Biostat 602 Winter 2017

Lecture Set 8

Point Estimation

Methods for Evaluating Estimators

Reading: CB 7.3.1–7.3.2

Methods for Evaluating Estimators

Bias

Definition: Suppose $\hat{\theta}$ is an estimator for θ, then the bias of θ is defined as

$$\text{Bias}(\theta) = \text{E}(\hat{\theta}) - \theta$$

If the bias is equal to 0, then $\hat{\theta}$ is an unbiased estimator for θ.

Example 1: Let $X_1, \cdots, X_n$ be iid samples from a distribution with mean μ. Let $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$ be an estimator of μ. Its bias is

$$\begin{aligned} \text{Bias}(\mu) &= \text{E}(\overline{X}) - \mu \\ &= \text{E}\left(\frac{1}{n}\sum_{i=1}^n X_i\right) - \mu = \frac{1}{n}\sum_{i=1}^n \text{E}\left(X_i\right) - \mu = \mu - \mu = 0 \end{aligned}$$

Therefore $\overline{X}$ is an unbiased estimator for μ.

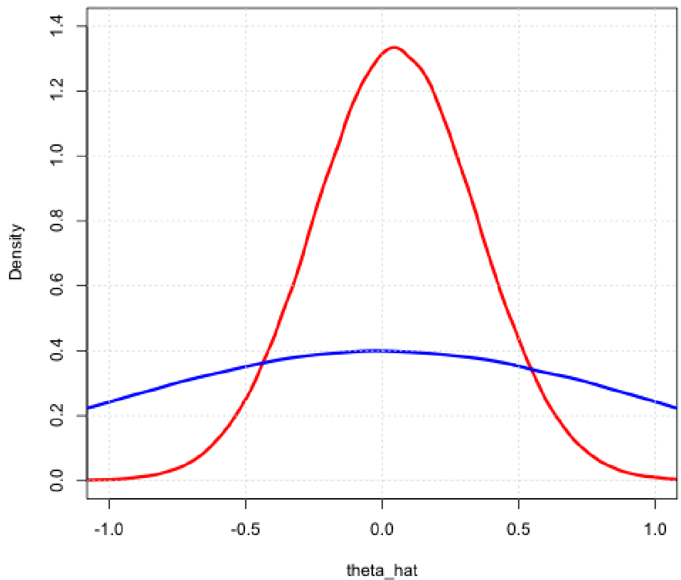
Example 2: Let $X_1, \cdots, X_n$ be iid samples from a distribution with mean μ and variance σ^2. Define

$$\begin{aligned} \hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2 \\ S^2 &= \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2 \end{aligned}$$

to be estimators of σ^2. Is either $\hat{\sigma}^2$ or S^2 an unbiased estimator? Which one?

How important is unbiasedness?

The Bias-Variance Trade-off



- $\hat{\theta}_1$ (blue) is unbiased but has a chance to be very far away from $\theta = 0$.
- $\hat{\theta}_2$ (red) is biased but more likely to be closer to the true θ than $\hat{\theta}_1$.

Mean Squared Error (MSE)

Definition: Mean Squared Error (MSE) of an estimator $\hat{\theta}$ is defined as

$$\text{MSE}(\hat{\theta}) = \text{E}[(\hat{\theta} - \theta)]^2$$

Note that

$$\begin{aligned}
\text{MSE}(\hat{\theta}) &= \text{E}[\hat{\theta} - \text{E}(\hat{\theta}) + \text{E}(\hat{\theta}) - \theta]^2 \\
&= \text{E}[\hat{\theta} - \text{E}(\hat{\theta})^2] + \text{E}[\text{E}(\hat{\theta}) - \theta]^2 + 2\text{E}[\hat{\theta} - \text{E}(\hat{\theta})]\text{E}[\text{E}(\hat{\theta}) - \theta] \\
&= \text{E}[\hat{\theta} - \text{E}(\hat{\theta})^2] + [\text{E}(\hat{\theta}) - \theta]^2 + 2[\text{E}(\hat{\theta}) - \text{E}(\hat{\theta})]\text{E}[\text{E}(\hat{\theta}) - \theta] \\
&= \text{Var}(\hat{\theta}) + \text{Bias}^2(\theta)
\end{aligned}$$

MSE, as a measure of performance, combines both bias and variance. So looking for an estimator that minimizes MSE for all $\theta \in \Omega$ would tantamount to searching for one which is on target on an average without ever going too far away.

Question: Is it possible to find an estimator that uniformly minimizes the MSE?

Example 3: Let $X_1, \cdots, X_n$ be a random sample from $\mathcal{N}(\mu, 1)$ Define

$$\hat{\mu}_1 = 1, \quad \hat{\mu}_2 = \overline{X}.$$

$$\begin{aligned} \text{MSE}(\hat{\mu}_1) &= \text{E}(\hat{\mu}_1 - \mu)^2 = (1-\mu)^2 \\ \text{MSE}(\hat{\mu}_2) &= \text{E}(\overline{X} - \mu)^2 = \text{Var}(\overline{X}) = \frac{1}{n} \end{aligned}$$

- Suppose that the true $\mu = 1$, then $\text{MSE}(\mu_1) = 0 < \text{MSE}(\mu_2)$, and no estimator can beat μ_1 in terms of MSE when true $\mu = 1$.
- Therefore, we cannot find an estimator that is uniformly the best in terms of MSE across all $\theta \in \Omega$ among all estimators
- Restrict the class of estimators, and find the "best" estimator within the small class.

Example 4: Let $X_1, \cdots, X_n$ be an iid random sample from $\mathcal{N}(\mu, \sigma^2)$. Define

$$
\begin{aligned}
\hat{\sigma}^2_{MLE} &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 \\
S^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2
\end{aligned}
$$

as estimators of σ^2. Which one has smaller MSE?

Solution: We shall use the following properties based on a i.i.d. random sample from a Normal Distribution (see pages 16-17 of Lecture Set 1; Theorem 5.3.1 C& B):

1. $E(S^2) = \sigma^2$.
2. $\overline{X}$andS^2 are independently distributed.
3. $\overline{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$.
4. $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$.

Example 5: Let $X_1, \cdots, X_n$ be iid $Poisson(\lambda)$. Let $\overline{X}$ and S^2 be the sample mean and variance, respectively. Since for Poisson distribution, mean = variance, both $\overline{X}$ and S^2 are unbiased for λ. Which estimator is better than the other (i.e. has smaller variance)?

Note that $Var(\overline{X}) = \lambda/n$, but $Var(S^2)$ is cumbersome and involves calculation of fourth moment.

Is there an alternative way to show which one is better?

We shall come back to this problem later.

Uniformly Minimum Variance Unbiased Estimator

Definition: $W^*(\mathbf{X})$ is the *best unbiased estimator*, or *uniformly minimum variance unbiased estimator (UMVUE)* of $\tau(\theta)$ if,

1. $\mathrm{E}[W^*(\mathbf{X})|\theta] = \tau(\theta)$ for all θ (unbiased)
2. and $\mathrm{Var}[W^*(\mathbf{X})|\theta] \leq \mathrm{Var}[W(\mathbf{X})|\theta]$ for all θ, where W is any other unbiased estimator of $\tau(\theta)$ (minimum variance).

First we develop some tools that will facilitate identification of best unbiased estimators. One of the key results towards that is an inequality called **Cramer-Rao (CR) Inequality** which we describe next.

Idea:

- CR inequality provides the lower bound of variances of any unbiased estimator of $\tau(\theta)$, say $B(\theta)$.
- If W^* is an unbiased estimator of $\tau(\theta)$ and satisfies $\mathrm{Var}[W^*(\mathbf{X})|\theta] = B(\theta)$, then W^* is the best unbiased estimator.

Some Terminology

Score and Fisher Information Number

Let $X_1, X_2, \ldots, X_n$ be random variables with joint pdf/pmf given by $f_{\mathbf{X}}(\mathbf{x}|\theta)$. The log-likelihood is defined by

$$l(\theta) = \log f_{\mathbf{X}}(\mathbf{x}|\theta).$$

Score Function: $l'(\theta) = \frac{\partial}{\partial\theta} \log f_{\mathbf{X}}(\mathbf{x}|\theta)$

Fisher Information Number: $I_n(\theta) = \mathrm{E}\left[\left\{\frac{\partial}{\partial\theta} \log f_{\mathbf{X}}(\mathbf{x}|\theta)\right\}^2\right]$

If $X_1, X_2, \ldots, X_n$ is a i.i.d. random sample from a well-behaved pdf/pmf (e.g. support of pdf/pmf does not depend on the parameter) then we have the following simplifications

$$\begin{aligned}
\mathrm{E}\left[l'(\theta)\right] &= \sum_{i=1}^{n} \frac{\partial}{\partial\theta} \log f_X(x_i|\theta) = 0 \\
I_n(\theta) &= \mathrm{E}\left[\left\{\frac{\partial}{\partial\theta} \log f_{\mathbf{X}}(\mathbf{x}|\theta)\right\}^2\right] \qquad \leftarrow \text{based on all data} \\
&= n\mathrm{E}\left[\left\{\frac{\partial}{\partial\theta} \log f_X(x|\theta)\right\}^2\right] \\
&= nI(\theta) = n \times (\text{Information based on a single observation})
\end{aligned}$$

where $f_X(x|\theta)$ is the pdf/pmf based on a single observation.

Example 6: Let $X_1, X_2, \ldots, X_n$ be a random sample from $Bernoulli(p)$. Obtain the score function and the information number.

Example 7: Let $X_1, X_2, \ldots, X_n$ be a random sample from $Exponential(\theta)$. Obtain the score function and the information number.

Cramer-Rao inequality

Theorem 7.3.9: Let $X_1, \cdots, X_n$ be a collection of random variables with joint pdf/pmf of $f_{\mathbf{X}}(\mathbf{x}|\theta)$. Suppose $W(\mathbf{X})$ is an unbiased estimator of $\tau(\theta)$ with finite variance, i.e.

$$\mathrm{E}[W(\mathbf{X})|\theta] = \tau(\theta),\ \forall \theta \in \Omega, \quad \mathrm{Var}[W(\mathbf{X})|\theta] < \infty.$$

If

$$\frac{d}{d\theta}\mathrm{E}[\log f_{\mathbf{X}}(\mathbf{X}|\theta)] \;=\; \mathrm{E}\left(\frac{\partial}{\partial\theta}\log f_{\mathbf{X}}(\mathbf{x}|\theta)\right) = 0$$

and

$$\frac{d}{d\theta}\mathrm{E}[W(\mathbf{X})|\theta] \;=\; \frac{d}{d\theta}\int_{x\in\mathcal{X}} W(\mathbf{x}) f_{\mathbf{X}}(\mathbf{x}|\theta)d\mathbf{x} = \int_{x\in\mathcal{X}} W(\mathbf{x})\frac{\partial}{\partial\theta} f_{\mathbf{X}}(\mathbf{x}|\theta)d\mathbf{x}$$

Then, a lower bound of $\mathrm{Var}[W(\mathbf{X})|\theta]$ is

$$\mathrm{Var}[W(\mathbf{X})] \geq \frac{[\tau'(\theta)]^2}{\mathrm{E}\left[\{\frac{\partial}{\partial\theta}\log f_{\mathbf{X}}(\mathbf{X}|\theta)\}^2\right]}$$

Corollary 7.3.10: If $X_1, \cdots, X_n$ are iid samples from pdf/pmf $f_X(x|\theta)$, and the assumptions in the above Cramer-Rao theorem hold, then the lower-bound of $\mathrm{Var}[W(\mathbf{X})]$ becomes

$$\mathrm{Var}[W(\mathbf{X})] \;\geq\; \frac{[\tau'(\theta)]^2}{n\mathrm{E}\left[\{\frac{\partial}{\partial\theta}\log f_X(X|\theta)\}^2\right]}$$

Remarks

1. The bound on the right of the inequality (called CRLB) is a uniform lower bound for the variance of all unbiased estimators of $\tau(\theta)$. Thus, if one finds an unbiased estimator whose variance satisfies CRLB, the search for best unbiased estimator is complete.

2. The proof of the CR inequality hinges on interchangeability of differentiation and integration. When we deal with pmf's for discrete distributions, integration is replaced by summation.

3. While the function f need not be differentiable with respect to $\mathbf{x}$ (e.g. pmf), it must be differentiable with respect to θ.

A Computational Tool

There is a computational simplification of Fisher Information number under mild regularity conditions.

Lemma 7.3.11: If $f_X(x|\theta)$ satisfies the two interchangeability conditions

$$\begin{aligned} \frac{d}{d\theta}\int_{x\in\mathcal{X}} f_X(x|\theta)dx &= \int_{x\in\mathcal{X}} \frac{\partial}{\partial\theta} f_X(x|\theta)dx \\ \frac{d}{d\theta}\int_{x\in\mathcal{X}} \frac{\partial}{\partial\theta} f_X(x|\theta)dx &= \int_{x\in\mathcal{X}} \frac{\partial^2}{\partial\theta^2} f_X(x|\theta)dx \end{aligned}$$

<u>which are true for exponential family</u>, then

$$I(\theta) = \mathrm{E}\left[\left\{\frac{\partial}{\partial\theta}\log f_X(X|\theta)\right\}^2\right] = -\mathrm{E}\left[\frac{\partial^2}{\partial\theta^2}\log f_X(X|\theta)\right]$$

Example 3 revisited: Let $X_1, \cdots, X_n$ be a iid random sample from Poisson(λ). Obtain the best unbiased estimator (if it exists) of λ. Define

$$\hat{\lambda}_1 = \overline{X}, \quad \hat{\lambda}_2 = s^2_{\mathbf{X}}.$$

Both $\hat{\lambda}_1$, $\hat{\lambda}_2$ are unbiased estimators of λ. In fact

$$\hat{\lambda}_a = a\overline{X} + (1-a)s^2_{\mathbf{X}}$$

is an unbiased estimator of λ for any $0 \le a \le 1$. Which one to choose? Since $\tau(\lambda) = \lambda$, the Cramer-Rao lower bound is $1/I_n(\lambda) = 1/\left[nI(\lambda)\right]$.

$$\begin{aligned}
I(\lambda) &= \mathrm{E}\left[\left\{\frac{\partial}{\partial\lambda}\log f_X(X|\lambda)\right\}^2\right] = -\mathrm{E}\left[\frac{\partial^2}{\partial\lambda^2}\log f_X(X|\lambda)\right] \\
&= -\mathrm{E}\left[\frac{\partial^2}{\partial\lambda^2}\log\frac{e^{-\lambda}\lambda^X}{X!}\right] = -\mathrm{E}\left[\frac{\partial^2}{\partial\lambda^2}\left(-\lambda + X\log\lambda - \log X!\right)\right] \\
&= \mathrm{E}\left[\frac{X}{\lambda^2}\right] = \frac{1}{\lambda^2}\mathrm{E}(X) = \frac{1}{\lambda}
\end{aligned}$$

Therefore, the Cramer-Rao lower bound is

$$\mathrm{Var}[W(\mathbf{X})] \ge \frac{1}{nI(\lambda)} = \frac{\lambda}{n}$$

where W is any unbiased estimator of λ.

$$\mathrm{Var}(\hat{\lambda}_1) = \mathrm{Var}(\overline{X}) = \frac{\mathrm{Var}(X)}{n} = \frac{\lambda}{n}$$

Therefore, $\hat{\lambda}_1 = \overline{X}$ is the best unbiased estimator of λ. With a lengthy calculation (need calculation of fourth moment), it is possible to show that

$$\mathrm{Var}(\hat{\lambda}_2) > \frac{\lambda}{n}$$

(details omitted), so $\hat{\lambda}_2$ is not the best unbiased estimator.

With and without Lemma 7.3.11

With Lemma 7.3.11

$$I(\lambda) = -\mathrm{E}\left[\tfrac{\partial^2}{\partial\lambda^2}\log f_X(X|\lambda)\right] = -\mathrm{E}\left[\tfrac{\partial^2}{\partial\lambda^2}\left(-\lambda + X\log\lambda - \log X!\right)\right] = \tfrac{1}{\lambda}$$

Without Lemma 7.3.11

$$\begin{aligned}
I(\lambda) &= \mathrm{E}\left[\left\{\frac{\partial}{\partial\lambda}\log f_X(X|\lambda)\right\}^2\right] \\
&= \mathrm{E}\left[\left\{\frac{\partial}{\partial\lambda}\left(-\lambda + X\log\lambda - \log X!\right)\right\}^2\right] \\
&= \mathrm{E}\left[\left\{-1 + \frac{X}{\lambda}\right\}^2\right] \\
&= \mathrm{E}\left[1 - 2\frac{X}{\lambda} + \frac{X^2}{\lambda^2}\right] \\
&= 1 - 2\frac{\mathrm{E}(X)}{\lambda} + \frac{\mathrm{E}(X^2)}{\lambda^2} \\
&= 1 - 2\frac{\mathrm{E}(X)}{\lambda} + \frac{\mathrm{Var}(X) + [\mathrm{E}(X)]^2}{\lambda^2} \\
&= 1 - 2\frac{\lambda}{\lambda} + \frac{\lambda + \lambda^2}{\lambda^2} = \frac{1}{\lambda}
\end{aligned}$$

Example 8: Let $X_1, \ldots, X_n$ be a random sample from $Bernoulli(p)$. Find the best unbiased estimator of p.

Example 9: Let $X_1, \ldots, X_n$ be a random sample from $\mathcal{N}(\mu, 1)$. Find the best unbiased estimator of μ.