ST2132 Distribution of ML Estimators

Semester 1 2022/2023

Main Result

$\hat{\theta}_n$: ML estimator of $\theta \in \Theta \subset \mathbb{R}^p$, based on either

1. IID RV's $X_1, \ldots, X_n$ with density $f(x|\theta)$. $\mathcal{I}(\theta)$: Fisher information in any X_i.

   or
2. $(X_1, \ldots, X_r) \sim$ Multinomial$(n, \mathbf{p}(\theta))$. $\mathcal{I}(\theta)$: Fisher information in Multinomial$(1,\mathbf{p}(\theta))$.

As $n \to \infty$, the distribution of

$$\sqrt{n\mathcal{I}(\theta)}(\underline{\hat{\theta}_n - \theta})$$

converges to N$(\mathbf{0},\mathbf{I}_p)$.

$$\boxed{\underset{p\times 1}{\hat{\theta}_n} \sim N\left(\underset{p\times 1}{\theta}, \frac{\underset{p\times p}{\mathcal{I}(\theta)^{-1}}}{n}\right)}$$

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

Required technical conditions hold in almost all applications.

Consequences

- For large n, approximately

$$\hat{\theta}_n \sim \mathrm{N}\left(\theta, \frac{\mathcal{I}(\theta)^{-1}}{n}\right)$$

- ML estimators are asymptotically unbiased, and consistent: $\hat{\theta}_n \to \theta$.
- Approximate CIs for θ can be constructed.

Poisson and Bernoulli

- $X_1, \ldots, X_n$ IID Poisson(λ). $\hat{\lambda} = \bar{X}$. $\mathcal{I}(\lambda) = 1/\lambda$. For large n, approximately

$$\hat{\lambda} \sim \mathrm{N}\left(\lambda, \frac{\lambda}{n}\right)$$

- $X_1, \ldots, X_n$ IID Bernoulli(p). $\hat{p} = \bar{X}$. $\mathcal{I}(p) = 1/p(1-p)$. For large n, approximately

$$\hat{p} \sim \mathrm{N}\left(p, \frac{p(1-p)}{n}\right)$$

- These also follow directly from CLT.

Normal distribution

- $X_1, \ldots, X_n$ IID $\mathrm{N}(\mu, \sigma^2)$.

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

$\mathrm{var}(\hat{\sigma}^2)$

- Variance of ML Estimators slide 5:

$$\mathcal{I}(\mu, \sigma) = \begin{bmatrix} 1/\sigma^2 & 0 \\ 0 & 2/\sigma^2 \end{bmatrix}$$

$\mathrm{var}(\hat{\sigma}) \approx \frac{\sigma^2}{2n}$

- For large n, approximately

$$\begin{bmatrix} \hat{\mu} \\ \hat{\sigma} \end{bmatrix} \sim \mathrm{N}\left(\begin{bmatrix} \mu \\ \sigma \end{bmatrix}, \begin{bmatrix} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{\sigma^2}{2n} \end{bmatrix} \right)$$

Distribution of $\hat{\mu}$ and independence are exact.

HWE trinomial

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

$$p_1 = (1-\theta)^2, \quad p_2 = 2\theta(1-\theta), \quad p_3 = \theta^2$$

$$\hat{\theta} = \frac{X_2 + 2X_3}{2n} \sim \text{Binomial}(2n, \theta)$$

- The information in a Trinomial(1,$\mathbf{p}$) distribution is

$$\mathcal{I}(\theta) = \frac{2}{\theta(1-\theta)} \quad \text{from Bin}(2, \theta)$$

- For large n, approximately,

$$\hat{\theta} \sim \mathrm{N}\left(\theta, \frac{\theta(1-\theta)}{2n}\right)$$

Also follows directly from CLT.

Trinomial distribution

- $\mathbf{X} \sim \text{Trinomial}(n, (p_1, p_2, p_3))$. Let $\theta = (p_1, p_2)$.

$$\hat{p}_i = \frac{X_i}{n} \qquad i=1,2,3$$

- The information in a $\text{Trinomial}(1, (p_1, p_2, p_3))$ distribution is

$$\mathcal{I}(p_1, p_2) = \begin{bmatrix} \frac{1}{p_1} + \frac{1}{p_3} & \frac{1}{p_3} \\ \frac{1}{p_3} & \frac{1}{p_2} + \frac{1}{p_3} \end{bmatrix}$$

- For large n, approximately

$$\begin{bmatrix} \hat{p}_1 \\ \hat{p}_2 \end{bmatrix} \sim \mathrm{N}\left(\begin{bmatrix} p_1 \\ p_2 \end{bmatrix}, \frac{1}{n} \begin{bmatrix} p_1(1-p_1) & -p_1 p_2 \\ -p_1 p_2 & p_2(1-p_2) \end{bmatrix} \right)$$

$$\begin{bmatrix} \hat{p}_1 \\ \hat{p}_2 \\ \hat{p}_3 \end{bmatrix} = \hat{p}$$

implying that $\hat{p}$ is also approximately normal.

Gamma distribution

- $X_1, \ldots, X_n$ IID Gamma(α, λ). The ML estimators $\hat{\alpha}$ and $\hat{\lambda}$ cannot be expressed algebraically.

- The Fisher information is

$$\mathcal{I}(\alpha, \lambda) = \begin{bmatrix} \psi'(\alpha) & -\frac{1}{\lambda} \\ -\frac{1}{\lambda} & \frac{\alpha}{\lambda^2} \end{bmatrix}$$

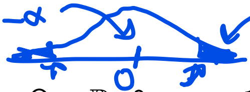
where $\psi(\alpha)$ is the digamma function.

- For large n, approximately

$$\begin{bmatrix} \hat{\alpha} \\ \hat{\lambda} \end{bmatrix} \sim \mathrm{N}\left(\begin{bmatrix} \alpha \\ \lambda \end{bmatrix}, \frac{\mathcal{I}(\alpha, \lambda)^{-1}}{n} \right)$$

Normal approximation for ML estimator

$\Pr(Z \geq z_{\frac{\alpha}{2}}) = \frac{\alpha}{2}$



$\frac{\alpha}{2}$

$\hat{\theta}_n$: ML estimator of $\theta \in \Theta \subset \mathbb{R}$. $0 < \alpha < 1$.

$I(\theta)$ is a scalar

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

$\frac{X-\mu}{\sigma}$

- For large n,

$$1 - \alpha \approx \Pr\left(-z_{\frac{\alpha}{2}} \leq \frac{\hat{\theta}_n - \theta}{\sqrt{\mathcal{I}(\theta)^{-1}/n}} \leq z_{\frac{\alpha}{2}}\right)$$

- Hence

$$1 - \alpha \approx \Pr\left(\hat{\theta}_n - z_{\frac{\alpha}{2}}\sqrt{\frac{\mathcal{I}(\theta)^{-1}}{n}} \leq \theta \leq \hat{\theta}_n + z_{\frac{\alpha}{2}}\sqrt{\frac{\mathcal{I}(\theta)^{-1}}{n}}\right)$$

Confidence interval

- For large n, the random interval

$$\left(\hat{\theta}_n - z_{\frac{\alpha}{2}}\sqrt{\frac{\mathcal{I}(\theta)^{-1}}{n}}, \hat{\theta}_n + z_{\frac{\alpha}{2}}\sqrt{\frac{\mathcal{I}(\theta)^{-1}}{n}}\right)$$

covers θ with probability of about $1-\alpha$.

- Data give the ML **estimate** of θ. ← realisation of $\hat{\theta}_n$

SE is approximated by bootstrap: replacing θ by its ML estimate in $\sqrt{\frac{\mathcal{I}(\theta)^{-1}}{n}}$.

Then

$$(\text{estimate} - z_{\frac{\alpha}{2}}\,\text{SE}, \text{estimate} + z_{\frac{\alpha}{2}}\,\text{SE})$$

is an approximate $(1-\alpha)$-CI for θ.

Poisson

ML estimate of λ is $\bar{x}$. $\mathcal{I}(\lambda)^{-1} = \lambda$.

Bootstrap approximation:

$$\mathrm{SE} = \sqrt{\frac{\lambda}{n}} \approx \sqrt{\frac{\bar{x}}{n}}$$

For large n, an approximate $(1-\alpha)$-CI for λ is

$$\left(\bar{x} - z_{\alpha/2}\sqrt{\frac{\bar{x}}{n}}, \bar{x} + z_{\alpha/2}\sqrt{\frac{\bar{x}}{n}}\right)$$

Used in Parameter Estimation I slide 7.

Normal distribution

$x_1, \ldots, x_n$ realisations of IID $\mathrm{N}(\mu, \sigma^2)$ RV's, n large. ML estimates of μ and σ are $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$ and $\hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2}$.

$\hat{\sigma}$

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

  $: \mathrm{var}(\bar{X})$

  $\approx \mathrm{var}(\hat{\sigma})$

  ML estimator RV.

  SEs of $\bar{x}$ and $\hat{\sigma}$ estimated as $\hat{\sigma}/\sqrt{n}$ and $\hat{\sigma}/\sqrt{2n}$.

  bootstrap

- Approximate $(1-\alpha)$-CI:

$$\begin{aligned} \mu &: \left(\bar{x} - z_{\frac{\alpha}{2}}\frac{\hat{\sigma}}{\sqrt{n}}, \bar{x} + z_{\frac{\alpha}{2}}\frac{\hat{\sigma}}{\sqrt{n}}\right) \\ \sigma &: \left(\hat{\sigma} - z_{\frac{\alpha}{2}}\frac{\hat{\sigma}}{\sqrt{2n}}, \hat{\sigma} + z_{\frac{\alpha}{2}}\frac{\hat{\sigma}}{\sqrt{2n}}\right) \end{aligned}$$

  s is not used. No big deal, since n is large.

Scope of asymptotic normality of ML estimators

- Given IID normal RV's, let $\hat{\sigma}$ be the ML estimator of σ, so $\hat{\sigma}^2$ is the ML estimator of σ^2.

  Both $\hat{\sigma}$ and $\hat{\sigma}^2$ are asymptotically normal, though for a given n, one will likely be closer to normal than the other.

- More generally, let $\hat{\theta}$ be the ML estimator of θ. For any $h : \Theta \to \mathbb{R}$, $h(\hat{\theta})$ is the ML estimator of $h(\theta)$. For large n, $h(\hat{\theta})$ is approximately normal.

  strictly ↑ or strictly ↓

- In the normal case, let $h(x) = 1/x$. Then $1/\hat{\sigma}$ is also asymptotically normal.

  $\sigma \in \mathbb{R}_+$

Another revisit to rainfall data

- ML estimates of α and λ are 0.44 and 1.96. Estimated SEs are 0.03 and 0.25.

- Assuming $n = 227$ is large enough, approximate 95%-CI:

$$\begin{aligned}\alpha &: \quad 0.44 \pm 1.96 \times 0.03 \approx (0.38, 0.50) \\ \lambda &: \quad 1.96 \pm 1.96 \times 0.25 \approx (1.47, 2.45)\end{aligned}$$

  Parameter Estimation II slide 20: bias in 1.96 is about 0.04. Bias-corrected 95%-CI for λ : $(1.43, 2.41)$.

Multinomial distribution (1)

$\mathbf{X} \sim \text{Multinomial}(n, (p_1, \ldots, p_r))$. ML estimator $\hat{\mathbf{p}} = \mathbf{X}/n$.

$\theta = (p_1, \ldots, p_{r-1})$. $\mathcal{I}(\theta)$ is the information in a Multinomial(1,$\mathbf{p}$) distribution, given on Variance of ML Estimators slide 12.

For large n, approximately,

$(r-1) \times (r-1)$

$r = 3$

$$\theta = \begin{bmatrix} p_1 \\ p_2 \end{bmatrix} \qquad \hat{\theta} \sim \mathrm{N}\left(\theta, \frac{\mathcal{I}(\theta)^{-1}}{n}\right) \qquad \begin{bmatrix} p_1 \\ \vdots \\ p_{r-1} \end{bmatrix}$$

$\text{var}(\hat{\theta}) = \frac{\mathcal{I}(\theta)^{-1}}{n}$, with (i, j)-entry:

$$= \begin{bmatrix} p_1(1-p_1) & -p_1p_2 \\ -p_2p_1 & p_2(1-p_2) \end{bmatrix}$$

$$\begin{cases} \dfrac{p_i(1-p_i)}{n}, & i = j \\ \cancel{\dfrac{p_ip_j}{n}}, & i \neq j \end{cases}$$

$-\dfrac{p_i p_j}{n}$

Multinomial distribution (2)

$$\begin{bmatrix} p_1 & & 0 \\ & \ddots & \\ 0 & & p_r \end{bmatrix} \qquad pp' = \begin{bmatrix} p_1^2 & p_1p_2 & \cdots & p_1p_r \\ p_rp_1 & p_r^2 & \cdots & \end{bmatrix}$$

- Distribution of $\hat{\theta}$ implies $\hat{\mathbf{p}}$ also has an approximate normal distribution, with expectation $\mathbf{p}$ and variance

columns add to 0

$r \times r$

$$\mathrm{var}(\hat{\mathbf{p}}) = \frac{1}{n}(\mathrm{diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}')$$

- $\mathrm{var}(\hat{\mathbf{p}})$ has $\mathrm{var}(\hat{\theta})$ at its top left. The additional entries are such that each row and column of $\mathrm{var}(\hat{\mathbf{p}})$ sums to 0. What is the rank of $\mathrm{var}(\hat{\mathbf{p}})$? $r-1$

- Large-sample CI for p_i can be constructed, and looks like one based on the binomial distribution.

Conclusion: ML vs MOM

- Both MOM amd ML estimators are consistent: bias goes to 0 as $n \to \infty$.
- MOM uses only sample moments to estimate parameter. ML uses all information contained in the density function. Hence ML estimates tend to have smaller bias and SE.
- The asymptotic properties of ML estimators are powerful and important. For large n, the SE can be estimated without Monte Carlo, and a good CI for the parameter is available.
- MOM estimators may not be asymptotically normal, so it is more difficult to construct a CI. However, it is easier to compute, so is sometimes useful.

Conclusion: Population mean vs parameter

SRS of size n from a large population with mean μ and variance σ^2. $\hat{\mu} = \bar{X}$.

$\hat{\theta}_n$ ML estimator of θ based on n IID RV's or a multinomial RV with n trials.

Below, the approximation is better for larger n.

Estimator	E	var	*Distribution*
$\hat{\mu}$	μ	σ^2/n	$\approx$ Normal
$\hat{\theta}_n$	$\approx \theta$	$\approx \mathcal{I}(\theta)^{-1}/n$	$\approx$ Normal

How large should n be for $\hat{\theta}_n$ to be normally distributed? Generally never. Monte Carlo can be used to check how close it is to normal.

ML in other models (1)

- ML estimation works in other statistical models, such as when the random variables are independent but not identically distributed, and beyond. Details in future modules.

- **Multiple Regression**. Suppose that

$$Y = \underset{n \times p}{X} \underset{p \times 1}{\beta} + \epsilon$$

X: $n \times p$ matrix of known constants,
β: $p \times 1$ vector of unknown constants,
ϵ: $n \times 1$ random vector, with IID N(0,σ^2) components.

- Given realisation $\underset{n \times 1}{\underline{y}}$, how to estimate β and σ^2 by ML?

ML in other models (2)

- For $0 < p < 1$, the log odds is $\log \frac{p}{1-p}$.
- **Logistic Regression**. $Y_i \sim \text{Bernoulli}(p_i)$ are independent for $i = 1, \ldots, n$. Let θ be vector of log odds: $\theta_i = \log \frac{p_i}{1-p_i}$.
  Suppose that
  $$\theta = X\beta$$
  X: $n \times p$ matrix of known constants.
  β: $p \times 1$ vector of unknown constants.
- Given realisations $y_1, \ldots, y_n$, how to estimate β by ML?
- Markov chain, time series, etc.