Lec 5 MATH 341/691

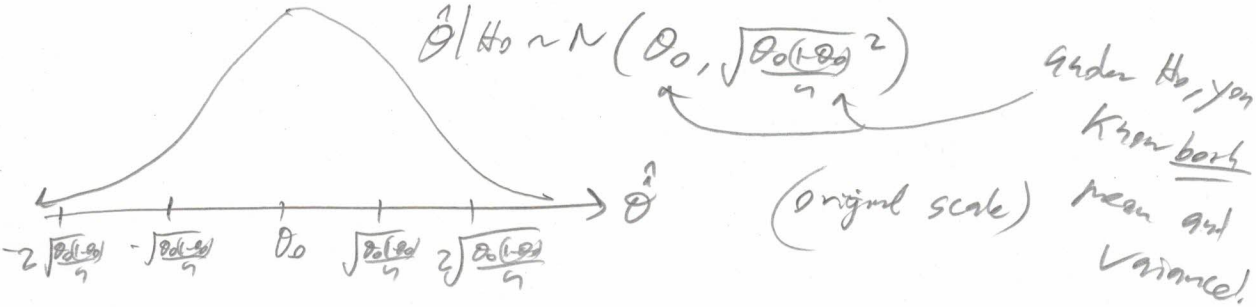
Before we do max. lik. est., let's employ the CLT to see why its important for statistics. Def: "asymptotically normal estimator means:

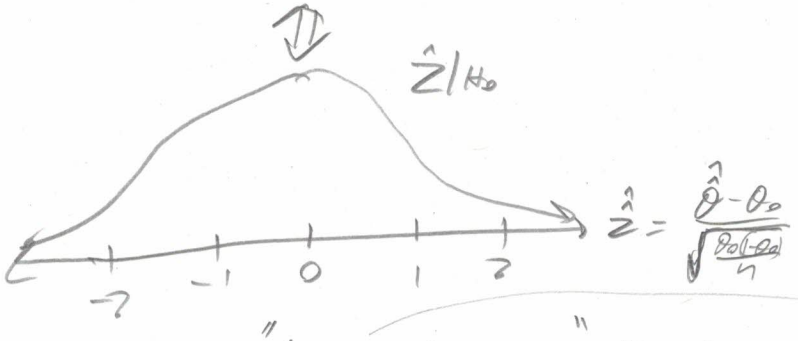
$$\frac{\hat{\theta} - E[\hat{\theta}]}{SE[\hat{\theta}]} \xrightarrow{d} N(0,1)$$

Consider DGP: $X_1, \dots, X_n \overset{iid}{\sim} Ber(\theta)$ and $\hat{\theta} = \bar{X}$. By the CLT, $\frac{\bar{X} - \theta}{\sqrt{\frac{\theta(1-\theta)}{n}}} \xrightarrow{d} N(0,1)$

and thus $\bar{X}$ is asymptotically normal. Once an estimator is asymptotically distr, you know with "large n" $\frac{\bar{X} - \theta}{\sqrt{\frac{\theta(1-\theta)}{n}}} \dot{\sim} N(0,1)$. You can use this fact to construct an "approximate test". $\Rightarrow \bar{X} \dot{\sim} N\left(\theta, \sqrt{\frac{\theta(1-\theta)}{n}}^2\right)$

Previously $\hat{\theta}|H_0 \sim Binom(n, \theta_0)$ exactly. Now → i.e. the level is approximately α





→ $\hat{\theta}$ is a proportion of 1's out of n.

This test is called the "1-prop z-test". It is approximate.

In my opinion, there is no reason to use it! The exact test is better! It performs poorly if θ_0 is close to 0 or 1. Otherwise it performs okay. E.g.

$H_0: \theta = .529$ $\quad RET_\alpha = (-2, 2)$
$H_A: \theta \neq .529$ $\Rightarrow SE[\hat{\theta}] = \sqrt{\frac{.529(1-.529)}{21}} = .499 \Rightarrow \hat{z} = \frac{.476 - .529}{.499} = -.096 \in RET_\alpha \Rightarrow$ Fail to Reject

$\alpha = 5\%$ $\quad n = 21$
$\hat{\theta} = \frac{10}{21} = .476$

Type II error? Interpret?

Procedures for finding estimators

DGP: $X_1, \ldots, X_n \overset{iid}{\sim} p(x; \theta_1, \ldots, \theta_k)$ for discrete θ ... or $f(x; \theta_1, \ldots, \theta_k)$ for cont.

DGP $\xrightarrow{\text{procedure}}$ $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k$, a means of estimation of all parameters of interest

e.g. $\hat{\theta}_j^{MM} = \beta_j(\hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_k)$

this is justified since $\hat{\mu}_\ell \approx E[X^\ell]$

and all parameters are functions of the moments

Although the idea goes back to 1800's . . .

Popularized by RA Fisher between 1912-1922

Another method for finding estimators is the "maximum likelihood method". Here's how this goes for iid data:

$$X_1, \dots, X_n \overset{iid}{\sim} DGP(\theta_1, \dots, \theta_K) = p(x; \theta_1, \dots, \theta_K) \text{ if discrete (PMF)}$$
$$= f(x; \theta_1, \dots, \theta_K) \text{ if continuous (PDF)}$$

I'll use this notation only arbitrarily

Due to independence and identical distributedness,

$$\underbrace{f(x_1, \dots, x_n; \theta_1, \dots, \theta_K)}_{\text{joint density function}} = \prod_{i=1}^{n} f(x_i; \theta_1, \dots, \theta_K)$$

The $\theta_1, \dots, \theta_K$ are the parameters you would need to calculate the density for the data at any point in the support. But we don't know $\theta_1, \dots, \theta_K$ and we are trying to estimate it. So, we do the following conceptual inversion:

$\mathcal{L}$ — stats In Prob 341/641; prob In Prob 340/640

$$\prod_{i=1}^{n} \mathcal{L}(\theta_1 \dots \theta_K; x_i) = \mathcal{L}(\theta_1, \dots, \theta_K; x_1, \dots, x_n) = f(x_1, \dots, x_n; \theta_1, \dots, \theta_K) = \prod_{i=1}^{n} f(x_i; \theta_1, \dots, \theta_K)$$

- $\prod$ on the left: if iid DGP
- $\theta_1, \dots, \theta_K$ in $\mathcal{L}$: settings you wish to compute
- $x_1, \dots, x_n$ in $\mathcal{L}$: givens
- $x_1, \dots, x_n$ in f: inputs/settings you wish to compute
- $\theta_1, \dots, \theta_K$ in f: givens
- $\prod$ on the right: if iid DGP

Note: $\mathcal{L} \geq 0$ for continuous (probability for a density!)
$\mathcal{L} \in (0,1)$ for discrete

Now we vary $\theta_1, \dots, \theta_K$ and see which gives us the most likely values. The value is called the "maximum likelihood estimate" (MLE).

value at which the max is achieved

$$\hat{\theta}^{MLE}_{1,\dots,\theta_K} := \underset{\theta_1,\dots,\theta_K}{\operatorname{argmax}}(\mathcal{L}) = \underset{\theta_1,\dots,\theta_K}{\operatorname{argmax}}\left\{\prod_{i=1}^{n} f(x_i;\theta_1,\dots,\theta_K)\right\}$$

if iid

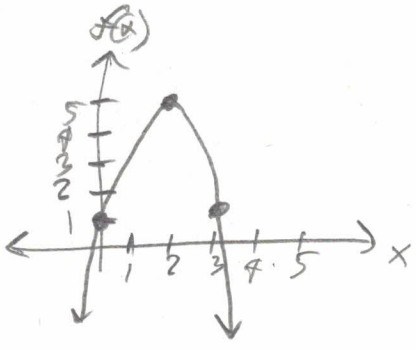
What is argmax?

go back to precalc...

$-x^2+4x-4$

$-(x^2-4x+4)$

let $f(x) = -x^2+4x+1 = -(x-2)^2+5$



What is $\max\{f(x)\} = 5$

$\operatorname{argmax}\{f(x)\} = 2 \Rightarrow \operatorname{argmax}\{f(x)\} := \{x : f(x) = \max\{f(x)\}\}$

How to find argmax? Set $f'(x) = 0$ and solve

$f'(x) = -2x+4 \overset{set}{=} 0 \Rightarrow x_* = 2$. If you're responsible, you should check $f''(x_*)$ to ensure it's < 0 to ensure it's a max (and not a min).

$f''(x) = -2 \Rightarrow f''(2) = -2 < 0$ ✓

Note: the argmax is unaffected by a strictly increasing function

Proof: let $g(x)$ be a function whose derivative is always positive $g'(x) > 0 \ \forall x$

Find $\operatorname{argmax}\{g(f(x))\}$ $\frac{d}{dx}[g(f(x))] = \underbrace{g'(f(x))}_{>0} f'(x) \overset{set}{=} 0$

$\Rightarrow f'(x) \overset{set}{=} 0$ solution is $x_* = \operatorname{argmax}\{f(x)\}$ #

Second derivative test...

$$\frac{d^2}{dx^2}[g(f(x))]\Big|_{x=x_*} = \frac{d}{dx}\underbrace{[g'(f(x))\, f'(x)]}_{<0}\Big|_{x=x} = g'(f(x))\, f''(x) + f'(x)^2 g''(f(x))\Big|_{x=x_*}$$

$$= \underbrace{g'(f(x_*))}_{>0}\underbrace{f''(x)}_{<0} + \underbrace{f'(x_*)^2}_{=0}\overset{?}{g''(f(x))} \leq 0$$

a smooth increasing function in (log). [A]

Let us use the ln function. The log makes everything easier since products become sums, and sums are easy to take derivatives of.

$$l := \ln(\mathcal{L}) = \ln\left(\prod_{i=1}^{n} f(x_i;\theta_1,\ldots,\theta_K)\right) = \sum_{i=1}^{n} \ln\left(f(x_i;\theta_1,\ldots,\theta_K)\right)$$

To get the MLE's, we take $\underset{\theta\in\Theta}{\operatorname{argmax}}\{l\}$ (param space). If local max's exist, we use the following system of eq's.

$$\sum_{i=1}^{n} \frac{\partial}{\partial\theta_1}\left[\ln\left(f(x_i;\theta_1,\ldots,\theta_K)\right)\right] \overset{set}{=} 0$$

$$\sum \frac{\partial}{\partial\theta_2}\left[\ \ldots\ \right] \overset{set}{=} 0$$

$$\vdots$$

$$\sum \frac{\partial}{\partial\theta_K}\left[\ \ldots\ \right] \overset{set}{=} 0.$$

If local max's don't exist, we need to work harder.

---

e.g. $X_1,\ldots,X_n \overset{iid}{\sim} \text{Bern}(\theta)$ ($K=1$ dimension). Here:

$$\sum_{i=1}^{n} \frac{\partial}{\partial\theta}\left[\ln\left(p(x_i;\theta)\right)\right] = \sum_{i=1}^{n} \frac{\partial}{\partial\theta}\left[\ln\left(\theta^{x_i}(1-\theta)^{1-x_i}\right)\right]$$

$$= \sum_{i=1}^{n} \frac{\partial}{\partial\theta}\left[x_i\ln(\theta) + (1-x_i)\ln(1-\theta)\right] = \sum_{i=1}^{n} \frac{x_i}{\theta} - \frac{1-x_i}{1-\theta}$$

$$= \frac{\sum x_i}{\theta} - \frac{n-\sum x_i}{1-\theta} \overset{set}{=} 0 \Rightarrow (1-\theta)\left(\sum x_i\right) = (\theta)\left(n-\sum x_i\right)$$

$$\Rightarrow \sum x_i - \theta\sum x_i = \theta n - \theta\sum x_i \Rightarrow \sum x_i = \theta\left(\sum x_i + n - \sum x_i\right)$$

$$\Rightarrow \hat{\theta}^{MLE} = \frac{\sum x_i}{n} = \bar{X}$$ which was our MM estimate and the naive estimate.

e.g. $X_1 \dots X_n \overset{iid}{\sim} N(\theta_1, \theta_2)$. Find MLE's $\hat{\theta}_1^{MLE}$, $\hat{\theta}_2^{MLE}$

Let's do θ_1 first:

$$\sum_{i=1}^{n} \frac{\partial}{\partial \theta_1}\left[\ln\left(f(x_i;\theta_1,\theta_2)\right)\right] = \sum_{i=1}^{n} \frac{\partial}{\partial \theta_1}\left[\ln\left(\frac{1}{\sqrt{2\pi\theta_2}} e^{-\frac{1}{2\theta_2}(x_i-\theta_1)^2}\right)\right]$$

$$= \sum_{i=1}^{n} \frac{\partial}{\partial \theta_1}\left[-\frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln(\theta_2) - \frac{1}{2\theta_2}(x_i-\theta_1)^2\right]$$

$$-\frac{(x_i^2 - 2x_i\theta_1 + \theta_1^2)}{2\theta_2} = -\frac{x_i^2}{2\theta_2} + \frac{x_i\theta_1}{\theta_2} - \frac{\theta_1^2}{2\theta_2}$$

$$= \sum_{i=1}^{n} \frac{x_i}{\theta_2} - \frac{\theta_1}{\theta_2} = \frac{\sum x_i}{\theta_2} - \frac{n\theta_1}{\theta_2} \overset{set}{=} 0 \Rightarrow \hat{\theta}_1^{MLE} = \bar{X}$$

Let's do θ_2

$$\sum \frac{\partial}{\partial \theta_2}\left[-\frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln(\theta_2) - \frac{1}{2\theta_2}(x_i-\theta_1)^2\right]$$

$$= \sum -\frac{1}{2\theta_2} + \frac{1}{2\theta_2^2}(x_i-\theta_1)^2 = -\frac{n}{2\theta_2} + \frac{\sum(x_i-\theta_1)^2}{2\theta_2^2} \overset{set}{=} 0$$

Systems eq's

$$\Rightarrow \sum(x_i-\theta_1)^2 = n\theta_2 \Rightarrow \theta_2 = \frac{1}{n}\sum(x_i-\theta_1)^2 \Rightarrow \hat{\theta}_2^{MLE} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{X})^2$$

Since w is the same for $\hat{\theta}$ and $\hat{\theta}$, define the MLE "maximum likelihood estimator" as the rv that realizes $\hat{\theta}$, $\hat{\theta}^{MLE}$.

$\hat{\theta}^{MLE}$ is a value, a pt. estimate and $\hat{\theta}^{MLE}$ is its sampling distr.

let's do a difficult example. I'm not expecting you to be able to do this yourself. Only follow along. The results of this will launch us into a whole new topic. Let $X_1, \dots, X_n \overset{iid}{\sim} U(0, \theta)$.

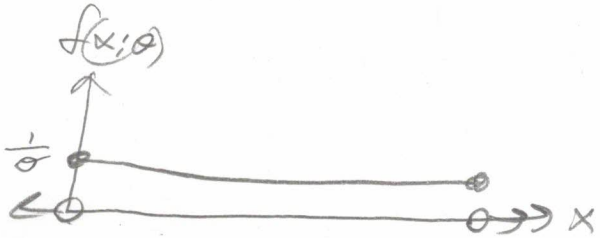
We already derived the silly $\hat{\theta}^{MM} = 2\bar{X}$. Can the MLE do "better"?

$$\sum \frac{\partial}{\partial \theta}\left[\ln(f(x;\theta))\right] = \sum \frac{\partial}{\partial \theta}\left[\ln\left(\frac{1}{\theta}\right)\right] = -\sum \frac{1}{\theta} = -\frac{n}{\theta} \overset{\text{set}}{=} 0$$

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

NO SOL!!!!

Something is wrong! Yes, this MLE is not a local max. It's a global max.

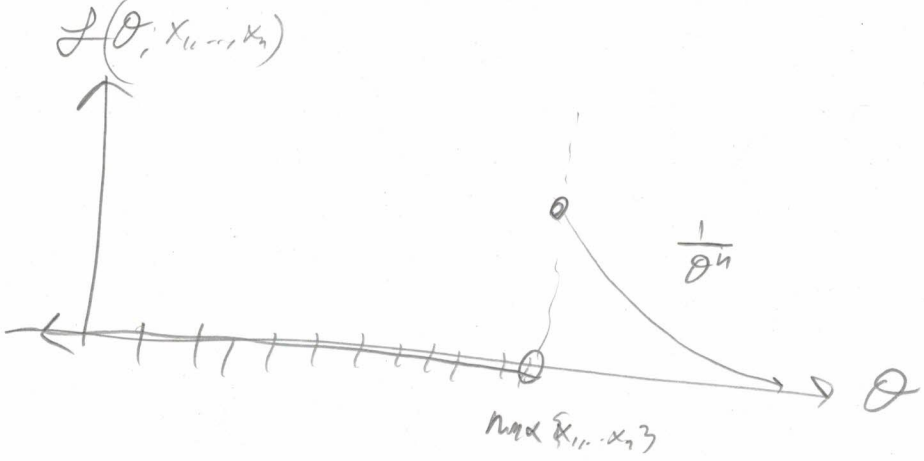


$$f(x;\theta) = \begin{cases} \frac{1}{\theta} & \text{if } 0 \le x \le \theta \\ 0 & \text{o/w} \end{cases}$$

$$f(x;\theta) = \frac{1}{\theta} \mathbb{1}_{x \in [0,\theta]}$$

$$\Rightarrow \prod_{i=1}^{n} f(x;\theta) = \frac{1}{\theta^n} \mathbb{1}_{x_1 \in [0,\theta]} \mathbb{1}_{x_2 \in [0,\theta]} \cdot \dots \cdot \mathbb{1}_{x_n \in [0,\theta]}$$

$$= \mathcal{L}(\theta; x_1, \dots x_n) = \begin{cases} \frac{1}{\theta^n} \underbrace{\mathbb{1}_{\theta > \max(x_1, \dots x_n)}}_{\mathbb{1}_{\theta > x_1} \mathbb{1}_{\theta > x_2} \cdots \mathbb{1}_{\theta > x_n}} & x_{(n)} \\ 0 & \text{o/w} \end{cases}$$



Where is the maximum likelihood estimate? $\hat{\theta}^{MLE} = \max\{x_1, \ldots x_n\}$

$$Var[\hat{\theta}^{MLE}] = \theta^2 \frac{n}{(n+1)^2(n+2)}$$ (proof in Math 390 class)

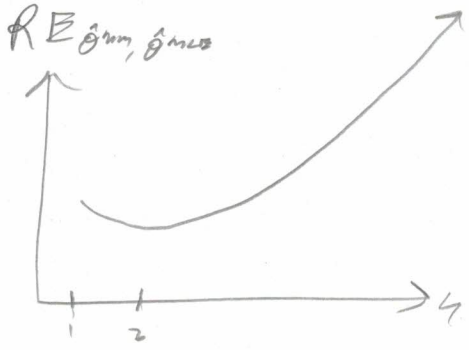
$X \sim U(a,b) \Rightarrow Var(X) = \frac{(b-a)^2}{12}$

$$Var[\hat{\theta}^{mm}] = Var(2\bar{X}) = 4\,Var[\bar{X}] = \frac{4}{n} Var[X] = \frac{4}{n}\frac{\theta^2}{12} = \frac{\theta^2}{3n}$$

$\forall n$, $Var[\hat{\theta}^{MLE}] < Var[\hat{\theta}^{mm}]$

Relative Efficiency (RE)

$$RE = \frac{Var[\hat{\theta}^{mm}]}{Var[\hat{\theta}^{MLE}]} = \frac{\frac{\theta^2}{3n}}{\theta^2 \frac{n}{(n+1)^2(n+2)}} = \frac{(n+1)^2(n+2)}{3n^2}$$



$\hat{\theta}^{MLE}$ gets better and better!!

Should we use

$\frac{MSE[\hat{\theta}^{mm}]}{MSE[\hat{\theta}^{MLE}]}$? Yes But $\hat{\theta}^{mm}$ is unbiased.

we can prove $\hat{\theta}^{MLE}$ has very small bias

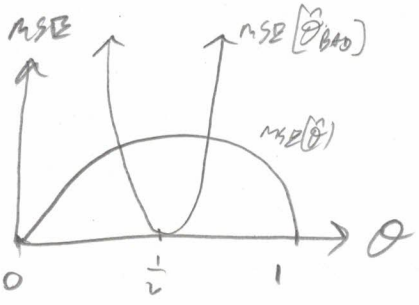
which goes to 0 as $n \to \infty$, so the picture is the same.

There are lots of estimators! Questions:

(1) Is there a theoretical minimum MSE when estimating θ for a given DGP?

(2) If so, is there a procedure to find this estimator with theoretically minimal MSE?

The answer to # is NO. Why? Recall $\hat{\theta}_{BAD} = \frac{1}{2}$ vs $\hat{\theta} = \bar{X}$ for DGP: $X_1, \ldots, X_n \overset{iid}{\sim} \text{Bern}(\theta)$



The class of estimators is a set which is crucial. Intuitive and the choice of θ matters!

So let's limit the set and ask again.

(1) Among all unbiased estimators, is there a theoretical minimum Var = MSE? Yes! It is called the Cramer-Rao Lower Bound (CRLB) proven in 1945-1946. This estimator is called the "uniformly minimum variance unbiased estimator" (UMVUE)

(2) Is there a way to derive the UMVUE from a DGP?

No! But if you find a $\hat{\theta}$ with the CRLB $\Rightarrow \hat{\theta}$ is the UMVUE!