Minimum Variance Unbiased Estimation

Heading toward the Cramér-Rao Lower bound. The variance of an unbiased estimator cannot be any lower than this. If the variance of an unbiased estimator equals the C-R lower bound, it is a <u>Minimum</u> <u>variance</u> <u>unbias</u>ed <u>Estimator</u>.

It's a long story with a big payoff.

We often view the (log) likelihood as a function of the observed (fixed) values $x_1, \dots x_n$

But it can also be viewed as a function of $X_1, \dots X_n$

$\ell(\theta, \underset{\sim}{X})$ is a random variable.

$$\ell'(\theta, \underset{\sim}{X}) = \frac{\partial}{\partial\theta} \ln \prod_{i=1}^{n} f(X_i \mid \theta) = \frac{\partial}{\partial\theta} \sum_{i=1}^{n} \ln f(X_i \mid \theta)$$

$$= \sum_{i=1}^{n} \underbrace{\frac{\partial}{\partial\theta} \ln f(X_i \mid \theta)}_{Y_i} = \sum_{i=1}^{n} Y_i = S$$

Sum of iid random variables: Think CLT...

$X \sim f(x \mid \theta)$, $Y = \frac{\partial}{\partial\theta} \ln f(X \mid \theta)$

$$E(Y) = \int_{-\infty}^{\infty} \left(\frac{\partial}{\partial\theta} \ln f(x \mid \theta)\right) f(x \mid \theta)\, dx$$

Now, $\int_{-\infty}^{\infty} f(x \mid \theta)\, dx = 1 \implies \underbrace{\frac{\partial}{\partial\theta} 1}_{=0} = \frac{\partial}{\partial\theta} \int f(x \mid \theta)\, dx$

suppose

$$= \int \frac{\partial}{\partial\theta} f(x \mid \theta)\, dx = \int \frac{\frac{\partial}{\partial\theta} f(x \mid \theta)}{f(x \mid \theta)} f(x \mid \theta)\, dx$$

(Noting $\frac{\partial}{\partial\theta} \ln f(x \mid \theta) = \frac{1}{f(x \mid \theta)} \frac{\partial}{\partial\theta} f(x \mid \theta)$)

$$= \int \left(\frac{\partial}{\partial\theta} \ln f(x \mid \theta)\right) f(x \mid \theta)\, dx = E(Y) = 0$$

So that $E(\ell'(\theta, \underset{\sim}{X})) = E\left(\sum_{i=1}^{n} Y_i\right) = \sum_{i=1}^{n} E(Y_i) = 0$

Expected value of the derivative of the LL is zero

Have $E(Y) = E\left(\frac{\partial}{\partial\theta} \ln f(X|\theta)\right) = 0$

$$Var(Y) = E(Y^2) - E(Y)^2 = E(Y^2) = E\left(\left(\frac{\partial}{\partial\theta} \ln f(X|\theta)\right)^2\right)$$

Had $0 = \frac{\partial}{\partial\theta}\int f(x|\theta)\,dx = \int \frac{\partial}{\partial\theta} f(x|\theta)\,dx = \int \frac{\partial}{\partial\theta} \ln f(x|\theta)\, f(x|\theta)\,dx$

Differentiate again

$$\frac{\partial}{\partial\theta} 0 = 0 = \frac{\partial}{\partial\theta}\int \frac{\partial}{\partial\theta} \ln f(x|\theta)\, f(x|\theta)\,dx$$

$$\overset{\text{suppose}}{=} \int \frac{\partial}{\partial\theta}\left\{\left(\frac{\partial}{\partial\theta} \ln f(x|\theta)\right)\left(f(x|\theta)\right)\right\} dx \qquad u'v + v'u$$

$$= \int \left[\frac{\partial^2}{\partial\theta^2} \ln f(x|\theta) \cdot f(x|\theta) + \frac{\partial}{\partial\theta} f(x|\theta) \cdot \frac{\partial}{\partial\theta} \ln f(x|\theta)\right] dx$$

$$= \int \frac{\partial^2}{\partial\theta^2} \ln f(x|\theta) \cdot f(x|\theta)\,dx + \int \frac{\partial}{\partial\theta} f(x|\theta) \cdot \frac{\partial}{\partial\theta} \ln f(x|\theta)\,dx$$

$$= 0 \implies -E\left(\frac{\partial^2}{\partial\theta^2} \ln f(X|\theta)\right) = \int \frac{\partial}{\partial\theta} f(x|\theta) \cdot \frac{\partial}{\partial\theta} \ln f(x|\theta)\,dx$$

$$= \int \frac{\partial}{\partial\theta} f(x|\theta)\, \frac{\frac{\partial}{\partial\theta} f(x|\theta)}{f(x|\theta)}\,dx$$

$$= \int \frac{\frac{\partial}{\partial\theta} f(x|\theta)}{f(x|\theta)}\, \frac{\frac{\partial}{\partial\theta} f(x|\theta)}{f(x|\theta)}\, f(x|\theta)\,dx$$

$$= \int \left(\frac{\partial}{\partial\theta} \ln f(x|\theta)\right)^2 f(x|\theta)\,dx = E\left[\left(\frac{\partial}{\partial\theta} \ln f(X|\theta)\right)^2\right]$$

$$= E(Y^2) = Var(Y)$$

Have $E\left(\frac{\partial}{\partial\theta}\ln f(X|\theta)\right) = 0 = E(Y)$

$$\mathrm{var}\left(\frac{\partial}{\partial\theta}\ln f(X|\theta)\right) = \mathrm{Var}(Y)$$

$$= E\left[\left(\frac{\partial}{\partial\theta}\ln f(X|\theta)\right)^2\right] = -E\left(\frac{\partial^2}{\partial\theta^2}\ln f(X|\theta)\right)$$

$= I(\theta)$ The "Fisher information" in one observation.

Why <u>information</u> ?

Fisher noticed some likelihoods are more concentrated at the MLE

Downward curvature (measured by 2nd derivative) greater, then $l(\theta)$ carries more information about θ

$$\mathrm{Var}(l'(\theta, \underset{\sim}{X})) = \mathrm{Var}\left(\sum_{i=1}^{n} Y_i\right) \overset{\text{iid}}{=} \sum_{i=1}^{n} \mathrm{Var}(Y_i)$$

$= \sum_{i=1}^{n} I(\theta) = n\,I(\theta)$ Information in the whole sample

Have $S = \sum_{i=1}^{n} Y_i = \ell'(\theta, \underset{\sim}{X})$, $E(S)=0$, $Var(S) = n I(\theta)$

Cauchy-Schwarz Inequality says for <u>any</u> RVs
S & T $\quad (Cov(S,T))^2 \leq Var(S)\, Var(T)$

Let T be an unbiased estimator of θ:

$E(T) = \theta$ for all $\theta \in \Omega$

Cauchy-Schwarz says $Var(T) \geq \dfrac{Cov(S,T)^2}{Var(S)} = \dfrac{Cov(S,T)^2}{n I(\theta)}$

$Cov(S,T) = E(ST) - E(S)E(T) = E(ST) - 0$

<u>Consider</u> $\quad E(T) = \theta \implies \frac{\partial}{\partial\theta}\theta = \frac{\partial}{\partial\theta} E(T)$

$$\implies 1 = \frac{\partial}{\partial\theta} E(T) = \frac{\partial}{\partial\theta} \int \cdots \int t(\underset{\sim}{x}) \prod_{i=1}^{n} f(x_i|\theta)\, dx_1\, dx_2 \cdots dx_n$$

$$\overset{\text{suppose}}{=} \int \cdots \int \frac{\partial}{\partial\theta}\, t(\underset{\sim}{x})\, f_{\underset{\sim}{X}}(\underset{\sim}{x}|\theta)\, d\underset{\sim}{x}$$

$$= \int \cdots \int t(\underset{\sim}{x})\, \frac{\partial}{\partial\theta} f_{\underset{\sim}{X}}(\underset{\sim}{x}|\theta)\, d\underset{\sim}{x}$$

$$= \int \cdots \int t(\underset{\sim}{x})\, \frac{\frac{\partial}{\partial\theta} f_{\underset{\sim}{X}}(\underset{\sim}{x}|\theta)}{f_{\underset{\sim}{X}}(\underset{\sim}{x}|\theta)}\, f_{\underset{\sim}{X}}(\underset{\sim}{x}|\theta)\, d\underset{\sim}{x}$$

$$= \int \cdots \int t(\underset{\sim}{x}) \frac{\partial}{\partial\theta} \ln f(\underset{\sim}{x}|\theta) \cdot f(\underset{\sim}{x}|\theta)\, d\underset{\sim}{x}$$

$$= \int \cdots \int t(x) \left(\frac{\partial}{\partial\theta} \ln \prod_{i=1}^{n} f(x_i|\theta) \right) f(\underset{\sim}{x}|\theta)\, d\underset{\sim}{x}$$

$$= \int \cdots \int t(\underset{\sim}{x}) \left(\sum_{i=1}^{n} \frac{\partial}{\partial\theta} \ln f(x_i|\theta) \right) f(\underset{\sim}{x}|\theta)\, dx$$

$$= \int \cdots \int t(\underset{\sim}{x})\, s(\underset{\sim}{x})\, f(\underset{\sim}{x}|\theta)\, d\underset{\sim}{x}$$

$$= E(ST) = 1 = Cov(S, T)$$, so

$$Var(T) \geq \frac{1^2}{n\, I(\theta)}$$

This is the Cramér-Rao lower bound.

For any model that satisfies the regularity conditions

For any unbiased estimator T,

$$Var(T) \geq \frac{1}{n\, I(\theta)}$$, where

$$I(\theta) = E\left(\frac{\partial}{\partial\theta} \ln f(X|\theta)\right)^2 = -E\left(\frac{\partial^2}{\partial\theta^2} \ln f(X|\theta)\right)$$

Ex $X_1, \ldots X_n \overset{iid}{\sim}$ Poisson (λ) $\quad \hat{\lambda} = \bar{X}_n$

$$E(\bar{X}) = \mu = \lambda, \quad var(\bar{X}) = \frac{\lambda}{n}$$

Compare CR lower bound $\frac{1}{n I(\theta)}$

Two formulas for $I(\theta)$

1) $E\left(\frac{\partial}{\partial\theta} \ln f(x|\theta)\right)^2 = E\left(\frac{\partial}{\partial\lambda} \ln \frac{e^{-\lambda}\lambda^x}{x!}\right)^2$

$$= E\left(\frac{\partial}{\partial\lambda}[-\lambda + x\ln\lambda - \ln x!]\right)^2$$

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

$$= \frac{1}{\lambda^2} E(x-\lambda)^2 = \frac{1}{\lambda^2} Var(X) = \frac{\lambda}{\lambda^2} = \frac{1}{\lambda}$$

So lower bound for variance is $\frac{1}{n I(\lambda)} = \frac{1}{n(1/\lambda)} = \frac{\lambda}{n}$

$= Var(\hat{\lambda}) = Var(\bar{X})$, so

No other unbiased estimator can have a smaller variance, for <u>any</u> λ.

It's a <u>uniformly</u> <u>minimum</u> <u>variance</u> <u>unbiased</u> <u>estimator</u> (UMVUE).

2) The other formula for $I(\theta)$

$$I(\theta) = -E\left(\frac{\partial^2}{\partial\theta^2} \ln f(x \mid \theta)\right)$$

$$= -E\left(\frac{\partial^2}{\partial\lambda^2} \ln \frac{e^{-\lambda}\lambda^x}{x!}\right)$$

$$= -E\left(\frac{\partial^2}{\partial\lambda^2}\left(-\lambda + x\ln\lambda - \ln x!\right)\right)$$

$$= -E\left(\frac{\partial}{\partial\lambda}\left(-1 + x\lambda^{-1} - 0\right)\right)$$

$$= -E\left(x(-1)\lambda^{-2}\right) = -(-1)\frac{E(X)}{\lambda^2} = \frac{\lambda}{\lambda^2} = \frac{1}{\lambda}$$

MUCH EASIER!

Again, lower bound is $\dfrac{1}{n I(\theta)} = \dfrac{1}{n(1/\lambda)}$

$$= \frac{\lambda}{n} = Var(\bar{X}_n)$$

An example where the C-R lower bound does not apply

For the shifted exponential example of lecture 1

$$f(x|\theta) = e^{-(x-\theta)} I(x \geq \theta), \text{ and}$$

MOM $\hat{\Theta}_1 = \bar{X}_n - 1$, $E(\hat{\Theta}_1) = \theta$, $var(\hat{\Theta}_1) = \frac{1}{n}$

$\hat{\Theta}_3 = Min(X_i) - \frac{1}{n}$ $E(\hat{\Theta}_3) = \theta$, $var(\hat{\Theta}_3) = \frac{1}{n^2}$

The regularity conditions do not hold, because

$$\frac{d}{d\theta} 1 = \frac{d}{d\theta} \int f(x|\theta)\, dx = \frac{d}{d\theta} \int_\theta^\infty e^{-(x-\theta)}\, dx$$

$$\neq \int_\theta^\infty \frac{d}{d\theta} e^{-(x-\theta)}\, dx$$

Try it anyway. $I(\theta) \stackrel{?}{=} -E\left(\frac{d^2}{d\theta^2} \ln f(X|\theta)\right)$

$$= -E\left(\frac{d^2}{d\theta^2} \ln e^{-(X-\theta)}\right) = -E\left(\frac{d^2}{d\theta^2}(-1)(X-\theta)\right)$$

$$= E\left(\frac{d}{d\theta} 1\right) = E(0) = 0$$

Lower bound $= \dfrac{1}{n I(\theta)}$?

Trying the other formula,

$$I(\theta) \stackrel{?}{=} E\left(\frac{\partial}{\partial\theta} \ln f(X|\theta)\right)^2$$

$$= E\left(\frac{\partial}{\partial\theta}\left(-(X-\theta)\right)\right)^2$$

$$= E\left(\frac{\partial X}{\partial\theta} - \frac{\partial\theta}{\partial\theta}\right)^2 = E(0-1)^2 = 1$$

Gives lower bound of

$$\frac{1}{nI(\theta)} = \frac{1}{n} = Var(\hat{\Theta}_1) \quad \text{MOM}$$

But $\hat{\Theta}_3 = Min(X_i) - \frac{1}{n}$ is unbiased

with $Var(\hat{\Theta}_3) = \frac{1}{n^2}$

So don't use C-R lower bound when the support depends on Θ.

A central Limit Theorem for the MLE

Under the same regularity conditions as the Cramér - Rao inequality,

$$Z_n = \frac{\sqrt{n}\,(\widehat{\Theta}_n - \theta)}{\sqrt{1/I(\theta)}} \xrightarrow{d} Z \sim N(0,1)$$

Comments

- If $\widehat{\Theta}_n \sim N(\theta, \frac{1}{n\,I(\theta)})$ Exactly,

$$Z = \frac{\widehat{\Theta}_n - \theta}{\sqrt{1/n\,I(\theta)}} = \frac{\sqrt{n}\,(\widehat{\Theta}_n - \theta)}{\sqrt{1/I(\theta)}} \sim N(0,1)$$

Exactly

So it's useful to say $\widehat{\Theta}_n$ is asymptotically $N(\theta, \frac{1}{n\,I(\theta)})$.

- MLES reach the CR lower bound, at least approximately for large samples
- They are hard to beat.
- This result does not depend on an explicit formula for $\widehat{\theta}$.

More comments

$$Z_n = \frac{\sqrt{n}\,(\hat{\Theta}_n - \theta)}{\sqrt{1/I(\theta)}} \xrightarrow{d} Z \sim N(0,1)$$

OR

$$Z_n = \frac{\sqrt{n}\,(\hat{\Theta}_n - \theta)}{\sqrt{1/I(\hat{\Theta}_n)}} \xrightarrow{d} Z \sim N(0,1)$$

Ex $X_1, \dots, X_n \overset{iid}{\sim} \text{Geometric}(\theta)$

Want to test $H_0: \theta \le \frac{1}{2}$ vs $H_1: \theta > \frac{1}{2}$

$$\hat{\Theta}_n = \frac{1}{1 + \bar{X}_n}$$

A random sample of size $n = 100$ yields $\bar{x} = 0.85$ and $\hat{\theta} = \frac{1}{1+\bar{x}} = 0.54$

Obtain the p-value and carry out the test at $\alpha = 0.05$

$$I(\theta) = -E\left(\frac{\partial^2}{\partial\theta^2} \ln f(x\mid\theta)\right)$$

$$= -E\left(\frac{\partial^2}{\partial\theta^2} \ln\left((1-\theta)^x \theta\right)\right)$$

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

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

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

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

Formula sheet

$$\overset{\downarrow}{=} \quad \frac{(1-\theta)/\theta}{(1-\theta)^2} + \frac{1}{\theta^2} = \frac{1}{\theta(1-\theta)} + \frac{1}{\theta^2}$$

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

Testing $H_0: \theta \le \frac{1}{2}$ vs $H_1: \theta > \frac{1}{2}$ using
$\hat{\theta} = 0.54$, $n = 100$, $I(\theta) = \dfrac{1}{\theta^2(1-\theta)}$

$$I(\theta_0) = I(\tfrac{1}{2}) = \frac{1}{(\frac{1}{2})^2(1-\frac{1}{2})} = \frac{1}{\frac{1}{4}\cdot\frac{1}{2}} = 8$$

$$Z_n = \frac{\sqrt{100}\,(0.54-0.5)}{\sqrt{1/8}} = 1.13$$

p-value $= 1 - \text{pnorm}(1.13) = 0.129$

Don't reject H_0. Conclude $\theta \le \frac{1}{2}$

Another way, based on ordinary CLT

$$E(X_i) = \mu \overset{\substack{\text{Formula sheet}\\ \downarrow}}{=} \frac{1-\theta}{\theta} = \frac{1}{\theta} - 1$$ A decreasing function

$$H_0: \theta \le \tfrac{1}{2} \iff \frac{1}{\theta} - 1 \ge \frac{1}{1/2} - 1 = 2 - 1 = 1$$

So test $H_0: \mu \ge 1$ vs $H_1: \mu < 1$ with a lower tailed z-test

$$Z = \frac{\sqrt{n}\,(\bar{X}-1)}{\sqrt{(1-\theta_0)/\theta_0^2}} = \frac{\sqrt{100}\,(0.85-1)}{\sqrt{2}}$$

$$= -1.06, \quad p = \text{pnorm}(-1.06) = 0.145$$

Similar p-value, same conclusion

Rao - Blackwell Theorem

A systematic way to improve on the variance of an unbiased estimator (if it's possible).

Theorem (C.R. Rao & David Blackwell)

Let W be an unbiased estimator of θ & let T be sufficient for θ. Then

$\hat{\Theta} = E(W|T)$ is unbiased & $Var(\hat{\Theta}) \le Var(W)$

$\parallel$

$\int w \, f(w|T) \, dw$

A random variable

Proof uses

$$E(X) = E\{E(X|Y)\}$$

&

$$Var(X) = Var\{E(X|Y)\} + E\{Var(X|Y)\}$$

$E(X) = E\{E(X|Y)\}$ & $Var(X) = Var\{E(X|Y)\} + E\{Var(X|Y)\}$

Proof Let $E(W) = \theta$, T sufficient for θ,

$\hat{\Theta} = E(W|T)$

(a) $E(\hat{\Theta}) = E\{E(W|T)\} = E(W) = \theta$

(b) $Var(W) = Var\{E(W|T)\} + \underbrace{E\{Var(W|T)\}}_{\geq 0}$

$\geq Var\{\hat{\Theta}\}$ ▨

Ex $X_1, \dots, X_n \overset{iid}{\sim} B(\theta)$, $T = \sum_{i=1}^n X_i$ is sufficient

$W = X_1$, $E(W) = \theta$ unbiased

$\hat{\theta} = E(W|T=t)$ A scalar, but substitute T for t at the end to make it a RV

$$E(X_1 | \sum_{i=1}^n X_i = t) = P(X_1 = 1 | \sum_{i=1}^n X_i = t) = \frac{P(X_1 = 1, \sum_{i=1}^n X_i = t)}{P(\sum_{i=1}^n X_i = t)}$$

$$= \frac{P(X_1 = 1, \sum_{i=2}^n X_i = t-1)}{P(\sum_{i=1}^n X_i = t)} \overset{iid}{=} \frac{P(X_1 = 1) P(\sum_{i=2}^n X_i = t-1)}{P(\sum_{i=1}^n X_i = t)}$$

$$= \frac{\theta \binom{n-1}{t-1} \theta^{t-1} (1-\theta)^{n-t}}{\binom{n}{t} \theta^{t} (1-\theta)^{n-t}} = \frac{\dfrac{(n-1)!}{(t-1)!\,(n-t)!}}{\dfrac{n!}{t!\,(n-t)!}}$$

$$= \frac{t!\,(n-1)!}{(t-1)!\,n!} = \frac{t}{n}$$ substitute T for t to make it a R.V.

$$\frac{T}{n} = \frac{\sum_{i=1}^{n} X_i}{n} = \bar{X}_n = \hat{\Theta}_n$$

This example shows that even the simplest examples require hard work to calculate.

Main Moral: Make any (unbiased) estimator a function of a sufficient statistic if possible.