

# Stat 709: Mathematical Statistics

## Lecture 32

Jun Shao

Department of Statistics  
University of Wisconsin  
Madison, WI 53706, USA

## Lecture 32: Information inequality

### Theorem 3.3 (Cramér-Rao lower bound)

Let  $X = (X_1, \dots, X_n)$  be a sample from  $P \in \mathcal{P} = \{P_\theta : \theta \in \Theta\}$ , where  $\Theta$  is an open set in  $\mathcal{R}^k$ .

Suppose that  $T(X)$  is an estimator with  $E[T(X)] = g(\theta)$  being a differentiable function of  $\theta$ ;  $P_\theta$  has a p.d.f.  $f_\theta$  w.r.t. a measure  $\nu$  for all  $\theta \in \Theta$ ; and  $f_\theta$  is differentiable as a function of  $\theta$  and satisfies

$$\frac{\partial}{\partial \theta} \int h(x) f_\theta(x) d\nu = \int h(x) \frac{\partial}{\partial \theta} f_\theta(x) d\nu, \quad \theta \in \Theta, \quad (1)$$

for  $h(x) \equiv 1$  and  $h(x) = T(x)$ .

Then

$$\text{Var}(T(X)) \geq \left[ \frac{\partial}{\partial \theta} g(\theta) \right]^\tau [I(\theta)]^{-1} \frac{\partial}{\partial \theta} g(\theta), \quad (2)$$

where

$$I(\theta) = E \left\{ \frac{\partial}{\partial \theta} \log f_\theta(X) \left[ \frac{\partial}{\partial \theta} \log f_\theta(X) \right]^\tau \right\}$$

is assumed to be positive definite for any  $\theta \in \Theta$ .

## Discussion

Suppose that we have a lower bound for the variances of all unbiased estimators of  $\vartheta$ .

If there is an unbiased estimator  $T$  of  $\vartheta$  whose variance is always the same as the lower bound, then  $T$  is a UMVUE of  $\vartheta$ .

Although this is not an effective way to find UMVUE's, it provides a way of assessing the performance of UMVUE's.

## Proof of Theorem 3.3

We prove the univariate case ( $k = 1$ ) only.

When  $k = 1$ , (2) reduces to

$$\text{Var}(T(X)) \geq \frac{[g'(\theta)]^2}{E \left[ \frac{\partial}{\partial \theta} \log f_{\theta}(X) \right]^2}.$$

From the Cauchy-Schwartz inequality, we only need to show that

$$E \left[ \frac{\partial}{\partial \theta} \log f_{\theta}(X) \right]^2 = \text{Var} \left( \frac{\partial}{\partial \theta} \log f_{\theta}(X) \right)$$

## Proof of Theorem 3.3 (continued)

and

$$g'(\theta) = \text{Cov} \left( T(X), \frac{\partial}{\partial \theta} \log f_{\theta}(X) \right).$$

From condition (1) with  $h(x) = 1$ ,

$$E \left[ \frac{\partial}{\partial \theta} \log f_{\theta}(X) \right] = \int \frac{\partial}{\partial \theta} f_{\theta}(X) d\nu = \frac{\partial}{\partial \theta} \int f_{\theta}(X) d\nu = 0.$$

From condition (1) with  $h(x) = T(x)$ ,

$$E \left[ T(X) \frac{\partial}{\partial \theta} \log f_{\theta}(X) \right] = \int T(x) \frac{\partial}{\partial \theta} f_{\theta}(X) d\nu = \frac{\partial}{\partial \theta} \int T(x) f_{\theta}(X) d\nu,$$

which  $= g'(\theta)$ .

The  $k \times k$  matrix

$$I(\theta) = E \left\{ \frac{\partial}{\partial \theta} \log f_{\theta}(X) \left[ \frac{\partial}{\partial \theta} \log f_{\theta}(X) \right]^{\tau} \right\}$$

is called the *Fisher information matrix*.

## Proof of Theorem 3.3 (continued)

and

$$g'(\theta) = \text{Cov} \left( T(X), \frac{\partial}{\partial \theta} \log f_{\theta}(X) \right).$$

From condition (1) with  $h(x) = 1$ ,

$$E \left[ \frac{\partial}{\partial \theta} \log f_{\theta}(X) \right] = \int \frac{\partial}{\partial \theta} f_{\theta}(X) d\nu = \frac{\partial}{\partial \theta} \int f_{\theta}(X) d\nu = 0.$$

From condition (1) with  $h(x) = T(x)$ ,

$$E \left[ T(X) \frac{\partial}{\partial \theta} \log f_{\theta}(X) \right] = \int T(x) \frac{\partial}{\partial \theta} f_{\theta}(X) d\nu = \frac{\partial}{\partial \theta} \int T(x) f_{\theta}(X) d\nu,$$

which  $= g'(\theta)$ .

The  $k \times k$  matrix

$$I(\theta) = E \left\{ \frac{\partial}{\partial \theta} \log f_{\theta}(X) \left[ \frac{\partial}{\partial \theta} \log f_{\theta}(X) \right]^{\tau} \right\}$$

is called the *Fisher information matrix*.

The greater  $I(\theta)$  is, the easier it is to distinguish  $\theta$  from neighboring values and, therefore, the more accurately  $\theta$  can be estimated.

Thus,  $I(\theta)$  is a measure of the information that  $X$  contains about  $\theta$ .

The inequality in (2) is called *information inequalities*.

The following result is helpful in finding the Fisher information matrix.

### Proposition 3.1

- (i) If  $X$  and  $Y$  are independent with the Fisher information matrices  $I_X(\theta)$  and  $I_Y(\theta)$ , respectively, then the Fisher information about  $\theta$  contained in  $(X, Y)$  is  $I_X(\theta) + I_Y(\theta)$ .

In particular, if  $X_1, \dots, X_n$  are i.i.d. and  $I_1(\theta)$  is the Fisher information about  $\theta$  contained in a single  $X_i$ , then the Fisher information about  $\theta$  contained in  $X_1, \dots, X_n$  is  $nI_1(\theta)$ .

- (ii) Suppose that  $X$  has the p.d.f.  $f_\theta$  that is twice differentiable in  $\theta$  and that (1) holds with  $h(x) \equiv 1$  and  $f_\theta$  replaced by  $\partial f_\theta / \partial \theta$ .

Then

$$I(\theta) = -E \left[ \frac{\partial^2}{\partial \theta \partial \theta^\tau} \log f_\theta(X) \right].$$

The greater  $I(\theta)$  is, the easier it is to distinguish  $\theta$  from neighboring values and, therefore, the more accurately  $\theta$  can be estimated.

Thus,  $I(\theta)$  is a measure of the information that  $X$  contains about  $\theta$ .

The inequality in (2) is called *information inequalities*.

The following result is helpful in finding the Fisher information matrix.

### Proposition 3.1

- (i) If  $X$  and  $Y$  are independent with the Fisher information matrices  $I_X(\theta)$  and  $I_Y(\theta)$ , respectively, then the Fisher information about  $\theta$  contained in  $(X, Y)$  is  $I_X(\theta) + I_Y(\theta)$ .

In particular, if  $X_1, \dots, X_n$  are i.i.d. and  $I_1(\theta)$  is the Fisher information about  $\theta$  contained in a single  $X_i$ , then the Fisher information about  $\theta$  contained in  $X_1, \dots, X_n$  is  $nI_1(\theta)$ .

- (ii) Suppose that  $X$  has the p.d.f.  $f_\theta$  that is twice differentiable in  $\theta$  and that (1) holds with  $h(x) \equiv 1$  and  $f_\theta$  replaced by  $\partial f_\theta / \partial \theta$ .

Then

$$I(\theta) = -E \left[ \frac{\partial^2}{\partial \theta \partial \theta^\tau} \log f_\theta(X) \right].$$

The greater  $I(\theta)$  is, the easier it is to distinguish  $\theta$  from neighboring values and, therefore, the more accurately  $\theta$  can be estimated.

Thus,  $I(\theta)$  is a measure of the information that  $X$  contains about  $\theta$ .

The inequality in (2) is called *information inequalities*.

The following result is helpful in finding the Fisher information matrix.

### Proposition 3.1

- (i) If  $X$  and  $Y$  are independent with the Fisher information matrices  $I_X(\theta)$  and  $I_Y(\theta)$ , respectively, then the Fisher information about  $\theta$  contained in  $(X, Y)$  is  $I_X(\theta) + I_Y(\theta)$ .

In particular, if  $X_1, \dots, X_n$  are i.i.d. and  $I_1(\theta)$  is the Fisher information about  $\theta$  contained in a single  $X_i$ , then the Fisher information about  $\theta$  contained in  $X_1, \dots, X_n$  is  $nI_1(\theta)$ .

- (ii) Suppose that  $X$  has the p.d.f.  $f_\theta$  that is twice differentiable in  $\theta$  and that (1) holds with  $h(x) \equiv 1$  and  $f_\theta$  replaced by  $\partial f_\theta / \partial \theta$ .

Then

$$I(\theta) = -E \left[ \frac{\partial^2}{\partial \theta \partial \theta^\tau} \log f_\theta(X) \right].$$

## Proof

Result (i) follows from the independence of  $X$  and  $Y$  and the definition of the Fisher information.

Result (ii) follows from the equality

$$\frac{\partial^2}{\partial \theta \partial \theta^\tau} \log f_\theta(X) = \frac{\frac{\partial^2}{\partial \theta \partial \theta^\tau} f_\theta(X)}{f_\theta(X)} - \frac{\partial}{\partial \theta} \log f_\theta(X) \left[ \frac{\partial}{\partial \theta} \log f_\theta(X) \right]^\tau.$$

## Example 3.9

Let  $X_1, \dots, X_n$  be i.i.d. with the Lebesgue p.d.f.  $\frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right)$ , where  $f(x) > 0$  and  $f'(x)$  exists for all  $x \in \mathcal{R}$ ,  $\mu \in \mathcal{R}$ , and  $\sigma > 0$  (a location-scale family).

Let  $\theta = (\mu, \sigma)$ . Then, the Fisher information about  $\theta$  contained in  $X_1, \dots, X_n$  is (exercise)

$$I(\theta) = \frac{n}{\sigma^2} \begin{pmatrix} \mathbb{E} \int \frac{[f'(x)]^2}{f(x)} dx & \int \frac{f'(x)[xf'(x)+f(x)]}{f(x)} dx \\ \int \frac{f'(x)[xf'(x)+f(x)]}{f(x)} dx & \int \frac{[xf'(x)+f(x)]^2}{f(x)} dx \end{pmatrix}.$$

## Proof

Result (i) follows from the independence of  $X$  and  $Y$  and the definition of the Fisher information.

Result (ii) follows from the equality

$$\frac{\partial^2}{\partial \theta \partial \theta^\tau} \log f_\theta(X) = \frac{\frac{\partial^2}{\partial \theta \partial \theta^\tau} f_\theta(X)}{f_\theta(X)} - \frac{\partial}{\partial \theta} \log f_\theta(X) \left[ \frac{\partial}{\partial \theta} \log f_\theta(X) \right]^\tau.$$

## Example 3.9

Let  $X_1, \dots, X_n$  be i.i.d. with the Lebesgue p.d.f.  $\frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right)$ , where  $f(x) > 0$  and  $f'(x)$  exists for all  $x \in \mathcal{R}$ ,  $\mu \in \mathcal{R}$ , and  $\sigma > 0$  (a location-scale family).

Let  $\theta = (\mu, \sigma)$ . Then, the Fisher information about  $\theta$  contained in  $X_1, \dots, X_n$  is (exercise)

$$I(\theta) = \frac{n}{\sigma^2} \begin{pmatrix} \mathbb{C} \int \frac{[f'(x)]^2}{f(x)} dx & \int \frac{f'(x)[xf'(x)+f(x)]}{f(x)} dx \\ \int \frac{f'(x)[xf'(x)+f(x)]}{f(x)} dx & \int \frac{[xf'(x)+f(x)]^2}{f(x)} dx \end{pmatrix}.$$

## Remarks

- Note that  $I(\theta)$  depends on the particular parameterization.
- If  $\theta = \psi(\eta)$  and  $\psi$  is differentiable, then the Fisher information that  $X$  contains about  $\eta$  is

$$\frac{\partial}{\partial \eta} \psi(\eta) I(\psi(\eta)) \left[ \frac{\partial}{\partial \eta} \psi(\eta) \right]^{\tau}.$$

- However, the Cramér-Rao lower bound in (2) is not affected by any one-to-one reparameterization.
- If we use inequality (2) to find a UMVUE  $T(X)$ , then we obtain a formula for  $\text{Var}(T(X))$  at the same time.
- On the other hand, the Cramér-Rao lower bound in (2) is typically not sharp.
- Under some regularity conditions, the Cramér-Rao lower bound is attained iff  $f_{\theta}$  is in an exponential family; see Propositions 3.2 and 3.3 and the discussion in Lehmann (1983, p. 123).
- Some improved information inequalities are available (see, e.g., Lehmann (1983, Sections 2.6 and 2.7)).

## Proposition 3.2.

Suppose that the distribution of  $X$  is from an exponential family  $\{f_\theta : \theta \in \Theta\}$ , i.e., the p.d.f. of  $X$  w.r.t. a  $\sigma$ -finite measure is

$$f_\theta(x) = \exp\{[\eta(\theta)]^\tau T(x) - \xi(\theta)\} c(x), \quad (3)$$

where  $\Theta$  is an open subset of  $\mathcal{R}^k$ .

- (i) The regularity condition (1) is satisfied for any  $h$  with  $E|h(X)| < \infty$  and

$$I(\theta) = -E \left[ \frac{\partial^2}{\partial \theta \partial \theta^\tau} \log f_\theta(X) \right].$$

- (ii) If  $\underline{I}(\eta)$  is the Fisher information matrix for the natural parameter  $\eta$ , then the variance-covariance matrix  $\text{Var}(T) = \underline{I}(\eta)$ .
- (iii) If  $\bar{I}(\vartheta)$  is the Fisher information matrix for the parameter  $\vartheta = E[T(X)]$ , then  $\text{Var}(T) = [\bar{I}(\vartheta)]^{-1}$ .

## Proof

- (i) This is a direct consequence of Theorem 2.1.
- (ii) The p.d.f. under the natural parameter  $\eta$  is

$$f_{\eta}(x) = \exp \{ \eta^{\tau} T(x) - \zeta(\eta) \} c(x).$$

From Theorem 2.1,  $E[T(X)] = \frac{\partial}{\partial \eta} \zeta(\eta)$ .

The result follows from

$$\frac{\partial}{\partial \eta} \log f_{\eta}(x) = T(x) - \frac{\partial}{\partial \eta} \zeta(\eta).$$

- (iii) Since  $\vartheta = E[T(X)] = \frac{\partial}{\partial \eta} \zeta(\eta)$ ,

$$\underline{l}(\eta) = \frac{\partial \vartheta}{\partial \eta} \bar{l}(\vartheta) \left( \frac{\partial \vartheta}{\partial \eta} \right)^{\tau} = \frac{\partial^2}{\partial \eta \partial \eta^{\tau}} \zeta(\eta) \bar{l}(\vartheta) \left[ \frac{\partial^2}{\partial \eta \partial \eta^{\tau}} \zeta(\eta) \right]^{\tau}.$$

By Theorem 2.1 and the result in (ii),

$$\frac{\partial^2}{\partial \eta \partial \eta^{\tau}} \zeta(\eta) = \text{Var}(T) = \underline{l}(\eta).$$

Hence

$$\bar{l}(\vartheta) = [\underline{l}(\eta)]^{-1} \underline{l}(\eta) [\underline{l}(\eta)]^{-1} = [\underline{l}(\eta)]^{-1} = [\text{Var}(T)]^{-1}.$$

A direct consequence of Proposition 3.2(ii) is that the variance of any linear function of  $T$  in (3) attains the Cramér-Rao lower bound.

The following result gives a necessary condition for  $\text{Var}(U(X))$  of an estimator  $U(X)$  to attain the Cramér-Rao lower bound.

### Proposition 3.3

Assume that the conditions in Theorem 3.3 hold with  $T(X)$  replaced by  $U(X)$  and that  $\Theta \subset \mathcal{R}$ .

(i) If  $\text{Var}(U(X))$  attains the Cramér-Rao lower bound in (2), then

$$a(\theta)[U(X) - g(\theta)] = g'(\theta) \frac{\partial}{\partial \theta} \log f_{\theta}(X) \quad \text{a.s. } P_{\theta}$$

for some function  $a(\theta)$ ,  $\theta \in \Theta$ .

(ii) Let  $f_{\theta}$  and  $T$  be given by (3).

If  $\text{Var}(U(X))$  attains the Cramér-Rao lower bound, then  $U(X)$  is a linear function of  $T(X)$  a.s.  $P_{\theta}$ ,  $\theta \in \Theta$ .

A direct consequence of Proposition 3.2(ii) is that the variance of any linear function of  $T$  in (3) attains the Cramér-Rao lower bound.

The following result gives a necessary condition for  $\text{Var}(U(X))$  of an estimator  $U(X)$  to attain the Cramér-Rao lower bound.

### Proposition 3.3

Assume that the conditions in Theorem 3.3 hold with  $T(X)$  replaced by  $U(X)$  and that  $\Theta \subset \mathcal{R}$ .

(i) If  $\text{Var}(U(X))$  attains the Cramér-Rao lower bound in (2), then

$$a(\theta)[U(X) - g(\theta)] = g'(\theta) \frac{\partial}{\partial \theta} \log f_{\theta}(X) \quad \text{a.s. } P_{\theta}$$

for some function  $a(\theta)$ ,  $\theta \in \Theta$ .

(ii) Let  $f_{\theta}$  and  $T$  be given by (3).

If  $\text{Var}(U(X))$  attains the Cramér-Rao lower bound, then  $U(X)$  is a linear function of  $T(X)$  a.s.  $P_{\theta}$ ,  $\theta \in \Theta$ .

## Example 3.10

Let  $X_1, \dots, X_n$  be i.i.d. from the  $N(\mu, \sigma^2)$  distribution with an unknown  $\mu \in \mathcal{R}$  and a known  $\sigma^2$ .

Let  $f_\mu$  be the joint distribution of  $X = (X_1, \dots, X_n)$ .

Then

$$\frac{\partial}{\partial \mu} \log f_\mu(X) = \sum_{i=1}^n (X_i - \mu) / \sigma^2.$$

Thus,  $I(\mu) = n/\sigma^2$ .

Consider the estimation of  $\mu$ .

It is obvious that  $\text{Var}(\bar{X})$  attains the Cramér-Rao lower bound in (2).

Consider now the estimation of  $\vartheta = \mu^2$ .

Since  $E\bar{X}^2 = \mu^2 + \sigma^2/n$ , the UMVUE of  $\vartheta$  is  $h(\bar{X}) = \bar{X}^2 - \sigma^2/n$ .

A straightforward calculation shows that

$$\text{Var}(h(\bar{X})) = \frac{4\mu^2\sigma^2}{n} + \frac{2\sigma^4}{n^2}.$$

## Example 3.10

Let  $X_1, \dots, X_n$  be i.i.d. from the  $N(\mu, \sigma^2)$  distribution with an unknown  $\mu \in \mathcal{R}$  and a known  $\sigma^2$ .

Let  $f_\mu$  be the joint distribution of  $X = (X_1, \dots, X_n)$ .

Then

$$\frac{\partial}{\partial \mu} \log f_\mu(X) = \sum_{i=1}^n (X_i - \mu) / \sigma^2.$$

Thus,  $I(\mu) = n/\sigma^2$ .

Consider the estimation of  $\mu$ .

It is obvious that  $\text{Var}(\bar{X})$  attains the Cramér-Rao lower bound in (2).

Consider now the estimation of  $\vartheta = \mu^2$ .

Since  $E\bar{X}^2 = \mu^2 + \sigma^2/n$ , the UMVUE of  $\vartheta$  is  $h(\bar{X}) = \bar{X}^2 - \sigma^2/n$ .

A straightforward calculation shows that

$$\text{Var}(h(\bar{X})) = \frac{4\mu^2\sigma^2}{n} + \frac{2\sigma^4}{n^2}.$$

## Example 3.10

Let  $X_1, \dots, X_n$  be i.i.d. from the  $N(\mu, \sigma^2)$  distribution with an unknown  $\mu \in \mathcal{R}$  and a known  $\sigma^2$ .

Let  $f_\mu$  be the joint distribution of  $X = (X_1, \dots, X_n)$ .

Then

$$\frac{\partial}{\partial \mu} \log f_\mu(X) = \sum_{i=1}^n (X_i - \mu) / \sigma^2.$$

Thus,  $I(\mu) = n/\sigma^2$ .

Consider the estimation of  $\mu$ .

It is obvious that  $\text{Var}(\bar{X})$  attains the Cramér-Rao lower bound in (2).

Consider now the estimation of  $\vartheta = \mu^2$ .

Since  $E\bar{X}^2 = \mu^2 + \sigma^2/n$ , the UMVUE of  $\vartheta$  is  $h(\bar{X}) = \bar{X}^2 - \sigma^2/n$ .

A straightforward calculation shows that

$$\text{Var}(h(\bar{X})) = \frac{4\mu^2\sigma^2}{n} + \frac{2\sigma^4}{n^2}.$$

## Example 3.10 (continued)

On the other hand, the Cramér-Rao lower bound in this case is  $4\mu^2\sigma^2/n$ .

Hence  $\text{Var}(h(\bar{X}))$  does not attain the Cramér-Rao lower bound. The difference is  $2\sigma^4/n^2$ .

## Remarks

- Condition (1) is a key regularity condition for the results in Theorem 3.3 and Proposition 3.3.
- If  $f_\theta$  is not in an exponential family, then (1) has to be checked.
- Typically, it does not hold if the set  $\{x : f_\theta(x) > 0\}$  depends on  $\theta$  (Exercise 37).
- More discussions can be found in Pitman (1979).

## Example 3.10 (continued)

On the other hand, the Cramér-Rao lower bound in this case is  $4\mu^2\sigma^2/n$ .

Hence  $\text{Var}(h(\bar{X}))$  does not attain the Cramér-Rao lower bound. The difference is  $2\sigma^4/n^2$ .

## Remarks

- Condition (1) is a key regularity condition for the results in Theorem 3.3 and Proposition 3.3.
- If  $f_\theta$  is not in an exponential family, then (1) has to be checked.
- Typically, it does not hold if the set  $\{x : f_\theta(x) > 0\}$  depends on  $\theta$  (Exercise 37).
- More discussions can be found in Pitman (1979).