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

Lecture 1 – January 8

Lecturer: John Duchi *Scribe: Han Wu*

Warning: *these notes may contain factual errors*

Reading: VDV Chapter 2.1, 2.2

Outline of lecture 1:

- Administrative basic stuff
- Overview of the course
- Basic notions of convergence: Probability, Distributions and CLTs

Course Website: stanford.edu/class/stats300b

Grading:

5% Scribe notes
60% Problem sets (weekly, every Thursday)
35% Final

Overview: What is this course about?

1. Convergence of random variables, random vectors, estimators and functions.

2. Understanding various notions of optimality and quality of estimators and tests.

What you need to be happy/ get through this class:

1. Stat 300a (good to have but not strictly necessary).

2. Probability at stat 310a level. e.g. Convergence of distribution, Helly Selection Theorem etc.

3. Analysis at Math 171 level. e.g. Compactness, metric spaces etc.

Part I of the course:

Finite dimensional problems and statistic models.

Example 1: One example problem is that we have $X_i \overset{\text{iid}}{\sim} P_\theta, X_i \in \mathbb{R}^d$, where d is fixed. We want to understand the estimators of parameter $\theta \in \mathbb{R}^d$ and tests in this regime. ♣

Part II of the course:

Optimality and comparisons of estimators

In this part, we will try to understand when an estimator $\hat{\theta}$ of θ is good or optimal. Can we compare estimators or tests?

Part III of the course:

Infinite dimensional quantities and uniform convergence. Concentration inequalities, and uniform laws i.e.

$$\frac{1}{n}\sum_{i=1}^{n} f(x_i) \to \mathbb{E}[f(x)]$$

uniformly in $f \in \mathscr{F}$

Basic theory of convergence of random variables:

In this part we will go through basic definitions, Continuous Mapping Theorem and Slutsky Lemmas.

For now, assume $X_i \in \mathbb{R}^d, d < \infty$. We first give the definition of various convergence of random variables.

Definition 0.1. *(Convergence in probability) We call $X_n \overset{p}{\to} X$ (X_n converges to X in probability) if*

$$\lim_{n\to\infty} \mathbb{P}(||X_n - X|| \geq \epsilon) = 0, \ \forall \epsilon > 0$$

In a general metric space, with metric ρ, the above definition becomes

$$\lim_{n\to\infty} \mathbb{P}(\rho(X_n, X) \geq \epsilon) = 0, \ \forall \epsilon > 0$$

Definition 0.2. *(Weak convergence or convergence in distribution)*

We say

$$X_n \overset{d}{\to} X$$

if for $\forall x \in \mathbb{R}^d$,

$$\mathbb{P}(X_n \leq x) \to \mathbb{P}(X \leq x)$$

at all $X \in \mathbb{R}^d$ such that $x \to \mathbb{P}(X \leq x)$ is continuous.

Note: In the above definition $\mathbb{P}(X \leq x) = \mathbb{P}(X \in (-\infty, x_1] \times \cdots \times (-\infty, x_d])$

We also have a general definition(for example in Polish space) for convergence in distribution.

Definition 0.3.

$$X_n \overset{d}{\to} X$$

if and only if for all bounded continuous function f,

$$\mathbb{E}[f(X_n)] \to \mathbb{E}[f(X)]$$

Below is the definition of L^p convergence.

Definition 0.4. *(Convergence in the p^{th} mean)*
We say that

$$X_n \overset{L^p}{\to} X$$

if

$$\lim_{n\to\infty} \mathbb{E}[||X_n - X||^p] = 0$$

Finally, we give the definition of almost surely convergence for random variables.

Definition 0.5. *(X_n converges almost surely to X)*
We say that

$$X_n \overset{a.s.}{\to} X$$

if

$$\mathbb{P}(\lim_{n\to\infty} X_n \neq X) = 0$$

i.e.

$$\mathbb{P}(\limsup_{n\to\infty} ||X_n - X|| \geq \epsilon) = 0, \ \forall \epsilon > 0$$

Standard implications:

For the various types of convergence above, we have the following relationships.

$$X_n \overset{a.s.}{\to} X \Rightarrow X_n \overset{p}{\to} X \Rightarrow X_n \overset{d}{\to} X$$
$$X_n \overset{L^p}{\to} X \Rightarrow X_n \overset{p}{\to} X$$

No relations between covergence in L^p and convergence almost surely in either direction. No reversed implication between weak convergence and covergence in L^p.

Note: All proofs above and below can be found in Van der Vaart Chapter 2.

Example 2: Let $X_i \overset{\text{iid}}{\sim} P$, $cov(X_i) = \Sigma = \mathbb{E}[(X_i - \mu)(X_i - \mu)^T]$, $\mu = \mathbb{E}[X_i]$. Then

$$\frac{1}{n}\sum_{i=1}^{n} X_i \overset{a.s.}{\to} \mu$$

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} (X_i - \mu) \overset{d}{\to} \mathsf{N}(0, \Sigma)$$

(The second line is the CLT) ♣

Basic Convergence Theorems:

Theorem 1. *(Continuous Mapping Theorem) Let g be continuous on a set B such that $\mathbb{P}(X \in B) = 1$ then*

$$X_n \xrightarrow{p} X \Rightarrow g(X_n) \xrightarrow{p} g(X)$$
$$X_n \xrightarrow{a.s.} X \Rightarrow g(X_n) \xrightarrow{a.s.} g(X)$$
$$X_n \xrightarrow{d} X \Rightarrow g(X_n) \xrightarrow{d} g(X)$$

For the heuristics of the third line: If g is continuous, then $f \circ g$ is continuous and bounded for any continuous bounded f. Thus,

$$\mathbb{E}[f(g(X_n))] \to \mathbb{E}[f(g(x))]$$

Another important theorem we will need is Slutsky's Theorem.

Theorem 2. *(Slutsky's Theorem)*

(1) If c is constant, then

$$X_n \xrightarrow{d} c \Leftrightarrow X_n \xrightarrow{p} c$$

(2) If $X_n \xrightarrow{d} X$, $d(X_n, Y_n) \xrightarrow{p} 0$, then

$$Y_n \xrightarrow{d} X$$

(3) If $X_n \xrightarrow{d} X$, $Y_n \xrightarrow{p} c$ (c constant), then

$$\begin{pmatrix} X_n \\ Y_n \end{pmatrix} \xrightarrow{d} \begin{pmatrix} X \\ c \end{pmatrix}$$

The Slutsky's theorem allows us to ignore low order terms in convergence. Also, the following example shows that stronger impliations over part (3) may not be true.

Example 3: If $X_n \xrightarrow{d} \mathsf{N}(0, I)$, then $-X_n \xrightarrow{d} \mathsf{N}(0, I)$.

However,

$$\begin{pmatrix} X_n \\ -X_n \end{pmatrix} \xrightarrow{d} \begin{pmatrix} Z \\ -Z \end{pmatrix}$$

where $Z \sim \mathsf{N}(0, I)$ instead of $\mathsf{N}(0, I)$. ♣

Sketch of Proof

(1) The " $\Leftarrow$ " direction is trivial and given in the previous sections. For " $\Rightarrow$ " direction of the theorem, take

$$f(x) = ||x - c|| \wedge 1 = \min\{||x - c||, 1\}$$

then

$$\mathbb{E}[f(x_n)] \to \mathbb{E}[f(c)] = 0$$

i.e.

$$\mathbb{E}[||x_n - c|| \wedge 1] \to 0$$

which implies convergence in probability

(2) Let f be 1-Lipschitz with range [0,1], then we have for any $\epsilon > 0$

$$|\mathbb{E}[f(Y_n)] - \mathbb{E}[f(X_n)]| \leq \epsilon \mathbb{E}1\{d(X_n, Y_n) \leq \epsilon\} + 2E1\{d(X_n, Y_n) > \epsilon\}$$

which implies $E[f(Y_n)]$ and $E[f(X_n)]$ have the same limit. The result follows from Portmanteau.

(3) We have

$$\begin{pmatrix} X_n \\ Y_n \end{pmatrix} - \begin{pmatrix} X \\ c \end{pmatrix} = \begin{pmatrix} 0 \\ Y_n - c \end{pmatrix} \xrightarrow{p} 0$$

By part (2),

$$\begin{pmatrix} X_n \\ c \end{pmatrix} \xrightarrow{d} \begin{pmatrix} X \\ c \end{pmatrix} \Rightarrow \begin{pmatrix} X_n \\ Y_n \end{pmatrix} \xrightarrow{d} \begin{pmatrix} X \\ c \end{pmatrix}$$

The left part of above implication follows from Portmanteau.

□

Consequences of Slutsky's Theorem:

If $X_n \xrightarrow{d} X$, and $Y_n \xrightarrow{d} c$, then

$$X_n + Y_n \xrightarrow{d} X + c$$

$$Y_n X_n \xrightarrow{d} cX$$

If $c \neq 0$,

$$\frac{X_n}{Y_n} \xrightarrow{d} \frac{X}{c}$$

Proof Apply Continuous Mapping Theorem and Slutsky's Theorem and the statements can be proved. □

Note: For the third line of convergence, if $c \in \mathbb{R}^{d \times d}$ is a matrix, then (2) still holds. Moreover, if $\det(c) \neq 0$, (3) holds but

$$Y_n^{-1} X_n \xrightarrow{d} c^{-1} X$$

because $c \to c^{-1}$ is continuous when $\det(c) \neq 0$.

Example 4: (T-like statistics) Let $X_i \overset{\text{iid}}{\sim} P$, $\text{Cov}(X_i) = \Sigma \succ 0$. Define

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

$$S_n = \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu_n)(X_i - \mu_n)^T$$

$$T_n = \frac{1}{\sqrt{n}} S_n^{-\frac{1}{2}} \sum_{i=1}^{n} (X_i - \mu_n)$$

Then $T_n \xrightarrow{d} \mathsf{N}(0, I)$.

The reason is that

$$\mu_n \xrightarrow{p} \mathbb{E}[X]$$
$$S_n \xrightarrow{p} \Sigma$$

and

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - \mu) \xrightarrow{d} N(0, \Sigma)$$

Apply Slutsky's Theorem,

$$T_n - \frac{1}{\sqrt{n}} \Gamma^{-\frac{1}{2}} \sum_{i=1}^{n} (X_i - \mu) \xrightarrow{p} 0$$

♣

Big-O Notation:

In this part we introduce the big-o and little-o notation in probability.

Let X_n be random vectors, and R_n be $\mathbb{R}$-valued random variables. We say that $X_n = o_p(R_n)$ if $\exists$ random vectors Y_n such that

$$X_n = Y_n R_n$$
$$Y_n \xrightarrow{p} 0$$

This is called "little o-pea".

We say that $X_n = O_p(R_n)$ if $\exists$ random vectors Y_n where $Y_n = O_p(1)$. $Y_n = O_p(1)$ means that $\{Y_n\}$ is uniformly tight. i.e.

$$\lim_{M \to \infty} \sup_{n \in \mathbb{N}} \mathbb{P}(||Y_n|| \geq M) = 0$$

or $\forall \epsilon > 0$, $\exists$ M such that

$$\mathbb{P}(||Y_n|| \geq M) \leq \epsilon, \forall\ n$$

Comsequences:

With the definition above, we can get the following properties and lemma.

$$o_p(1) + o_p(1) = o_p(1)$$
$$O_p(1) + O_p(1) = O_p(1)$$

Lemma 3. *Let function $R : \mathbb{R}^d \to \mathbb{R}^k$, with $R(0) = 0$, and $X_n \xrightarrow{p} 0$. Then*

(1) If $R(h) = o(||h||^p)$ as $h \to 0$, then

$$R(X_n) = o_p(||X_n||^p)$$

(2) If $R(h) = O(||h||^p)$ as $h \to 0$, then

$$R(X_n) = O_p(||X_n||^p)$$

Proof Define

$$g(h) = \begin{cases} \dfrac{R(h)}{||h||^p}, if\ h \neq 0 \\ \quad 0, if\ h = 0 \end{cases}$$

(1) Then $g(h) \to 0$ as $h \to 0$. Thus, g is continuous at 0 and $X_n \overset{p}{\to} 0$. Apply Continuous Mapping Theorem(CMT), we get

$$g(X_n) \overset{p}{\to} 0$$

(2) $\exists\ M$, $\delta > 0$ such that $||g(h)|| \leq M$, $\forall ||h|| \leq \delta$. Then

$$\mathbb{P}(||g(X_n)|| > M) \leq \mathbb{P}(||X_n|| > \delta) \to 0$$

so

$$g(X_n) = O_p(1)$$

□