Probabilidad

Daniel Fraiman

Maestría en Ciencia de Datos, Universidad de San Andrés

CONVERGENCIA DE VARIABLES ALEATORIAS

Convergencia

Definición:convergencia en probabilidad

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

$$\lim_{n\to\infty} \mathbb{P}\left(|Z_n - Z| > \epsilon\right) = 0 \qquad \forall \epsilon > 0$$

Definición: convergencia en distribución

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

$$\mathit{lim}_{n\to\infty} \mathbb{P}\left(Z_n \leq t\right) = \mathbb{P}\left(Z \leq t\right) \qquad \forall t \text{ punto de continuidad}$$

$$\mathit{lim}_{n\to\infty} F_{Z_n}(t) = F_Z(t) \qquad \forall t \text{ punto de continuidad}$$

Ley de Grandes Números

Ley de los Grandes Números (LGN)

LGN:¿Un resultado innato o adquirido?

Supongamos que queremos conocer las chances de ganar a un juego, $\mathbb{P}\left(Ganar\right) = p$. 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 $X_1, X_2, \ldots X_n$ donde $X_i \sim Bernoulli(p)$. Sabemos que $\mathbb{E}\left(X_i\right) = p$ pero desconocemos el valor de p.
- Calculamos $\overline{X}_n = \frac{X_1 + \cdots + X_n}{n}$. Sabemos que $\mathbb{E}\left(\overline{X}_n\right) = \mathbb{E}\left(X_i\right) = p$.
- Finalmente estimamos a p con $\overline{X}_n$. ¿Por qué? ¿Qué pasa a medida que n crece?

Promedio de variables aleatorias

Dadas $X_1, X_2, \ldots, X_n$ variables aleatorias independientes tales que

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

definimos la variable aleatoria promedio

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

Recordemos que vimos que

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

En este caso, no conocemos la distribución exacta de $\overline{X}_n$ pero sabemos que la esperanza es la misma y que la varianza tiende a cero cuando $n \to \infty$. 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 $\overline{X}_n = \frac{X_1+\cdots+X_n}{n}$. Entonces, para todo $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\left(|\overline{X}_n - \mu| \geq \varepsilon\right) = 0$$

es decir

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

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:

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

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

Teorema Central del Límite

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

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 $S_n = X_1 + \cdots + X_n$ y $Z_n = \frac{S_n - n\mu}{\sqrt{n}\sigma} = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma}$. Entonces, para todo $x \in \mathbb{R}$,

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

o equivalentemente

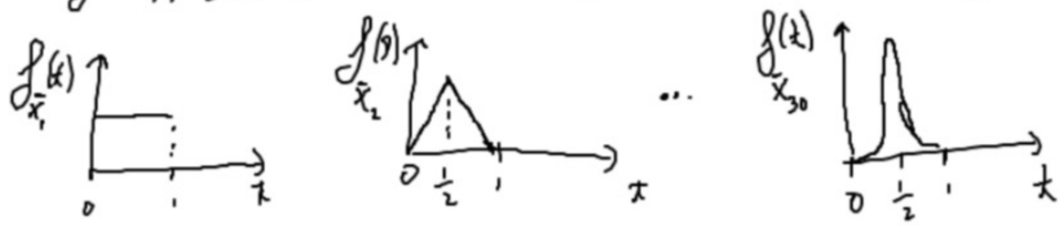
$$Z_n \xrightarrow{D} Z \sim N(0,1)$$

O sea, el TCL dice que $\lim_{n\to\infty} \mathbb{P}(Z_n \leq x) = \Phi(x)$. (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.

Sj: $X_1, X_2, \ldots, X_n$ v.a. iid con $X_i \sim U[0,1]$ $\Rightarrow$ $\bar{X}_n = \frac{X_1 + \cdots + X_n}{n}$



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?

Tenemos $X_1, \ldots, X_{400}$ 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

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

Teorema Central del Límite

Ejemplo $\overline{X}_n$

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?

Teorema Central del Límite

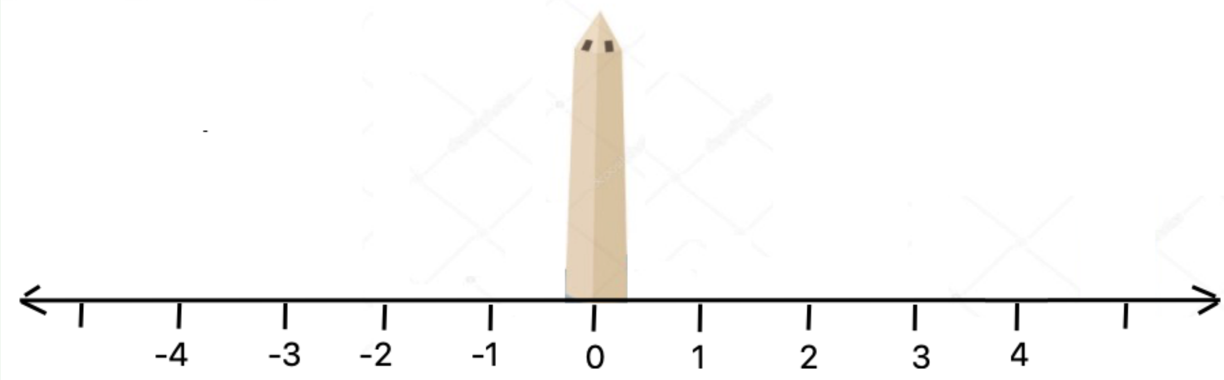
Ejemplo $\overline{X}_n$

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

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

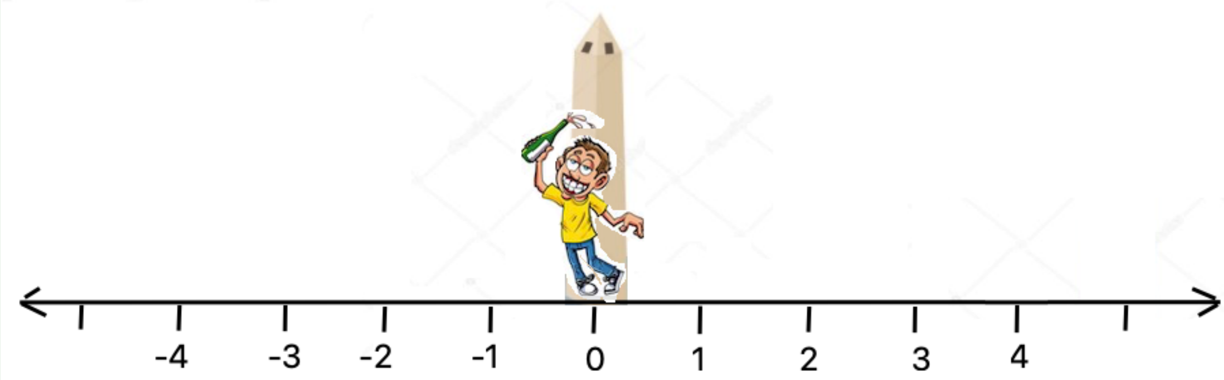
Teorema Central del Límite

Aplicación: Random walk o paseo del borracho



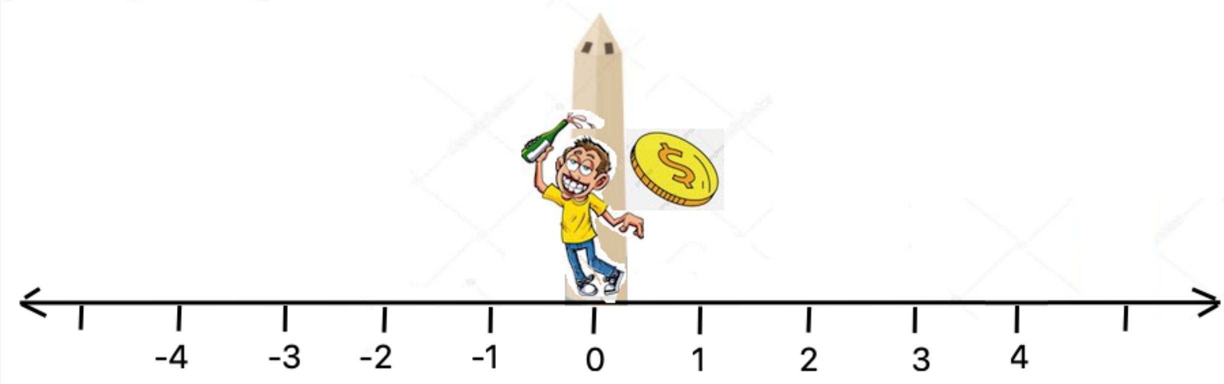
Teorema Central del Límite

Aplicación: Random walk o paseo del borracho



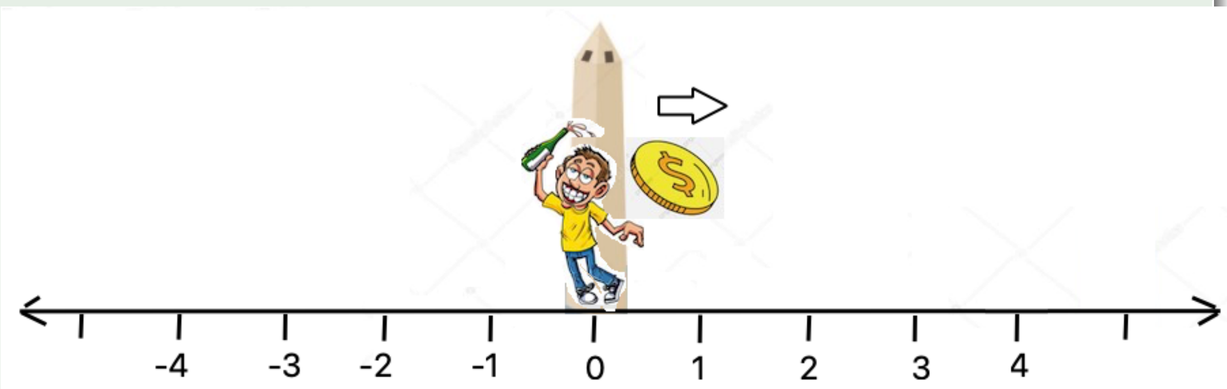
Teorema Central del Límite

Aplicación: Random walk o paseo del borracho



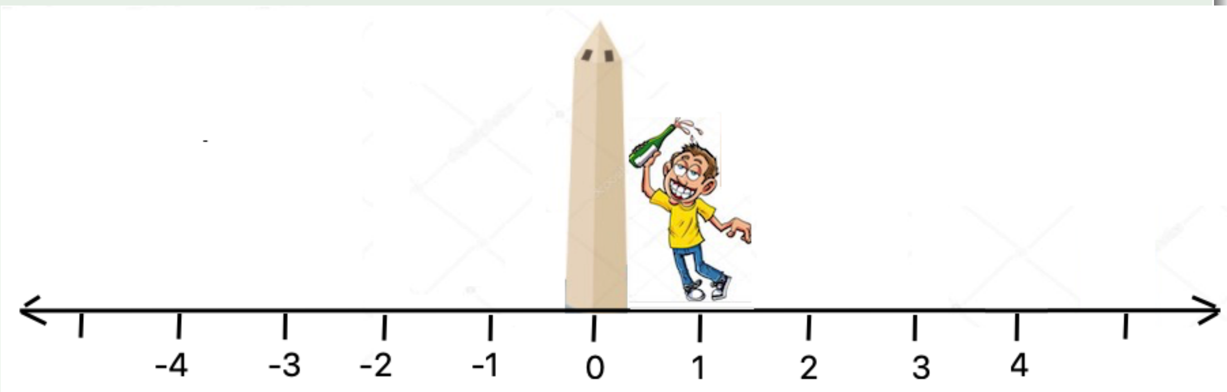
Teorema Central del Límite

Aplicación: Random walk o paseo del borracho



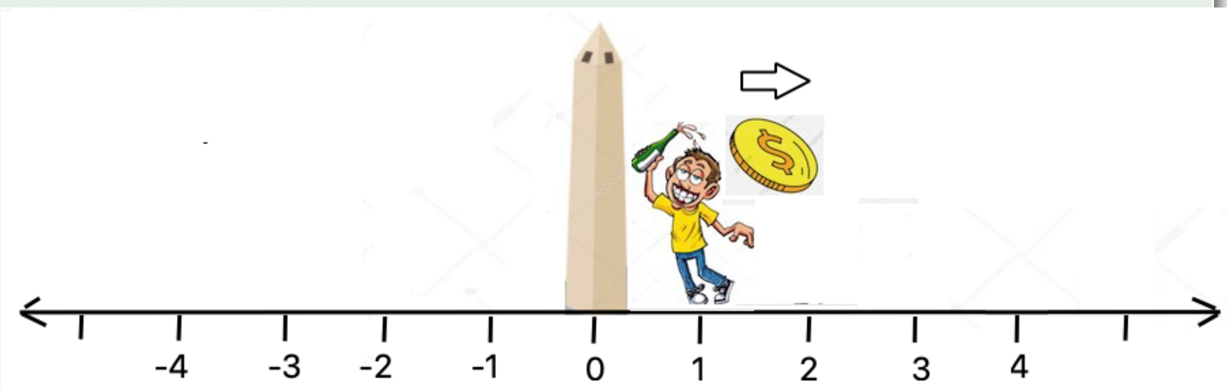
Teorema Central del Límite

Aplicación: Random walk o paseo del borracho



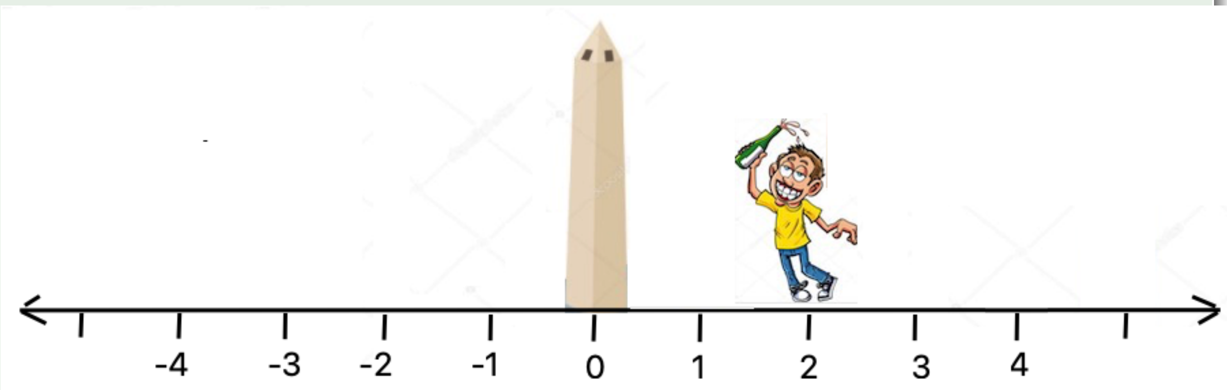
Teorema Central del Límite

Aplicación: Random walk o paseo del borracho



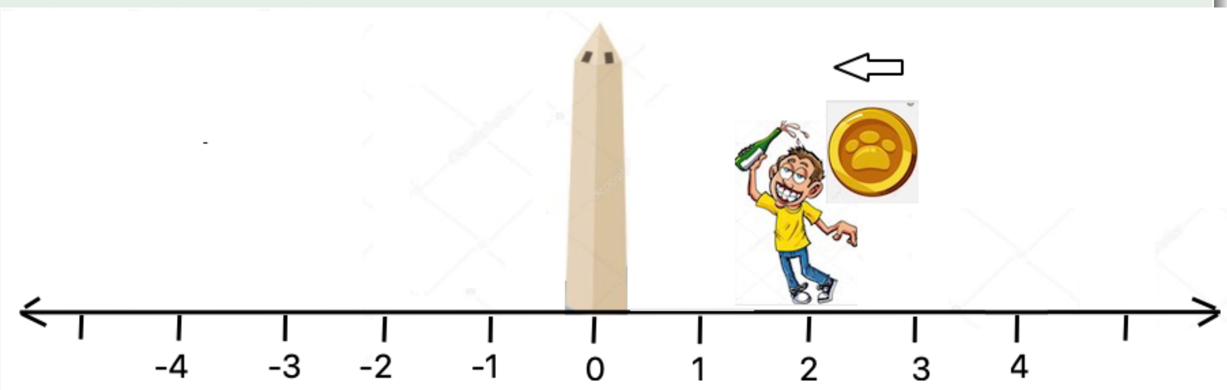
Teorema Central del Límite

Aplicación: Random walk o paseo del borracho



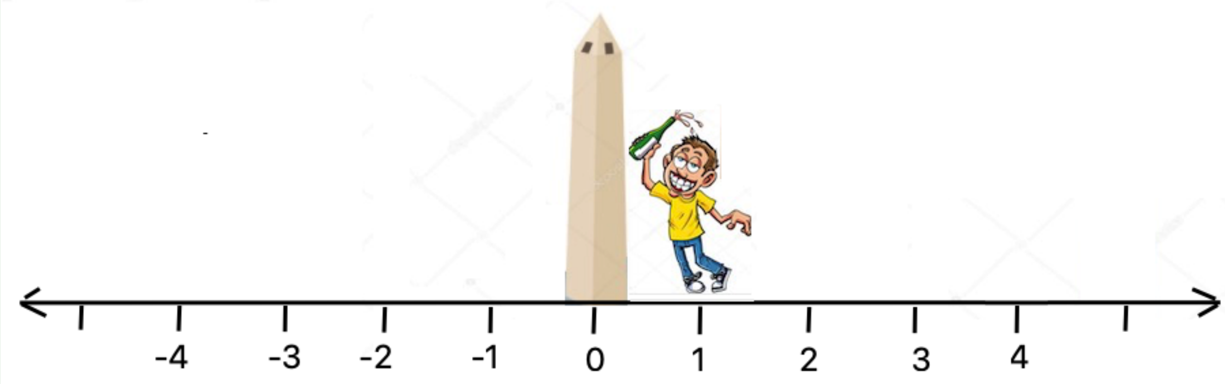
Teorema Central del Límite

Aplicación: Random walk o paseo del borracho



Teorema Central del Límite

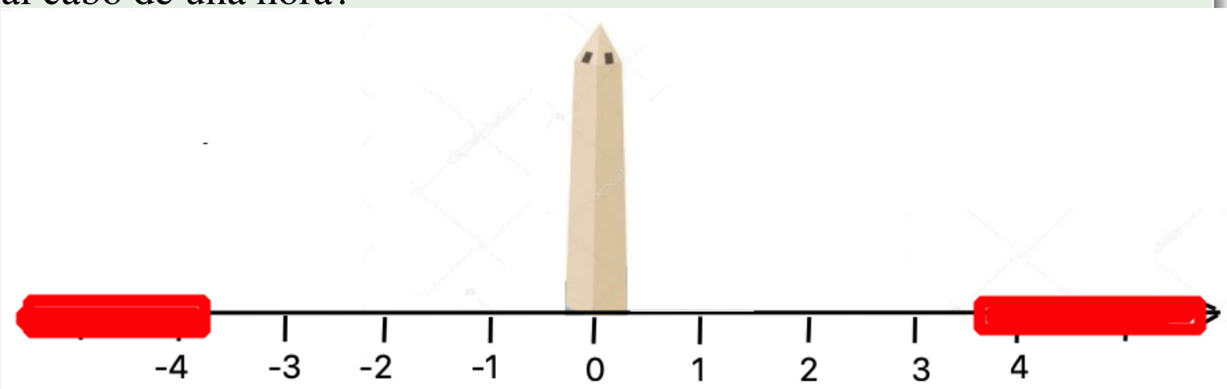
Aplicación: Random walk o paseo del borracho



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 $X_1, X_2, \ldots, X_n$ 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 $X \sim Bi(n, p)$, vimos que $X = X_1 + \cdots + X_n$ con $X_1, \ldots, X_n$ 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 $X \sim Bi(100, 0{,}2)$ entonces $E(X) = 20$, $Var(X) = 16$. Por ser $n = 100$ grande, resulta $X \approx Y \sim N(20, 16)$.

?‘$P(X \leq 25)$?.

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 queres que esté el verdadero valor en ese intervalo?

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

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