Stats 300b: Theory of Statistics | **Winter 2017**

Lecture 2 – January 12

Lecturer: John Duchi | *Scribe: Yi Liu*

Warning: *these notes may contain factual errors*

Reading: A.W. van der Vaart. Asymptotic Statistics. Chapter 2 and Chapter 3

1. Prohorov Theorem
2. Portmanteau Lemma
3. Delta Method

1 Prohorov Theorem

Definition 1.1. *A collection of random vectors $\{X_\alpha\}_{\alpha \in A}$ is uniformly tight if for all $\epsilon > 0$, there exists M such that*

$$\sup_\alpha \mathbb{P}(\|X_\alpha\| \geq M) \leq \epsilon$$

Remark A single random vector is tight since $\lim_{n\to\infty} \mathbb{P}(\|X\| > M) = 0$

Remark If X_n converges in distribution to X, then $\{X_n\}_{n\in\mathbb{N}}$ is uniformly tight
Proof Fix a number M such that $\mathbb{P}(\|X\| \geq M) < \epsilon$. By the portmanteau lemma $\mathbb{P}(\|X_n\| \geq M)$ exceeds $\mathbb{P}(\|X\| \geq M)$ arbitrarily little for sufficient large n. Thus there exists N such that $P(\|X\| \geq M) < 2\epsilon$, for all $n \geq N$. Because each of the finitely many variables X_n with $n < N$ is tight, the value fo M can be increased, if necessary, to ensure that $\mathbb{P}(\|X_n\| \geq M) < 2\epsilon$ for every n. □

Theorem 1 (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. $\forall$ sequences $\{X_n\}_{n\in\mathbb{N}} \subset \{X_\alpha\}_{\alpha\in A}$, there exist n_k, a subsequence and a random vector X such that X_{nk} converges in distribution in X.*

Proof d-dimensional analogue of Helly selection □

Example 1: Random variable bounded in expectation
Let $\{X_\alpha\}_{\alpha\in A}$ satisfy $\mathbb{E}(\|X_\alpha\|) < M < \infty$, for $\alpha \in A$. Then $\{X_\alpha\}_{\alpha\in A}$ is uniformly tight.
Proof By markov inequality,

$$\mathbb{P}(\|X_\alpha\| \geq C) \leq \frac{\mathbb{E}(\|X_\alpha\|^p))}{C^p} \leq \frac{M^p}{C^p} \to 0$$

as $C \to \infty$ □

♣

2 Portmanteau Theorem

Theorem 2. *Portmanteau Theorem Let X_n, X be random vectors, then the following are equivalent.*

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

Remark We call a collection of function $\mathcal{F}$a determining class if $E(f(X_n)) \to E(f(X))$ for all $f \in \mathcal{F}$ if and only if $X_n \to X$. For example, the characteristic function are determining class
Remark The function has to be bounded in this case. Consider the following counter example.

Let $g(x) = x$ and

$$X_n = \begin{cases} n^2 & \text{with probability} \frac{1}{n} \\ 0 & \text{otherwise} \end{cases}$$

X_n converges in probability to 0 and therefore converges in distribution to 0. However,

$$\mathbb{E}(g(X_n)) = n \to \infty$$

3 Delta Method

Prompt: Suppose you have a sequence of statistics T_n that estimate a parameter θ and you know that $r_n(T_n - \theta)$ converges in distribution to T when $r_n \to \infty$. (Here we understand r_n as the rate) Suppose ϕ is smooth in the neighborhood of θ, it is possible to say anything about $\phi(T_n)$?

Theorem 3. *Delta Method Let $r_n \to \infty$ and $\phi : \mathbb{R}^d \to \mathbb{R}^k$ be differentiable at θ and assume that $r_n(T_n - \theta)$ in distribution to 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 of derivatives $[\phi'(\theta)]_{ij} = \frac{\partial \phi_i(\theta)}{\partial \theta_j}$

Proof By taylor theorem, we have that

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

as $t \to \theta$
This is equivalent to

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

where $R(t-\theta) = o(\|t-\theta\|)$
we know that $r_n(T_n - \theta) = O_p(1)$ because of the convergence in distribution.

$$r_n R(T_n - \theta) = r_n O_p(\|T_n - \theta\|) = O_p(\|r_n(T_n - \theta)\|) = o_p(1)$$

which converges in probability to 0. Now put everything together and use equation i

$$r_n(\phi(T_n) - \phi(\theta)) + r_n\phi'(\theta)(T_n - \theta) \xrightarrow{p} 0$$

This concludes the proof for part 2)
Now, we note that $r_n\phi'(\theta)(T_n - \theta) \xrightarrow{d} \phi'(\theta)T$ so we apply Slutsky lemma to get that

$$r_n(\phi(T_n) - \phi(\theta)) \xrightarrow{d} \phi'(\theta)T$$

□

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

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

Now, assume that $\Delta\phi(\theta) = \theta$ So

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

Now if $\|\theta\|^2 = 0$, all we get is that

$$\sqrt{n}(\frac{1}{2}\|\frac{1}{n}\sum X_i\|^2) \xrightarrow{p} 0$$

♣

Remark A generalization of the delta method to higher order problems can be done with taylor expansion with careful use of O_p and o_p

Remark If $\phi'(\theta) = 0$, the higher order expansion allows us to get more powerful results and faster rate fo convergences

Theorem 4. *Let $r_n \to 0$ be deterministic, Suppose that $r_n(T_n - \theta) \xrightarrow{d} T$, Let $\phi : \mathbb{R} \to \mathbb{R}$ be twice differentiable at θ such that $\nabla\phi(\theta) = 0$, then we get*

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

where $\Delta^2\phi(\theta)$ is the Hessian Matrix

Proof As before we will do taylor expansion

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

where $R(h) = o(\|h\|^2)$ as $h \to \infty$

$$r_n^2 R(T_n = \theta) = r_n^2 O_p(\|T_n - \theta\|^2) \xrightarrow{p} 0$$

So we have

$$r_n^2(\phi(T_n) - \phi(\theta)) = \frac{1}{2}r_n(T_n-\theta)^T\nabla^2\phi(\theta)r_n(T_n-\theta)) + o_p(1)$$

So

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

by continuous mapping □

Example 3: KL-divergences and likelihood for Bernoulli random variable.

$$D_{\text{kl}}(P\|Q) = \int p\log(\frac{p}{q})d\mu$$

where p, q are density with respect to μ We know that

$$D_{\text{kl}}(P\|Q) \geq 0$$

and

$$D_{\text{kl}}(P\|Q) = 0$$

if and only if $p = q$ Here we consider the Bernoulli: P_θ where $X \in \{0, 1\}$

$$X_n = \begin{cases} 1 & \text{with probability}\theta \\ 0 & \text{with probability } 1-\theta \end{cases}$$

Consider the estimator

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

By CLT, we know that

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathsf{N}(0.\theta(1-\theta))$$

Now, we compute the KL-distance between

$$D_{\text{kl}}\left(P_{\hat{\theta}}\|P_\theta\right)$$

$$\phi(t) = D_{\text{kl}}(P_t\|P_\theta) = t\log(\frac{t}{\theta}) + (1-t)\log(\frac{1-t}{1-\theta})$$

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

We know that

$$\phi'(\theta) = 0$$

so we try

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

$$\phi''(\theta) = \frac{1}{\theta(1-\theta)}$$

now, take $r_n = \sqrt{n}$,$T_n = \hat{\theta}$, we have that

$$nD_{\text{kl}}\left(P_{\hat{\theta}} \| P_{\theta}\right) \xrightarrow{d} \frac{1}{2}\frac{1}{\theta(1-\theta)}T^2 = \frac{1}{2}Z^2 = \frac{1}{2}\chi_1^2$$

♣