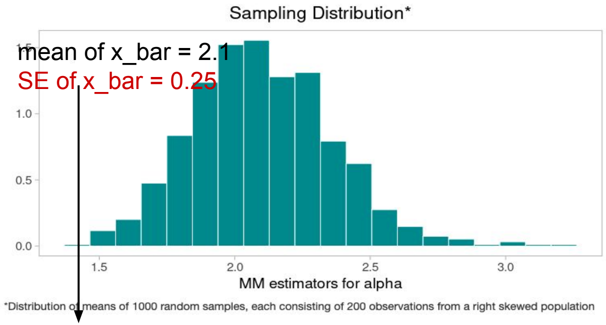
More Examples on MLE

8.5 of Rice - Maximum Likelihood Estimators

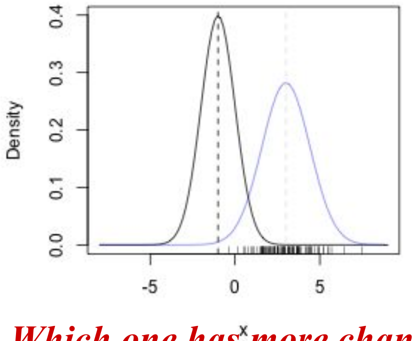
06/29/2021

Berkeley
UNIVERSITY OF CALIFORNIA

In the previous lecture,



Standard error of the estimate



Which one has more chance to generate such a sample?

- The Bootstrap simulation of the **sample distribution** of the MM estimator (Page 10 & 11):
  - Mimicking the population distribution $f(x\,|\,\theta)$ by $f\left(x\,|\,\hat{\theta}_{MM}\right)$.
  - Mimicking the theoretical SE $\sqrt{h(\theta, n)}$ by $\sqrt{h\left(\hat{\theta}_{MM}, n\right)}$.

  This procedure can be generalized to MLE.
  Coding practices in HW2 and Lab 3.

- Consistency of MM estimators:

$$\hat{\theta}_{MM} \xrightarrow{p} \theta, \text{ as } n \to \infty.$$

  Compare with unbiasedness.

- Introduced maximum likelihood estimators (MLE):
  - MM estimators can give unrealistic estimate when n is small.
  - Maximize the likelihood over a meaning set of θ.
  - Derived MLE for $N\left(\mu, \sigma^2\right)$ and $\text{Poisson}(\lambda)$.

Maximum likelihood estimators

Example 1. Let $X_1, \ldots, X_n$ be i.i.d $\mathrm{Gamma}(\alpha, \beta)$. Find the MLE for α and β.

Under the i.i.d assumption,

$$\hat{\theta}_{MLE} = \arg\max_{\theta \in \Theta} \prod_{i=1}^{n} f(X_i \mid \theta).$$

Solution. The probability density function is $f(x \mid \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$.

$$L(\alpha, \beta) = \prod_{i=1}^{n} f(X_i \mid \alpha, \beta) = \frac{\beta^{n\alpha}}{\Gamma^n(\alpha)} \left(\prod_{i=1}^{n} X_i\right)^{\alpha-1} e^{-\beta \sum_{i=1}^{n} X_i}$$

$$l(\alpha, \beta) = \log(\alpha, \beta) = n\alpha \log \beta - n \log \Gamma(\alpha) + (\alpha - 1) \sum_{i=1}^{n} \log X_i - \beta \sum_{i=1}^{n} X_i.$$

$$\begin{cases} \dfrac{\partial l}{\partial \alpha} = n \log \beta - n \dfrac{\Gamma'(\alpha)}{\Gamma(\alpha)} + \sum_{i=1}^{n} \log X_i = 0 \\ \dfrac{\partial l}{\partial \beta} = n \dfrac{\alpha}{\beta} - \sum_{i=1}^{n} X_i = 0 \end{cases}$$

$$\beta = \frac{\alpha}{\bar{X}_n}$$

Need to solve

$$\Rightarrow n \log \alpha - n \log \bar{X}_n - n \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} + \sum_{i=1}^{n} \log X_i = 0$$

$$\hat{\alpha}_{MM} = 2.104$$
$$\hat{\beta}_{MM} = 2.298$$

Maximum likelihood estimators - think more about where to search

$(0,1)$

Binomial (k, p)

$\hat{k}_{MLE}$ $\hat{p}_{MLE}$

Definition. For the i.i.d samples $X_1, \ldots, X_n$, we vary the value of θ *in a meaningful set* to evaluate its likelihood

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

and $L(\theta)$ is called the likelihood function. The maximum likelihood estimator of θ is the particular value that maximizes the likelihood.

$\{x : f(x|\theta) > 0\}$ the support of density function

$X_{(1)} = \min_i \{X_1, \cdots X_n\}$

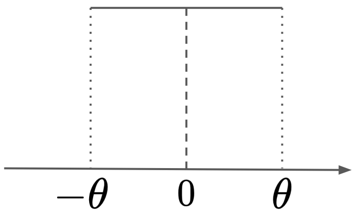
$X_{(n)} = \max_i \{X_1, \cdots X_n\}$

Maximum likelihood estimators

Example 2. Let $X_1, \ldots, X_n$ be i.i.d $U(-\theta, \theta)$. Find the MM estimator and MLE for θ.

Solution: The density for $U(-\theta,\theta)$ is

$$f(x|\theta) = \frac{1}{2\theta}, \quad x \in (-\theta, \theta)$$



$$\mu = EX = \int_{-\theta}^{\theta} \frac{x}{2\theta} dx = 0.$$

$$\mu_2 = EX^2 = \int_{-\theta}^{\theta} \frac{x^2}{2\theta} dx = \frac{\theta^3}{3}.$$

$$\Rightarrow \frac{1}{n}\sum_{i=1}^{n} X_i^2 = \frac{\hat{\theta}_{MM}^3}{3}$$

$$\Rightarrow \hat{\theta}_{MM} = \left(\frac{3}{n}\sum_{i=1}^{n} X_i^2\right)^{1/3}.$$

For MLE, let's look into the set where θ is likely to generate a sample of $X_1, \cdots X_n$.

$$-\theta \le X_i \le \theta, \quad i = 1, \cdots, n$$

$$\Leftrightarrow -\theta \le X_{(1)} < X_{(n)} \le \theta$$

$$\Leftrightarrow \theta \ge -X_{(1)}, \quad \theta \ge X_{(n)}$$

$$\Leftrightarrow \theta \ge \max\{-X_{(1)}, X_{(n)}\}.$$

$$L(\theta) = \prod_{i=1}^{n} f(x_i|\theta) = \left(\frac{1}{2\theta}\right)^n.$$

$$\hat{\theta}_{MLE} = \max\{-X_{(1)}, X_{(n)}\}.$$

Maximum likelihood estimators

$\hat{\theta}_{MLE} = \max\{-X_{(1)}, X_{(n)}\}$

t

Example 2 cont'd. Let $X_1, \ldots, X_n$ be i.i.d $U(-\theta, \theta)$. Find the MLE for θ .

Can you work out the theoretical sampling distribution of $\hat{\theta}_{MLE}$? *(3.7 of Rice)*

$V = X_{(1)}$ and $U = X_{(n)}$ have the joint density

$$f(u, v) = n(n-1) f(v) f(u)[F(u) - F(v)]^{n-2}, \qquad u \geq v$$

$$= n(n-1) \frac{1}{2\theta} \frac{1}{2\theta} \left[\frac{u-v}{2\theta} \right]$$

$$= \frac{n(n-1)}{8\theta^3} (u-v).$$

$$P(\hat{\theta}_{MLE} \leq t) = P(\max\{-X_{(1)}, X_{(n)}\} \leq t) = P(-X_{(1)} \leq t, \ X_{(n)} \leq t)$$

$$= P(X_{(1)} \geq -t, \ X_{(n)} \leq t) = \int_{-t}^{\theta} \int_{-\theta}^{t} f(u,v) \, du \, dv.$$

Maximum likelihood estimators

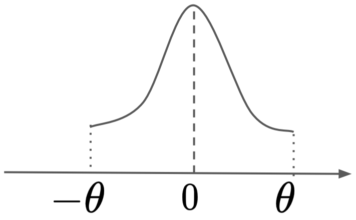
$$f(u,v) = n(n-1) f(u) f(v) \left[F(u) - F(v) \right]^{n-}$$

$$= n(n-1) \frac{3u^2}{2\theta^3} \frac{3v^2}{2\theta^3} \int_v^u \frac{3t^2}{2\theta^3}\, dt$$

$$= \ldots$$

Example 3. Let $X_1, \ldots, X_n$ be i.i.d from a population with density

$$f(x \mid \theta) = \begin{cases} \frac{3x^2}{2\theta^3}, & \text{if } -\theta \le x \le \theta, \\ 0, & \text{otherwise.} \end{cases}$$



Solution: $-\theta \le X_1, \cdots X_{(n)} \le \theta$

$$\Leftrightarrow \underbrace{\theta \ge \max\{-X_{(1)}, X_{(n)}\}}.$$

$$L(\theta) = \prod_{i=1}^n f(X_i \mid \theta) = \prod_{i=1}^n \frac{3X_i^2}{2\theta^3} = \frac{3^n \prod_{i=1}^n X_i^2}{2^n \theta^{3n}}.$$

$$\hat{\theta}_{MLE} = \underbrace{\max\{-X_{(1)}, X_{(n)}\}}. \Leftarrow$$

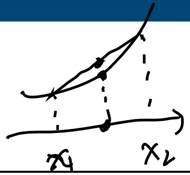
Large Sample Theories for MLE

8.5.2 of Rice

06/29/2021

How good are the MLE?

Jensen's inequality:
φ convex
$$\frac{\varphi(x_1)+\varphi(x_2)}{2} \geq \varphi\left(\frac{x_1+x_2}{2}\right)$$



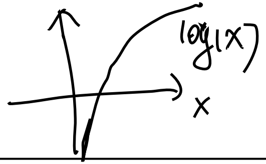
φ convex $\Rightarrow E\,\varphi(X) \geq \varphi(E(X))$

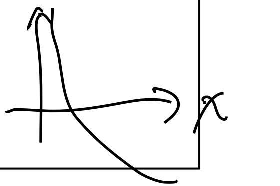
Lemma A. Suppose $X \sim f(x \mid \theta_0)$. For any other θ, the average difference between $\log f(X \mid \theta_0)$ and $\log f(X \mid \theta)$ is called the Kullback-Leibler divergence. Moreover,

$$KL(\theta_0, \theta) = E_{\theta_0}\left(\log \frac{f(X \mid \theta_0)}{f(X \mid \theta)}\right) \geq 0.$$

The KL divergence equals to zero only when $f(x \mid \theta) \equiv f(x \mid \theta_0)$.

φ is linear





A measure of how one probability distribution is different from a second, reference probability distribution, or relative entropy.

Proof. Since $-\log(x)$ is a convex function of x, we know by Jensen's inequality that

$$-E_{\theta_0}\left[\log \frac{f(X\mid\theta)}{f(X\mid\theta_0)}\right] \geq -\log\left[E_{\theta_0} \frac{f(X\mid\theta)}{f(X\mid\theta_0)}\right]$$

$$= -\log\left[\int \cancel{f(x\mid\theta_0)} \frac{f(x\mid\theta)}{\cancel{f(x\mid\theta_0)}}\,dx\right]$$

$$= -\log\left[\int f(x\mid\theta)\,dx\right]$$

$$\underset{f(x\mid\theta)\text{ is a density}}{=} -\log 1 = 0.$$

$N(\mu, \sigma^2)$

$f(x \mid \mu, \sigma^2)$

$e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

$\mu_1 \neq \mu_2$ will give different densities

Identifiable if different θ's produce different densities.

How good are the MLE?

Theorem B. Under the i.i.d and a few other assumptions, $\hat{\theta}_{MLE}$ is a consistent estimator of θ:

$$\hat{\theta}_{MLE} \xrightarrow{p} \theta, \text{ as } n \to \infty.$$

1. *$f(x \mid \theta)$ is identifiable;* $\leftarrow$
2. *$l(\theta)$ is differentiable;*
3. *$\{x : f(x \mid \theta) > 0\}$ does not depend on θ.*

Proof*. Recall that the log-likelihood function is $l(\theta) = \log\left(\prod_{i=1}^{n} f(X_i \mid \theta)\right) = \sum_{i=1}^{n} \log f(X_i \mid \theta)$.

$$\boxed{\frac{1}{n}\sum_{i=1}^{n} \log f(X_i \mid \theta) \xrightarrow{p} E_{\theta_0} \log f(X \mid \theta)}$$

Law of large numbers $\Leftarrow$ if $X_1, \cdots X_n$ are i.i.d then $f(X_1 \mid \theta), \cdots f(X_n \mid \theta)$ are also i.i.d

$$\frac{1}{n} l(\theta) - \frac{1}{n} l(\theta_0) \approx E_{\theta_0} \log f(X \mid \theta) - E_{\theta_0} \log f(X \mid \theta_0)$$

$$= E_{\theta_0} \log \frac{f(X \mid \theta)}{f(X \mid \theta_0)} = -KL(\theta_0, \theta) \leq 0.$$

The equality is only attained when $f(x \mid \theta) \equiv f(x \mid \theta_0) \Rightarrow \theta = \theta_0$

identifiability

$$\frac{\max l(\theta)}{n} - \frac{1}{n} l(\theta_0) \leq 0.$$

$\hat{\theta}_{MLE} \approx \theta_0 \Leftarrow$

How good are the MLE? - Stronger than consistency

If $\hat\theta_n$ is MLE, then $l'(\hat\theta_n) = 0$.

Denote $l_i(\theta) = \log f(X_i \mid \theta)$, then $l(\theta) = \sum_{i=1}^n l_i(\theta)$.

Use Taylor expansion to examine $l_i'(\theta)$.

$$l_i'(\hat\theta_n) - l_i'(\theta_0) = l_i''(\theta_0)(\hat\theta_n - \theta_0) + \underbrace{e_i}_{\text{infinitely small}}$$

sum over i

$$\Leftrightarrow \sum_{i=1}^n l_i'(\hat\theta_n) - \sum_{i=1}^n l_i'(\theta_0) = \sum_{i=1}^n l_i''(\theta_0)\cdot(\hat\theta_n - \theta_0) + \underbrace{\sum_{i=1}^n e_i}_{E}$$

$$\Leftrightarrow l'(\hat\theta_n) - l'(\theta_0) = l''(\theta_0)\cdot(\hat\theta_n - \theta_0) + E$$

$$\Leftrightarrow \hat\theta_n - \theta_0 = \frac{-(l'(\hat\theta_n) - l'(\theta_0) + E)}{-(l''(\theta_0))}$$

$$\Rightarrow \sqrt{n}\left(\hat\theta_n - \theta_0\right) = \frac{\sqrt{n}l'(\theta_0) - \cancel{\sqrt{n}l'(\hat\theta_n)} + \cancel{E_n}}{-l''(\theta_0)} \approx \frac{\frac{1}{n}\sqrt{n}l'(\theta_0)}{-\frac{1}{n}l''(\theta_0)}$$

Numerator

$$\frac{1}{\sqrt{n}}l'(\theta_0) = \sqrt{n} \times \frac{1}{n}\sum_{i=1}^n \frac{\partial}{\partial\theta}\log f(X_i \mid \theta_0)$$

$$\sqrt{n}\left[\frac{1}{n}\sum_{i=1}^n \frac{\partial}{\partial\theta}\log f(X_i \mid \theta_0) - \underbrace{E\frac{\partial}{\partial\theta}\log f(X\mid\theta_0)}_{=0}\right]$$

$$\xrightarrow{CLT} N\left(0, \underbrace{\operatorname{var}\frac{\partial}{\partial\theta}\log f(X\mid\theta_0)}_{=I(\theta_0)}\right)$$

Denominator

$$-\frac{1}{n}l''(\theta_0) = -\frac{1}{n}\sum_{i=1}^n \frac{\partial^2}{\partial\theta^2}\log f(X_i \mid \theta_0)$$

$$\xrightarrow{P} \underbrace{-E_{\theta_0}\frac{\partial^2}{\partial\theta^2}\log f(X_i\mid\theta_0)}_{=I(\theta_0)}$$

How good are the MLE? - *One parameter model*

(*) We need to show that $E_\theta\left[\dfrac{\frac{\partial^2}{\partial\theta^2} f(X|\theta)}{f(X|\theta)}\right] = 0$.

Note $E_\theta\left[\dfrac{\frac{\partial^2}{\partial\theta^2} f(X|\theta)}{f(X|\theta)}\right] = \int \dfrac{\frac{\partial^2}{\partial\theta^2} f(x|\theta)}{\cancel{f(x|\theta)}} \cancel{f(x|\theta)}\, dx$

$= \int \dfrac{\partial^2}{\partial\theta^2} f(x|\theta)\, dx$

$= \dfrac{\partial^2}{\partial\theta^2} \underbrace{\int f(x|\theta)\, dx}_{=1} = 0$

Definition. The Fisher information of θ is defined by

$$I(\theta) = E_\theta\left[\frac{\partial}{\partial\theta}\log f(X\,|\,\theta)\right]^2.$$

Lemma C. Under appropriate conditions on $f(x\,|\,\theta)$, the Fisher information can also be written as

The exchangability of differentiation & integration

$$I(\theta) = -E_\theta\left[\frac{\partial^2}{\partial\theta^2}\log f(X\,|\,\theta)\right].$$

Proof. We begin by examining the ~~1st~~ 2nd derivative of $\log f(x|\theta)$:

$\dfrac{\partial}{\partial\theta}\log f(x|\theta) = \dfrac{\frac{\partial}{\partial\theta} f(x|\theta)}{f(x|\theta)}$,

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

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

Therefore,

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

$\overset{(*)}{=} 0 - I(\theta)$

see above

Cont'd from Page 11:

✪ Using the results from Page 11, we know

$$\sqrt{n}\left\{\frac{1}{n}\ell'(\theta) - E_{\theta_0}\left[\frac{\partial}{\partial\theta}\log f(X|\theta_0)\right]\right\} \to N\left(0, \mathrm{var}_{\theta_0}\left(\frac{\partial}{\partial\theta}\log f(X|\theta_0)\right)\right).$$

Note: $E_{\theta_0}\left[\frac{\partial}{\partial\theta}\log f(X|\theta_0)\right] = \int \frac{\frac{\partial}{\partial\theta} f(x|\theta_0)}{\cancel{f(x|\theta_0)}} \cancel{f(x|\theta_0)}\, dx$

$$= \boxed{\int \frac{\partial}{\partial\theta} f(x|\theta_0)\, dx = \frac{\partial}{\partial\theta} \underbrace{\int f(x|\theta_0)\, dx}_{=1}} = 0.$$

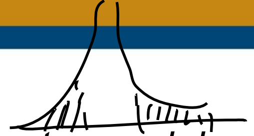
Another exchangeability ✓ assumption

$$\mathrm{var}_{\theta_0}\left[\frac{\partial}{\partial\theta}\log f(X|\theta_0)\right] \underset{\mathrm{var}X = EX^2 - (EX)^2}{=} \underbrace{E_{\theta_0}\left[\frac{\partial}{\partial\theta}\log f(X|\theta_0)\right]^2}_{= I(\theta_0) \text{ by definition}} - \left\{\underbrace{E_{\theta_0}\left[\frac{\partial}{\partial\theta}\log f(X|\theta_0)\right]}_{=0}\right\}^2$$

$$= I(\theta_0).$$

Therefore, $\boxed{\frac{1}{\sqrt{n}}\ell'(\theta) \to N(0, I(\theta_0))}$.

✪ For the denominator, $\boxed{-\frac{1}{n}\sum_{i=1}^{n}\ell''(\theta_0) \xrightarrow{P} E_{\theta_0}\left[\frac{\partial^2}{\partial\theta^2}\log f(X|\theta_0)\right] \overset{\text{Lemma C}}{=} I(\theta_0)}$



$P(|\bar{X}_n - \theta| > \varepsilon) \to 0$

Asymptotic Normality of MLE - *One parameter model*

To sum up,

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \approx \frac{\frac{1}{\sqrt{n}} l'(\theta_0)}{-\frac{1}{n} l''(\theta_0)} \approx$$

in which

$$\frac{1}{\sqrt{n}} l'(\theta_0) \to N(0, I(\theta_0)),$$

and

$$-\frac{1}{n} l''(\theta_0) = -\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2}{\partial \theta^2} \log f(X_i \mid \theta_0) \xrightarrow{p} I(\theta_0).$$

$$\frac{N(0, I(\theta_0))}{I(\theta_0)} = N\left(0, \frac{1}{I(\theta_0)}\right)$$

1. $f(x \mid \theta)$ is identifiable;
2. $l(\theta)$ is differentiable; Exchangeability
3. $\{x : f(x \mid \theta) > 0\}$ does not depend on θ; Assumption
4. $\int \frac{\partial}{\partial \theta} f(x \mid \theta) dx = \frac{\partial}{\partial \theta} \int f(x \mid \theta) dx$,
   $\int \frac{\partial^2}{\partial \theta^2} f(x \mid \theta) dx = \frac{\partial^2}{\partial \theta^2} \int f(x \mid \theta) dx$;

This can be easily satisfied by sufficiently smooth densities

Rate of convergence : $n^{-1/2}$

$SE(\hat{\theta}_n) = \sqrt{\frac{1}{n I(\theta_0)}} = (n I(\theta_0))^{-1/2}$

Theorem D. Under the i.i.d and a few other assumptions, the MLE $\hat{\theta}_n$ has asymptotic normality:

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \xrightarrow{d} N\left(0, \frac{1}{I(\theta_0)}\right).$$

$$\hat{\theta}_n - \theta_0 \to N\left(0, \frac{1}{n I(\theta_0)}\right)$$