MATH 341/641 Lec 6

$\hat{\theta}_{MM}$ and $\hat{\theta}_{MLE}$ have different MSE's.

Is there a fundamental limit of estimation? A minimum MSE?

In general, no; the class of possible estimators is too large!

However, if you limit it to only the set of unbiased estimators, there is a bound called the Cramer-Rao Lower Bound (CRLB) discovered in 1945-46.

Unbiased Estimators that achieve the CRLB are called a "Uniformly minimum variance unbiased estimator" (UMVUE).

In order to prove the CRLB, we begin with the Covariance Inequality: For all r.v.'s A, B,

$$\Rightarrow \mathrm{Var}[A] \geq \frac{\mathrm{Cov}[A,B]^2}{\mathrm{Var}[B]} = \frac{\left(E[AB] - E[A]E[B]\right)^2}{E[B^2] - E[B]}$$

i.e. a lower bound on the variance of any r.v.

(from MATH 390)

Consider DGP: $X_1, \ldots X_n \overset{iid}{\sim} p(x;\theta)$ or $f(x;\theta)$

and any unbiased estimator $\hat{\theta}$.

Let $A = \hat{\theta}$, $B = S$, the "score function" for parameter θ

~~let $Q = \hat{\theta}$.~~ ~~And~~ Define the "score function" $S := \frac{\partial}{\partial \theta}\left[\ln f(x_1, \dots, x_n; \theta)\right]$ (def 1)

~~$\Rightarrow E(\hat{\theta}) = \theta$ assumption of unbiasedness~~

A property of the DGP, not any estimator.

by chain rule $= \dfrac{\frac{\partial}{\partial \theta}\left[f(x_1, \dots, x_n; \theta)\right]}{f(x_1, \dots, x_n; \theta)}$ (def 2)

by iid $= \frac{\partial}{\partial \theta}\left[\ln\left(\prod_{i=1}^{n} f(x_i; \theta)\right)\right]$ (def 3)

precalc $= \frac{\partial}{\partial \theta}\left[\sum_{i=1}^{n} \ln f(x_i; \theta)\right]$ (def 4)

linearity of derivative operator $= \sum_{i=1}^{n} \frac{\partial}{\partial \theta}\left[\ln\left(f(x_i; \theta)\right)\right]$ (def 5)

From def 1,
since $\mathcal{L} = f(x_1, \dots, x_n; \theta)$, $l := \ln(\mathcal{L})$ $= \frac{\partial}{\partial \theta}\left[l(\theta; x_1, \dots, x_n)\right]$ (def 6)

def of l' $= l'(\theta; x_1, \dots, x_n)$ (def 7)

by def 4 and linearity of deriv. operator $= \frac{\partial}{\partial \theta} \sum l(\theta; x_i) = \sum_{i=1}^{n} l'(\theta; x_i)$ (def 8)

All X's capital letters since S is a r.v.

We need to find $E[\hat{\theta} S]$, $E[S^2]$, $E[S]$ to prove the CRLB formula.

~~since $Cov(\hat{\theta}, S) := E(\hat{\theta} S) - E(\hat{\theta}) E(S)$ and $Var(S) = E(S^2) - E(S)^2$~~

~~We start with $E[S]$, then $E(S^2)$ then $E(\hat{\theta} S)$ and the substitute to find:~~

~~$$Var[\hat{\theta}] \geq \frac{E(QS) - E(Q)E(S)}{E(S^2) - E(S)^2}$$~~

by def S

by def of expectation of a r.v.

lowercase x's now

4

$$E[S] = E\left[\frac{\frac{\partial}{\partial\theta}(f(X_1 \ldots X_n;\theta))}{f(X_1 \ldots X_n;\theta)}\right] = \int_{S_X} \cdots \int_{S_X} \frac{\frac{\partial}{\partial\theta}[f(x_1,\ldots,x_n;\theta)]}{\cancel{f(x_1,\ldots,x_n;\theta)}} \cancel{f(x_1,\ldots,x_n;\theta)}\, dx_1 \ldots dx_n$$

Assume derivative and integral can be interchanged (Assum 1).

by definition of JDF/JMF

$$= \frac{\partial}{\partial\theta}\left[\int_{S_X} \cdots \int_{S_X} f(x_1,\ldots,x_n;\theta)\, dx_1 \ldots dx_n\right] = \frac{\partial}{\partial\theta}(1) = 0 \quad \text{(Fact 1a)}$$

$\Rightarrow Var(\hat{\theta}) \geq \frac{E[\hat{\theta} S]}{E[S^2]}$ here on our way!

def

by linearity of expectation

by iid

$$\Rightarrow E[S] = E[\ell'(\theta; X_1,\ldots,X_n)] = 0 \quad \text{and} \quad E[S] = E\left[\sum \ell'(\theta;X)\right] = n\,E[\ell'(\theta;X)] = 0$$

~~Sometimes in textbooks, single say~~

$$\Rightarrow E[\ell'(\theta;X)] = 0$$

def θ

$$(a_1 + \ldots + a_n)^2 = \sum a_i^2 + \sum_{i \neq j} a_i a_j \quad \text{(Fact 1b)}$$

$$Var[S] = E(S^2) - \cancel{E(S)^2}^{\,0} = E\left[\left(\sum_{i=1}^{n} \ell'(\theta;X_i)\right)^2\right] =$$

linearity of expectation

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

by iid

Recall $E[UV] = E[U]E[V]$ if U, V indep

$I(\theta) = E[S^2]$

$$= n\,E[\ell'(\theta;X_i)^2] + \sum_{i \neq j} \underbrace{E[\ell'(\theta;X_i)]}_{0}\,\underbrace{E[\ell'(\theta;X_j)]}_{0} = \underbrace{n\,E[\ell'(\theta;X_i)^2]}_{I_n(\theta)}$$

(by fact 1b)

$$\Rightarrow Var(\hat{\theta}) \geq \frac{E[\hat{\theta} S]}{n I(\theta)}$$

Sometimes Fisher Information is defined this way

Now ~~to~~ solve for $E[\hat{\theta} S] \stackrel{\text{def } S}{=} E\left[\hat{\theta}\, \frac{\frac{\partial}{\partial\theta}\left[f(X_1,\dots X_n;\theta)\right]}{f(X_1,\dots X_n;\theta)}\right]$

def of expectation of vector rv

$$= \int_{S_X} \cdots \int_{S_X} \hat{\theta}\, \frac{\frac{\partial}{\partial\theta}\left[f(x_1,\dots x_n;\theta)\right]}{\cancel{f(x_1,\dots x_n;\theta)}}\, \cancel{f(x_1,\dots x_n;\theta)}\, dx_1 \dots dx_n$$

Assumption 1

$$= \frac{\partial}{\partial\theta}\left[\int_{S_X} \cdots \int_{S_X} \hat{\theta}\, f(x_1,\dots x_n;\theta)\, dx_1 \dots dx_n\right] \stackrel{\text{def of expectation of vector rv}}{=} \frac{\partial}{\partial\theta}\left[E[\hat{\theta}]\right] \stackrel{\text{assumption of unbiasedness of } \hat{\theta}}{=} \frac{\partial}{\partial\theta}(\theta) = 1$$

we're done

Constant shrinking at rate n^{-1}!

$$\Rightarrow \operatorname{Var}[\hat{\theta}] \geq \frac{1}{n I(\theta)} = \frac{I(\theta)^{-1}}{n}$$

So Fisher Information is super important! It's a metric about the distribution itself that tells you how to make degree θ is able to be estimated. We will let it sink in for now by using it and then recognize it conceptually later...

Let's prove some estimators are UMVUE's! First, let's get a more convenient expression for Fisher Information:

$$I(\theta) := E\left[\ell'(\theta;X)^2\right] \stackrel{\text{Assumption 1 AND lots of work for HW}}{=} \cdots = E\left[-\ell''(\theta;X)\right]$$

Ex: $X_1, \dots X_n \stackrel{iid}{\sim} \text{Bern}(\theta)$, $\hat{\theta} = \bar{X}$, $\operatorname{Var}[\bar{X}] = \frac{\theta(1-\theta)}{n}$. Is this the optimal UMVUE? Let's compute the CRLB.

$$f(\theta;x) = p(x;\theta) = \theta^x (1-\theta)^{1-x}$$

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

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

$$I(\theta) = E[\;] = \frac{E(X)}{\theta^2} + \frac{E(1-X)}{(1-\theta)^2} = \frac{\cancel{\theta}}{\theta^{\cancel{2}}} + \frac{\cancel{1-\theta}}{(1-\theta)^{\cancel{2}}} = \frac{1}{\theta} + \frac{1}{1-\theta}$$

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

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

$$CRLB = \frac{I(\theta)^{-1}}{n} = \frac{\theta(1-\theta)}{n} \Rightarrow \hat{\theta} = \bar{X} \text{ is the UMVUE!}$$

Hooray!

Remember, there's no way to find the $\hat{\theta}$ with variance = CRLB. But if we find $\hat{\theta}$ with variance = CRLB, we found the best one!!

$X_1, \ldots, X_n \overset{iid}{\sim} N(\theta_1, \theta_2)$, $\hat{\theta}_1 = \bar{X}$. $Var[\bar{X}] = \frac{\theta_2}{n}$. Is it optimal?

$$f(\theta; x) = f(x; \theta) = \frac{1}{\sqrt{2\pi\theta_2}} e^{-\frac{1}{2\theta_2}(x-\theta_1)^2}$$

$$-\frac{x^2}{2\theta_2} + \frac{x\theta_1}{\theta_2} - \frac{\theta_1^2}{2\theta_2}$$

$$\ell(\theta; x) = -\tfrac{1}{2}\ln(2\pi) - \tfrac{1}{2}\ln(\theta_2) - \frac{1}{2\theta_2}(x-\theta_1)^2$$

$$\ell'(\theta; x) = \frac{x}{\theta_2} - \frac{\theta_1}{\theta_2}$$

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

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

no X!

$$I(\theta) = E[\cdot] = \frac{1}{\theta_2} \Rightarrow I(\theta)^{-1} = \theta_2$$

$$CRLB = \frac{I(\theta)^{-1}}{n} = \frac{\theta_2}{n} \Rightarrow \hat{\theta}_1 = \bar{X} \text{ is a UMVUE!}$$

Hooray!

Hw: θ_2, n $\overset{iid}{\sim} N(\theta_1, \theta_2)$

Hw: Binomial

Let's do a demo now to try and understand $\mathcal{L}, \ell, \ell', -\ell''$, $I(\theta)$ and how they're all related.

Inverse Fisher Information $I(\theta)^{-1}$ of a DGP measures a fundamental limit on the difficulty of estimating θ. If $I(\theta)^{-1}$ is large there's not a lot of "information" in X about θ. $\Leftrightarrow$ If $I(\theta)$ large $\Rightarrow$ there's a lot of information in X about θ.

The last of the 3 goals of statistical inference: confidence sets

Point estimation focused on best guess of θ e.g. $\hat{\theta} = .476$

Confidence sets focuses on a range of possible θ's e.g. [.475, .477] or [.42, .53]. The confidence sets answers the question "how sure are you of this pt. estimate $\hat{\theta}$? If set has a tight band $\Rightarrow$ more sure of $\hat{\theta}$. If wide bands $\Rightarrow$ more not sure of $\hat{\theta}$.

Define: an "interval estimate" is

$$\left[w_L(x_1, \ldots x_n),\ w_U(x_1, \ldots x_n)\right] = \left[\hat{\theta}_L, \hat{\theta}_U\right]$$

where w_L, w_U are two statistical functions s.t. $w_L < w_U$ for all possible datasets.

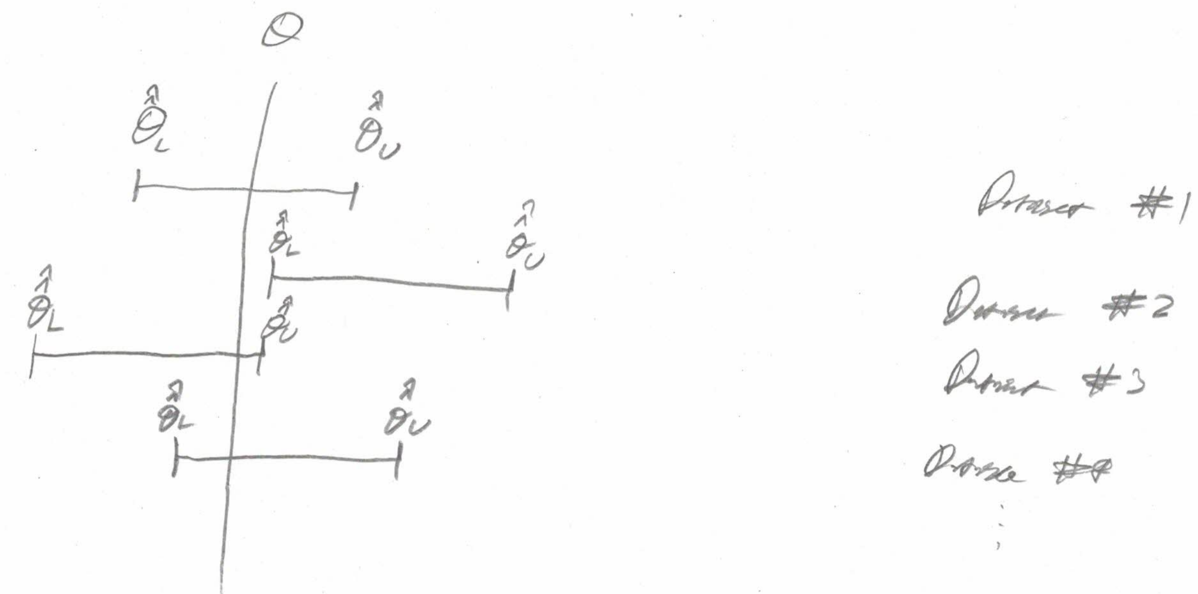
The An "interval estimator" is $\left[w_L(X_1, \ldots X_n),\ w_U(X_1, \ldots X_n)\right]$ which is a random interval $= \left[\hat{\theta}_L, \hat{\theta}_U\right]$

Define: The "coverage probability" of an interval estimator is

$$P\left(\theta \in \left[\hat{\theta}_L, \hat{\theta}_U\right] \mid \theta\right)$$

why given θ? Because you need to know the true distr's of $\hat{\theta}_L, \hat{\theta}_U$ to compute coverage explicitly.

Coverage prob. is best illustrated as follows:



The coverage prob. is computed over every possible dataset. If this were every dataset, cov. prob = 75%.

Def: A two-sided confidence interval estimator with cov. prob. $1-\alpha$ for param θ

$$\hat{CI}_{\theta,1-\alpha} := \left[\hat{\theta}_{\frac{\alpha}{2}}, \hat{\theta}_{1-\frac{\alpha}{2}}\right]$$

A two-sided confidence interval estimate (just "confidence interval") corresponding to the above confidence interval estimator is

$$\hat{CI}_{\theta,1-\alpha} := \left[\hat{\theta}_{\frac{\alpha}{2}}, \hat{\theta}_{1-\frac{\alpha}{2}}\right]$$