STA732
Statistical Inference

Lecture 06: Information Inequality

Yuansi Chen

Spring 2023

Duke University

https://www2.stat.duke.edu/courses/Spring23/sta732.01/

Recap from Lecture 05

- **Convex loss** and Jensen's inequality
- **Rao-Blackwell Theorem** allows us to improve an estimator using sufficient statistics
- **UMVU** exists and is unique when the estimand is U-estimable and complete sufficient statistics exist

Goal of Lecture 06

1. Second thoughts about bias
2. Log-likelihood, score and Fisher information
3. Cramér-Rao lower bound
4. Hammersley-Chapman-Robbins ineq

Chap. 4.2, 4.5-4.6 in Keener or Chap. 2.5 in Lehmann and Casella

Second thoughts about bias

Def. Admissible

An estimator δ is called inadmissible if there exists δ^* which has a better risk:
$R(\theta, \delta^*) \leq R(\theta, \delta)$ for all $\theta \in \Omega$, with $R(\theta_1, \delta^*) < R(\theta_1, \delta)$ for some $\theta_1 \in \Omega$.

We also say that δ^* dominates δ

Uniform distribution example from last lecture

$X_1, \ldots, X_n$ are i.i.d. from the uniform distribution on $(0, \theta)$.
$T = \max\{X_1, \ldots, X_n\}$ is complete sufficient.

- We have derived that $\frac{n+1}{n}T$ is UMVU for estimating θ.
- Among estimators in the form of mutiple of T, is the UMVU estimator admissible?

Gaussian sequence model example

$X_i \sim \mathcal{N}(\mu_i, 1), i = 1, \ldots, n$, independent. Want to estimate $\|\mu\|_2^2$, where $\mu = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_n \end{pmatrix}$

- Find a UMVU estimator $\|X\|_2^2 - n$
- Can we find a better estimator (if $\mu = 0$)?

Thoughts about unbiased estimators

- A UMVU estimator is not necessarily admissible!
- It might even be absurd (Ex 4.7 in Keener)
- It is a good estimator to start with, but in general we shall not insist on UMVU

Log-likelihood, score and Fisher information

Log-likelihood

Suppose X has distribution from a family $\mathscr{P} = \{P_\theta, \theta \in \Omega\}$. Assume each distribution has density p_θ and shares the common support $\{x \mid p_\theta(x) > 0\}$. The log-likelihood is

$$\ell(\theta; X) = \log p_\theta(X)$$

Score

Def. Score

The score is defined as the gradient of the log-likelihood with respect to the parameter vector

$$\nabla \ell(\theta; X)$$

Remark

- can treat it as "local sufficient statistics", for $\xi \approx 0$

$$\begin{aligned} p_{\theta_0+\xi} &= \exp \ell(\theta_0 + \xi; x) \\ &\approx \exp\left[\xi^\top \nabla \ell(\theta_0; x)\right] \cdot p_{\theta_0}(x) \end{aligned}$$

- indicates the sensitivity to infinitesimal changes to θ.

Expected value of score is zero

Under enough regularity conditions, we have

$$\mathbb{E}_\theta\left[\nabla \ell(\theta; X)\right] = 0$$

Proof:

$$1 = \int \exp \ell(\theta; x) d\mu(x)$$

Taking derivative (under regularity conditions) implies

$$0 = \int \frac{\partial}{\partial \theta_j} \ell(\theta; x) \cdot \exp \ell(\theta; x) d\mu(x)$$

Def. Fisher information

For θ taking values in $\mathbb{R}^s$, the Fisher information is a $s \times s$ matrix

$$\begin{aligned} I(\theta) &= \mathrm{Cov}_\theta\left(\nabla \ell(\theta; X)\right) \\ &= \mathbb{E}_\theta\left[-\nabla^2 \ell(\theta; X)\right] \end{aligned}$$

why are the two definitions equivalent?

Cramér-Rao lower bound

Cramér-Rao lower bound in 1-dimension case

Consider an estimator $\delta(X)$ which is unbiased for $g(\theta)$. Then

$$g(\theta) = \mathbb{E}_\theta \delta$$

Under enough regularity

$$g'(\theta) = \int \delta(x)\ell'(\theta;x)e^{\ell(\theta;x)}d\mu(x) = \mathbb{E}_\theta \delta \ell'$$

Thm 4.9 in Keener

Let $\mathscr{P} = \{P_\theta : \theta \in \Omega\}$ be a dominated family with densities p_θ differentiable. Under enough regularity conditions ($\mathbb{E}_\theta l' = 0$, $\mathbb{E}_\theta \delta^2 < \infty$, g' well defined), we have

$$\mathrm{Var}_\theta(\delta) \geq \frac{[g'(\theta)]^2}{I(\theta)}, \theta \in \Omega$$

called Cramér-Rao lower bound or information lower bound proof idea: Cauchy Schwarz inequality

Cramér-Rao lower bound in high dimension

For $\theta \in \mathbb{R}^s$, we have

$$\mathrm{Var}_\theta(\delta) \geq \nabla g(\theta)^\top I(\theta)^{-1} \nabla g(\theta)$$

Interpretation of the Cramér-Rao lower bound

- To estimate $g(\theta)$, no unbiased estimator can have smaller variance than $\nabla g(\theta)^\top I(\theta)^{-1} \nabla g(\theta)$
- For a unbiased estimator δ, we always have the lower bound of the form for any random variable ψ

$$\mathrm{Var}_\theta(\delta) \geq \frac{\mathrm{Cov}_\theta^2(\delta, \psi)}{\mathrm{Var}_\theta(\psi)}$$

What is a good ψ?

Example: Cramér-Rao lower bound for i.i.d. samples

Suppose $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} p_\theta^{(1)}, \theta \in \Omega$. The joint density is

$$p_\theta(x) = \prod_{i=1}^{n} p_\theta^{(1)}(x_i)$$

What is the relationship between Fisher information for n i.i.d. observations and that for a single observation?

Efficiency

CRLB is not always attainable

Def. efficiency

The efficiency of an unbiased estimator δ is

$$\text{eff}_\theta(\delta) = \frac{\text{CRLB}}{\mathrm{Var}_\theta(\delta)}$$

Remark

- According to the definition and the Cramér-Rao lower bound, for "regular" unbiased estimators, $\text{eff}_\theta(\delta) \leq 1$
- Efficiency 1 is rarely achieved in finite samples, but usually we can approach it asymptotically as $n \to \infty$

Hammersley-Chapman-Robbins Inequality

Motivation behind Hammersley-Chapman-Robbins Inequality

The Cramér-Rao lower bound requires the differentiation under integral, thus requires regularity conditions so that the differentiation is well-defined.

We can get a more general statement if we replace $\nabla \ell(\theta; X)$ with the corresponding finite difference.

Hammersley-Chapman-Robbins Inequality (1)

Recall that by Cauchy-Schwarz, for a unbiased estimator δ, we always have the lower bound of the form for any random variable ψ

$$\mathrm{Var}_\theta(\delta) \geq \frac{\mathrm{Cov}_\theta^2(\delta, \psi)}{\mathrm{Var}_\theta(\psi)}$$

- In CRLB, we took $\psi = \nabla \ell(\theta; X)$
- Here we take

$$\frac{p_{\theta+\epsilon}(X)}{p_\theta(X)} - 1 = \exp\left(\ell(\theta+\epsilon; X) - \ell(\theta; X)\right) - 1$$

$\approx \epsilon^\top \nabla \ell(\theta; X)$ for small ϵ

Hammersley-Chapman-Robbins Inequality (2)

We verify that

- $\mathbb{E}\left[\frac{p_{\theta+\epsilon}(X)}{p_\theta(X)} - 1\right] = 0$
-

$$\begin{aligned}
\mathrm{Cov}_\theta\left(\delta(X), \frac{p_{\theta+\epsilon}(X)}{p_\theta(X)} - 1\right) &= \int \delta(x)\left(\frac{p_{\theta+\epsilon}(x)}{p_\theta(x)} - 1\right) p_\theta(x) d\mu(x) \\
&= \mathbb{E}_{\theta+\epsilon}[\delta] - \mathbb{E}_\theta[\delta] \\
&= g(\theta+\epsilon) - g(\theta)
\end{aligned}$$

Hence HCRI:

$$\mathrm{Var}_\theta(\delta) \geq \frac{\left(g(\theta+\epsilon) - g(\theta)\right)^2}{\mathbb{E}\left[\left(\frac{p_{\theta+\epsilon}(x)}{p_\theta(x)} - 1\right)^2\right]}$$

CRLB follows from taking $\epsilon \to 0$, but taking sup over ϵ can give better bounds

Example 1: exponential family

What is the Cramér-Rao lower bound for the exponential family?

Example 2: curved exponential family

What is the Cramér-Rao lower bound for the curved exponential family?

$$p_\theta(x) = \exp(\eta(\theta)^\top T(x) - B(\theta))h(x), \quad \theta \in \mathbb{R}, T(x) \in \mathbb{R}^s$$

Summary

- Restricting to unbiased estimators have nice theory: UMVU theory. But it is not always admissible in terms of total risk
- Score and Fisher information
- Cramér-Rao lower bound and its variant

What is next?

- Equivariance

Thank you