Lecture 3: Unbiasedness, UMVUE, and Cramer-Rao information inequality

We now focue on the MSE criterion.

As we discussed earlier, another criterion is needed for the purpose of finding an estimator of certain optimal property.

The Bayes risk is one example.

In the next few lectures, we will study the unbiasedness criterion.

Definition 7.3.2.

The bias of an estimator $T(X)$ of $g(\theta)$ is the function of θ defined by $E_\theta[T(X)] - g(\theta)$. An estimator $T(X)$ of $g(\theta)$ is unbiased if its bias is 0, i.e., $E_\theta[T(X)] = g(\theta)$ for all $\theta \in \Theta$.

An unbiased estimator can be thought of an estimator that has no systematic estimation error: the center of the distribution of $T(X)$ is what we want to estimate.

Note that the MSE of an estimator T is equal to

$$\begin{aligned} E_\theta[T(X)-g(\theta)]^2 &= E_\theta[T-E_\theta(T)]^2+[E_\theta(T)-g(\theta)]^2 \\ &= \mathrm{Var}_\theta(T)+[\text{the bias of } T(X)]^2 \end{aligned}$$

Hence, an estimator T is unbiased iff MSE = the variance of T.

Example 7.3.3.

Let $X_1,...,X_n$ be iid from $N(\mu,\sigma^2)$ with unknown $\theta=(\mu,\sigma^2)$.
From Theorem 5.2.6, $\bar{X}$ is an unbiased estimator of μ and S^2 is an unbiased estimator of σ^2.
Since $\bar{X}$ is normally distributed and $(n-1)S^2/\sigma^2$ is chi-square, the MSE's are

$$E_\theta(\bar{X}-\mu)^2=\mathrm{Var}_\theta(\bar{X})=\frac{\sigma^2}{n} \qquad E_\theta(S^2-\sigma^2)^2=\mathrm{Var}_\theta(S^2)=\frac{2\sigma^4}{n-1}$$

The unbiasedness of $\bar{X}$ and S^2 and the MSE of $\bar{X}$ remain the same if the normality assumption is dropped, whereas the MSE of S^2 is not the same if the normality assumption is dropped (Exercise 5.8).
The sample standard deviation, $S=\sqrt{S^2}$, is not an unbiased estimator of the population standard deviation σ.

Under the normality assumption, since $(n-1)S^2/\sigma^2$ is chi-square with degrees of freedom $n-1$,

$$\frac{\sqrt{n-1}}{\sigma}E_\theta(S) = E_\theta\left(\frac{\sqrt{n-1}S}{\sigma}\right) = \frac{1}{\Gamma(\frac{n-1}{2})2^{\frac{n-1}{2}}}\int_0^\infty \sqrt{x}x^{\frac{n-1}{2}-1}e^{-x/2}dx$$

Using the fact that $\sqrt{x}x^{\frac{n-1}{2}-1}e^{-x/2} = x^{\frac{n}{2}-1}e^{-x/2}$ is the kernel for the chi-square with degrees of freedom n, we obtain that

$$\frac{\sqrt{n-1}}{\sigma}E_\theta(S) = \frac{\Gamma(\frac{n}{2})2^{\frac{n}{2}}}{\Gamma(\frac{n-1}{2})2^{\frac{n-1}{2}}} = \frac{\sqrt{2}\Gamma(\frac{n}{2})}{\Gamma(\frac{n-1}{2})}$$

i.e.,

$$E_\theta(S) = \frac{\sqrt{2}\Gamma(\frac{n}{2})}{\sqrt{n-1}\,\Gamma(\frac{n-1}{2})}\sigma = k_{n-1}\sigma$$

For example, $E_\theta(S) = \frac{\sqrt{2}}{\sqrt{\pi}}\sigma$ when $n=2$, and $E_\theta(S) = \frac{\sqrt{\pi}}{2}\sigma$ when $n=3$.

Since $[E_\theta(S)]^2 < E_\theta(S^2) = \sigma^2$,

$$k_n < 1 \quad \text{for all } n = 2,3,...$$

i.e., $E_\theta(S) < \sigma$ is always true and S underestimates σ with a negative bias $E_\theta(S) - \sigma = (k_{n-1} - 1)\sigma$.

What is the MSE of S as an estimator of σ?

$$E_\theta(S-\sigma)^2 = \mathrm{Var}_\theta(S) + [(k_{n-1}-1)\sigma]^2 = E_\theta(S^2) - [E_\theta(S)]^2 + [(k_{n-1}-1)\sigma]^2$$
$$= \sigma^2 - k_{n-1}^2\sigma^2 + (k_{n-1}-1)^2\sigma^2 = 2(1-k_{n-1})\sigma^2$$

There is a trade-off between variance and bias.
We want an estimator having small MSE (for that purpose, sometimes we give up exact unbiasedness), but we do not want a systematic error trend (such as always underestimate or always overestimate).

Example 7.3.4.

Let $X_1, ..., X_n$ be iid from $N(\mu, \sigma^2)$ with unknown $\theta = (\mu, \sigma^2)$.

The MLE of σ^2 is $\widehat{\sigma}^2 = n^{-1}\sum_{i=1}^n (X_i - \bar{X})^2 = \frac{n-1}{n} S^2$.

We now compare the biases and MSE's of $\widehat{\sigma}^2$ and S^2.

$$E_\theta(\widehat{\sigma}^2) = \frac{n-1}{n} E_\theta(S^2) = \frac{n-1}{n}\sigma^2$$

i.e., the bias of $\widehat{\sigma}^2$ is $-\sigma^2/n$.

$$\mathrm{Var}_\theta(\widehat{\sigma}^2) = \frac{(n-1)^2}{n^2}\mathrm{Var}_\theta(S^2) = \frac{2(n-1)\sigma^4}{n^2}$$

The MSE of $\widehat{\sigma}^2$ is then

$$\begin{aligned} E_\theta(\widehat{\sigma}^2 - \sigma^2)^2 &= \mathrm{Var}_\theta(\widehat{\sigma}^2) + \frac{\sigma^4}{n^2} \\ &= \frac{2(n-1)\sigma^4}{n^2} + \frac{\sigma^4}{n^2} = \frac{2n-1}{n^2}\sigma^2 \\ &< \frac{2}{n-1}\sigma^4 = \mathrm{Var}_\theta(S^2) \\ &= E_\theta(S^2 - \sigma^2)^2 \end{aligned}$$

That is, the MSE of $\widehat{\sigma}^2$ is always smaller than that of S^2, although $\widehat{\sigma}^2$ is biased and S^2 is unbiased.

In this example we don't want to conclude that $\widehat{\sigma}^2$ is better than S^2.

First, $\widehat{\sigma}^2$ always underestimates.

Second, it can be argued that the MSE, while a reasonable criterion for location parameters, is not so reasonable for scale parameters; the MSE penalizes equally for overestimation and underestimation, which is fine in the location case but may not be good for the non-symmetric problem of estimating a scale parameter.

How to find unbiased estimators

- Try some simple statistics; e.g, the sample moment m_j is unbiased for the population moment μ_j for any j; try the order statistics.
- Note that a linear function of an unbiased estimator is unbiased for the same linear function of the parameter. However, nonlinear functions of unbiased estimators are no longer unbiased.
- Let $X_1, ..., X_n$ be iid random variables from a cdf F and $t \in \mathscr{R}$. Let $I(X \leq t) = 1$ if $X \leq t$ and $I(X \leq t) = 0$ otherwise. Then an unbiased estimator of $F(t)$ is

$$\widehat{F}_n(t) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \leq t)$$

which is the empirical cdf based on $X_1, ..., X_n$.

- Let m be a fixed positive integer, $h(x_1, ..., x_m)$ be a symmetric function, and $X_1, ..., X_n$ be iid.
  Let $g(\theta) = E_\theta[h(X_1, ..., X_m)]$ be the parameter of interest.
  Then an unbiased estimator of θ is the so called U-statistic

$$U(X) = \binom{n}{m}^{-1} \sum_c h(X_{i_1}, ..., X_{i_m}),$$

  where $\sum_c$ denotes the summation over the $\binom{n}{m}$ combinations of m distinct elements $\{i_1, ..., i_m\}$ from $\{1, ..., n\}$.

  When $m = 1$, this is a kind of sample mean.

  For example, if we want to estimate μ^k, $\mu = E(X_1)$, then $h(x_1, ..., x_k) = x_1 \cdots x_k$ and $E_\theta(X_1 \cdots X_k) = \mu^k$.

  The sample variance is a U-statistic with $h(x_1, x_2) = (x_1 - x_2)^2/2$, $m = 2$:

$$U(X) = \frac{2}{n(n-1)} \sum_{i<j} \frac{(X_i - X_j)^2}{2} = \frac{1}{n(n-1)} \sum_{i \neq j} \frac{(X_i - X_j)^2}{2}$$

$$= \frac{1}{2n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} (X_i - X_j)^2$$

$$= \frac{1}{2n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} (X_i^2 + X_j^2 - 2X_i X_j)$$

$$= \frac{1}{2n(n-1)} \left(\sum_{i=1}^{n} \sum_{j=1}^{n} X_i^2 + \sum_{i=1}^{n} \sum_{j=1}^{n} X_j^2 - 2 \sum_{i=1}^{n} \sum_{j=1}^{n} X_i X_j \right)$$

$$= \frac{1}{n(n-1)} \left(n \sum_{i=1}^{n} X_i^2 - \sum_{i=1}^{n} X_i \sum_{j=1}^{n} X_j \right)$$

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

The UMVUE

We now try to find the best unbiased estimator in a given problem.

The MSE of an unbiased estimator is simply its variance.

Hence, the best unbiased estimator can be defined as follows.

Definition 7.3.7.

An estimator is a best unbiased estimator or a uniform minimum variance unbiased estimator (UMVUE) of $g(\theta)$ if it has the smallest variance (and hence the MSE) among all unbiased estimators of $g(\theta)$.

Finding a best unbiased estimator is not easy.
There are three main approaches.

Approach 1: using a lower bound.

If we can show that, for any unbiased estimator $T(X)$ of $g(\theta)$,

$$\mathrm{Var}_\theta(T(X)) \geq b(\theta) \qquad \text{all } \theta \in \Theta$$

and we can find an estimator $T^*(X)$ that achieves the variance lower bound, i.e., $\mathrm{Var}_\theta(T^*) = b(\theta)$ for all $\theta \in \Theta$, then T^* is a UMVUE.

Theorem 7.3.9 (Cramér-Rao inequality)

Let $\theta \in \Theta \subset \mathscr{R}$ and X be a sample with joint pdf or pmf $f_\theta(x)$ satisfying

$$\int_{\mathscr{X}} \frac{\partial f_\theta(x)}{\partial \theta} dx = \frac{\partial}{\partial \theta} \int_{\mathscr{X}} f_\theta(x) dx = 0 \qquad \theta \in \Theta$$

Let $T(X)$ be any estimator satisfying $\mathrm{Var}_\theta(T) < \infty$ and

$$\frac{\partial}{\partial\theta} E_\theta(T) = \int_{\mathscr{X}} T(x) \frac{\partial f_\theta(x)}{\partial\theta} dx \qquad \theta \in \Theta$$

Then,

$$\mathrm{Var}_\theta(T) \geq \frac{[\frac{\partial}{\partial\theta} E_\theta(T)]^2}{E_\theta\left[\frac{\partial}{\partial\theta} \log f_\theta(X)\right]^2} \qquad \theta \in \Theta$$

Proof.

Consider the Cauchy-Schwartz inequality,

$$\mathrm{Var}(Z) \geq \frac{[\mathrm{Cov}(Z, Y)]^2}{\mathrm{Var}(Y)}$$

Let $Z = T(X)$ and $Y = \frac{\partial}{\partial\theta} \log f_\theta(X)$.

$$E_\theta(Y) = E_\theta\left[\frac{\partial}{\partial\theta} \log f_\theta(X)\right] = \int_{\mathscr{X}} \frac{\partial f_\theta(x)}{\partial\theta} dx = 0$$

by the condition on $f_\theta(x)$.

$$E_\theta\left[\frac{\partial}{\partial\theta}\log f_\theta(X)\right]^2 = E_\theta(Y^2) = \mathrm{Var}_\theta(Y)$$

It remains to show that $\mathrm{Cov}_\theta(Z,Y) = E_\theta(ZY) = \frac{\partial}{\partial\theta}E_\theta(T)$.

From the condition on T,

$$\begin{aligned} E_\theta(ZY) &= E_\theta\left[T\frac{\partial}{\partial\theta}\log f_\theta(X)\right] = \int_{\mathscr{X}} T(x)\frac{\partial}{\partial\theta}f_\theta(x)dx \\ &= \frac{\partial}{\partial\theta}\int_{\mathscr{X}} T(x)f_\theta(x)dx = \frac{\partial}{\partial\theta}E_\theta(T) \end{aligned}$$

- There is a multivariate extension of Theorem 7.3.9: If $\theta \in \mathscr{R}^k$, then

$$\mathrm{Var}_\theta(T) \geq \left[\frac{\partial}{\partial\theta}E_\theta(T)\right]' [I(\theta)]^{-1} \frac{\partial}{\partial\theta}E_\theta(T)$$

  where

$$I(\theta) = E_\theta\left\{\frac{\partial}{\partial\theta}\log f_\theta(X)\left[\frac{\partial}{\partial\theta}\log f_\theta(X)\right]'\right\}$$

- The $k \times k$ matrix $I(\theta)$ is called the *Fisher information matrix*. The greater $I(\theta)$ is, the easier it is to distinguish θ from neighboring values, and the more accurately θ can be estimated.

- $I(\theta)$ is a measure of the information about θ contained in X.
- The Cramér-Rao inequality is one type of information inequality.
- If X and Y are independent with the Fisher information matrices $I_X(\theta)$ and $I_Y(\theta)$, respectively, then the Fisher information about θ contained in (X, Y) is $I_X(\theta) + I_Y(\theta)$.
  In particular, if $X_1, ..., X_n$ are iid and $I_1(\theta)$ is the Fisher information about θ contained in a single X_i, then the Fisher information about θ contained in $X_1, ..., X_n$ is $nI_1(\theta)$.
- Note that $I(\theta)$ depends on the particular parameterization.
  If $\theta = \psi(\eta)$ and ψ is differentiable, then the Fisher information about η contained in X is
$$\frac{\partial}{\partial \eta}\psi(\eta) I(\psi(\eta)) \left[\frac{\partial}{\partial \eta}\psi(\eta)\right]'.$$
- If T is an unbiased estimator of $g(\theta)$, then the extended Theorem 7.3.9 says that
$$\mathrm{Var}_\theta(T) \geq \left[\frac{\partial g(\theta)}{\partial \theta}\right]' [I(\theta)]^{-1} \frac{\partial g(\theta)}{\partial \theta}$$
  If the equality holds for T for all $\theta \in \Theta$, then T is a UMVUE of $g(\theta)$.

The following lemma may simplify the calculation of the matrix $I(\theta)$.

Lemma 7.3.11.

If $f_\theta(x)$ is twice differentiable in θ,

$$\frac{\partial}{\partial\theta}E_\theta\left[\frac{\partial}{\partial\theta}\log f_\theta(X)\right]=\int_{\mathscr{X}}\frac{\partial}{\partial\theta}\left[\left(\frac{\partial}{\partial\theta}\log f_\theta(x)\right)f_\theta(x)\right]dx,$$

and the condition on f_θ in Theorem 7.3.9 holds, then

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

Proof.

Under the conditions,

$$E_\theta\left[\frac{\frac{\partial^2}{\partial\theta\partial\theta'}f_\theta(X)}{f_\theta(X)}\right]=\int_{\mathscr{X}}\frac{\partial^2}{\partial\theta\partial\theta'}f_\theta(x)dx=\frac{\partial^2}{\partial\theta\partial\theta}\int_{\mathscr{X}}f_\theta(x)dx=0$$

Then the result follows from

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

The conditions in Theorem 7.3.9 and Lemma 7.3.11 are satisfied when $f_\theta(x)$ is from an exponential family.

Example 7.3.14.

Let $X_1, ..., X_n$ be iid from $N(\mu, \sigma^2)$, $\theta \in (\mu, \sigma^2) \in \mathscr{R} \times (0, \infty)$.
If $f_\theta(x_i)$ is the pdf of X_i, then

$$\log f_\theta(x_i) = -\frac{(x_i-\mu)^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)$$

$$\frac{\partial \log f_\theta(x_i)}{\partial \mu} = \frac{x_i-\mu}{\sigma^2} \qquad \frac{\partial \log f_\theta(x_i)}{\partial \sigma^2} = -\frac{(x_i-\mu)^2}{2\sigma^4} - \frac{1}{2\sigma^2}$$

$$\frac{\partial^2 \log f_\theta(x_i)}{\partial \mu^2} = -\frac{1}{\sigma^2} \qquad \frac{\partial^2 \log f_\theta(x_i)}{\partial \sigma^4} = -\frac{(x_i-\mu)^2}{\sigma^6} + \frac{1}{2\sigma^4}$$

$$\frac{\partial^2 \log f_\theta(x_i)}{\partial \mu \partial \sigma^2} = -\frac{x_i-\mu}{\sigma^4} \qquad E_\theta\left[\frac{\partial^2 \log f_\theta(X_i)}{\partial \mu \partial \sigma^2}\right] = -\frac{E_\theta(X_i-\mu)}{\sigma^4} = 0$$

$$E_\theta\left[\frac{\partial^2 \log f_\theta(X_i)}{\partial \sigma^4}\right] = -\frac{E_\theta(X_i-\mu)^2}{\sigma^6} + \frac{1}{2\sigma^4} = -\frac{1}{\sigma^2}\sigma^6 + \frac{1}{2\sigma^4} = -\frac{1}{2\sigma^4}$$

Hence, the Fisher information contained in X_i is

$$I_1 = \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{bmatrix}$$

The Fisher information contained in $X = (X_1, ..., X_n)$ is $I_n(\theta) = nI_1(\theta)$

$$[I_n(\theta)]^{-1} = \begin{bmatrix} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{2\sigma^4}{n} \end{bmatrix}$$

Since $\bar{X}$ is unbiased for μ, the first component of θ, and $\mathrm{Var}_\theta(\bar{X}) = \sigma^2/n$, which equals the first diagonal element of $[I_n(\theta)]^{-1}$, i.e., the equality in the Cramér-Rao inequality holds.

Hence, $\bar{X}$ is the UMVUE.

Note that S^2 is unbiased for σ^2.

Since it has a chi-square distribuition, $\mathrm{Var}_\theta(S^2) = 2\sigma^4/(n-1)$, which is larger than the second diagonal element in $[I_n(\theta)]^{-1}$.

We cannot conclude that S^2 is a UMVUE using this approach.