MLE III

STAT 24400
Lecture 11
February 11, 2016

Recall where we are:

Maximum Liklihood

θ parameter (to be found)

X or $X_1, \dots, X_n$ data

$L(\theta) = f(x|\theta)$ or $f(x_1, \dots, x_n|\theta)$ liklihood function (the model considered as function of θ)

MLE: $\hat{\theta} = \hat{\theta}(X) = \hat{\theta}(x_1, \dots, x_n)$ max's $L(\theta)$ depends on data, estimates θ.

Evaluation of estimators:

Judge by performance — how concentrated is their dist. around θ?

Measure by: $MSE(\hat{\theta}) = E(\hat{\theta}(x) - \theta)^2$

$= Var(\hat{\theta}(X)) + (Bias)^2$

$Bias(\hat{\theta}(X)) = E(\hat{\theta}(X)) - \theta$

So what can we say about the distribution of the random var $\hat{\theta}$?

Fisher's Approximation Theorem

$$\boxed{\text{If the MLE can be found from solving } \frac{d}{d\theta} L(\theta) = 0 \ \left(\text{or } \frac{d}{d\theta} \log L(\theta) = 0\right)^{\dagger}, \text{ Then } \hat{\theta} \text{ has an approximately}^{*} \ \mathcal{N}(\theta, \tau_n^2) \text{ distribution.}}$$

(* when n, the number of data points, is large).

($\dagger$ How could the MLE not be found from $\frac{d}{d\theta} L(\theta) = 0$? In general, because of differentiability problems, including the max not being in the interior of the domain.

This implies that:

$\hat{\theta}$ is approximately unbiased

$MSE(\hat{\theta})$ " " τ_n^2

How to find τ_n^2?

Consider first the iid case.

For X_i iid,

$$L(\theta) = \prod_{i=1}^{n} f(x_i|\theta),$$

$$\gamma_n^2 = \frac{\gamma^2}{n}, \quad \text{where}$$

$$\frac{1}{\gamma^2} = E\left[\frac{\partial}{\partial\theta} \log f(x_i|\theta)\right]^2$$

or

$$\frac{1}{\gamma^2} = -E\left[\frac{\partial^2}{\partial\theta^2} \log f(x_i|\theta)\right]$$

These are the same:

$$\int f(x|\theta) = 1, \text{ so}$$

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

It is identically true that

$$\frac{\partial}{\partial\theta} f(x|\theta) = \left[\frac{\partial}{\partial\theta} \log f(x|\theta)\right] f(x|\theta)$$

so

$$0 = \frac{\partial}{\partial\theta} \int f(x|\theta)\,dx = \int \left[\frac{\partial}{\partial\theta} \log f(x|\theta)\right] f(x|\theta)\,dx$$

daring assumption, justified with rigorous smoothness conditions not considered here.

Take a second derivative, and get

$$0 = \frac{\partial}{\partial\theta} \int \left[\frac{\partial}{\partial\theta} \log f(x|\theta)\right] f(x|\theta)\,dx$$

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

$$\boxed{0 = E\left[\frac{\partial^2}{\partial\theta^2} \log f(x|\theta)\right] + E\left(\frac{\partial}{\partial\theta} \log f(x|\theta)\right)^2}$$

Remark:

$$\frac{1}{\gamma_n^2} = E\left[\frac{\partial}{\partial\theta} \log f(x|\theta)\right]^2 = -E\left[\frac{\partial^2}{\partial\theta^2} \log f(x|\theta)\right]$$

is sometimes written

$$\frac{1}{\gamma_n^2} = I.$$

The textbook (Rice) uses that notation and never says why.

Here is the reason.

Imagine we are considering unbiased estimators, $\hat{\theta}$

so that $MSE(\hat{\theta}) = Var(\hat{\theta}) + \cancel{Bias}^{\to 0}$

and $MSE(\tilde{\theta}) = Var(\tilde{\theta}) + \cancel{Bias}^{\to 0}$

Both estimators are conditioned against the same data. We are likely to prefer $\hat{\theta}$ if it has a smaller variance than $\tilde{\theta}$, so

$$\text{eff}(\hat{\theta}, \tilde{\theta}) = \frac{Var(\tilde{\theta})}{Var(\hat{\theta})} > 1$$

and if $Var(\tilde{\theta}) = \frac{c_1}{n}$ and $Var(\hat{\theta}) = \frac{c_2}{n}$, we could use a smaller sample with $\tilde{\theta}$.

There is an upper limit to efficiency: <u>[Cramer - Rao Inequality]</u>

Let $X_1, \dots X_n$ be iid with density $f(x|\theta)$. Let $T = t(X_1, \dots, X_n)$ be an unbiased estimate of θ. Then

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

$I(\theta)$ is sometimes called "Fisher Information".

It is the **most** information you can squeeze out of a set of data. (Implicitly, it also reveals certain desirable math properties of the MLE).

Now, back to finding $\frac{1}{\gamma_n^2} = I$

Fisher's Approx. Thm:

If the MLE can be found from solving $\frac{d}{d\theta} \log L(\theta) = 0$, then

$\hat{\theta} \sim \mathcal{N}(\theta, \gamma_n^2)$.

1] Indep. Case $\left(L(\theta) = \prod_{i=1}^{n} f(x_i|\theta)\right)$

so $\gamma_n^2 = \frac{\gamma^2}{n}$, and

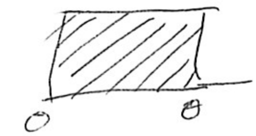
$$\frac{\gamma^2}{n} = E\left(\frac{d}{d\theta} \log f(x|\theta)\right)^2 = -E\left(\frac{d^2}{d\theta^2} \log f(x|\theta)\right)$$

2] More General Case,

$$\frac{1}{\gamma_n} = E\left[\frac{d}{d\theta} \log f(x_1, \ldots x_n|\theta)\right]^2 = -E\left[\frac{d^2}{d\theta^2} \log f(x_1, \ldots, x_n|\theta)\right]$$

[if indep, $E\left[\frac{d}{d\theta} \log f(x_1, \ldots x_n|\theta)\right]^2 = nE\left[\frac{d}{d\theta} \log f(x_1|\theta)\right]^2$]

Ex: Estimate the right hand boundary of uniform dist over $0 \le x < \theta$



$$f(x_i|\theta) = \begin{cases} \frac{1}{\theta} & 0 < x_i < \theta \\ 0 & \text{otherwise} \end{cases}$$

$\hat{\theta} = \max(x_i)$, but $\frac{d}{d\theta} L(\theta) \neq 0$ ever!!

Ex: Failure time; parametrize by λ (not $\bar{x}$!)

$$f(x_i|\lambda) = \begin{cases} \lambda e^{-\lambda x_i} & x_i > 0 \\ 0 & \text{otherwise} \end{cases}$$

$\hat{\lambda} = \frac{1}{\bar{x}}$, NOT unbiased.

(but found from $\frac{d}{d\lambda} \log L(\lambda) = 0$)

Finding γ^2

$$\frac{d}{d\lambda} \log f(x_1|\lambda) = \frac{d}{d\lambda}(\log \lambda - \lambda x_1) = \frac{1}{\lambda} - x_1$$

$$\frac{1}{\gamma^2} = E\left(\frac{1}{\lambda} - x_1\right)^2 = Var(x_1) = \frac{1}{\lambda^2}$$

$\longrightarrow \hat{\lambda}$ is approximately dist. $\mathcal{N}\left(\lambda, \frac{\lambda^2}{n}\right)$

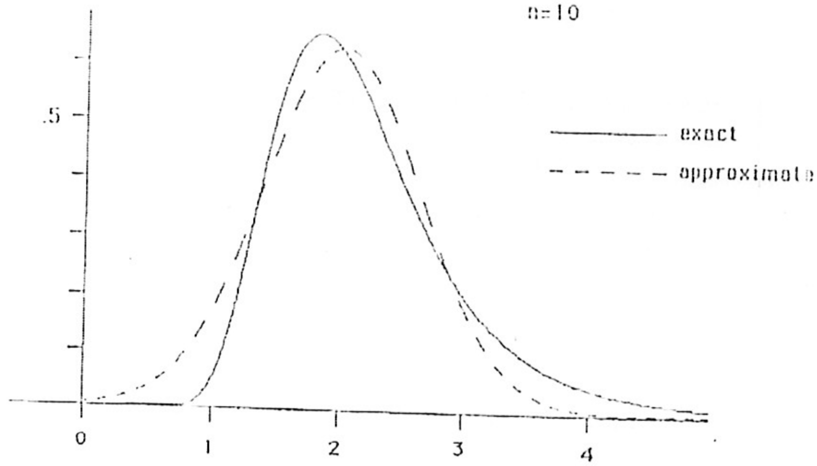
or take the 2^{nd} deriv of $\log f(x_1|\lambda)$:

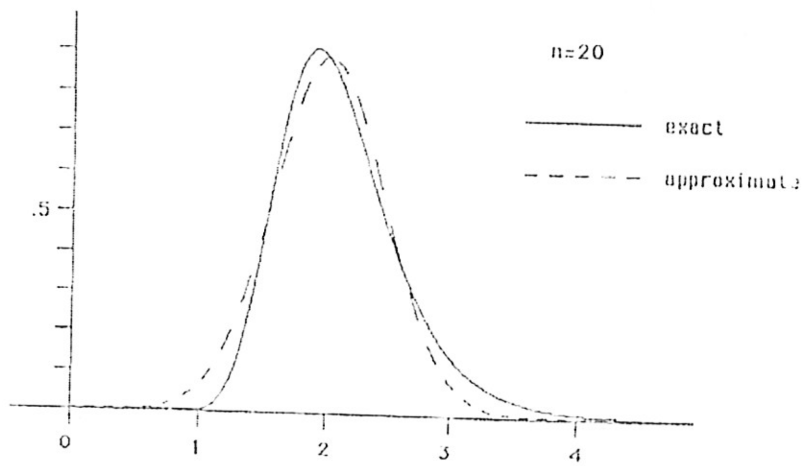
$$\frac{d^2}{d\lambda^2}(\quad) = \frac{d}{d\lambda}\left(\frac{1}{\lambda} - x_1\right) = \frac{-1}{\lambda^2} \qquad E\left(-\frac{1}{\lambda^2}\right) = \frac{-1}{\lambda^2}$$

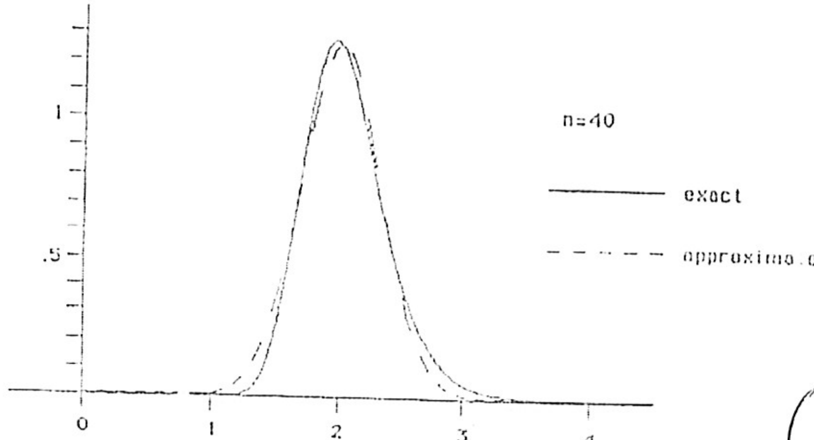
Figure 5.13

Distributions for $\hat{\lambda} = \frac{1}{\bar{X}}$

$\lambda = 2$







Ex: Binomial "Independent case"

$X_1, \ldots, X_n \quad P(X_i = 1|\theta) = 1 - P(X_i = 0|\theta) = \theta$

$$p(x_1, \ldots, x_n|\theta) = \prod_{i=1}^{n} p(x_i|\theta) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i}$$

$$\frac{d}{d\theta}\log p(x_i|\theta) = \frac{d}{d\theta}\left(x_i \log\theta + (1-x_i)\log(1-\theta)\right)$$

$$= \frac{x_i}{\theta} - \frac{(1-x_i)}{1-\theta}$$

$$= \frac{x_i(1-\theta) - \theta(1-x_i)}{\theta(1-\theta)}$$

$$= \frac{x_i - \theta}{\theta(1-\theta)}$$

$$E\left(\frac{d}{d\theta}\log p(x_i|\theta)\right)^2 = \frac{E(x_i-\theta)^2}{(\theta(1-\theta))^2} \qquad E(x_i-\theta)^2 \to Var(x_i) = \theta(1-\theta)$$

$$= \frac{1}{\theta(1-\theta)}$$

$$\Rightarrow \tau_n^2 = \frac{\theta(1-\theta)}{n}$$

or

$$\frac{d^2}{d\theta^2}\log p(x_i|\theta) = -\frac{x_i}{\theta^2} - \frac{(1-x_i)}{(1-\theta)^2}$$

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

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

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

$$E\left[\frac{d^2}{d\theta^2}\log p(x_i|\theta)\right]$$

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

$$= \frac{1}{\theta(1-\theta)}$$

Ex: Binomial "General Case"

X = # Successes n trials

$$p(x|\theta) = \binom{n}{x}\theta^{x}(1-\theta)^{n-x} = L(\theta)$$

$$\frac{d}{d\theta}\log L(\theta) = \frac{d}{d\theta}\left[\log\binom{n}{x} + x\log\theta + (n-x)\log(1-\theta)\right]$$

$$= \frac{x}{\theta} - \frac{n-x}{1-\theta} = \frac{X-n\theta}{\theta(1-\theta)}$$

$$E\left[\frac{d}{d\theta}\log L(\theta)\right]^2 = \frac{E(X-n\theta)^2}{(\theta(1-\theta))^2} \quad \text{Var}(X)$$

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

$$\Longrightarrow \quad \tau_n^2 = \frac{\theta(1-\theta)}{n}$$

Note: The variance (here also the MSE) of the MLE $\frac{X}{n}$ is exactly $\frac{\theta(1-\theta)}{n}$!

Ex. Genetic Linkage in Corn

$n = 3839$ seedlings in 4 classes

	Green	White
Starchy	1997	906
Sugary	904	32

$$= \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

Probs:

	Green	White	
Starchy	$(2+\theta)/4$	$(1-\theta)/4$	$3/4$
Sugary	$(1-\theta)/4$	$\theta/4$	$1/4$
	$3/4$	$1/4$	

θ = a coeff of linkage, unknown.
n indep trials, each with these probabilities ($\theta = \frac{1}{4}$ if no linkage)

Liklihood:

$$L(\theta) = \left(\frac{2+\theta}{4}\right)^a \left(\frac{1-\theta}{4}\right)^b \left(\frac{1-\theta}{4}\right)^c \left(\frac{\theta}{4}\right)^d$$

$$= (2+\theta)^a (1-\theta)^{b+c} \theta^d \Big/ 4^n$$

$$\log L(\theta) = a \log(2+\theta) + (b+c)\log(1-\theta) + d \log \theta$$

$$\frac{d}{d\theta} \log L(\theta) = \frac{a}{2+\theta} - \frac{b+c}{1-\theta} + \frac{d}{\theta}.$$

set $= 0$: $\quad \dfrac{a}{2+\theta} - \dfrac{b+c}{1-\theta} + \dfrac{d}{\theta} = 0$

$\Rightarrow$ quadratic in θ

$\hat{\theta} = 0.0357$ (only positive root)

Find γ_n:

(use general case)

Find $\frac{d^2}{d\theta^2} \log L(\theta)$

$$\frac{d}{d\theta} \log L(\theta) = \frac{a}{2+\theta} - \frac{b+c}{1-\theta} + \frac{d}{\theta}$$

$$\frac{d^2}{d\theta^2} \log L(\theta) = \frac{-a}{(2+\theta)^2} - \frac{b+c}{(1-\theta)^2} - \frac{d}{\theta^2}$$

Take expectations: $E(a) = \frac{2+\theta}{4} \cdot n$

$E(b) = E(c) = \frac{1-\theta}{4} \cdot n \qquad E(d) = \frac{\theta}{4} \cdot n$

$$-E\left(\frac{d^2}{d\theta^2} \log L(\theta)\right) = \frac{n}{4}\left[\frac{1}{2+\theta} + \frac{2}{1-\theta} + \frac{1}{\theta}\right]$$

So, $$\gamma_n^2 = \frac{4}{n}\left(\frac{1}{1/(2+\theta) + 2/(1-\theta) + 1/\theta}\right)$$

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

Plugging in $\tilde{\theta}$, we get

$$\approx \frac{.13}{n}$$

Now let's do it for the indep

case (1): $\begin{pmatrix} a & c \\ b & d \end{pmatrix} = \sum_{i=1}^{n} \begin{pmatrix} x_{i1} & x_{i2} \\ x_{i3} & x_{i4} \end{pmatrix}$

Poss X_i	$f(x_i \mid \theta)$	$\log f(x_i \mid \theta)$	$\frac{d}{d\theta} \log f$	$\frac{d^2}{d\theta^2} \log f$
$\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$	$\frac{(2+\theta)}{4}$	$\log(2+\theta)+c$	$\frac{1}{(2+\theta)}$	$\frac{-1}{(2+\theta)^2}$
$\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$	$\frac{(1-\theta)}{4}$	$\log(1-\theta)+c$	$\frac{-1}{(1-\theta)}$	$\frac{-1}{(1-\theta)^2}$
$\begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}$	$\frac{(1-\theta)}{4}$	$\log(1-\theta)+c$	$\frac{-1}{(1-\theta)}$	$\frac{-1}{(1-\theta)^2}$
$\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$	$\frac{\theta}{4}$	$\log(\theta)+c$	$+\frac{1}{\theta}$	$-\frac{1}{\theta^2}$

$$E\left(\frac{d^2}{d\theta^2} \log f\right) = \frac{2+\theta}{4} \cdot \left(\frac{-1}{(2+\theta)^2}\right) + \frac{1-\theta}{4}\left(\frac{-1}{(1-\theta)^2}\right)$$

$$+ \frac{(1-\theta)}{4}\left(\frac{-1}{(1-\theta)^2}\right) + \frac{\theta}{4}\left(\frac{-1}{\theta^2}\right)$$

$$= \frac{1}{4(2+\theta)} - \frac{1}{2(1-\theta)} - \frac{1}{4\theta} = \frac{-1}{\gamma^2}$$

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

Another estimator:

$$\hat{\theta}^* = \frac{a - b - c + d}{n}$$

$$E(\hat{\theta}^*) = \frac{E(a) - E(b) - E(c) + E(d)}{n}$$

$$= \frac{n\left(\frac{2+\theta}{4}\right) - n\left(\frac{1-\theta}{4}\right) - n\left(\frac{1-\theta}{4}\right) + n\left(\frac{\theta}{4}\right)}{n}$$

$$= \theta \qquad \underline{\text{unbiased}}$$

It can be shown (and will, shortly) that

$$Var(\hat{\theta}^*) = \frac{1-\theta^2}{n} = \frac{(1-\theta)(1+\theta)}{n}$$

<u>Then</u>

$$\frac{Var(\text{MLE } \hat{\theta})}{Var(\hat{\theta}^*)} \simeq \frac{2\theta(1-\theta)(2+\theta)}{(1+2\theta)(1-\theta)(1+\theta)}$$

$$= \frac{2\theta^2 + 4\theta}{2\theta^2 + 3\theta + 1}$$

$$< 1 \quad \text{since } \theta < 1, \text{ so } 4\theta < 3\theta + 1$$

Cramer-Rao in action...