# SDS 387
# Linear Models

## Fall 2024

## Lecture 9 - Tue, Sep 24, 2024

Instructor: Prof. Ale Rinaldo

■ CLT : Lindeberg - Feller (LF)

r.v.'s on the same row are indep.

Let $\{X_{j,n}\}$ $1 \leq j \leq n$, $n = 1, 2, \ldots$ be a triangular array

↳ different notation than last time (I wrote $X_{n,j}$)

$n$ : index for $n^{th}$ row of the triangular array

s.t. $\mathbb{E}[X_{j,n}] = 0$ $\forall j, \forall n$ and let

$$B_n^2 = \sum_{j=1}^n \sigma^2_{n,j} \qquad \sigma^2_{n,j} = \mathrm{Var}[X_{n,j}]$$

Then

$$\frac{\sum_{j=1}^n X_{j,n}}{B_n} \xrightarrow{d} N(0,1) \quad \text{if}$$

(LF)

$$\frac{1}{B_n^2} \sum_{j=1}^n \mathbb{E}\left[X_{j,n}^2 \, 1\{|X_{j,n}| > \varepsilon B_n\}\right] \to 0$$

as $n \to \infty$, $\forall \varepsilon > 0$.

- Conversely, if $\frac{\sum_{j=1}^{n} X_{j,n}}{B_n} \to N(0,1)$ and

$$\max_{j=1,\dots,n} \frac{\sigma_{j,n}^2}{B_n^2} \to 0 \quad \text{as } n\to\infty, \text{ then (LF) holds.}$$

- The proof is based on use of ch.f. See, e.g., Petrov.

- A stronger condition than (LF) is Lyapunov's:

$\exists \delta>0$ s.t.

$$\frac{1}{B_n^{2+\delta}} \sum_{j=1}^{n} \mathbb{E}\left[|X_{j,n}|^{2+\delta}\right] \to 0 \quad \text{as } n\to\infty$$

<u>Example</u> $Y_j = \text{Bernoulli}(p_j)$ independent

Let $X_j = Y_j - p_j$ $\left(\mathbb{E}[X_j]=0\right)$.

Under what conditions on the sequence $\{p_j\}$ do we have a CLT?

Use Lyapunov with $\delta=1$ (we control 3rd moment)

Let $\sigma_j^2 = \text{Var}[X_j] = p_j(1-p_j)$. Then

$\mathbb{E}\left[|X_j|^3\right] \le \sigma_j^2$. So

$2+\delta$

$$\frac{\sum_{j=1}^{n} \mathbb{E}\left[|X_j|^3\right]}{B_n^3} \leq \frac{B_n^2}{B_n^3} = \frac{1}{B_n}$$

$$B_n^2 = \sum_{j=1}^{n} \sigma_j^2 = \sum_{j=1}^{n} p_j(1-p_j)$$

This quantity will vanish as $n \to \infty$ if

$$\sum_{j=1}^{n} p_j(1-p_j) \to \infty$$

- The multivariate case: $d > 1$ but fixed (not increasing with $n$!)

Consider a triangular array of centered $d$-dimensional random vectors $\{X_{j,n}\}$ $j \leq n$ s.t. $\mathrm{Var}[X_{j,n}]$ exists. Let

$$Y_{j,n} = \left(\sum_{j=1}^{n} \mathrm{Var}[X_{j,n}]\right)^{-1/2} X_{j,n}$$

If

$$\text{(LF)} \quad \lim_{n\to\infty} \sum_{j=1}^{n} \mathbb{E}\left[\|Y_{j,n}\|^2 \, 1\{\|Y_{j,n}\| > \varepsilon\}\right] = 0 \quad \forall \varepsilon > 0$$

Then

$$\sum_{j=1}^{n} Y_{j,n} \xrightarrow{d} N_d(0, I)$$

Pf/ Use Cramer-Wold device. Let's first consider points $t \in \mathbb{S}^{d-1} = \{x \in \mathbb{R}^d : \|x\| = 1\}$

So we need to show that, $\forall t \in \mathbb{S}^{d-1}$

$$t^T \sum_{j=1}^{n} Y_{j,n} \xrightarrow{d} N(0,1)$$

(if $\|t\| \neq 1$, the limiting distribution will be $N(0, \|t\|^2)$)

First, notice that for $t \in \mathbb{S}^{d-1}$,

$$\sum_{j=1}^{n} \mathrm{Var}\left[t^T Y_{j,n}\right] = 1$$

So, using the LF for univariate CLT we only need to look at the sequence

$$\sum_{j=1}^{n} \mathbb{E}\left[\underbrace{\left(t^T Y_{j,n}\right)^2}_{\leq \underbrace{\|t\|^2}_{=1} \|Y_{j,n}\|^2} \mathbb{1}\left\{ |t^T Y_{j,n}| > \varepsilon \right\}\right]$$

arbitrary $> 0$

and show that it is vanishing as $n \to \infty$.

But this easily follows from bounding the last expression by

$$\sum_{j=1}^{n} \mathbb{E}\left[\|Y_{j,n}\|^2 \, \mathbb{1}\{\|Y_{j,n}\| > \varepsilon\}\right] \to 0$$

by assumption.

Check the calculations for the case $\|t\| \neq 1$. $\blacksquare$

# BERRY-ESSEEN BOUNDS → See Petrov's book

This $\underset{\text{finite sample}}{\wedge}$ result indicates how fast the CLT approximation works.

Let $X_1, X_2, \ldots$ be independent r.v.'s s.t. $\mathbb{E}[X_i] = 0$

$\mathrm{Var}[X_i] < \sigma_i^2 < \infty$ and $\mathbb{E}[|X_i|^3] < \infty$, all $i$.

↳ stronger condition than (LF)

Then

$$\sup_{x \in \mathbb{R}} \left| \mathbb{P}\left( \frac{\sum_{i=1}^n X_i}{B_n} \le x \right) - \Phi(x) \right| \le C \frac{\sum_{i=1}^n \mathbb{E}[|X_i|^3]}{B_n^{3/2}}$$

$B_n = \sqrt{\sum_{i=1}^n \sigma_i^2}$

$\Phi$: cdf of $N(0,1)$

$C$: universal constant $< \frac{1}{2}$

- When $\sigma_i^2 = \sigma^2$ all $i$ and $\mathbb{E}[|X_i|^3] \le \mu_3$ all $i$ then the RHS of this bound is upper bounded by

$$C \frac{n \mu_3}{n^{3/2} \sigma^{3/2}} = \frac{C}{\sqrt{n}} \frac{\mu_3}{\sigma^3}$$

- Example: $X_1, X_2, \ldots$ independent with $X_i \sim \text{Bernoulli}(p_i)$

where $p_i \in [\varepsilon, 1-\varepsilon]$ all $i$.

Then, we saw that,

$$\mathbb{E}[|X_i - p_i|^3] \le p_i(1-p_i)$$ so the

Berry-Esseen bound is

$$\frac{\sum_{i=1}^{n} p_i(1-p_i)}{\left(\sum_{i=1}^{n} p_i(1-p_i)\right)^{3/2}} \leq \frac{1}{\sqrt{n\,\varepsilon(1-\varepsilon)}}$$

We can allow $\varepsilon = \varepsilon_n$ to also depend on $n$ : $\varepsilon_n \to 0$

CLT works as long as $n\,\varepsilon_n(1-\varepsilon_n) \to \infty$

So $\varepsilon_n$ can $\downarrow 0$ but slower than $1/n$.

- Weakest version of Berry - Esseen is (see Petrov).

  Assume only that $\mathbb{E}\left[|X_i|^{2+\delta}\right] < \infty$ all $i$, $\delta \in (0,1]$

  Then

$$\sup_{x\in\mathbb{R}} \left| \mathbb{P}\left( \frac{\sum X_i}{B_n} \leq x \right) - \Phi(x) \right| \leq C \frac{\sum_{i=1}^{n} \mathbb{E}\left[|X_i|^{2+\delta}\right]}{B_n^{1+\delta/2}}$$

- General way of thinking about CLT: replace each $X_i$ by a $Z_i \sim N(\mathbb{E}[X_i], \mathrm{Var}[X_i])$

  Then $\underset{=}{\text{well-behaved}}_n$ functions of the $X_i$'s have a distribution that is close to that of the same function of the $Z_i$'s.

- In general we can frame this task (i.e. the task of demonstrating a Gaussian approximation) as follows.

Let $\mathcal{F}$ be a class of functions. Then we want to bound:

$$\sup_{f \in \mathcal{F}} \left| \mathbb{E}\left[f(\xrightarrow{\text{independent}} X_1, \dots, X_n)\right] - \mathbb{E}\left[f(\xrightarrow{\text{independent}} Z_1, \dots, Z_n)\right] \right|$$

where $Z_n \sim N(\mathbb{E}[X_n], \mathrm{Var}[X_n])$

For example, take $\mathcal{F} = \{ 1(-\infty, x],\ x \in \mathbb{R} \}$ and bound

$$\sup_{f \in \mathcal{F}} \left| \mathbb{E}\left[f\left(\frac{\sum X_n}{B_n}\right)\right] - \mathbb{E}[f(Z)] \right|$$

Berry - Esseen bound

- We can take $\mathcal{F}$ to be any class of functions we like.