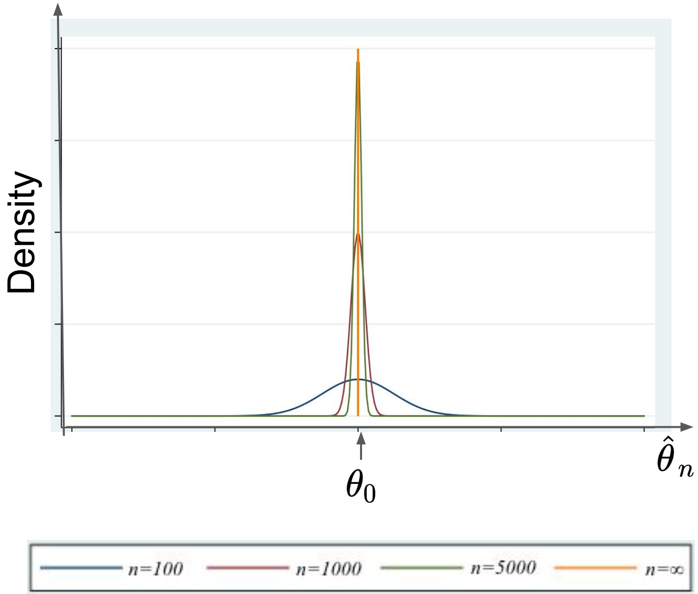
Large Sample Theories for MLE (*cont'd*)

8.5.2 of Rice

06/30/2021

Berkeley
UNIVERSITY OF CALIFORNIA

In the previous lecture,



- More examples on MLE:
  - For $\mathrm{Gamma}(\alpha, \beta)$, we need to solve numerically.
  - For $U(-\theta, \theta)$, be careful about where to search for $\hat{\theta}_{MLE}$.
- The consistency of MLE:

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

  - KL divergence: $KL(\theta_0, \theta) = E_{\theta_0}\left[\log \frac{f(X \mid \theta_0)}{f(X \mid \theta)}\right] \geq 0.$
- The asymptotic normality of MLE (*one-parameter model*):

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

  - Fisher information: $I(\theta) = E_\theta\left[\frac{\partial}{\partial \theta} \log f(X \mid \theta)\right]^2.$
  - Exchangeability assumptions.

Asymptotic Normality of MLE - *One parameter model*

Poisson(λ)

To sum up,

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

in which

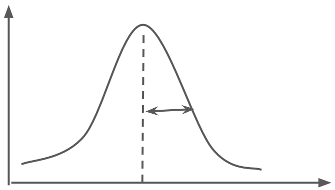
$$\frac{1}{\sqrt{n}} l'(\theta_0) \rightarrow N(0,\, I(\theta_0)),$$

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

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

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

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



θ_0

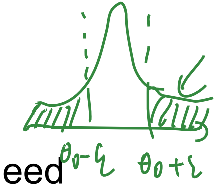
$$\mathrm{SE}(\hat{\theta}_n) \approx \frac{1}{\sqrt{n I(\theta_0)}}$$

Fisher information $\nearrow$, $\mathrm{SE}(\hat{\theta}_n) \searrow$, accuracy of $\hat{\theta}_n \nearrow$

Asymptotic Normality vs. Consistency

$P(|\hat{\theta}_n - \theta_0| > \varepsilon) \to 0$

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

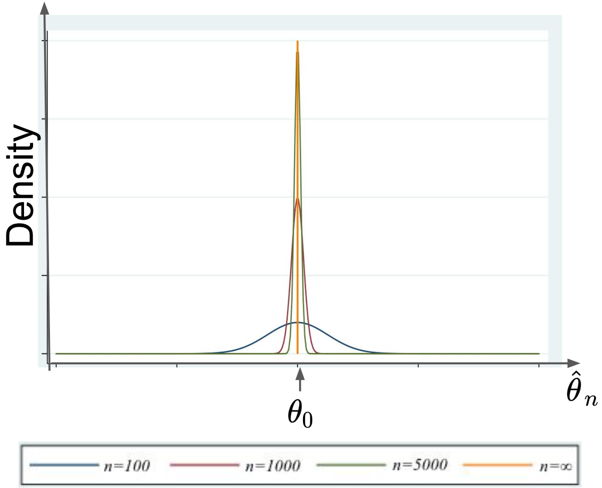


- Symmetric? X
- Rate of convergence? X
- What **sample size** do you need to ensure a certain degree of precision?

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

$$\boxed{\frac{1}{\sqrt{nI(\theta_0)}} \leq 0.1 \Rightarrow n \geq \frac{100}{I(\theta_0)}}$$

Useful for experimental design



Asymptotic Normality of MLE - *Multiple parameter model*

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

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

$$[\mathcal{I}(\theta)]_{i,j} = \mathbf{E}\left[\left(\frac{\partial}{\partial \theta_i} \log f(X|\theta)\right)\left(\frac{\partial}{\partial \theta_j} \log f(X|\theta)\right)\middle|\theta\right]$$

Similarly to Lemma C, we have

$$[\mathcal{I}(\theta)]_{i,j} = -\mathbf{E}\left[\frac{\partial^2}{\partial \theta_i\, \partial \theta_j} \log f(X|\theta)\middle|\theta\right].$$

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

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

Multivariate normal

Asymptotic Normality of MLE - *Multiple parameter model*

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2, & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2, & \sigma_2^2 \end{pmatrix} \right)$$

$X \sim N(\mu_1, \sigma_1^2)$, $Y \sim (\mu_2, \sigma_2^2)$

Example 4. Let $X_1, \ldots, X_n$ be i.i.d $N(\mu, \sigma^2)$. Find $\mathcal{I}(\mu, \sigma)$.

Solution. We know that $\log f(x \mid \mu, \sigma) = \log \frac{1}{\sqrt{2\pi}} - \log \sigma - \frac{1}{2\sigma^2}(x-\mu)^2$.

$(\hat{\mu}_{MLE}, \hat{\sigma}_{MLE})$ is at least asymptotically independent of each other

$$\frac{\partial \log f}{\partial \mu} = \frac{1}{\sigma^2}(x-\mu), \quad \frac{\partial \log f}{\partial \sigma} = -\frac{1}{\sigma} + \frac{1}{\sigma^3}(x-\mu)^2$$

$$\frac{\partial^2 \log f}{\partial \mu^2} = -\frac{1}{\sigma^2}, \quad \frac{\partial^2 \log f}{\partial \sigma^2} = \frac{1}{\sigma^2} - \frac{3}{\sigma^4}(x-\mu)^2, \quad \frac{\partial^2 \log f}{\partial \mu \partial \sigma} = \frac{\partial^2 \log f}{\partial \sigma \partial \mu} = -\frac{2}{\sigma^3}(x-\mu)$$

$$\Rightarrow I(\mu, \sigma) = \begin{pmatrix} \frac{1}{\sigma^2} & \frac{2}{\sigma^3} \underbrace{E(X-\mu)}_{=0} \\ \frac{2}{\sigma^3} E(X-\mu) & \frac{3}{\sigma^4} \underbrace{E(X-\mu)^2}_{=\sigma^2} - \frac{1}{\sigma^2} \end{pmatrix} = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{3}{\sigma^2} - \frac{1}{\sigma^2} \end{pmatrix} = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{2}{\sigma^2} \end{pmatrix}$$

$$I^{-1}(\mu, \sigma) = \begin{pmatrix} \sigma^2 & 0 \\ 0 & \frac{\sigma^2}{2} \end{pmatrix}$$

Therefore, $\sqrt{n} \left\{ \begin{pmatrix} \hat{\mu}_{MLE} \\ \hat{\sigma}_{MLE} \end{pmatrix} - \begin{pmatrix} \mu \\ \sigma \end{pmatrix} \right\} \xrightarrow{d} N \left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & 0 \\ 0 & \sigma^2/2 \end{pmatrix} \right\}$

MLE vs. MM estimators

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \begin{pmatrix} d & -b \\ -c & a \end{pmatrix} \frac{1}{ad-bc}.$$

Example 1 *cont'd*. Let $X_1, \ldots, X_n$ be i.i.d $\mathrm{Gamma}(\alpha, \beta)$. Calculate $\mathcal{I}(\alpha, \beta)$.

Solution. We know $\log f(x \mid \alpha, \beta) = \alpha \log \beta - \log \Gamma(\alpha) + (\alpha - 1) \log x - \beta x.$

$$\frac{\partial \log f}{\partial \alpha} = \log \beta - [\log \Gamma(\alpha)]' + \log x, \quad \frac{\partial \log f}{\partial \beta} = \frac{\alpha}{\beta} - x.$$

$$\frac{\partial^2 \log f}{\partial \alpha^2} = -[\log \Gamma(\alpha)]'', \quad \frac{\partial^2 \log f}{\partial \beta^2} = -\frac{\alpha}{\beta^2}, \quad \frac{\partial^2 \log f}{\partial \beta \partial \alpha} = \frac{\partial^2 \log f}{\partial \alpha \partial \beta} = \frac{1}{\beta}.$$

$$I(\alpha, \beta) = \begin{pmatrix} [\log \Gamma(\alpha)]'' & -\frac{1}{\beta} \\ -\frac{1}{\beta} & \frac{\alpha}{\beta^2} \end{pmatrix}, \quad I^{-1}(\alpha, \beta) = \begin{pmatrix} \frac{\alpha}{\beta^2} & \frac{1}{\beta} \\ \frac{1}{\beta} & [\log \Gamma(\alpha)]'' \end{pmatrix} \frac{1}{\frac{\alpha [\log \Gamma(\alpha)]''}{\beta^2} - \frac{1}{\beta^2}}.$$

$$\hat{\alpha}_{MLE} \sim N\left(\alpha, \frac{\alpha/\beta^2}{\frac{\alpha[\log\Gamma(\alpha)]''}{\beta^2} - \frac{1}{\beta^2}}\right) = N\left(\alpha, \frac{1}{[\log\Gamma(\alpha)]'' - 1/\alpha}\right)$$

Therefore, $\sqrt{n}\left[\begin{pmatrix} \hat{\alpha}_{MLE} \\ \hat{\beta}_{MLE} \end{pmatrix} - \begin{pmatrix} \alpha \\ \beta \end{pmatrix}\right] \xrightarrow{d} N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, I^{-1}(\alpha, \beta)\right).$

$[\log \Gamma(\alpha)]'$ is digamma
$[\log \Gamma(\alpha)]''$ is trigamma

MLE vs. MM estimators

Example 1 *cont'd*. Let $X_1, \ldots, X_n$ be i.i.d $\text{Gamma}(\alpha, \beta)$. Estimate the SE for $\hat{\alpha}_{MM}$ and $\hat{\alpha}_{MLE}$.

$$\hat{\alpha}_{MM} = \frac{\bar{X}_n^2}{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2}$$

- *No theoretical sampling distribution;*
- *No asymptotic normality;*
- *We have to obtain SE of $\hat{\alpha}_{MM}$ via bootstrapping (Page 10, Lecture 3).*

$SE(\hat{\alpha}_{MM}) = 0.260$

$\hat{\alpha}_{MLE}$ is the solution to

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

$$\text{SE}(\hat{\alpha}_{MLE}) \approx \frac{1}{\sqrt{n\{\text{trigamma}(\alpha)\text{-}1/\alpha\}}}$$

α_{MLE} α_{MLE}

$\hat{\alpha}_{MLE} = 2.172$

$SE(\hat{\alpha}_{MLE}) = 0.203$ $\leftarrow$

MLE vs. MM estimators

Example 1 *cont'd*. Let $X_1, \ldots, X_n$ be i.i.d $\text{Gamma}(\alpha, \beta)$. Estimate the SE for $\hat{\alpha}_{MM}$ and $\hat{\alpha}_{MLE}$.

$$\hat{\alpha}_{MM} = \frac{\bar{X}_n^2}{\frac{1}{n}\sum_{i=1}^n \left(X_i - \bar{X}_n\right)^2}$$

- *No theoretical sampling distribution;*
- *No asymptotic normality;*
- *We have to obtain SE of* $\hat{\alpha}_{MM}$ *via bootstrapping (Page 10, Lecture 3).*

$\hat{\alpha}_{MLE}$ is the solution to

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

$$\text{SE}(\hat{\alpha}_{MLE}) \approx \frac{1}{\sqrt{n\{\text{trigamma}(\alpha)\text{-}1/\alpha\}}}$$

$$SE(\hat{\beta}_{MLE}) \approx \sqrt{\frac{\beta^2 \, \text{trigamma}(\alpha)}{\alpha \cdot \text{trigamma}(\alpha) - 1}}$$

$$\hat{\alpha}_{MM} = 2.104$$
$$\text{SE}(\hat{\alpha}_{MM}) = 0.270$$

$$\hat{\beta}_{MM} = 2.298$$
$$\text{SE}(\hat{\beta}_{MM}) = 0.310$$

$$\hat{\alpha}_{MLE} = 2.172$$
$$\text{SE}(\hat{\alpha}_{MLE}) = 0.203$$

$$\hat{\beta}_{MLE} = 2.278$$
$$\text{SE}\left(\hat{\beta}_{MLE}\right) = 0.239$$

More efficient!

Bootstrap confidence intervals using MLE

8.5.3 of Rice

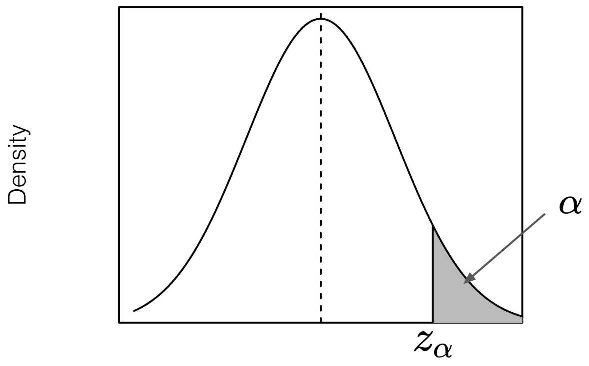
06/30/2021

Confidence interval for μ

Let z_α be the number such that $\Phi(z_\alpha) = 1 - \alpha$.

(Lecture 2, Thm D). Under the i.i.d assumption,

$$P\left(\bar{X}_n - \frac{z_{\alpha/2}\sigma}{\sqrt{n}} \leq \mu \leq \bar{X}_n + \frac{z_{\alpha/2}\sigma}{\sqrt{n}}\right) \approx 1 - \alpha.$$



$$1 - \alpha = 0.99,\ z_{\alpha/2} = 2.576$$

$$1 - \alpha = 0.95,\ z_{\alpha/2} = 1.96$$

$$1 - \alpha = 0.90,\ z_{\alpha/2} = 1.645$$

Confidence interval for θ

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

$\sqrt{n}(\bar{X}_n - \mu) \to N(0, \sigma^2)$

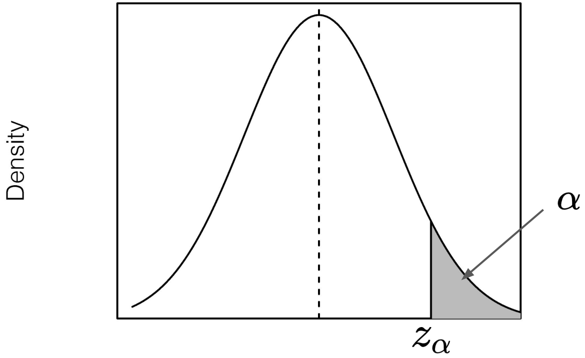
Let z_α be the number such that $\Phi(z_\alpha) = 1 - \alpha$.

Corollary D. Under the i.i.d and a few other assumptions,

$$P\left(\hat{\theta}_n - \frac{z_{\alpha/2}\sqrt{1/I(\theta_0)}}{\sqrt{n}} \le \theta \le \hat{\theta}_n + \frac{z_{\alpha/2}\sqrt{1/I(\theta_0)}}{\sqrt{n}}\right) \approx 1 - \alpha.$$

Or Bootstrap plug-in:

$$P\left(\hat{\theta}_n - \frac{z_{\alpha/2}}{\sqrt{nI\left(\hat{\theta}_n\right)}} \le \theta \le \hat{\theta}_n + \frac{z_{\alpha/2}}{\sqrt{nI\left(\hat{\theta}_n\right)}}\right) \approx 1 - \alpha.$$



$1 - \alpha = 0.99,\ z_{\alpha/2} = 2.576$

$1 - \alpha = 0.95,\ z_{\alpha/2} = 1.96$

$1 - \alpha = 0.90,\ z_{\alpha/2} = 1.645$

Calculate MLE for λ : $L(\lambda) = \lambda^n e^{-\lambda \sum_{i=1}^n x_i}$

$\ell(\lambda) = n \log \lambda - \lambda \sum_{i=1}^n x_i$

$\frac{\partial \ell}{\partial \lambda} = \frac{n}{\lambda} - \sum_{i=1}^n x_i = 0$

$\Rightarrow \boxed{\hat{\lambda}_{MLE} = \frac{1}{\bar{X}_n}}$

Confidence interval for θ

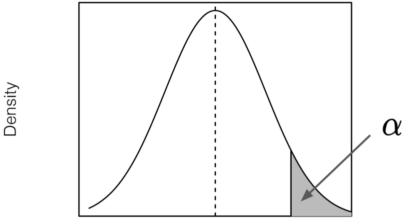
Example 5. Suppose the population follows the exponential distribution $\mathrm{Exp}(\lambda)$. Find the MLE estimator $\hat{\lambda}_{MLE}$ and a 95% bootstrap confidence interval for λ if $\bar{X}_n = 1.031$ and $n = 100$.

Lec 3 Example 6

$\hat{\lambda}_{mm} = \frac{1}{\hat{\mu}} = \frac{1}{\bar{X}_n}$

$SE(\hat{\lambda}_{mm}) \approx \frac{\hat{\lambda}_{mm}}{\sqrt{n}}$ ←

Bootstrap estimate



$1 - \alpha = 0.99,\ z_{\alpha/2} = 2.576$

$1 - \alpha = 0.95,\ z_{\alpha/2} = 1.96$

$1 - \alpha = 0.90,\ z_{\alpha/2} = 1.645$

$f(x|\lambda) = \lambda e^{-\lambda x}$, $\log f(x|\lambda) = \log \lambda - \lambda x$

$\frac{\partial \log f}{\partial \lambda} = \frac{1}{\lambda} - x$, $\frac{\partial^2 \log f}{\partial \lambda^2} = -\frac{1}{\lambda^2}$.

Thus, $I(\lambda) = -E\left(\frac{\partial^2 \log f}{\partial \lambda^2}\right) = \frac{1}{\lambda^2}$.

$\rightarrow \hat{\lambda}_{MLE} = \frac{1}{\bar{X}_n}$

$\sqrt{n}(\hat{\lambda}_{MLE} - \lambda) \xrightarrow{d} N\left(0, \frac{1}{I(\lambda)}\right) = N(0, \lambda^2)$

$var(\hat{\lambda}_{MLE}) \approx \frac{\lambda^2}{n}$, $SE(\hat{\lambda}_{MLE}) \approx \frac{\lambda}{\sqrt{n}} \approx \frac{\hat{\lambda}_{MLE}}{\sqrt{n}}$.

95% CI should be: $\hat{\lambda}_{MLE} \pm \frac{z_{\alpha/2}}{\sqrt{n I(\hat{\lambda}_{MLE})}} = \frac{1}{1.031} \pm \frac{1.96}{\sqrt{100 \times 1/\hat{\lambda}_{MLE}^2}} = 0.9699 \pm 0.1901 = [0.780, 1.160]$.

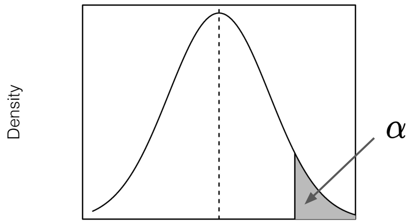
Confidence interval for θ

$\hat{\alpha}_{MLE} \sim N\left(\alpha, \frac{1}{n(\text{trigamma}(\alpha) - 1/\alpha)}\right)$

Example 1 *cont'd*. Let $X_1, \ldots, X_n$ be i.i.d $\text{Gamma}(\alpha, \beta)$. Calculate the 95% bootstrap confidence intervals separately for α and β.

$$\hat{\alpha}_{MLE} \pm \frac{z_{0.05/2}}{\sqrt{n\{\text{trigamma}(\alpha)\text{-}1/\alpha\}}} = 2.172 \pm \frac{1.96}{\sqrt{200\{\text{trigamma}(2.172) - 1/2.172\}}} = [1.175, 2.569].$$

$$\hat{\beta}_{MLE} \pm \frac{z_{0.05/2}\sqrt{\beta^2 * \text{trigamma}(\alpha)}}{\sqrt{n\{\alpha * \text{trigamma}(\alpha)\text{-}1\}}} = 2.278 \pm 0.469 = [1.809, 2.747].$$



$1 - \alpha = 0.99,\ z_{\alpha/2} = 2.576$

$1 - \alpha = 0.95,\ z_{\alpha/2} = 1.96$

$1 - \alpha = 0.90,\ z_{\alpha/2} = 1.645$

Don't confuse the α's!

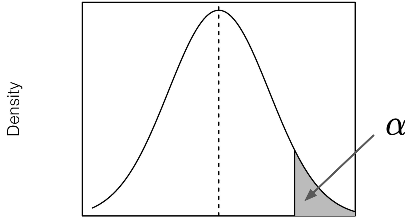
Confidence interval for θ - *Bonferroni correction*

$P\left(L_\alpha < \alpha < U_\alpha,\ L_\beta < \beta < U_\beta\right)$
$\geq 1-\alpha$

Example 1 *cont'd*. Let $X_1, \ldots, X_n$ be i.i.d $\text{Gamma}(\alpha, \beta)$. Calculate the 95% <u>simultaneous</u> confidence intervals for α and β.

$$\hat{\alpha}_{MLE} \pm \frac{z_{0.05/4}}{\sqrt{n\{\text{trigamma}(\alpha)\text{-}1/\alpha\}}}$$

$$\hat{\beta}_{MLE} \pm \frac{z_{0.05/4}\sqrt{\beta^2 * \text{trigamma}(\alpha)}}{\sqrt{n\{\alpha * \text{trigamma}(\alpha)\text{-}1\}}}$$



$1-\alpha = 0.99,\ z_{\alpha/2} = 2.576$
$1-\alpha = 0.95,\ z_{\alpha/2} = 1.96,\ z_{\alpha/4} = 2.241$
$1-\alpha = 0.90,\ z_{\alpha/2} = 1.645$

If one establishes m confidence intervals, and wishes to have an overall confidence level of $1-\alpha$, each individual confidence interval can be adjusted to the level of $1-\frac{\alpha}{m}$.

for each CI, $m=2$, $\to 1-\frac{\alpha}{2}$

$z_{\alpha/2} \to z_{\alpha/4}$.

α: $2.172 \pm 0.4541 = [1.716, 2.626]$ ∈

β: $2.278 \pm 0.536 = [1.742, 2.814]$.

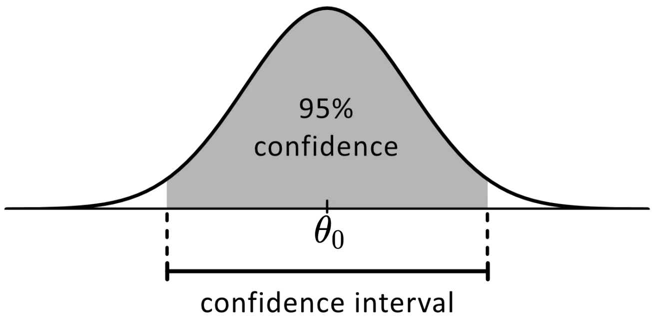
Short summary of CI (*more to come*)

$\frac{1}{n}(X_1+\cdots+X_n) = \text{Gamma}\left(n\alpha, \frac{\beta}{n}\right)$

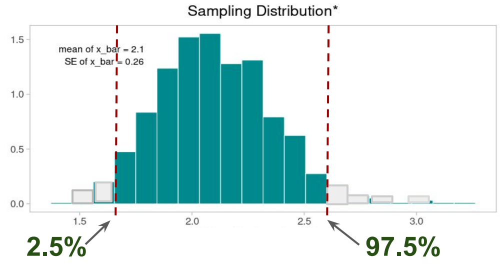
$\frac{1}{n}(X_1+\cdots+X_n) = N\left(\mu, \frac{\sigma^2}{n}\right)$

- $\bar{X}_n$ and $\hat{\sigma}_n^2$ for population mean and variance;
- MM estimators;
- Maximum likelihood estimators.

Sampling distribution of an estimator



1. *Exact sampling distribution;*
   e.g. MLE for $U(-\theta, \theta)$,
   $\bar{X}_n$ *for* $\text{Gamma}(\alpha, \beta)$ *and* $N(\mu, \sigma^2)$
2. *Asymptotic normality;*
   $\bar{X}_n$, $g(\bar{X}_n)$, *MLE*
3. *Parametric or non-parametric bootstrapping (Page 10, Lecture 3 & Lab 3).*



Efficiency and Cramer-Rao inequality

8.7 of Rice

06/30/2021

How to evaluate an estimator?

MM , MLE ← consistency

$E(\hat{\theta}_n) \to \theta$, $n \to \infty$

For any estimator $\hat{\theta}_n = g(X_1, \ldots, X_n)$,

$$\mathrm{MSE}(\hat{\theta}_n) = \underbrace{\mathrm{Var}\left(\hat{\theta}_n\right)}_{\text{variance}} + \underbrace{\left[E\left(\hat{\theta}_n\right) - \theta\right]^2}_{\text{bias}}$$

Definition. Given two estimators $\hat{\theta}_n$ and $\breve{\theta}_n$, the efficiency of $\hat{\theta}_n$ relative to $\breve{\theta}_n$ is defined to be

$$\mathrm{eff}\left(\breve{\theta}_n,\, \hat{\theta}_n\right) = \frac{\mathrm{Var}\left(\breve{\theta}_n\right)}{\mathrm{Var}\left(\hat{\theta}_n\right)} = \left[\frac{\mathrm{SE}\left(\breve{\theta}_n\right)}{\mathrm{SE}\left(\hat{\theta}_n\right)}\right]^2.$$

$$\begin{aligned} \hat{\alpha}_{MM} &= 2.104 \\ \mathrm{SE}(\hat{\alpha}_{MM}) &= 0.270 \end{aligned}$$

$$\begin{aligned} \hat{\alpha}_{MLE} &= 2.172 \\ \mathrm{SE}(\hat{\alpha}_{MLE}) &= 0.203 \end{aligned}$$

$$\Longrightarrow \quad \mathrm{eff}(\hat{\alpha}_{MM}, \hat{\alpha}_{MLE}) = \left(\frac{0.270}{0.203}\right)^2 = 1.769$$

*Is there a lower bound for the variance of **any** estimate?*

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

Cauchy-Schwartz inequality:

$\sum x_i^2 \sum y_i^2 \geq (\sum x_i y_i)^2$ continuous $EX^2 EY^2 \geq (EXY)^2$

Equality holds if and only if $y_i = a x_i + b$, for every i.

$\Rightarrow \sqrt{\text{var} X}\sqrt{\text{var} Y} \geq |\text{cov}(X,Y)|$.

Cramer-Rao inequality

Theorem E. Suppose the population has a density $f(x\,|\,\theta)$. Under the i.i.d and a few other assumptions, let $\hat{\delta}_n = g(X_1, \ldots, X_n)$ be any estimator. Define $\psi(\theta) = E_\theta\left(\hat{\delta}_n\right)$. Then

$$\text{Var}_\theta\left(\hat{\delta}_n\right) \geq \frac{[\psi'(\theta)]^2}{nI(\theta)},$$

where the lower bound is attained if and only if $\frac{\partial}{\partial\theta} l(\theta) = a(\theta)\left[\hat{\delta}_n - \psi(\theta)\right]$.

$E(X) = \lambda$.
$\text{var}(X) = \lambda$.

Example 6. Let $X_1, \ldots, X_n$ be i.i.d $\text{Poisson}(\lambda)$. Consider $\bar{X}_n$

$X \sim \text{Poisson}(\lambda)$, $Y \sim \text{Poisson}(\lambda)$
$X \perp\!\!\!\perp Y$ $X+Y \sim \text{Poisson}(2\lambda)$

$l(\lambda) = (\sum X_i)\log\lambda - n\lambda - \sum_{i=1}^{n} \log(X_i!)$.

$\psi(\lambda) = E\bar{X}_n = \lambda$, $\psi'(\lambda) = 1$

$$\frac{\partial l(\lambda)}{\partial\lambda} = \frac{\sum X_i}{\lambda} - n = \frac{1}{\lambda}\left[\frac{\sum X_i}{n} - \lambda\right] = \frac{1}{\lambda}(\bar{X}_n - \lambda).$$

From this theorem, we know that the lower bound is attained, i.e. $\text{var}(\bar{X}_n) = \frac{1}{nI(\lambda)} = \frac{\lambda}{n}$.

$I(\lambda) = \frac{1}{\lambda}$.

To verify $\text{var}(\bar{X}_n) = \frac{\lambda}{n}$, we look at the sampling distribution of $\bar{X}_n$: $X_1 + \cdots + X_n \sim \text{Poisson}(n\lambda) \rightarrow \text{var}(\bar{X}_n) = \frac{\text{var}(X_1+\cdots+X_n)}{n^2} = \frac{n\lambda}{n^2}$ $\frac{\lambda}{n}$.

Cramer-Rao inequality

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

$|\mathrm{cov}(X,Y)| \le \sqrt{\mathrm{var}X}\sqrt{\mathrm{var}Y}$

Theorem E. Suppose the population has a density $f(x \mid \theta)$. Under the i.i.d and a few other assumptions, let $\hat{\delta}_n = g(X_1, \ldots, X_n)$ be any estimator. Define $\psi(\theta) = E_\theta\left(\hat{\delta}_n\right)$. Then

$$\mathrm{Var}_\theta\left(\hat{\delta}_n\right) \geq \frac{[\psi'(\theta)]^2}{I(\theta)},$$

where the lower bound is attained if and only if $\frac{\partial}{\partial\theta} l(\theta) = a(\theta)\left[\hat{\delta}_n - \psi(\theta)\right]$.

Exchangeability assumption

$\vec{x} = (x_1, \cdots x_n)$

Proof*. 1. $\psi'(\theta) = \frac{\partial}{\partial\theta} E_\theta(\hat{\delta}_n) = \frac{\partial}{\partial\theta}\int \hat{\delta}_n f(x_1, \cdots x_n|\theta)\, d\vec{x} = \int \hat{\delta}_n \frac{\partial}{\partial\theta} f(x_1, \cdots x_n|\theta)\, d\vec{x}$

$= \int \hat{\delta}_n \frac{\frac{\partial}{\partial\theta} f(x_1, \cdots x_n|\theta)}{f(x_1, \cdots x_n|\theta)} \cdot f(x_1, \cdots x_n|\theta)\, d\vec{x}$

$= \int \hat{\delta}_n \frac{\partial}{\partial\theta}[\log f(x_1, \cdots x_n|\theta)]\, f(x_1, \cdots x_n|\theta)\, d\vec{x}$

$= E_\theta\left(\hat{\delta}_n \frac{\partial}{\partial\theta}[\log f(X_1, \cdots X_n|\theta)]\right)$

Cramer-Rao inequality

$$\mathrm{cov}(X,Y) = EXY - (EX)(EY)$$

Proof cont'd.

2. $\left|\mathrm{cov}\left(\delta_n, \frac{\partial}{\partial\theta}[\log f(x_1,\cdots x_n|\theta)]\right)\right| \le \sqrt{\mathrm{var}(\delta_n)}\cdot\sqrt{\mathrm{var}\left[\frac{\partial}{\partial\theta}\log f(x_1,\cdots x_n|\theta)\right]}$

$$\mathrm{cov}\left(\delta_n, \frac{\partial}{\partial\theta}[\log f(x_1,\cdots x_n|\theta)]\right) = E\left\{\underbrace{\delta_n\cdot\frac{\partial}{\partial\theta}[\log f(x_1,\cdots x_n|\theta)]}_{=\psi'(\theta)}\right\} - (E\delta_n)\underbrace{\left\{E\frac{\partial}{\partial\theta}\log f(x_1,\cdots x_n|\theta)\right\}}_{=0}$$

3. $$E\frac{\partial}{\partial\theta}\log f(x_1,\cdots x_n|\theta) \overset{\text{i.i.d}}{=} E\frac{\partial}{\partial\theta}\log\left[\prod_{i=1}^n f(x_i|\theta)\right] = E\frac{\partial}{\partial\theta}\left[\sum_{i=1}^n \log f(x_i|\theta)\right]$$

$$= \sum_{i=1}^n E\left(\frac{\partial}{\partial\theta}\log f(x_i|\theta)\right) \overset{\text{i.i.d}}{=} nE\left[\frac{\partial}{\partial\theta}\log f(x|\theta)\right]$$

$$= n\int\left(\frac{\partial}{\partial\theta}\log f(x|\theta)\right) f(x|\theta)\,d\theta = n\int\frac{\frac{\partial}{\partial\theta}f(x|\theta)}{\cancel{f(x|\theta)}}\cancel{f(x|\theta)}\,d\theta$$

$$= n\int\frac{\partial}{\partial\theta}f(x|\theta)\,d\theta \overset{\text{Exchangeability}}{=} n\frac{\partial}{\partial\theta}\int f(x|\theta)\,d\theta = 0.$$

Plug the results back into the Cauchy-Schwartz inequality, and we have:

$$|\psi'(\theta)| \le \sqrt{\operatorname{var}\hat{\delta}_n}\ \sqrt{\operatorname{var}\left(\frac{\partial}{\partial\theta}\log f(x_1,\cdots x_n|\theta)\right)}.$$

4. $\operatorname{var}\left[\frac{\partial}{\partial\theta}\log f(X_1,\cdots X_n|\theta)\right] = E\left[\frac{\partial}{\partial\theta}\log f(x_1,\cdots x_n|\theta)\right]^2 - \underbrace{\left[E\frac{\partial}{\partial\theta}\log f(x_1,\cdots x_n|\theta)\right]^2}_{=0 \text{ from step } 3}$

$\operatorname{var}X = EX^2 - (EX)^2$

$$E\left[\frac{\partial}{\partial\theta}\log f(X_1,\cdots X_n|\theta)\right]^2$$

$$= E\left[\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\log f(X_i|\theta)\right]^2 = E\left[\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\partial}{\partial\theta}\log f(X_i|\theta)\cdot\frac{\partial}{\partial\theta}\log f(X_j|\theta)\right]$$

expand the summation

$EXY = EXEY$ if X, Y are indep

$$= \sum_{i=1}^{n} E\left[\frac{\partial}{\partial\theta}\log f(X_i|\theta)\right]^2 + \sum_{i\ne j} E\left[\frac{\partial}{\partial\theta}\log f(X_i|\theta)\cdot\frac{\partial}{\partial\theta}\log f(X_j|\theta)\right]$$

independent

$$\stackrel{iid}{=} n\,\underbrace{E\left[\frac{\partial}{\partial\theta}\log f(X|\theta)\right]^2}_{=I(\theta)} + \sum_{i\ne j} \underbrace{E\left[\frac{\partial}{\partial\theta}\log f(X_i|\theta)\right]}_{=0}\cdot E\left[\frac{\partial}{\partial\theta}\log f(X_j|\theta)\right]_{=0}$$

By definition of Fisher Information,

$$E_\theta\left[\frac{\partial}{\partial\theta}\log f(x_1,\cdots x_n|\theta)\right]^2 = n\,I(\theta).$$

5. Plug the above result back into the Cauchy-Schwartz inequality, and we get:

$$|\psi'(\theta)| \le \sqrt{\mathrm{var}(\hat{\delta}_n)}\sqrt{n\,I(\theta)}$$

$$\Leftrightarrow \quad \mathrm{var}(\hat{\delta}_n) \ge \frac{[\psi'(\theta)]^2}{n\,I(\theta)}$$

The lower bound is only attained when Cauchy inequality attains its lower bound:

$$\frac{\partial}{\partial\theta}\underbrace{\log f(x_1,\cdots x_n|\theta)}_{=\ell(\theta)} = a(\theta)\cdot\hat{\delta}_n + b(\theta)$$

$$\Leftrightarrow \frac{\partial}{\partial\theta}\ell(\theta) = a(\theta)\hat{\delta}_n + b(\theta)$$

$$E\left(\frac{\partial}{\partial\theta}\ell(\theta)\right) = 0$$

$$\frac{\partial}{\partial\theta}\ell(\theta) = a(\theta)\left[\hat{\delta}_n - \psi(\theta)\right]$$

$E_\theta(\hat{\delta}_n)$

Cramer-Rao inequality

Definition. An estimator whose variance achieves the CR lower bound is said to be **efficient**. If the *asymptotic variance* of an estimator is equal to the lower bound, then it is said to be **asymptotically efficient**.

Theorem F. Under the assumptions that ensure asymptotic normality (*see Theorem D of this lecture*), MLE are always asymptotically efficient.

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

$\psi(\theta) = E_\theta \hat{\theta}_n \approx \theta$, $\psi'(\theta) = 1$.

$\operatorname{var}(\hat{\theta}_n) \approx \frac{1}{nI(\theta)} = \frac{[\psi'(\theta)]^2}{nI(\theta)}$ ← The C-R lower bound is attained asymptotically.

Tomorrow...

- More examples on Cramer-Rao lower bound;
- Exact sampling distribution for $\bar{X}_n$ and $\hat{\sigma}_n^2$ under $N(\mu, \sigma^2)$;
- Sufficiency.