ST2132 Variance of ML Estimators

Semester 1 2022/2023

Large-sample variance of ML estimator

Let $\hat{\theta}_n$ be the ML estimator of $\theta \in \Theta \subset \mathbb{R}$, based on IID RV's $X_1, \ldots, X_n$ with density $f(x|\theta)$.

- In virtually all applications, as $n \to \infty$,

$$\mathrm{var}(\hat{\theta}_n) \approx \frac{\mathcal{I}(\theta)^{-1}}{n}$$

  where the Fisher information $\mathcal{I}(\theta)$ is determined by $f(x|\theta)$.

- For large n, Monte Carlo may not be required to estimate the SE of an ML estimate. Bootstrap is still needed.

- The MOM estimate is given by a formula more often than the ML estimate, but the opposite is true for SE, for large n.

Fisher information: Poisson

$X \sim \text{Poisson}(\lambda)$:

$$f(x) = \frac{\lambda^x e^{-\lambda}}{x!}, x = 0, 1, \ldots$$

$$\log f(X) = X \log \lambda - \lambda - \log X!$$

$$\frac{\mathrm{d} \log f(X)}{\mathrm{d}\lambda} = \qquad\qquad \frac{\mathrm{d}^2 \log f(X)}{\mathrm{d}\lambda^2} =$$

The Fisher information is

$$\mathcal{I}(\lambda) = -\mathrm{E}\left(\frac{\mathrm{d}^2 \log f(X)}{\mathrm{d}\lambda^2}\right) =$$

Fisher information: Bernoulli

$X \sim \text{Bernoulli}(p)$:

$$f(x) = p^x(1-p)^{(1-x)}, \qquad x = 0, 1$$

$$\log f(X) =$$

$$\frac{\mathrm{d} \log f(X)}{\mathrm{d}p} = \qquad\qquad \frac{\mathrm{d}^2 \log f(X)}{\mathrm{d}p^2} =$$

The Fisher information is

$$\mathcal{I}(p) = -\mathrm{E}\left(\frac{\mathrm{d}^2 \log f(X)}{\mathrm{d}p^2}\right) =$$

Fisher information: normal

$X \sim \mathrm{N}(\mu, \sigma^2)$, $\theta = (\mu, \sigma)$.

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}, \qquad -\infty < x < \infty$$

$$\log f(X) = -\frac{\log 2\pi}{2} - \log\sigma - \frac{(X-\mu)^2}{2\sigma^2}$$

$$\frac{\mathrm{d}\log f(X)}{\mathrm{d}\theta} = \qquad\qquad \frac{\mathrm{d}^2 \log f(X)}{\mathrm{d}\theta^2} =$$

$$\mathcal{I}(\theta) =$$

Definition: Fisher information

- Let X have density $f(x|\theta)$, $\theta \in \Theta \subset \mathbb{R}^p$. The Fisher information is the $p \times p$ matrix

$$\mathcal{I}(\theta) = -\mathrm{E}\left[\frac{\mathrm{d}^2 \log f(X)}{\mathrm{d}\theta^2}\right]$$

- $\mathcal{I}(\theta)$ is symmetric, with (i, j)-entry

$$-\mathrm{E}\left[\frac{\partial^2 \log f(X)}{\partial\theta_i \partial\theta_j}\right]$$

which is

$$-\int_{-\infty}^{\infty} \frac{\partial^2 \log f(x)}{\partial\theta_i \partial\theta_j} f(x)\,\mathrm{d}x, \qquad \sum_x \frac{\partial^2 \log f(x)}{\partial\theta_i \partial\theta_j} f(x)$$

according to whether X is continuous or discrete.

Interpretation

- $\mathcal{I}(\theta)$ measures the information about θ in one sample $X \sim f(x|\theta)$.
- Let $X \sim \text{Poisson}(\lambda)$. The larger λ, the less information in X.
- For which value of p does a Bernoulli(p) sample have the least information?
- Guess: How much information about p is in n IID Bernoulli(p) samples?

n IID samples

- IID $X_1, \ldots, X_n$ with density $f(x|\theta)$ can be regarded as a sample from $\mathbf{X} = (X_1, \ldots, X_n)$ with random joint density

$$g(\mathbf{X}|\theta) = f(X_1|\theta) \cdots f(X_n|\theta)$$

- The information in $\mathbf{X}$,

$$-\mathrm{E}\left[\frac{\mathrm{d}^2 \log g(\mathbf{X})}{\mathrm{d}\theta^2}\right]$$

is $n\mathcal{I}(\theta)$, where $\mathcal{I}(\theta)$ is the information in any one of the X's.

Binomial

$X \sim \text{Binomial}(n, p)$: $f(x) = \binom{n}{x} p^x (1-p)^{n-x}, x = 0, 1, \ldots, n.$

$$\log f(X) = \log \binom{n}{X} + X \log p + (n - X) \log(1 - p)$$

$$\frac{\mathrm{d} \log f}{\mathrm{d} p} =$$

$$\frac{\mathrm{d}^2 \log f}{\mathrm{d} p^2} =$$

$$\mathcal{I}(p) =$$

Bernoulli vs binomial

- One binomial(n, p) sample has the same information about p as n IID Bernoulli(p) samples.
- Surprising? The binomial sample only tells us the number of successes, $\sum_{i=1}^{n} X_i$, while the n Bernoulli samples tell us the full sequence of successes and failures: $X_1, \dots, X_n$.
- Unsurprising? The ML estimators of p are identical.
- Similarly, the multinomial$(n, (p_1, \dots, p_r))$ has the same information about $(p_1, \dots, p_r)$ as n ______________________ samples.

HWE trinomial distribution

$\mathbf{X} = (X_1, X_2, X_3) \sim$ multinomial(n,$\mathbf{p}$), where

$$p_1 = (1-\theta)^2, p_2 = 2\theta(1-\theta), p_3 = \theta^2, \qquad 0 < \theta < 1$$

$$\mathcal{I}(\theta) = \frac{2n}{\theta(1-\theta)}$$

Name two other sets of samples (number of IID samples, and the distribution) that have the same information.

Multinomial data

- $\mathbf{X} \sim \text{multinomial}(n, (p_1, \ldots, p_r))$, $\theta = (p_1, \ldots, p_{r-1})$.

$$\log f(\mathbf{X}) = \sum_{i=1}^{r} X_i \log p_i$$

$$\begin{aligned}
\frac{\partial \log f(X)}{\partial p_i} &= && 1 \leq i \leq r-1 \\
\frac{\partial^2 \log f(X)}{\partial p_i^2} &= && 1 \leq i \leq r-1 \\
\frac{\partial^2 \log f(X)}{\partial p_i \partial p_j} &= && 1 \leq i \neq j \leq r-1
\end{aligned}$$

- (i, j)-entry of $\mathcal{I}(\theta)$:

$$\begin{array}{ll} \frac{n}{p_i} + \frac{n}{p_r}, & i = j \\ \frac{n}{p_r}, & i \neq j \end{array}$$

Multinomial parameterisation

- The general trinomial distribution has two parameters: p_1, p_2, or any other set of two probabilities. The HWE trinomial is defined by a single parameter.
- They are opposite ends of a collection of multinomial distributions with the same n and r, but of varying number of independent parameters.
- Any such distribution can be described as

$$\text{Multinomial}(n,\mathbf{p}(\theta))$$

$$\mathbf{p}(\theta) = (p_1(\theta), \ldots, p_r(\theta))$$

$$\theta \in \Theta \subset \mathbb{R}^k, 1 \leq k \leq r-1.$$

Summary: Fisher information

- Let X have mass/density $f(x|\theta)$, $\theta \in \Theta \subset \mathbb{R}^p$. The Fisher information at θ in X is the $p \times p$ matrix

$$-\mathrm{E}\left[\frac{\mathrm{d}^2 \log f(X)}{\mathrm{d}\theta^2}\right]$$

  X can be a random vector.

- Information in n IID samples is n times that in one sample.

- A binomial(n, p) sample has the same information as n IID Bernoulli(p) samples. Similarly, a multinomial$(n,\mathbf{p})$ has the same information as n IID multinomial$(1,\mathbf{p})$ samples.

Variance of ML estimator

- $X_1, \ldots, X_n$ IID Poisson(λ). What is $\mathcal{I}(\lambda)$, the information in any X_i? What is the variance of the ML estimator $\hat{\lambda} = \bar{X}$?

- $X_1, \ldots, X_n$ IID Bermoulli(p). What is $\mathcal{I}(p)$, the information in any X_i? What is the variance of the ML estimator $\hat{p} = \bar{X}$?

- What is your conjecture?

Variance of ML estimator: normal

- $X_1, \ldots, X_n$ are IID normal(μ, ν). Let $\theta = (\mu, \nu)$. What is $\mathcal{I}(\theta)$, the information in any X_i?

- What is the variance of the ML estimator $\hat{\theta} = (\bar{X}, \hat{\sigma}^2)$?

- Revise conjecture?

Gamma distribution

$X \sim \text{Gamma}(\alpha, \lambda)$, $\theta = (\alpha, \lambda)$.

$$\log f(X) = \alpha \log \lambda + (\alpha - 1) \log X - \lambda X - \log \Gamma(\alpha)$$

$$\frac{\partial \log f(X)}{\partial \alpha} = \log \lambda + \log X - \psi(\alpha), \qquad \frac{\partial \log f(X)}{\partial \lambda} = \frac{\alpha}{\lambda} - X$$

where $\psi(\alpha) = \frac{\mathrm{d}}{\mathrm{d}\alpha} \log \Gamma(\alpha) = \frac{\Gamma'(\alpha)}{\Gamma(\alpha)}$ is the digamma function.

$$\frac{\partial^2 \log f(X)}{\partial \alpha^2} = -\psi'(\alpha), \qquad \frac{\partial^2 \log f(X)}{\partial \lambda^2} = -\frac{\alpha}{\lambda^2}$$

$$\frac{\partial^2 \log f(X)}{\partial \alpha \partial \lambda} = \frac{\partial^2 \log f(X)}{\partial \lambda \partial \alpha} = \frac{1}{\lambda}$$

$$\mathcal{I}(\theta) = -\mathrm{E}\left[\frac{\mathrm{d}^2 \log f(X)}{\mathrm{d}\theta^2}\right] = \begin{bmatrix} \psi'(\alpha) & -\frac{1}{\lambda} \\ -\frac{1}{\lambda} & \frac{\alpha}{\lambda^2} \end{bmatrix}$$

Revisiting rainfall again

- Based on 227 realisations, the ML estimate of $\theta = (\alpha, \lambda)$ is (0.44,1.96). Let $\hat{\alpha}$ and $\hat{\lambda}$ be the ML estimators. Assuming 227 is large enough,

$$\text{var}(\hat{\alpha}, \hat{\lambda}) \approx \frac{\mathcal{I}(\alpha, \lambda)^{-1}}{227}$$

- Bootstrap approximation:

$$\text{var}(\hat{\alpha}, \hat{\lambda}) \approx \frac{\mathcal{I}(0.44, 1.96)^{-1}}{227} \approx \begin{bmatrix} 0.0011 & 0.0051 \\ 0.0051 & 0.0610 \end{bmatrix}$$

- SE in 0.44 is $\text{SD}(\hat{\alpha}) \approx \sqrt{0.0011} \approx 0.03$.
  SE in 1.96 is $\text{SD}(\hat{\lambda}) \approx \sqrt{0.0610} \approx 0.25$.
  They are very close to the Monte Carlo approximations 0.03 and 0.26 (Parameter Estimation II slide 20).

Conclusion

- $X_1, \ldots, X_n$ IIR RV's with density $f(x|\theta)$. ML estimator is $\hat{\theta}$. As $n \to \infty$,

$$\text{var}(\hat{\theta}) \approx \frac{\mathcal{I}(\theta)^{-1}}{n}$$

where $\mathcal{I}(\theta)$ is the Fisher information in any X_i (one sample).

- For large n, SE in ML estimate can be approximated without Monte Carlo.

- Some technical conditions are required, but they are true in almost all applications. For technical details, see Rice's book.

Special case: multinomial

The previous result applies to multinomial data, provided the Fisher information is suitably defined.

$\mathbf{X} \sim \text{Multinomial}(n,\mathbf{p}(\theta)), \theta \in \Theta \subset \mathbb{R}^k$ with $1 \leq k \leq r-1$. For large n,

$$\text{var}(\hat{\theta}) \approx \frac{\mathcal{I}(\theta)^{-1}}{n}$$

where $\mathcal{I}(\theta)$ is the information in Multinomial$(1,\mathbf{p}(\theta))$.

- $X \sim \text{Binomial}(n,p)$, $\text{var}(\hat{p}) = \frac{p(1-p)}{n}$, $\frac{1}{p(1-p)}$ the information in Binomial$(1,p)$ = Bernoulli(p).
- HWE trinomial: $\text{var}(\hat{\theta}) = \frac{\theta(1-\theta)}{2n}$, $\frac{1}{\theta(1-\theta)}$ the information in Trinomial$(1,\mathbf{p}(\theta))$.