

Probabilidad

Daniel Fraiman

Maestría en Ciencia de Datos, Universidad de San Andrés

CONVERGENCIA DE VARIABLES ALEATORIAS

Convergencia

DOS FORMAS DE QUE CONVERJA UNA SUCESSION DE v.a.

Definición: convergencia en probabilidad

Un sucesión Z_1, Z_2, Z_3, \dots de v.a. converge en *probabilidad* a Z (otra v.a. ó un número), $Z_n \xrightarrow{P} Z$, si

$$\lim_{n \rightarrow \infty} \mathbb{P}(|Z_n - Z| > \epsilon) = 0 \quad \forall \epsilon > 0$$

Definición: convergencia en distribución

Un sucesión Z_1, Z_2, Z_3, \dots de v.a. converge en *distribución* a Z (otra v.a.), $Z_n \xrightarrow{D} Z$, si

$$\lim_{n \rightarrow \infty} \mathbb{P}(Z_n \leq t) = \mathbb{P}(Z \leq t) \quad \forall t \text{ punto de continuidad}$$

$$\lim_{n \rightarrow \infty} F_{Z_n}(t) = F_Z(t) \quad \forall t \text{ punto de continuidad}$$

LEY DE GRANDES NÚMEROS

Ley de los Grandes Números (LGN)

LEY DE GRANDES NUMEROS → “El promedio de una sucesión/secuencia de variables aleatorias converge a la Esperanza Matemática”

LGN: ¿Un resultado innato o adquirido?

Supongamos que queremos conocer las chances de ganar a un juego, $\mathbb{P}(\text{Ganar}) = p$. Pero hacer el cálculo es muy difícil. ¿Qué harían?

- Jugamos muchas veces al juego, digamos n veces. Si ganamos en el i -ésimo juego $X_i = 1$ y si perdemos $X_i = 0$. Y así tenemos X_1, X_2, \dots, X_n donde $X_i \sim \text{Bernoulli}(p)$. Sabemos que $\mathbb{E}(X_i) = p$ pero desconocemos el valor de p .
- Calculamos $\bar{X}_n = \frac{X_1 + \dots + X_n}{n}$. Sabemos que $\mathbb{E}(\bar{X}_n) = \mathbb{E}(X_i) = p$.
- Finalmente **estimamos a p con \bar{X}_n** . ¿Por qué? ¿Qué pasa a medida que n crece?

Promedio de variables aleatorias

Dadas X_1, X_2, \dots, X_n variables aleatorias independientes tales que

$$E(X_i) = \mu \quad \text{y} \quad \text{Var}(X_i) = \sigma^2 \quad \forall i,$$

definimos la variable aleatoria promedio

$$\bar{X}_n = \frac{X_1 + \dots + X_n}{n}.$$

Recordemos que vimos que

$$E(\bar{X}_n) = \mu \quad \text{y} \quad \text{Var}(\bar{X}_n) = \frac{\sigma^2}{n}.$$

En este caso, no conocemos la distribución exacta de \bar{X}_n pero sabemos que **la esperanza es la misma** y que **la varianza tiende a cero cuando $n \rightarrow \infty$** . Por lo tanto....

Ley de los Grandes Números (LGN)

Ley de los Grandes Números

Sea $\{X_i\}_{i=1}^{\infty}$ una sucesión de variables aleatorias independientes tales que $E(X_i) = \mu$ y $Var(X_i) = \sigma^2$ para todo i . Sea $\bar{X}_n = \frac{X_1 + \dots + X_n}{n}$. Entonces, para todo $\varepsilon > 0$,

$$\lim_{n \rightarrow \infty} \mathbb{P}(|\bar{X}_n - \mu| \geq \varepsilon) = 0$$

es decir

$$\bar{X}_n \xrightarrow{P} \mu$$

“el promedio converge en probabilidad a la esperanza(μ)”

La esperanza es la media de una variable aleatoria

En el infinito, la media aritmética, converge a la esperanza matemática

Los promedios y las sumas, a medida que “n” crece, tienden a tener una distribución que se asemeja a la normal

TEOREMA CENTRAL DEL LÍMITE

Teorema Central del Límite

Preliminares

Sea $\{X_i\}_{i=1}^{\infty}$ una sucesión de variables aleatorias independientes idénticamente distribuidas ($X_i \sim F$) tales que $E(X_i) = \mu$ y $Var(X_i) = \sigma^2$ para todo i . Consideremos:

- La suma de las primeras n variables:

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

Sabemos que $\mathbb{E}(S_n) = n\mu$, y $Var(S_n) = n\sigma^2$

- El promedio de las primeras n variables:

$$\bar{X}_n = \frac{X_1 + \cdots + X_n}{n}.$$

Sabemos que $\mathbb{E}(\bar{X}_n) = \mu$, y $Var(\bar{X}_n) = \frac{\sigma^2}{n}$

La varianza de un promedio es σ^2/n

Teorema Central del Límite


El Teorema Central del Límite (TCL) nos dirá cómo se comportan (qué ley tienen) las variables aleatorias S_n y \bar{X}_n cuando $n \rightarrow \infty$.

Teorema Central del Límite

Teorema Central del Límite

Sea $\{X_i\}_{i=1}^{\infty}$ una sucesión de variables aleatorias independientes e idénticamente distribuidas con $E(X_i) = \mu$ y $Var(X_i) = \sigma^2$ para todo i . Sean $S_n = X_1 + \dots + X_n$ y $Z_n = \frac{S_n - n\mu}{\sqrt{n}\sigma} = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}$. Entonces, para todo $x \in \mathbb{R}$,

$$\lim_{n \rightarrow \infty} F_{Z_n}(x) = \Phi(x).$$

Normal(0, 1) 

o equivalentemente

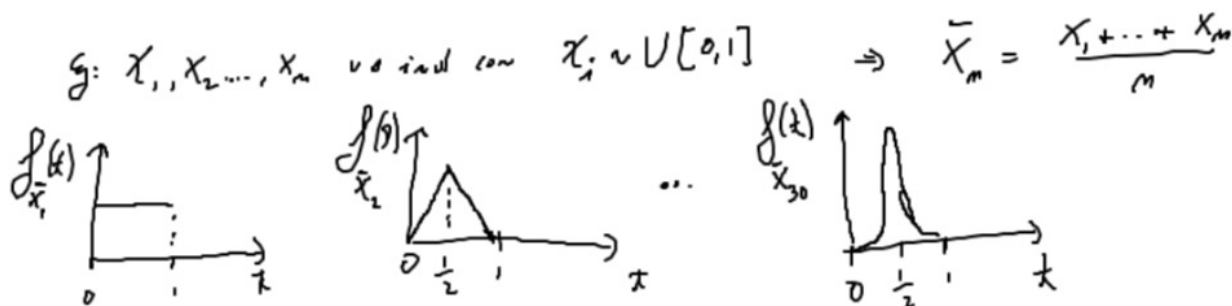
$$Z_n \xrightarrow{D} Z \sim N(0, 1)$$

O sea, el TCL dice que $\lim_{n \rightarrow \infty} \mathbb{P}(Z_n \leq x) = \Phi(x)$. (Recordemos que Φ es la función de distribución acumulada de una variable $N(0, 1)$).

Otra forma de indicarlo es decir que si n es *suficientemente grande*, $P(Z_n \leq x) \approx \Phi(x)$.

Teorema Central del Límite

La distribución de la suma y del promedio de v.a. iid se aproxima a una Normal.



Teorema Central del Límite

Ejemplo S_n

Una compañía aérea modela los pesos de las valijas de sus clientes con variables aleatorias independientes idénticamente distribuidas de media 20 kg y varianza 100. La compañía decide no pesar el equipaje y permite despachar una valija por persona. Si en un avión viajan 400 personas y la bodega soporta 10000 kg, ¿cuál es la probabilidad de sobrecarga?

Como tienen cierta varianza y cierta esperanza y son independientes (las valijas entre sí), entonces vale el TCL

Tenemos X_1, \dots, X_{400} v. a. i. i. d., donde cada X_i representa *peso de la valija i*. Sabemos que $E(X_i) = 20$ y que $Var(X_i) = 100$.

Queremos calcular $P(S_{400} > 10000)$. Como 400 es grande, usamos TCL.

Teorema Central del Límite

Ejemplo S_n

S_{400} = Suma de $X_1 + X_2 + \dots + X_{400}$, donde X_i = peso de la valija i

$$\begin{aligned} P(S_{400} > 10000) &= P\left(\frac{S_{400} - 400 \cdot 20}{\sqrt{400} \cdot 10} > \frac{10000 - 400 \cdot 20}{\sqrt{400} \cdot 10}\right) \\ &= P(Z_{400} > 10) = 1 - P(Z_{400} \leq 10) \\ &\underset{TCL}{\approx} 1 - \Phi(10) \cong 0. \end{aligned}$$

$1 - \text{Pnorm}(10000, (400 \cdot 20), (100 \cdot 400^{1/2}))$
 $E(S_n) = n \cdot \mu = n \cdot 20$
 $Var(sN) = n \cdot \sigma^2 = n \cdot 100$

Teorema Central del Límite

Ejemplo \bar{X}_n

La ganancia semanal de una empresa (en miles de USD) está dada por una variable aleatoria W con $E(W) = 64$ y $Var(W) = 144$.

Considerando independencia entre las semanas,

- 1 ¿Cuál es la probabilidad de que la ganancia promedio en 1 año (52 semanas) sea mayor a USD 65000?

TCL para promedios, ya que sirve para sumas y promedios

Teorema Central del Límite

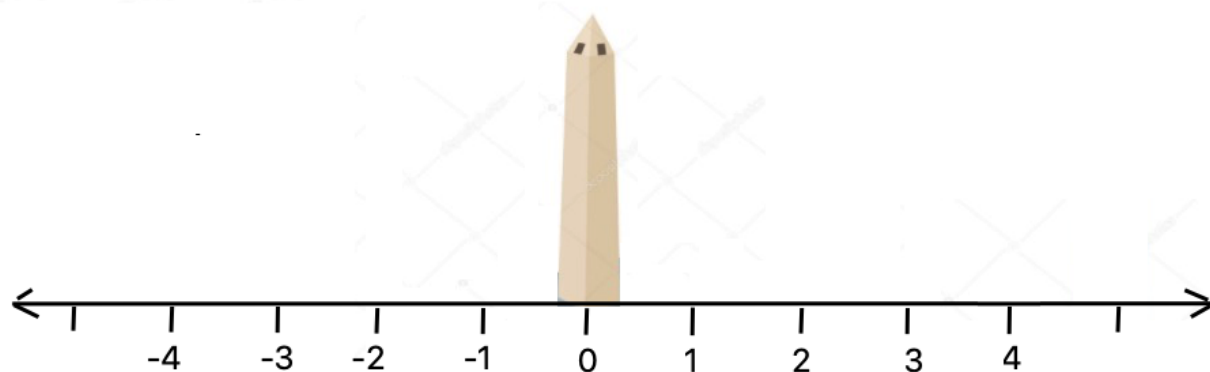
Ejemplo \bar{X}_n

Tenemos W_1, \dots, W_{52} v. a. iid, cada una representa la ganancia de una semana.

$$\begin{aligned} P(\bar{W}_{52} > 65) &= P\left(\frac{\bar{W}_{52} - 64}{12/\sqrt{52}} > \frac{65 - 64}{12/\sqrt{52}}\right) \\ &= P(Z_{52} > \sqrt{52}/12) \approx 1 - \Phi(0,6) = 0,2743. \end{aligned}$$

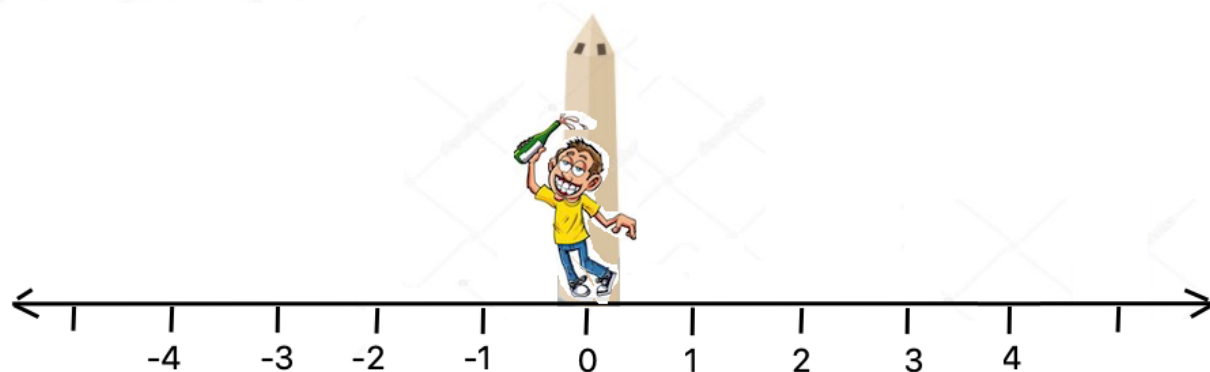
Teorema Central del Límite

Aplicación: Random walk o paseo del borracho



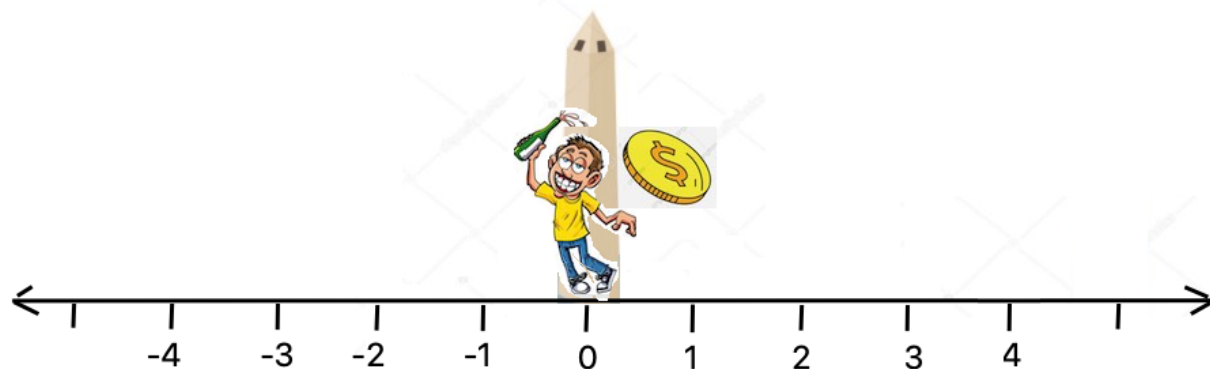
Teorema Central del Límite

Aplicación: Random walk o paseo del borracho



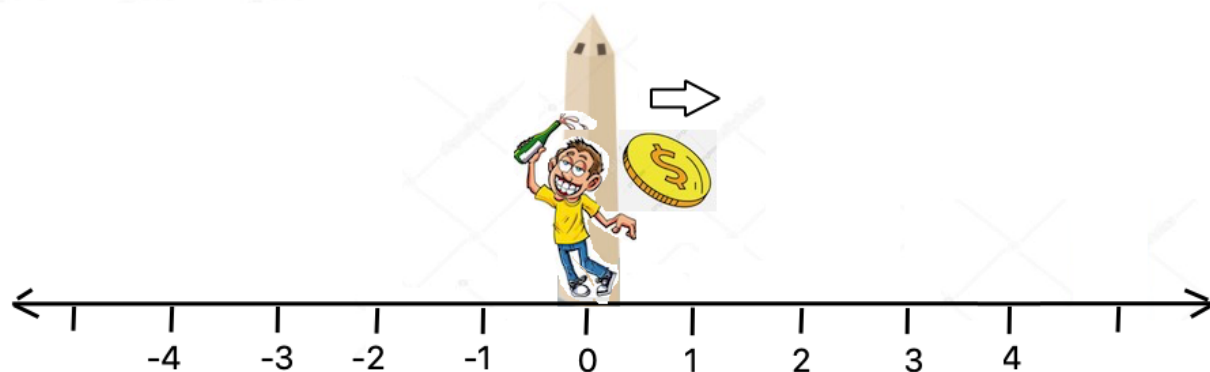
Teorema Central del Límite

Aplicación: Random walk o paseo del borracho



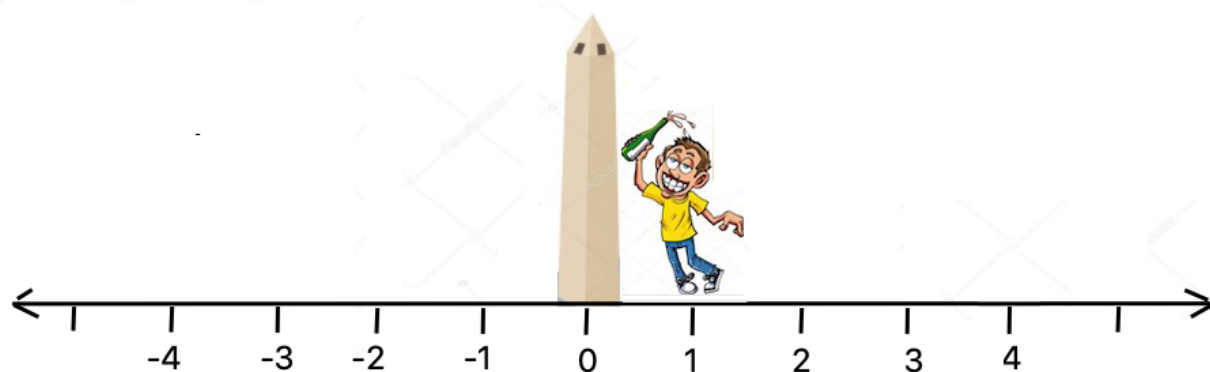
Teorema Central del Límite

Aplicación: Random walk o paseo del borracho



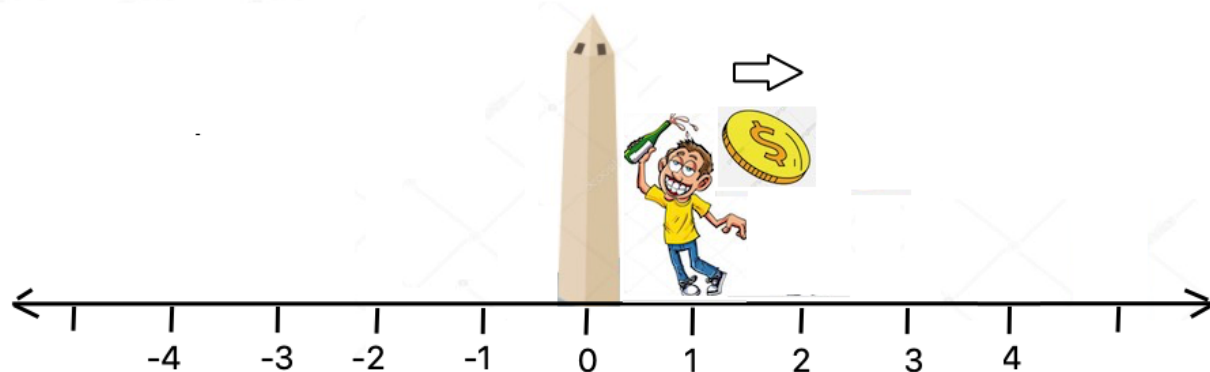
Teorema Central del Límite

Aplicación: Random walk o paseo del borracho



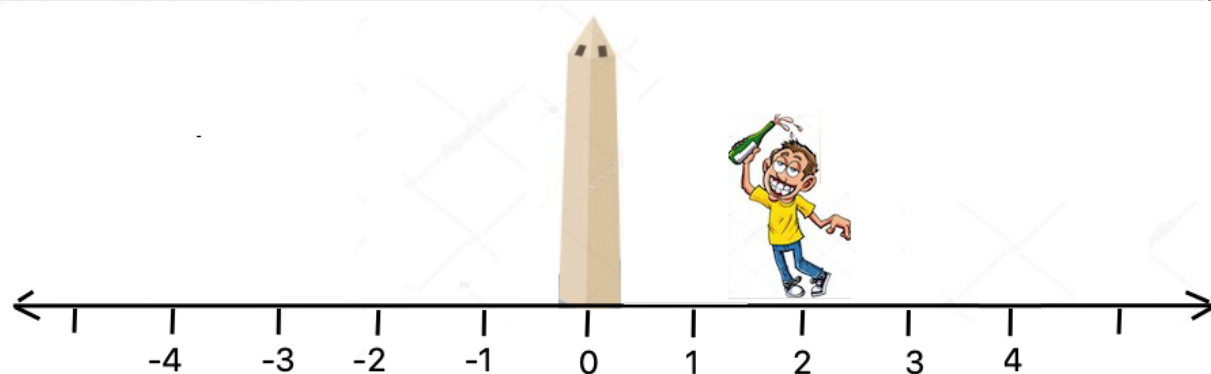
Teorema Central del Límite

Aplicación: Random walk o paseo del borracho



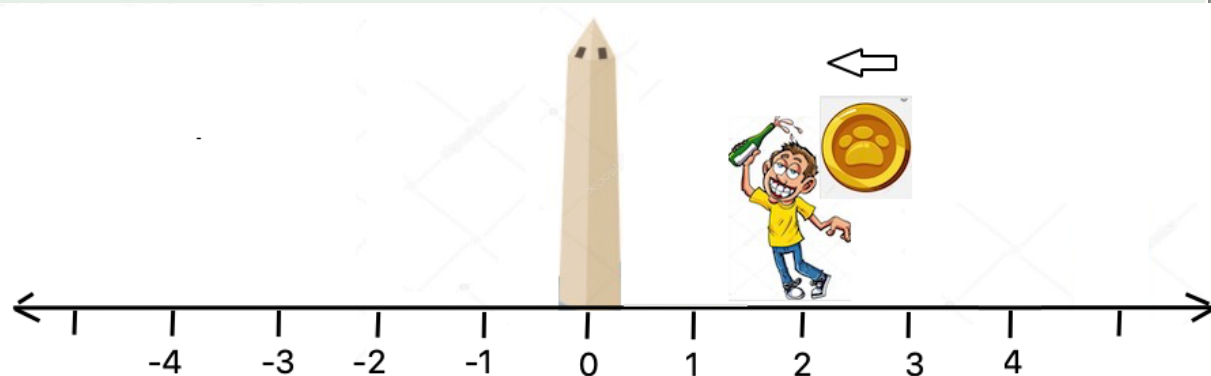
Teorema Central del Límite

Aplicación: Random walk o paseo del borracho



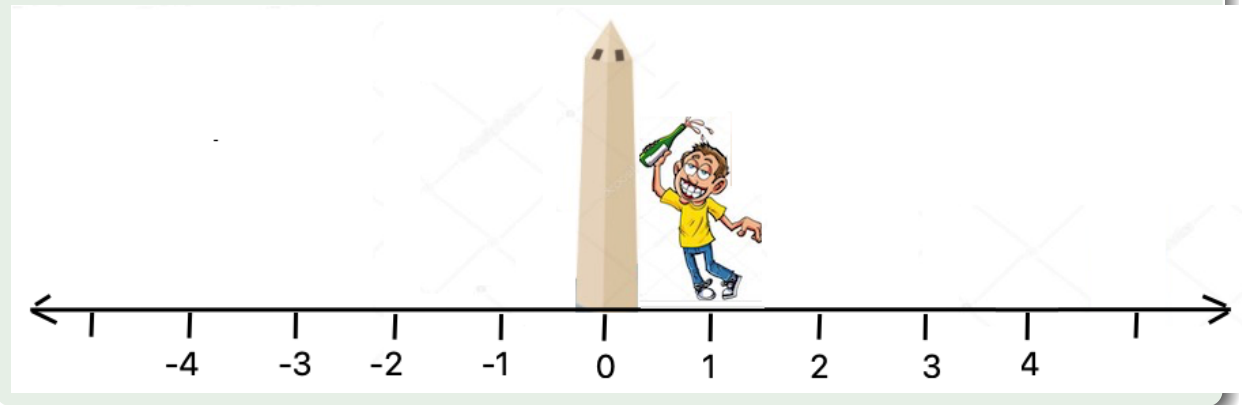
Teorema Central del Límite

Aplicación: Random walk o paseo del borracho



Teorema Central del Límite

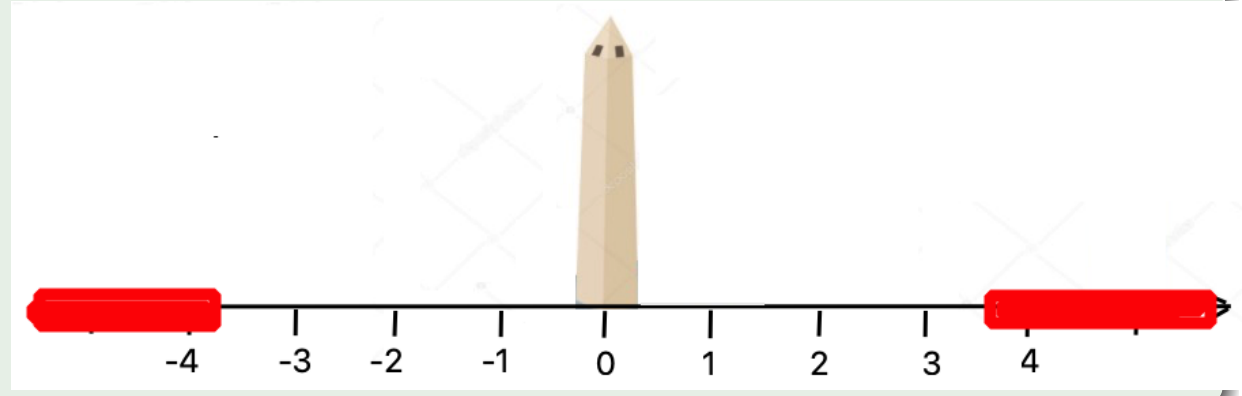
Aplicación: Random walk o paseo del borracho



Teorema Central del Límite

Aplicación: Random walk o paseo del borracho

Sabiendo que el borracho tarda 1 minuto en caminar una cuadra, ¿cuál es la probabilidad de que se encuentre a más de 3 cuadras del obelisco al cabo de una hora?



Teorema Central del Límite

Aplicación: Random walk o paseo del borracho

S_n =posición del borracho al cabo de n minutos.

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

Donde X_1, X_2, \dots, X_n son iid, con $X_i = \{1, -1\}$ y

$\mathbb{P}(X_i = 1) = \mathbb{P}(X_i = -1) = 1/2$. ¿cuál es la probabilidad de que se

encuentre a más de 3 cuadras del obelisco al cabo de una hora?=
 $\mathbb{P}(|S_{60}| > 3)$?

Teorema Central del Límite: Aproximación normal a la binomial

Si $X \sim \text{Bi}(n, p)$, vimos que $X = X_1 + \cdots + X_n$ con X_1, \dots, X_n v. a. i. i. d. *Bernoulli*(p). O sea, $X = S_n$ es una suma de variables independientes idénticamente distribuidas. Si n es grande, por TCL, X se aproxima por una variable normal:

$$X \approx Y \sim N(np, np(1 - p)).$$

Una Binomial con n grande se aproxima por una Normal.

Ejemplo

Si $X \sim \text{Bi}(100, 0.2)$ entonces $E(X) = 20$, $\text{Var}(X) = 16$. Por ser $n = 100$ grande, resulta $X \approx Y \sim N(20, 16)$.

¿ $P(X \leq 25)$? `pnorm(q = 25, mean = 20, sd = sqrt(16)) = 0.894`

Teorema Central del Límite

Ejemplo

Queremos estimar la proporción de gente, p que hoy votaría a un candidato.

¿Con que error? ¿Decir $0,35 \pm 0,02$ está ok? o $\pm 0,01$?

¿Y con qué confianza quieres que esté el verdadero valor en ese intervalo?

$\mathbb{P}(|\bar{X}_n - p| < 0,02) \geq 0,95$ ¿0.95 está bien? ¿o necesitas 0.99?

Hallar n tal que $\mathbb{P}(|\bar{X}_n - p| < 0,02) \geq 0,95$.

$X \text{ dist Binom}(10000, 0.1)$

$P(x = 8932) = \binom{10000}{8932} (0.1)^{8932} (0.9)^{10000-8932}$

Una binomial es una suma de experimentos independientes de bernoulli $X = y_1 + y_2 + \dots + y_{10000}$

$P(X < 8932) = P(S_{10000} < 8932) \rightarrow$ aplicamos TCL y lo aproximamos a una normal

En conclusion, una binomial se puede aproximar a una Normal por TCL, donde en los casos donde tengo una matematica mas complicada, puedo aproximar su resultado a partir de la distribucion Normal