MATH888: High-dimensional probability and statistics Fall 2021

Lecture 5 — September 17, 2021

Sebastien Roch, UW-Madison *Scribe: Zhuoyan Xu*

1 Overview

In the previous lecture we introduced some basic notations in point estimation and the definition of unbiasedness, mean squared error and Cramér-Rao Lower Bound. We introduced the main theorem of CR lower bound and talked about the intuition behind it.

In this lecture we mainly prove the Cramér-Rao Lower Bound and illustrate the property of it.

2 Cramér-Rao Lower Bound main theory

In this section we consider the special discrete case of random variables. For general results, see *Theorem 6.6* in Lehmann and Casella [1]. We redefine the following setting:

1. The *sample space* $\mathcal{X}$ is finite, i.e. $|\mathcal{X}| < \infty$.
2. The *parameter space* $\Theta \subseteq \mathbb{R}$ is an open set.
3. The *family of distribution*
$$\mathcal{P} = \{ \ p(\ \cdot \ ; \theta) \ : \ \theta \in \Theta \ \}$$
satisfies $p(x, \theta) > 0$ and $\frac{\partial}{\partial \theta} p(x, \theta)$ exists for $\forall x \in \mathcal{X}$ and $\forall \theta \in \Theta$.

Theorem 1. *Let $\hat{\vartheta}^{(n)}$, $n \in \mathbb{N}$ be an unbiased estimators of $g(\theta)$ i.e. $\mathbb{E}[\hat{\vartheta}^{(n)}] = g(\theta)$ for all $n \in \mathbb{N}$, where $g : \mathbb{R} \to \mathbb{R}$ is a continuously differentiable function, then:*

$$Var\left[\hat{\vartheta}^{(n)}\right] \geq \frac{[g'(\theta)]^2}{n\mathbb{E}\left[\frac{\partial}{\partial \theta} \log p(X_1; \theta)\right]^2}$$

where $\mathbb{E}\left[\frac{\partial}{\partial \theta} \log p(X_1; \theta)\right]^2 =: I(\theta)$ is the Fisher Information for θ.

Remark 2. *The Fisher information*

$$I(\theta) := \mathbb{E}\left[\left(\frac{\partial}{\partial \theta} \log p(x; \theta)\right)^2\right] = \mathbb{E}\left[\left(\frac{\frac{\partial}{\partial \theta} p(x; \theta)}{p(x; \theta)}\right)^2\right]$$

quantifies the expected relative rate of change of the likelihood with respect to a small perturbation in θ. It can be roughly seen as the relative "derivative" of pdf with respect to θ. A larger fisher information indicates the steep change in log-likelihood function in changing of parameter θ. It makes it easier to distinguish two likelihood function with different values of θ. In this sense, $I(\theta)$ captures information about the parameter θ. A larger Fisher information value also leads to a lower Cramér-Rao bound.

To illustrate the theorem of Cramér-Rao bound, we consider following example.

Example 3. *Suppose we have true parameter θ^* for some distribution p, define $\hat{\vartheta}^{(n)} = \theta^*$. Then we have*

$$\mathbb{E}[\hat{\vartheta}^{(n)}] = \theta^* \textit{and } \textit{Var}[\hat{\vartheta}^{(n)}] = 0$$

The estimator in this example has lower variance than CR lower bound, but it's not an unbiased estimator indeed. Recall the definition of unbiasedness: An estimator $\hat{\theta}$ is said to be unbiased if $\text{bias}(\hat{\theta}, \theta) = 0$ for all $P \in \mathcal{P}$. An unbiased estimator must have zero bias for *all* possible distributions. In this example, the constant estimator $\hat{\vartheta}^{(n)}$ has zero bias for any P with $\theta(P) = \theta^*$, but this is not an unbiased estimator for any P with a different value of θ.

Now we begin our proof of the main theorem.

Proof. Without abuse of notation, we define $\boldsymbol{X} = (X_1, \ldots, X_n)$ as random vector where $X_1, \ldots, X_n \overset{\text{iid}}{\sim} p(X, \theta)$, $\boldsymbol{x} = (x_1, \ldots, x_n)$ as the realization of $\boldsymbol{X}$.

We have $p^{(n)}(\boldsymbol{X}, \theta) = \prod_{i=1}^n p(X_i, \theta)$. Recall

$$\text{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}X)(Y - \mathbb{E}Y)] = \mathbb{E}XY - \mathbb{E}X\mathbb{E}Y)$$

for any function $\psi(\boldsymbol{X}, \theta)$, by Cauchy-Schwarz inequality, we have

$$\begin{aligned}
|\text{Cov}(\hat{\vartheta}, \psi(\boldsymbol{X}, \theta))|^2 &\leq \text{Var}(\hat{\vartheta})\text{Var}(\psi(\boldsymbol{X}, \theta)) \\
\text{Var}(\hat{\vartheta}) &\geq \frac{|\text{Cov}(\hat{\vartheta}, \psi(\boldsymbol{X}, \theta))|^2}{\text{Var}(\psi(\boldsymbol{X}, \theta)))}
\end{aligned}$$

We choose $\psi(\boldsymbol{X}, \theta) = \frac{\partial}{\partial \theta} \log p^{(n)}(\boldsymbol{X}, \theta)$, then

$$\begin{aligned}
\mathbb{E}(\psi) &= \sum_{x \in \mathcal{X}} p^{(n)}(\boldsymbol{X}, \theta) \frac{\partial}{\partial \theta} \log p^{(n)}(\boldsymbol{X}, \theta) \\
&= \sum_{x \in \mathcal{X}} p^{(n)}(\boldsymbol{X}, \theta) \frac{\frac{\partial}{\partial \theta} p^{(n)}(\boldsymbol{X}, \theta)}{p^{(n)}(\boldsymbol{X}, \theta)} \\
&= \frac{\partial}{\partial \theta} \left(\sum_{x \in \mathcal{X}} p^{(n)}(\boldsymbol{X}, \theta) \right) \\
&= 0
\end{aligned}$$

Since $X_1, \ldots, X_n \overset{\text{iid}}{\sim} p(X_1, \theta)$, we have:

$$\begin{aligned}
\text{Var}(\psi) &= \text{Var}\left(\frac{\partial}{\partial \theta} \log p^{(n)}(\boldsymbol{X}, \theta) \right) \\
&= \text{Var}\left(\sum_{i=1}^n \frac{\partial}{\partial \theta} \log p(X, \theta) \right) \\
&= n\text{Var}\left(\frac{\partial}{\partial \theta} \log p(X, \theta) \right) \\
&= n\mathbb{E}\left[\left(\frac{\partial}{\partial \theta} \log p(X, \theta) \right)^2 \right]
\end{aligned}$$

Consider covariance, we have:

$$\begin{aligned}
\operatorname{Cov}(\hat{\vartheta}, \psi) &= \operatorname{Cov}\left[\hat{\vartheta}, \frac{\partial}{\partial \theta} \log p^{(n)}(\boldsymbol{X}, \theta)\right] \\
&= \mathbb{E}\left[\hat{\vartheta} \cdot \frac{\partial}{\partial \theta} \log p^{(n)}(\boldsymbol{X}, \theta)\right] \\
&= \sum_{x \in \mathcal{X}} \left[p^{(n)}(\boldsymbol{X}, \theta) \cdot \hat{\vartheta} \cdot \frac{\frac{\partial}{\partial \theta} p^{(n)}(\boldsymbol{X}, \theta)}{p^{(n)}(\boldsymbol{X}, \theta)}\right] \\
&= \frac{\partial}{\partial \theta}\left[\sum_{x \in \mathcal{X}} p^{(n)}(\boldsymbol{X}, \theta) \cdot \hat{\vartheta}\right] \\
&= \frac{\partial}{\partial \theta}[g(\theta)] \\
&= g'(\theta)
\end{aligned}$$

plug in previous equation, we have desired result:

$$\operatorname{Var}\left[\hat{\vartheta}^{(n)}\right] \geq \frac{[g'(\theta)]^2}{n \mathbb{E}\left[\frac{\partial}{\partial \theta} \log p(X_1, \theta)\right]^2}$$

□

Example 4. *Consider Bernoulli distribution:*

- $\mathcal{X}$ *is* $= \{0,1\}$ Θ $=$ *(0,1)*
- $p(X, \theta) = \theta^X (1-\theta)^{1-X} > 0$ *for* $\forall x \in \mathcal{X}$ *and* $\forall \theta \in \Theta$.
- $\frac{\partial}{\partial \theta} p(x, \theta) = \frac{X}{\theta} - \frac{1-X}{1-\theta}$ *exists for* $\forall x \in \mathcal{X}$ *and* $\forall \theta \in \Theta$.
- $g(\theta) = \theta$.

We have Fisher information

$$\begin{aligned}
I(\theta) &= \mathbb{E}\left[\left(\frac{\partial}{\partial \theta} \log p(x; \theta)\right)^2\right] \\
&= \mathbb{E}\left[\left(\frac{\partial}{\partial \theta}[X \log(\theta) + (1-X) \log(1-\theta)]\right)^2\right] \\
&= \mathbb{E}\left[\left(\frac{X}{\theta} - \frac{1-X}{1-\theta}\right)^2\right] \\
&= \mathbb{E}\left[\frac{(X-\theta)^2}{(\theta(1-\theta))^2}\right] \\
&= \frac{1}{\theta(1-\theta)}
\end{aligned}$$

We have

$$Var\left[\hat{\vartheta}^{(n)}\right] \geq \frac{\theta(1-\theta)}{n}$$

Consider empirical mean estimator $\hat{\theta}^{(n)} = \frac{1}{n}\sum_{i=1}^{n} X_i$ *, we have*

$$\mathbb{E}\hat{\theta}^{(n)} = \theta$$
$$Var(\hat{\theta}^{(n)}) = \frac{\theta(1-\theta)}{n}$$

achieves the CR lower bound.

Remark 5. *We may not always find such unbiased estimator, consider* $g(\theta) = 1/\theta$ *in previous example, the unbiased estimator does not exist, see details on [2].*

References

[1] Lehmann, E. L. and Casella, G. (1998). *Theory of point estimation.* Springer.

[2] Uon-existence of unbiased estimator: `https://math.stackexchange.com/questions/681638/for-the-binomial-distribution-why-does-no-unbiased-estimator-exist-for-1-p`