STAT 5010: Advanced Statistical Inference

Lecturer: Tony Sit Lecture # 7
Scribe: Xiaoxuan Xia

1 Fisher Information

Recap:
Example (UMVUE for normal population variance)
Let $X_1, ..., X_n \sim^{i.i.d} N(\mu, \sigma^2)$ with both μ and σ^2 are unknown. Define $s^2 = \Sigma_{i=1}^n (X_i - \bar{X})^2$.

In this setting, $s^2/(n-1)$ is the UMNUE for σ^2. The MLE for σ^2 is s^2/n which has a lower mean squared error. In fact, the shrunk estimator $s^2/(n+1)$ has an even lower mean squared error. Therefore, neither UMVUE nor the MLE is admissible.

Question 1:
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 $g_1(\theta) + g_2(\theta)$?

If our underlying family of distributions has a complete sufficient statistic, then the answer if yes. (Because Lehman-scheffe Theorem).
Otherwise,...

Theorem 1: (TPE 2.1.7) (Characterization of UMVUEs)
Let $\Delta = \{\delta : E_\theta(\delta^2) < \infty\}$. Then $\delta_0 \in \Delta$ is UMVU for $g(\theta) = E(\delta_0)$ if an d only if $E(\delta_0(\theta), u) = 0$ for every $u \in \mathcal{U} = \{E(u) = 0\}$.

Proof 1: If δ_0 is an 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,

$$\begin{aligned} Var(\delta_\lambda) &= Var(\delta_0) + \lambda^2 Var(U) + 2\lambda cov(\delta_0, u) \\ &\geqslant Var(\delta_0) \end{aligned}$$

Consider the quadratic form $q(\lambda) = \lambda^2 Var(U) + 2\lambda cov(\delta_0, u)$.
The form q has the roots $\lambda = 0$ and $-2cov(\delta_0, u)/var(u)$.
If the roots are distinct, then the form must be negative at some point, which would violate the inequality above.
Hence, $-2cov(\delta_0, u)/var(u) = 0$ in which case, $E(u\delta_0) = cov(\delta_0, u) = 0$.

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

$$Var(\delta_0) = cov(\delta_0, \delta) \leqslant \sqrt{Var(\delta_0) Var(\delta)}$$

by Cauchy-Schwarz inequality. Hence, $Var(\delta_0) \leqslant Var(\delta)$ for any arbitrary unbiased estimator δ and δ_0. Hence, δ_0 is an UMVUE for $g(\theta)$.

Answer for Question 1: $\forall u \in \mathcal{U}$, $E((\delta_1 + \delta_2)u) = E(\delta_1 u) + E(\delta_2 u) = 0$. Therefore, $\delta_1 + \delta_2$ is an UMVUE for $g_1(\theta) + g_1(\theta))$.

2 Variance Bound and Information

Recall: $Cov(X,Y) \leq \sqrt{Var(X)Var(Y)}$

Using the covariance inequality, if δ is an unbiased estimator for $g(\theta)$ and ψ is an arbitrary random variable, then

$$Var_\theta(\delta) \geq \frac{Cov_\theta^2(\delta,\psi)}{Var_\theta(\psi)} \tag{1}$$

The trick here is to choose 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 $g(\theta)$.

Question 2: How to find proper ψ?

Let $\mathcal{P} = \{p_\theta : \theta \in \Omega\}$ be a dominated family with densities $p_\theta : \theta \in \Omega \in \mathbb{R}$.

To begin, $E_{\theta+\Delta}(\delta) - E_\theta(\delta)$ gives the same value $g(\theta+\Delta) - g(\theta)$, for any unbiased δ.
Here, Δ must be chosen so that $\theta + \Delta \in \Omega$.
Next, we write $E_{\theta+\Delta}(\delta) - E_\theta(\delta)$ as a covariance under p_θ.
This step involves the use of the "likelihood ratio". We assume here that $p_{\theta+\Delta}(x) = 0$ whenever $p_\theta(x) = 0$.
Define $L(x) = \frac{p_{\theta+\Delta}(x)}{p_\theta(x)}$ when $p_\theta(x) > 0$, and $L(x) = 1$ otherwise. We have

$$L(x)p_\theta(x) = \frac{p_{\theta+\Delta}(x)}{p_\theta(x)}p_\theta(x) = p_{\theta+\Delta}(x), a.e.x$$

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

$$\begin{aligned} E_{\theta+\Delta}h(x) = \int hp_{\theta+\Delta}d\mu &= \int hLp_\theta d\mu \\ &= E_\theta(L(x)h(x)). \end{aligned}$$

Take $h = 1$, $E_\theta L = 1$ (because $\int \frac{p_{\theta+\Delta}(x)}{p_\theta(x)}p_\theta(x)dx = \int_{\theta+\Delta} dx = 1$).
Take $h = \delta$, $E_{\theta+\Delta}\delta = E_\theta(L\delta)$, so if we define $\boldsymbol{\psi(x) = L(x) - 1}$ (**answer for Question 2**), then we can see that

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

and

$$E_{\theta+\Delta}(\delta) - E_\theta(\delta) = E_\theta(L\delta) - E_\theta(\delta) = E_\theta(\psi\delta) = Cov_\theta(\delta,\psi)$$

($E_\theta(\psi\delta) = Cov_\theta(\delta,\psi)$ because $\psi = L - 1$).

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 of Equation 1 can be written as:

$$Var_\theta(\delta) \geq \frac{\{g(\theta + \Delta) - g(\theta)\}^2}{Var_\theta(\psi)} = \frac{\{g(\theta + \Delta)\}^2}{E_\theta\left(\frac{p_{\theta+\Delta}(x)}{p_\theta(x)} - 1\right)^2}, \tag{2}$$

which is known as the *Hammersley−Chapman−Robbins inequality*.

Under suitable conditions, we can show that

$$\lim_{\Delta \to 0} \frac{\{\frac{g(\theta+\Delta)-g(\theta)}{\Delta}\}^2}{E_\theta\left(\frac{\{p_{\theta+\Delta}(x)-p_\theta(x)\}/\Delta}{p_\theta(x)}\right)^2} = \frac{(g^{'}(\theta)^2)}{E_\theta\left(\frac{\partial p_\theta(x)/\partial\theta}{p_\theta(x)}\right)^2}. \tag{3}$$

The denominator here is known as **Fisher Information**, denoted as $I(\theta)$ and is given by

$$I(\theta) = E_\theta\left(\frac{\partial \log p_\theta(x)}{\partial\theta}\right)^2 \tag{4}$$

With enough regularity to interchange integration and differentiation,

$$\begin{aligned} 0 = \frac{\partial}{\partial\theta}(1) &= \frac{\partial}{\partial\theta}\int p_\theta(x)d\mu(x) = \int \frac{\partial}{\partial\theta}p_\theta(x)d\mu(x) \\ &= \int \frac{\partial \log p_\theta(x)}{\partial\theta}p_\theta(x)d\mu(x) = E_\theta\left(\frac{\partial \log p_\theta(x)}{\partial\theta}\right) \end{aligned}$$

and so

$$I(\theta) = E_\theta\left(\frac{\partial \log p_\theta(x)}{\partial\theta}\right)^2 - \left\{E_\theta\left(\frac{\partial \log p_\theta(x)}{\partial\theta}\right)\right\}^2 = Var\left(\frac{\partial \log p_\theta(x)}{\partial\theta}\right). \tag{5}$$

Furthermore, since

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

We can see that

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

$$\Rightarrow I(\theta) = -E_\theta\left(\frac{\partial^2 \log p_\theta(x)}{\partial\theta^2}\right) \tag{6}$$

Therefore,

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

Theorem 2 Let $\mathcal{P} = \{p_\theta : \theta \in \Omega\}$ be a dominated family with Ω 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

$$Var_\theta(\delta) \geq \frac{\{g^{'}(\theta)\}^2}{I(\theta)}, \theta \in \Omega \tag{7}$$

This result is called the **Cramer−Rao**, or **information bound**.

Example (Exponential Families)
Let $\mathcal{P}$ be a one parameter exponential family in canonical form and density p_η given by

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

Then,

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

By the previous results, we have

$$I(\eta) = Var_\eta(T(x) - A^{'}(\eta)) = Var_\eta(T(x)) = A^{''}(\eta) \tag{8}$$

because $\frac{\partial^2 \log p_\eta(x)}{\partial \eta^2} = -A^{''}(\eta)$.

If the family is parameterized instead by $\mu = A^{'}(\eta) = E_\eta(T(x))$.
Then,

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

and so, because $A^{''}(\eta) = Var(T)$, we have

$$I(\mu) = \frac{1}{Var_\eta(T)} \tag{9}$$

observe also that because T is UMVUE for μ. The lower bound variance $Var_\mu(\delta) \geq 1/I(\mu)$ for an unbiased estimator δ of μ is sharp.

Example (Location Family)
Suppose q is an 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 $\theta + \varepsilon$ is called a **location family**.

$$\begin{aligned}\int g(x)dP_\theta(x) &= E_\theta(g(x)) = E_\theta(g(\theta + \varepsilon)) \\ &= \int g(\theta + \varepsilon)f(\varepsilon)d\varepsilon = \int g(x)f(x-\theta)dx\end{aligned}$$

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

$$
\begin{aligned}
I(\theta) &= E_\theta \left(\frac{\partial \log f(x-\theta)^2}{\partial \theta} \right)^2 = E\left(-\frac{f^{'}(x-\theta)}{f(x-\theta)} \right)^2 \\
&= E\left(\frac{f^{'}(\varepsilon)}{f(\varepsilon)} \right)^2 = \int \frac{\{f^{'}(x)\}^2}{f(x)} dx
\end{aligned} \tag{10}
$$

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

Question 3: If two (or more) independent vectors are observed, what is the total Fisher Information?

Answer to Question 3:
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 is

$$
I_X(\theta) = Var_\theta \left(\frac{\partial \log p_\theta(x)}{\partial \theta} \right).
$$

Correspondingly, the Fisher Information from Y is

$$
I_Y(\theta) = Var_\theta \left(\frac{\partial \log q_\theta(y)}{\partial \theta} \right).
$$

Then

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

Suppose we have $X_1, ..., X_n \overset{i.i.d}{\sim} p_\theta$, $I_{\mathbf{x}} = I_{X_1}(\theta) + \cdots + I_{X_n}(\theta) = nI_{X_1}(\theta)$.
Then

$$
Var_\theta(\delta) \geq \frac{g^{'}(\theta)}{nI(\theta)}. \tag{11}
$$

Multi-dimensional Fisher Information:
Suppose $\boldsymbol{\theta}$ takes values in $\mathbb{R}^k$, then the Fisher Information will become a matrix defined in regular case by

$$
\begin{aligned}
\{\boldsymbol{I}(\theta)\}_{i,j} &= E_\theta\left(\frac{\partial \log p_\theta(x)}{\partial \theta_i}\frac{\partial \log p_\theta(x)}{\partial \theta_j}\right) \\
&(E_\theta(\nabla_\theta \log p_\theta(x)) = 0) \\
&= Cov_\theta\left(\frac{\partial \log p_\theta(x)}{\partial \theta_i}, \frac{\partial \log p_\theta(x)}{\partial \theta_j}\right) \\
&= -E_\theta\left(\frac{\partial^2 \log p_\theta(x)}{\partial \theta_i \partial \theta_j}\right).
\end{aligned}
$$

$$
\begin{aligned}
\boldsymbol{I}(\boldsymbol{\theta}) &= E_{\boldsymbol{\theta}}(\{\nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(x)\}\{\nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(x)\}^\top) \\
&= Cov(\nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(x)) = -E_{\boldsymbol{\theta}}\nabla^2_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(x)
\end{aligned}
\tag{12}
$$

Where $\nabla_{\boldsymbol{\theta}}$ is the gradient with respect to $\boldsymbol{\theta}$ and $\nabla^2_{\boldsymbol{\theta}}$ is the Hessian matrix for the second order derivatives.

The lower bound for the variance of an unbiased estimator δ of $g(\boldsymbol{\theta})$, where $g : \Omega \to \mathbb{R}$ is

$$
Var_{\boldsymbol{\theta}}(\delta) \geq \{\nabla g(\boldsymbol{\theta})\}^T I^{-1}(\boldsymbol{\theta})\{\nabla g(\boldsymbol{\theta})\}. \tag{13}
$$

3 Next Lecture

Average Risk Optimality:
Originally, we have

$$
R(\theta, \delta) = E_\theta(L(\theta, \delta(x)))
$$

The average risk:

$$
r(\Lambda, \delta) = \int R(\theta, \delta) d\Lambda(\theta)
$$

where, $\Lambda(\theta)$ is the prior distribution on θ.
We aim to find

$$
r(\Lambda, \delta^*) \leq r(\Lambda, \delta).
$$