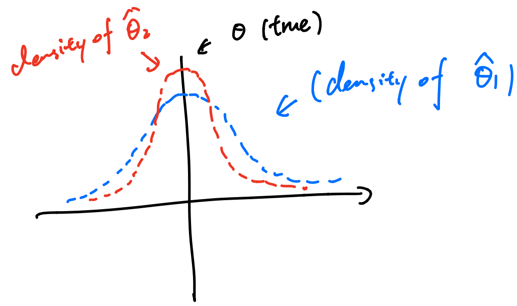
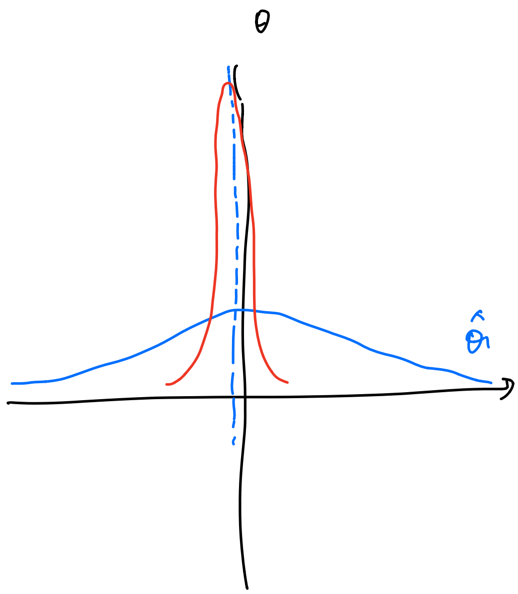
Feb 19, 2019

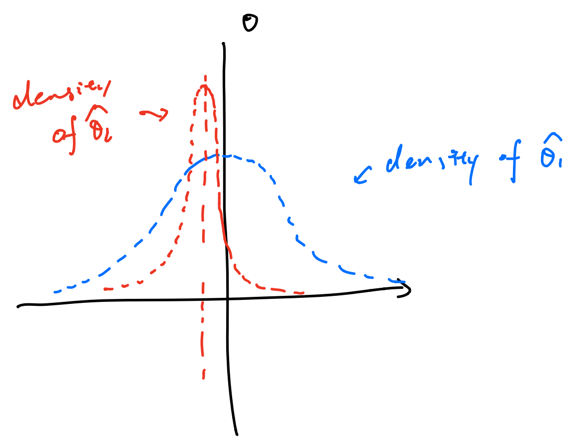
- Method of Moments 1
- MLE 2
- Bayesian $\longrightarrow$ Posterior $\begin{cases} \text{Mean} & 3 \\ \text{Median} & 4 \\ \text{Mode} & 5 \end{cases}$

θ $\hat{\theta}_1$ $\hat{\theta}_2$

$\cancel{\hat{\theta}_1 = \bar{X}_n \sim N(\mu, \frac{\sigma_0^2}{n})}$







Example 7.12. Suppose $X_1, X_2, ..., X_n$ are iid $\mathcal{N}(\mu, \sigma^2)$, where $-\infty < \mu < \infty$ and $\sigma^2 > 0$.

- If σ^2 is known, a conjugate prior for μ is

$$\mu \sim \mathcal{N}(\xi, \tau^2), \qquad \xi, \tau^2 \text{ known.}$$

- If μ is known, a conjugate prior for σ^2 is

$$\sigma^2 \sim \text{IG}(a, b) \qquad a, b \text{ known.}$$

7.3 Methods of Evaluating Estimators

7.3.1 Bias, variance, and MSE

Definition: Suppose $W = W(\mathbf{X})$ is a point estimator. We call W an **unbiased estimator** of θ if

$$E_\theta(W) = \theta \quad \text{for all } \theta \in \Theta.$$

$$\boxed{E_\theta(W) - \theta}$$

↳ bias

More generally, we call W an unbiased estimator of $\tau(\theta)$ if

$$E_\theta(W) = \tau(\theta) \quad \text{for all } \theta \in \Theta.$$

Definition: The **mean-squared error** (**MSE**) of a point estimator $W = W(\mathbf{X})$ is

$$\begin{aligned} \text{MSE}_\theta(W) &= E_\theta[(W - \theta)^2] \\ &= \text{var}_\theta(W) + [E_\theta(W) - \theta]^2 \\ &= \text{var}_\theta(W) + \text{Bias}_\theta^2(W), \end{aligned}$$

$$\begin{aligned} &= E_\theta\left[\left(W - E_\theta(W) + E_\theta(W) - \theta\right)^2\right] \\ &= E_\theta\left[(W - E_\theta(W))^2 + (E_\theta(W) - \theta)^2 + 2(W - E_\theta(W))(E_\theta(W) - \theta)\right] \\ &= E_\theta\left[(W - E_\theta(W))^2\right] + E_\theta\left[(E_\theta(W) - \theta)^2\right] + 2E_\theta\left[(W - E_\theta(W))(E_\theta(W) - \theta)\right] \end{aligned}$$

where $\text{Bias}_\theta(W) = E_\theta(W) - \theta$ is the **bias** of W as an estimator of θ. Note that if W is an unbiased estimator of θ, then for all $\theta \in \Theta$,

$$E_\theta(W) = \theta \implies \text{Bias}_\theta(W) = E_\theta(W) - \theta = 0.$$

In this case,

$$\text{MSE}_\theta(W) = \text{var}_\theta(W).$$

Remark: In general, the MSE incorporates two components:

- $\text{var}_\theta(W)$; this measures **precision**
- $\text{Bias}_\theta(W)$; this measures **accuracy**.

Obviously, we prefer estimators with small MSE because these estimators have small bias (i.e., high accuracy) and small variance (i.e., high precision).

Example 7.13. Suppose $X_1, X_2, ..., X_n$ are iid $\mathcal{N}(\mu, \sigma^2)$, where $-\infty < \mu < \infty$ and $\sigma^2 > 0$; i.e., both parameters unknown. Set $\boldsymbol{\theta} = (\mu, \sigma^2)$. Recall that our "usual" sample variance estimator is

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

$$MSE(S^2) = \frac{2}{n-1}\sigma^4$$

and for all $\boldsymbol{\theta}$,

$$\begin{aligned} E_{\boldsymbol{\theta}}(S^2) &= \sigma^2 \\ \text{var}_{\boldsymbol{\theta}}(S^2) &= \frac{2\sigma^4}{n-1}. \end{aligned}$$

Consider the "competing estimator:"

$$MSE(S_b^2) = \frac{2n-1}{n^2}\sigma^4$$

$$S_b^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2,$$

which recall is the MOM and MLE of σ^2.

$$E_\theta\left[\frac{n}{n-1} S_b^2\right] = \sigma^2$$

Note that

$$S_b^2 = \left(\frac{n-1}{n}\right)S^2 \implies E_{\boldsymbol{\theta}}(S_b^2) = E_{\boldsymbol{\theta}}\left[\left(\frac{n-1}{n}\right)S^2\right] = \left(\frac{n-1}{n}\right)E_{\boldsymbol{\theta}}(S^2) = \left(\frac{n-1}{n}\right)\sigma^2.$$

That is, the estimator S_b^2 is biased; it **underestimates** σ^2 on average.

Comparison: Let's compare S^2 and S_b^2 on the basis of MSE. Because S^2 is an unbiased estimator of σ^2,

$$\text{MSE}_{\boldsymbol{\theta}}(S^2) = \text{var}_{\boldsymbol{\theta}}(S^2) = \frac{2\sigma^4}{n-1}.$$

The MSE of S_b^2 is

$$\text{MSE}_{\boldsymbol{\theta}}(S_b^2) = \text{var}_{\boldsymbol{\theta}}(S_b^2) + \text{Bias}^2_{\boldsymbol{\theta}}(S_b^2).$$

The variance of S_b^2 is

$$\begin{aligned} \text{var}_{\boldsymbol{\theta}}(S_b^2) &= \text{var}_{\boldsymbol{\theta}}\left[\left(\frac{n-1}{n}\right)S^2\right] \\ &= \left(\frac{n-1}{n}\right)^2 \text{var}_{\boldsymbol{\theta}}(S^2) = \left(\frac{n-1}{n}\right)^2 \frac{2\sigma^4}{n-1} = \frac{2(n-1)\sigma^4}{n^2}. \end{aligned}$$

The bias of S_b^2 is

$$\text{E}_{\boldsymbol{\theta}}(S_b^2 - \sigma^2) = \text{E}_{\boldsymbol{\theta}}(S_b^2) - \sigma^2 = \left(\frac{n-1}{n}\right)\sigma^2 - \sigma^2.$$

Therefore,

$$\text{MSE}_{\boldsymbol{\theta}}(S_b^2) = \underbrace{\frac{2(n-1)\sigma^4}{n^2}}_{\text{var}_{\boldsymbol{\theta}}(S_b^2)} + \underbrace{\left[\left(\frac{n-1}{n}\right)\sigma^2 - \sigma^2\right]^2}_{\text{Bias}^2_{\boldsymbol{\theta}}(S_b^2)} = \left(\frac{2n-1}{n^2}\right)\sigma^4.$$

Finally, to compare $\text{MSE}_{\boldsymbol{\theta}}(S^2)$ with $\text{MSE}_{\boldsymbol{\theta}}(S_b^2)$, we are left to compare the constants

$$\frac{2}{n-1} \quad \text{and} \quad \frac{2n-1}{n^2}.$$

Note that the ratio

$$\frac{\frac{2n-1}{n^2}}{\frac{2}{n-1}} = \frac{2n^2-3n+1}{2n^2} < 1,$$

for all $n \geq 2$. Therefore,

$$\text{MSE}_{\boldsymbol{\theta}}(S_b^2) < \text{MSE}_{\boldsymbol{\theta}}(S^2),$$

showing that S_b^2 is a "better" estimator than S^2 on the basis of MSE.

Discussion: In general, how should we **compare** two competing estimators W_1 and W_2?

- If both W_1 and W_2 are unbiased, we prefer the estimator with the smaller variance.
- If either W_1 or W_2 is biased (or perhaps both are biased), we prefer the estimator with the smaller MSE.

There is no guarantee that one estimator, say W_1, will **always** beat the other for all $\theta \in \Theta$ (i.e., for all values of θ in the parameter space). For example, it may be that W_1 has smaller MSE for some values of $\theta \in \Theta$, but larger MSE for other values.

Remark: In some situations, we might have a biased estimator, but we can calculate its bias. We can then "adjust" the (biased) estimator to make it unbiased. I like to call this "making biased estimators unbiased." The following example illustrates this.

$\frac{\theta}{2} = \frac{1}{n}\sum X_i$

Example 7.14. Suppose that $X_1, X_2, ..., X_n$ are iid $\mathcal{U}[0, \theta]$, where $\theta > 0$. We know (from Example 7.4) that the MLE of θ is $X_{(n)}$, the maximum order statistic. It is easy to show that

$$E_\theta(X_{(n)}) = \left(\frac{n}{n+1}\right)\theta.$$

The MLE is biased because $E_\theta(X_{(n)}) \neq \theta$. However, the estimator

$$\left(\frac{n+1}{n}\right)X_{(n)},$$

an "adjusted version" of $X_{(n)}$, is unbiased.

Remark: In the previous example, we might compare the following estimators:

$$\begin{aligned} W_1 = W_1(\mathbf{X}) &= \left(\frac{n+1}{n}\right)X_{(n)} \\ W_2 = W_2(\mathbf{X}) &= 2\overline{X}. \end{aligned}$$

The estimator W_1 is an unbiased version of the MLE. The estimator W_2 is the MOM (which is also unbiased). I have calculated

$$\text{var}_\theta(W_1) = \frac{\theta^2}{n(n+2)} \quad \text{and} \quad \text{var}_\theta(W_2) = \frac{\theta^2}{3n}.$$

It is easy to see that $\text{var}_\theta(W_1) \leq \text{var}_\theta(W_2)$, for all $n \geq 2$. Therefore, W_1 is a "better" estimator on the basis of this variance comparison. Are you surprised?

Curiosity: Might there be another unbiased estimator, say $W_3 = W_3(\mathbf{X})$ that is "better" than both W_1 and W_2? If a better (unbiased) estimator does exist, how do we find it?

7.3.2 Best unbiased estimators

Goal: Consider the class of estimators

$$\mathcal{C}_\tau = \{W = W(\mathbf{X}) : E_\theta(W) = \tau(\theta) \ \ \forall \theta \in \Theta\}.$$

That is, $\mathcal{C}_\tau$ is the collection of all unbiased estimators of $\tau(\theta)$. Our goal is to find the (unbiased) estimator $W^* \in \mathcal{C}_\tau$ that has the smallest variance.

Remark: On the surface, this task seems somewhat insurmountable because $\mathcal{C}_\tau$ is a very large class. In Example 7.14, for example, both $W_1 = \left(\frac{n+1}{n}\right) X_{(n)}$ and $W_2 = 2\overline{X}$ are unbiased estimators of θ. However, so is the convex combination

$$W_a = W_a(\mathbf{X}) = a\left(\frac{n+1}{n}\right) X_{(n)} + (1-a)2\overline{X},$$

$E[W_1] = \theta$

$E[W_2] = \theta$

$W_1 \in C_\theta$, $W_2 \in C_\theta$

for all $a \in (0, 1)$.

$W_a = aW_1 + (1-a) W_2$

$E[W_a]$

$= a\theta + (1-a)\theta$

$= \theta$

$W_a \in C_\theta$

Remark: It seems that our discussion of "best" estimators starts with the restriction that we will consider only those that are unbiased. If we did not make a restriction like this, then we would have to deal with too many estimators, many of which are nonsensical. For example, suppose $X_1, X_2, ..., X_n$ are iid Poisson(θ), where $\theta > 0$.

- The estimators $\overline{X}$ and S^2 emerge as candidate estimators because they are unbiased.
- However, suppose we widen our search to consider all possible estimators and then try to find the one with the smallest MSE. Consider the estimator $\widehat{\theta} = 17$.
  - If $\theta = 17$, then $\widehat{\theta}$ can never be beaten in terms of MSE; its MSE = 0.
  - If $\theta \neq 17$, then $\widehat{\theta}$ may be a terrible estimator; its MSE = $(17 - \theta)^2$.
- We want to exclude nonsensical estimators like this. Our solution is to restrict attention to estimators that are unbiased.

Definition: An estimator $W^* = W^*(\mathbf{X})$ is a **uniformly minimum variance unbiased estimator** (**UMVUE**) of $\tau(\theta)$ if

1. $E_\theta(W^*) = \tau(\theta) \quad$ for all $\theta \in \Theta$
2. $\text{var}_\theta(W^*) \leq \text{var}_\theta(W)$, for all $\theta \in \Theta$, where W is any other unbiased estimator of $\tau(\theta)$.

Note: This definition is stated in full generality. Most of the time (but certainly not always), we will be interested in estimating θ itself; i.e., $\tau(\theta) = \theta$. Also, as the notation suggests, we assume that $\tau(\theta)$ is a scalar parameter and that estimators are also scalar.

Discussion/Preview: How do we find UMVUEs? We start by noting the following:

- UMVUEs may not exist.
- If a UMVUE does exist, it is unique (we'll prove this later).

We present **two approaches** to find UMVUEs:

Approach 1: Determine a **lower bound**, say $B(\theta)$, on the variance of any unbiased estimator of $\tau(\theta)$. Then, if we can find an unbiased estimator W^* whose variance attains this lower bound, that is,

$$\text{var}_\theta(W^*) = B(\theta),$$

for all $\theta \in \Theta$, then we know that W^* is UMVUE.

Approach 2: Link the notion of being "best" with that of sufficiency and completeness.

Theorem 7.3.9 (Cramér-Rao Inequality). Suppose $\mathbf{X} \sim f_{\mathbf{X}}(\mathbf{x}|\theta)$, where

1. the support of $\mathbf{X}$ is free of all unknown parameters
2. for any function $h(\mathbf{x})$ such that $E_\theta[h(\mathbf{X})] < \infty$ for all $\theta \in \Theta$, the interchange

$$\frac{d}{d\theta} \int_{\mathbb{R}^n} h(\mathbf{x}) f_{\mathbf{X}}(\mathbf{x}|\theta) d\mathbf{x} = \int_{\mathbb{R}^n} \frac{\partial}{\partial\theta} h(\mathbf{x}) f_{\mathbf{X}}(\mathbf{x}|\theta) d\mathbf{x}$$

   is justified; i.e., we can interchange the derivative and integral (derivative and sum if $\mathbf{X}$ is discrete).

For any estimator $W(\mathbf{X})$ with $\text{var}_\theta[W(\mathbf{X})] < \infty$, the following inequality holds:

$$\text{var}_\theta[W(\mathbf{X})] \geq \frac{\left\{\frac{d}{d\theta} E_\theta[W(\mathbf{X})]\right\}^2}{E_\theta\left\{\left[\frac{\partial}{\partial\theta} \ln f_{\mathbf{X}}(\mathbf{X}|\theta)\right]^2\right\}}.$$

The quantity on the RHS is called the **Cramér-Rao Lower Bound** (**CRLB**) on the variance of the estimator $W(\mathbf{X})$.

Remark: Note that in the statement of the CRLB in Theorem 7.3.9, we haven't said exactly what $W(\mathbf{X})$ is an estimator for. This is to preserve the generality of the result; Theorem 7.3.9 holds for any estimator with finite variance. However, given our desire to restrict attention to unbiased estimators, we will usually consider one of these cases:

- If $W(\mathbf{X})$ is an unbiased estimator of $\tau(\theta)$, then the numerator becomes

$$\left[\frac{d}{d\theta}\,\tau(\theta)\right]^2 = [\tau'(\theta)]^2.$$

$\mathrm{Var}_\theta[W(X)] \geq \dfrac{[\tau'(\theta)]^2}{E_\theta\left[\left(\frac{\partial}{\partial\theta}\ln f_X(x|\theta)\right)^2\right]}$

- If $W(\mathbf{X})$ is an unbiased estimator of $\tau(\theta) = \theta$, then the numerator equals 1.

$\mathrm{Var}_\theta[W(X)] \geq \dfrac{1}{E_\theta\left[\left(\frac{\partial}{\partial\theta}\ln f_X(x|\theta)\right)^2\right]}$

Important special case (Corollary 7.3.10): When $\mathbf{X}$ consists of $X_1, X_2, ..., X_n$ which are **iid** from the population $f_X(x|\theta)$, then the denominator in Theorem 7.3.9

$$E_\theta\left\{\left[\frac{\partial}{\partial\theta}\ln f_{\mathbf{X}}(\mathbf{X}|\theta)\right]^2\right\} \quad = \quad nE_\theta\left\{\left[\frac{\partial}{\partial\theta}\ln f_X(X|\theta)\right]^2\right\},$$

G

or, using other notation,

$$I_n(\theta) = nI_1(\theta).$$

We call $I_n(\theta)$ the **Fisher information** based on the sample $\mathbf{X}$. We call $I_1(\theta)$ the **Fisher information** based on one observation X.

CRLB of unbiased estimator of θ

Lemma 7.3.11 (Information Equality): Under fairly mild assumptions (which hold for exponential families, for example), the Fisher information based on one observation

$$I_1(\theta) = E_\theta\left\{\left[\frac{\partial}{\partial\theta}\ln f_X(X|\theta)\right]^2\right\} = -E_\theta\left[\frac{\partial^2}{\partial\theta^2}\ln f_X(X|\theta)\right].$$

$\dfrac{1}{-n\,E_\theta\left[\frac{\partial^2}{\partial\theta^2}\ln f_X(x|\theta)\right]}$

The second expectation is often easier to calculate.

Preview: In Chapter 10, we will investigate the large-sample properties of MLEs. Under certain regularity conditions, we will show an MLE $\widehat{\theta}$ satisfies

$$\sqrt{n}(\widehat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, \sigma^2_{\widehat{\theta}}),$$

where the asymptotic variance

$$\sigma^2_{\widehat{\theta}} = \frac{1}{I_1(\theta)}.$$

This is an extremely useful (large-sample) result; e.g., it makes getting large-sample CIs and performing large-sample tests straightforward. Furthermore, an analogous large-sample result holds for vector-valued MLEs. If $\widehat{\boldsymbol{\theta}}$ is the MLE of a $k \times 1$ dimensional parameter $\boldsymbol{\theta}$, then

$$\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} \mathrm{mvn}_k(\mathbf{0}, \boldsymbol{\Sigma}),$$

where the asymptotic variance-covariance matrix (now, $k \times k$)

$$\boldsymbol{\Sigma} = [I_1(\boldsymbol{\theta})]^{-1}$$

is the inverse of the $k \times k$ Fisher information matrix $I_1(\boldsymbol{\theta})$.