sets a lower bound on the variance of any unbiased estimator[a]. It is useful as follows:

1. If we find an estimator that achieves the CRLB, then we know that we have found a Minimum Variance Unbiased Estimator (MVUE).

2. THe CRLB can provide a benchmark against which we can compare the performance of any unbiased estimator.

3. The CRLB can be used to rule-out impossible estimators.

4. The theory behind the CRLB can tell us if an estimator exists that achieves the lower bound (not discussed here)

[a] R. Nowak, C. Scott, The Cramer-Rao Lower Bound, cnx.rice.edu/content/m11429/latest

Theorem: The Cramer-Rao Inequality

Let $W_1,\ldots,W_n$ be a random sample from $f_W(w,\theta)$, where $f_W(w,\theta)$ has continuous first-order and second-order partial derivatives at all but a finite set of points. Suppose the set of w's for which $f_W(w,\theta)\neq 0$ does not depend on θ.
Let $\hat{\theta}=h(W_1,\ldots,W_n)$ be an unbiased estimator of θ. Then,

$$Var(\ \hat{\theta}\ ) \geq \frac{1}{n\,E\left[\left(\frac{\partial \ln f_W(w,\theta)}{\partial\theta}\right)^2\right]}$$

and

$$Var(\ \hat{\theta}\ ) \geq \frac{1}{n\,E\left[\frac{\partial^2 \ln f_W(w,\theta)}{\partial^2\theta}\right]}$$

Example 5.5.1 Let $X_1, X_2, \ldots, X_n$ denote the total number of successes in each of n independent trials, where p=Prob. of success at any given trial is unknown parameter.
We have that

$$p_{X_\ell}(k\,;\,p) = p^k(1-p)^k, \quad k=0,1; \quad 0<p<1$$

Let $X = X_1+X_2+\cdots+X_n$=total # of successes. Define $\hat{p}=X/n$.

a) Show that $\hat{p}$ is unbiased.
b) Compare $Var(\hat{p})$ with the CRLB for p_X.

ANSWER
(a): $E[\hat{p}] = E[X/n] = E[X]/n = np/n = p$.

(b) We have,

$$\begin{aligned} Var(\hat{p}) &= Var(X/n) = \tfrac{1}{n^2}Var(X) \\ &= \tfrac{1}{n^2}Var(X_1+\cdots+X_n) \\ &= \tfrac{1}{n^2}np\,(1-p) = \tfrac{1}{n}\,p\,(1-p) \end{aligned}$$

b) We use the second form of the CRLB.
Note that

$$\ln\, p_{X_\ell}(X_\ell;p) = X_\ell \ln p + (1-X_\ell)\ln(1-p)$$

Then

$$\frac{\partial \ln\, p_{X_\ell}(X_\ell;p)}{\partial p} = \frac{X_\ell}{p} - \frac{1-X_\ell}{1-p}$$

and

$$\frac{\partial^2 \ln\, p_{X_\ell}(X_\ell;p)}{\partial p^2} = \frac{X_\ell}{p^2} - \frac{1-X_\ell}{(1-p)^2}$$

Take expected value in above equation to get:

$$E\left[\frac{\partial^2 \ln\, p_{X_\ell}(X_\ell;p)}{\partial p^2}\right] = \frac{p}{p^2} - \frac{1-p}{(1-p)^2} = -\frac{1}{p(1-p)}$$

Substituting in the formula for the CRLB we get

$$\frac{1}{-n\left(-\frac{1}{p(1-p)}\right)} = \frac{p(1-p)}{n}$$

CONCLUSION: $Var(\hat{p})$ equals the CRLB.

Definition

1. $\hat{\theta}^*$ is a best or Minimum Variance Unbiased Estimator if it is unbiased and for all unbiased estimators θ,

$$Var(\hat{\theta}^*) \leq Var(\hat{\theta})$$

2. An unbiased estimator $\hat{\theta}$ is efficient if the variance of $\hat{\theta}$ equals the CRLB.

3. The efficiency of an estimator $\hat{\theta}$ is the ratio of the CRLB to $Var(\hat{\theta})$.

Example:

The estimator $\hat{\theta}$ in Example 5.5.1 is both best and efficient, and its efficiency is 1.

Example 5.5.2

Let $Y_1, \ldots, Y_n$ be a random sample from

$$f_Y(y;\theta) = 2y/\theta^2, \quad 0 < y < \theta$$

Can the CRLB be applied to the estimator

$$\hat{\theta} = 3/2 \cdot \overline{Y}$$

ANSWER: No. Reason: $f_Y(y;\theta) > 0$ on a set that depends on θ, thus violating one of the hypotheses of the theorem that gives the inequality.

Limitations of the CRLB

One limitation we already discussed, the fact that the "domain" of the pdf does not depend on the parameter.

Another limitation is that most estimators (including maximum likelihood estimators) are biased in finite samples. There is a version of the CRLB for biased estimators, but it is of limited practical value, since it contains a term that depends on an unknown quantity.

The CRLB is more useful in large samples for a class of estimators ("consistent") that have the property that they are asymptotically unbiased. It can be proved that under very general conditions, maximum-likelihood estimators are in this class.

Consistency will be studied in Section 5.7

5.6 Sufficiency

Consider a coin flipped 4 times.
The prob. p of success (head) is unknown.
Suppose that $(1,1,0,0)$ was observed.
An estimate for p is

$$\hat{p} = \frac{\#\text{successes}}{\#\text{trials}} = \frac{2}{4} = 0.5$$

Let's compute some conditional probabilities:

$$P((1,1,0,0) \mid \hat{p} = 0.5)$$

$$= \frac{P((1,1,0,0) \text{ and } \hat{p} = 0.5)}{P(\hat{p} = 0.5)}$$

$$= \frac{P(1,1,0,0)}{2 \text{ successes in 4 trials}}$$

$$= \frac{p^2(1-p)^2}{\binom{4}{2}p^2(1-p)^2}$$

$$= \binom{4}{2}^{-1}$$

Hence

$$P((1,1,0,0) \mid \hat{p} = 0.5) = \binom{4}{2}^{-1}$$

Similarly, we may verify that

$$\begin{aligned} P((1,0,1,0) \mid \hat{p} = 0.5) &= \binom{4}{2}^{-1} \\ P((1,0,0,1) \mid \hat{p} = 0.5) &= \binom{4}{2}^{-1} \\ P((0,1,1,0) \mid \hat{p} = 0.5) &= \binom{4}{2}^{-1} \\ P((0,1,0,1) \mid \hat{p} = 0.5) &= \binom{4}{2}^{-1} \\ P((0,0,1,1) \mid \hat{p} = 0.5) &= \binom{4}{2}^{-1} \end{aligned}$$

So knowing what a particular outcome is does not add any additional information to what we know about p, once we have been informed that $\hat{p} = 0.5$. This motivates the following

Definition Let $W_1, \ldots, W_n$ be a random sample from $f_W(w;\theta)$. The estimator $\hat{\theta} = h(W_1, \ldots, W_n)$ is said to be sufficient for θ if for all θ and all possible sample points, the conditional pdf of $W_1, \ldots, W_n$ given $\hat{\theta}$ does not depend on θ.

The Fisher-Neyman Criterion

Theorem 5.6.1 Let $W_1, \ldots, W_n$ be a random sample from $f_W(w;\theta)$. Then,

$\hat{\theta} = h(W_1, \ldots, W_n)$ is sufficient for θ

if and only if

the joint pdf of the W_ℓ's factors into a product of the pdf for $\hat{\theta}$ times a second function that does not depend on θ, that is,

$$\prod_{\ell=1}^{n} f_W(w_\ell;\theta) = f_{\hat{\theta}}(\hat{\theta};\theta) \cdot s(w_1, \ldots, w_n)$$

COMMENT

If $\hat{\theta}$ is sufficient for θ, then any one-to-one function of $\hat{\theta}$ (for ex., $k\hat{\theta}$ or $\hat{\theta} + k$) is also sufficient.

Example 5.6.1 Let $X_1, \ldots, X_n$ be a random sample of n Bernoulli RVs with unknown parameter p. The pdf of X_ℓ is

$$p_{X_\ell}(k;p) = p^k(1-p)^{1-k}, \quad k = 0, 1, \quad 0 \le p \le 1$$

Is $\hat{p} = \sum_{\ell=1}^{n} X_\ell$ sufficient for p?

ANSWER: The Fisher-Neyman condition requires that we compute three terms:

$$\text{(F-N)} \quad \prod_{\ell=1}^{n} f_W(w_\ell;\theta) = f_{\hat{\theta}}(\hat{\theta};\theta) \cdot s(w_1, \ldots, w_n)$$

(a) the product on the LHS of "=",

(b) the pdf of $\hat{\theta}$, and,

(c) the function s.

We now proceed to do this.

$$\text{(F-N)} \quad \prod_{\ell=1}^{n} f_W(w_\ell;\theta) = f_{\hat{\theta}}(\hat{\theta};\theta) \cdot s(w_1, \ldots, w_n)$$

(a) The product on the LHS of F-N is

$$p_{X_1}(k_1;p) \cdots p_{X_n}(k_n;p)$$

$$= p^{k_1}(1-p)^{1-k_1} \cdots p^{k_n}(1-p)^{1-k_n}$$

$$= p^k(1-p)^{1-k}, \text{ where } k = k_1 + \cdots + k_n$$

(b) Note that $\hat{p}$ is binomial since it is the number of successes in n independent trials. Then,

$$f_{\hat{p}}(k;p) = \binom{n}{k} p^k(1-p)^{1-k}$$

(c) Choose the function $s(\cdot)$ in F-N to be

$$s(k_1, \ldots, k_n) = \binom{n}{k}^{-1}, \text{ where } k = k_1 + \cdots + k_n$$

Then F-N theorem holds, hence $\hat{p}$ is sufficient.

Example 5.6.2 Let $Y_1, \ldots, Y_n$ be a random sample from the uniform pdf

$$f_Y(y;\theta) = 1/\theta, \quad 0 \le y \le \theta$$

Knowing that $\hat{\theta} = Y_{max}$ is the MLE for θ, determine if $\hat{\theta}$ is sufficient.

ANSWER: Here is F-N (for reference):

$$\text{(F-N)} \quad \prod_{\ell=1}^{n} f_W(w_\ell;\theta) = f_{\hat{\theta}}(\hat{\theta};\theta) \cdot s(w_1, \ldots, w_n)$$

The L-H-S of F-N is

$$f_{Y_1}(y_1;p) \cdots f_{Y_n}(y_n;p) = \frac{1}{\theta^n}$$

Recall from Ex. 5.4.2 that

$$f_{Y_{\max}}(u,\theta) = \frac{n\,u^{n-1}}{\theta^n}, \quad 0 \le u \le \theta$$

Finally, we may set $s(y_1, \ldots, y_n) := \frac{1}{n\,y_{\max}^{n-1}}$.

$$\frac{1}{\theta^n} = \frac{n\,(y_{\max})^{n-1}}{\theta^n} \cdot \frac{1}{n\,y_{\max}^{n-1}}$$

A Result that is Easier to Use than F-N

The Factorization Theorem 5.6.2 Let $W_1, \ldots, W_n$ be a random sample from $f_W(w;\theta)$. Then,

$\hat{\theta} = h(W_1, \ldots, W_n)$ is sufficient for θ

if and only if

there are functions $g(\hat{\theta},\theta)$ and $u(w_1, \ldots, w_n)$ such that

$$\prod_{\ell=1}^{n} f_W(w_\ell;\theta) = g(\hat{\theta},\theta) \cdot u(w_1, \ldots, w_n)$$

Example 5.6.3 A random sample of size n is drawn from the pdf

$$f_Y(y;\theta) = \theta y^{\theta-1}, \quad 0 < y < 1, \quad \theta > 0$$

Use the Factorization Theorem to find an estimator that is sufficient for θ.

ANSWER:

$$f_{Y_1}(y_1;\theta) \cdots f_{Y_n}(y_n;\theta)$$

$$= \theta y_1^{\theta-1} \cdots \theta y_n^{\theta-1}$$

$$= \theta^n (y_1 \cdots y_n)^{\theta-1}$$

By staring at the last expression, we see that $\hat{\theta} := y_1 \cdots y_n$

could be used as an estimator:

$$f_{Y_1}(y_1;\theta) \cdots f_{Y_n}(y_n;\theta)$$

$$= \theta^n (y_1 \cdots y_n)^{\theta-1} \quad = \theta^n \left(\hat{\theta}\right)^{\theta-1} \cdot 1$$

$$= g(\hat{\theta},\theta) \cdot u(y_1, \ldots, y_n)$$

By the Factorization Thm., $\hat{\theta}$ is sufficient.

Example 5.6.4: Why MLEs are preferred to Method-of-Moments Estimators

GIVEN: an MLE $\hat{\theta}_{MLE}$ for θ based on a random sample of size n drawn from a pdf $f_W(w,\theta)$.

GIVEN: a sufficient estimator $\hat{\theta}_s$ for θ.

CLAIM: $\hat{\theta}_{MLE}$ is a function of $\hat{\theta}_s$.

Idea of Proof: Consider the likelihood function

$$L(\theta) = \prod_{\ell=1}^{n} f_{W_\ell}(w_\ell,\theta)$$

From the Factorization Theorem we have

$$L(\theta) = g(\hat{\theta}_s,\theta) \cdot u(w_1, \ldots, w_n)$$

From above eqn. & since $\hat{\theta}_{MLE}$ maximizes $L(\theta)$, $\hat{\theta}_{MLE}$ maxi - mizes $g(\hat{\theta}_s,\theta)$. But any θ that maximizes $g(\hat{\theta}_s,\theta)$ is a fn. of $\hat{\theta}_s$.

Sufficient Estimators are more Efficient

Consider estimators for θ based on a random sample of size n drawn from $f_W(w,\theta)$.

A Theorem of Rao-Blackwell states that
given estimators,

- θ_1 unbiased and sufficient, and,
- θ_2 biased, not sufficient,

then necessarily $Var(\theta_1) < Var(\theta_2)$, that is, θ_1 is more efficient. Thus to search for highly efficient estimators, it suffices to search among sufficient estimators.

Moreover, a result of Lehman and Scheffé says that under very general conditions, there is only one sufficient estimator. If this is the case, then finding one sufficient estimator gives the best unbiased estimator.

5.7 Asymptotically unbiased estimators

Consider estimators $\hat{\theta}_n$ based on a random sample of size n taken from a pdf $f_Y(y;\theta)$. We say that $\hat{\theta}_n$ is asymptotically unbiased if

$$\lim_{n\to\infty} E(\hat{\theta}_n) = \theta, \quad \text{for all } \theta$$

EXAMPLE: A random sample of size n is drawn from a normal pdf. Set

$$\hat{\sigma}_n = \frac{1}{n}\sum_{\ell=1}^{n}(Y_\ell - \overline{Y})^2$$

Let us prove that $\hat{\sigma}_n$ is asymptotically unbiased.

$$\lim_{n\to\infty} E(\hat{\sigma}_n)$$

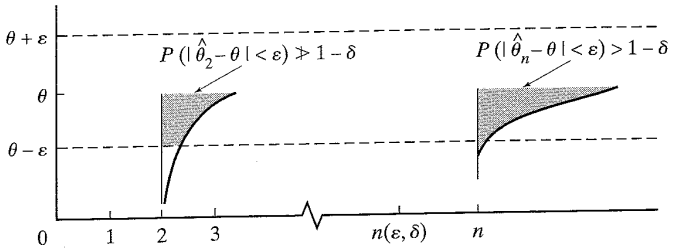
$$= \lim_{n\to\infty} E\left(\frac{1}{n}\sum_{\ell=1}^{n}(Y_\ell - \overline{Y})^2\right)$$

$$= \lim_{n\to\infty} \frac{n-1}{n}\sigma^2 \quad = \quad \sigma$$

Section 5.7: Consistent Estimators

Definition An estimator $\hat{\theta}_n = h(W_1,\ldots,W_n)$ is consistent for θ if it converges in probability to θ, that is,

$$\text{for all } \epsilon > 0, \quad \lim_{n\to\infty} P\left(|\hat{\theta}_n - \theta| < \epsilon\right) = 1$$



Problem 5.7.1 Let $Y_1,\ldots,Y_n$ be a random sample from the uniform distribution over $[0,\theta]$. Set $\hat{\theta}_n = Y_{max}$. Is $\hat{\theta}_n$ consistent?

ANSWER: Recall pdf of Y_{max} (p. 182) is:

$$f_{Y_{max}}(y) = \frac{n}{\theta}\left(\frac{y}{\theta}\right)^{n-1}, \quad 0 \le y \le \theta$$

Consider $\epsilon > 0$ arbitrary but fixed. Then,

$$P\left(|\hat{\theta}_n - \theta| < \epsilon\right) \quad = \quad P\left(\theta - \epsilon < \hat{\theta}_n < \theta\right)$$

$$= \int_{\theta-\epsilon}^{\theta} \frac{n}{\theta}\left(\frac{y}{\theta}\right)^{n-1} dy$$

$$= \left.\frac{y^n}{\theta^n}\right|_{\theta-\epsilon}^{\theta} = 1 - \left(\frac{\theta-\epsilon}{\theta}\right)^n$$

Then

$$\lim_{n\to\infty} P\left(|\hat{\theta}_n - \theta| < \epsilon\right) = \lim_{n\to\infty} 1 - \left(\frac{\theta-\epsilon}{\theta}\right)^n = 1$$

So YES, $\hat{\theta}_n$ is consistent.