- Mean squared error: $\textsc{mse} = \mathbb{E}\left[(\widehat{\theta}_n - \theta)^2\right] = \mathsf{bias}(\widehat{\theta}_n)^2 + \mathbb{V}\left[\widehat{\theta}_n\right]$
- $\lim_{n\to\infty} \mathsf{bias}(\widehat{\theta}_n) = 0 \wedge \lim_{n\to\infty} \mathsf{se}(\widehat{\theta}_n) = 0 \implies \widehat{\theta}_n$ is consistent
- Asymptotic normality: $\dfrac{\widehat{\theta}_n - \theta}{\mathsf{se}} \xrightarrow{\text{D}} \mathcal{N}(0,1)$
- Slutzky's Theorem often lets us replace $\mathsf{se}(\widehat{\theta}_n)$ by some (weakly) consistent estimator $\widehat{\sigma}_n$.

11.2 Normal-Based Confidence Interval

Suppose $\widehat{\theta}_n \approx \mathcal{N}\left(\theta, \widehat{\mathsf{se}}^2\right)$. Let $z_{\alpha/2} = \Phi^{-1}(1-(\alpha/2))$, i.e., $\mathbb{P}\left[Z > z_{\alpha/2}\right] = \alpha/2$ and $\mathbb{P}\left[-z_{\alpha/2} < Z < z_{\alpha/2}\right] = 1-\alpha$ where $Z \sim \mathcal{N}(0,1)$. Then

$$C_n = \widehat{\theta}_n \pm z_{\alpha/2}\widehat{\mathsf{se}}$$

11.3 Empirical distribution

Empirical Distribution Function (ECDF)

$$\widehat{F}_n(x) = \frac{\sum_{i=1}^n I(X_i \le x)}{n}$$

$$I(X_i \le x) = \begin{cases} 1 & X_i \le x \\ 0 & X_i > x \end{cases}$$

Properties (for any fixed x)

- $\mathbb{E}\left[\widehat{F}_n\right] = F(x)$
- $\mathbb{V}\left[\widehat{F}_n\right] = \dfrac{F(x)(1-F(x))}{n}$
- $\textsc{mse} = \dfrac{F(x)(1-F(x))}{n} \xrightarrow{\text{D}} 0$
- $\widehat{F}_n \xrightarrow{\text{P}} F(x)$

Dvoretzky-Kiefer-Wolfowitz (DKW) inequality ($X_1, \dots, X_n \sim F$)

$$\mathbb{P}\left[\sup_x \left|F(x) - \widehat{F}_n(x)\right| > \varepsilon\right] = 2e^{-2n\varepsilon^2}$$

Nonparametric $1-\alpha$ confidence band for F

$$L(x) = \max\{\widehat{F}_n - \epsilon_n, 0\}$$
$$U(x) = \min\{\widehat{F}_n + \epsilon_n, 1\}$$
$$\epsilon = \sqrt{\frac{1}{2n}\log\left(\frac{2}{\alpha}\right)}$$

$$\mathbb{P}\left[L(x) \le F(x) \le U(x)\ \forall x\right] \ge 1-\alpha$$

11.4 Statistical Functionals

- Statistical functional: $T(F)$
- Plug-in estimator of $\theta = (F)$: $\widehat{\theta}_n = T(\widehat{F}_n)$
- Linear functional: $T(F) = \int \varphi(x)\, dF_X(x)$
- Plug-in estimator for linear functional:

$$T(\widehat{F}_n) = \int \varphi(x)\, d\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^n \varphi(X_i)$$

- Often: $T(\widehat{F}_n) \approx \mathcal{N}\left(T(F), \widehat{\mathsf{se}}^2\right) \implies T(\widehat{F}_n) \pm z_{\alpha/2}\widehat{\mathsf{se}}$
- p^{th} quantile: $F^{-1}(p) = \inf\{x : F(x) \ge p\}$
- $\widehat{\mu} = \bar{X}_n$
- $\widehat{\sigma}^2 = \dfrac{1}{n-1}\displaystyle\sum_{i=1}^n (X_i - \bar{X}_n)^2$
- $\widehat{\kappa} = \dfrac{\frac{1}{n}\sum_{i=1}^n (X_i - \widehat{\mu})^3}{\widehat{\sigma}^3}$
- $\widehat{\rho} = \dfrac{\sum_{i=1}^n (X_i - \bar{X}_n)(Y_i - \bar{Y}_n)}{\sqrt{\sum_{i=1}^n (X_i - \bar{X}_n)^2}\sqrt{\sum_{i=1}^n (Y_i - \bar{Y}_n)^2}}$

12 Parametric Inference

Let $\mathfrak{F} = \left\{f(x;\theta) : \theta \in \Theta\right\}$ be a parametric model with parameter space $\Theta \subset \mathbb{R}^k$ and parameter $\theta = (\theta_1, \dots, \theta_k)$.

12.1 Method of Moments

j^{th} moment

$$\alpha_j(\theta) = \mathbb{E}\left[X^j\right] = \int x^j\, dF_X(x)$$

j^{th} sample moment

$$\widehat{\alpha}_j = \frac{1}{n}\sum_{i=1}^n X_i^j$$

Method of Moments estimator (MoM)

$$\begin{aligned} \alpha_1(\theta) &= \widehat{\alpha}_1 \\ \alpha_2(\theta) &= \widehat{\alpha}_2 \\ \vdots &= \vdots \\ \alpha_k(\theta) &= \widehat{\alpha}_k \end{aligned}$$

Properties of the MoM estimator

- $\widehat{\theta}_n$ exists with probability tending to 1
- Consistency: $\widehat{\theta}_n \xrightarrow{\text{P}} \theta$
- Asymptotic normality:

$$\sqrt{n}(\widehat{\theta} - \theta) \xrightarrow{\text{D}} \mathcal{N}(0, \Sigma)$$

where $\Sigma = g\mathbb{E}\left[YY^T\right]g^T$, $Y = (X, X^2, \ldots, X^k)^T$, $g = (g_1, \ldots, g_k)$ and $g_j = \frac{\partial}{\partial\theta}\alpha_j^{-1}(\theta)$

12.2 Maximum Likelihood

Likelihood: $\mathcal{L}_n : \Theta \to [0, \infty)$

$$\mathcal{L}_n(\theta) = \prod_{i=1}^{n} f(X_i; \theta)$$

Log-likelihood

$$\ell_n(\theta) = \log \mathcal{L}_n(\theta) = \sum_{i=1}^{n} \log f(X_i; \theta)$$

Maximum likelihood estimator (MLE)

$$\mathcal{L}_n(\widehat{\theta}_n) = \sup_{\theta} \mathcal{L}_n(\theta)$$

Score function

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

Fisher information

$$I(\theta) = \mathbb{V}_\theta\left[s(X; \theta)\right]$$

$$I_n(\theta) = nI(\theta)$$

Fisher information (exponential family)

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

Observed Fisher information

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

Properties of the MLE

- Consistency: $\widehat{\theta}_n \xrightarrow{\text{P}} \theta$
- Equivariance: $\widehat{\theta}_n$ is the MLE $\implies \varphi(\widehat{\theta}_n)$ is the MLE of $\varphi(\theta)$
- Asymptotic optimality (or efficiency), i.e., smallest variance for large samples. If $\widetilde{\theta}_n$ is any other estimator, the asymptotic relative efficiency is:
  1. $\mathsf{se} \approx \sqrt{1/I_n(\theta)}$
  $$\frac{(\widehat{\theta}_n - \theta)}{\mathsf{se}} \xrightarrow{\text{D}} \mathcal{N}(0, 1)$$
  2. $\widehat{\mathsf{se}} \approx \sqrt{1/I_n(\widehat{\theta}_n)}$
  $$\frac{(\widehat{\theta}_n - \theta)}{\widehat{\mathsf{se}}} \xrightarrow{\text{D}} \mathcal{N}(0, 1)$$
- Asymptotic optimality

$$\text{ARE}(\widetilde{\theta}_n, \widehat{\theta}_n) = \frac{\mathbb{V}\left[\widehat{\theta}_n\right]}{\mathbb{V}\left[\widetilde{\theta}_n\right]} \le 1$$

- Approximately the Bayes estimator

12.2.1 Delta Method

If $\tau = \varphi(\widehat{\theta})$ where φ is differentiable and $\varphi'(\theta) \neq 0$:

$$\frac{(\widehat{\tau}_n - \tau)}{\widehat{\mathsf{se}}(\widehat{\tau})} \xrightarrow{\text{D}} \mathcal{N}(0, 1)$$

where $\widehat{\tau} = \varphi(\widehat{\theta})$ is the MLE of τ and

$$\widehat{\mathsf{se}} = \left|\varphi'(\widehat{\theta})\right| \widehat{\mathsf{se}}(\widehat{\theta}_n)$$

12.3 Multiparameter Models

Let $\theta = (\theta_1, \ldots, \theta_k)$ and $\widehat{\theta} = (\widehat{\theta}_1, \ldots, \widehat{\theta}_k)$ be the MLE.

$$H_{jj} = \frac{\partial^2 \ell_n}{\partial\theta^2} \qquad H_{jk} = \frac{\partial^2 \ell_n}{\partial\theta_j \partial\theta_k}$$

Fisher information matrix

$$I_n(\theta) = -\begin{bmatrix} \mathbb{E}_\theta\left[H_{11}\right] & \cdots & \mathbb{E}_\theta\left[H_{1k}\right] \\ \vdots & \ddots & \vdots \\ \mathbb{E}_\theta\left[H_{k1}\right] & \cdots & \mathbb{E}_\theta\left[H_{kk}\right] \end{bmatrix}$$

Under appropriate regularity conditions

$$(\widehat{\theta} - \theta) \approx \mathcal{N}(0, J_n)$$

with $J_n(\theta) = I_n^{-1}$. Further, if $\widehat{\theta}_j$ is the j^{th} component of θ, then

$$\frac{(\widehat{\theta}_j - \theta_j)}{\widehat{\mathsf{se}}_j} \xrightarrow{\text{D}} \mathcal{N}(0,1)$$

where $\widehat{\mathsf{se}}_j^2 = J_n(j,j)$ and $\text{Cov}\left[\widehat{\theta}_j, \widehat{\theta}_k\right] = J_n(j,k)$

12.3.1 Multiparameter delta method

Let $\tau = \varphi(\theta_1, \ldots, \theta_k)$ and let the gradient of φ be

$$\nabla\varphi = \begin{pmatrix} \dfrac{\partial\varphi}{\partial\theta_1} \\ \vdots \\ \dfrac{\partial\varphi}{\partial\theta_k} \end{pmatrix}$$

Suppose $\nabla\varphi\big|_{\theta=\widehat{\theta}} \neq 0$ and $\widehat{\tau} = \varphi(\widehat{\theta})$. Then,

$$\frac{(\widehat{\tau} - \tau)}{\widehat{\mathsf{se}}(\widehat{\tau})} \xrightarrow{\text{D}} \mathcal{N}(0,1)$$

where

$$\widehat{\mathsf{se}}(\widehat{\tau}) = \sqrt{\left(\widehat{\nabla}\varphi\right)^T \widehat{J}_n \left(\widehat{\nabla}\varphi\right)}$$

and $\widehat{J}_n = J_n(\widehat{\theta})$ and $\widehat{\nabla}\varphi = \nabla\varphi\big|_{\theta=\widehat{\theta}}$.

12.4 Parametric Bootstrap

Sample from $f(x;\widehat{\theta}_n)$ instead of from $\widehat{F}_n$, where $\widehat{\theta}_n$ could be the MLE or method of moments estimator.

13 Hypothesis Testing

$$H_0 : \theta \in \Theta_0 \qquad \text{versus} \qquad H_1 : \theta \in \Theta_1$$

Definitions

- Null hypothesis H_0
- Alternative hypothesis H_1
- Simple hypothesis $\theta = \theta_0$
- Composite hypothesis $\theta > \theta_0$ or $\theta < \theta_0$
- Two-sided test: $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$
- One-sided test: $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$
- Critical value c
- Test statistic T
- Rejection region $R = \{x : T(x) > c\}$
- Power function $\beta(\theta) = \mathbb{P}[X \in R]$
- Power of a test: $1 - \mathbb{P}[\text{Type II error}] = 1 - \beta = \inf_{\theta\in\Theta_1} \beta(\theta)$
- Test size: $\alpha = \mathbb{P}[\text{Type I error}] = \sup_{\theta\in\Theta_0} \beta(\theta)$

	Retain H_0	Reject H_0
H_0 true	$\surd$	Type I Error (α)
H_1 true	Type II Error (β)	$\surd$ (power)

p-value

- p-value $= \sup_{\theta\in\Theta_0} \mathbb{P}_\theta[T(X) \geq T(x)] = \inf\big\{\alpha : T(x) \in R_\alpha\big\}$
- p-value $= \sup_{\theta\in\Theta_0} \underbrace{\mathbb{P}_\theta[T(X^\star) \geq T(X)]}_{1-F_\theta(T(X)) \quad \text{since } T(X^\star)\sim F_\theta} = \inf\big\{\alpha : T(X) \in R_\alpha\big\}$

p-value	evidence
< 0.01	very strong evidence against H_0
$0.01 - 0.05$	strong evidence against H_0
$0.05 - 0.1$	weak evidence against H_0
> 0.1	little or no evidence against H_0

Wald test

- Two-sided test
- Reject H_0 when $|W| > z_{\alpha/2}$ where $W = \dfrac{\widehat{\theta} - \theta_0}{\widehat{\mathsf{se}}}$
- $\mathbb{P}\left[|W| > z_{\alpha/2}\right] \to \alpha$
- p-value $= \mathbb{P}_{\theta_0}[|W| > |w|] \approx \mathbb{P}[|Z| > |w|] = 2\Phi(-|w|)$

Likelihood ratio test

- $T(X) = \dfrac{\sup_{\theta\in\Theta} \mathcal{L}_n(\theta)}{\sup_{\theta\in\Theta_0} \mathcal{L}_n(\theta)} = \dfrac{\mathcal{L}_n(\widehat{\theta}_n)}{\mathcal{L}_n(\widehat{\theta}_{n,0})}$
- $\lambda(X) = 2\log T(X) \xrightarrow{\text{D}} \chi^2_{r-q}$ where $\sum_{i=1}^k Z_i^2 \sim \chi^2_k$ and $Z_1, \ldots, Z_k \overset{iid}{\sim} \mathcal{N}(0,1)$
- p-value $= \mathbb{P}_{\theta_0}[\lambda(X) > \lambda(x)] \approx \mathbb{P}\left[\chi^2_{r-q} > \lambda(x)\right]$