208711 Statistical Theory 1 **Fall 2019**

Lecture 22: November 1

Lecturer: Donlapark Pornnopparath

22.1 Information and variance bounds

From the previous lectures, we have discussed that an unbiased estimator with the lowest variance can be obtained, given that a complete sufficient statistic is known. We now consider more general situation i.e. the estimator might or might not be unbiased. First, we make an estimate on how low the variance of an estimator can be. For any two variables X and Y with $a = \mathbb{E}[X]$ and $b = \mathbb{E}[Y]$, Cauchy-Schwarz inequality yields

$$\begin{aligned}
\mathrm{Cov}(X,Y) &= \int (x-a)(y-b) f_X(x) f_Y(y)\ dxdy \\
&\leq \left(\int\int (x-a)^2 f_X(x) f_Y(y)\ dxdy \right)^{1/2} \left(\int\int (y-b)^2 f_X(x) f_Y(y)\ dxdy \right)^{1/2} \\
&= \left(\int (x-a)^2 f_X(x)\ dx \int f_Y(y)\ dy \right)^{1/2} \left(\int (y-b)^2 f_Y(y)\ dy \int f_X(x)\ dxdy \right)^{1/2} \\
&= [\mathrm{Var}(X)]^{1/2} [\mathrm{Var}(Y)]^{1/2}.
\end{aligned}$$

Replacing X with an unbiased estimator δ of $g(\theta)$, it follows that

$$\mathrm{Var}(\delta) \geq \frac{[\mathrm{Cov}(\delta, Y)]^2}{\mathrm{Var}(Y)}, \tag{22.1}$$

To get a uniform lower bound for $\mathrm{Var}[\delta]$ we have to come up with a clever choice of Y so that the right-hand side does not depend on δ. The variance lower bound that we will talk about relies on the concept of *information*. Statistically, this quantity should measure how much the random sample X tells us about the parameter θ. One way to do this is by looking at the shape of the likelihood function of X: the one with a narrow and sharply peaked likelihood can usually tell us more about θ than the one with shallow and spread-out likelihood. Since these discriminating characteristics can be measured by the second derivative of the likelihood function, we came up with the following definition of information:

Definition 22.1 (Fisher information)**.** Let X be a random variable from a distribution pdf $f(x|\theta)$which is twice-differentiable in θ. Then the Fisher information is given by

$$I(\theta) = \mathbb{E}_\theta \left[-\frac{\partial^2}{\partial \theta^2} \log \mathcal{L}(\theta|X) \right] = \mathbb{E}_\theta \left[-\frac{\partial^2}{\partial \theta^2} \log f(X|\theta) \right]. \tag{22.2}$$

Example 22.2. Let $X_1, X_2 \ldots, X_n$ be a random sample from Bernoulli(p), and we want to estimate p using the statistic $Y = \sum_i X_i$ which is distributed as Bin(n, p). We start with computing

$$\begin{aligned}
\frac{\partial^2}{\partial p^2} \log \binom{n}{y} p^y (1-p)^{n-y} &= \frac{\partial^2}{\partial p^2} \left[\log \binom{n}{y} + y \log p + (n-y) \log(1-p) \right] \\
&= -\frac{y}{p^2} - \frac{n-y}{(1-p)^2}.
\end{aligned}$$

Therefore, the Fisher information of Y is

$$I(p) = \mathbb{E}_p \left[\frac{y}{p^2} + \frac{n-y}{(1-p)^2} \right] = \frac{np}{p^2} + \frac{n-np}{(1-p)^2} = \frac{n}{p(1-p)},$$

which tells us that Y gives us more information about p at $p = 0$ or $p = 1$ than those around $p = 0.5$. ◇

We can write Fisher information in another form, assuming some regularity condition on the pdf.

Proposition 22.3. Let X be a random variable with twice-differentiable pdf $f(x|\theta)$. In addition, if we have that

$$\frac{d^2}{d\theta^2} \int f(x|\theta)\ dx = \frac{d}{d\theta} \int \frac{\partial}{\partial\theta} f(x|\theta)\ dx = \int \frac{\partial^2}{\partial\theta^2} f(x|\theta)\ dx \tag{22.3}$$

then the following identity holds:

$$I(\theta) = \mathbb{E}_\theta \left[-\frac{\partial^2}{\partial\theta^2} \log f(X|\theta) \right] = \mathbb{E}_\theta \left[\left(\frac{\partial}{\partial\theta} \log f(X|\theta) \right)^2 \right]. \tag{22.4}$$

Proof: Since $\int f(x|\theta)\ dx$ is a constant, so

$$\begin{aligned} 0 = \frac{d^2}{d\theta^2} \int f(x|\theta)\ dx &= \frac{d}{d\theta} \int \frac{\partial}{\partial\theta} f(x|\theta)\ dx \\ &= \frac{d}{d\theta} \int \frac{\frac{\partial}{\partial\theta} f(x|\theta)}{f(x|\theta)} f(x|\theta)\ dx \\ &= \frac{d}{d\theta} \int \left[\frac{\partial}{\partial\theta} \log f(x|\theta) \right] f(x|\theta)\ dx. \end{aligned}$$

By passing the derivative inside the integral and applying the product rule,

$$\begin{aligned} 0 &= \int \left[\frac{\partial^2}{\partial\theta^2} \log f(x|\theta) \right] f(x|\theta)\ dx + \int \frac{\partial}{\partial\theta} \log f(x|\theta) \frac{\partial}{\partial\theta} f(x|\theta)\ dx \\ &= -I(\theta) + \int \left[\frac{\partial}{\partial\theta} \log f(x|\theta) \right]^2 f(x|\theta)\ dx \\ &= -I(\theta) + \mathbb{E}_\theta \left[\left(\frac{\partial}{\partial\theta} \log f(X|\theta) \right)^2 \right], \end{aligned}$$

from which we obtain (22.4). ■

Intuitively, the more information about the true parameter we can get from an estimator δ should correspond to the lower variance of δ, and so we might anticipate the inverse relationship between $\mathrm{Var}_\theta(\delta)$ and $I(\theta)$. This is indeed the case, as shown by the main variance bound in this section.

Theorem 22.4 (Cramér-Rao inequality)**.** Consider a random sample $\boldsymbol{X}$ from a distribution with pdf $f(x|\theta)$. Let $\delta = \delta(\boldsymbol{X})$ be an estimator with $\mathbb{E}_\theta[\delta] = g(\theta)$. If $\mathrm{Var}_\theta(\delta) < \infty$ and for any integrable function $h(x)$,

$$\frac{d}{d\theta}\int h(x)f(x|\theta)\ dx = \int h(x)\frac{\partial}{\partial\theta}f(x|\theta)\ dx \tag{22.5}$$

for all $\theta \in \Theta$, then

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

where $I(\theta)$ is defined to be the rightmost expression in (22.4). If $f(x|\theta)$ also satisfies the regularity condition (22.3), then we can take $I(\theta)$ to be as defined in (22.2).

Proof: The idea is to apply inequality (22.1) and choose a random variable Y in order to deal with the troublesome term $\mathbb{E}_\theta[\delta Y]$ on the right-hand side. By picking $Y = \partial \log f(x|\theta)/\partial\theta$, we have $\mathbb{E}_\theta Y^2 = I(\theta)$ and it follows from (22.5) that

$$\begin{aligned}
\mathbb{E}_\theta\left[\delta Y\right] &= \int \delta(x)\left[\frac{\partial}{\partial\theta}\log f(x|\theta)\right] f(x|\theta)\ dx \\
&= \int \delta(x)\left[\frac{\partial f(x|\theta)/\partial\theta}{f(x|\theta)}\right] f(x|\theta)\ dx \\
&= \int \delta(x)\frac{\partial}{\partial\theta}f(x|\theta)\ dx \\
&= \frac{d}{d\theta}\int \delta(x)f(x|\theta)\ dx \\
&= g'(\theta).
\end{aligned}$$

In addition, Y has zero mean which also follows from (22.5):

$$\begin{aligned}
\mathbb{E}_\theta Y &= \int \left[\frac{\partial}{\partial\theta}\log f(x|\theta)\right] f(x|\theta)\ dx \\
&= \int \left[\frac{\partial f(x|\theta)/\partial\theta}{f(x|\theta)}\right] f(x|\theta)\ dx \\
&= \int \frac{\partial}{\partial\theta}f(x|\theta)\ dx \\
&= \frac{d}{d\theta}\int f(x|\theta)\ dx \\
&= 0,
\end{aligned}$$

since $\int f(x|\theta)\ dx = 1$ for all θ. We now employ the inequality (22.1) on δ and Y.

$$\begin{aligned}
\mathrm{Var}_\theta(\delta) \geq \frac{\left[\mathrm{Cov}(\delta, Y)\right]^2}{\mathrm{Var}_\theta(Y)} &= \frac{(\mathbb{E}_\theta[\delta Y] - \mathbb{E}_\theta\delta\mathbb{E}_\theta Y)^2}{\mathbb{E}_\theta Y^2 - (\mathbb{E}_\theta Y)^2} \\
&= \frac{(\mathbb{E}_\theta[\delta Y])^2}{\mathbb{E}_\theta Y^2} \\
&= \frac{[g'(\theta)]^2}{I(\theta)}
\end{aligned}$$

■

We give here one application of Cramér-Rao inequality.

Example 22.5. Let $X_1, X_2, \ldots, X_n$ be iid from Pois(λ). We have already seen that the sample mean $\overline{X}$ is a UMVU of λ via Rao-Blackwellization. An alternate way to show this is to use Cramér-Rao inequality. First, we have $g(\lambda) = \lambda$, so $g'(\lambda) = 1$. To compute the Fisher information, we start with

$$\log f(X_1, X_2, \ldots, X_n|\theta) = \log \prod_i f(X_i|\theta) = \sum_i \log f(X_i|\theta),$$

and therefore

$$\begin{aligned}
I(\lambda) &= \mathbb{E}_\lambda \left[-\frac{\partial^2}{\partial \lambda^2} \log f(X_1, X_2, \ldots, X_n|\theta) \right] \\
&= \mathbb{E}_\lambda \left[-\sum_i \frac{\partial^2}{\partial \lambda^2} \log f(X_i|\theta) \right] \\
&= n\mathbb{E}_\lambda \left[-\frac{\partial^2}{\partial \lambda^2} \log f(X|\theta) \right] \\
&= n\mathbb{E}_\lambda \left[-\frac{\partial^2}{\partial \lambda^2} \log \frac{e^{-\lambda}\lambda^X}{X!} \right] \\
&= n\mathbb{E}_\lambda \left[-\frac{\partial^2}{\partial \lambda^2} (-\lambda + X \log \lambda - \log X!) \right] \\
&= n\mathbb{E}_\lambda \left[\frac{X}{\lambda^2} \right] \\
&= \frac{n}{\lambda}.
\end{aligned}$$

By Cramér-Rao, $\text{Var}_\lambda \geq \lambda/n$, which implies that $\overline{X}$ is UMVU. $\diamond$