Statistical Inference

Yuki Shiraito

POLSCI 599 Statistical Methods in Political Research I
University of Michigan

Statistical Inference: Overview

- Statistical model:
  1. Assumption about the world, F_X
  2. Data, $(X_1, \ldots, X_n)$, form a random sample from F_X
- *Estimand*, what we want to know about F_X:
  1. Population moments, e.g., $\mathbb{E}[X], \mathbb{V}(X)$
  2. Parameters of distribution, θ if F_X is written as $F_X(x; \theta)$
- *Estimation*:
  - Define an estimator or statistic, $T_n = r(X_1, \ldots, X_n)$
- Sampling distribution:
  - *Theoretical* distribution of T_n across samples
  - Only one realization of T_n in one sample
  - Theoretical because it depends on F_X (and θ)

- Exact inference:
  - Given sample size n
  - Sampling distribution of T_n derived from F_X
- Approximate inference:
  - Asymptotics: Convergence as $n \to \infty$
  - Sampling distribution of $\lim_{n\to\infty} T_n$ approximated via LLN and CLT

Method of Moments Estimator

- Method of moments estimator: Let θ be a vector of k estimands and suppose that the kth moment of F_X is written as a function of θ, $\mathbb{E}[X^k] = \eta_k(\theta)$. The *method of moments (MM) estimator* $\widehat{\theta}_{MM}$ is the solution for θ to the system of equations

$$\begin{aligned} \eta_1(\theta) &= M_1, \\ &\vdots \\ \eta_k(\theta) &= M_k. \end{aligned}$$

- MM estimator of the population mean μ and variance σ^2:
  1. $\widehat{\mu}_n = \frac{1}{n}\sum_{i=1}^n X_i$
  2. $\widehat{\sigma^2}_n = \frac{1}{n}\sum_{i=1}^n X_i^2 - \left(\frac{1}{n}\sum_{j=1}^n X_j\right)^2 = \frac{1}{n}\sum_{i=1}^n \left(X_i - \frac{1}{n}\sum_{j=1}^n X_j\right)^2$
- Intuitive: Replace population moments with sample moments
- Simple: Not necessarily require assumptions on distribution F_X
- What if more equations than estimands?
  - E.g., Poisson distribution: $\lambda = \mathbb{E}[X] = \mathbb{V}(X)$
  - Incorporate p.(d.)f. into estimator $\rightsquigarrow$ MLE
  - Finding the "best" value $\rightsquigarrow$ GMM

Exact Inference on MM Estimator

- We know the mean and variance of $\widehat{\mu}_n$:
  - $\mathbb{E}[\widehat{\mu}_n] = \mu, \mathbb{V}(\widehat{\mu}_n) = \sigma^2/n$
  - For any n, without any parametric assumptions
- $\widehat{\mu}_n$ is unbiased: $\mathbb{E}[\widehat{\mu}_n] = \mu$
- Is $\widehat{\sigma^2}_n$ unbiased?

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 = \widehat{\sigma^2}_n + (\widehat{\mu}_n - \mu)^2 \Leftrightarrow \sigma^2 = \mathbb{E}\left[\widehat{\sigma^2}_n\right] + \frac{\sigma^2}{n}$$

- $\widehat{\sigma^2}_n$ is biased: $\mathbb{E}\left[\widehat{\sigma^2}_n\right] = \frac{n-1}{n}\sigma^2$
- Unbiased variance:
  - $s_n^2 \equiv \frac{n}{n-1}\widehat{\sigma^2}_n = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \widehat{\mu}_n)$
  - $\mathbb{E}[s_n^2] = \frac{n}{n-1}\mathbb{E}[\widehat{\sigma^2}_n] = \sigma^2$
  - Again for any n, without any parametric assumptions
- Can we find the sampling distribution of σ_n^2 or s_n^2?

Exact Sampling Distribution of MM Estimator

- We need parametric assumptions for further exact inference:
  - Well known if F_X is Gaussian $\rightsquigarrow$ t-statistic
  - F_X is Bernoulli
- $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2) \Rightarrow \widehat{\mu}_n \sim \mathcal{N}(\mu, \sigma^2/n)$
- Sampling distribution of sample/unbiased variance:

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \widehat{\mu}_n)^2 = \frac{n-1}{\sigma^2}s_n^2 = \frac{n}{\sigma^2}\widehat{\sigma^2}_n \sim \chi^2_{n-1}$$

  - χ^2_{n-1} is the *Chi-squared distribution with degrees of freedom* $n-1$
  - $\chi^2_{n-1} = \text{Gamma}\left(\frac{n-1}{2}, \frac{1}{2}\right)$
  - Sum of the squares of $n-1$ independent Gaussian r.v.s

Proof. Assume $\mu = 0$ since $X_i + \mu - (\widehat{\mu}_n + \mu)$ for any μ

1. $\frac{1}{\sigma^2}\sum_{i=1}^{n} X_i^2 = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \widehat{\mu}_n)^2 + \frac{n}{\sigma^2}\widehat{\mu}_n^2$
2. If $X_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$, $\sum_{i=1}^{n}(X_i - \widehat{\mu}_n)^2$ and $\widehat{\mu}_n^2$ are independent
3. We know $\frac{1}{\sigma^2}\sum_{i=1}^{n} X_i^2 \sim \chi^2_n$ and $\frac{n}{\sigma^2}\widehat{\mu}_n^2 \sim \chi^2_1$
4. We use the m.g.f.s of these to get the m.g.f. of $\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \widehat{\mu}_n)^2$

t-statistic

- Now we know the sampling distributions of $\widehat{\mu}_n$ and $\frac{n-1}{\sigma^2}s_n^2$
- Problem: Two unknown parameters, μ and σ^2
- t-statistic: For a fixed number θ, the *t-statistic* is defined as

$$\mathcal{T}_n(\theta) \equiv \frac{\sqrt{n}(\widehat{\mu}_n - \theta)}{\sqrt{s_n^2}}$$

- Student's t-distribution: Let $Z \sim \mathcal{N}(0,1)$ and $V \sim \chi^2_{n-1}$. Then, $\frac{Z}{\sqrt{V/(n-1)}}$ follows the *t-distribution with* $n-1$ *degrees of freedom*.
- If $\theta = \mu$, $\mathcal{T}_n(\theta)$ follows Student's t-distribution:

$$\frac{\overbrace{\sqrt{\frac{n}{\sigma^2}}(\widehat{\mu}_n - \mu)}^{\sim\mathcal{N}(0,1)}}{\left(\underbrace{\frac{n-1}{\sigma^2}s_n^2}_{\sim\chi^2_{n-1}}/(n-1)\right)^{\frac{1}{2}}} = \mathcal{T}_n(\mu)$$

Hypothesis Testing: The t-test

- Assuming $\theta = \mu$, we know how likely a value of $\mathcal{T}_n(\theta)$ is
- Hypothesis Testing:
  1. Have a hypothesis that $\mu = \theta_0$ (*null hypothesis*)
  2. Compute $\mathcal{T}_n(\theta_0)$
  3. Is the value of $\mathcal{T}_n(\theta_0)$ consistent with the null?
- What "consistent" means:
  - The value of $\mathcal{T}_n(\theta_0)$ is "not unlikely" under the null
  - The sampling distribution $\rightsquigarrow$ how likely a value is
  - Range from the $\frac{\alpha}{2}$ quantile to the $1 - \frac{\alpha}{2}$ quantile is not unlikely
- Testing procedure:
  - Reject (accept) the null if $\mathcal{T}_n(\theta_0) \notin (\in) [t^*_{n-1,\frac{\alpha}{2}}, t^*_{n-1,1-\frac{\alpha}{2}}]$
  - $t^*_{n-1,\delta}$: the δ quantile of the t-distribution
- $T_n(\theta)$ is an r.v. $\rightsquigarrow$ error is always possible
  - Type I error (false positive): Reject the null when $\mu = \theta_0$
  - Type II error (false negative): Accept the null when $\mu \neq \theta_0$
- Probability of Type I error across samples is α
- Multiple testing problem:
  - If you run testing many times, you reject the null in some tests

Interval Estimation

- We know $\widehat{\mu}_n$ (estimator) is not exactly equal to μ (estimand)
- Instead of one value, use an interval to account for randomness
- Interval estimation:
  - Get the inverse of the acceptance region of the t-test

$$\mathcal{T}_n(\theta_0) \lesseqgtr t^*_{n-1,\delta} \Leftrightarrow \theta_0 \gtreqless \widehat{\mu}_n - \frac{\sqrt{s_n^2}}{\sqrt{n}} t^*_{n-1,\delta} \Leftrightarrow \theta_0 \gtreqless \widehat{\mu}_n + \frac{\sqrt{s_n^2}}{\sqrt{n}} t^*_{n-1,1-\delta}$$

  - $\left[\widehat{\mu}_n + \frac{\sqrt{s_n^2}}{\sqrt{n}} t^*_{n-1,\frac{\alpha}{2}},\ \widehat{\mu}_n + \frac{\sqrt{s_n^2}}{\sqrt{n}} t^*_{n-1,1-\frac{\alpha}{2}}\right]$ is the confidence interval

- *Confidence intervals are random intervals*
  - Bounds have sampling distributions
  - Intervals vary across samples
  - $\mathbb{P}\left(\widehat{\mu}_n + \frac{\sqrt{s_n^2}}{\sqrt{n}} t^*_{n-1,\frac{\alpha}{2}} \leq \mu \leq \widehat{\mu}_n + \frac{\sqrt{s_n^2}}{\sqrt{n}} t^*_{n-1,1-\frac{\alpha}{2}}\right) = 1 - \alpha$
  - Correct: Across samples, the C.I.s cover μ with probability $1 - \alpha$
  - Wrong: A given C.I. contains μ with probability $1 - \alpha$

Approximate Inference

- In exact inference, we need to assume data distribution F_X
- We may not know a reasonable parametric assumption on F_X
- Derived sampling distribution may not be in a well known family
- Asymptotic inference: Sampling distributions are approximated by the limit as sample size n approaches ∞
- What is the limit of r.v.s?
- Limit of a sequence: Let $\{x_n\}_{n=1}^{\infty}$ be a sequence of real numbers. The limit of sequence x_n, written by $\lim_{n\to\infty} x_n = x$ or $x_n \to x$, is
$$\lim_{n\to\infty} x_n = x \overset{\text{def.}}{\iff} \forall \varepsilon > 0 \; \exists N \text{ s.t. } |x_n - x| < \varepsilon \text{ for } n > N$$
- Convergence in probability: Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of r.v.s. X_n converges in probability to an r.v. X if and only if for any $\varepsilon > 0$
$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| \geq \varepsilon) = 0.$$
We write $X_n \xrightarrow{p} X$ or $\operatorname{plim}_{n\to\infty} X_n = X$
  - Recall that X_n is random but $\mathbb{P}(|X_n - X| \geq \varepsilon)$ is not
  - Consider $\mathbb{P}(|X_n - X| \geq \varepsilon)$ as a sequence, and its limit is 0
  - X can be a constant

Law of Large Numbers

- Does estimator T_n approach estimand θ as n approaches ∞?
- Consistency: T_n is a *consistent estimator* of θ if $T_n \xrightarrow{p} \theta$
  - As n increases, probability that T_n is away from θ vanishes
  - Consistency neither implies or is implied by unbiasedness
- Weak law of large numbers (LLN): Let $X_1, \ldots, X_n$ form a random sample from F_X with a finite second moment. Then,

$$\overline{X}_n \equiv \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{p} \mathbb{E}[X]$$

  - Powerful tool to establish consistency of an estimator
  - $\widehat{\mu}_n$ is a consistent estimator of μ
- Continuous mapping theorem: Let g be a continuous function. Then, $X_n \xrightarrow{p} X$ implies that $g(X_n) \xrightarrow{p} g(X)$
  - If F_X has a finite fourth moment, $\widehat{\sigma^2}_n$ is a consistent estimator of σ^2

$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 \xrightarrow{p} \mathbb{E}[X^2], \quad \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2 \xrightarrow{p} (\mathbb{E}[X])^2$$

Proof of LLN

- Markov inequality: For any r.v. X and constant $a > 0$,

$$\mathbb{P}(|X| \geq a) \leq \frac{\mathbb{E}[|X|]}{a}$$

Proof.

1. $1\{|X|/a \geq 1\} \leq |X|/a$
2. $\underbrace{\mathbb{E}\left[1\{|X|/a \geq 1\}\right]}_{\mathbb{P}(|X| \geq a)} \leq \mathbb{E}\left[|X|\right]/a$

- Chebychev inequality: If X have finite variance, for any $a > 0$

$$\mathbb{P}\left(|X - \mathbb{E}[X]| \geq a\right) \leq \frac{\mathbb{V}(X)}{a^2}$$

Proof.

$$\mathbb{P}\left(|X - \mathbb{E}[X]| \geq a\right) = \mathbb{P}\left((X - \mathbb{E}[X])^2 \geq a^2\right) \leq \frac{\mathbb{E}\left[(X - \mathbb{E}[X])^2\right]}{a^2}$$

- Proof of LLN: By Chebychev,

$$\mathbb{P}\left(|\bar{X}_n - \mathbb{E}[X]| \geq \varepsilon\right) \leq \frac{\mathbb{V}(\bar{X}_n)}{\varepsilon^2} = \frac{\mathbb{V}(X)}{n\varepsilon^2} \to 0 \text{ as } n \to \infty$$

Central Limit Theorem

- Consistency is not about the distribution of T_n
- Sampling distribution at the limit: *Asymptotic distribution*
- Convergence in distribution: A sequence of r.v.s, $\{X_n\}_{n=1}^{\infty}$ converges in distribution to r.v. X if and only if
$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$$
at all points x where $F_X(x)$ is continuous. We write $X_n \xrightarrow{d} X$
- Central limit theorem (CLT): Let $X_1, \ldots, X_n$ form a random sample from F_X with m.g.f. $M_X(t)$. Then,
$$\frac{\sqrt{n}(\overline{X}_n - \mathbb{E}[X])}{\sqrt{(\mathbb{V}(X))}} \xrightarrow{d} \mathcal{N}(0,1)$$
    - Whatever F_X is, $\overline{X}_n$ follows the Gaussian!
    - Powerful tool to establish *asymptotic normality* of estimators
    - $\widehat{\mu}_n$ is asymptotically Normal $\rightsquigarrow$ tests and C.I.s with large n
- Slutzky's theorem: If $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} c$ for constant c. Then,
$$X_n + Y_n \xrightarrow{d} X + c, \quad X_n Y_n \xrightarrow{d} cX$$

Proof of CLT

- Proof here assumes that F_X has its m.g.f.
- CLT holds under much weaker conditions (c.f. **DS** 6.3)
- **Proof.**
  1. WLOG, $\mathbb{E}[X] = 0$ and $\mathbb{V}(X) = 1 \Rightarrow M_X'(0) = 0$ and $M_X''(0) = 1$
  2. The m.g.f. of $\sqrt{n}\overline{X}_n = \sum_{i=1}^n X_i/\sqrt{n}$ is
  $$\mathbb{E}[e^{t\sum_{i=1}^n X_i/\sqrt{n}}] = M_X\left(\frac{t}{\sqrt{n}}\right)^n$$
  3. Its limit is the indeterminate form
  4. Take the limit of the log and exponentiate
  $$\lim_{n\to\infty} n \log M_X\left(\frac{t}{\sqrt{n}}\right) = \lim_{y\to 0} \frac{\log M_X(yt)}{y^2} = \lim_{y\to 0} \frac{tM_X'(yt)}{2yM_X(yt)} = \frac{t}{2}\lim_{y\to 0}\frac{M_X'(yt)}{y}$$
  $$= \frac{t^2}{2}\lim_{y\to 0} M''(yt) = \frac{t^2}{2}$$
  Second and fourth equalities hold by L'Hôpital's rule
  5. $e^{t^2/2}$ is the standard Gaussian's m.g.f.

Asymptotic Tests and C.I.s

- CLT + Slutzky $\rightsquigarrow$ asymptotic distribution of a test statistic
- *Z*-test: Under the null hypothesis that $\mathbb{E}[X] = \theta_0$,

$$Z_n(\theta_0) \equiv \frac{\sqrt{n}(\widehat{\mu}_n - \theta_0)}{\sqrt{s_n^2}} \xrightarrow{d} \mathcal{N}(0, 1)$$

  because $s_n^2 \xrightarrow{p} \mathbb{V}(X)$
  - Reject (accept) the null if $Z_n(\theta_0) \notin (\in) \left(z^*_{\frac{\alpha}{2}}, z^*_{1-\frac{\alpha}{2}}\right)$
  - z^*_δ: δ quantile of the standard Gaussian distribution
- Asymptotic confidence intervals:

$$z^*_{\frac{\alpha}{2}} \leq Z_n(\theta_0) \leq z^*_{1-\frac{\alpha}{2}} \Leftrightarrow \widehat{\mu}_n + \frac{\sqrt{s_n^2}}{\sqrt{n}} z^*_{\frac{\alpha}{2}} \leq \theta_0 \leq \widehat{\mu}_n + \frac{\sqrt{s_n^2}}{\sqrt{n}} z^*_{1-\frac{\alpha}{2}}$$

- Asymptotics are useful: Binary, discrete, skewed, etc.
- Asymptotics are not always correct
  - Probability of Type I error is not exactly α
  - Probability that C.I.s cover μ is not exactly $1 - \alpha$
  - Sample size is always finite:
    1. Consistent estimator can be biased
    2. Asymptotic approximation can be poor

The Delta Method

- Recall the Nigeria survey example:
  - X_i: True answer, 1 if contact and 0 otherwise
  - W_i: Dice roll, $1, 2, 3, 4, 5, 6$
  - Y_i: Observed response, 1 if yes and 0 if no
- Estimand is $\mathbb{E}[X_i]$, true probability of contact with armed groups
- We only observe Y_i. Can we estimate $\mathbb{E}[X_i]$?
- $\mathbb{E}[Y_i] = \mathbb{P}(W_i = 6) + \mathbb{P}(W_i \in \{2, 3, 4, 5\})\mathbb{E}[X_i] = 1/6 + 2\mathbb{E}[X_i]/3$
- Consistent estimator of $\mathbb{E}[Y_i]$: $\overline{Y}_n \xrightarrow{p} \mathbb{E}[Y_i]$ by LLN
- Use the continuous mapping theorem!
$$\widehat{\mu}_X \equiv \frac{3}{2}\left(\overline{Y}_n - \frac{1}{6}\right) \xrightarrow{p} \mathbb{E}[X_i]$$
- The Delta Method: For a differentiable function g s.t. $g'(\mu) \neq 0$,
$$\frac{\sqrt{n}(g(\widehat{\mu}_n) - g(\mu))}{|g'(\mu)|\sqrt{\mathbb{V}(X)}} \xrightarrow{d} \mathcal{N}(0, 1)$$
  Proof uses Taylor approximation
- You can derive the asymptotic distribution of $\widehat{\mu}_X$

Maximum Likelihood Estimator

- For some data sets, you want to make parametric assumptions
  - E.g., Binary indicator for Dem support $\rightsquigarrow$ Bernoulli
- For Bernoulli r.v.s, we know:
  1. $\mathbb{E}[X] = p$
  2. $\mathbb{V}(X) = p(1-p)$
- We only need to estimate p, parameter of the distribution
- Maximum Likelihood Estimator (MLE): For a random sample $X_i \overset{\text{i.i.d.}}{\sim} f_X(x;\theta)$, the *maximum likelihood estimator* of θ is given by

$$\widehat{\theta}_{MLE} \equiv \underset{\theta}{\operatorname{argmax}}\, L_n(\theta) = \underset{\theta}{\operatorname{argmax}} \prod_{i=1}^{n} f_X(X_i;\theta)$$

- Log-likelihood:

$$\ell_n(\theta) \equiv \log \prod_{i=1}^{n} f_X(X_i;\theta) = \sum_{i=1}^{n} \log f_X(X_i;\theta)$$

- Log is monotone $\rightsquigarrow$ MLE maximizes the log-likelihood, too
- Differentiation is much easier as product becomes summation

Consistency and Invariance of MLE

- MLE is consistent: Under "regularity conditions," $\widehat{\theta}_{\text{MLE}} \xrightarrow{p} \theta$
- MLE is invariant: If g is a one-to-one function,
  1. $g(\widehat{\theta}_{\text{MLE}})$ is the MLE of $g(\theta)$
  2. Hence, $g(\widehat{\theta}_{\text{MLE}}) \xrightarrow{p} g(\theta)$

- Overdispersion:
  - $X_i \overset{\text{i.i.d.}}{\sim} \text{Bern}(p)$
    - $\widehat{p}_{\text{MLE}} = \operatorname{argmax}_p \sum_{i=1}^n \{X_i \log p + (1 - X_i) \log(1-p)\} = \overline{X}_n \xrightarrow{p} p$
    - $\widehat{\mathbb{V}(X)}_{\text{MLE}} = \overline{X}_n(1 - \overline{X}_n) \xrightarrow{p} \mathbb{V}(X) = p(1-p)$
  - $X_i \overset{\text{i.i.d.}}{\sim} \text{Pois}(\lambda)$
    - $\widehat{\lambda}_{\text{MLE}} = \operatorname{argmax}_\lambda \sum_{i=1}^n \{X_i \log \lambda - \lambda\} = \overline{X}_n \xrightarrow{p} \lambda$
    - $\widehat{\mathbb{V}(X)}_{\text{MLE}} = \overline{X}_n \xrightarrow{p} \mathbb{V}(X) = \lambda$
  - $\widehat{\sigma^2}_n = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X}_n)^2 \xrightarrow{p} \mathbb{V}(X)$ under no parametric assumptions
  - $\widehat{\sigma^2}_n \gg \widehat{\mathbb{V}(X)}_{\text{MLE}}$ suggests parametric assumption is inappropriate

Fisher Information

- MLE is asymptotically normal:
$$\sqrt{n}(\widehat{\theta} - \theta) \xrightarrow{d} \mathcal{N}\left(0, \mathcal{I}(\theta)^{-1}\right)$$
where $\mathcal{I}(\theta)^{-1}$ is the *Fisher information*

- Score: $s_n(\theta) \equiv \frac{\partial}{\partial \theta}\ell_n(\theta) = \sum_{i=1}^n \frac{\partial}{\partial \theta} \log f_X(X_i; \theta) = \sum_{i=1}^n \frac{\frac{\partial}{\partial \theta} f_X(X_i;\theta)}{f_X(X_i;\theta)}$
- Expected score for each i is zero:
$$\mathbb{E}\left[s_i(\theta)\right] = \int \frac{\frac{\partial}{\partial \theta} f_X(x_i; \theta)}{f_X(x_i; \theta)} f_X(x_i; \theta) dx_i = \frac{\partial}{\partial \theta} \underbrace{\int f_X(x_i; \theta) dx_i}_{=1} = 0$$

- Fisher information: $\mathcal{I}(\theta) \equiv \mathbb{E}\left[s_i(\theta)s_i(\theta)^\top\right] = \mathbb{V}(s_i(\theta))$
- Information equality: For *Hessian* $\mathbf{H}_i(\theta) \equiv \frac{\partial^2}{\partial \theta \partial \theta^\top} \log f_X(X_i; \theta)$,
$$\mathbb{E}\left[\mathbf{H}_i(\theta)\right] = -\mathbb{E}[s_i(\theta)s_i(\theta)^\top] + \frac{\partial}{\partial \theta^\top} \underbrace{\frac{\partial}{\partial \theta} \int f_X(x_i; \theta) dx_i}_{=0} = -\mathcal{I}(\theta)$$

Asymptotic Normality of MLE

- Score function evaluated at MLE is zero: $s_n(\widehat{\theta}_{\text{MLE}}) = 0$
- Taylor expansion of $s_n(\widehat{\theta}_{\text{MLE}})$ around θ:

$$0 = s_n(\widehat{\theta}_{\text{MLE}}) \approx s_n(\theta) + \left(\sum_{i=1}^{n} \mathbf{H}_i(\theta)\right)(\widehat{\theta}_{\text{MLE}} - \theta)$$

$$\sqrt{n}(\widehat{\theta}_{\text{MLE}} - \theta) \approx -\left(\sum_{i=1}^{n} \mathbf{H}_i(\theta)\right)^{-1} \sqrt{n} s_n(\theta)$$

$$= \underbrace{\left(-\frac{1}{n}\sum_{i=1}^{n} \mathbf{H}_i(\theta)\right)^{-1}}_{\xrightarrow{\text{p}} \mathcal{I}(\theta)} \underbrace{\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} s_i(\theta)\right)}_{\xrightarrow{\text{d}} \mathcal{N}(0, \mathcal{I}(\theta))} \xrightarrow{\text{d}} \mathcal{N}\left(0, \mathcal{I}(\theta)^{-1}\right)$$

- Estimated asymptotic variance of MLE:

$$\mathbb{V}\left(\widehat{\theta}_{\text{MLE}}\right) \approx \frac{1}{n}\left(\mathbb{E}\left[-\mathbf{H}_i(\widehat{\theta}_{\text{MLE}})\right]\right)^{-1} \approx \frac{1}{n}\mathbb{E}\left[s_i(\widehat{\theta}_{\text{MLE}}) s_i(\widehat{\theta}_{\text{MLE}})^\top\right]$$

- Hypothesis tests and C.I.s: Replace $\sqrt{s_n^2}$ with $\text{se}(\widehat{\theta}_{\text{MLE}})$

Asymptotic Efficiency of MLE

- Cramér-Rao Lower Bound (univariate): Let $X_1, \dots, X_n$ form a random sample from $f_X(x;\theta)$ and T_n be an estimator of θ. Then,

$$\mathbb{V}(T_n) \geq \frac{\left(\frac{\partial}{\partial\theta}\mathbb{E}[T_n]\right)^2}{n\mathcal{I}(\theta)}$$

Proof.

$$\frac{\partial}{\partial\theta}\mathbb{E}[T_n] = \mathbb{E}\left[T_n \frac{\partial}{\partial\theta}\sum_{i=1}^{n}\log f_X(X_i;\theta)\right] = \text{Cov}\left(T_n, s_n(\theta)\right)$$

- Cauchy-Schwarz inequality: For r.v.s X and Y with finite variance,

$$\text{Cov}(X,Y)^2 \leq \mathbb{V}(X)\mathbb{V}(Y)$$

- Implication of Cauchy-Shwarz: $\text{Cov}\left(T_n, s_n(\theta)\right)^2 \leq \mathbb{V}(T_n)\underbrace{\mathbb{V}\left(s_n(\theta)\right)}_{n\mathcal{I}(\theta)}$
- MLE is asymptotically efficient: MLE achieves CRLB as $n \to \infty$
- MLE has the minimum asymptotic variance