Lecture 15: The central limit Thm

The central limit theorem is most likely one of the most fundamental results in probability and statistics.

It is traditionally taught with Fourier methods (i.e. characteristic functions) but, in a way, this method doesn't really extend well for things like martingales, etc.

The method of proof we will give is based on Lindeberg's proof & has proven to be much more powerful and conceptual.

Taylor with remainder for C_c^∞ functions

For $x \in \mathbb{R}$ and $f \in C_c^\infty(\mathbb{R})$ Taylor's Theorem gives

$$f(x+\Delta x) = f(x) + f'(x)\Delta x + \frac{f''(x)}{2}\Delta x^2 + R(x,\Delta x)$$

where $R_f(x,\Delta x) = \frac{f'''(x_*)}{3!}\Delta x^3$

since $|R_f(x,\Delta x)| = \underbrace{\left|\frac{f'''(x_*)}{3!}\Delta x\right|}_{\text{this term} \to 0 \text{ as } \Delta x \to 0 \text{ uniformly in } x \text{ since } \|f'''\|_\infty < \infty}\Delta x^2$

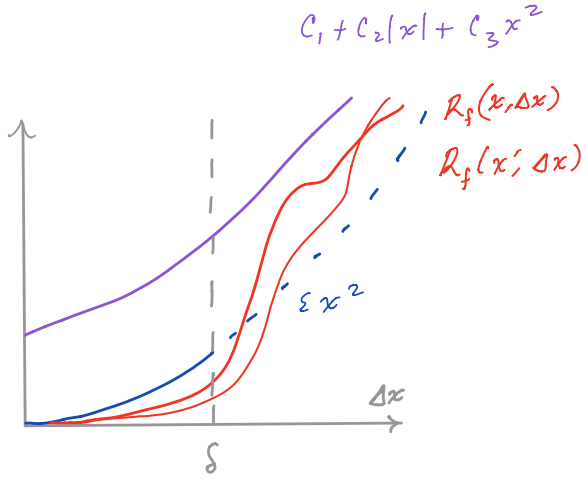
$\therefore \forall \varepsilon > 0 \ \exists \delta > 0$ s.t.

$$|\Delta x| \le \delta \Rightarrow |R_f(x,\Delta x)| \le \varepsilon(\Delta x)^2$$

But we also have $\max(\|f\|_\infty, \|f'\|_\infty, \|f''\|_\infty) < \infty$

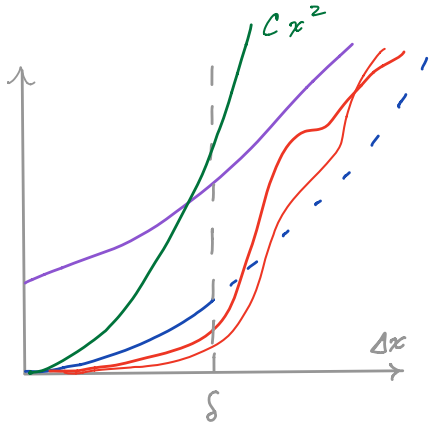
$\therefore |\Delta x| \ge \delta \Rightarrow |R_f(x,\Delta x)| \le C_1 + C_2|\Delta x| + C_3\Delta x^2$

Here is the picture



Notice, however, we can find a C large enough so that

$$x \ge \delta \Rightarrow C_1 + C_2 x + C_3 x^2 \le C x^2$$



$\therefore \forall \varepsilon > 0 \ \exists \delta, C > 0$ s.t.

$$f(x+\Delta x) = f(x) + f'(x)\Delta x + \frac{f''(x)}{2}\Delta x^2 + R_f(x,\Delta x)$$

$$\sup_{x \in \mathbb{R}} |R_f(x,\Delta x)| \le \varepsilon(\Delta x)^2 + C(\Delta x)^2 \mathbb{I}_{\{|\Delta x| \ge \delta\}}$$

Lindeberg's Method for CLT

As a warmup suppose $X_1, X_2, \ldots$ and $Z_1, Z_2, \ldots$ are all independent r.v.s s.t. $EX_i = EZ_i$ &

$$EX_i^2 = EZ_i^2 < \infty$$

Let $g : \mathbb{R}^n \to \mathbb{R}$ be a smooth function and notice

$$\begin{aligned} &|Eg(X_1,\ldots,X_n) - Eg(Z_1,\ldots,Z_n)| \\ &\quad \pm Eg(Z_1, X_2, \ldots X_n) \\ &\quad \pm Eg(Z_1, Z_2, X_3, \ldots, X_n) \\ &\quad \vdots \\ &\quad \pm Eg(Z_1, Z_2, \ldots, Z_{n-1}, X_n) \end{aligned}$$

$$\leq \sum_{i=1}^{n} \left| Eg(\ldots Z_{i-1}, X_i, X_{i+1}, \ldots) - Eg(\ldots Z_{i-1}, Z_i, X_{i+1}, \ldots) \right|$$

$$= \sum_{i=1}^{n} |Eg_i(X_i) - Eg_i(Z_i)|$$

where g_i depends on $\ldots, Z_{i-1}$ & $X_{i+1}, \ldots$

Now let $f \in C_c^\infty(\mathbb{R})$ and look what happens when we set

$$g(x_1,\ldots,x_n) = f(x_1 + \cdots + x_n)$$

$$\therefore g_i(x) = f(\ldots + Z_{i-1} + x + X_{i+1} + \cdots)$$

& $g_i(X_i) - g_i(0)$

$$\begin{aligned} &= f(\ldots + Z_{i-1} + X_i + X_{i+1} + \cdots) \\ &\quad - f(\ldots + Z_{i-1} + 0 + X_{i+1} + \cdots) \\ &= f'(ZX_{(i)})X_i + \frac{f''(ZX_{(i)})}{2}X_i^2 + R_f(ZX_{(i)}, X_i) \end{aligned}$$

where $ZX_{(i)} := (\ldots + Z_{i-1} + X_{i+1} + \cdots)$

We also have

$$\begin{aligned} &g_i(Z_i) - g_i(0) \\ &= f'(ZX_{(i)})Z_i + \frac{f''(ZX_{(i)})}{2}Z_i^2 + R_f(ZX_{(i)}, Z_i) \end{aligned}$$

Therefore

$$\begin{aligned} &Eg_i(X_i) - Eg_i(Z_i) \quad \leftarrow \pm Eg_i(0) \\ &= E\left[f'(ZX_{(i)})(X_i - Z_i)\right] \\ &\quad + E\left[\frac{f''(ZX_{(i)})}{2}(X_i^2 - Z_i^2)\right] \\ &\quad + ER_f(ZX_{(i)}, X_i) - ER_f(ZX_{(i)}, Z_i) \end{aligned}$$

But (X_i, Z_i) is indep of $ZX_{(i)}$ so

$$E\left[f'(ZX_{(i)})(X_i - Z_i)\right] = 0$$

$$E\left[\frac{f''(ZX_{(i)})}{2}(X_i^2 - Z_i^2)\right] = 0$$

since $E(X_i - Z_i) = E(X_i^2 - Z_i^2) = 0$

$\therefore \quad \forall \varepsilon > 0 \ \exists \delta, C > 0$ s.t.

$$|E g_i(X_i) - E g_i(Z_i)|$$

$$\leq E\left|R_f(\sqrt{n}X_{(i)}, X_i)\right| + E\left|R_f(\sqrt{n}X_{(i)}, Z_i)\right|$$

$$\leq \varepsilon E X_i^2 + C E\left(X_i^2 I_{|X_i| \geq \delta}\right)$$

$$+ \varepsilon E Z_i^2 + C E\left(Z_i^2 I_{|Z_i| \geq \delta}\right)$$

Lets put these results together into a lemma.

Lemma 1:

If $X_1, X_2, \ldots$ $Z_1, Z_2, \ldots$ are indep r.v.s s.t. $EX_i = EZ_i$ & $EX_i^2 = EZ_i^2$ and $f \in C_c^\infty(\mathbb{R})$ then $\forall \varepsilon > 0$ $\exists \delta, C > 0$ s.t.

$$\left|E f(X_1 + \cdots + X_n) - E f(Z_1 + \cdots + Z_n)\right|$$

$$\leq 2\varepsilon \sum_{i=1}^n E X_i^2 + C \sum_{i=1}^n E X_i^2 I_{|X_i| \geq \delta} + C \sum_{i=1}^n E Z_i^2 I_{|Z| \geq \delta}$$

Now the key "reason" the limit in the CLT is Gaussian is the following lemma.

Lemma 2:

If $Z, Z_1, \ldots, Z_n \overset{iid}{\sim} N(0, \sigma^2)$ then

$$Z \overset{\mathcal{D}}{=} \frac{Z_1}{\sqrt{n}} + \cdots + \frac{Z_n}{\sqrt{n}}.$$

Theorem: (The central limit theorem)

If $X_1, X_2, \ldots$ are iid r.v.s with $EX_i = 0$ and $EX_i^2 = \sigma^2 < \infty$, then

$$\frac{S_n}{\sqrt{n}} \xrightarrow{\mathcal{D}} Z \sim N(0, \sigma^2)$$

where $S_n := X_1 + \cdots + X_n$.

Proof:

Since $\mathrm{var}\left(\frac{S_n}{\sqrt{n}}\right) = \sigma^2$ chebyshev's Thm implies

$$\left|\frac{S_n}{\sqrt{n}}\right| = \mathcal{O}_p(1) \Longrightarrow \left\{\frac{S_n}{\sqrt{n}}\right\}_{n \geq 1} \text{ is tight}$$

Now $\mathbb{R}$ is locally compact and therefore Portmanteau V implies it will be sufficient to show that $\forall f \in C_c^\infty(\mathbb{R})$

$$\left|E f\left(\frac{S_n}{\sqrt{n}}\right) - E f(Z)\right| \longrightarrow 0$$

$$\Big\| \text{ Lemma 2}$$

$$\left|E f\left(\frac{X_1}{\sqrt{n}} + \cdots + \frac{X_n}{\sqrt{n}}\right) - E f\left(\frac{Z_1}{\sqrt{n}} + \cdots + \frac{Z_n}{\sqrt{n}}\right)\right| \quad (*)$$

($\cdots$: iid $N(0, \sigma^2)$)

Now let $\varepsilon > 0$.

By Lemma 1, $\exists \delta, C > 0$ s.t.

$$(*) \leq 2\varepsilon \sum_{i=1}^n E\frac{X_i^2}{n} + C \sum_{i=1}^n E\frac{X_i^2}{n} I_{|X_i| \geq \sqrt{n}\delta} + C \sum_{i=1}^n E\frac{Z_i^2}{n} I_{|Z_i| \geq \sqrt{n}\delta}$$

$$= 2\varepsilon\sigma^2 + C E X_1^2 I_{|X_1| \geq \sqrt{n}\delta} + C E Z_1^2 I_{|Z_1| \geq \sqrt{n}\delta}$$

since $X_i \sim X_1$, $Z_i \sim Z_1$ & $EX_1^2 = EZ_1^2 = \sigma^2$

Now by the DCT we get

$$E X_1^2 I_{|X_1| \geq \sqrt{n}\delta} \xrightarrow{n\to\infty} 0$$

$$E Z_1^2 I_{|Z_1| \geq \sqrt{n}\delta} \xrightarrow{n\to\infty} 0.$$

Therefore

$$\limsup_{n\to\infty} \left| E f\left(\frac{S_n}{\sqrt{n}}\right) - E f(Z) \right| \leq 2\varepsilon\sigma^2$$

Since $\varepsilon > 0$ was arbitrary

$$\left| E f\left(\frac{S_n}{\sqrt{n}}\right) - E f(Z) \right| \xrightarrow{n\to\infty} 0$$

as was to be shown. QED.

Corollary:

If Z is a r.v. that satisfies

(i) $\text{var}(Z) := \sigma^2 < \infty$

(ii) $Z \overset{d}{=} \frac{Z_1}{\sqrt{n}} + \cdots + \frac{Z_n}{\sqrt{n}}$

where $Z_1, \ldots, Z_n$ are indep copies of Z

then Z must be Gaussian:

$$Z \sim N(0, \sigma^2)$$

Proof:

Properties (i) & (ii) were the only properties of Z we use in the proof of the CLT. By uniqness of distributional limits it must be the only distribution with this property. QED

Theorem: (CLT for random vectors)

If $X_1, X_2, \ldots$ are iid d-dimensional r.v.s s.t.

$$E(X_i) = 0$$

$$E(X_i X_i^T) = \Sigma \in \mathbb{R}^{d\times d}$$

Then

$$\frac{S_n}{\sqrt{n}} \xrightarrow{d} Z \sim N_d(0, \Sigma)$$

where $S_n := X_1 + X_2 + \cdots + X_n$.

Proof:

One can use the same method of proof as in the CLT... but it requires d-dimensional Taylor approximations. Instead lets use the Cramér-Wold device & show

$$\forall k \in \mathbb{R}^d \quad \left\langle k, \frac{S_n}{\sqrt{n}} \right\rangle \xrightarrow{d} \langle k, Z \rangle.$$

Now since

$$\left\langle k, \frac{S_n}{\sqrt{n}} \right\rangle = \frac{\langle k, X_1\rangle}{\sqrt{n}} + \cdots + \frac{\langle k, X_n\rangle}{\sqrt{n}} \quad (\text{iid})$$

where

$$E\langle k, X_i\rangle = \langle k, E X_i\rangle = 0$$

$$E\langle k, X_i\rangle^2 = k^T \Sigma k < \infty$$

$$\langle k, Z\rangle \sim N(0, k^T \Sigma k).$$

the univariate CLT applies.

$$\therefore \quad \left\langle k, \frac{S_n}{\sqrt{n}} \right\rangle \xrightarrow{d} \langle k, Z\rangle \quad \text{as } n\to\infty$$

as was to be shown. QED

Theorem: (Lindeberg CLT generalization)

Let $X_{i,n}$ be a triangular array of independent r.v.s.

X_{11}

X_{12}, X_{22}

X_{13}, X_{23}, X_{33}

$\vdots \qquad \ddots$

which satisfy

(a) $EX_{in} = 0, \quad \forall n \; \forall i \le n$

(b) $\sum_{i=1}^{n} EX_{in}^2 = \sigma^2, \quad \forall n$

(c) $\sum_{i=1}^{n} EX_{in}^2 I_{|X_{in}| \ge \delta} \xrightarrow{n \to \infty} 0, \quad \forall \delta > 0$

Then

$$X_{1n} + \cdots + X_{nn} \xrightarrow[n \to \infty]{\mathcal{D}} N(0, \sigma^2).$$

Proof:

This proof is nearly identical to the proof we gave for the univariate CLT. I'll leave the details for a homework.

QED

Kolmogorov's 3 series Theorem

Stein's Method