Teorema Central del Límite

Caso 1:

Situación:

Un juego consiste en lanzar dos dados numerados y ganar tantos pesos (en miles) como indique el dado de mayor puntuación.

Preguntas:

1. Si juego 1 vez, ¿cuál es la probabilidad de ganar menos de $4.000?
2. Si juego 10 veces, ¿cuál es la probabilidad de ganar menos de $40000?
3. Si juego 100 veces, ¿cuál es la probabilidad de ganar menos de $400000?

Dado 1 : sale 2
Dado 2 : sale 4 } gana $4000

Si sale (3,3) ⇒ gana $3000
(5,2) ⇒ gana $5000

Posibles resultados:

Primer Lanzamiento	Segundo Lanzamiento					
	1	2	3	4	5	6
1	1	2	3	4	5	6
2	2	2	3	4	5	6
3	3	3	3	4	5	6
4	4	4	4	4	5	6
5	5	5	5	5	5	6
6	6	6	6	6	6	6

36 resultados posibles

Sea X : ganancia (en miles de pesos).

X	p(x)
1	1/36
2	3/36
3	5/36
4	7/36
5	9/36
6	11/36

Resulta $E(X) \approx 4.47, \sigma^2 \approx 1.97$

(1) La probabilidad que gane menos de \$4000 si juego 1 sola vez es $P(X<4) = P(X=1)+P(X=2)+P(X=3)$

$$= \frac{1}{36} + \frac{3}{36} + \frac{5}{36} = \frac{1}{4} = 0.25$$

Supongamos ahora que juego 2 veces. ¿Cuál es la probabilidad que gane menos de \$7000?

Sean X_1 = ganancia (en miles) en primer jugada

X_2 = " " " segunda jugada.

La respuesta a la pregunta planteada es

$$P(X_1 + X_2 < 7) = ¿?$$

No conocemos la distribución de $X_1 + X_2$, sin embargo con un poco de trabajo podríamos conocerla

¿Y si quisiera responder las preguntas 2 y 3?

2 $P(X_1 + \dots + X_{10} < 40) = ¿?$

3 $P(X_1 + X_2 + \dots + X_{100} < 400) = ¿?$

donde X_i = ganancia en la i-ésima jugada

Notemos que las v.a X_i son independientes entre sí, y todos tienen la misma distribución: son indep. e idénticamente distribuidos (i.i.d)

Sea $S_n = \sum_{i=1}^{n} X_i = X_1 + X_2 + \dots + X_n$

$S_{10} = X_1 + X_2 + \dots + X_{10}$ $\quad$ $S_{100} = X_1 + X_2 + \dots + X_{100}$

Reescribiendo las preguntas:

(2) $P(S_{10} < 40) = ¿?$ (3) $P(S_{100} < 400) = ¿?$

$$\mu_{S_n} = E(S_n) = E(X_1 + \dots + X_n) = \sum_{i=1}^{n} E(X_i) = \sum_{i=1}^{n} \mu = n\mu$$

$E(X_i) = \mu \quad \forall i = 1, \dots, n$

(son idénticamente distribuidos)

Como son independientes, e id. dist.

$$\sigma^2_{S_n} = Var(S_n) = Var(X_1 + \dots + X_n) = \sum_{i=1}^{n} Var(X_i) = n\sigma^2$$

Luego $\boxed{E(S_n) = n\mu \qquad Var(S_n) = n\sigma^2}$

Conocemos la esperanza y la varianza de S_n, pero no tenemos una expresión o fórmula para la distribución de S_n.

En este ejemplo en particular, como

$\mu = E(X) \simeq 4{,}47$ y $\sigma^2 = Var(X) = 1{,}97$ $\Rightarrow$

$\mu_{S_{10}} = E(S_{10}) = 10 \times \mu \simeq 44{,}7$ $\quad \sigma^2_{S_{10}} = Var(S_{10}) = 10 \times \sigma^2 \simeq 19{,}7$

$\mu_{S_{100}} = E(S_{100}) = 100 \times \mu \simeq 447$ y $\sigma^2_{S_{100}} = Var(S_{100}) = 197$

Hicimos algunas simulaciones y graficamos el histograma para

a) $X_1 + X_2$

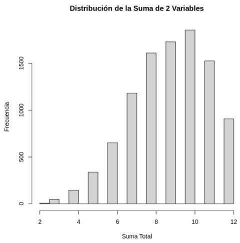
b) $S_{10} = X_1 + X_2 + \dots + X_{10}$

c) $S_{100} = X_1 + \dots + X_{100}$

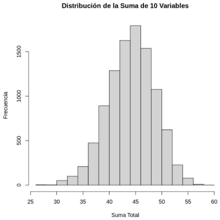
$\left\{ \right.$ X_i i.i.d con la dist. del ejemplo

No conocemos las distribuciones para $S_2 = X_1 + X_2$, $S_{10} = X_1 + \dots + X_{10}$ ni para $S_{100} = X_1 + X_2 + \dots + X_{100}$.

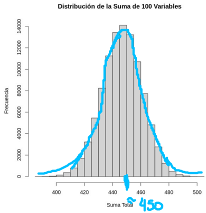
Pero sus histogramas serían algo así (para 10000 simulaciones)



$S_2 = X_1 + X_2$



$S_{10} = X_1 + \dots + X_{10}$



$S_{100} = X_1 + \dots + X_{100}$

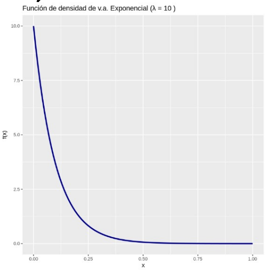
Observamos que la distribución de S_2 es asimétrica, con cola hacia la izquierda. La dist. de S_{10} es ligeramente asimétrica, también con cola pesada hacia la izq.

Sin embargo la dist. de S_{100} es más simétrica entorno a su media, y si trazáramos una curva por encima del histograma, se asemejaría bastante al gráfico de una v.a con densidad normal.

$\boxed{\underline{\text{Caso 2}}:}$ Supongamos que $X_i \sim \mathcal{E}(10)$ (distribución exponencial, $\lambda = 10$)

$$\Rightarrow \mu = E(X_i) = \frac{1}{\lambda} = 0{,}1 \; ; \; \sigma^2 = Var(X_i) = \frac{1}{\lambda^2} = 0{,}01$$

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{si } x \geq 0 \\ 0 & \text{c.c} \end{cases}$$



Sean $X_1, \dots, X_n$ v.a. i.i.d (variables aleatorias indep. e idénticamente distribuidas)

$$X_i \sim \mathcal{E}(10) \quad \forall i=1,\dots,n \implies \begin{cases} \mu_{X_i} = \mu = E(X_i) = \frac{1}{\lambda} & (*) \\ \sigma^2_{X_i} = \sigma^2 = Var(X_i) = \frac{1}{\lambda^2} \end{cases}$$

$$\text{Sea } S_n = \sum_{i=1}^{n} X_i \implies \begin{cases} \mu_{S_n} = E(S_n) = n\mu \overset{(*)}{=} n/\lambda \\ \sigma^2_{S_n} = Var(S_n) = n\sigma^2 = n/\lambda^2 \end{cases}$$

$$S_2 = X_1 + X_2$$

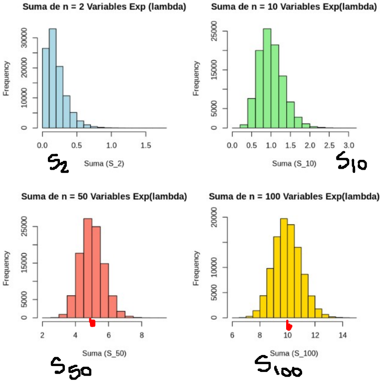
$$E(S_2) = \frac{2}{10} = 0{,}2$$

$$\sigma^2_{S_2} = \frac{2}{100} = 0{,}02$$

$$\begin{cases} S_{10} = \sum_{i=1}^{10} X_i \\ E(S_{10}) = \frac{10}{10} = 1 \\ \sigma^2_{S_{10}} = \frac{10}{100} = 0{,}1 \end{cases}$$

$$\begin{cases} S_{100} = \sum_{i=1}^{100} X_i \\ \mu_{S_{100}} = E(S_{100}) = \frac{100}{10} = 10 \\ \sigma^2_{S_{100}} = Var(S_{100}) = \frac{100}{100} = 1 \end{cases}$$

Como los X_i son v.a.i.i.d $\Rightarrow$ podemos calcular de manera exacta μ_{S_n} y $\sigma^2_{S_n}$. Sin embargo NO sabemos, en general, cuál es la distribución de S_n. Recurriendo nuevamente a simulaciones obtenemos los siguientes gráficos:



Observaciones:

Si trazáramos una línea por encima de los histogramas, vemos que para S_2 y S_{10} los "densidades" son asimétricos, con colas pesados hacia la derecha.

Para $n=50$ y $n=100$ los "densidades" se asemejan a la de una v.a con densidad normal. Observamos además que en estos dos últimos casos, parecen ser simétricos alrededor de un número cercano a las respectivas medias $\mu_{S_{50}} = 5$ y $\mu_{S_{100}} = 10$ (recordemos Ley de los grandes números: $P(|\bar{X} - \mu| > \varepsilon) \xrightarrow[n\to\infty]{} 0$ o equivalentemente: $P\left(|S_n - n\mu| \leq n\varepsilon\right) \xrightarrow[n\to\infty]{} 1$

Conclusiones (empíricas):

Analizamos dos casos:

$\begin{cases} \text{Caso 1: Suma de v.a.i.i.d. discretas} \\ \text{Caso 2: Suma de v.a.i.i.d. continuas} \end{cases}$

En ambos casos pudimos observar que para n suficientemente grande, la distribución de la suma S_n se "asemeja" a la de una v.a con distribución Normal, con media μ_{S_n} (y varianza $\sigma^2_{S_n}$). Esto es, intuitivamente, lo que formalmente se enuncia como Teorema Central del límite (y cuya prueba es una demostración matemática y rigurosa)

Aplicación:

Recordemos el caso 1: se lanzan dos dados numerados y se ganan tantos pesos (en miles) según sea la puntuación del dado de mayor numeración.

Pregunta 3: Si juego 100 veces, ¿cuál es la probabilidad de ganar menos de $400000?

$$P(X_1 + \dots + X_{100} < 400) = P(S_{100} < 400) =$$

$$= P\left(\frac{S_{100} - \mu_{S_{100}}}{\sigma_{S_{100}}} < \frac{400 - \mu_{S_{100}}}{\sigma_{S_{100}}} \right) = \boxtimes$$

Como S_{100} tiene "aproximadamente" dist. normal

$$\mathcal{N}(\mu_{S_{100}}, \sigma^2_{S_{100}})$$

entonces $\dfrac{S_{100} - \mu_{S_{100}}}{\sigma_{S_{100}}}$ tiene, aprox., dist $\mathcal{N}(0,1)$

Entonces:

$$\boxed{*} := P\left(\frac{S_{100} - \mu_{S_{100}}}{\sigma_{S_{100}}} < \frac{400 - \mu_{S_{100}}}{\sigma_{S_{100}}} \right) \approx \Phi\left(\frac{400 - \mu_{S_{100}}}{\sigma_{S_{100}}} \right)$$

$\approx$ es aprox.

$$= \Phi\left(\frac{400 - 447}{\sqrt{197}} \right) = 0{,}0004$$

Φ func. de dist. acum. de v.a. normal estándar

Luego, la prob. de ganar menos de \$400000 si juego 100 veces es aproximadamente: 0,0004

$$P(S_{100} < 400) \approx 0{,}0004$$

Teorema Central del Límite

- *Teorema Central de Límite* o (Teorema del Límite Central): conjunto de teoremas con variaciones acerca del comportamiento de la distribución de la suma (o promedio) de variables aleatorias. En ellos se afirma que, bajo ciertas condiciones, la distribución de probabilidad de la suma de un número "grande" de variables aleatorias es, aproximadamente, una distribución normal.
- Pólya (1920) lo denominó *Teorema "Central" del Límite* por el rol fundamental (central) que cumple este teorema en la Teoría de Probabilidades, ya que, entre otras cosas, justifica porqué en muchas aplicaciones es válido asumir normalidad en las variables o porqué las distribuciones normales son tan comunes.

Un poco de historia

- Primera versión impresa se debe a De Moivre (ppos siglo 18). En su libro "The Doctrine of Chances" aproxima a la distribución binomial (para el caso especial $p = 1/2$) por una curva "suave" que hoy se conoce como "normal"
- Científicos de la época, observaron que muchos fenómenos naturales tenían una distribución aproximadamente normal. En 1809 Gauss desarrolló la fórmula de la distribución normal y mostró que ajustaba perfectamente a la distribución de los errores cometidos en las observaciones astronómicas.
- Laplace generaliza el resultado de De Moivre al caso p arbitrario en lo que hoy se conoce como *Teorema de De Moivre - Laplace*.

- Teorema Central del Límite de Laplace (s. 19): bajo ciertas condiciones la suma de un número considerable de variables aleatorias mutuamente independientes e idénticamente distribuidas puede aproximarse por una normal. Sólo lo demostró para distribuciones discretas y para ciertas distribuciones continuas.
- Primeras demostraciones rigurosas: Tshebyshev (1887), Markov(1898) (momentos) y Liapunov (1901) (funciones características. Liapunov estableció además condiciones de suficiencia (*Condiciones de Lyapunov*).
- Lindeberg (1922), propuso condiciones que son hasta cierto punto, necesarias. Esta demostración se simplifica mediante el uso de funciones características, idea propuesta por Lévy (*Teorema Central del Límite de Lindeberg-Lévy*).
- En 1937 Feller demuestra el *Teorema Central del Límite de Lindeberg-Feller* donde establece las condiciones necesarias para la validez de este resultado.

Distribuciones Muestrales

Sea $X_1, X_2, \ldots, X_n$ muestra aleatoria de v.a.i.i.d. con $E(X_i) = \mu$, $Var(X_i) = \sigma^2$, $\forall i = 1, \ldots, n$.

1. $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ es la media muestral de la muestra $X_1, X_2, \ldots, X_n$.
2. $S_n = \sum_{i=1}^{n} X_i$ es el total muestral de la muestra $X_1, X_2, \ldots, X_n$.

- $\mu_{\bar{X}} = E(\bar{X}) = \mu$
- $\sigma^2_{\bar{X}} = Var(\bar{X}) = \frac{\sigma^2}{n}$
- $\mu_{S_n} = E(S_n) = n\mu$
- $\sigma^2_{S_n} = Var(S_n) = n\sigma^2$

Teorema 9.4: **TCL de Lindenberg-Levy**

Sean $X_1, X_2, \ldots$ variables aleatorias independientes e indénticamente distribuidas, con $E(X_i) = \mu,\ \mathrm{Var}(X_i) = \sigma^2.$ Sea $S_n = X_1 + X_2 + \ldots + X_n.$ Entonces

$$\frac{Sn - n\mu}{\sigma\sqrt{n}} \underset{\rightarrow}{D} N(0,1)$$

Demostración Apunte

Para probar el teorema de Lindenberg - Levy necesitaremos el siguiente resultado que es consecuencias de los teoremas de Helly Bray(necesidad) y de Continuidad de Levy (suficiencia)):

Teorema 9.5

Sean $X, X_1, \ldots$ v.a y $\phi_X, \phi_{X_1}, \ldots$ sus funciones características. Entonces

$$X_n \underset{\rightarrow}{D} X \Leftrightarrow \phi_{X_n}(x) \rightarrow_{n\rightarrow\infty} \phi_X(x)$$

sin demostración

Teorema 9.6 (DeMoivre- Laplace):

Sea S_n el número de éxitos en n ensayos Bernoulli independientes, con probabilidad p de éxito en cada ensayo. Entonces

$$\frac{Sn - np}{\sqrt{np(1-p)}} \xrightarrow{D} N(0,1)$$

Dem: Ejercicio (usar TCL)

El teorema de DeMoivre- Laplace dice que si $X \sim \mathcal{B}(n,p) \Rightarrow Z = \frac{X-np}{\sqrt{np(1-p)}}$ tiene *aproximadamente* distribución normal estándar.

La aproximación es buena siempre que:

- $np > 5$
- $n(1-p) > 5$

Ejemplo 1

En la fabricación de ciertos chips semiconductores se produce un 2% de chips defectuosos. Suponga que los chips son independientes entre sí y que se examina un lote que contiene 1000 chips.
¿Cuál es la probabilidad que más de 25 chips sean defectuosos?

<u>Respuesta</u>
Si $X = n^o$ de chips defectuosos en ese lote $\Rightarrow X \sim \mathcal{B}(1000, 0.02)$

$$P(X > 25) = \sum_{k=25}^{1000} \left(\begin{array}{c} 1000 \\ k \end{array} \right) 0.02^k (0.8)^{1000-k}$$

Aproximaciones de la Distribución Normal

Continuación Ejemplo 1

- $np = 1000 \cdot 0.02 = 20 > 5$
- $n(1-p) = 1000 \cdot 0.98 = 980 > 5$
- Además $\sqrt{np(1-p)} = 4.43$

$\Rightarrow Z = \frac{X-np}{\sqrt{np(1-p)}} = \frac{X-20}{4.43}$ tiene aproximadamente distribución $N(0,1)$

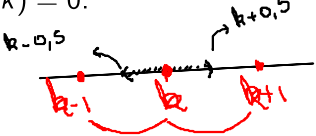
Entonces:

$$P(X > 25) = P\left(\frac{X-20}{4.43} > \frac{25-20}{4.43}\right) \approx P(Z > 1.13) = 1 - \Phi(1.13) = 0.13$$

Correción de la aproximación por la Distribución Normal

Continuación Ejemplo 1 ¿ Cuál es la probabilidad que se encuentren exactamente 25 chips defectuosos? Es decir, $P(X = 25) = ?$

Si quisiéramos aproximar la distribucï¿½n de una va Binomial (X, discreta) por una va Normal (Z, continua), debemos tener en cuenta que Z es continua, entonces $P(Z = k) = 0$.



Supongamos $X \sim \mathcal{B}(n, p)$, $k \in [0, n]$

$$P(X = k) = P(k - 0.5 < X < k + 0.5)$$

X discreta

$$= P\left(\frac{k-0.5-np}{\sqrt{npq}} < \frac{X-np}{\sqrt{npq}} < \frac{k+0.5-np}{\sqrt{npq}}\right)$$

$$\approx \Phi\left(\frac{k+0.5-np}{\sqrt{npq}}\right) - \Phi\left(\frac{k-0.5-np}{\sqrt{npq}}\right)$$

Luego, $P(X = 25) \approx \Phi\left(\frac{25+0.5-20}{4.43}\right) - \Phi\left(\frac{25-0.5-20}{4.43}\right) = 0.048$

La probabilidad exacta, es $P(X = 25) = 0.044$

En general, si se va a aproximar una va X_D discreta, por una va X_C continua debe hacerse una corrección por continuidad

Corrección por continuidad

Dados $a, b \in \mathbb{R}$,

X_D	X_C
$X_D = a$	$a - 0.5 < X_C < a + 0.5$
$X_D < a$	$X_C < a - 0.5$
$X_D \leq a$	$X_C < a + 0.5$
$X_D > a$	$X_C > a + 0.5$
$X_D \geq a$	$X_C > a - 0.5$
$a < X_D < b$	$a + 0.5 < X_C < b - 0.5$
$a \leq X_D < b$	$a - 0.5 < X_C < b - 0.5$
$a < X_D \leq b$	$a + 0.5 < X_C < b + 0.5$
$a \leq X_D \leq b$	$a - 0.5 < X_C < b + 0.5$

Continuación Ejercicio 1:
Usar la corrección a la aproximación por una normal y calcular nuevamente $P(X > 25)$.

1. SIN corrección por continuidad,
   $P(X > 25) \approx P(X_N > 25) = 0.13$
2. CON corrección por continuidad,
   $P(X > 25) \approx P(X_N > 25 + 0.5)$
   $= P(Z > \frac{25.5-20}{4.43}) = 1 - \Phi(\frac{25.5-20}{4.43}) = 0.107$
3. Probabilidad Exacta $P(X > 25) = 0.109$

Ejemplo 2:

Cierto equipo consta de 30 instrumentos electrónicos $C_1, C_2, \dots, C_{30}$ que se usan de la siguiente manera: si C_1 falla, entonces comienza a trabajar C_2, cuando éste falla, recién comienza a trabajar C_3 y así sucesivamente hasta C_{30}. Supongamos que el tiempo de falla $C_i \sim \mathcal{E}(\lambda)$, con $\lambda = 0.1$ por hora. (exponencial de parámetro $\lambda = 0.1$)

$E(C_i) = \frac{1}{\lambda} = 10 \qquad Var(C_i) = \frac{1}{\lambda^2} = 100$

¿Cuál es la probabilidad que el equipo dure ~~más~~ más de 350h?

El tiempo de falla del equipo será la suma de los tiempos de fallos de cada una de los 30 componentes

$$P(\underbrace{C_1 + \dots + C_{30} > 350}_{\text{el equipo dura más de 350 h}}) = P(S_{30} > 350) = \Delta$$

$\mu_{S_{30}} = E(S_{30}) = 30 \cdot E(C_i) = 300$, $\sigma^2_{S_{30}} = Var(S_{30}) = 30\, Var(C_i) = 3000$

$S_{30} \underset{\sim}{\sim} \mathcal{N}(\mu_{S_{30}}, \sigma^2_{S_{30}}) \Rightarrow \dfrac{S_{30} - \mu_{S_{30}}}{\sigma^2_{S_{30}}} \underset{\sim}{\sim} \mathcal{N}(0,1)$

↓ tiene aprox. dist

$$\Delta = P(S_{30} > 350) = P\left(\frac{S_{30} - \mu_{S_{30}}}{\sigma_{S_{30}}} > \frac{350 - \mu_{S_{30}}}{\sigma_{S_{30}}}\right)$$

$$= P\left(\frac{S_{30} - \mu_{S_{30}}}{\sigma_{S_{30}}} > \frac{350 - 300}{\sqrt{3000}}\right) \underset{\sim}{\sim} P\left(Z > \frac{350-300}{\sqrt{3000}}\right)$$

↳ $Z \sim N(0,1)$

$$= 1 - \Phi\left(\frac{50}{\sqrt{3000}}\right) = 0{,}18$$

La prob. que el equipo dure más de 350h es, aproximadamente, 0,18.

NO se hace corrección por continuidad pues los C_i son continuas.

Teorema Central del Límite de Lindeberg Sean $X_1, X_2, \ldots$ variables aleatorias independientes e indénticamente distribuidas, con $E(X_i) = \mu_i,\ \mathrm{Var}(X_i) = \sigma_i^2$ tales que $\sigma_i < \infty$ y al menos un $\sigma_i > 0$. Sea $S_n = X_1 + X_2 + \ldots + X_n$, $s_n = \sqrt{\mathrm{Var(S_n)}} = \sqrt{\sigma_1^2 + \ldots + \sigma_n^2}$. Entonces, si se satisfacen las condiciones de Lindeberg,

$$\frac{Sn - E(S_n)}{s_n} \xrightarrow{D} N(0,1)$$

Condiciones de Lindeberg:

$$\forall \epsilon > 0, \lim_{n \to \infty} \frac{1}{s_n} \sum_{k=1}^{n} \int_{|x-\mu_k| > \epsilon s_n} (x - \mu_k)^2 dF_{X_k}(x) = 0$$

Bibliografía

James B.*Probabilidade: um curso em nivel intermediario* IMPA Rio de Janeiro (1983)
Chung, K.L, *A course in Probability Theory*, Academic Press, 2nd Ed.,(1974)