Lecture 10 10/7/20

Continued from last class...

$E[AB] = E[A]E[B] \leftarrow$ if A, B are independent

$$= \sum_{i=1}^{n} E[\ell'(\theta; x_i)^2] + \sum_{i \neq j} E[\ell'(\theta; x_i)\,\ell'(\theta; x_j)]$$

due to $x_1, \ldots, x_n$ iid

$$= \underbrace{n\,\underbrace{E[\ell'(\theta; x)^2]}_{I(\theta)}}_{I_n(\theta)} + \underbrace{\sum_{i \neq j} E[\ell'(\theta; x_i)]^{\to 0}\, E[\ell'(\theta; x_j)]^{\to 0}}_{0}$$

$$E[\hat{\theta} S] \overset{\text{lot z}}{=} E\left[\hat{\theta}\,\frac{\frac{\partial}{\partial\theta}[f(x_1, \ldots, x_n; \theta)]}{f(x_1, \ldots, x_n; \theta)}\right] = \int \cdots \int \hat{\theta}\,\frac{\frac{\partial}{\partial\theta}[f(x_1, \ldots, x_n; \theta)]}{f(x_1, \ldots, x_n; \theta)}\, f(x_1, \ldots, x_n; \theta)\, dx_1 \ldots dx_n$$

$\hat{\theta} = w(x_1, \ldots, x_n)$

Assuming we can interchange differentiation and integration...

$$= \frac{\partial}{\partial\theta}\Bigg[\underbrace{\int \cdots \int \hat{\theta} f(x_1, \ldots, x_n; \theta)\, dx_1, \ldots, dx_n}_{E[\hat{\theta}] = \theta}\Bigg] = \frac{\partial}{\partial\theta}[\theta] = 1$$

Putting it all together...

$$Var[\hat{\theta}] \geq \frac{(E[\hat{\theta} S]^{1} - E[\hat{\theta}]\,\cancel{E[S]}^{0})^2}{\underbrace{E[S^2]}_{nI(\theta)} - \cancel{E[S]}^2_{0}} = \frac{1}{nI(\theta)} = \frac{I(\theta)^{-1}}{n} \quad \text{CRLB.}$$

This allows you to compute the variance of the best estimator (UMVUE) for most iid DGPs (which means you can then assess if an estimator is a UMVUE). How? You calculate the CRLB and calculate the variance of the estimator. If the two are the same, then it is truly the best. Let's do some examples. First we need a fact...

$$I(\theta) := E[\ell'(\theta; x)^2] \underset{\text{HW}}{=} \ldots = E[-\ell''(\theta; x)]$$

Need to assume differentiation and integration can be interchanged just like in the proof of the CRLB.

$X_1, \dots, X_n \overset{iid}{\sim} Bern(\theta)$. $\hat{\theta} = \bar{X}$. Is this $\hat{\theta}$ the UMVUE?

$$\mathcal{L}(\theta; x) = \theta^x (1-\theta)^{1-x}$$

$$\ell(\theta; x) = x \ln(\theta) + (1-x)\ln(1-\theta)$$

$$\ell'(\theta; x) = \frac{x}{\theta} - \frac{1-x}{1-\theta}$$

$$\ell''(\theta; x) = \frac{-x}{\theta^2} - \frac{1-x}{(1-\theta)^2}$$

$$-\ell''(\theta; x) = \frac{x}{\theta^2} + \frac{1-x}{(1-\theta)^2}$$

$$I(\theta) = E[-\ell''(\theta; x)]$$

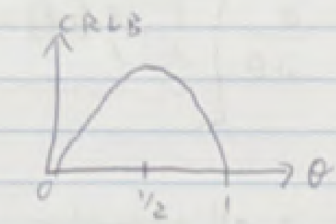
$$= E\left[\frac{x}{\theta^2} + \frac{1-x}{(1-\theta)^2}\right]$$

$$= \frac{E[x]}{\theta^2} + \frac{1-E[x]}{(1-\theta)^2}$$

$$= \frac{\theta}{\theta^2} + \frac{1-\theta}{(1-\theta)^2} = \frac{1}{\theta} + \frac{1}{1-\theta}$$

$$I(\theta) = \frac{1}{\theta(1-\theta)} \Rightarrow I(\theta)^{-1} = \theta(1-\theta)$$

$$CRLB = \frac{\theta(1-\theta)}{n} \Rightarrow$$



$$Var[\hat{\theta}] = Var[\bar{X}] = \frac{Var[X]}{n} = \frac{\theta(1-\theta)}{n}$$

$\hat{\theta}$ is the UMVUE!

$X_1, \dots, X_n \overset{iid}{\sim} N(\theta, \theta_2)$. $\hat{\theta} = \bar{X}$. Is this the UMVUE?

$$\mathcal{L}(\theta; x) = \frac{1}{\sqrt{2\pi\theta_2}} e^{-\frac{1}{2\theta_2}(x-\theta)^2}$$

$$\ell(\theta; x) = -\frac{1}{2}\ln(2\pi\theta_2) - \frac{1}{2\theta_2}(x^2 - 2\theta x + \theta^2) = -\frac{1}{2}\ln(2\pi\theta_2) - \frac{x^2}{2\theta_2} + \frac{\theta x}{\theta_2} - \frac{\theta^2}{2\theta_2}$$

$$\ell'(\theta; x) = \frac{x}{\theta_2} - \frac{\theta}{\theta_2}$$

$$\ell''(\theta; x) = -\frac{1}{\theta_2} \Rightarrow -\ell''(\theta; x) = \frac{1}{\theta_2}$$

$$I(\theta) = E\left[\frac{1}{\theta_2}\right] = \frac{1}{\theta_2} \Rightarrow I(\theta)^{-1} = \theta_2 \Rightarrow CRLB = \frac{\theta_2}{n}$$

$$Var[\hat{\theta}] = \frac{Var[x]}{n} = \frac{\theta_2}{n} \Rightarrow \hat{\theta} \text{ is the UMVUE!}$$

Where did we come from so far? We started with the question "given a DGP, how do we come up with an estimator for θ?" We had two procedures (1) MM and (2) MLE. Then we observed that sometimes they have different performances (in MSE). And we asked "what's the best performance?" Assuming an estimator is unbiased, we proved the best performance is given by the CRLB formula. If an estimator has the CRLB variance, it is the UMVUE (i.e. the very very best).

Let's go back to testing. Let's say you found the MM or the MLE and you want to test Ha. What do you need to do this? You need the "sampling distribution" (the distribution of $\hat{\theta}$) either approximately (for an approximate test) or exactly (for an exact test). We need to derive it...

Definition: an estimator $\hat{\theta}$ is "asymptotically normal" if:

$$\hat{\theta}_{std} = \frac{\hat{\theta} - \theta}{SE[\hat{\theta}]} \xrightarrow{d \leftarrow \text{convergence}} N(0,1)$$

This means as n gets large the $\hat{\theta}$-standardized distribution looks more and more like the $Z \sim N(0,1)$.

Is this possible to use the above as-is?
Hardly ever. Here's why.

* DGP $\overset{iid}{\sim}$ Bern(θ), $\hat{\theta} = \bar{X}$, $SE[\hat{\theta}] = \sqrt{\frac{\theta(1-\theta)}{n}}$

$$\overset{\text{by CLT}}{\Rightarrow} \frac{\hat{\theta} - \theta}{\sqrt{\frac{\theta(1-\theta)}{n}}} \xrightarrow{d} N(0,1)$$

What's wrong with the above expression? You do not know θ. In a testing setting, the null hypothesis will assume it. But in general, it is unknown.

In general, $SE[\hat{\theta}](\theta_1, \ldots, \theta_K)$ A quantity you need to know is a function of the things you can never know.

* DGP $\overset{iid}{\sim} N(\theta, \theta_2)$ $\hat{\theta} = \bar{X}$, $SE = \frac{\theta_2}{\sqrt{n}}$ ← unknown

We need an estimate of the Standard Error without assuming we know the θ's:

$\hat{SE}[\hat{\theta}](\hat{\theta}_1, ..., \hat{\theta}_K)$ ← function of estimates which come from the data. $\hat{SE}$ - is an estimate of SE.

* DGP $\overset{iid}{\sim}$ Bern (θ) $\hat{\theta} = \bar{X}$, $SE[\hat{\theta}] \approx \hat{SE}[\hat{\theta}] = \sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}$

Wouldn't it be nice if the following were true...

$$\frac{\hat{\theta} - \theta}{\hat{SE}[\hat{\theta}]} \xrightarrow{d} N(0,1)$$

This is true if the estimates employed in $\hat{SE}$ are "consistent."
Definition for this class: an estimator $\hat{\theta}$ is consistent if you can estimate it for any degree of precision you wish given large enough sample size (n).

$$\hat{\theta} \xrightarrow{p} \theta$$

this type of convergence is called "convergence in probability" and it's done at the end of 368. But we're not going to need to know it.

Here are two technical theorems.

Thm 5.5.4 p 233 C&B. Let $\hat{A}$ be a r.v. and c is a constant.
if $\hat{A} \xrightarrow{p} c$ then $h(\hat{A}) \xrightarrow{p} h(c)$ for h continuous

$$\Rightarrow \frac{\hat{A}}{c} = h(\hat{A}) \xrightarrow{p} h(c) = \frac{c}{c} = 1 \Rightarrow \underbrace{\frac{\hat{A}}{c} \xrightarrow{p} 1}_{\text{Fact 2}}$$

$$\hat{SE}[\hat{\theta}] = h(\hat{\theta}) \xrightarrow{p} h(\theta) = SE[\hat{\theta}] \Rightarrow \underbrace{\hat{SE}[\hat{\theta}] \xrightarrow{p} SE[\hat{\theta}]}_{\text{Fact 1}}$$

if $\hat{\theta} \xrightarrow{p} \theta$

Slutsky's Thm (Thm 5.5.17 p 239-240 C&B). Let $\hat{A}, \hat{B}$ be r.v.'s.

If $\hat{A} \xrightarrow{p} c$, $\hat{B} \xrightarrow{d} B \Rightarrow \hat{A}\hat{B} \xrightarrow{d} cB$.

$\hat{A} \xrightarrow{p} 1$ (by Fact 1 & 2)

$$\frac{\hat{\theta}-\theta}{\widehat{SE}[\hat{\theta}]} = \underbrace{\frac{SE[\hat{\theta}]}{\widehat{SE}[\hat{\theta}]}}_{\hat{A}} \underbrace{\frac{\hat{\theta}-\theta}{SE[\hat{\theta}]}}_{\hat{B}} \xrightarrow{d} 1N(0,1) = N(0,1)$$

Assume $\hat{B} \xrightarrow{d} N(0,1)$

We just proved that if $\hat{\theta}$ is asymptotically normal, the $\hat{\theta}_{std}$ with a consistent estimate of its standard error is ALSO asymptotically normal.

One of the most fundamental results in this class is the following: Under some technical conditions,

1. $\hat{\theta}^{MM}, \hat{\theta}^{MLE}$ are consistent
2. $\hat{\theta}^{MM}, \hat{\theta}^{MLE}$ are asymptotically normal where: $SE[\hat{\theta}^{MM}]$, $\widehat{SE}[\hat{\theta}^{MM}]$ are too difficult for this class but...

$$SE[\hat{\theta}^{MLE}] = \sqrt{\frac{I(\theta)^{-1}}{n}} \text{ i.e. the CRLB!!}$$

$$\widehat{SE}[\hat{\theta}^{MLE}] = \sqrt{\frac{I(\hat{\theta}^{MLE})^{-1}}{n}} \ *$$

3. $\hat{\theta}^{MLE}$ is called "asymptotically efficient" because as n gets large, it provides the SMALLEST possible variance. The MM does not...

Proof next class...