STAT 24400 Lecture 15
Section 8.5.2 Large Sample Theory for MLEs
Section 8.5.3 Confidence Intervals from MLEs

Yibi Huang
Department of Statistics
University of Chicago

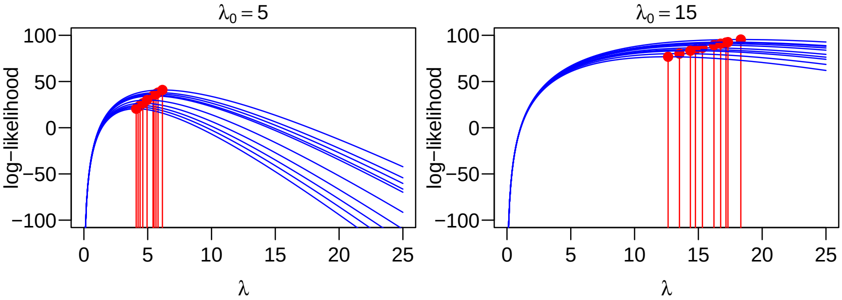
Accuracy of the MLE

Example: Suppose $X_1, \dots, X_{50} \overset{\text{iid}}{\sim}$ Exponential(λ_0).

Recall the log likelihood for i.i.d. Exponential(λ) is

$$\ell(\lambda) = n\log(\lambda) - n\lambda\overline{X}$$

Here is a plot of the log likelihood function $\ell(\lambda)$, and the MLE, over 10 trials:



$\Rightarrow$ higher curvature of $\ell(\lambda)$ around the true value λ_0 leads to a more accurate estimate

Curvature of a Function (Calculus Review)

For a sufficiently smooth function $g(u)$, if u_0 is a local maximum or minimum of $g(u)$, then $g'(u_0) = 0$ and its Taylor expansion around $u = u_0$ would be

$$\begin{aligned} g(u) &\approx g(u_0) + \overbrace{g'(u_0)}^{=0}(u - u_0) + \frac{g''(u_0)}{2}(u - u_0)^2, \\ &\approx g(u_0) + \frac{g''(u_0)}{2}(u - u_0)^2, \quad \text{for } u \approx u_0. \end{aligned}$$

The curvature of $g(u)$ at a local maximum or or minimum $u = u_0$ is reflected by its second derivative at u_0,

$$g''(u_0) = \frac{d^2}{du^2} g(u)\Big|_{u=u_0}$$

- $g''(u_0) > 0$ if $g(u)$ has a upward concavity at u_0
- $g''(u_0) < 0$ if $g(u)$ has a downward concavity at u_0
- The greater the magnitude of $g''(u_0)$, the greater the curvature

Curvature of the Log Likelihood

For $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x \mid \theta)$ for an unknown parameter θ, recall the log likelihood for θ is

$$\ell(\theta) = \sum_{i=1}^{n} \log f(X_i \mid \theta).$$

Its second derivative is

$$\frac{\partial^2}{\partial\theta^2}\ell(\theta) = \sum_{i=1}^{n} \frac{\partial^2}{\partial\theta^2} \log f(X_i \mid \theta).$$

By LLN, as $n \to \infty$,

$$\frac{1}{n}\frac{\partial^2}{\partial\theta^2}\ell(\theta) = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2}{\partial\theta^2} \log f(X_i \mid \theta) \longrightarrow \mathrm{E}\left[\frac{\partial^2}{\partial\theta^2} \log f(X \mid \theta)\right].$$

where the expected value is taken with respect to X.

Thus the accuracy of MLE can be reflected by $\mathrm{E}\left[\frac{\partial^2}{\partial\theta^2} \log f(X_i \mid \theta)\right]$.

Fisher Information

(From this point on, we assume there is only a single parameter θ.)

For a PDF/PMF $f(X \mid \theta)$ with a single parameter θ, the *Fisher information* for θ is defined as:

$$\mathcal{I}(\theta) = -\operatorname{E}\left[\frac{\partial^2}{\partial\theta^2}\log f(X_i \mid \theta)\right]$$

- Usually, $\frac{\partial^2}{\partial\theta^2}\log f(X_i \mid \theta) < 0$ as the log likelihood generally has a downward concavity. We add the minus sign $-$ to get rid of the sign and ensure that $\mathcal{I}(\theta) > 0$
- $\mathcal{I}(\theta)$ reflects the curvature of the log likelihood. The greater the value of $\mathcal{I}(\theta)$, the less variability of the MLE $\hat{\theta}$.
- $\mathcal{I}(\theta)$ measures the amount of information that an observed random variable $X \sim f(X \mid \theta)$ carries about an unknown parameter θ.

Examples: Fisher Information $\mathcal{I}(\theta) = \mathrm{E}\left(-\frac{\partial^2}{\partial\theta^2}\log f(X \mid \theta)\right)$

Ex1: Exponential(λ):

- PDF $f(x \mid \lambda) = \lambda e^{-\lambda x}$
- $\log f(X \mid \lambda) = \log(\lambda) - \lambda X$
- $\frac{\partial}{\partial\lambda}\log f(X \mid \lambda) = \dfrac{1}{\lambda} - X$
- $\frac{\partial^2}{\partial\lambda^2}\log f(X \mid \lambda) = -1/\lambda^2$
- $\mathcal{I}(\lambda) = \mathrm{E}\left(-\frac{\partial^2}{\partial\lambda^2}\log f(X \mid \lambda)\right) = \mathrm{E}(1/\lambda^2) = 1/\lambda^2$

Ex2: N(μ, σ^2) with σ^2 known:

- PDF: $f(x \mid \mu) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(X-\mu)^2/2\sigma^2}$
- $\log f(X \mid \mu) = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{(X-\mu)^2}{2\sigma^2}$
- $\frac{\partial}{\partial\mu}\log f(X \mid \mu) = (X-\mu)/\sigma^2$
- $\frac{\partial^2}{\partial\mu^2}\log f(X \mid \mu) = -1/\sigma^2$
- $\mathcal{I}(\mu) = -\mathrm{E}\left(\frac{\partial^2}{\partial\mu^2}\log f(X \mid \mu)\right) = 1/\sigma^2$

Examples: Fisher Information

Ex3: Bernoulli(p):

- PMF: $f(x \mid p) = p^X (1-p)^{1-X}$
- $\log f(X \mid p) = X \log(p) + (1-X)\log(1-p)$
- $\frac{\partial}{\partial p} \log f(X \mid p) = \frac{X}{p} - \frac{1-X}{1-p}$
- $\frac{\partial^2}{\partial p^2} \log f(X \mid p) = -\frac{X}{p^2} - \frac{1-X}{(1-p)^2}$
- $\mathcal{I}(p) = -\,\mathrm{E}\left(\frac{\partial^2}{\partial p^2} \log f(X \mid p)\right) = \frac{\mathrm{E}(X)}{p^2} + \frac{1-\mathrm{E}(X)}{(1-p)^2} = \frac{1}{p(1-p)}$

Asymptotic (Large Sample) Distribution of the MLE

Fisher information determines the (approx) variance of the MLE.

Informally: if $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x \mid \theta_0)$ and $\hat{\theta}$ is the MLE,

$$\text{the distribution of } \hat{\theta} \text{ is approx. } N\left(\theta_0,\ \frac{1}{n\mathcal{I}(\theta_0)}\right)$$

Asymptotic (Large Sample) Distribution of the MLE

Fisher information determines the (approx) variance of the MLE.

Informally: if $X_1, \dots, X_n \overset{\text{iid}}{\sim} f(x \mid \theta_0)$ and $\hat{\theta}$ is the MLE,

$$\text{the distribution of } \hat{\theta} \text{ is approx. } N\left(\theta_0,\ \frac{1}{n\mathcal{J}(\theta_0)}\right)$$

More formally: under some regularity conditions ($f(x \mid \theta)$ is a smooth function of θ),

$$\sqrt{n\mathcal{J}(\theta_0)} \cdot (\hat{\theta} - \theta_0) \quad \text{converge in distribution to} \quad \mathsf{N}(0, 1)$$

This means that the CDF converges — i.e., for all fixed x,

$$\mathrm{P}\left(\sqrt{n\mathcal{J}(\theta_0)} \cdot (\hat{\theta} - \theta_0) \leq x\right) \to \Phi(x) \quad \text{as } n \to \infty.$$

The same holds with $\mathcal{J}(\hat{\theta})$ in place of $\mathcal{J}(\theta_0)$:

$$\sqrt{n\mathcal{J}(\hat{\theta})} \cdot (\hat{\theta} - \theta_0) \quad \text{converge in distribution to} \quad \mathsf{N}(0, 1)$$

Asymptotic Distribution of the MLE — Examples

- Exponential(λ): $\hat{\lambda} = 1/\overline{X}$ and $\mathcal{I}(\lambda) = 1/\lambda^2$, so:

$$\hat{\lambda} \approx \mathsf{N}\left(\lambda_0, \frac{\lambda_0^2}{n}\right) \text{ or } \approx \mathsf{N}\left(\lambda_0, \frac{\widehat{\lambda}^2}{n}\right)$$

Asymptotic Distribution of the MLE — Examples

- Exponential(λ): $\hat{\lambda} = 1/\overline{X}$ and $\mathcal{I}(\lambda) = 1/\lambda^2$, so:

$$\hat{\lambda} \approx \mathsf{N}(\lambda_0, \tfrac{\lambda_0^2}{n}) \text{ or } \approx \mathsf{N}(\lambda_0, \tfrac{\widehat{\lambda}^2}{n})$$

- $\mathsf{N}(\mu, \sigma^2)$ with σ^2 known: $\hat{\mu} = \overline{X}$ and $\mathcal{I}(\mu) = 1/\sigma^2$ so:

$$\hat{\mu} \approx \mathsf{N}(\mu_0, \frac{\sigma^2}{n})$$

(In this case we know this is the exact distribution!)

Asymptotic Distribution of the MLE — Examples

- Exponential(λ): $\hat{\lambda} = 1/\overline{X}$ and $\mathcal{I}(\lambda) = 1/\lambda^2$, so:

$$\hat{\lambda} \approx \mathsf{N}(\lambda_0, \tfrac{\lambda_0^2}{n}) \text{ or } \approx \mathsf{N}(\lambda_0, \tfrac{\widehat{\lambda}^2}{n})$$

- $\mathsf{N}(\mu, \sigma^2)$ with σ^2 known: $\hat{\mu} = \overline{X}$ and $\mathcal{I}(\mu) = 1/\sigma^2$ so:

$$\hat{\mu} \approx \mathsf{N}(\mu_0, \frac{\sigma^2}{n})$$

(In this case we know this is the <u>exact</u> distribution!)

- Bernoulli(p): $\hat{p} = \overline{X}$ and $\mathcal{I}(p) = \frac{1}{p(1-p)}$, so:

$$\hat{p} \approx \mathsf{N}(p_0, \tfrac{p_0(1-p_0)}{n}) \text{ or } \approx \mathsf{N}(p_0, \tfrac{\hat{p}(1-\hat{p})}{n})$$

A Counter Example: Asymptotic Distribution of the MLE

For Uniform$[0, \theta]$:

- PDF $f(x \mid \theta) = \frac{1}{\theta}$, $0 \leq x \leq \theta$
- In this case the regularity conditions do not hold.
  $\log(f(X \mid \theta))$ is not a smooth function of θ,

$$\log(f(X \mid \theta)) = \begin{cases} -\log\theta & \text{if } \theta > X \\ \log(0) = -\infty & \text{if } \theta < X \end{cases}$$

- Recall in L14, we showed that $\hat{\theta}_{\text{MLE}} = X_{(n)}$ and calculated

$$\mathrm{Var}(\hat{\theta}) = \frac{n\theta^2}{(n+1)^2(n+2)} = \mathcal{O}(\frac{1}{n^2})$$

while asymptotic normality of the MLE would yield $\mathrm{Var}(\hat{\theta}) = \mathcal{O}(\frac{1}{n})$

A Counter Example: Asymptotic Distribution of the MLE

For Uniform$[0, \theta]$:

- PDF $f(x \mid \theta) = \frac{1}{\theta}$, $0 \leq x \leq \theta$
- In this case the regularity conditions do not hold. $\log(f(X \mid \theta))$ is not a smooth function of θ,

$$\log(f(X \mid \theta)) = \begin{cases} -\log\theta & \text{if } \theta > X \\ \log(0) = -\infty & \text{if } \theta < X \end{cases}$$

- Recall in L14, we showed that $\hat{\theta}_{\text{MLE}} = X_{(n)}$ and calculated

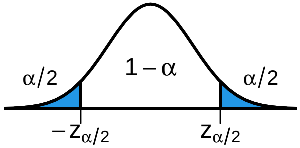
$$\mathrm{Var}(\hat{\theta}) = \frac{n\theta^2}{(n+1)^2(n+2)} = \mathcal{O}(\frac{1}{n^2})$$

while asymptotic normality of the MLE would yield $\mathrm{Var}(\hat{\theta}) = \mathcal{O}(\frac{1}{n})$

In fact, no approximation is needed here, since we actually know the exact distribution of the MLE in this case (via order statistics)

Confidence Intervals Based on MLE

We can use asymptotic normality of the MLE to construct a *confidence interval for θ_0*, where θ_0 is the true value of θ.
Let $z_{\alpha/2}$ be the value so that



$$\mathrm{P}(|Z| \le z_{\alpha/2}) = 1 - \alpha \text{ for } Z \sim N(0,1).$$

$$\sqrt{n\mathcal{J}(\hat{\theta})} \cdot (\hat{\theta} - \theta_0) \to \mathsf{N}(0,1)$$

$$\Rightarrow \quad \mathrm{P}\left(\left|\sqrt{n\mathcal{J}(\hat{\theta})} \cdot (\hat{\theta} - \theta_0)\right| < z_{\alpha/2}\right) \approx 1 - \alpha$$

$$\Rightarrow \quad \mathrm{P}\left(\hat{\theta} - z_{\alpha/2} \cdot \frac{1}{\sqrt{n\mathcal{J}(\hat{\theta})}} < \theta_0 < \hat{\theta} + z_{\alpha/2} \cdot \frac{1}{\sqrt{n\mathcal{J}(\hat{\theta})}}\right) \approx 1 - \alpha$$

We have approximately $(1-\alpha)$ confidence that θ_0 lies in the interval below

$$\hat{\theta} \pm z_{\alpha/2} \cdot \frac{1}{\sqrt{n\mathcal{J}(\hat{\theta})}}.$$

Example — Confidence Interval for Normal Mean

$X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{N}(\mu, \sigma^2)$ for unknown $\mu \in \mathbb{R}$ (σ^2 is known)

- The MLE is $\hat{\mu} = \overline{X}$
- The Fisher information is $\mathcal{I}(\mu) = \frac{1}{\sigma^2}$
- Therefore, $\hat{\mu} \approx \mathsf{N}\left(\mu_0, \frac{\sigma^2}{n}\right)$ and an approx $(1 - \alpha)$ conf. int. is:

$$\overline{X} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

- In fact, we know this distribution and conf. int. are <u>exact</u> for this case

Examples — Confidence Interval for Exponential

$X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \text{Exponential}(\lambda)$ for unknown $\lambda > 0$

- The MLE is $\hat{\lambda} = \frac{1}{\overline{X}}$
- The Fisher information is $\mathcal{I}(\lambda) = \frac{1}{\lambda^2}$
- Therefore, $\hat{\lambda} \approx \mathsf{N}(\lambda_0, \frac{\lambda_0^2}{n})$ and an approx. $(1-\alpha)$ conf. int. is:

$$\hat{\lambda} \pm z_{\alpha/2} \cdot \frac{\hat{\lambda}}{\sqrt{n}} = \left(\hat{\lambda} - z_{\alpha/2} \cdot \frac{\hat{\lambda}}{\sqrt{n}}, \hat{\lambda} + z_{\alpha/2} \cdot \frac{\hat{\lambda}}{\sqrt{n}}\right)$$

Example — Bernoulli p

$X_1, \dots, X_n \overset{\text{iid}}{\sim} \text{Bernoulli}(p)$ for unknown $p \in (0,1)$

- The MLE is $\hat{p} = \overline{X}$
- The Fisher information is $\mathcal{I}(p) = \frac{1}{p(1-p)}$
- Therefore, $\hat{p} \approx \text{N}(p_0, \frac{p_0(1-p_0)}{n}) \approx \text{N}(p_0, \frac{\hat{p}(1-\hat{p})}{n})$ and an approx. $(1-\alpha)$ conf. int. is:

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \left(\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

Cramer-Rao Lower Bound (CRLB)

Many possible estimators for parameters (MME, MLE, etc). Is there a best one?

Theorem: Let $X_1, \dots, X_n$ be i.i.d. with PDF/PMF $f(x \mid \theta)$. Let $T = t(X_1, \dots, X_n)$ be an unbiased estimate for θ. Then, under smoothness assumptions on $f(x \mid \theta)$,

$$\mathrm{Var}(T) \geq \frac{1}{n\mathcal{J}(\theta)}.$$

For the MLE $\hat{\theta}$ of θ, recall $\hat{\theta}$ is approx. $N\left(\theta,\ \frac{1}{n\mathcal{J}(\theta)}\right)$.

- The MLE is (asymptotically) unbiased
- The MLE's variance is (asymptotically) $\frac{1}{n\mathcal{J}(\theta)}$
- The MLE thus (asymptotically) achieves the CRLB

Is the MLE optimal?

- Not necessarily...
  there might be biased estimators with a smaller MSE

Lemma for the Proof of CRLB

If $\log f(X \mid \theta)$ is a smooth function of θ, it can be shown that

1. $\mathrm{E}\left(\frac{\partial}{\partial\theta} \log f(X \mid \theta)\right) = 0$
2. the Fisher information $\mathcal{I}(\theta)$ can also be calculated as

$$\mathcal{I}(\theta) = \mathrm{E}\left(\left(\frac{\partial}{\partial\theta} \log f(X \mid \theta)\right)^2\right)$$

The two points above combined also implies that

$$\mathrm{Var}\left(\frac{\partial}{\partial\theta} \log f(X \mid \theta)\right) = \mathrm{E}\left(\left(\frac{\partial}{\partial\theta} \log f(X \mid \theta)\right)^2\right) = \mathcal{I}(\theta).$$

since $\mathrm{E}\left(\frac{\partial}{\partial\theta} \log f(X \mid \theta)\right) = 0$

Proof of $\mathrm{E}\left(\frac{\partial}{\partial\theta}\log f(X \mid \theta)\right) = 0$

The proof is done for continuous X. The discrete case can be done similarly.

$$
\begin{aligned}
\mathrm{E}\left(\frac{\partial}{\partial\theta}\log f(X \mid \theta)\right) &= \int \frac{\partial}{\partial\theta}\log f(x \mid \theta) f(x \mid \theta)\mathrm{d}x \\
&= \int \frac{\frac{\partial}{\partial\theta} f(x \mid \theta)}{f(x \mid \theta)} f(x \mid \theta)\mathrm{d}x \\
&= \int \frac{\partial}{\partial\theta} f(x \mid \theta)\mathrm{d}x \\
&= \frac{\partial}{\partial\theta}\underbrace{\int f(x \mid \theta)\mathrm{d}x}_{=1} \quad \left(\begin{array}{c}\text{assume it's okay to swap the order}\\ \text{of integration \& differentiation}\end{array}\right) \\
&= \frac{\partial}{\partial\theta} 1 = 0
\end{aligned}
$$

Proof that $\mathcal{J}(\theta) = \mathrm{E}\left(\left(\frac{\partial}{\partial\theta}\log f(X \mid \theta)\right)^2\right)$

From the proof in the previous page, we've obtained that

$$0 = \int \frac{\partial}{\partial\theta}\log f(x \mid \theta) f(x \mid \theta)\mathrm{d}x.$$

Taking another derivative of the preceding expressions, and swapping the order of differentiation and integration, we have

$$0 = \underbrace{\int \frac{\partial^2}{\partial\theta^2}\log f(x \mid \theta) f(x \mid \theta)\mathrm{d}x}_{=I} + \underbrace{\int \frac{\partial}{\partial\theta}\log f(x \mid \theta) \cdot \frac{\partial}{\partial\theta} f(x \mid \theta)\mathrm{d}x}_{=II}$$

where

$$I = \mathrm{E}\left[\frac{\partial^2}{\partial\theta^2}\log f(X_i \mid \theta)\right] = -\mathcal{J}(\theta).$$

$$
\begin{aligned}
II &= \int \frac{\partial}{\partial\theta} \log f(x \mid \theta) \cdot \frac{\partial}{\partial\theta} f(x \mid \theta) \mathrm{d}x \\
&= \int \frac{\partial}{\partial\theta} \log f(x \mid \theta) \cdot \underbrace{\frac{\frac{\partial}{\partial\theta} f(x \mid \theta)}{f(x \mid \theta)}}_{=\frac{\partial}{\partial\theta} \log f(x|\theta)} f(x \mid \theta) \mathrm{d}x \\
&= \int \left[\frac{\partial}{\partial\theta} \log f(x \mid \theta) \right]^2 f(x \mid \theta) \mathrm{d}x \\
&= \mathrm{E} \left(\left(\frac{\partial}{\partial\theta} \log f(X \mid \theta) \right)^2 \right)
\end{aligned}
$$

As $I + II = 0$, and $I = -\mathcal{I}(\theta)$, we have

$$
\mathcal{I}(\theta) = -I = II = \mathrm{E} \left(\left(\frac{\partial}{\partial\theta} \log f(X \mid \theta) \right)^2 \right).
$$

Proof of CRLB

Let

$$Z = \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \log f(X_i \mid \theta) = \sum_{i=1}^{n} \frac{\frac{\partial}{\partial \theta} f(X_i \mid \theta)}{f(X_i \mid \theta)}.$$

As shown in the Lemma that $\mathrm{Var}\left(\frac{\partial}{\partial \theta} \log f(X \mid \theta)\right) = \mathcal{I}(\theta)$, we have

$$\mathrm{Var}(Z) = n\mathcal{I}(\theta).$$

The lemma also asserts $\mathrm{E}(Z) = 0$ since $\mathrm{E}\left(\frac{\partial}{\partial \theta} \log f(X \mid \theta)\right) = 0$

Recall that $T = t(X_1, \dots, X_n)$ is an unbiased estimate for θ. We have

$$[\mathrm{Cov}(Z, T)]^2 \le \mathrm{Var}(Z)\,\mathrm{Var}(T).$$

It remains to show that $\mathrm{Cov}(Z, T) = 1$, then CRLB would follows since

$$\mathrm{Var}(T) \ge \frac{[\mathrm{Cov}(Z, T)]^2}{\mathrm{Var}(Z)} = \frac{1}{n\mathcal{I}(\theta)}.$$

See p.301 of the textbook for the rest of the proof.