36-752: Advanced Probability **Spring 2018**

Lecture 20: April 12

Lecturer: Alessandro Rinaldo *Scribe: Theresa Gebert*

Note: *LaTeX template courtesy of UC Berkeley EECS dept.*

20.1 Review

Recall the delta method from last lecture.

Theorem 20.1 *If $r_n(X_n - \theta) \xrightarrow{D} X$ in $\mathbb{R}^d$, and $f : \mathbb{R}^d \to \mathbb{R}^k$ is a function differentiable at θ, then*

$$r_n(f(X_n) - f(\theta)) \xrightarrow{D} f'(\theta)X.$$

20.2 Delta Method

Definition 20.2 *The* **variance stabilizing transformation** *is the delta method where $k = d = 1$, so that*

$$r_n(f(X_n) - f(\theta)) \xrightarrow{D} \mathcal{N}(0, f'(\theta)^2\sigma_\theta^2).$$

Example: Consider

$$(X_1, Y_1), \ldots, (X_n, Y_n) \overset{i.i.d.}{\sim} \mathcal{N}\left(0, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right).$$

We want to estimate $\rho = \operatorname{corr}(X, Y)$. We know

$$\rho_n = \frac{\frac{1}{n}\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\frac{1}{n}\sum_i (X_i - \bar{X})^2}\sqrt{\frac{1}{n}\sum_i (Y_i - \bar{Y})^2}}.$$

Then $\sqrt{n}(\rho_n - \rho) \xrightarrow{D} \mathcal{N}(0, (1-\rho^2)^2)$. When constructing a confidence interval, we have to know ρ. One solution could be that we choose $f : \theta \to \int \frac{1}{\sigma_\theta}$, the anti-derivative. For ρ_n, choose $f(\rho) = \int \frac{1}{1-\rho^2} = \frac{1}{2}\log\left(\frac{1+\rho}{1-\rho}\right) = \tanh^{-1}(\rho)$. By the delta method we have that $\sqrt{n}(\tanh^{-1}(\rho_n) - \tanh^{-1}(\rho)) \xrightarrow{D} \mathcal{N}(0, 1)$.

For $\alpha \in (0, 1)$, an asymptotic confidence interval for ρ is:

$$\left\{\rho : \tanh^{-1}(\rho) \in \left[\tanh^{-1}(\rho_n) - \frac{z_{\alpha/2}}{\sqrt{n}}, \tanh^{-1}(\rho_n) + \frac{z_{\alpha/2}}{\sqrt{n}}\right]\right\}$$

which is equal to

$$\left[\tanh\left(\tanh^{-1}(\rho_n) - \frac{z_{\alpha/2}}{\sqrt{n}}\right), \tanh\left(\tanh^{-1}(\rho_n) + \frac{z_{\alpha/2}}{\sqrt{n}}\right)\right].$$

Example: Now consider what happens if $\nabla f(\theta)$ or $f'(\theta) = 0$. Consider $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mathcal{N}(\theta, \Sigma)$, and $f(\theta) = \frac{||\theta||^2}{2}$. If $\sqrt{n}(\bar{X}_n - \theta) \xrightarrow{D} \mathcal{N}(0, \Sigma)$, then $\sqrt{n}(f(\bar{X}_n) - f(\theta)) \xrightarrow{D} \mathcal{N}(0, \theta^T \Sigma \theta)$. If $\theta = 0$, then $\sqrt{n} f(\bar{X}_n) \xrightarrow{D} 0$, i.e. $\sqrt{n} f(\bar{X}_n) \xrightarrow{P} 0$.

Theorem 20.3 *If $f : \mathbb{R}^d \to \mathbb{R}$ is a function that is twice differentiable at θ, and if $r_n(X_n - \theta) \xrightarrow{D} X$ and $\nabla f(\theta) = 0$, then*

$$r_n^2(f(X_n) - f(\theta)) \xrightarrow{D} \frac{1}{2} X^T H_\theta X$$

where H_θ is the Hessian of f at θ, i.e. $(H_\theta)_{i,j} = \frac{\delta^2 f}{\delta x_i \delta x_j}|_\theta$.

Example: Consider $X_1, X_2, \ldots \overset{i.i.d.}{\sim}$ Expo(2), then $E(X_i) = \frac{1}{2}$ for all i. Let $Y_n = \min\{X_1, \ldots, X_n\} \sim$ Expo($2n$). Then $n(Y_n - 0) \xrightarrow{D} X_1$. Let $f(y) = \cos(y)$, which means $f'(y) = -\sin(y)$. Then $n(\cos(Y_n) - 1) \xrightarrow{P} 0$. Since we have that $f''(y) = -\cos(y)$, we know

$$n^2(\cos(Y_n) - 1) = -\frac{1}{2} n^2 Y_n + o_P(1) \xrightarrow{D} -\frac{1}{2} X_1^2.$$

20.3 Characteristic Functions

We will soon discuss central limit theorems, but in order to do so, we need to introduce characteristic functions first. That is because we will be checking the convergence condition for only this small class of functions. Recall the definition of convergence in distribution, where we have that $X_n \xrightarrow{D} X$ if $E(f(X_n)) \to E(f(X))$ for all $f \in \mathcal{F}$. We only need to check this for a small class $\mathcal{F}$ of characteristic functions.

Definition 20.4 *The* **characteristic function** *of a random variable X is the function*

$$t \in \mathbb{R} \to \phi(t) = E(e^{itX})$$

where $i^2 = -1$ and so $e^{i\mu} = cos(\mu) + isin(\mu)$.

Remark 20.5 *Recall that $\phi(t)$ is continuous and bounded. In addition, if $X \in \mathbb{R}^d$, then we have that*

$$t \in \mathbb{R}^d \to \phi(t) = E(e^{it^T X}).$$

Example: Consider our favorite $Z \sim \mathcal{N}(0, 1)$, then $\phi(t) = \exp\{-t^2/2\}$.

What follows are some useful properties of characteristic functions (chf).

20.3.1 Properties of Characteristic Functions

1. $\phi(0) = 1$, $|\phi(t)| \leq 1$;
2. $\phi(-t) = \phi(t)$;
3. $|\phi(t+h) - \phi(t)| \leq E(|e^{itX} - 1|)$ i.e. ϕ is uniformly continuous;
4. the chf of $aX + b$ is $\phi_{aX+b} = e^{itb} \phi_X(at)$;

5. if $E(|X|^r) < \infty$ then $\phi^{(k)}(0) = i^k E(X^k)$.

Two things to remember about characteristic functions:

1. **Inversion formula and uniqueness.** Using ϕ, we can reconstruct the original CDF.

   Theorem 20.6 *$X \stackrel{D}{=} Y$ if and only if $\phi_X(t) = \phi_Y(t)$ for all t, $X, Y, t \in \mathbb{R}$.*

   Lemma 20.7 *(Cramer-Wold.) $X \stackrel{D}{=} Y$ for $X, Y \in \mathbb{R}^d$ if and only if $a^T X \stackrel{D}{=} a^T Y$ for all $a \in \mathbb{R}^d$.*

2. Continuity theorem.

   Theorem 20.8 *Let $\{P_n\}$ be a sequence of probability measures in $(\mathbb{R}^d, \mathcal{B}^d)$ and P be a probability measure in $(\mathbb{R}^d, \mathcal{B}^d)$ with chfs $\{\phi_n\}$ and ϕ. Then $P_n \stackrel{D}{\to} P$ if and only if $\phi_n(t) \to \phi(t)$ for all $t \in \mathbb{R}^d$.*

   Corollary 20.9 *In $\mathbb{R}^d$, $X_n \stackrel{D}{\to} X$ if and only if $\alpha^T X_n \stackrel{D}{\to} \alpha^T X$ for all $\alpha \in \mathbb{R}^d$.*

   Note that these are special properties of $\mathbb{R}^d$.

20.4 Central Limit Theorem

Theorem 20.10 *(Lindeberg-Feller Central Limit Theorem for triangular arrays.) Let $\{r_n\}$ be a monotonically increasing sequence of integers. For each n, let $X_{n,1}, \ldots, X_{n,r_n}$ be independent random variables with $E(X_{n,k}) = 0$ and $Var(X_{n,k}) = \sigma_{n,k}^2$. Let $\sigma_n^2 = \sum_k \sigma_{n,k}^2$ and $S_n = \sum_k X_{n,k}$. Assume the Lindeberg-Feller condition is true:*

$$\frac{1}{\sigma_n^2} \sum_{k=1}^{r_n} E(|X_{n,k}^2| \mathbb{I}\{|X_{n,k}| > \epsilon \sigma_n\}) \to 0, \forall\ \epsilon > 0,$$

then

$$\frac{S_n}{\sigma_n} \stackrel{D}{\to} \mathcal{N}(0,1).$$

The vanilla Central Limit Theorem follows from this more general theorem.

Example: If $X_1, \ldots, X_n \stackrel{i.i.d.}{\sim} (\mu, \sigma^2)$, and the Lindeberg-Feller condition holds:

$$\frac{1}{\sigma^2} E(|X_1^2| \mathbb{I}\{|X_1| > \epsilon \sqrt{n} \sigma\}) \to 0, \forall\ n,$$

then

$$\sqrt{n}(X_n - \mu) \stackrel{D}{\to} \mathcal{N}(0, \sigma^2).$$

Example: Now consider a counterexample. Let $X_{n,k} \sim$ Bern$(1/k)$ and $r_n = n$. The Lindeberg-Feller condition is almost necessary. If $\max_k P(|X_{n,k}| > \epsilon \sigma_n) \to 0$ as $n \to \infty$ for all n, and if $S_n/\sigma_n \stackrel{D}{\to} \mathcal{N}(0,1)$, then the Lindeberg-Feller condition holds.

Lastly, we consider the multivariate case.

Theorem 20.11 *(Multivariate Central Limit Theorem.) Let $\{r_n\}$ be a monotonically increasing sequence of integers. For each n, let $X_{n,1}, \ldots, X_{n,r_n}$ be independent random variables in $\mathbb{R}^d$ with mean zero. If for all $\epsilon > 0$,*

$$\sum_{k=1}^{r_n} E(||X_{n,k}||^2 \mathbb{I}\{||X_{n,k}|| > \epsilon\}) \to 0,$$

and

$$\sum_{k=1}^{n} cov(X_{n,k}) \to \Sigma$$

then

$$\sum_{k=1}^{n} X_{n,k} \xrightarrow{D} \mathcal{N}(0, \Sigma).$$

Example: Consider $X_1, \ldots, X_n \overset{i.i.d.}{\sim} (\theta, \Sigma)$. Then $\sqrt{n}(\bar{X}_n - \theta) \xrightarrow{D} \mathcal{N}(0, \Sigma)$.

Example: Another classic example is used in linear regression. Consider $Y = X\beta + \epsilon$, where $\epsilon = (\epsilon_1, \ldots, \epsilon_n) \overset{i.i.d.}{\sim} (0, \sigma^2)$. Then the ordinary least squares (OLS) solution is $\hat{\beta} = (X^T X)^{-1} X^T Y$ and $\text{Cov}(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$. Then $E(\hat{\beta}) = \beta$. $(X^T X)^{1/2}(\hat{\beta} - \beta) = (X^T X)^{-1/2} X^T \epsilon = \sum_{i=1}^n A^{(i)} \epsilon_i$ where $A = (X^T X)^{-1/2} X^T$, and $\text{Cov}(\Sigma_i A^{(i)} \epsilon_i) = \sigma^2 \mathcal{L}_d$. We need to verify that:

$$\sum_{i=1}^{n} ||A^{(i)}||^2 E(\epsilon_i^2 \mathbb{I}\{||A^{(i)}|| \cdot |\epsilon_i| > \eta\}) \to 0$$

for all $\eta > 0$.