STATS310A - Lecture 16

Persi Diaconis
Scribed by Michael Howes

11/16/21

Contents

1 Announcements

Final homeowork uploaded to Canvas and due November 30.

- Read chpaters 25,26.
- Do 25 / 1,3 and 26 / 1,3,12-14.
- The hints contain surprises.

2 The central limit theorem

We are in the process of proving Lindeberg's version of the central limit theorem. Recall that we have a triangular array of random variables $\{X_{n,i}\}$ where $i = 1, \dots, k_n$ and $n = 1, 2, \dots$. We assume that the array has independent rows. That is, for each n, $\{X_{n,i}\}_{i=1}^{k_n}$ are independent. Assume also that

$$\mathbb{E}[X_{n,i}] = 0 \quad \text{and} \quad \sigma_{i,n}^2 = \mathrm{Var}(X_{n,i}) < \infty.$$

Define $S_n = \sum_{i=1}^{k_n} X_{n,i}$ and $s_n^2 = \sum_{i=1}^{k_n} \sigma_{i,n}^2 = \mathrm{Var}(S_n)$.

Definition 1. A triangular array with independent rows $\{X_{n,i}\}$ is said to satisfy *Lindeberg's condtion* if for all $\varepsilon > 0$

$$\frac{1}{s_n^2} \sum_{i=1}^{k_n} \int_{\{|X_{n,i}| > \varepsilon s_n\}} |X_{n,i}|^2 d\mathbb{P} \xrightarrow{n} 0.$$

Lindeberg's version of the central limit theorem is:

Theorem 1 (Lindeberg)**.** *Let $\{X_{n,i}\}$ be a triangular array with independent rows. If $\{X_{n,i}\}$ satisfy Lindeberg's condition, then for all $x \in \mathbb{R}$,*

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

where $\Phi(x) = \mathbb{P}(Z \leq x)$ for $Z \sim \mathcal{N}(0,1)$.

Proof. To prove this we will use the portmanteau theorem. Let $C_c^\infty(\mathbb{R})$ be the class of infinitely differentiable functions on $\mathbb{R}$ with compact support. By the portmanteau theorem it suffices to show that for all $f \in C_c^\infty(\mathbb{R})$, $\mathbb{E}[f(S_n/s_n)] \stackrel{n}{\to} \mathbb{E}[f(Z)]$ where $Z \sim \mathcal{N}(0,1)$. Thus fix such an f. Define $Z_{n,i}$ is be independent random variables such that $Z_{n,i} \sim \mathcal{N}(0, \sigma_{n,i}^2)$. Let $Z_n = \sum_{i=1}^{k_n} Z_{n,i}$, Then

$$Z = \frac{1}{s_n} Z_n \sim \mathcal{N}(0,1).$$

The idea behind the proof is to swap out $X_{n,i}$ for $Z_{n,i}$ one at a time. With this in mind, define

$$T_{n,i} = X_{n,1} + \ldots + X_{n,i-1} + Z_{n,i} + \ldots + Z_{n,k_n}.$$

Note that X_i, Z_i are independent of $T_{n,i}$ for each i. Furthermore we have

$$S_n = T_{n,k_n} + X_{n,k_n} \quad \text{and} \quad Z_n = T_{n,1} + Z_{n,1}.$$

And also

$$T_{n,i} + Z_{n,i} = T_{n,i-1} + X_{n,i-1},$$

for $i = 2, \ldots, k_n$. Thus by telescoping we have

$$f\left(\frac{S_n}{s_n}\right) - f\left(\frac{Z_n}{s_n}\right) = \sum_{i=1}^{k_n} f\left(\frac{T_{n,i} + X_{n,i}}{s_n}\right) - f\left(\frac{T_{n,i} + Z_{n,i}}{s_n}\right).$$

And so

$$\left|\mathbb{E}\left[f\left(\frac{S_n}{s_n}\right)\right] - \mathbb{E}\left[f(Z)\right]\right| \leq \sum_{i=1}^{k_n} \left|\mathbb{E}\left[f\left(\frac{T_{n,i} + X_{n,i}}{s_n}\right)\right] - \mathbb{E}\left[f\left(\frac{T_{n,i} + Z_{n,i}}{s_n}\right)\right]\right| \tag{1}$$

We will now use Taylor's approximation to bound each of the terms in the above sum. For $x, h \in \mathbb{R}$, define

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

Since all deriviates of f are bounded, Taylor's approximation with remainder says that there exists $k > 0$ such that for all h and x

$$g(h) \leq k \min\{|h|^3, |h|^2\}.$$

Thus for all $x, h_1, h_2 \in \mathbb{R}$ we have

$$\begin{aligned}\left|f(x+h_1) - f(x+h_2) - f'(x)(h_1 - h_2) - \frac{1}{2} f''(x)(h_1^2 - h_2^2)\right| &= |g(h_1) - g(h_2)| \\ &\leq |g(h_1)| + |g(h_2)|.\end{aligned}$$

We wish to apply this to equation (1) with $x = \frac{T_{n,i}}{s_n}$, $h_1 = \frac{X_{n,i}}{s_n}$ and $h_2 = \frac{Z_{n,i}}{s_n}$. Thus we need to add the high order terms $f'(x)(h_1 - h_2)$ and $\frac{1}{2} f''(x)(h_1^2 - h_2^2)$. Since $X_{n,i}$ and $Z_{n,i}$ have the same mean and variance and $X_{n,i}, Z_{n,i}$ are independent of $T_{n,i}$, we have

$$\begin{aligned}
&\left|\mathbb{E}\left[f\left(\frac{T_{n,i}+X_{n,i}}{s_n}\right)\right]-\mathbb{E}\left[f\left(\frac{T_{n,i}+Z_{n,i}}{s_n}\right)\right]\right| \\
=&\left|\mathbb{E}\left[f\left(\frac{T_{n,i}+X_{n,i}}{s_n}\right)\right]-\mathbb{E}\left[f\left(\frac{T_{n,i}+Z_{n,i}}{s_n}\right)\right]-\mathbb{E}\left[f'\left(\frac{T_{n,i}}{s_n}\right)\left(\frac{X_{n,i}}{s_n}-\frac{Z_{n,i}}{s_n}\right)\right]\right. \\
&\left.-\frac{1}{2}\mathbb{E}\left[f''\left(\frac{T_{n,i}}{s_n}\right)\left(\frac{X_{n,i}^2}{s_n^2}-\frac{Z_{n,i}^2}{s_n^2}\right)\right]\right| \\
\leq&\mathbb{E}\left[g\left(\frac{X_{n,i}}{s_n}\right)+g\left(\frac{Z_{n,i}}{s_n}\right)\right].
\end{aligned}$$

Thus combining this with equation (1), we have

$$\begin{aligned}
\left|\mathbb{E}\left[f\left(\frac{S_n}{s_n}\right)\right]-\mathbb{E}[f(Z)]\right| &\leq \sum_{i=1}^{k_n}\mathbb{E}\left[g\left(\frac{X_{n,i}}{s_n}\right)+g\left(\frac{Z_{n,i}}{s_n}\right)\right] \\
&= \sum_{i=1}^{k_n}\mathbb{E}\left[g\left(\frac{X_{n,i}}{s_n}\right)\right]+\sum_{i=1}^{k_n}\mathbb{E}\left[g\left(\frac{Z_{n,i}}{s_n}\right)\right] \\
&= (I)+(II).
\end{aligned}$$

We will deal with the sum (I) first. Recall that $g(h) \leq k\min\{h^2, h^3\}$. Thus we will split $\mathbb{E}\left[g\left(\frac{X_{n,i}}{s_n}\right)\right]$ into two regions where we will use two different bounds. For each $\varepsilon > 0$, we have

$$\begin{aligned}
(I) &= \sum_{i=1}^{k_n}\mathbb{E}\left[g\left(\frac{X_{n,i}}{s_n}\right)\right] \\
&= \sum_{i=1}^{k_n}\int_{\{X_{n,i}\leq\varepsilon s_n\}} g\left(\frac{X_{n,i}}{s_n}\right)d\mathbb{P}+\sum_{i=1}^{k_n}\int_{\{X_{n,i}>\varepsilon s_n\}} g\left(\frac{X_{n,i}}{s_n}\right)d\mathbb{P} \\
&\leq k\sum_{i=1}^{k_n}\int_{\{X_{n,i}\leq\varepsilon s_n\}}\left|\frac{X_{n,i}}{s_n}\right|^3 d\mathbb{P}+k\sum_{i=1}^{k_n}\int_{\{X_{n,i}>\varepsilon s_n\}}\left|\frac{X_{n,i}}{s_n}\right|^2 d\mathbb{P} \\
&= \frac{k}{s_n^2}\sum_{i=1}^{k_n}\int_{\{X_{n,i}\leq\varepsilon s_n\}}\left|\frac{X_{n,i}}{s_n}\right||X_{n,i}|^2 d\mathbb{P}+\frac{k}{s_n^2}\sum_{i=1}^{k_n}\int_{\{X_{n,i}>\varepsilon s_n\}}|X_{n,i}|^2 d\mathbb{P} \\
&\leq \frac{k\varepsilon}{s_n^2}\sum_{i=1}^{k_n}\int_{\{X_{n,i}\leq\varepsilon s_n\}}|X_{n,i}|^2 d\mathbb{P}+\frac{k}{s_n^2}\sum_{i=1}^{k_n}\int_{\{X_{n,i}>\varepsilon s_n\}}|X_{n,i}|^2 d\mathbb{P} \\
&\leq \frac{k\varepsilon}{s_n^2}\sum_{i=1}^{k_n}\mathrm{Var}(X_{n,i})+\frac{k}{s_n^2}\sum_{i=1}^{k_n}\int_{\{X_{n,i}>\varepsilon s_n\}}|X_{n,i}|^2 d\mathbb{P} \\
&= k\varepsilon+\frac{k}{s_n^2}\sum_{i=1}^{k_n}\int_{\{X_{n,i}>\varepsilon s_n\}}|X_{n,i}|^2 d\mathbb{P}.
\end{aligned}$$

By Lindeberg's condtion we have that the second term goes to zero for all $\varepsilon > 0$. Thus

$$\overline{\lim_n}\sum_{i=1}^{k_n}\mathbb{E}\left[g\left(\frac{X_{n,i}}{s_n}\right)\right] \leq k\varepsilon.$$

And so the sum (I) goes to 0 as n goes to infinity. For sum (II), we can get a similar bound. If we split each expectation into two regions, then we again get

$$(II) = \sum_{i=1}^{k_n} \mathbb{E}\left[g\left(\frac{Z_{n,i}}{s_n}\right)\right] \leq k\varepsilon + \frac{k}{s_n^2}\sum_{i=1}^{k_n}\int_{\{Z_{n,i}>\varepsilon s_n\}} |Z_{n,i}|^2 d\mathbb{P}.$$

Note that it thus suffices to prove that $\{Z_{n,i}\}$ satisfy Lindeberg's condtion. This is because if $\{Z_{n,i}\}$ satisfies Lindeberg's condtion, then sum (II) will go to zero by the same argument we used for (I).

We have only assumed that Lindeberg's condition holds of $\{X_{n,i}\}$ but we can prove this implies Lindeberg's condition holds for $\{Z_{n,i}\}$. Note that for all $\varepsilon' > 0$ we have

$$\begin{aligned}
\frac{\sigma_{n,i}^2}{s_n^2} &= \frac{1}{s_n^2}\int |X_{n,i}|^2 d\mathbb{P} \\
&= \frac{1}{s_n^2}\int_{\{|X_{n,i}|\leq\varepsilon' s_n\}} |X_{n,i}|^2 d\mathbb{P} + \frac{1}{s_n^2}\int_{\{|X_{n,i}|>\varepsilon' s_n\}} |X_{n,i}|^2 d\mathbb{P} \\
&\leq (\varepsilon')^2 + \frac{1}{s_n^2}\int_{\{|X_{n,i}|>\varepsilon' s_n\}} |X_{n,i}|^2 d\mathbb{P}.
\end{aligned}$$

Thus, by Lindeberg's condtion on $\{X_{n,i}\}$, we have

$$\overline{\lim_n} \max_{1\leq i\leq k_n} \left\{\frac{\sigma_{n,i}^2}{s_n^2}\right\} \leq \varepsilon'.$$

It follows that $\max_{1\leq i\leq k_n}\left\{\frac{\sigma_{n,i}}{s_n}\right\}$ goes to zero as n goes to infinity. Let $Z \sim \mathcal{N}(0,1)$. To show that $\{Z_{n,i}\}$ satisfy Lindeberg's condition let $\varepsilon > 0$ be given. We then have

$$\begin{aligned}
\frac{1}{s_n^2}\sum_{i=1}^{k_n}\int_{\{Z_{n,i}>\varepsilon s_n\}} |Z_{n,i}|^2 d\mathbb{P} &= \frac{1}{s_n^2}\sum_{i=1}^{k_n}\int_{\{\sigma_{n,i}|Z|>\varepsilon s_n\}} |\sigma_{n,i}Z|^2 d\mathbb{P} \\
&\leq \frac{1}{s_n^2}\sum_{i=1}^{k_n}\int_{\{\sigma_{n,i}|Z|>\varepsilon s_n\}} |\sigma_{n,i}Z|^2 \frac{\sigma_{n,i}|Z|}{\varepsilon s_n} d\mathbb{P} \\
&\leq \frac{\mathbb{E}[|Z|^3]}{\varepsilon s_n^3}\sum_{i=1}^{k_n}\sigma_{n,i}^3 \\
&\leq \frac{\mathbb{E}[|Z|^3]}{\varepsilon s_n^3}\max_{1\leq i\leq k_n}\{\sigma_{n,i}\}\sum_{i=1}^{k_n}\sigma_{n,i}^2 \\
&= \frac{\mathbb{E}[|Z|^3]}{\varepsilon}\max_{1\leq i\leq k_n}\left\{\frac{\sigma_{n,i}}{s_n}\right\} \\
&\to 0.
\end{aligned}$$

Thus $\{X_{n,i}\}$ satisfies Lindeberg's condtion and we are done. □

3 Comments

3.1 The main idea and generalizations

In some sense our proof was elementary but the idea is *very general*. Any random variable Z can be expressed as

$$Y = U(X_1, \ldots, X_n),$$

where X_i are independent and U is a function. If U is smooth and we have bounds on the derivatives of U, then Y will be close to

$$U(Z_1, \ldots, Z_n),$$

where $Z_1, \ldots, Z_n$ are independent normal. Sourav Chatterjee writes about this in "A generalization of Lindeberg's principle" in the Annals of Probability. The only property of the normal distribution that we really used was that normals have finite third moment and that sums of independent normals are normal.

3.2 Comments on the theorem and proof

There is a converse to Lindeberg's theorem (Lindeberg-Feller). Suppose $s_n \to \infty$ and $\frac{\sigma_{n,i}}{n} \to 0$. Under this assumption, if $\frac{S_n}{s_n}$ converges weakly to a normal random variable, then Lindeberg's condtion holds. This is proved in the textbook.

Our version of the central limit theorem is a limit theorem. It does not have an error bound or tell us anything about finite n. It is possible to get explicit error bounds and Lindeberg did do this. For notation, define

$$S(x) = \begin{cases} x^3 & \text{if } |x| \leq 1, \\ x^2 & \text{if } |x| \geq 1. \end{cases}$$

There exists a constant $C > 0$ such that for all $X_1, X_2, \ldots$ independent with mean 0 and variance σ_i^2, then

$$\sup_{x \in \mathbb{R}} \left| \mathbb{P}\left(\frac{S_n}{s_n} \leq x\right) - \Phi(x) \right| \leq C \left(\sum_{i=1}^{n} l_i \right)^{1/4},$$

where $l_i = \mathbb{E}\left[S\left(\frac{X_n}{s_n}\right)\right]$. See S.D. Chatterji "Lindeberg's central limit theorem à la Hausdorff."

The "right" convergence rate was proved in the Berry-Essen theorem.

Theorem 2. *The exists a constant $C \in (0.4097, 0.4748)$ such that if X_i are i.i.d. with mean μ, variance σ^2 and finite third moment, then*

$$\sup_{x \in \mathbb{R}} \left| \mathbb{P}\left(\frac{S_n}{s_n} \leq x\right) - \Phi(x) \right| \leq \frac{c\mathbb{E}[|X_1|^3]}{\sigma^3 \sqrt{n}}.$$

Thus when X_i has a third moment, the convergence is at a rate of $\frac{1}{\sqrt{n}}$.

3.3 The noraml heuristic

In our central limit thoerem the rows were independent. In many cases if $X_{n,i}$ are not too wild and not too dependent, then $\frac{S_n}{s_n} \implies \mathcal{N}(0,1)$. This is the normal heuristic and multiple examples were discussed previously.

3.4 Different proof techniques

There are many ways to prove the central limit theorem. For example

(a) Lindeberg coupling (our proof).

(b) Stein's method (showing $\mathbb{E}[Wf(W)] \approx \mathbb{E}[f'(W)]$ where $W = \frac{S_n}{s_n}$).

(c) The method of moments (Laplace's proof). We need the result:

Theorem 3. *If Q_n are a sequence of probabilities on $\mathbb{R}$ and*

$$\int x^j Q_n(dx) \to \mathbb{E}[Z^j],$$

for $j = 1, 2, 3, \ldots$, then $Q_n((-\infty, x]) \to \Phi(x)$ for all x.

These ideas are used in physics.

(d) Fourier analysis and characteristic functions (we will discuss these ideas next week).

(e) Entropy. On $\mathbb{R}$ the distribution with a fixed variance σ^2 and the largest entropy is $\mathcal{N}(0, \sigma^2)$ where the entropy of a distribution Q with density q is defined to be

$$\mathbb{E}_Q[-\log(q(X))].$$

Convoluting two distributions increases entropy. Thus $\frac{S_n}{s_n}$ has increasing entropy and fixed variance and so it should be approaching the normal distribution. Turning this idea into a proof is Linnik's argument.