

STA 522, Spring 2021

Introduction to Theoretical Statistics II

Lecture 7

Department of Biostatistics
University at Buffalo

15 March, 2021

AGENDA

- ▶ Wrap up discussions of the Bayesian approach to Statistics
- ▶ Methods of evaluating estimators
- ▶ Cramér-Rao Lower Bound

Review: Bayesian approach to statistics

- ▶ The parameter θ is considered a random variable. Consider a **prior distribution** for θ before observing any data.
- ▶ After drawing a sample find the likelihood function for θ , and use Bayes' Rule to update the prior with the likelihood function to get the **posterior distribution**:

$$\pi(\theta | \underline{x}) = \frac{f(\underline{x} | \theta)\pi(\theta)}{m(\underline{x})} = \frac{f(\underline{x}, \theta)}{m(\underline{x})},$$

where $m(\underline{x})$ is the marginal distribution of \underline{X} :

$$m(\underline{x}) = \int f(\underline{x} | \theta)\pi(\theta) d\theta.$$

- ▶ The posterior distribution combines information from prior and likelihood.
- ▶ One Bayesian point estimator of θ is given by the posterior mean $E(\theta | \underline{X})$ (can also use posterior median, posterior mode, etc.)

Example (Binomial Bayes estimation): Let

$X_1, X_2, \dots, X_n \sim \text{iid Bernoulli}(p)$, and let $Y = \sum_{i=1}^n X_i$. Then

$Y \sim \text{binomial}(n, p)$. Assume the prior distribution on p to be $\text{beta}(\alpha, \beta)$.

The posterior distribution of p is

$$p \mid Y \sim \text{Beta}(y + \alpha, n - y + \beta)$$

The posterior mean is:

$$\begin{aligned}\hat{p}_B &= E(p \mid Y) \\ &= \frac{y + \alpha}{n + \alpha + \beta} \\ &= \underbrace{\left(\frac{n}{n + \alpha + \beta} \right)}_{=\text{sample mean}} \underbrace{\left(\frac{y}{n} \right)}_{=\text{sample mean}} + \underbrace{\left(\frac{\alpha + \beta}{n + \alpha + \beta} \right)}_{=\text{prior mean}} \underbrace{\left(\frac{\alpha}{\alpha + \beta} \right)}_{=\text{prior mean}}\end{aligned}$$

Conjugate Family

Definition: Let \mathcal{F} denote the class of pdfs or pmfs $f(x | \theta)$, indexed by θ . A class Π of prior distributions is a **conjugate family** for \mathcal{F} if the posterior distribution is in the class Π for all $f \in \mathcal{F}$, all priors in Π , and all $x \in \mathcal{X}$.

- ▶ The beta family is conjugate for the binomial family, which is why it was chosen as the prior distribution in the previous example.
- ▶ the gamma family is conjugate for the Poisson family.
- ▶ the normal family is its own conjugate.

Example (Normal Bayes Estimator): Let $X \sim N(\theta, \sigma^2)$, and suppose that the prior distribution on θ is $N(\mu, \tau^2)$ where σ^2 , μ and τ^2 are all known.

The posterior distribution of θ is also normal (Exercise 7.22; Homework) with

$$E(\theta | x) = \frac{\tau^2}{\sigma^2 + \tau^2}x + \frac{\sigma^2}{\sigma^2 + \tau^2}\mu = \frac{1/\sigma^2}{1/\sigma^2 + 1/\tau^2}x + \frac{1/\tau^2}{1/\sigma^2 + 1/\tau^2}\mu$$

and

$$\text{Var}(\theta | x) = \frac{\sigma^2\tau^2}{\sigma^2 + \tau^2} = \frac{1}{1/\sigma^2 + 1/\tau^2}$$

Using the posterior mean, a Bayes point estimator is given by $E(\theta | X)$.

Note that the Bayes estimator is again a linear combination of prior and sample means.

Method of Evaluating Estimators

- ▶ There may exist multiple estimators for the same problem, obtained from different approaches, e.g., method of moments, maximum likelihood, Bayesian approach etc.
- ▶ We want to compare these estimators and possibly obtain the “best” estimator.

Definition: The **mean squared error (MSE)** of an estimator W of a parameter θ is the function of θ defined by

$$\text{MSE} = \text{MSE}_{\theta}(W) = E_{\theta} [(W - \theta)^2] .$$

Note: **mean absolute error**, defined as

$$E_{\theta} [|W - \theta|] ,$$

is an alternative for measuring the performance of an estimator.

Definition: The **bias** of a point estimator W of a parameter θ is the difference between the expected value of W and θ ; that is,

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

An estimator whose bias is identically equal to 0 *as a function of θ* is called **unbiased** and satisfies $E_\theta(W) = \theta$ for all θ .

Note that $\text{MSE} = E_\theta [(W - \theta)^2] = \text{Var}_\theta(W) + [\text{Bias}_\theta(W)]^2$

- For an unbiased estimator, we have

$$\text{MSE} = E_\theta [(W - \theta)^2] = \text{Var}_\theta(W)$$

- Unbiased is a good property for an estimator to have, but it can be misleading.

Example: Let $X_1, X_2, \dots, X_n \sim \text{iid Bernoulli}(p)$. Let $W = X_1$. Since $E(W) = p$, W is unbiased, but doesn't use all the data. Note that for W ,

$$\text{MSE}(W) = \text{Var}(W) = p(1 - p).$$

Example: Let $X_1, X_2, \dots, X_n \sim \text{iid } N(\mu, \sigma^2)$. Since $E(\bar{X}) = \mu$ and $E(S^2) = \sigma^2$, \bar{X} and S^2 are unbiased for μ and σ^2 .

Thus, the mean squared errors (see Thms. 5.2.6 and 5.3.1) are

$$\text{MSE}(\bar{X}) = \text{Var}(\bar{X}) = \frac{\sigma^2}{n}$$

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

Recall the MLE for σ^2 is $\hat{\sigma}^2 = \frac{n-1}{n} S^2$.

Note that

$$\begin{aligned} E(\hat{\sigma}^2) &= E\left(\frac{n-1}{n} S^2\right) = \frac{n-1}{n} \sigma^2 \implies [\text{Bias}(\hat{\sigma}^2)]^2 = \frac{\sigma^4}{n^2} \\ \text{Var}(\hat{\sigma}^2) &= \text{Var}\left(\frac{n-1}{n} S^2\right) = \frac{2(n-1)\sigma^4}{n^2} \end{aligned}$$

Therefore

$$\text{MSE}(\hat{\sigma}^2) = \text{Var}(\hat{\sigma}^2) + [\text{Bias}(\hat{\sigma}^2)]^2 = \frac{2n-1}{n^2} \sigma^4$$

Note that

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

which implies $\text{MSE}(\hat{\sigma}^2) < \text{MSE}(S^2)$.

Example (7.3.5; Contd.): Let $X_1, X_2, \dots, X_n \sim \text{iid Bernoulli}(p)$. Then $Y = \sum_{i=1}^n X_i \sim \text{binomial}(n, p)$. Recall that we have developed two estimators for p , the MLE and the Bayes estimator:

$$\hat{p} = \frac{Y}{n} = \bar{X}$$

$$\hat{p}_B = \frac{Y + \alpha}{\alpha + \beta + n}$$

We have

$$\text{MSE}_p(\hat{p}) = \frac{p(1-p)}{n}$$

$$\begin{aligned}\text{MSE}_p(\hat{p}_B) &= \text{Var}_p(\hat{p}_B) + (\text{Bias}_p(\hat{p}_B))^2 \\ &= \text{Var}_p\left(\frac{Y + \alpha}{\alpha + \beta + n}\right) + \left[\mathbb{E}_p\left(\frac{Y + \alpha}{\alpha + \beta + n}\right) - p\right]^2 \\ &= \frac{np(1-p)}{(\alpha + \beta + n)^2} + \left(\frac{np + \alpha}{\alpha + \beta + n} - p\right)^2\end{aligned}$$

Choosing $\alpha = \beta = \frac{\sqrt{n}}{2}$ makes the MSE of \hat{p}_B constant as a function of p . Under this choice the MSEs are as follows:

$$\text{MSE}(\hat{p}) = \frac{p(1-p)}{n}$$

$$\text{MSE}(\hat{p}_B) = \frac{n}{4(n + \sqrt{n})^2}$$

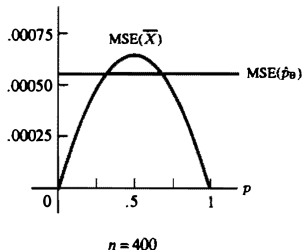
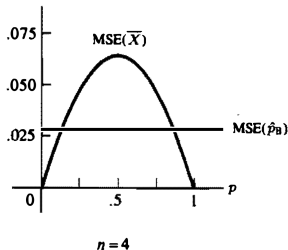


Figure 7.3.1. Comparison of MSE of \hat{p} and \hat{p}_B for sample sizes $n = 4$ and $n = 400$ in Example 7.3.5

Finding the “Best” Estimator

- ▶ Depends on what “best” means.
- ▶ Depends on the value of the parameter.
- ▶ There may be situations where a biased estimator (like \hat{p}_B) may be better.
- ▶ We first define “best” in relation to the variance of an unbiased estimator.

Definition: An estimator W^* is a **best unbiased estimator** of $\tau(\theta)$ if it satisfies the following:

- (a) W^* is unbiased, i.e., $E_{\theta}(W^*) = \tau(\theta)$ for all θ ; and
- (b) among all unbiased estimators, the variance (or MSE) of W^* is a minimum, i.e., for any other estimator W with $E_{\theta}(W) = \tau(\theta)$, we have

$$\text{MSE}(W^*) = \text{Var}_{\theta}(W^*) \leq \text{Var}_{\theta}(W) = \text{MSE}(W)$$

for all θ . W^* may also be called a **uniform minimum variance unbiased estimator (UMVUE)** of $\tau(\theta)$.

Cramér-Rao Inequality

- ▶ It is usually hard to determine if a UMVUE exists.
- ▶ However, there is a lower bound on the variance of any unbiased estimator.
- ▶ So if we can find an unbiased estimator that achieves this lower bound, we know it must be UMVUE.

Theorem (7.3.9; Cramér-Rao Lower Bound)

Let $\underline{X} = (X_1, X_2, \dots, X_n)$ have pdf $f(\underline{x} \mid \theta)$, and let $W(\underline{X})$ be any estimator satisfying

(a) $\frac{d}{d\theta} E_{\theta} [W(\underline{X})] = \int_{\mathcal{X}} \frac{\partial}{\partial \theta} [W(\underline{x})f(\underline{x} \mid \theta)] d\underline{x}$; and

(b) $\text{Var}_{\theta} [W(\underline{X})] < \infty$.

Then

$$\text{Var}_{\theta}(W(\underline{X})) \geq \frac{\left[\frac{d}{d\theta} E_{\theta} [W(\underline{X})] \right]^2}{E_{\theta} \left[\left[\frac{\partial}{\partial \theta} \log f(\underline{X} \mid \theta) \right]^2 \right]}.$$

Proof: Note that for any two random variables U and V

$$[\text{Cov}(U, V)]^2 \stackrel{\text{Cauchy-Schwarz}}{\leq} \text{Var}(U) \text{Var}(V) \implies \text{Var}(U) \geq \frac{[\text{Cov}(U, V)]^2}{\text{Var}(V)}$$

Take $U \equiv W(\underline{X})$ and $V \equiv \frac{\partial}{\partial \theta} \log f(\underline{X} | \theta)$. Note that

$$E(V) = E_{\theta} \left(\frac{\partial}{\partial \theta} \log f(\underline{X} | \theta) \right) = \int_{\mathcal{X}} \frac{\partial}{\partial \theta} \log f(\underline{x} | \theta) f(\underline{x} | \theta) d\underline{x} = \frac{\partial}{\partial \theta} E(1) = 0$$

So, $\text{Cov}(U, V) = E(UV)$ and $\text{Var}(V) = E(V^2)$. Note that

$$E(UV) = \int_{\mathcal{X}} W(\underline{x}) \frac{\partial}{\partial \theta} f(\underline{x} | \theta) d\underline{x} = \frac{d}{d\theta} E_{\theta}(W(\underline{X}))$$

Therefore, combining we get

$$\text{Var}_{\theta}(W(\underline{X})) = \text{Var}(U) \geq \frac{[\text{Cov}(U, V)]^2}{\text{Var}(V)} = \frac{[E(UV)]^2}{E(V^2)} = \frac{\left[\frac{d}{d\theta} E_{\theta}[W(\underline{X})] \right]^2}{E_{\theta} \left[\left[\frac{\partial}{\partial \theta} \log f(\underline{X} | \theta) \right]^2 \right]}$$

Corollary: Cramér-Rao Lower Bound, iid case

If the assumptions of Theorem 7.3.9 are satisfied, and, additionally, if X_1, X_2, \dots, X_n are iid with pdf $f(x \mid \theta)$, then

$$\text{Var}_\theta(W(\underline{X})) \geq \frac{\left[\frac{d}{d\theta} \mathbb{E}_\theta[W(\underline{X})]\right]^2}{n \mathbb{E}_\theta \left[\left[\frac{\partial}{\partial \theta} \log f(X_1 \mid \theta)\right]^2\right]}.$$

Proof: Homework. See p. 337 in the textbook.

Notes

- ▶ If $W(\underline{X})$ is unbiased for θ , then the numerator is

$$\left[\frac{d}{d\theta} \mathbb{E}_\theta[W(\underline{X})]\right]^2 = 1.$$

- ▶ The denominator is a function of the density, not the data.

Fisher Information

Definition: $E_{\theta} \left[\left[\frac{\partial}{\partial \theta} \log f(X | \theta) \right]^2 \right]$ is called the **Fisher information** of the sample.

Lemma 7.3.11: Calculating the Fisher Information

If $f(x | \theta)$ satisfies

$$\frac{d}{d\theta} E_{\theta} \left[\frac{\partial}{\partial \theta} \log f(X | \theta) \right] = \int_{\mathcal{X}} \frac{\partial}{\partial \theta} \left[\left[\frac{\partial}{\partial \theta} \log f(x | \theta) \right] f(x | \theta) \right] dx$$

(which is always true for an exponential family), then

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

Example: Let $X_1, X_2, \dots, X_n \sim \text{iid gamma}(\alpha, \beta)$, and assume α is known. Show that $W(\underline{X}) = \frac{1}{\alpha} \bar{X}$ attains the Cramér-Rao lower bound for an unbiased estimator of β , and hence is UMVUE.

First note that $\text{Var}_\beta(W(\underline{X})) = \text{Var}_\beta\left(\frac{1}{\alpha} \bar{X}\right) = \frac{1}{\alpha^2} \text{Var}_\beta(\bar{X}) = \frac{\beta^2}{\alpha n}$

Now obtain the CR lower bound. We have $f(x | \beta) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta}$ which implies $\log f(x | \beta) = \log \Gamma(\alpha) - \alpha \log \beta + (\alpha - 1) \log x - \frac{x}{\beta}$. Then

$$\frac{\partial}{\partial \beta} \log f(x | \beta) = -\frac{\alpha}{\beta} + \frac{x}{\beta^2}; \quad \frac{\partial^2}{\partial \beta^2} \log f(x | \beta) = \frac{\alpha}{\beta^2} - \frac{2x}{\beta^3}$$

Therefore

$$\mathbb{E}_\beta \left[\left[\frac{\partial}{\partial \beta} \log f(X_1 | \beta) \right]^2 \right] = -\mathbb{E}_\beta \left[\frac{\partial^2}{\partial \beta^2} \log f(X_1 | \beta) \right] = \mathbb{E}_\beta \left[\frac{2X_1}{\beta^3} - \frac{\alpha}{\beta^2} \right] = \frac{\alpha}{\beta^2}$$

Of course, $W(\underline{X}) = \frac{1}{\alpha} \bar{X}$ being unbiased for β means $\frac{d}{d\beta} \mathbb{E}_\beta [W(\underline{X})] = 1$.

Hence, the CR lower bound is (iid):

$$\frac{\left[\frac{d}{d\beta} \mathbb{E}_\beta [W(\underline{X})] \right]^2}{n \mathbb{E}_\beta \left[\left[\frac{\partial}{\partial \beta} \log f(X_1 | \beta) \right]^2 \right]} = \frac{1}{\alpha/\beta^2} = \frac{\beta^2}{\alpha} = \text{Var}_\beta(W(\underline{X}))$$

Example Let $X_1, X_2, \dots, X_n \sim \text{iid uniform}(0, \theta)$. The assumptions in CR inequality does not hold (verify; see p. 340 in the textbook). We will show that there exists an unbiased estimator of θ whose variance is uniformly smaller than the CRLB.

Note that here $\frac{\partial}{\partial \theta} \log f(x | \theta) = -\frac{1}{\theta} \implies E_{\theta} \left[\left[\frac{\partial}{\partial \theta} \log f(X_1 | \theta) \right]^2 \right] = \frac{1}{\theta^2}$.

So, CR lower bound $= \frac{\theta^2}{n}$.

Consider $Y = X_{(n)}$. $f_Y(y) = ny^{n-1}/\theta^n$, $0 < y < \theta$ so that

$$E_{\theta}(Y) = \int_0^{\theta} \frac{ny^n}{\theta^n} dy = \frac{n}{n+1} \theta \implies E_{\theta} \left(\underbrace{\frac{n+1}{n} Y}_{=U} \right) = \theta$$

i.e., U is an unbiased estimator of θ . We have

$$\text{Var}_{\theta}(U) = \text{Var}_{\theta} \left(\frac{n+1}{n} Y \right) = \frac{1}{n(n+2)} \theta^2 < \frac{1}{n} \theta^2$$

NOTE: In general, if the range of the pdf depends on the parameter, the Cramér-Rao Theorem will not be applicable.

Attainment

- ▶ There is no guarantee that the bound given in the Cramér-Rao Inequality is sharp. That is, our best unbiased estimator may not achieve the CRLB.
- ▶ Problem: when do we stop searching?

Homework

- ▶ Read p. 330 – 342.
- ▶ Exercises: TBA.

->

->