Teorema central del límite: Lindeberg

1 Introducción

Recordemos que una sucesión de variables aleatorias $\{X_n,\ n \geq 1\}$ converge en distribución (o debilmente) a una variable aleatoria X si

$$\lim_{n\to\infty} F_n(x) = F(x)\ \forall x \text{ de continuidad de F}, \tag{1}$$

siendo F_n la función de distribución de X_n y F la de X. La definición (1) la podemos reescribir como

$$\lim_{n\to\infty} \mathbb{E}(1_{(-\infty,x]}(X_n)) = \mathbb{E}(1_{(-\infty,x]}(X))\ \forall x \text{ de continuidad de F}. \tag{2}$$

Para probar el TCL comenzaremos por dar otras caracterizaciones equivalentes de la convergencia en distribución.

Theorem 1.1. *Sean $\{X_n : n \geq 1\}$ y X variables aleatorias.*

a) Si $\mathbb{E}(f(X_n)) \to \mathbb{E}(f(X))$, $\forall f$ continua y acotada, entonces X_n converge en distribución a X.

b) Si $\mathbb{E}(f(X_n)) \to \mathbb{E}(f(X))$, $\forall f$, continua, con derivadas continuas y acotadas de todo orden, entonces X_n converge en distribución a X.

Proof. a) Para todo $\epsilon > 0$ podemos aproximar $1_{(-\infty,x]}(t)$ por una función continua y acotada dada por

$$f_\epsilon(t) = \begin{cases} 1 & \text{si } t \leq x \\ 1 - \frac{t-x}{\epsilon} & \text{si } x \leq t \leq x+\epsilon \\ 0 & \text{si } t > x+\epsilon. \end{cases}$$

Tenemos por un lado que para todo $\epsilon > 0$

$$\limsup_n F_n(x) = \limsup_n \mathbb{E}(1_{(-\infty,x]}(X_n)) \leq \limsup_n \mathbb{E}(f_\epsilon(X_n)) = \mathbb{E}(f_\epsilon(X)) \leq F(x+\epsilon).$$

Usando

$$g_\epsilon(t) = \begin{cases} 1 & \text{si } t \leq x - \epsilon \\ 1 - \frac{t-x+\epsilon}{\epsilon} & \text{si } x - \epsilon \leq t \leq x \\ 0 & \text{si } t > x, \end{cases}$$

obtenemos de modo análogo

$$\liminf_n F_n(x) = \liminf_n \mathbb{E}(1_{(-\infty,x]}(X_n)) \geq \liminf_n \mathbb{E}(g_\epsilon(X_n)) = \mathbb{E}(g_\epsilon(X)) \geq F(x-\epsilon).$$

Luego, $\lim_n F_n(x) = F(x)$ si F es continua en x.

b) Basta considerar una función análoga cambiandola entre x y $x+\epsilon$ por una función regular que cumpla con las condiciones. Por ejemplo definimos

$$\psi_\epsilon(t) = \begin{cases} 1 & \text{si } t \leq x \\ \psi(\frac{t-x}{\epsilon}) & \text{si } x \leq t \leq x + \epsilon \\ 0 & \text{si } t > x + \epsilon, \end{cases}$$

con

$$\psi(t) = \frac{1}{a}\int_t^1 e^{-1/s(1-s)}ds,\ a = \int_0^1 e^{-1/s(1-s)}ds.$$

Luego argumentar como antes. Terminar como ejercicio. □

2 Teorema central del límite. Lindeberg

En muchos problemas reales se puede suponer que los errores de medición son una suma de varios errores independientes pequeños debidos a distintos factores que contribuyen al error total. En ese marco es que se supone que los errores son aproximadamente normales. Sin embargo, para que esto se cumpla, los pequeños errores deben cumplir con condiciones muy precisas dadas por el siguiente teorema.

La siguiente demostración - debida a Lindeberg - muestra que si se cumple la condición de Lindeberg dada en (3) y las variables son independientes al sumarlas y normalizarlas adecuadamente la suma de ellas (la convolución de sus distribuciones) converge a una distribución normal.

Theorem 2.1. *Tenemos un sistema triangular de variables aleatorias, o sea para cada n tenemos variables $X_{n1}, \ldots, X_{nk_n}$ con varianza finita σ_{nj}^2, $j = 1, \ldots k_n$ (con cada n puedo cambiarlas todas) que verifican:*

1. *Son independientes dentro de cada fila*

2. $\mathbb{E}(X_{nj}) = 0$ *(todas tienen media cero)*

3. *Sean* $S_n = X_{n1} + \ldots, X_{nk_n}$, $s_n^2 = var(S_n) = \sum_{j=1}^{k_n} \sigma_{nj}^2$. *Si se cumple la condición de Lindeberg dada por*

$$\frac{1}{s_n^2} \sum_{j=1}^{k_n} \mathbb{E}(X_{nj}^2 1_{|X_{nj}|>\epsilon s_n}) \to 0, \ \forall \epsilon > 0, \tag{3}$$

$\frac{S_n}{s_n}$ *converge en distribución a* Z *con distribución* $N(0,1)$.

Proof. Basta probar que $\mathbb{E}(f(\frac{S_n}{s_n})) \to \mathbb{E}(f(Z))$ para toda función f continua, acotada y con derivadas continuas y acotadas.

Fijemos f en la clase y definamos

$$g(h) = \sup_x |f(x+h) - f(x) - f'(x)h - f''(x)h^2/2| \leq Cmin(h^2, h^3). \tag{4}$$

(Verificar usando la fórmula de Taylor).

Truco: usaremos h^2 para h "grande" y h^3 para h pequeño.

Como f está fija en el argumento que sigue, también lo está la constante C. De la definición de g tenemos que

$$|[f(x+h_1)-f(x+h_2)]-[f'(x)(h_1-h_2)+\frac{1}{2}f''(x)(h_1^2-h_2^2)]| \leq g(h_1)+g(h_2). \tag{5}$$

Para cada n sean $Z_{n1}, \ldots, Z_{nk_n}$ variables aleatorias normales con media 0 y la misma varianza que las originales, $var(Z_{nj}) = \sigma_{nj}^2$, que verifiquen que

$$\{X_{n1}, \ldots, X_{nk_n}, Z_{n1}, \ldots Z_{nk_n}\}, \tag{6}$$

son independientes.

La idea de la demostración es ir cambiando de a una las variables originales por las normales, es decir pasando de

$$\mathbb{E}(f(\frac{X_{n1} + \ldots, X_{nk_n}}{s_n})) \text{ a } \mathbb{E}(f(\frac{X_{n1} + \ldots, X_{n(k_n-1)} + Z_{nk_n}}{s_n})),$$

y así sucesivamente hasta terminar en $\mathbb{E}(f(Z))$ con $Z \sim N(0,1)$.

$$\begin{aligned} &\mathbb{E}(f(\frac{X_{n1} + \ldots, X_{nk_n}}{s_n})) \\ &\mathbb{E}(f(\frac{X_{n1} + \ldots, X_{n(k_n-1)} + Z_{nk_n}}{s_n})) \\ &\ldots \\ &\ldots \\ &\mathbb{E}(f(\frac{Z_{n1} + \ldots, Z_{nk_n}}{s_n})) = \mathbb{E}(f(Z))) \end{aligned}$$

con todas normales independientes y ver que en cada paso la esperanza cambia poco. Como la combinación de normales independientes es normal el resultado es válido para las normales.

Si

$$U_{nj} = \sum_{i=1}^{j-1} X_{ni} + \sum_{i=j+1}^{k_n} Z_{ni}, \quad 1 \leq j \leq k_n,$$

como

$$U_{nk_n} + X_{nk_n} = S_n \ \text{ (todas X's) y , } \ U_{n1} + Z_{n1} \sim N(0, s_n^2).$$

Cambiando de a una con una suma telescópica tenemos que

$$|\mathbb{E}(f(\frac{S_n}{s_n}) - \mathbb{E}(f(Z))| \leq \sum_{j=1}^{k_n} |\mathbb{E}(f(\frac{U_{nj} + X_{nj}}{s_n}) - f(\frac{U_{nj} + Z_{nj}}{s_n})|. \tag{7}$$

Si tomamos en la ecuación (5)

$$x = \frac{U_{nj}}{s_n}, h_1 = \frac{X_{nj}}{s_n}, h_2 = \frac{Znj}{s_n},$$

como por la independencia de las variables en (6) tenemos que

1. $$\mathbb{E}(f'(\frac{U_{nj}}{s_n})(\frac{X_{nj} - Z_{nj}}{s_n})) = \mathbb{E}(f'(\frac{U_{nj}}{s_n}))\frac{1}{s_n}\mathbb{E}(X_{nj} - Z_{nj}) = 0,$$

   (independencia + media cero)

2. $$\mathbb{E}(f''(\frac{U_{nj}}{s_n})(\frac{X_{nj}^2}{s_n^2} - \frac{Z_{nj}^2}{s_n^2})) = \mathbb{E}(f''(\frac{U_{nj}}{s_n}))\frac{1}{s_n^2}\mathbb{E}(X_{nj}^2 - Z_{nj}^2) = 0,$$

   (independencia + igual varianza),

resulta que

$$|\mathbb{E}(f(\frac{S_n}{s_n})) - \mathbb{E}(f(Z))| \leq \sum_{j=1}^{k_n} \mathbb{E}\left(g(\frac{X_{nj}}{s_n}) + g(\frac{Z_{nj}}{s_n})\right). \tag{8}$$

Por tanto bastará con probar que

A)

$$\sum_{j=1}^{k_n} \mathbb{E}\left(g(\frac{X_{nj}}{s_n})\right) \to 0$$

B)

$$\sum_{j=1}^{k_n} \mathbb{E}\left(g(\frac{Z_{nj}}{s_n})\right) \to 0$$

Demostración de A)

Dado $\epsilon > 0$, descomponemos

$$\sum_{j=1}^{k_n} \mathbb{E}\left(g(\frac{X_{nj}}{s_n})\right) = I + II,$$

con

$$I = \sum_{j=1}^{k_n} \mathbb{E}\left(g(\frac{X_{nj}}{s_n})1_{|X_{nj}|\leq \epsilon s_n}\right)$$

$$II = \sum_{j=1}^{k_n} \mathbb{E}\left(g(\frac{X_{nj}}{s_n})1_{|X_{nj}|> \epsilon s_n}\right)$$

Por un lado tenemos que acotando un $|X_{nj}|$ por ϵs_n,

$$I \leq C\sum_{j=1}^{k_n} \mathbb{E}\left(\frac{|X_{nj}|^3}{s_n^3}1_{|X_{nj}|\leq \epsilon s_n}\right) \leq C\sum_{j=1}^{k_n} \frac{\epsilon s_n}{s_n^3}\mathbb{E}(X_{nj}^2) = \frac{C\epsilon}{s_n^2}\sum_{j=1}^{k_n}\sigma_{nj}^2 = C\epsilon.$$

$$II \leq C\sum_{j=1}^{k_n} \mathbb{E}\left(\frac{X_{nj}^2}{s_n^2}1_{|X_{nj}|> \epsilon s_n}\right) \to 0,$$

por la hipótesis de la condición de Lindeberg.

Demostración de B).

Si hacemos lo mismo y descomponemos en las sumas I y II (ahora con las Z_{nj}) la acotación para el término I es idéntica pues vale tanto para las X_{nj} como para las Z_{nj}.

Para completar la demostración resta entonces probar que las Z_{nj} también cumplen con la condición de Lindeberg, o sea que

$$III = \frac{1}{s_n^2}\sum_{j=1}^{k_n} \mathbb{E}(Z_{nj}^2 1_{|Z_{nj}|> \epsilon s_n}) \to 0. \tag{9}$$

$$III \leq \frac{1}{s_n^2}\sum_{j=1}^{k_n} \mathbb{E}\left(\frac{|Z_{nj}|^3}{\epsilon s_n}1_{|Z_{nj}|> \epsilon s_n}\right) \leq \frac{1}{s_n^2}\sum_{j=1}^{k_n} \mathbb{E}\left(\frac{|Z_{nj}|^3}{\epsilon s_n}\right) = \frac{1}{\epsilon s_n^3}\sum_{j=1}^{k_n} \mathbb{E}(|\sigma_{nj}Z|^3)$$

$$= \frac{1}{\epsilon s_n^3} \sum_{j=1}^{k_n} \sigma_{nj}^3 \mathbb{E}(|Z|^3).$$

Luego hemos reducido el problema a probar que

$$\frac{1}{\epsilon s_n^3} \sum_{j=1}^{k_n} \sigma_{nj}^3 \to 0. \tag{10}$$

$$\frac{1}{\epsilon s_n^3} \sum_{j=1}^{k_n} \sigma_{nj}^3 \leq \max_j \frac{\sigma_{nj}}{s_n} \frac{1}{s_n^2} \sum_{j=1}^{k_n} \sigma_{nj}^2 = \max_j \frac{\sigma_{nj}}{s_n}.$$

Finalmente bastará probar que $\max_j \frac{\sigma_{nj}^2}{s_n^2} \to 0$.

$$\frac{\sigma_{nj}^2}{s_n^2} = \frac{\mathbb{E}(X_{nj}^2)}{s_n^2} = \frac{1}{s_n^2}\mathbb{E}(X_{nj}^2 1_{|X_{nj}|\leq \epsilon s_n}) + \frac{1}{s_n^2}\mathbb{E}(X_{nj}^2 1_{|X_{nj}|>\epsilon s_n}) \leq$$

$$\frac{\epsilon^2 s_n^2}{s_n^2} + \frac{1}{s_n^2}\mathbb{E}(X_{nj}^2 1_{|X_{nj}|>\epsilon s_n}).$$

Luego,

$$\frac{\max_j \sigma_{nj}^2}{s_n^2} \leq \epsilon^2 + \max_j \mathbb{E}(X_{nj}^2 1_{|X_{nj}|>\epsilon s_n}) \leq \epsilon^2 + \frac{1}{s_n^2} \sum_{j=1}^{k_n} \mathbb{E}(X_{nj}^2 1_{|X_{nj}|>\epsilon s_n}) < 2\epsilon^2,$$

si $n \geq n_0$ pues el segundo sumando tiende a cero por la condición de Lindeberg. □

3 Versión clásica del TCL

Corollary 1. *Sean $\{X_n : n \geq 1\}$ variables aleatorias independientes e igualmente distribuídas, $\mathbb{E}(X_i) = 0$, $var(X_i) = \sigma^2 < \infty$. Sea $s_n^2 = n\sigma^2$. Entonces S_n/s_n converge en distribución a $Z \sim N(0,1)$.*

La condición de Lindeberg nos queda

$$\frac{1}{n\sigma^2} \sum_{j=1}^{n} \mathbb{E}(X_j^2 1_{|X_j|>\epsilon\sqrt{n}\sigma}) = \frac{1}{\sigma^2}\mathbb{E}(X_1^2 1_{|X_1|>\epsilon\sqrt{n}\sigma}) \to 0, \tag{11}$$

por convergencia dominada, pues $X_1^2 1_{|X_1|>\epsilon\sqrt{n}\sigma} \leq X_1^2$, $\mathbb{E}(X_1^2) < \infty$ y $\lim_n X_1^2 1_{|X_1|>\epsilon\sqrt{n}\sigma} = 0$.

4 Ejemplos de aplicación

1). Aproximación normal a la Binomial

Como una variable aleatoria $X \sim Bi(n,p)$ es la suma de n variables X_i iid Bernoulli(p),

$$X = \sum_{j=1}^{n} X_j, \;\; \mathbb{E}(X_j) = p, \;\; j = 1, \ldots, n, \;\; var(X_j) = p(1-p),$$

podemos aplicar el TCL.

Supongamos que tiramos 100 veces una moneda equilibrada y salieron 60 veces cara. ¿Debemos sospechar que la moneda no es equilibrada?

$X = S_n = X_1 + \ldots X_n$, $\mathbb{E}(S_n) = 100 \times 0.5 = 50$, $var(S_n) = 100 \times 0.5 \times 0.5 = 2.5$, $\sigma = 5$.

$$P(X \geq 60) = P(X - 50 \geq 60 - 50) = P(\frac{S_n - 50}{5} > 10/5) \sim 1 - \Phi(2) = 0.0228.$$

2) Tamaño de una muestra para una encuesta.

Supongamos que queremos hacer una encuesta para saber que porcentaje de la población de Montevideo en condiciones de votar lo haría por el candidato A.

Supondremos además que el tamaño del universo N respecto del de la muestra n es suficientemente grande de modo que muestrar con o sin reposición es aproximadamente lo mismo.

Luego podemos suponer que cada individuo en la muestra responde

$$X_i = \begin{cases} 1 & \text{si responde que si} \\ 0 & \text{si responde que no} \end{cases}$$

Tenemos que $X_i = 1$ si votaría al candidato A, con $\{X_1, \ldots, X_n\}$ iid. $p = P(X_i = 1) = p$ es la proporción desconocida que queremos estimar.

Queremos cierta precisión en la estimación que podemos imponer, por ejemplo que $P(|\bar{p} - p| < 0.05)$, donde $\bar{p} = 1/n \sum_{i=1}^{n} X_i$.

¿Cual debería ser el tamaño de la muestra para cumplir con dicha condición?

Usando la aproximación normal, tenemos que

$$P(|\bar{p} - p| < 0.05) = P\left(\frac{\sqrt{n}|\bar{p} - p|}{\sqrt{p(1-p)}} < \frac{0.05\sqrt{n}}{\sqrt{p(1-p)}}\right) \sim P\left(|Z| < \frac{0.05\sqrt{n}}{\sqrt{p(1-p)}}\right),$$

con $Z \sim N(0.1)$.

Por otro lado, como $P(|Z| < 1.96) = 0.95$ deberá ser $\frac{0.05\sqrt{n}}{\sqrt{p(1-p)}} \geq 1.96$, o sea que

$$\sqrt{n} \geq 1.96 \frac{\sqrt{p(1-p)}}{0.05},$$

pero no conocemos p. Como $p(1-p) \leq 1/4$, $\sqrt{p(1-p)} \leq 1/2$, luego basta con que $\sqrt{n} \geq \frac{1.96}{0.05} \times 1/2 = 20 \times 1.96/2$, y por tanto $n \geq (19.6)^2 \sim 400$.

Remark 1. *Observemos que en la condición $P(|\bar{p} - p| < 0.05)$ estamos pidiendo un error absoluto y no un error relativo que sería $P(|\bar{p} - p| < 0.05 \times p)$. No es lo mismo decir que $0.6 - 0.05 \leq \hat{p} \leq 0.6 + 0.05$ que $0.07 - 0.05 \leq \hat{p} \leq 0.07 + 0.05$. Para ello necesitamos tamaños muestrales mucho mas grandes.*

En este caso, sería

$$P(|\bar{p} - p| < 0.05 \times p) \sim P\left(|Z| < \frac{0.05 \times p\sqrt{n}}{\sqrt{p(1-p)}}\right),$$

y por tanto $\frac{0.05\sqrt{np}}{\sqrt{(1-p)}} \geq 1.96$, y

$$\sqrt{n} \geq 1.96 \times \sqrt{1-p}/(0.05 \times \sqrt{p}).$$

Si $p = 0.06$ por ejemplo, nos queda $\sqrt{n} \geq 158$ y por tanto $n \geq 25100$. ´