STSCI 5080
Probability Models and Inference
Lecture 21: MLE and Confidence Intervals

November 13, 2018

Setting

- Let $\{f_\theta \mid \theta \in \Theta\}$ be a class of pmfs/pdfs where $\Theta \subset \mathbb{R}^k$, and suppose that

  $$X_1, \ldots, X_n \sim f_\theta \text{ i.i.d.}$$

  for some $\theta \in \Theta$.

- The likelihood function is

  $$L_n(\theta) = \prod_{i=1}^{n} f_\theta(X_i).$$

- The log likelihood function is

  $$\ell_n(\theta) = \log L_n(\theta).$$

- The MLE is a maximizer of the log likelihood function:

  $$\ell_n(\widehat{\theta}) = \max_{\theta \in \Theta} \ell_n(\theta).$$

In the one-dimensional case ($k = 1$), the MLE is obtained by solving the first order condition (FOC) w.r.t. θ:

$$\ell_n'(\theta) = 0.$$

Functions of MLE

Definition

Let $\widehat{\theta}$ be the MLE of θ. Then the MLE of $g(\theta)$ is $g(\widehat{\theta})$.

List of MLEs

- $Po(\lambda)$: $\widehat{\lambda} = \overline{X}$.
- $N(\mu, \sigma_0^2)$ (where σ_0^2 is known): $\widehat{\mu} = \overline{X}$.
- $N(0, \sigma^2)$: $\widehat{\sigma}^2 = n^{-1} \sum_{i=1}^{n} X_i^2$.
- $Ex(\lambda)$: $\widehat{\lambda} = 1/\overline{X}$.
- $Bin(n, p)$ (where $X \sim Bin(n, p)$): $\widehat{p} = X/n$.

Convergence in probability and in distribution

- Let Y_n and Y be random variables with cdfs F_n and F, respectively.
- Y_n converges in probability to Y, denoted as $Y_n \xrightarrow{P} Y$, if
$$\lim_{n\to\infty} P(|Y_n - Y| > \varepsilon) = 0$$
for any $\varepsilon > 0$.
- Y_n converges in distribution to Y, denoted as $Y_n \xrightarrow{d} Y$, if
$$\lim_{n\to\infty} F_n(x) = F(x)$$
for any continuity point of F. If $Y \sim N(0, \sigma^2)$ e.g., we also write
$$Y_n \xrightarrow{d} N(0, \sigma^2).$$

Asymptotic properties of MLE

Definition

Suppose $k = 1$ (i.e., θ is one-dim.). An estimator $\widehat{\theta}_n = \widehat{\theta}_n(X_1, \ldots, X_n)$ is consistent for θ if

$$\widehat{\theta}_n \xrightarrow{P} \theta$$

as $n \to \infty$ whatever the value of θ is.
The estimator $\widehat{\theta}_n$ is asymptotically normal if

$$\sqrt{n}(\widehat{\theta}_n - \theta) \xrightarrow{d} N(0, \sigma^2(\theta))$$

as $n \to \infty$, where $\sigma^2(\theta) > 0$.

List of asymptotic distributions of MLEs

- $Po(\lambda)$: $\widehat{\lambda}_n = \overline{X}_n$.

$$\sqrt{n}(\widehat{\lambda}_n - \lambda) \xrightarrow{d} N(0, \lambda).$$

- $N(\mu, \sigma_0^2)$ (where σ_0^2 is known): $\widehat{\mu}_n = \overline{X}_n$.

$$\sqrt{n}(\widehat{\mu}_n - \mu) \sim N(0, \sigma_0^2).$$

- $N(0, \sigma^2)$: $\widehat{\sigma}_n^2 = n^{-1} \sum_{i=1}^n X_i^2$. ??

- $Ex(\lambda)$: $\widehat{\lambda}_n = 1/\overline{X}_n$.

$$\sqrt{n}(\widehat{\lambda}_n - \lambda) \xrightarrow{d} N(0, \lambda^2).$$

- $Bin(n, p)$ (where $X_n \sim Bin(n, p)$): $\widehat{p}_n = X_n/n$.

$$\sqrt{n}(\widehat{p}_n - p) \xrightarrow{d} N(0, p(1-p)).$$

General case (not included in Final)

Theorem

In general, the MLE $\widehat{\theta}_n$ is consistent and asymptotically normal under suitable regularity conditions:

$$\widehat{\theta}_n \xrightarrow{P} \theta,$$
$$\sqrt{n}(\widehat{\theta}_n - \theta) \xrightarrow{d} N(0, 1/I(\theta)),$$

where $I(\theta)$ is the Fisher information:

$$I(\theta) = E_\theta\left[-\frac{\partial^2 \log f_\theta(X_1)}{\partial\theta^2}\right].$$

Regularity conditions

1. The set $\{x \mid f_\theta(x) > 0\}$ does not depend on θ.
2. The true parameter θ is not on the boundary of Θ.
3. The Fisher information $I(\theta)$ is positive.
4. A few more technical conditions.

Regularity conditions

1. The set $\{x \mid f_\theta(x) > 0\}$ does not depend on θ.
2. The true parameter θ is not on the boundary of Θ.
3. The Fisher information $I(\theta)$ is positive.
4. A few more technical conditions.

MLE may not be asymptotically normal

Example

Let

$$X_1, \ldots, X_n \sim U[0, \theta] \text{ i.i.d.}$$

for some $\theta > 0$. The pdf of $U[0, \theta]$ is

$$f_\theta(x) = \begin{cases} \frac{1}{\theta} & \text{if } 0 \leq x \leq \theta \\ 0 & \text{otherwise} \end{cases}.$$

The likelihood function is

$$L_n(\theta) = \begin{cases} \frac{1}{\theta^n} & \text{if } \theta \geq X_{(n)} \\ 0 & \text{otherwise} \end{cases},$$

which is maximized at $X_{(n)}$. So the MLE is

$$\widehat{\theta}_n = X_{(n)}.$$

Example

The cdf of $X_{(n)}$ is

$$P_\theta(X_{(n)} \le x) = \{P_\theta(X_1 \le x)\}^n = \left(\frac{x}{\theta}\right)^n$$

for $0 \le x \le \theta$. Hence, for $x \ge 0$,

$$\begin{aligned} P_\theta\{n(\theta - X_{(n)}) \le x\} &= P_\theta(X_{(n)} \ge \theta - x/n) \\ &= 1 - \left(1 - \frac{x}{n\theta}\right)^n \\ &\to 1 - e^{-x/\theta}. \end{aligned}$$

On the other hand, $P_\theta\{n(\theta - X_{(n)}) \le x\} = 0$ for $x < 0$, and so

$$n(\theta - X_{(n)}) \xrightarrow{d} Ex(1/\theta).$$

Better estimator than MLE (not included in Final)

- Let
$$X_1, \ldots, X_n \sim U[0, \theta] \text{ i.i.d.}$$
- Consider to evaluate an estimator $\widehat{\theta}$ based on the MSE (mean squared error):
$$R(\theta, \widehat{\theta}) = E_\theta\{(\widehat{\theta} - \theta)^2\}$$
which can be decomposed as
$$R(\theta, \widehat{\theta}) = \underbrace{\{E_\theta(\widehat{\theta}) - \theta\}^2}_{\text{bias}} + \underbrace{\mathrm{Var}_\theta(\widehat{\theta})}_{\text{variance}}.$$
- The MLE of θ is $\widehat{\theta} = X_{(n)}$.

- We note that

$$E_\theta(\widehat{\theta}) = \frac{n}{n+1}\theta \quad \text{and} \quad \mathrm{Var}_\theta(\widehat{\theta}) = \frac{n}{(n+2)(n+1)^2}\theta^2.$$

Note: if $X \sim U[0,\theta]$, then $X/\theta \sim U[0,1]$. The MSE is

$$R(\theta,\widehat{\theta}) = \frac{2n+2}{(n+2)(n+1)^2}\theta^2.$$

- Consider instead an unbiased estimator

$$\widetilde{\theta} = \frac{n+1}{n}X_{(n)}$$

so that $E_\theta(\widetilde{\theta}) = \theta$ for any $\theta > 0$. We note that

$$R(\theta,\widetilde{\theta}) = \mathrm{Var}_\theta(\widetilde{\theta}) = \left(\frac{n+1}{n}\right)^2 \mathrm{Var}_\theta(X_{(n)}) = \frac{1}{n(n+2)}\theta^2.$$

Hence,

$$\lim_{n\to\infty} \frac{R(\theta, \widetilde{\theta})}{R(\theta, \widehat{\theta})} = \frac{1}{2}.$$

In terms of MSE, $\widetilde{\theta}$ is better than the MLE $\widehat{\theta}$.

Confidence intervals

Definition

Suppose that $k = 1$ (i.e., θ is one-dim.) and let $\alpha \in (0, 1)$.

- A data dependent interval $[A_n, B_n]$, where $A_n = A_n(X_1, \ldots, X_n)$ and $B_n = B_n(X_1, \ldots, X_n)$, is a confidence interval (CI) with level $1 - \alpha$ for θ if

$$\underbrace{P_\theta(A_n \le \theta \le B_n)}_{\text{coverage probability}} \ge 1 - \alpha$$

  for any $\theta \in \Theta$.

- The interval $[A_n, B_n]$ is a confidence interval with asymptotic level $1 - \alpha$ for θ if

$$\lim_{n \to \infty} P_\theta(A_n \le \theta \le B_n) \ge 1 - \alpha$$

  for any $\theta \in \Theta$.

- The parameter θ is not random, but the end points A_n and B_n are random variables!
- Common choices of α: $\alpha = 0.05$ or 0.01. "a confidence interval with level 95%".
- A CI should be small as long as the level is verified!

Example 21.1

Example

Let

$$X \sim Bin(n, p) \quad \text{for some } 0 < p < 1.$$

Is the interval $[0, 1]$ a CI with level 95 % (say) for p?

Example 21.1

Example

Let

$$X \sim Bin(n,p) \quad \text{for some } 0 < p < 1.$$

Is the interval $[0,1]$ a CI with level 95 % (say) for p?

Answer: Yes.

$$P_p(0 \leq p \leq 1) = 1 \geq 0.95.$$

Example 21.1

Example

Let

$$X \sim Bin(n,p) \quad \text{for some } 0 < p < 1.$$

Is the interval $[0,1]$ a CI with level 95 % (say) for p?

Answer: Yes.

$$P_p(0 \leq p \leq 1) = 1 \geq 0.95.$$

Is this CI practically useful?

Example 21.1

Example

Let

$$X \sim Bin(n, p) \quad \text{for some } 0 < p < 1.$$

Is the interval $[0, 1]$ a CI with level 95 % (say) for p?

Answer: Yes.

$$P_p(0 \leq p \leq 1) = 1 \geq 0.95.$$

Is this CI practically useful?

Answer: No!

Rule of thumb

We should construct a CI $[A_n, B_n]$ in such a way that

$$P_\theta(A_n \leq \theta \leq B_n) = 1 - \alpha \tag{*}$$

for any θ, or

$$\lim_{n\to\infty} P_\theta(A_n \leq \theta \leq B_n) = 1 - \alpha$$

for any θ if the requirement (*) is too stringent.

Example 21.2

Let

$$X_1, \ldots, X_n \sim N(\mu, \sigma_0^2) \text{ i.i.d.}$$

where μ is unknown but σ_0^2 is known. The MLE is

$$\widehat{\mu} = \overline{X} \sim N(\mu, \sigma_0^2/n).$$

Recall: a linear combination of independent normal random variables is normal! So

$$\widehat{\mu} = \mu + \sigma_0 Z/\sqrt{n} \quad \text{for some } Z \sim N(0, 1).$$

In other words,

$$\frac{\sqrt{n}(\widehat{\mu} - \mu)}{\sigma_0} = Z \sim N(0, 1).$$

We note that

$$
\begin{aligned}
& P_\mu\left\{\left|\frac{\sqrt{n}(\widehat{\mu}-\mu)}{\sigma_0}\right| \leq z\right\} \\
& = P(|Z| \leq z) = P(-z \leq Z \leq z) = P(Z \leq z) - P(Z \leq -z) \\
& = \Phi(z) - \Phi(-z) = 2\Phi(z) - 1,
\end{aligned}
$$

where $\Phi(z)$ is the cdf of $N(0,1)$ and recall that

$$\Phi(-z) = 1 - \Phi(z).$$

In addition, we note that

$$\left|\frac{\sqrt{n}(\widehat{\mu}-\mu)}{\sigma_0}\right| \leq z \Leftrightarrow \widehat{\mu} - \frac{z\sigma_0}{\sqrt{n}} \leq \mu \leq \widehat{\mu} + \frac{z\sigma_0}{\sqrt{n}}.$$

In conclusion, we have

$$P_\mu \left\{ \widehat{\mu} - \frac{z\sigma_0}{\sqrt{n}} \leq \mu \leq \widehat{\mu} + \frac{z\sigma_0}{\sqrt{n}} \right\} = 2\Phi(z) - 1.$$

We should choose z in such a way that

$$2\Phi(z) - 1 = 1 - \alpha,$$

which leads to

$$z_{\alpha/2} = \Phi^{-1}(1 - \alpha/2).$$

For example,

$$z_{\alpha/2} \approx \begin{cases} 1.96 & \text{if } \alpha = 0.05 \\ 2.58 & \text{if } \alpha = 0.01 \end{cases}.$$

Recap

CI for μ with level $1-\alpha$

A CI for μ with level $1-\alpha$ is given by

$$\left[\widehat{\mu} - \frac{z_{\alpha/2}\sigma_0}{\sqrt{n}}, \widehat{\mu} + \frac{z_{\alpha/2}\sigma_0}{\sqrt{n}}\right].$$

If $\alpha = 0.05$, this CI will be

$$\left[\widehat{\mu} - \frac{1.96\sigma_0}{\sqrt{n}}, \widehat{\mu} + \frac{1.96\sigma_0}{\sqrt{n}}\right].$$