Stats 300b: Theory of Statistics **Winter 2019**

Lecture 2 – January 11

Lecturer: John Duchi *Scribe: Xinkun Nie*

Warning: *these notes may contain factual errors*

Reading: VDV Chapter 2

1. Portmanteau and Prohorov's Theorems
2. Delta method and examples

1 Convergence recap

Definition 1.1. *A sequence of random variables $\{X_n\}$ converges in probability to a random variable X, denoted $X_n \overset{p}{\to} X$, if $P(d(X_n, X) > \varepsilon) \to 0$ for all $\varepsilon > 0$.*

Definition 1.2. *A sequence of random variables $\{X_n\}$ converges in distribution to a random variable X, denoted $X_n \overset{d}{\to} X$, if $P(X_n \le x) \to P(X \le x)$ for all continuity points x of the function $x \mapsto P(X \le x)$. This is equivalent to the assertion that $\mathbb{E}f(X_n) \to \mathbb{E}f(X)$ for all bounded continuous functions f.*

Theorem 1. *(*Slutsky's Theorem*).*

1. *If $d(X_n, Y_n) \overset{p}{\to} 0$, $X_n \overset{d}{\to} X$, then $Y_n \overset{d}{\to} X$.*
2. *If $X_n \overset{d}{\to} X$, $Y_n \overset{d}{\to} c$, then $(X_n, Y_n) \overset{d}{\to} (X, c)$.*

Remark If the limiting distribution of Y_n is not a constant, then the second part of the theorem does not necessarily hold. Because when Y is random and (X, c) is replaced by (X, Y), we must now specify the joint law of (X, Y).

Definition 1.3. *A collection $\{X_\alpha\}_{\alpha \in \mathcal{A}}$ is uniformly tight if or $\forall \epsilon > 0$, $\exists\, M < \infty$ such that*

$$\sup_{\alpha \in \mathcal{A}} \mathbb{P}(\|X_\alpha\| \ge M) \le \epsilon$$

Remark

1. A single random vector is tight
2. If $X_n \overset{d}{\to} X$ then $\{X_n\}$ is uniformly tight. To show this, let x be a continuity point of $\mathbb{P}(\|X\| \ge x)$, then $\mathbb{P}(\|X_n\| \ge x) \to \mathbb{P}(\|X\| \ge x)$. Choose x large enough such that $\mathbb{P}(\|X\| \ge x)$ is small.

Theorem 2. *(Prohorov's theorem)*

A collection of random vectors $\{X_\alpha\}_{\alpha \in A}$ is uniformly tight if and only if it is sequentially compact for weak convergence. i.e. for all sequences $\{X_n\}_{n \in \mathbb{N}} \subset \{X_\alpha\}_{\alpha \in A}$, there exists a subsequence n_k and a random vector X such that $X_{n_k} \overset{d}{\to} X$.

Remark In $\mathbb{R}^d$ this is Helley's selection theorem (i.e. CDFs F_n have convergent subsequences.)

Example 1: ("Easy" way to get uniformly tightness: Markov's inequality)
Let $\{X_\alpha\}_{\alpha\in A}$ satisfy $\mathbb{E}(\|X_\alpha\|^p) \le k < \infty$, for all $\alpha \in A$ and some $p > 0$. Then $\{X_\alpha\}_{\alpha\in A}$ is uniformly tight.

Proof By markov inequality,

$$\mathbb{P}(\|X_\alpha\| \ge M) \le \frac{\mathbb{E}(\|X_\alpha\|^p))}{M^p} \le \frac{k}{M^p} \to 0$$

as $M \to \infty$ □

♣

Theorem 3. *(*Portmanteau Theorem*). Let X_n, X be random vectors. The following are equivalent.*

1. *X_n converges in distribution to X*
2. *$\mathbb{E}(f(X_n)) \to \mathbb{E}(f(X))$ for all bounded and continuous f*
3. *$\mathbb{E}(f(X_n)) \to \mathbb{E}(f(X))$ for all one-Lipschitz f with $f \in [0,1]$*
4. *$\liminf_{n\to\infty} \mathbb{E}(f(X_n)) \ge E(f(X))$ for non-negative and continuous f.*
5. *$\liminf_{n\to\infty} \mathbb{P}(X_n \in O) \ge \mathbb{P}(X \in O)$ for all open sets O*
6. *$\limsup_{n\to\infty} \mathbb{P}(X_n \in C) \le \mathbb{P}(X \in C)$ for all closed sets C*
7. *$\lim_{n\to\infty} \mathbb{P}(X_n \in B) = \mathbb{P}(X \in B)$ for all sets B such that $\mathbb{P}(X \in \partial B) = 0$*

Remark We call a collection of functions $\mathcal{F}$ a determining class if $\mathbb{E}(f(X_n)) \to \mathbb{E}(f(X))$ for all $f \in \mathcal{F}$ if and only if $X_n \xrightarrow{d} X$. For example, from the theory of characteristic functions, we have a determining class $\mathcal{F} = \{x \mapsto e^{it^\top x} \;:\; t \in \mathbb{R}^d\}$.

Example 2: Fourier transforms or characteristic functions. Let $i = \sqrt{-1}$ and $f_t(x) = \exp(it^\top x)$ for $t \in \mathbb{R}^d$. Then

$$\mathbb{E}(f_t(X_n)) \to \mathbb{E}(f_t(X))\, \forall t \in \mathbb{R}^d \iff X_n \xrightarrow{d} X.$$

♣

2 Delta Method

Suppose we have a sequence of statistics T_n that estimate a parameter θ and we know that $r_n(T_n - \theta)$ converges in distribution to T, and $r_n \to \infty$. Intuitively, we think of r_n as the rate of convergence. Suppose a function ϕ is smooth in the neighborhood of θ. Is it possible to say anything about $\phi(T_n) - \phi(\theta)$?

Theorem 4. *(*Delta Method*). Let $r_n \to \infty$ and $\phi : \mathbb{R}^d \to \mathbb{R}^k$ be differentiable at θ and assume that $r_n(T_n - \theta) \xrightarrow{d} T$ for some random vector T. Then*

1. *$r_n(\phi(T_n) - \phi(\theta))$ converges in distribution to $\phi'(\theta)T$*

2. $r_n(\phi(T_n) - \phi(\theta)) - r_n\phi'(\theta)(T_n - \theta)$ *converges in probability to 0*

Here $\phi'(\theta) \in \mathbb{R}^{k\times d}$ *is the Jacobian Matrix* $[\phi'(\theta)]_{ij} = \frac{\partial\phi_i(\theta)}{\partial\theta_j}$

Proof By the definition of the derivative, we have that

$$\phi(t) = \phi(\theta) + \phi'(\theta)(t-\theta) + o(\|t-\theta\|),$$

i.e.

$$\phi(t) = \phi(\theta) + \phi'(\theta)(t-\theta) + R(\|t-\theta\|) \tag{1}$$

where $\lim_{h\to 0}\frac{R(h)}{h} = 0$. Since $r_n(T_n-\theta)$ converges in distribution, we know that $r_n(T_n-\theta) = O_p(1)$, which implies that $r_n\|T_n-\theta\| = O_p(1)$. We also have that $\|T_n-\theta\| = o_p(1)$, which implies $R(\|T_n-\theta\|) = o_p(\|T_n-\theta\|)$. Thus

$$r_nR(\|T_n-\theta\|) = r_no_p(\|T_n-\theta\|) = o_p(r_n\|T_n-\theta\|) = o_p(O_p(1)) = o_p(1).$$

Using this along with (1), we have the second part of the theorem. Noting that $r_n\phi'(\theta)(T_n-\theta) \xrightarrow{d} \phi'(\theta)T$, and applying Slutsky's theorem, we get the first part as well. □

Example 3: Let $X_i \overset{iid}{\sim} P$, $\mathbb{E}(X) = \theta \neq 0$, $\mathrm{Cov}(X) = \Gamma$ and $\phi(h) = \frac{1}{2}\|h\|^2$. Then

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{k}X_i - \theta\right) \xrightarrow{d} \mathsf{N}(0,\Gamma)$$

By the Delta Method, we have

$$\sqrt{n}\left(\frac{1}{2}\left\|\frac{1}{n}\sum X_i\right\|^2 - \frac{1}{2}\|\theta\|^2\right) \xrightarrow{d} \mathsf{N}(0,\theta^T\Gamma\theta).$$

Note if $\|\theta\|^2 = 0$, we actually have

$$\sqrt{n}\left(\frac{1}{2}\left\|\frac{1}{n}\sum X_i\right\|^2 - \frac{1}{2}\|\theta\|^2\right) \xrightarrow{p} 0.$$

So when $\theta = 0$, we would like to somehow adjust $r_n(\phi(T_n)-\phi(\theta))$ so that we get convergence to a non-trivial distribution. This is a precursor to the next section. ♣

Example 4: (Sample Variance). Let $X_1,\ldots,X_n$ be i.i.d with finite fourth moment. Let $\bar{X}_n = n^{-1}\sum_{i=1}^n X_i$, $S_n^2 = n^{-1}\sum_{i=1}^n(X_i-\bar{X}_n)^2$, and $\overline{X_n^2} = n^{-1}\sum_{i=1}^n X_i^2$. We want weak convergence of $\sqrt{n}(S_n^2-\sigma^2)$. First note that $S_n^2 = \overline{X_n^2} - (\bar{X}_n)^2 = \phi(\bar{X}_n,\overline{X_n^2})$, where $\phi(x,y) = y - x^2$. With $\alpha_i = \mathbb{E}X^i$, one can check that

$$\sqrt{n}\left(\begin{pmatrix}\bar{X}_n\\ \overline{X_n^2}\end{pmatrix} - \begin{pmatrix}\alpha_1\\ \alpha_2\end{pmatrix}\right) \xrightarrow{d} \mathsf{N}\left(0,\begin{pmatrix}\alpha_2-\alpha_1^2 & \alpha_3-\alpha_1\alpha_2\\ \alpha_3-\alpha_1\alpha_2 & \alpha_4-\alpha_2^2\end{pmatrix}\right).$$

Then by the Delta Method, we obtain

$$\sqrt{n}(S_n^2-\sigma^2) \xrightarrow{d} \mathsf{N}(0,\alpha_4-\alpha_2^2).$$

♣

3 Second Order Delta Method

Note that the Delta Method is just a Taylor expansion! So if $\phi'(\theta) = 0$, just look at higher order approximations. Usually in such settings, $\phi : \mathbb{R}^d \to \mathbb{R}$, and so $\phi'(\theta) = \nabla\phi(\theta) = 0 \in \mathbb{R}^d$.

Theorem 5. *(Second Order Delta Method). Let $\phi : \mathbb{R}^d \to \mathbb{R}$ be twice differentiable at θ, and $r_n(T_n - \theta) \xrightarrow{d} T$. Then if $\nabla\phi(\theta) = 0$, we have*

$$r_n^2(\phi(T_n) - \phi(\theta)) \xrightarrow{d} \frac{1}{2}T^T\nabla^2\phi(\theta)T.$$

Proof By definition,

$$\phi(t) = \phi(\theta) + \nabla\phi(\theta)^T(t-\theta) + \frac{1}{2}(t-\theta)^T\nabla^2\phi(\theta)(t-\theta) + R(\|t-\theta\|),$$

where $R(h) = o(\|h\|^2)$. Since $\nabla\phi(\theta) = 0$, we actually have

$$\phi(t) = \phi(\theta) + \frac{1}{2}(t-\theta)^T\nabla^2\phi(\theta)(t-\theta) + R(\|t-\theta\|). \tag{2}$$

Note $r_n^2R(\|T_n - \theta\|) = r_n^2 o_p(\|T_n - \theta\|^2) = o_p(\|r_n(T_n - \theta)\|^2)$. Since $r_n(T_n - \theta)$ converges in distribution, so does $\|r_n(T_n - \theta)\|^2$, and so $\|r_n(T_n - \theta)\|^2 = O_p(1)$. Thus

$$r_n^2R(\|T_n - \theta\|) = o_p(O_p(1)) = o_p(1). \tag{3}$$

Now by the continuous mapping theorem, we have that

$$\frac{1}{2}(r_n(T_n-\theta))^T\nabla^2\phi(\theta)(r_n(T_n-\theta)) \xrightarrow{d} \frac{1}{2}T^T\nabla^2\phi(\theta)T. \tag{4}$$

So combining (2), (3), (4) and using Slutsky's lemma, we get the desired convergence in distribution. □

Example 5: Estimating the parameter of a Bernoulli random variable.
Suppose $\theta \in (0,1)$, $X_i \sim \text{Bernoulli}(\theta)$. To estimate θ, we may use the sample mean $\hat{\theta}_n = n^{-1}\sum_{i=1}^n X_i$. Clearly, $\mathbb{E}\hat{\theta}_n = \theta$, $\text{Var}(\hat{\theta}_n) = \frac{\theta(1-\theta)}{n}$. Instead of using mean squared error to measure the performance of $\hat{\theta}_n$, let us use the Kullback-Leibler (KL) divergence (or the log loss). This is

$$D_{KL}(P \,\|\, Q) = \int dP \log\left(\frac{dP}{dQ}\right).$$

Let $P_t = \text{Bernoulli}(t)$, $t \in [0,1]$. So

$$D_{KL}(P_t \,\|\, P_\theta) = t\log\frac{t}{\theta} + (1-t)\log\frac{1-t}{1-\theta}.$$

Let $\phi(t) = D_{KL}(P_t \,\|\, P_\theta)$. Then

$$\phi'(t) = \log\frac{t}{1-t} - \log\frac{\theta}{1-\theta}.$$

Note $\phi'(\theta) = 0$. So we need the second derivative:

$$\phi''(t) = \frac{1}{t} + \frac{1}{1-t} = \frac{1}{t(1-t)},$$

and so $\phi''(\theta) = \frac{1}{\theta(1-\theta)}$. So by the second order Delta Method,

$$nD_{KL}(P_{\hat{\theta}_n} \,\|\, P_\theta) \xrightarrow{d} \frac{1}{2}\chi^2_{(1)}.$$

♣