STAT5010 Advanced Statistical Inference **Fall 2023**

Lecture 4: Cramér-Rao Information Bound

Lecturer: Tony Sit *Scribe: Cheukhei Chan and Yang Boyu*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

4.1 UMVUE

Q: Suppose we have δ_1 and δ_2 as UMVUEs for $g_1(\theta)$ and $g_2(\theta)$ respectively. Is $\delta_1 + \delta_2$ an UMVUE for $g_1(\theta) + g_2(\theta)$?

Theorem 4.1 (Characterization of UMVUEs; TPE 2.1.7) *Let $\Delta = \{\delta : E(\delta^2) < \infty\}$. Then $\delta_0 \in \Delta$ is UMVU for $g(\theta) = E(\delta_0)$ if and only if $E(\delta_0(\theta)U) = 0$ for every $U \in \mathcal{U} = \{U : E(U) = 0\}$*

Proof:
If δ_0 is a UMVUE, let's consider $\delta_\lambda = \delta_0 + \lambda U$ for $\lambda \in \mathbb{R}$ and $U \in \mathcal{U}$. Since δ_0 has minimal variance,

$$Var(\delta_\lambda) = Var(\delta_0) + \lambda^2 Var(U) + 2\lambda cov(\delta_0, U)$$
$$\geq Var(\delta_0)$$

Consider the quadratic form q(λ) = $\lambda^2 Var(U) + 2\lambda$ cov(δ_0, U).
The form q has the roots $\lambda = 0$ and 2cov(δ_0, U)/var(U).
If the roots are distinct, then the form must be negative at some point, which would violate the inequality above.
Hence, 2cov(δ_0, U)/var(U) = 0 in which case, E(Uδ_0) = cov(δ_0, U) = 0.

To prove the converse result, we assume that E(U δ_0) = 0 $\forall$ U $\in \mathcal{U}$ and consider any δ unbiased for g(θ). It follows that δ - $\delta_0 \in \mathcal{U}$, so E($\delta_0(\delta - \delta_0)$)=0. This implies that E($\delta_0\delta$) = E(${\delta_0}^2$) and subtracting E(δ_0)E(δ) on both sides, we obtain

$$Var(\delta_0) = cov(\delta_0,\delta) \leq \sqrt{Var(\delta_0)Var(\delta)} \quad byCauchy-Schwarzinequality.Hence, Var($$

$\delta_0) \leq Var(\delta)$ for any arbitrary unbiased estimator δ and δ_0. Hence, δ_0 is an UMVUE for g(θ).

To answer the question, $\forall$ U $\in$U$, E((\delta_1 + \delta_2)$U) = E($\delta_1$U) + E($\delta_2$U) = 0 ($\delta_1$, δ_2: UMVUEs)
$\Rightarrow \delta_1 + \delta_2$ is a UMVUE for $g_1(\theta) + g_2(\theta)$

4.2 Variance Bound and Information

Recall Cov(X,Y) $\leq \sqrt{Var(X)Var(Y}$
Given this inequality, if δ is an unbiased estimator for g(θ) and ψ is an arbitrary random variable, then

$$Var_\theta(\delta) \geq \text{Cov}^2\theta(\delta,\psi)/\text{Var}_\theta(\psi) \quad \{*\}$$

If we manage to find a suitable ψ so that the bound is meaningful in the sense that Cov$_\theta(\delta,\, \psi)$ is the same for all δ that are unbiased for g(θ).

Let P = $\{P_\theta : \theta \in \Theta\}$ be a dominated family with densities P$_\theta$:$\theta \in \Theta \in \mathbb{R}$. To begin, E$_{\theta+\Delta}(\delta) - E\theta(\delta)$ gives the same value g($\theta + \Delta$) - g(θ) for any unbiased δ. Hence, Δ must be chosen so that $\theta\, +\, \Delta \in \Theta$.

Next, we write E$_{\theta+\Delta}(\delta) - E_\theta(\delta)$ as a covariance under Pθ. This step involves the use of likelihood ratio. We assume here that Pθ + Δ(x) = 0 whenever P$_\theta(x) = 0$

Define L(x) = P$_{\theta+\Delta}(x)/P_\theta(x), where P\theta$(x) ¿ 0 and L(x) = 1 otherwise.

We have

$$L(x)P_\theta(x) = \frac{P_{\theta+\Delta}(x)}{P_\theta(x)}P_\theta(x) = P_{\theta+\Delta}(x) \quad a.s. \quad P.$$

and so, for any function h integrable under $P_{\theta+\Delta}$, we have

$$E_{\theta+\Delta}h(X) = \int hP_{\theta+\Delta}d\mu = \int hLP_\theta d\mu = E_\theta(L(X)h(X))$$

Take h = 1, $E_\theta(L(X)) = 1$ because E$_\theta(L(X)) = \int \frac{P_{\theta+\Delta}(x)}{P_\theta(x)}P_\theta(x)dx = \int P_{\theta+\Delta}(x)dx = 1$

Take h = $\delta, E_{\theta+\Delta}(\delta) = E_\theta(L\delta)$. So if we define ψ(x) = L(x) - 1, we can see that

$$E_\theta(\psi(X)) = E_\theta(L(X) - 1) = 1 - 1 = 0$$

and

$$E_{\theta+\Delta}(\delta) - E_\theta(\delta) = \mathrm{E}_\theta(L\delta) - E_\theta(\delta) = E_\theta(\psi\delta)$$

As a result,

$$Cov_\theta(\delta, \psi) = g(\theta + \Delta) - g(\theta)$$

for any unbiased estimator δ. With this particular choice of ψ, the inequality (*) can be rewritten as

$$Var_\theta(\delta) = \frac{\{g(\theta+\Delta) - g(\theta)\}^2}{Var_\theta(\psi)} = \frac{\{g(\theta+\Delta) - g(\theta)\}^2}{E_\theta(\frac{P_{\theta+\Delta}(x)}{P_\theta(x)} - 1)^2}, \quad (**)$$

which is known as the Hammersley-Chapman-Robbins Inequality.
Under suitable regularity conditions, we can show that

$$\frac{\lim_{\Delta\to 0}\{\frac{g(\theta+\Delta)-g(\theta)}{\Delta}\}^2}{\lim_{\Delta\to 0} E_\theta(\frac{\{P_{\theta+\Delta}(x)-P_\theta(x)\}/\Delta}{P_\theta(x)})^2} \to \frac{(g'(\theta)^2)}{E(\frac{\frac{\theta(x)}{\partial\theta}}{P_\theta((x)})^2} \quad (***)$$

The denominator on the RHS of (***) is known as Fisher information, denoted as I(θ) and is given by

$$I(\theta) = E_\theta(\frac{\partial log P_\theta(x)}{\partial\theta})^2$$

With enough regularity conditions to interchange integration and differentiation,

$$0 = \frac{\partial}{\partial\theta}(1) = \frac{\partial}{\partial\theta}\int P_\theta(x)d\mu(x) = \int \frac{\partial}{\partial\theta}P_\theta(x)d\mu(x)$$

$$= \int \frac{\partial log P_\theta(x)}{\partial\theta}P_\theta(x)d\mu(x) = E_\theta(\frac{\partial log P_\theta(x)}{\partial\theta})$$

and so,

$$I(\theta) = E_\theta(\frac{\partial log P_\theta(x)}{\partial\theta})^2 - (E_\theta(\frac{\partial log P_\theta(x)}{\partial\theta}))^2 = Var_\theta(\frac{\partial log P_\theta(x)}{\partial\theta}).$$

Furthermore, since

$$\int \frac{\partial^2 log P_\theta(x)}{\partial\theta^2}d\mu(x) = E_\theta(\frac{\partial^2 P_\theta(x)/\partial\theta^2}{P_\theta(x)}) = 0$$

We can see that

$$\frac{\partial^2 log P_\theta(x)}{\partial\theta^2} = \frac{\partial^2 P_\theta(x)/\partial\theta^2}{P_\theta(x)} - (\frac{\partial log P_\theta(x)}{\partial\theta})^2$$

$$\Longrightarrow I(\theta) = -E_\theta(\frac{\partial^2 log P_\theta(x)}{\partial\theta^2})$$

Therefore,

$$Var_\theta(\delta) \geq \frac{\{g'(\theta)\}^2}{I(\theta)}, \quad \theta \in \Theta$$

Theorem 4.2 *Let $P = \{P_\theta : \theta \in \Theta\}$ be a dominated family with Θ and open set in $\mathbb{R}$ and densities P_θ differentiable with respect to θ. If $E_\theta(\psi) = 0$ and $E_\theta(\delta^2) < \infty$, then*

$$Var_\theta(\delta) \geq \frac{\{g'(\theta)\}^2}{I(\theta)}, \quad \theta \in \Theta.$$

This result is called the Cramér-Rao / Information Bound.

Example 4.2.1 *Let $\mathcal{P}$ be a one-parameter exponential family in canonical form and density p_η given by*

$$p_\eta = exp(\eta T(x) - A(\eta))h(x),$$

then

$$\frac{\partial \log p_\eta(x)}{\partial\eta} = T(x) - A'(\eta).$$

By the previous results, we have

$$I(\eta) = Var_\eta(T(x) - A'(\eta)) = Var_\eta(T(x)) = A''(\eta).$$

If the family is parameterised instead by $\mu = A'(\eta) = \eta\mathbb{T}(\mathbb{X})$, *then*

$$A''(\eta) = I(\mu)(A''(\eta))^2,$$

and so, because $A''(\eta) = Var(T)$, *we have* $I(mu) = \frac{1}{Var_e ta(T)}$. *Observe also that because* T *is UMVUE for* μ, *the lower bound variance* $Var(\delta) \geq (I(\mu))^{-1}$ *for unbiased estimation* δ *of* μ *is ??*

Example 4.2.2 *Suppose* X *us ab absolutely continuous random variable with density* f. *The family of distributions* $\mathcal{P} = \{P_\theta : \theta \in \mathbb{R}\}$ *with* P_θ *the distribution of* $\theta + \epsilon$ *is called a location family.*

$$\begin{aligned}
\int g(x)\mathrm{d}P_\theta(x) &= \mathbb{E}\theta(g(X)) \\
&= \mathbb{E}_\theta(g(\theta + \epsilon)) \\
&= \int g(\theta + \epsilon) f(\epsilon)\mathrm{d}\epsilon \\
&= \int g(x) f(x - \theta)\mathrm{d}x.
\end{aligned}$$

So P_θ *has the density* $p_\theta = f(x - \theta)$. *The corresponding Fisher information for the family is*

$$\begin{aligned}
I(\theta) &= \mathbb{E}_\theta(\frac{\partial \log f(X - \theta)}{\partial \theta})^2 \\
&= \mathbb{E}_\theta(-\frac{f'(X - \theta)}{f(X - \theta)})^2 \\
&= \mathbb{E}(\frac{f'(\epsilon)}{f(\epsilon)})^2 \\
&= \int \frac{\{f'(x)\}^2}{f(x)}\mathrm{d}x \perp\!\!\!\perp \theta.
\end{aligned}$$

So, for the location families, $I(\theta)$ *is constant with respect to* θ.

If two (or more) independent vectors are observed, then the total Fisher information is the sum of the Fisher information provided by the individual observations.

Suppose X and Y are independent, and that X has density p_θ and Y has density q_θ. The Fisher information from X and Y are respectively

$$I_X(\theta) = Var_\theta(\frac{\partial \log p_\theta(X)}{\partial \theta}),$$

and

$$I_Y(\theta) = Var_\theta(\frac{\partial \log q_\theta(Y)}{\partial \theta}).$$

$$\begin{aligned}
I_{X,Y}(\theta) &= Var_\theta(\frac{\partial \log \{p_\theta(X) q_\theta(Y)\}}{\partial \theta}) \\
&= Var_\theta(\frac{\partial \log p_\theta(X)}{\partial \theta} + \frac{\partial \log q_\theta(Y)}{\partial \theta}) \\
&= Var_\theta(\frac{\partial \log p_\theta(X)}{\partial \theta}) + Var_\theta(\frac{\partial \log q_\theta(Y)}{\partial \theta}) \\
&= I_X(\theta) + I_Y(\theta).
\end{aligned}$$

$\Rightarrow$ If $X_1, \cdots, X_n \overset{\text{i.i.d.}}{\sim} P_\theta$, $I_X(\theta) = I_{X_1}(\theta + \cdots + I_{X_n}(\theta)) = nI_{X_1}(\theta)$.

$\Rightarrow Var_\theta(\delta) \geq \frac{g'(\theta)}{nI(\theta)}$

4.3 Multi-parameter Cramer-Rao Inequality

Suppose the following conditions hold:

1. $\Theta \subseteq \mathbb{R}^k$ is an open set,
2. $\{P_\theta : \theta \in \Theta\}$ have common support I,
3. $\frac{\partial p_\theta(X)}{\partial \theta_i}$ exists, $\forall i = 1, \cdots k, x \in I$ and is finite,
4. $\frac{\partial \int_x p_\theta(x) \mathrm{d}\mu}{\partial \theta_i} = \int_x \frac{\partial p_\theta(x)}{\partial \theta_i} \mathrm{d}\mu, \forall i = 1, \cdots, k,$
5. $\frac{\partial \int_x \delta(x) p_\theta(x) \mathrm{d}\mu}{\partial \theta_i} = \int_x \delta(x) \frac{\partial p_\theta(x)}{\partial \theta_i} \mathrm{d}\mu, \forall i = 1, \cdots, k$

Define the $k \times k$ information matrix $I(\theta)$ by

$$I_{ij}(\theta)_{i,j=1,\cdots,k}, \text{with} I_{ij}(\theta) = \mathbb{E}_\theta(\frac{\partial \log p_\theta(X)}{\partial \theta_i} \frac{\partial \log p_\theta(X)}{\partial \theta_j})$$

Specially, if $k = 1, I_{11} = \mathbb{E}_\theta(\frac{\partial \log p_\theta(X)}{\partial \theta})^2$. Assume that $I(\theta)$ is finite and positive definite, then

$$Var_\theta(\delta(X)) \geq \alpha^T I(\theta)^{-1} \alpha, \text{with} \alpha_i = \frac{\partial g(\theta)}{\partial \theta_i}.$$

Proof: Let $\psi_i(X) = \frac{\partial \log p_\theta(X)}{\partial \theta_i}$, then

$$\begin{aligned}
\mathbb{E}_\theta(\psi_i(X)) &= \int_x (\frac{\partial \log p_\theta(x)}{\partial \theta_i}) p_\theta(x) \mathrm{d}\mu(x) \\
&= \int_x \frac{\frac{\partial p_\theta(x)}{\partial \theta_i}}{p_\theta(x)} p_theta(x) \mathrm{d}\mu \\
&= \int_x \frac{\partial p_\theta(x)}{\partial \theta_i} \mathrm{d}\mu \\
&= \frac{\partial \int_x p_\theta(x) \mathrm{d}\mu}{\partial \theta_i} = 0
\end{aligned}$$

Fix a non-zero vector $(a_1, \cdots, a_k)$. Then $\mathbb{E}_\theta(\sum_{i=1}^k a_i \psi_i(X)) = 0$. Claim: $Var(\sum_{i=1}^k a_i \psi_i(X)) = a^T I(\theta) a$.

$$\begin{aligned}
Var(\sum_{i=1}^k a_i \psi_i(X)) &= \sum_{ij} a_i a_j Cov(\psi_i(X), \psi_j(X)) \\
&= \sum_{ij} a_i a_j \mathbb{E}(\psi_i(X) \psi_j(X)) \\
&= \sum_{i,j} a_i a_j I_{ij}(\theta) \\
&= a^T I(\theta) a.
\end{aligned}$$

Finally,

$$
\begin{aligned}
Cov(\delta(X), \sum_{i=1} ka_j\psi_i(X)) &= \sum_{i=1} ka_i Cov(\delta(X), \psi_i(X)) \\
&= \sum_{i=1} ka_i \mathbb{E}(\delta(X)\psi_i(X)) \\
&= \sum_{i=1} ka_i \int_x \delta(x) \frac{\partial \log p_\theta(x)}{\partial \theta_i} p_\theta(x) \mathrm{d}\mu \\
&= \sum_{i=1} ka_i \int_x \delta(x) \frac{\partial p_\theta(X)}{\partial \theta_i} \mathrm{d}x \\
&= \sum_{i=1} ka_i \frac{\partial \int_x \delta(x) p_\theta(x) \mathrm{d}\mu}{\partial \theta_i} \\
&= \sum_{i=1} ka_i \alpha_i(\theta).
\end{aligned}
$$

By Cauchy-Schwarz inequality,

$$
Var(\delta(X))Var(\sum_{i=1} ka_i\psi_i(X)) \geq Cov(\sum_{i=1} ka_i\psi_i(X), \delta(X)),
$$

$\Rightarrow$

$$
Var(\delta(X)) \geq \sup_{a \neq 0} \frac{(\sum_{i=1} ka_i\alpha_i(\theta))^2}{a^T I(\theta) a} = \alpha^T I(\theta)^{-1} \alpha.
$$

■

Example 4.3.1 *Suppose* $X_1, \cdots, X_n \overset{i.i.d.}{\sim} N(\mu, \sigma^2), \mu \in \mathbb{R}$, *and* $\sigma^2 > 0$. *We want to estimate* $g_1(\mu, \sigma^2) = \mu$ *and* $g_2(\mu, \sigma^2) = \sigma^2$. *Look at only unbiased estimators.*

Claim 1: $\frac{1}{n}\sum_{i=1}^n (X_i - \overline{X_n})^2$ *is UMVUE for* σ^2, $\overline{X_n} = \frac{1}{n}\sum_{i=1}^n X_i$.

Proof:

1. $(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$ *is complete and sufficient.*
2. $\frac{1}{n}\sum_{i=1}^n (X_i - \overline{X_n})^2$ *is a function of* $(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$. *This is because* $\frac{1}{n}\sum_{i=1}^n (X_i - \overline{X_n}) = \sum_{i=1}^n X_i^2 - n\overline{X_n}^2$.
3. $\frac{1}{\sigma^2} \sim \chi^2_{n-1} \Rightarrow \mathbb{E}(\frac{1}{n-1}\frac{1}{n}\sum_{i=1}^n (X_i - \overline{X_n})^2) = \frac{\sigma^2}{n-1}(n-1) = \sigma^2$

■

Claim 2: $\overline{X_n}$ *is UMVNE for* μ. $\mathbb{E}(\overline{X_n} - \mu)^2 = \sigma^2/n$. *Note that*$\mathbb{E}(\frac{1}{n}\sum_{i=1}^n (X_i - \overline{X_n})^2/(n-1) - \sigma^2) = Var(\frac{1}{n}\sum_{i=1}^n (X_i - \overline{X_n})^2/(n-1)) = \frac{2(n-1)\sigma^4}{(n-1)^2} = \frac{2\sigma^4}{(n-1)}$.

$$
\log p_{\mu,\sigma^2}(X) = -\frac{n}{2}\log \sigma^2 - \sum_{i=1}^n \frac{(X_i - \mu)^2}{2 \sim^2} + C_n,
$$

$$
\frac{\partial \log p_\theta(X)}{\partial \mu} = \sum_{i=1}^n \frac{X_i - \mu}{\sigma^2},
$$

$$\frac{\partial^2 \log p_\theta(X)}{\partial \mu^2} = -\frac{n}{\sigma^2},$$

$$\frac{\partial \log p_\theta(X)}{\partial \sigma^2} = -\frac{n}{\sigma^2} + \sum_{i=1}^{n} \frac{(X_i - \mu)^2}{2\sigma^4}$$

$$\frac{\partial^2 \log p_\theta(X)}{\partial (\sigma^2)^2} = \frac{n}{\sigma^4} - \sum_{i=1}^{n} \frac{(X_i - \mu)^2}{\sigma^6}$$

$$\frac{\partial^2 \log p_\theta(X)}{\partial \mu \partial \sigma^2} = -\sum_{i=1}^{n} \frac{X_i - \mu}{\sigma^4}$$

Therefore, the Fisher information matrix is

$$I = \begin{pmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{pmatrix}$$

$$I_{22} = -\frac{n}{2\sigma^4} + \mathbb{E}(\sum_{i=1}^{n} \frac{(X_i - \mu)^2}{\sigma^6}) = -\frac{n}{2\sigma^4} + \frac{n\sigma^2}{\sigma^6} = \frac{n}{2\sigma^4}$$

$\Rightarrow$ *CRLB for* $\mu = \begin{pmatrix} 1 & 0 \end{pmatrix} I^{-1} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \frac{1}{I_{11}} = \frac{\sigma^2}{n}.$

$\Rightarrow$ *CRLB for* $\sigma^2 = \begin{pmatrix} 0 & 1 \end{pmatrix} I^{-1} \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \frac{1}{I_{22}} = \frac{2\sigma^4}{n}.$