STATS 310A: Theory of Probability I — Autumn 2016/17

Lecture 17: November 28

Lecturer: Persi Diaconis — Scribes: Kenneth Tay

17.1 Lindeberg's Central Limit Theorem

We have the following set-up:

- A triangular array of random variables $\{X_{ni}\}$, $1 \leq i \leq k_n$, such that $\mathbb{E}X_{ni} = 0$, Var $X_{ni} = \sigma_{ni}^2 < \infty$.
- $S_n := \displaystyle\sum_{i=1}^{k_n} X_{ni}$.
- $s_n^2 := \displaystyle\sum_{i=1}^{k_n} \sigma_{ni}^2$.
- The random variables in each row are independent of each other, i.e. for every n, $\{X_{ni}\}_{i=1}^{k_n}$ is independent.

Lindeberg's version of the Central Limit Theorem is as follows:

Theorem 17.1 (Lindeberg's CLT) *Suppose that the Lindeburg condition holds: i.e. for every $\varepsilon > 0$,*

$$\lim_{n \to \infty} \frac{1}{s_n^2} \sum_{i=1}^{k_n} \int_{\{|X_{ni}| > \varepsilon s_n\}} X_{ni}^2 dP = 0.$$

Then for every $x \in \mathbb{R}$,

$$P\left\{ \frac{S_n}{s_n} \leq x \right\} \to \Phi(x).$$

To prove this theorem, we will introduce a few lemmas:

Lemma 17.2 *Let $\mathcal{C}_b^\infty = \{f : f : \mathbb{R} \to \mathbb{R}, f \text{ has bounded derivatives of all order}\}$.*

If F_n, F are distribution functions on $\mathbb{R}$ such that

$$\int_{-\infty}^{\infty} f dF_n \to \int_{-\infty}^{\infty} f dF$$

for all $f \in \mathcal{C}_b^\infty$, then $F_n \Rightarrow F$, i.e. F_n converges weakly to F.

We proved this lemma last lecture.

Lemma 17.3 *Given $f \in \mathcal{C}_b^\infty$, let*

$$g(h) := \sup_{x \in \mathbb{R}} \left| f(x+h) - f(x) - hf'(x) - \frac{h^2}{2} f''(x) \right|.$$

Then there exists a constant $k > 0$ (which depends only on f) such that $g(h) \leq \min(kh^3, kh^2)$.

This lemma follows directly from Taylor's Theorem with remainder.

Lemma 17.4 *For any h_1 and h_2,*

$$\left| f(x+h_1) - f(x+h_2) - f'(x)(h_1 - h_2) + \frac{f''(x)}{2}(h_1^2 - h_2^2) \right| \leq g(h_1) + g(h_2).$$

Let us now prove Lindeberg's CLT.

Proof: By Lemma 17.2, it is enough to show that for every $f \in \mathcal{C}_b^\infty$, $\mathbb{E}f\left(\frac{S_n}{s_n}\right) \to \mathbb{E}(f(Z))$, where Z is a standard normal random variable.

Central idea of the proof: Replace X_{ni} in S_n one at a time with $Z_{ni} \sim \mathcal{N}(0, \sigma_{ni}^2)$.

Let $T_{ni} := X_{n1} + \cdots + X_{n(i-1)} + Z_{n(i+1)} + \cdots + Z_{nk_n}$. Then

$$S_n = T_{nk_n} + X_{nk_n}, \qquad \sum_{i=1}^{k_n} Z_i = T_{n1} + Z_{n1}.$$

Note that $Z := \dfrac{1}{s_n} \displaystyle\sum_{i=1}^{k_n} Z_i \sim \mathcal{N}(0,1)$. Hence,

$$\begin{aligned}
\left| \mathbb{E}f\left(\frac{S_n}{s_n}\right) - \mathbb{E}f(Z) \right| &= \left| \sum_{i=1}^{k_n} \mathbb{E}f\left(\frac{T_{ni} + X_{ni}}{s_n}\right) - \mathbb{E}f\left(\frac{T_{ni} + Z_{ni}}{s_n}\right) \right| \\
&\leq \sum_{i=1}^{k_n} \left| \mathbb{E}f\left(\frac{T_{ni} + X_{ni}}{s_n}\right) - \mathbb{E}f\left(\frac{T_{ni} + Z_{ni}}{s_n}\right) \right|.
\end{aligned}$$

By independence, for any n and i, we have

$$\begin{aligned}
\mathbb{E}f'(T_{ni})(X_{ni} - Z_{ni}) &= 0, \\
\mathbb{E}f''(T_{ni})(X_{ni}^2 - Z_{ni}^2) &= 0.
\end{aligned}$$

Add these terms in and using the bound in Lemma 17.4, we have

$$\begin{aligned}\left|\mathbb{E}f\left(\frac{S_n}{s_n}\right)-\mathbb{E}f(Z)\right| &\le \sum_{i=1}^{k_n}\Bigg|\mathbb{E}f\left(\frac{T_{ni}+X_{ni}}{s_n}\right)-\mathbb{E}f\left(\frac{T_{ni}+Z_{ni}}{s_n}\right)-\mathbb{E}f'\left(\frac{T_{ni}}{s_n}\right)\left(\frac{X_{ni}}{s_n}-\frac{Z_{ni}}{s_n}\right)\\ &\qquad +\mathbb{E}\frac{1}{2}f''\left(\frac{T_{ni}}{s_n}\right)\left[\left(\frac{X_{ni}}{s_n}\right)^2-\left(\frac{Z_{ni}}{s_n}\right)^2\right]\Bigg|\\ &\le \sum_{i=1}^{k_n}\mathbb{E}g\left(\frac{X_{ni}}{s_n}\right)+\sum_{i=1}^{k_n}\mathbb{E}g\left(\frac{Z_{ni}}{s_n}\right)\\ &=: I+II.\end{aligned}$$

First consider I. We break the ith integral into $\{\omega : \frac{|X_{ni}|}{s_n}\le\varepsilon\}$ and its complement, and we use $g(h)\le kh^3$ on the part and $g(h)\le kh^2$ on the second:

$$\begin{aligned}I &\le k\sum_{i=1}^{k_n}\int_{\{|X_{ni}|/s_n\le\varepsilon\}}\frac{|X_{ni}|^3}{s_n^3}dP+k\sum_{i=1}^{k_n}\int_{\{|X_{ni}|/s_n>\varepsilon\}}\frac{|X_{ni}|^2}{s_n^2}dP\\ &\le k\varepsilon+\frac{k}{s_n^2}\sum_{i=1}^{k_n}\int_{\{|X_{ni}|/s_n>\varepsilon\}}|X_{ni}|^2dP.\end{aligned}$$

By the Lindeberg condition, for fixed ε, the second term goes to 0 as $n\to\infty$. Hence, I can be bounded above by a constant multiple of ε, with the constant only depending of f.

Next consider II. Splitting up the integral in the same way as I, we obtain

$$\begin{aligned}II &\le k\varepsilon+\frac{k}{s_n^2}\sum_{i=1}^{k_n}\int_{\{|Z_{ni}|/s_n>\varepsilon\}}|Z_{ni}|^2dP\\ &\le k\varepsilon+\frac{k}{\varepsilon s_n^3}\sum_{i=1}^{k_n}\mathbb{E}|Z_{ni}|^3\\ &= k\varepsilon+\frac{k\mathbb{E}|Z|^3}{\varepsilon s_n^3}\sum_{i=1}^{k_n}\sigma_{ni}^3.\end{aligned}$$

Note that for any $\varepsilon'>0$, we have the bound

$$\frac{\sigma_{ni}^2}{s_n^2}\le(\varepsilon')^2+\frac{1}{s_n^2}\int_{\{|X_{ni}|/s_n>\varepsilon'\}}|X_{ni}|^2dP\to(\varepsilon')^2$$

as $n\to\infty$. Since ε' is arbitrary, it means that $\max\dfrac{\sigma_{ni}}{s_n}\to 0$ as $n\to\infty$. Hence,

$$
\begin{aligned}
II &\leq k\varepsilon + \frac{k\mathbb{E}|Z|^3}{\varepsilon s_n^3}\sum_{i=1}^{k_n}\sigma_{ni}^3 \\
&\leq k\varepsilon + \frac{k\mathbb{E}|Z|^3}{\varepsilon s_n^3}(\max\sigma_{ni})\sum_{i=1}^{k_n}\sigma_{ni}^2 \\
&= k\varepsilon + \frac{k\mathbb{E}|Z|^3}{\varepsilon s_n}(\max\sigma_{ni}) \to 0
\end{aligned}
$$

as $n \to \infty$.

■

Remark: This is a nice example of a coupling proof, or a probabilistic proof.

17.2 CLT Variants

There is a converse of sorts for the theorem, known as the Lindeberg-Feller Theorem. (See Feller Vol II.)

Theorem 17.5 (Lyapunov) *Let X_{ni} be a triangular array such that the X_{ni}'s have mean 0, $\mathbb{E}|X_{ni}|^{2+\delta} < \infty$ for some $\delta > 0$.*

If Lyapunov's condition holds:

$$
\lim_{n\to\infty}\frac{1}{s_n^{2+\delta}}\sum_{i=1}^{k_n}\mathbb{E}|X_{ni}|^{2+\delta} = 0,
$$

then $P\left\{\frac{S_n}{s_n}\right\} \leq x) \to \Phi(x)$.

Proof: We just need to show that the Lindeberg condition holds. Given $\varepsilon > 0$,

$$
\begin{aligned}
\frac{1}{s_n^2}\sum_{i=1}^{k_n}\int_{\{|X_{ni}|/s_n > \varepsilon\}}|X_{ni}|^2 dP &= \frac{1}{s_n^2}\sum_{i=1}^{k_n}\int_{\{|X_{ni}|^\delta > \varepsilon^\delta s_n^\delta\}}|X_{ni}|^2 dP \\
&\leq \frac{1}{\varepsilon^\delta s_n^{2+\delta}}\sum_{i=1}^{k_n}\mathbb{E}|X_{ni}|^{2+\delta} \to 0.
\end{aligned}
$$

■

Theorem 17.6 ("Ordinary" CLT) *Suppose X_i are iid, mean 0, variance σ^2. Then $P\left\{\frac{S_n}{\sigma\sqrt{n}} \leq x\right\} \to \Phi(x)$.*

Proof: Again, we just need to show that the Lindeberg condition holds. Given $\varepsilon > 0$,

$$\begin{aligned}\frac{1}{s_n^2}\sum_{i=1}^{k_n}\int_{\{|X_{ni}|/s_n>\varepsilon\}}|X_{ni}|^2dP &= \frac{1}{n\sigma^2}\sum_{i=1}^{n}\int_{\{|X_{ni}|>\varepsilon\sigma\sqrt{n}\}}|X_{ni}|^2dP \\ &= \frac{1}{n\sigma^2}\cdot n\int_{|X_1|>\varepsilon\sigma\sqrt{n}}|X_1|^2dP \\ &\to 0\end{aligned}$$

as $n \to \infty$ by the Dominated Convergence Theorem. ∎

Example: Say an urn has R red balls and B black balls, where $R + B = N$. Sample the balls without replacement, and let

$$X_i = \begin{cases} 0 & \text{if ball is red,} \\ 1 & \text{if ball is black.} \end{cases}$$

If $\frac{n}{N} \to 0$, then $P\left\{\frac{S_n}{\sigma\sqrt{n}} \le x\right\} \to \Phi(x)$ for some value of σ.

General version (Hoefding Combinatorial CLT): Let $M_n = (m^{ij})_{1\le i,j\le n}$ be a square matrix of numbers. Let $S_n = \sum_{i=1}^{n} m^{i\pi(i)}$, where π is a random permutation (i.e. sum of a random diagonal). Under mild conditions on M_n, the CLT holds.

17.3 Bounds on CLT Approximation

Note that the proof for Lindeberg's CLT gives an actual bound for $\left|\mathbb{E}f\left(\frac{S_n}{s_n}\right) - \mathbb{E}f(Z)\right|$. The Berry-Esseen Theorem also gives such bounds:

Theorem 17.7 (Berry-Esseen) *Say $X_1, \ldots, X_n$ have mean 0, variance σ_i^2, and $\mathbb{E}|X_i|^3 = r_i$ are finite. Then*

$$\sup_{-\infty<x<\infty}\left|P\left\{\frac{S_n}{s_n} \le x\right\} - \Phi(x)\right| \le \frac{0.78R_n}{s_n^3},$$

where $R_n = \sum_{i=1}^n r_i$.

If the X_i's are iid, the RHS is $\frac{0.78\mathbb{E}|X_1|^3}{\sigma^3\sqrt{n}}$.

Alternatively, one can use Edgeworth expansions for better approximations than $\Phi(x)$, e.g.

$$P\left\{\frac{S_n}{s_n} \le x\right\} = \Phi(x) + \frac{H_1(x)}{\sqrt{n}} + O\left(\frac{1}{n}\right).$$