

STAT201B Lecture 8 Properties of MLE

Logic ▾

我们将分别介绍 MLE 的几个重要性质 (更多论述见 [STA3020 Lecture 7](#)):

1. Equivariance (invariance)
2. Consistency
3. Asymptotic normality
4. Asymptotic efficiency

对于后三个性质, 我们主要考虑 $\theta \in \Theta \subset \mathbb{R}$ 的情况

1 MLE 的 Equivariance

1.1 Definition: $g(\theta)$ 的 log-likelihood

若 θ 的 log-likelihood function 为 $l_n(\theta)$

则对于任意 function $g(\theta)$, $g(\theta)$ 的 log-likelihood function 被定义为:

$$l_g(\phi) := \sup_{\theta \in \Theta, g(\theta) = \phi} l(\theta)$$

Remark ▾

1. 换言之, $g(\theta) = \phi$ 时的 log-likelihood = 先限定 θ 满足 $g(\theta) = \phi$, 再 maximize $l(\theta)$
2. 这么定义是为了满足 MLE 的 equivariance (invariance) property

1.2 Theorem: MLE 的 equivariance

令:

1. $\tau = g(\theta)$ 为关于 θ 的函数
2. $\hat{\theta}_n$ 为 θ 的 MLE

则 $\hat{\tau}_n = g(\hat{\theta}_n)$ 为 τ 的 MLE

Proof: g 为 one-to-one mapping 时的证明 ▾

若 g 为 one-to-one, 则存在 inverse g^{-1} , 因此关于 τ 的 (induced) likelihood 可以被定义为 $\mathcal{L}^*(\tau) = \mathcal{L}(g^{-1}(\tau))$.

注意到对于任意 τ , 有

$$\mathcal{L}^*(\tau) = \mathcal{L}(g^{-1}(\tau)) \leq \mathcal{L}(\hat{\theta}_n) = \mathcal{L}^*(g(\hat{\theta}_n))$$

因此 $\hat{\tau} = g(\hat{\theta})$ maximizes \mathcal{L}^*

Proof: 更 general 的证明 ▾

Proof of Theorem.1.7. We proof by contradiction. If there exist some other $\theta_0 \neq \hat{\theta}_{MLE}$ such that $\phi_0 = g(\theta_0) \neq g(\hat{\theta}_{MLE})$, and $\phi_0 = g(\theta_0)$ is the MLE of $g(\theta)$. Then

$$\ell_g(\phi_0) = \max_{\theta \in \Theta, g(\theta) = \phi_0} \ell(\theta) > \ell_g(g(\hat{\theta}_{MLE})) = \max_{\theta \in \Theta, g(\theta) = g(\hat{\theta}_{MLE})} \ell(\theta) = \ell(\hat{\theta}_{MLE}),$$

which contradict with the fact that $\hat{\theta}_{MLE}$ being the MLE of $\ell(\theta)$. Hence $g(\hat{\theta}_{MLE})$ is the MLE of $g(\theta)$. \square

| 2 MLE 的 Consistency

| 2.1 Definition: MLE 的 consistency condition

1. **independent distributed**: $X_1, \dots, X_n \stackrel{i.i.d.}{\sim} f(x; \theta)$
2. **identifiability**: 若 $\theta \neq \theta'$, 则 $f(x; \theta) \neq f(x; \theta')$
3. **common support**: $\text{Support } \{x : f(x; \theta) > 0\}$ 与 θ 的取值无关
4. **interior**: parameter space Θ 包含一个 open set ω , true parameter value θ^* 为其 interior point
5. **differentiable**: function $f(x; \theta)$ 在 ω 内部关于 θ differentiable

⚠ Remark ▾

[Exponential family](#) 中的 distributions 满足该 condition

⚠ Remark: 和 STA3020 中的 condition 的联系 ▾

在 [STA3020 Lecture 7](#) 中, consistency condition 为:

1. **Separation condition**: Θ 为 compact 或 Θ 满足 $\sup_{\theta \in \Theta, |\theta - \theta^*| \geq \epsilon} \mathbb{E}[l(\theta)] < \mathbb{E}[l(\theta^*)]$, $\forall \epsilon > 0$ ($\mathbb{E}[l(\theta_0)]$ 严格大于其他 $\mathbb{E}[l(\theta)]$ 的 supremum)
2. **Convergence condition**: $\frac{1}{n} l_n(\theta)$ converges uniformly to $\frac{1}{n} \mathbb{E}[l_n(\theta)]$ in probability, 即

$$\sup_{\theta \in \Theta} \frac{1}{n} |l_n(\theta) - \mathbb{E}[l_n(\theta)]| \xrightarrow{p} 0$$

事实上, STAT201B 中给出的 consistency condition 和 STA3020 Lecture 7 中的 consistency condition 几乎等价:

Convergence condition 的证明:

令 θ^* 表示 θ 的 true value, 则有

$$\begin{aligned} \frac{1}{n} l_n(\theta) &= \frac{1}{n} \sum_{i=1}^n \log f(X_i; \theta) \\ &\xrightarrow{p} \mathbb{E}_{\theta^*}[\log f(X_1; \theta)] \quad \text{for any fixed } \theta \text{ by WLLN} \end{aligned}$$

Separation condition 的证明:

此处我们证明 $\mathbb{E}_{\theta^*}[\log f(X_1; \theta)]$ is (uniquely) maximized at $\theta = \theta^*$, 对于任意 $\theta \in \Theta$, 有:

$$\begin{aligned} \mathbb{E}_{\theta^*}[\log f(X_1; \theta)] - \mathbb{E}_{\theta^*}[\log f(X_1; \theta^*)] &= \mathbb{E}_{\theta^*} \left[\log \frac{f(X_1; \theta)}{f(X_1; \theta^*)} \right] \\ &\leq \log \left[\mathbb{E}_{\theta^*} \left[\frac{f(X_1; \theta)}{f(X_1; \theta^*)} \right] \right] \quad (\text{Jensen's inequality}) \\ &= \log \left[\int \frac{f(X_1; \theta)}{f(X_1; \theta^*)} \cdot f(X_1; \theta^*) dX_1 \right] \\ &= \log \left[\int f(X_1; \theta) dX_1 \right] \\ &= \log[1] \\ &= 0 \end{aligned}$$

结合 condition 中的 identifiability 和 interior, θ^* 为 unique maximizer

| 2.2 Theorem: MLE 的 consistency

若 MLE 的 consistency condition 满足,

则 $\hat{\theta}_{MLE}$ 为 true value θ^* 的一个 consistent estimator, 即 $\hat{\theta}_{MLE} \xrightarrow{p} \theta^*$

🔗 Proof: STAT201B 的证明思路 ▾

由于

$$\frac{l_n(\theta)}{n} \xrightarrow{p} \mathbb{E}_{\theta^*}[\log f(X_1; \theta)]$$

同时有

$$\hat{\theta}_n := \arg \max_{\theta \in \Theta} \frac{l_n(\theta)}{n} \quad \text{and} \quad \theta^* = \arg \max_{\theta \in \Theta} \mathbb{E}_{\theta^*}[\log f(X_1; \theta)]$$

且 $\hat{\theta}_n$ 和 θ^* 在 consistency condition 下均为 unique, 则

$$\hat{\theta}_n \xrightarrow{p} \theta^*$$

↪ Proof: STA3020 Lecture 7 中的证明 ↩

Proof of Theorem.1.3. Notice that,

$$\hat{\theta}_{MLE} = \arg \max_{\theta \in \Theta} \ell(\theta) = \arg \max_{\theta \in \Theta} \left[\frac{1}{n} \sum_{i=1}^n \log \frac{f(X_i|\theta)}{f(X_i|\theta_0)} \right] \triangleq \arg \max_{\theta \in \Theta} M(\theta)$$

Since we have

$$M(\theta) = \frac{1}{n} \sum_{i=1}^n \log \frac{f(X_i|\theta)}{f(X_i|\theta_0)} \rightarrow \mathbb{E} \left[\log \frac{f(X|\theta)}{f(X|\theta_0)} \right] = -KL(f_{\theta_0} \| f_{\theta}) < 0,$$

where the last inequality sign changes to the equal sign iff $f(x|\theta) \equiv f(x|\theta_0)$, a.s., i.e., $\theta = \theta_0$. Therefore, if Θ is compact, then for $\forall \epsilon > 0$, we have

$$\sup_{\theta \in \Theta, |\theta - \theta_0| \geq \epsilon} \mathbb{E} M(\theta) < \mathbb{E} M(\theta_0) = -KL(f_{\theta_0} \| f_{\theta_0}) = 0. \quad (4.1)$$

Similarly we can also conclude (4.1) under the separation condition (1.1). Now, if we denote

$$\delta = \mathbb{E} M(\theta_0) - \sup_{\theta \in \Theta, |\theta - \theta_0| \geq \epsilon} \mathbb{E} M(\theta) > 0,$$

Since $M(\theta)$ converges uniformly to $\mathbb{E} M(\theta)$ in probability according to (ii) of Condition.1.2. Therefore, there exists $N \in \mathbb{N}^+$ s.t. when $n \geq N$,

$$\begin{aligned} \mathbb{P}(|\hat{\theta}_{MLE} - \theta_0| \geq \epsilon) &= \mathbb{P}\left(\sup_{\theta \in \Theta, |\theta - \theta_0| \geq \epsilon} M(\theta) > M(\theta_0)\right) \\ &\leq \mathbb{P}\left(\sup_{\theta \in \Theta, |\theta - \theta_0| \geq \epsilon} [M(\theta) - \mathbb{E} M(\theta)] > [M(\theta_0) - \mathbb{E} M(\theta_0)] + \delta\right) \\ &\leq \mathbb{P}\left(\sup_{\theta \in \Theta, |\theta - \theta_0| \geq \epsilon} [M(\theta) - \mathbb{E} M(\theta)] > \frac{\delta}{2}\right) + \mathbb{P}\left([M(\theta_0) - \mathbb{E} M(\theta_0)] > -\frac{\delta}{2}\right) \\ &\leq 2\mathbb{P}\left(\sup_{\theta \in \Theta} |M(\theta) - \mathbb{E} M(\theta)| > \frac{\delta}{2}\right) \rightarrow 0. \end{aligned}$$

Hence $\hat{\theta}_{MLE} \xrightarrow{p} \theta_0$. □

3 MLE 的 Asymptotic Normality 和 Asymptotic Efficiency

↪ Logic ↩

关于 score function, fisher information, second Bartlett's identity 的更多论述, 见 [STA3020 Lecture 3](#)

3.1 Definition: Score function

Score function 被定义为 likelihood function 的一阶导数:

$$s(X; \theta) = \frac{\partial}{\partial \theta} \log f(X; \theta)$$

⚠ Remark ↩

1. 若 likelihood function 的 derivative 不存在, 则 score function 不存在
2. Score function 的实际意义是: log-likelihood 的 rate of changes at different values of θ

3.2 Definition: Fisher information

令:

1. observations 的数量为 n

2. log-likelihood function 为 $l_n(\theta)$

则 (n 个 observations 的) **Fisher information** 被定义为:

$$I_n(\theta) = V_\theta(s(X; \theta)) = V_\theta\left(\frac{\partial}{\partial \theta} l_n(\theta)\right)$$

若 X_1, \dots, X_n 为 independent 和 identically distributed, 则

$$\begin{aligned} I_n(\theta) &= V_\theta\left(\sum_{i=1}^n s(X_i; \theta)\right) \\ &= \sum_{i=1}^n V_\theta(s(X_i; \theta)) \\ &= nV_\theta(s(X_1; \theta)) \\ &= nI_1(\theta) \\ &:= nI(\theta) \end{aligned}$$

⚠ Remark: Fisher information 的实际意义 ∨

Fisher information 的实际意义是:

Variation of $\frac{\partial l_n(\theta)}{\partial \theta}$ when we observe different samples (对 X_1, \dots, X_n 求 variance)

若 fisher information 较小, 则表示 log-likelihood 在 θ 处的 rate of change 不会随着 X_1, \dots, X_n 的变化而变化很大, 因此就估计 MLE 而言, X_1, \dots, X_n 带来的信息量并不大

⚠ Remark ∨

在特定的 Fisher information regularity condition 下 (见 [STA3020 Lecture 3](#), exponential family 满足该 condition), 有:

1. score function $s(X; \theta)$ 是一个 unbiased estimator of 0, 即

$$\mathbb{E}[s(X; \theta)] = 0$$

此时, 有

$$I_n(\theta) = \mathbb{E}[s(X; \theta)^2]$$

2. Second Bartlett's identity

| 3.3 Theorem: Second Bartlett's identity

若 Fisher information regularity condition 成立,

则 (单个 observation 的) Fisher information 满足:

$$I(\theta) = -\mathbb{E}_\theta \left[\frac{\partial^2}{\partial \theta^2} \log f(X; \theta) \right]$$

⚡ Proof ∨

注意到:

$$\begin{aligned} LHS &= V_\theta\left(\frac{\partial}{\partial \theta} \log f(X; \theta)\right) \\ &= V_\theta\left[\frac{\frac{\partial f(X; \theta)}{\partial \theta}}{f(X; \theta)}\right] \\ &= \mathbb{E}_\theta \left[\left(\frac{\frac{\partial f(X; \theta)}{\partial \theta}}{f(X; \theta)} \right)^2 \right] - \left(\mathbb{E}_\theta \left[\frac{\frac{\partial f(X; \theta)}{\partial \theta}}{f(X; \theta)} \right] \right)^2 \\ &= \mathbb{E}_\theta \left[\left(\frac{\frac{\partial f(X; \theta)}{\partial \theta}}{f(X; \theta)} \right)^2 \right] - \left(\mathbb{E}_\theta \left[\frac{\frac{\partial f(X; \theta)}{\partial \theta}}{f(X; \theta)} \right] \right)^2 = \int \frac{\partial f(X; \theta)}{\partial \theta} \frac{1}{f(X; \theta)} dx = \frac{\partial}{\partial \theta} \left(\int f(X; \theta) \right) = 0 \end{aligned}$$

且:

$$\begin{aligned}
RHS &= -\mathbb{E}_\theta \left[\frac{\partial^2}{\partial \theta^2} \log f(X; \theta) \right] \\
&= -\mathbb{E}_\theta \left[\frac{\partial}{\partial \theta} \left(\frac{\frac{\partial}{\partial \theta} f(X; \theta)}{f(X; \theta)} \right) \right] \\
&= -\mathbb{E}_\theta \left[\frac{\frac{\partial^2 f(X; \theta)}{\partial \theta^2} \cdot f(X; \theta) - \left(\frac{\partial f(X; \theta)}{\partial \theta} \right)^2}{f(X; \theta)^2} \right] \\
&= -\mathbb{E}_\theta \left[\frac{\frac{\partial^2 f(X; \theta)}{\partial \theta^2}}{f(X; \theta)} \right] + \mathbb{E} \left[\left(\frac{\frac{\partial f(X; \theta)}{\partial \theta}}{f(X; \theta)} \right)^2 \right] \\
&= \mathbb{E} \left[\left(\frac{\frac{\partial f(X; \theta)}{\partial \theta}}{f(X; \theta)} \right)^2 \right] \quad (\text{理由同上})
\end{aligned}$$

因此,

$$LHS = RHS$$

Example

问题:

令 $X_1, \dots, X_n \stackrel{i.i.d.}{\sim} \text{Pois}(\lambda)$, 求 $I_n(\lambda)$

解答:

$$\begin{aligned}
f(x) &= \frac{\lambda^x \cdot e^{-\lambda}}{x!} \\
\Rightarrow \log f(x) &= x \cdot \log \lambda - \lambda - \log(x!) \\
\Rightarrow \frac{\partial \log f(x; \lambda)}{\partial \lambda} &= \frac{x}{\lambda} - 1 \\
\Rightarrow \frac{\partial^2 \log f(x; \lambda)}{\partial \lambda^2} &= -\frac{x}{\lambda^2} \\
\Rightarrow I(\lambda) &= -\mathbb{E}_\lambda \left[-\frac{x}{\lambda^2} \right] = \frac{1}{\lambda} \\
\Rightarrow I_n(\lambda) &= nI(\lambda) = \frac{n}{\lambda}
\end{aligned}$$

3.4 Definition: Observed Fisher information

令 X_1, \dots, X_n 为 observed samples, 则

1. (1 个 samples 的) **observed Fisher information** 被定义为:

$$I^{obs}(\theta) = -\frac{1}{n} \frac{\partial^2}{\partial \theta^2} \sum_{i=1}^n \log f(X_i; \theta)$$

2. (n 个 samples 的) **observed Fisher information** 被定义为:

$$I_n^{obs}(\theta) = -\frac{\partial^2}{\partial \theta^2} \sum_{i=1}^n \log f(X_i; \theta) = n \cdot I^{obs}(\theta)$$

Remark: Observed Fisher information 的实际意义

- Observed Fisher information $I_n^{obs}(\theta)$ 衡量了 log-likelihood $l_n(\theta)$ 在 θ 处的 curvature
- 特别的, $I_n^{obs}(\hat{\theta})$ 衡量了 MLE 处的 curvature: $l_n(\theta)$ 在 $\hat{\theta}$ 处越 peaked, likelihood 提供的 information 就越多
- $I(\theta)$ 衡量了该 quantity 的 average value

3.5 Theorem: MLE 的 asymptotic normality

若特定 conditions (MLE 的 CAN conditions, 见 [STA3020 Lecture 7](#)) 满足, 则 MLE $\hat{\theta}_n$ 满足 asymptotic normality:

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

若将 $I(\theta)$ 替换为 $I(\hat{\theta})$, 则 asymptotic normality 仍然成立:

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

可以由此构建 θ 的 approximate $1 - \alpha$ confidence interval

Remark

1. 对于来自 exponential family models 的 i.i.d observations, 其满足上述 conditions
2. 上述 asymptotic normality 中的 $I(\theta)$ 表示单个 sample 时的 Fisher Information (我们不希望 asymptotic distribution 中包含 n)

Proof

- 若 θ_0 是 Θ 的一个 interior, 且 $\hat{\theta}_{MLE}$ 为 θ_0 的 consistent estimator, 则 $\hat{\theta}_{MLE}$ 也是 Θ 的一个 interior (n 足够大时)
 $\Rightarrow l'(\hat{\theta}_{MLE}) = 0$
- 由 Taylor's expansion, 有
 $0 = l'(\hat{\theta}_{MLE}) = l'(\theta_0) + l''(\tilde{\theta})(\theta_0 - \hat{\theta}_{MLE})$ (关键步骤, 在 Delta method 中出现过)
 $(\Rightarrow \sqrt{n}(\hat{\theta}_{MLE} - \theta_0) = \sqrt{n} \cdot \frac{l'(\theta_0)}{l''(\tilde{\theta})} = \frac{\frac{1}{\sqrt{n}} l'(\theta_0)}{\frac{1}{n} l''(\tilde{\theta})}$, 接下来分别研究分子分母的渐近分布即可)
 其中 $\tilde{\theta}$ lies in between θ_0 和 $\hat{\theta}_{MLE}$ ($\tilde{\theta} \xrightarrow{P} \theta_0$)

- 由 CAN condition 的 ② 和 CLT, 有
 $\frac{1}{\sqrt{n}} l'(\theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{\partial \log f(X_i | \theta)}{\partial \theta} \Big|_{\theta=\theta_0} \xrightarrow{d} N(E[S(\theta_0 | X_1)], I(\theta_0)) \stackrel{d}{=} N(0, I(\theta_0))$
 交换微分与积分
- 由 CAN condition 的 ③ 和 WLLN, 有
 $\frac{1}{n} l''(\theta_0) = \frac{1}{n} \sum_{i=1}^n \frac{\partial^2 \log f(X_i | \theta)}{\partial \theta^2} \Big|_{\theta=\theta_0} \xrightarrow{P} E\left[\frac{\partial^2 \log f(X_1 | \theta)}{\partial \theta^2}\right] \Big|_{\theta=\theta_0} = -I(\theta_0)$
 交换微分与积分
- 综上, 由 Slutsky's theorem, 有
 $\sqrt{n}(\hat{\theta}_{MLE} - \theta_0) = \frac{\frac{1}{\sqrt{n}} l'(\theta_0)}{\frac{1}{n} l''(\tilde{\theta})} \xrightarrow{d} N(0, I(\theta_0)^{-1})$

Example

问题:

若 $X_1, \dots, X_n \stackrel{i.i.d.}{\sim} \text{Exp}(\theta)$, 求 θ 的 MLE 和 approximate 95% confidence interval

解答:

$$\begin{aligned}
f(X_1, \dots, X_n; \theta) &= \prod_{i=1}^n (\theta \cdot e^{-\theta X_i}) = \theta^n \cdot \exp \left\{ - \left(\sum_{i=1}^n X_i \right) \theta \right\} \\
\Rightarrow l_n(\theta) &= n \cdot \log(\theta) - \left(\sum_{i=1}^n X_i \right) \theta \\
\Rightarrow \frac{\partial}{\partial \theta} l_n(\theta) &= \frac{n}{\theta} - \sum_{i=1}^n X_i \\
\Rightarrow \frac{\partial^2}{\partial \theta^2} l_n(\theta) &= -\frac{n}{\theta^2} < 0 \\
\Rightarrow \begin{cases} \hat{\theta}_{MLE} = \frac{n}{\sum_{i=1}^n X_i} \\ I_n(\theta) = \mathbb{E} \left[-\frac{\partial^2}{\partial \theta^2} l_n(\theta) \right] = \frac{n}{\theta^2} \end{cases} \\
\Rightarrow \sqrt{n}(\hat{\theta}_n - \theta) &\sim \mathcal{N} \left(0, \frac{1}{I(\theta)} \right) \\
\Rightarrow \hat{\theta}_n &\sim \mathcal{N} \left(\theta, \frac{1}{I_n(\hat{\theta})} \right), \text{ where } I_n(\hat{\theta}) = \frac{n}{\hat{\theta}_n^2} = n\bar{X}^2 \\
\Rightarrow CI &: \frac{n}{\sum_{i=1}^n X_i} \pm z_{0.0025} \cdot \frac{1}{\sqrt{n\bar{X}}}
\end{aligned}$$

3.6 Theorem: MLE 的 asymptotic efficiency

若:

1. 特定 conditions 满足
2. $\tilde{\theta}_n$ 为其他某个 estimator, 满足 $\sqrt{n}(\tilde{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}(0, v(\theta))$

则:

$$v(\theta) \geq \frac{1}{I(\theta)}, \quad \forall \theta$$

4 Fisher Information Matrix 和 Delta Method

Logic ▾

关于 Fisher information matrix 的详细论述, 见 [STA3020 Lecture 5](#)

4.1 Definition: Fisher information matrix

若:

1. Parameter of interest 为 $\theta = (\theta_1, \dots, \theta_k)$
2. log-likelihood 的 Hessian matrix 为:

$$H_{jj} = \frac{\partial^2}{\partial \theta_j^2} l_n(\theta); \quad H_{jk} = \frac{\partial^2}{\partial \theta_j \partial \theta_k} l_n(\theta)$$

则 **Fisher information matrix** 被定义为:

$$I_n(\theta) = - \begin{bmatrix} E_{\theta}(H_{11}) & \cdots & E_{\theta}(H_{1k}) \\ E_{\theta}(H_{21}) & \cdots & E_{\theta}(H_{2k}) \\ \vdots & \vdots & \vdots \\ E_{\theta}(H_{k1}) & \cdots & E_{\theta}(H_{kk}) \end{bmatrix}$$

Example ▾

问题:

若 $X_1, \dots, X_n \stackrel{i.i.d.}{\sim} \mathcal{N}(\mu, \sigma^2)$, 求 $I_n(\mu, \sigma)$

解答:

$$\begin{aligned}
\mathcal{L}_n(\mu, \sigma) &= \left(\frac{1}{\sqrt{2\pi}\sigma} \right)^n \cdot \exp \left\{ -\frac{\sum_{i=1}^n (X_i - \mu)^2}{2\sigma^2} \right\} \\
\Rightarrow l_n(\mu, \sigma) &= \log \left(\frac{1}{\sqrt{2\pi}\sigma} \right)^n - n \log \sigma - \frac{\sum_{i=1}^n (X_i - \mu)^2}{2\sigma^2} \\
\Rightarrow \begin{cases} \frac{\partial}{\partial \mu} l_n(\mu, \sigma) = \frac{\sum_{i=1}^n (X_i - \mu)}{\sigma^2} \\ \frac{\partial}{\partial \sigma} l_n(\mu, \sigma) = -\frac{n}{\sigma} + \frac{\sum_{i=1}^n (X_i - \mu)^2}{\sigma^3} \end{cases} \\
\Rightarrow \begin{cases} \frac{\partial^2}{\partial \mu^2} l_n(\mu, \sigma) = -\frac{n}{\sigma^2} \\ \frac{\partial^2}{\partial \mu \partial \sigma} l_n(\mu, \sigma) = -\frac{2 \sum_{i=1}^n (X_i - \mu)}{\sigma^3} \\ \frac{\partial^2}{\partial \sigma \partial \mu} l_n(\mu, \sigma) = -\frac{2 \sum_{i=1}^n (X_i - \mu)}{\sigma^3} \\ \frac{\partial^2}{\partial \sigma^2} l_n(\mu, \sigma) = \frac{n}{\sigma^2} - \frac{3 \sum_{i=1}^n (X_i - \mu)^2}{\sigma^4} \end{cases} \\
\Rightarrow \begin{cases} \mathbb{E} \left[\frac{\partial^2}{\partial \mu^2} l_n(\mu, \sigma) \right] = -\frac{n}{\sigma^2} \\ \mathbb{E} \left[\frac{\partial^2}{\partial \mu \partial \sigma} l_n(\mu, \sigma) \right] = 0 \\ \mathbb{E} \left[\frac{\partial^2}{\partial \sigma \partial \mu} l_n(\mu, \sigma) \right] = 0 \\ \mathbb{E} \left[\frac{\partial^2}{\partial \sigma^2} l_n(\mu, \sigma) \right] = -\frac{2n}{\sigma^2} \end{cases} \\
\Rightarrow I_n(\mu, \sigma) &= \begin{bmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{2n}{\sigma^2} \end{bmatrix}
\end{aligned}$$

⚠ Remark

Fisher Information 会受到 reparametrization 的影响, 例如

$$I_n(\mu, \sigma^2) = \begin{bmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{bmatrix}$$

更多论述见 [STA3020 Lecture 5](#)

4.2 Multiparameter Delta Method

🔗 Logic

关于单变量 Delta Method 的更多论述, 见 [STA3020 Lecture 2](#), 内容包括:

- First order delta method
- Second order delta method

关于多变量 Delta Method 的更多论述, 见 [STA3020 Lecture 5](#)

令 $\tau = g(\theta_1, \dots, \theta_k)$ 为一个 differentiable function

若:

- $\nabla g = \left(\frac{\partial}{\partial \theta_1} g(\theta), \dots, \frac{\partial}{\partial \theta_k} g(\theta) \right)^T$ 为 g 的 gradient
 - $\hat{\theta}_n$ 处的 ∇g 为 $\hat{\nabla} g$, 且不为 0
 - $J_n(\hat{\theta}_n) = I_n(\hat{\theta}_n)^{-1}$
- 则:

$$\frac{\hat{\tau}_n - \tau}{\hat{s}e(\hat{\tau}_n)} \xrightarrow{d} \mathcal{N}(0, 1)$$

其中

$$\hat{s}e(\hat{\tau}_n) = \sqrt{(\hat{\nabla} g)^T J_n(\hat{\theta}_n) (\hat{\nabla} g)}$$

⚠ Remark ▾

若考虑从 \mathbb{R}^n 到 \mathbb{R}^m 的 mapping, 则需要将 gradient 替换为 Jacobian Matrix

≡ Example ▾

问题:

若 $X_1, \dots, X_n \stackrel{i.i.d.}{\sim} \mathcal{N}(\mu, \sigma^2)$, 求 $\tau = g(\mu, \sigma) = \mu/\sigma$ 的 MLE 和 limiting normal distribution

解答:

在之前的例子中, 我们已经得到了:

$$\begin{cases} \frac{\partial}{\partial \mu} l_n(\mu, \sigma) = \frac{\sum_{i=1}^n (X_i - \mu)}{\sigma^2} \\ \frac{\partial}{\partial \sigma} l_n(\mu, \sigma) = -\frac{n}{\sigma} + \frac{\sum_{i=1}^n (X_i - \mu)^2}{\sigma^3} \end{cases}$$

令其等于 0, 有

$$\begin{cases} \hat{\mu} = \frac{\sum_{i=1}^n X_i}{n} = \bar{X}_n \\ \hat{\sigma} = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X}_n)^2}{n}} \end{cases}$$

根据 MLE 的 equivariance property, 有

$$\hat{\tau} = \frac{\hat{\mu}}{\hat{\sigma}} = \frac{\bar{X}_n}{\sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X}_n)^2}{n}}}$$

在之前的例子中, 我们已经得到了:

$$I_n(\mu, \sigma) = \begin{bmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{2n}{\sigma^2} \end{bmatrix}$$

因此,

$$J_n(\mu, \sigma) = \begin{bmatrix} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{\sigma^2}{2n} \end{bmatrix}$$

由于

$$\nabla g = \left(\frac{\partial g}{\partial \mu}, \frac{\partial g}{\partial \sigma} \right) = \left(\frac{1}{\sigma}, -\frac{\mu}{\sigma^2} \right)$$

因此,

$$\begin{aligned} se(\hat{\tau}_n) &= \sqrt{(\nabla g)^T J_n(\hat{\theta}_n) (\nabla g)} \\ &= \sqrt{\left(\frac{1}{\sigma}, -\frac{\mu}{\sigma^2} \right) \begin{bmatrix} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{\sigma^2}{2n} \end{bmatrix} \left(\frac{1}{\sigma}, -\frac{\mu}{\sigma^2} \right)^T} \\ &= \sqrt{\frac{1}{n} + \frac{\mu^2}{2n\sigma^2}} \\ &= \sqrt{\frac{1}{n} + \frac{\tau^2}{2n}} \end{aligned}$$

因此,

$$\sqrt{n}(\hat{\tau} - \tau) \rightarrow \mathcal{N}\left(0, 1 + \frac{\tau^2}{2}\right)$$