STAT 24400
1/28/16
Lecture 8

Maximum Likelihood,
Continued. (from p. 13 of Lecture 7)

Example

Consider the polling example from the last two lectures, where the data $X \sim Bin(n, \theta)$. We want to estimate θ. Here are three possible estimators:

I. $\hat{\theta} = \frac{X}{n}$

$$E(\hat{\theta}) = \frac{E(X)}{n} = \frac{(n\theta)}{n} = \theta$$ unbiased

II $\theta^* = \frac{X+1}{n+2}$ where did this come from?

Well, if were Bayesians and the prior were Uniform(0,1). Then if $X = x$, the posterior

$\theta \sim Beta(x+1, n-x+1)$, and

$$E(\theta \mid X = x) = \frac{x+1}{n+2}$$

III. Consider a third estimator, the "stopped clock" estimator,

$\hat{\theta}^{**} = \frac{1}{2}$. It can be very accurate (or not!). It does not depend on data.

Note from the last lecture that we can write $\hat{\theta}^{*}$ as a weighted sum of the other estimators, so that

$$\hat{\theta}^{*} = \left(\frac{n}{n+2}\right)\hat{\theta} + \left(\frac{2}{n+2}\right)\hat{\theta}^{**}$$

Recall that $E(\hat{\theta}) = \theta$; it was unbiased.

$$E(\hat{\theta}^{*}) = \frac{E(x)+1}{n+2} = \frac{n\theta+1}{n+2}, \text{ Biased.}$$

$$E(\hat{\theta}^{**}) = 0.5 - \theta, \text{ Biased (Doh!)}$$

What about the mean error and mean squared error?

I. $\hat{\theta}$: mean error (see Stigler for details):

$$E\left[|\hat{\theta}-\theta|\right] = 2\binom{n-1}{\lfloor n\theta \rfloor}\theta^{\lfloor n\theta \rfloor+1}(1-\theta)^{n-\lfloor n\theta \rfloor}$$

($\lfloor n\theta \rfloor$ the largest integer smaller than $n\theta$)

mean squared error:

$$\begin{aligned}
MSE_{\hat{\theta}}(\theta) &= E(\hat{\theta}-\theta)^2 \\
&= E\left(\frac{X}{n}-\theta\right)^2 \\
&= Var\left(\frac{X}{n}\right) \\
&= \frac{Var(X)}{n^2} \\
&= \frac{n\theta(1-\theta)}{n^2} \\
&= \frac{\theta(1-\theta)}{n}
\end{aligned}$$

binomial dist

II $\hat{\theta}^*$: $E\left[|\theta^*-\theta|\right] = \sum_{k=0}^{n}\left|\left(\frac{k+1}{n+2}\right)-\theta\right| b(k;n,\theta)$

(evaluate numerically!)

$\hat{\theta}^*$, continued:

$$MSE_{\hat{\theta}^*} = Var(\hat{\theta}^* | \theta) + (Bias_{\hat{\theta}^*}(\theta))^2$$

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

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

[note: $Var(\hat{\theta}^*) < Var(\hat{\theta}) = \frac{n\theta(1-\theta)}{n^2}$]

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

III. $\hat{\theta}^{**}$:

mean error:

$$E(|\hat{\theta}^{**} - \theta|) = \left|\frac{1}{2} - \theta\right|$$

mean squared error:

$$MSE_{\hat{\theta}^{**}}(\theta) = \left(\frac{1}{2} - \theta\right)^2 = \frac{1}{4} - \theta(1-\theta)$$

interval where $\hat{\theta}^{**}$ is better than $\hat{\theta}$

$n =$	Mean Error	Mean Squared Error
1	$.29 < \theta < .71$	$.15 < \theta < .85$
4	$.40 < \theta < .60$	$.28 < \theta < .72$
25	$.46 < \theta < .54$	$.40 < \theta < .60$
100	$.48 < \theta < .52$	$.45 < \theta < .55$

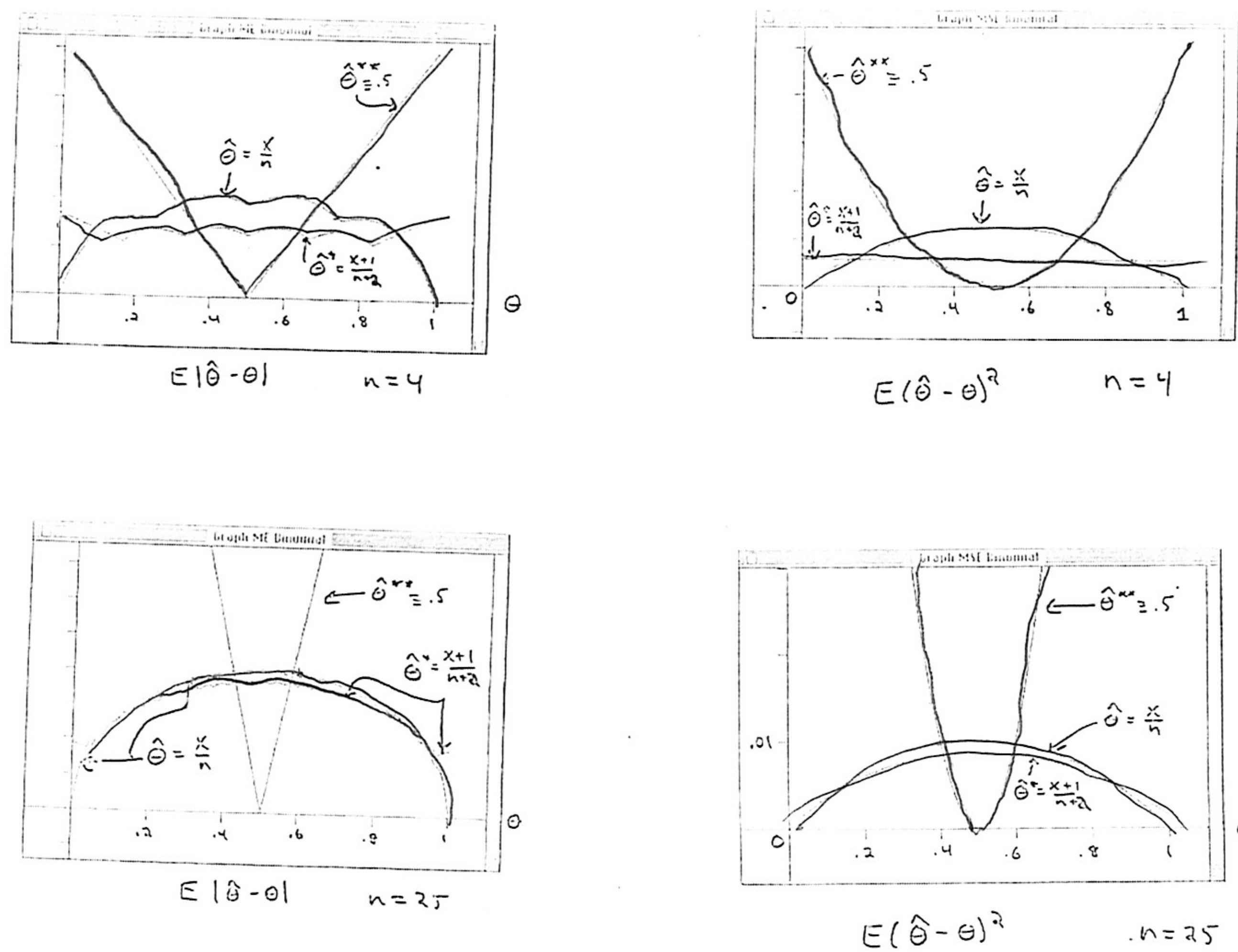

Graph MSE Binomial
$\hat{\theta} = \frac{X}{n}$
$\hat{\theta}^{*} = \frac{X+1}{n+2}$
$\hat{\theta}^{**} \equiv .5$
θ
$E|\hat{\theta} - \theta|$ $n = 4$
$E(\hat{\theta} - \theta)^2$ $n = 4$
$E|\hat{\theta} - \theta|$ $n = 25$
$E(\hat{\theta} - \theta)^2$ $n = 25$
.01

Figure 5.3

Where to get estimators?
One source is Maximum Likelihood

Recall Bayes: $f(\theta/x) \propto f(\theta) f(x/\theta)$

Might want the "most likely" θ:
θ that max's $f(\theta/x)$ (ie, max's $f(\theta)f(x/\theta)$)
But $f(\theta)$ is <u>not</u> <u>available</u>.
If $f(\theta)$ is flat (ie, approximately constant $\equiv$ not much prior info) could maximize $f(x/\theta)$ instead, ie find θ to make $f(x/\theta)$ as large as possible for the given data x — call <u>that</u> $\hat{\theta}$.

<u>Definition</u>. We call $L(\theta) = f(x/\theta)$ (or $L(\theta) = f(x_1, \ldots, x_n/\theta)$), viewed as a function of θ, the <u>Liklihood</u> <u>function</u>. The value of θ, say $\hat{\theta}$, for which $L(\theta)$ acheives its max is called the <u>maximum liklihood estimate</u> of θ.

Interpreting $L(\theta)$

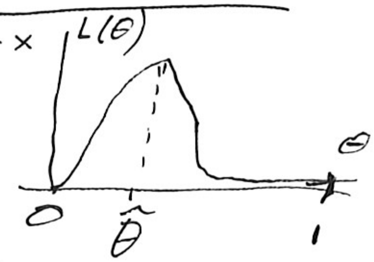
Bayesians: $L(\theta) = f(x|\theta)$ is proportional to the posterior distribution of θ given x

if $f(\theta) =$ constant.

Others: $L(\theta)$ is the probability (density) that we observe the data we actually observed *if* the true state of nature is θ.

Hence the MLE $\hat{\theta}$ is the "state of nature" that best explains our data, for which it is most likely.

Example: $L(\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$



Maxing $L(\theta)$ same as max'ing $\log(L(\theta)) = \log\binom{n}{x} + x\log\theta + (n-x)\log(1-\theta)$

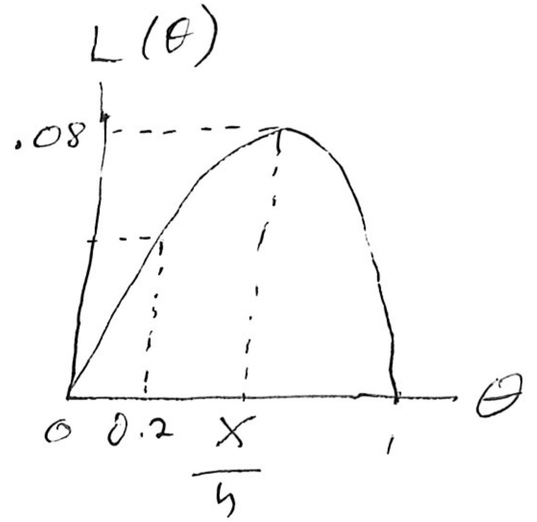
$$\frac{d}{d\theta}\log L(\theta) = 0 + \frac{x}{\theta} - \frac{(n-x)}{1-\theta};\ \text{set} = 0 \rightarrow \boxed{\hat{\theta} = \frac{x}{n}}$$

min or max ?

$$\frac{d^2}{d\theta^2}\log L(\theta) = -\frac{x}{\theta^2} - \frac{(n-x)}{(1-\theta)^2} < 0$$

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

is maximum for $\hat{\theta} = \frac{x}{n}$



$$\frac{d}{d\theta} L(\theta) = 0$$

solve, then:

$$\frac{d^2}{d\theta^2} L(\hat{\theta}) < 0$$

check

$L(\theta) = Pr$ (observed data | state of nature is θ) is largest for $\theta = \hat{\theta} = \frac{x}{n}$.

$\hat{\theta} = \frac{x}{n}$ "best" explains data.

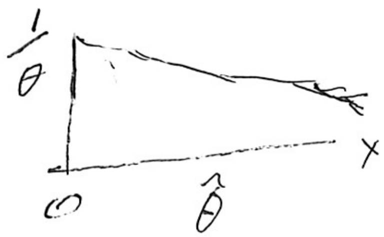
We are more likely to get 40 of 100 for Bernie if the true fraction were 0.4 than if it were any other value (.2, .5, etc). Note that $L(\hat{\theta})$ need not be large - $L(0.4) = 0.0812$ here.

Note: the MLE does not necessarily have good "sampling" properties - it does not necessarily have smallest MSE.

But: It can be proved to be often nearly best in a certain sense with large samples.

Example: Estimating Average Failure Time

Suppose a component has a constant probability of failure, so it lasts a time X with dist.



$$f(x/\theta) = \begin{cases} \frac{1}{\theta} e^{-x/\theta} & \theta > 0,\ x > 0 \\ 0 & \text{otherwise} \end{cases}$$

$$E(x) = \int_0^\infty x \cdot \frac{1}{\theta} e^{-x/\theta} dx = \theta$$

Problem: The mean failure time θ is unknown. n components are tested independently, to observe $X_1, X_2, \ldots, X_n$. Estimate θ.

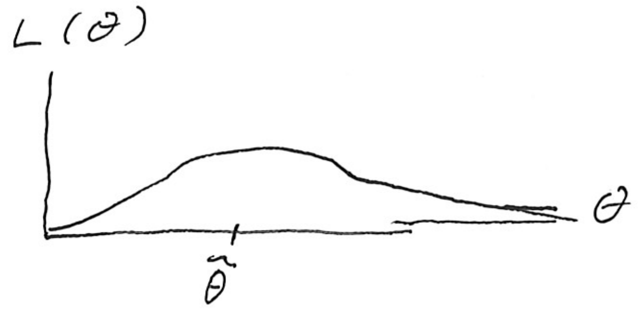
The joint density $f(x_1, \ldots, x_n / \theta) = f(x_1/\theta) f(x_2/\theta) \ldots f(x_n/\theta)$ (by independence)

$$= \frac{1}{\theta} e^{-\frac{x_1}{\theta}} \cdot \frac{1}{\theta} e^{-\frac{x_2}{\theta}} \cdots$$

$$= \frac{1}{\theta^n} \cdot e^{-\frac{(x_1 + \cdots + x_n)}{\theta}}$$

iff all x's > 0 (otherwise zero)

So $L(\theta) = \frac{1}{\theta^n} e^{-\Sigma x_i / \theta}$, $\theta > 0$



You can think of $L(\theta)$ as giving "relative liklihood" of data for different values of θ

want to max $L(\theta)$. So:

$$\max \log L(\theta) = -n \log \theta - \sum X_i / \theta$$

$$\frac{d}{d\theta} \log L(\theta) = -\frac{n}{\theta} + \frac{\sum X_i}{\theta^2}$$

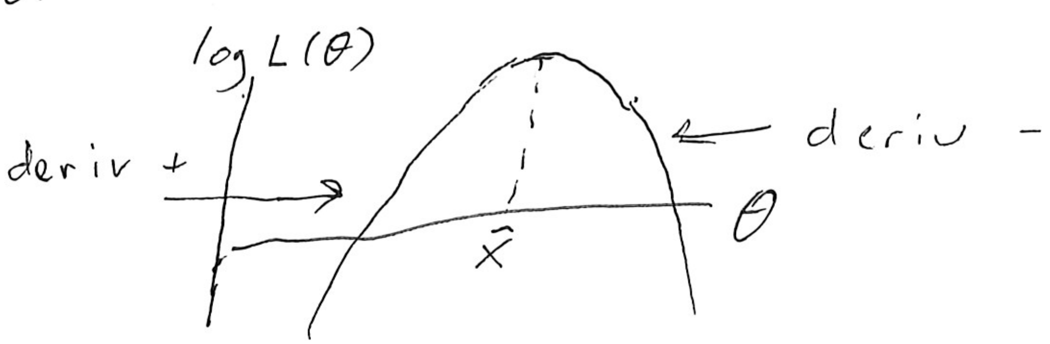
Set $= 0$:
$$-\frac{n}{\hat{\theta}} + \frac{\sum X_i}{\hat{\theta}^2} = 0$$

$$\Rightarrow \hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} X_i = \bar{X}.$$

Check:

$$\frac{d}{d\theta} \log L(\theta) = \frac{n}{\theta} \left(\frac{\bar{X}}{\theta} - 1 \right)$$

For $\theta < \bar{X}$ this is > 0
For $\theta > \bar{X}$ this is < 0



Summing up: $f(x_i|\theta) = \frac{1}{\theta} e^{-\frac{x_i}{\theta}} \quad x_i > 0$

Data: $X_1, \ldots, X_n$ indep

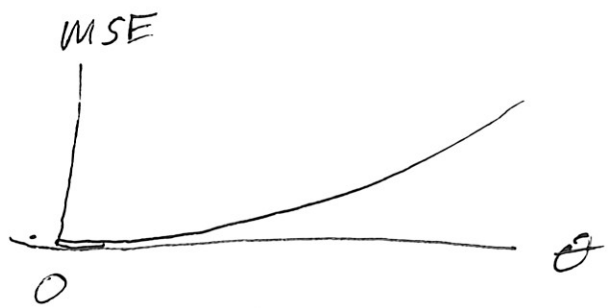
MLE $\hat{\theta} = \bar{X}$

$E(\hat{\theta}) = E(\bar{X}) = E(X_1) = \theta$

<u>unbiased</u>

So $MSE = Var(\hat{\theta}) = Var(\bar{X}) = \frac{Var(X_i)}{n}$

$= \boxed{\frac{\theta^2}{n}}$



So, the MSE <u>decreases</u> as n increases; it <u>increases</u> as θ increases

What about the exponential—

Say $f(x_i | \lambda) = \lambda e^{-\lambda x_i}$

Data: $X_1, \ldots, X_n$ indep.

i.e, same model $\Theta = \frac{1}{\lambda}$, $\lambda = \frac{1}{\theta}$

simply a different parametrization.

<u>Invariance of MLE</u>

Likelihood function for λ is $L(1/\lambda)$, max for $\hat{\lambda} = 1/\bar{X}$. In general, the MLE of $h(\theta)$ is $h(\hat{\theta})$.

BUT $\hat{\lambda}$ is not unbiased, because $E\left(\frac{1}{\bar{x}}\right) \neq \frac{1}{E(\bar{x})}$.

Issues:

(1) Finding MLE

→ $\frac{d}{d\theta} L(\theta) = 0$ solve

→ $\frac{d}{d\theta} \log L(\theta) = 0$ solve

→ numerical methods

→ algebraic ingenuity

(Next time:)

(2) Distribution of MLE

→ find exactly

→ Central Limit Theorem

→ Fisher's App

(3) Properties of MLE

→ unbiased? Not usually

→ Approximate var MSE (Fisher)

→ consider exact distribution