

Distribuciones Especiales, Media Muestral y TLC

Clase 5 — 14.310x: Data Analysis for Social Scientists

B.Sc. Cristian Lazo Quispe

BREIT (Brescia Institute of Technology) · Aporta — Grupo Breca

Programa SDS (Statistics and Data Science, MITx) · BREIT (Brescia Institute of Technology), iniciativa de Aporta —Grupo Breca—
en alianza con MIT IDSS.

Basado en MIT 14.310x (CC BY-NC-SA 4.0).

✉ clazoq@uni.pe · [in linkedin.com/in/cristian-lazo-quispe](https://www.linkedin.com/in/cristian-lazo-quispe) · github.com/CristianLazoQuispe

¿Qué veremos hoy?

- 1 Motivación: distribuciones en la práctica (MLOps)
- 2 Distribuciones discretas especiales
- 3 Distribuciones continuas especiales
- 4 Sumas de variables independientes
- 5 La media muestral $\bar{X}_n$
- 6 Teorema del Límite Central (TLC)
- 7 Introducción a la estimación
- 8 Simular en R
- 9 Manos a la obra

Contenido

1 Motivación: distribuciones en la práctica (MLOps)

2 Distribuciones discretas especiales

3 Distribuciones continuas especiales

4 Sumas de variables independientes

5 La media muestral $\bar{X}_n$

6 Teorema del Límite Central (TLC)

7 Introducción a la estimación

8 Simular en R

9 Manos a la obra

MLOps nivel 2: CI/CD pipeline automation

Tres niveles de madurez: nivel 0 (manual), nivel 1 (automatización del *pipeline* de ML) y nivel 2 (automatización de CI/CD). En el nivel 2 todo el ciclo —construir, probar, desplegar y reentrenar— se automatiza.

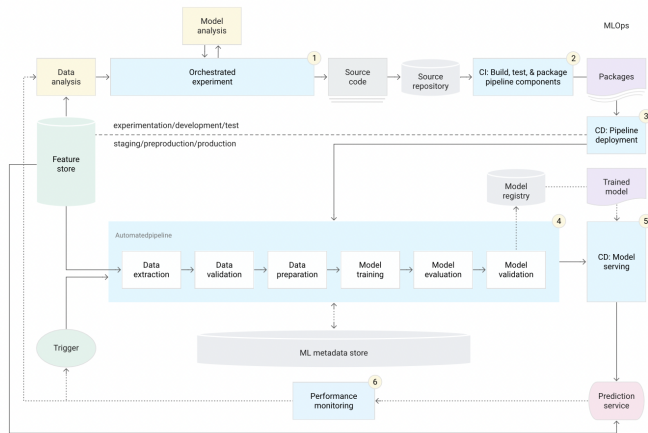
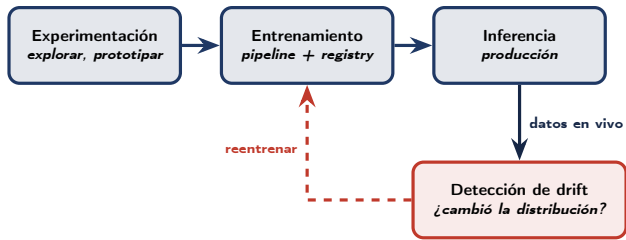
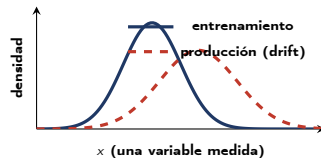


Figure 4. CI/CD and automated ML pipeline. Fuente: Google Cloud Architecture — cloud.google.com (clic para abrir).

Aplicación: detección de *drift* (puente con la Clase 4)



Si la distribución de los datos se mueve, el modelo se degrada $\Rightarrow$ se dispara el reentrenamiento.

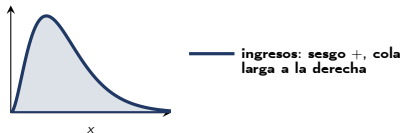


Lo medimos con la Clase 4

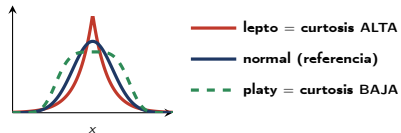
Comparamos los momentos de la distribución: media μ , varianza σ^2 , asimetría (skewness) y curtosis (kurtosis); más distancias entre distribuciones (KS, PSI).

Asimetría y curtosis en datos reales (lo que mide el drift)

Asimetría (skewness)



Curtosis (kurtosis)



Ejemplos (coef. γ_1)

- Ingresos/sueldos: cola a la derecha, $\gamma_1 \approx +2 \Rightarrow$ media $>$ mediana.
- Notas de un examen fácil: cola a la izquierda, $\gamma_1 < 0$.
- Estaturas: simétrica, $\gamma_1 \approx 0$.

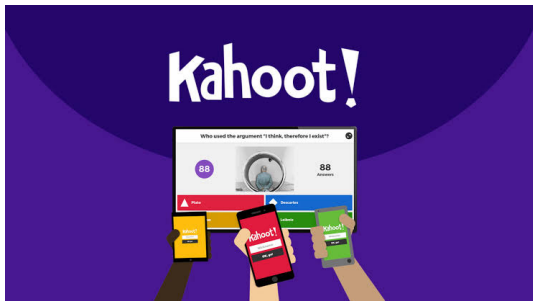
Ejemplos (exceso de curtosis γ_2)

Regla fácil: ALTA ($\gamma_2 > 0$) = pico puntiagudo + colas *pesadas*;
BAJA ($\gamma_2 < 0$) = chata + colas *finas*.

- Retornos de acciones: curtosis alta, $\gamma_2 \approx +3$ a $+6$ ("cisnes negros").
- Errores de medición: $\approx$ normal, $\gamma_2 \approx 0$.
- Uniforme $[a, b]$: curtosis baja, $\gamma_2 = -1.2$.

Estos momentos (Clase 4) son justo los que comparamos entre entrenamiento y producción para detectar drift.

Calentamiento: Kahoot de repaso (Clases 1–4)



Antes de entrar a la Clase 5

Repasamos **jugando** lo de las **Clases 1–4**: probabilidad y conteo, Bayes, variables aleatorias y esperanza/varianza.

- **30 preguntas**, de simple a complejo.
- Cuidado con las **alternativas trampa**: cazan el error típico.
- ¡Responder **rápido** suma más puntos!

Cómo unirse

Entra a **kahoot.it** (o la app) e ingresa el **PIN** que aparece en pantalla.

Contenido

- 1 Motivación: distribuciones en la práctica (MLOps)
- 2 Distribuciones discretas especiales
- 3 Distribuciones continuas especiales
- 4 Sumas de variables independientes
- 5 La media muestral $\bar{X}_n$
- 6 Teorema del Límite Central (TLC)
- 7 Introducción a la estimación
- 8 Simular en R
- 9 Manos a la obra

¿Por qué algunas distribuciones son “especiales”?

La idea

Unas pocas familias modelan una enorme variedad de fenómenos con *muy pocos parámetros*. Conocerlas es tener un **catálogo** de modelos listos para usar.

- **Bernoulli**: un único ensayo con dos resultados (éxito/fracaso).
- **Binomial**: número de éxitos en n ensayos independientes.
- **Geométrica**: número de ensayos hasta el *primer* éxito.
- **Poisson**: conteo de eventos raros en un intervalo (goles, accidentes).

Todas son **discretas**: se describen con una *pmf* $f_X(x) = \mathbb{P}(X = x)$.

Cuatro discretas en una tabla

Distribución	pmf	$\mathbb{E}[X]$	$\text{Var}(X)$	¿Cuándo?
Bernoulli(p)	$p^x(1-p)^{1-x}$	p	$p(1-p)$	1 ensayo sí/no
Binomial(n, p)	$\binom{n}{x} p^x(1-p)^{n-x}$	np	$np(1-p)$	# éxitos en n
Geométrica(p)	$(1-p)^{k-1}p$	$\frac{1}{p}$	$\frac{1-p}{p^2}$	# ensayos al 1.º éxito
Poisson(λ)	$\frac{\lambda^k e^{-\lambda}}{k!}$	λ	λ	conteo de eventos raros

Relaciones útiles

- Bernoulli(p) = Binomial(1, p); suma de n Bernoulli iid = Binomial(n, p).
- Binomial $\rightarrow$ Poisson cuando n es grande y p pequeño, con $\lambda = np$.

Demo resuelto: Binomial (tiros de Steph Curry)

Problema

Un jugador encesta cada tiro libre con probabilidad $p = 0.6$, de forma independiente. Lanza $n = 10$ tiros.

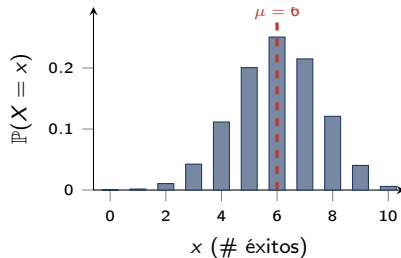
Se pide: $\mathbb{P}(X = 6)$, $\mathbb{E}[X]$ y $\text{Var}(X)$.

Solución

$X \sim \text{Binomial}(10, 0.6)$, así que

$$\mathbb{P}(X = 6) = \binom{10}{6} (0.6)^6 (0.4)^4 \approx 0.251.$$

$$\mathbb{E}[X] = np = 6, \quad \text{Var}(X) = np(1 - p) = 2.4.$$



Barras: pmf Binomial(10, 0.6).

Línea roja: la media $\mu = np = 6$.

Demo resuelto: Poisson (goles por partido)

Problema

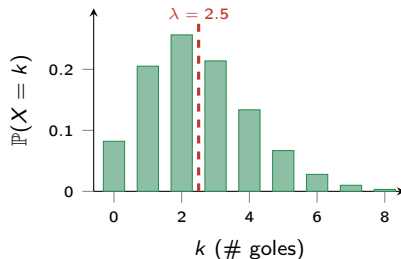
El número de goles por partido es Poisson con tasa media $\lambda = 2.5$. **Se pide:** $\mathbb{P}(X = 0)$ (partido sin goles), $\mathbb{E}[X]$ y $\text{Var}(X)$.

Solución

$X \sim \text{Poisson}(2.5)$, con $f(k) = \frac{\lambda^k e^{-\lambda}}{k!}$:

$$\mathbb{P}(X = 0) = e^{-2.5} \approx 0.082.$$

El parámetro gobierna media y varianza:
 $\mathbb{E}[X] = \text{Var}(X) = \lambda = 2.5$.

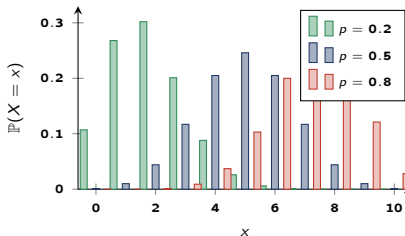


Barras: pmf Poisson(2.5).

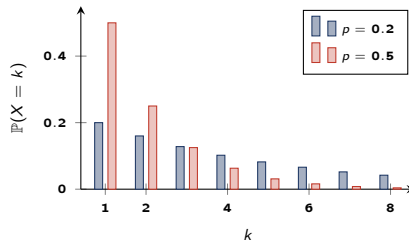
Asimétrica: nunca es negativa.

El efecto de los parámetros: Binomial y Geométrica

Binomial(10, p): ¿dónde cae la masa?



Geométrica(p): decaimiento

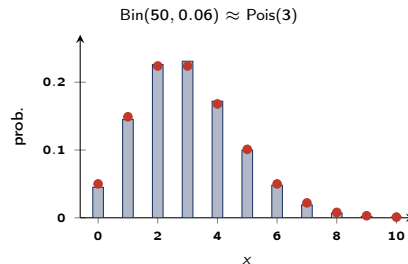
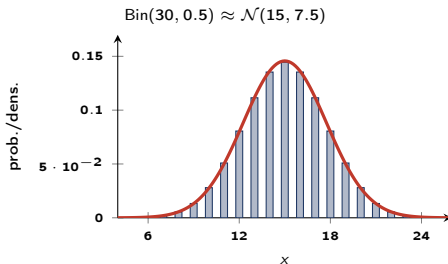


Qué ver

Binomial: p mueve el pico ($p < 0.5$ a la izquierda, $p > 0.5$ a la derecha; $p = 0.5$ simétrica).

Geométrica: cada barra es $(1 - p)$ veces la anterior $\Rightarrow$ decae; el valor más probable es siempre $k = 1$.

Aproximaciones: la Binomial “se vuelve” Normal o Poisson



Dos atajos clásicos

- **Normal** si n es grande y p no muy extremo: $\text{Bin}(n, p) \approx \mathcal{N}(np, np(1 - p))$.
- **Poisson** si n grande y p pequeño (eventos raros): $\text{Bin}(n, p) \approx \text{Pois}(np)$.

Contenido

- 1 Motivación: distribuciones en la práctica (MLOps)
- 2 Distribuciones discretas especiales
- 3 Distribuciones continuas especiales**
- 4 Sumas de variables independientes
- 5 La media muestral $\bar{X}_n$
- 6 Teorema del Límite Central (TLC)
- 7 Introducción a la estimación
- 8 Simular en R
- 9 Manos a la obra

Tres continuas que aparecen en todas partes

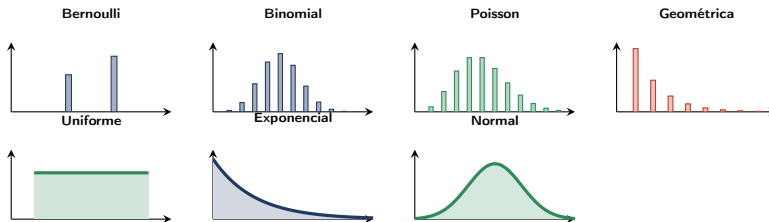
Ahora X es **continua**: se describe con una *pdf* f_X y $\mathbb{P}(a \leq X \leq b) = \int_a^b f_X(x) dx$.

Distribución	pdf	$\mathbb{E}[X]$	$\text{Var}(X)$	¿Cuándo?
Uniforme $[a, b]$	$\frac{1}{b-a}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	todo igual de probable
Exponencial(λ)	$\lambda e^{-\lambda x}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	espera entre eventos
Normal(μ, σ^2)	$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	μ	σ^2	“campana”, sumas/promedios

La estrella: la Normal

Simétrica, unimodal, “acampanada”, colas delgadas, soporte $\mathbb{R}$. Toda combinación lineal de normales *sigue siendo normal*.

Galería visual: ¿qué distribución uso?



Una frase por distribución

Bernoulli: un sí/no. **Binomial:** # éxitos en n intentos. **Poisson:** conteo de eventos en un intervalo. **Geométrica:** # intentos hasta el 1.^{er} éxito (decae). **Uniforme:** todo el rango igual de probable. **Exponencial:** tiempo de espera (decae). **Normal:** campana (sumas/promedios).

Estandarizar: pasar a la Normal estándar Z

Definición 1 (Estandarización (the Z -score))

Si $X \sim N(\mu, \sigma^2)$, entonces

$$Z = \frac{X - \mu}{\sigma} \sim N(0, 1), \quad \mathbb{E}[Z] = 0, \text{ Var}(Z) = 1.$$

¿Para qué sirve?

La integral de la campana no tiene forma cerrada. Estandarizamos para usar **una sola tabla** (o software) de la Normal estándar Φ .

Trucos de tabla (usando $\Phi(z) = \mathbb{P}(Z \leq z)$)

$$\mathbb{P}(Z > z) = 1 - \Phi(z), \quad \mathbb{P}(a \leq Z \leq b) = \Phi(b) - \Phi(a).$$

Demo resuelto: la Normal (estandarización Z)

Problema

Las notas de un examen son

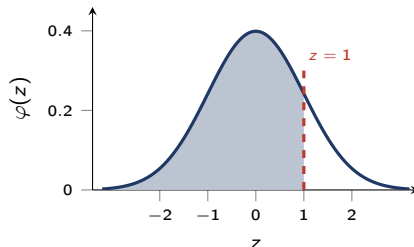
$X \sim N(\mu = 70, \sigma^2 = 64)$ (es decir $\sigma = 8$). Se pide:
 $\mathbb{P}(X \leq 78)$.

Solución (estandarizar)

$$Z = \frac{X - \mu}{\sigma} = \frac{78 - 70}{8} = 1.0, \text{ por tanto}$$

$$\mathbb{P}(X \leq 78) = \Phi(1) \approx 0.841.$$

Cerca del 84 % saca 78 o menos.



Área sombreada = $\Phi(1) \approx 0.841$:
la probabilidad a la izquierda de $z = 1$.

Ejemplo con gráfico: llegadas a una oficina (Poisson)

Problema

A una ventanilla del Estado llegan en promedio $\lambda = 5$ personas cada 10 min, de forma independiente.

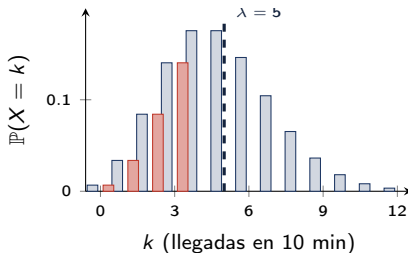
$X \sim \text{Pois}(5)$.

Se pide: $\mathbb{P}(X \leq 3)$.

Solución

$$\mathbb{P}(X \leq 3) = e^{-5} \left(1 + 5 + \frac{25}{2} + \frac{125}{6} \right) \approx 0.265.$$

Las barras rojas ($k = 0, 1, 2, 3$) suman esa probabilidad. Media y varianza: $\mathbb{E}[X] = \text{Var}(X) = 5$.



Rojo: $\mathbb{P}(X \leq 3) \approx 0.265$.

Ejemplo con gráfico: espera del bus (Exponencial) y $\mathbb{P}(X > t)$

Problema

El tiempo X (min) hasta el próximo bus es $\text{Exp}(\lambda)$ con $\lambda = 1/8$ (media $\mathbb{E}[X] = 8$ min).

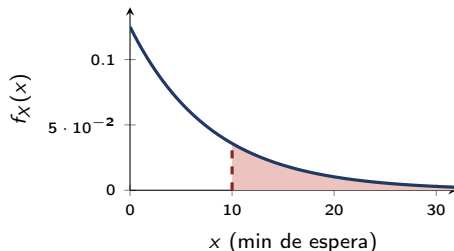
Se pide: $\mathbb{P}(X > 10)$ (esperar más de 10 min).

Solución (la “cola” de la cdf)

$$\mathbb{P}(X > t) = e^{-\lambda t}$$

$$\mathbb{P}(X > 10) = e^{-10/8} = e^{-1.25} \approx 0.287.$$

Es el área roja a la derecha de $x = 10$.



Área roja = $\mathbb{P}(X > 10) \approx 0.287$.

Contenido

- 1 Motivación: distribuciones en la práctica (MLOps)
- 2 Distribuciones discretas especiales
- 3 Distribuciones continuas especiales
- 4 Sumas de variables independientes**
- 5 La media muestral $\bar{X}_n$
- 6 Teorema del Límite Central (TLC)
- 7 Introducción a la estimación
- 8 Simular en R
- 9 Manos a la obra

Sumas de independientes: ¿qué familia resulta?

Propiedades reproductivas (la suma “no cambia de familia”)

Sumandos independientes	Condición	Suma $S = \sum_i X_i$
$X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$	—	$\mathcal{N}(\sum \mu_i, \sum \sigma_i^2)$
$X_i \sim \text{Pois}(\lambda_i)$	—	$\text{Pois}(\sum \lambda_i)$
$X_i \sim \text{Bernoulli}(p)$	mismo p	$\text{Bin}(k, p)$
$X_i \sim \text{Bin}(n_i, p)$	mismo p	$\text{Bin}(\sum n_i, p)$
$X_i \sim \text{Exp}(\lambda)$	i.i.d.	$\text{Gamma/Erlang}(k, \lambda)$
$X_i \sim \text{Gamma}(\alpha_i, \lambda)$	misma λ	$\text{Gamma}(\sum \alpha_i, \lambda)$

Suma “cruzada” (combinación lineal)

Con pesos a, b y normales independientes:

$$aX + bY \sim \mathcal{N}(a\mu_X + b\mu_Y, a^2\sigma_X^2 + b^2\sigma_Y^2).$$

Incluso la *diferencia* suma varianzas: $\text{Var}(X - Y) = \sigma_X^2 + \sigma_Y^2$.

Ojo

La Uniforme no es reproductiva: suma de dos uniformes ya es *triangular*. ¡Ese es el primer paso del TLC!

¿Cómo se calcula $\mathbb{E}[XY]$? (y por qué importa la independencia)

Esperanza de un producto y covarianza

Con la distribución *conjunta* $p(x, y)$ ó $f(x, y)$:

$$\mathbb{E}[XY] = \sum_x \sum_y x y p(x, y) \quad \text{ó} \quad \mathbb{E}[XY] = \iint x y f(x, y) dx dy.$$

$$\text{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X] \mathbb{E}[Y]$$

Si X, Y son independientes

La conjunta se factoriza $f(x, y) = f_X(x)f_Y(y)$, entonces

$$\mathbb{E}[XY] = \mathbb{E}[X] \mathbb{E}[Y] \Rightarrow \text{Cov}(X, Y) = 0.$$

(La recíproca *no* siempre vale: covarianza 0 no implica independencia.)

Varianza de una suma

$$\text{Var}(X + Y) = \text{Var } X + \text{Var } Y + 2 \text{Cov}(X, Y).$$

Con independencia el término cruzado se anula $\Rightarrow$ es lo que da $\text{Var}(\bar{X}) = \sigma^2/n$.

Contenido

- 1 Motivación: distribuciones en la práctica (MLOps)
- 2 Distribuciones discretas especiales
- 3 Distribuciones continuas especiales
- 4 Sumas de variables independientes
- 5 La media muestral $\bar{X}_n$**
- 6 Teorema del Límite Central (TLC)
- 7 Introducción a la estimación
- 8 Simular en R
- 9 Manos a la obra

La media muestral (the sample mean)

Definición 2 (Media muestral)

Para una muestra aleatoria $X_1, \dots, X_n$ (iid), $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$. Es función de v.a., así que **ella misma es una variable aleatoria**.

Sus dos momentos

Si cada X_i tiene media μ y varianza σ^2 :

$$\mathbb{E}[\bar{X}_n] = \mu$$

$$\text{Var}(\bar{X}_n) = \frac{\sigma^2}{n}$$

La media *no* cambia (no hace falta independencia); la varianza *se encoge* al crecer n (aquí sí usamos independencia).

Error estándar: la precisión del promedio

Definición 3 (Error estándar (standard error))

$$SE(\bar{X}_n) = \sqrt{\text{Var}(\bar{X}_n)} = \frac{\sigma}{\sqrt{n}}.$$

Es la desviación estándar de la media muestral.

La lección clave: la regla del $\sqrt{n}$

- El error estándar baja con $\sqrt{n}$, *no* con n .
- Para **reducir el error a la mitad** hay que **cuadruplicar** el tamaño de muestra ($n \rightarrow 4n$).
- Más datos $\Rightarrow$ promedio más concentrado alrededor de μ .

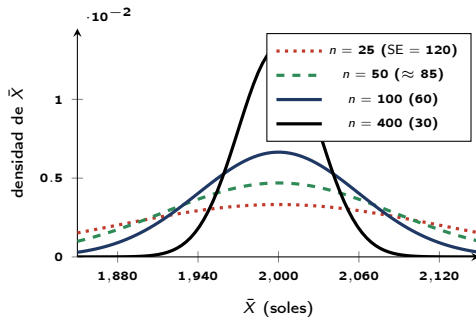
Visual: el error estándar *angosta* la distribución de $\bar{X}$

La idea

Con $\mu = 2000$, $\sigma = 600$ (ingresos), la distribución de $\bar{X}_n$ siempre está **centrada** en μ , pero al crecer n se vuelve **más alta y angosta**: $SE = \sigma/\sqrt{n}$ encoge.

Cuidado con el $\sqrt{n}$

De $n = 25$ a $n = 100$ ($4\times$ datos) el SE baja a la mitad ($120 \rightarrow 60$), no a la cuarta parte.



Demo resuelto: $\mathbb{E}[\bar{X}_n]$ y $\text{SE}(\bar{X}_n)$

Problema

El gasto diario en transporte de un hogar limeño tiene $\mu = 12$ soles y $\sigma = 6$ soles. Tomamos una muestra de $n = 100$ hogares.

Se pide: $\mathbb{E}[\bar{X}_n]$, $\text{Var}(\bar{X}_n)$ y el error estándar $\text{SE}(\bar{X}_n)$.

Solución

$$\mathbb{E}[\bar{X}_n] = \mu = 12 \text{ soles}, \quad \text{Var}(\bar{X}_n) = \frac{\sigma^2}{n} = \frac{36}{100} = 0.36,$$

$$\text{SE}(\bar{X}_n) = \frac{\sigma}{\sqrt{n}} = \frac{6}{\sqrt{100}} = \frac{6}{10} = 0.6 \text{ soles.}$$

Con $n = 400$ el error caería a $6/20 = 0.3$: la mitad, pero al costo de $4\times$ datos.

Contenido

- 1 Motivación: distribuciones en la práctica (MLOps)
- 2 Distribuciones discretas especiales
- 3 Distribuciones continuas especiales
- 4 Sumas de variables independientes
- 5 La media muestral $\bar{X}_n$
- 6 Teorema del Límite Central (TLC)**
- 7 Introducción a la estimación
- 8 Simular en R
- 9 Manos a la obra

El Teorema del Límite Central (Central Limit Theorem)

Teorema 1 (TLC)

Sea $X_1, \dots, X_n$ una muestra aleatoria de cualquier distribución con media μ y varianza σ^2 finitas. Entonces, para n grande,

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} N(0, 1), \quad \text{equivalentemente}$$

$$\bar{X}_n \approx N\left(\mu, \frac{\sigma^2}{n}\right).$$

¿Por qué es tan importante?

No importa cuán “rara” sea la distribución de partida: el **promedio** de una muestra grande es aproximadamente **normal**. Es la base de casi toda la estadística inferencial.

¿Por qué funciona el TLC? (intuición)

La idea de “cancelación”

$\bar{X}_n$ suma muchas contribuciones pequeñas e independientes y divide entre n . Cada dato trae su idiosincrasia (unos arriba de μ , otros abajo); al promediar, esas desviaciones **se cancelan** en gran parte. Lo que sobrevive es un ruido neto, simétrico, con **forma de campana**.

Por qué se “borra” la forma original

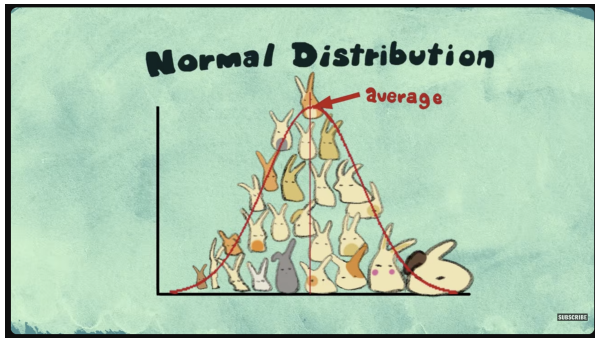
Ningún dato manda: todos pesan $1/n$, y ese peso $\rightarrow 0$. Por eso la forma de partida (plana, sesgada, con dos jorobas) *desaparece* y emerge la Normal.

Morphing del dado

Promedio de dados:

- $n = 1$: plana (uniforme).
- $n = 2$: triangular.
- $n = 5$: ya se acampana.
- $n = 30$: casi Normal.

Video recomendado: el TLC en imágenes



Clic en la imagen para abrir el video (YouTube).

Bunnies, Dragons and the “Normal” World

The New York Times — Central Limit Theorem.

- Muchos efectos pequeños e independientes (los conejitos) se **suman**...
- ...y su distribución toma **forma de campana**: la *Normal*.
- Es la intuición visual del **TLC** que acabamos de ver.

▶ youtube.com/watch?v=jvoxEYmQHNM

Video recomendado: la animación del TLC



Clic en la imagen para abrir el video (YouTube).

But what is the Central Limit Theorem?

3Blue1Brown — intuición visual del TLC.

- Al **sumar** cada vez más copias ($X_1, X_1+X_2, \dots$) la forma se **transforma** hacia la *campana*.
- No importa la distribución de partida: el límite es la **Normal** $\frac{1}{\sqrt{2\pi}} e^{-x^2/2}$.
- Animación que muestra el **morphing** hacia $N(0, 1)$.

📺 youtu.be/zeJD6dqJ5lo

El rol de $\sqrt{n}$ y errores comunes del TLC

La regla del $\sqrt{n}$

El ancho típico de $\bar{X}_n$ es $SE = \sigma/\sqrt{n}$. Para **duplicar la precisión** (mitad de SE) hacen falta $4\times$ datos; para $10\times$ precisión, $100\times$ datos.

Cuidado: tres errores frecuentes

- El TLC **no** normaliza los *datos*, sino el *promedio* $\bar{X}_n$ (o la suma). Datos sesgados siguen sesgados.
- No confundir σ (dispersión de un dato) con $SE = \sigma/\sqrt{n}$ (dispersión del promedio, mucho menor).
- “ n grande” depende de la asimetría: ≈ 30 como regla, más si hay colas pesadas; **falla** si la varianza es infinita.

Demo resuelto: aproximación normal del promedio

Problema

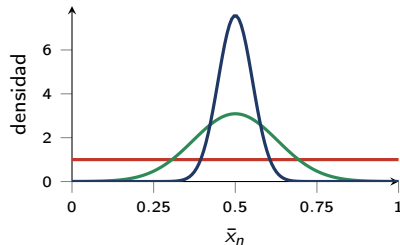
Los X_i son uniformes en $[0, 1]$ (¡plana, nada acampanada!), con $\mu = 0.5$ y $\sigma^2 = \frac{1}{12}$. Se pide: la distribución aproximada de $\bar{X}_n$ para $n = 1, 5, 30$.

Solución (TLC)

$\mathbb{E}[\bar{X}_n] = 0.5$ siempre, y $\text{Var}(\bar{X}_n) = \frac{1}{12n}$. Por el TLC,

$$\bar{X}_n \approx N\left(0.5, \frac{1}{12n}\right).$$

Al crecer n , la densidad se **concentra** y se vuelve **acampanada** alrededor de 0.5.



Roja: $n = 1$ (uniforme, plana).

Verde: $n = 5$. Azul: $n = 30$.

Más $n \Rightarrow$ más alta y angosta.

Demo resuelto: $\mathbb{P}(\bar{X}_n > c)$ paso a paso

Problema

Tiempo de atención por cliente: $\mu = 4$ min, $\sigma = 3$ min (¡muy sesgado!). Atendemos $n = 36$ clientes.
Se pide: $\mathbb{P}(\bar{X}_n > 5)$.

Solución (estandarizar con SE, no con σ)

- 1 TLC: $\bar{X}_n \approx N(4, \frac{9}{36}) = N(4, 0.25)$, con $SE = \frac{3}{6} = 0.5$.
- 2 $Z = \frac{5 - 4}{0.5} = 2.0$.
- 3 $\mathbb{P}(\bar{X}_n > 5) \approx \mathbb{P}(Z > 2) = 1 - \Phi(2) \approx 0.023$.

Interpretación

Solo $\approx 2.3\%$. *Un* cliente de 9 min es normal; que el *promedio* de 36 suba 1 min completo es raro, porque $SE = 0.5$ es $6\times$ menor que σ .

Demo resuelto: aproximación Normal a una proporción

Problema (una encuesta)

Encuestamos a $n = 400$ personas; la aprobación real es $p = 0.55$. La proporción muestral $\hat{p}$ es un *promedio de Bernoulli*. **Se pide:** $\mathbb{P}(\hat{p} < 0.50)$.

Solución (TLC para proporciones)

Con $\mu = p$ y $\sigma^2 = p(1 - p)$:

$$\hat{p} \approx N\left(p, \frac{p(1-p)}{n}\right), \quad SE = \sqrt{\frac{0.55 \cdot 0.45}{400}} \approx 0.0249.$$

$$Z = \frac{0.50 - 0.55}{0.0249} \approx -2.01, \quad \mathbb{P}(\hat{p} < 0.50) \approx \Phi(-2.01) \approx 0.022.$$

Interpretación

Aunque la verdad sea 55 %, hay $\approx 2.2\%$ de ver un “falso empate” por azar. Así nace el *margen de error*: con $n = 400$, $\approx 2 SE \approx \pm 5$ puntos.

Contenido

- 1 Motivación: distribuciones en la práctica (MLOps)
- 2 Distribuciones discretas especiales
- 3 Distribuciones continuas especiales
- 4 Sumas de variables independientes
- 5 La media muestral $\bar{X}_n$
- 6 Teorema del Límite Central (TLC)
- 7 Introducción a la estimación**
- 8 Simular en R
- 9 Manos a la obra

Estimador vs. parámetro

Definición 4 (Parámetro y estimador)

Un **parámetro** (*parameter*) θ es una constante que indexa una familia de distribuciones (μ, σ^2 de la normal; λ de la exponencial; p de la binomial). Un **estimador** (*estimator*) $\hat{\theta}$ es una función de la muestra usada para aproximar θ .

Tres palabras que no hay que confundir

- **Parámetro** θ : el valor verdadero (fijo, desconocido).
- **Estimador** $\hat{\theta}$: la *regla* (v.a., tiene distribución).
- **Estimación** (estimate): el *número* al meter los datos en la regla.

Distribución muestral y sesgo (bias)

Distribución muestral (sampling distribution)

Como $\hat{\theta}$ es una v.a., *tiene su propia distribución*. Sobre ella juzgamos si un estimador es bueno (¿está centrado?, ¿es preciso?).

Definición 5 (Inssegadez (unbiasedness))

$\hat{\theta}$ es **inssegado** para θ si $\mathbb{E}[\hat{\theta}] = \theta$ para todo θ . El **sesgo** es

$$\text{Bias}(\hat{\theta}) = \mathbb{E}[\hat{\theta}] - \theta.$$

Conexión con lo de hoy

La media muestral $\bar{X}_n$ es **inssegada** para μ , porque $\mathbb{E}[\bar{X}_n] = \mu$.

¿Qué es un buen estimador? El tiro al blanco

Dos virtudes (la diana tiene centro = θ)

- **Poco sesgo** = *apuntar bien*: en promedio, los dardos caen centrados en el blanco ($\mathbb{E}[\hat{\theta}] = \theta$).
- **Poca varianza** = *pulso firme*: los dardos caen juntos, no dispersos.

Las cuatro combinaciones

- Insesgado + poca var.: **el ideal**.
- Insesgado + mucha var.: centrado pero disperso.
- Sesgado + poca var.: agrupado, zona equivocada.
- Sesgado + mucha var.: lo peor.

Un solo número: el MSE

$$\text{MSE} = \text{sesgo}^2 + \text{varianza}.$$

“Qué tan lejos del centro caen los dardos, en promedio.”

Demo resuelto: estimar θ con $X \sim \text{Unif}[0, \theta]$

Problema

Sea $X \sim \text{Unif}[0, \theta]$, de modo que $\mathbb{E}[X] = \theta/2$. Proponemos el estimador $\hat{\theta} = 2\bar{X}_n$. **Se pide:** mostrar que es **insesgado**.

Solución

Usando $\mathbb{E}[\bar{X}_n] = \mathbb{E}[X] = \theta/2$ y la linealidad de la esperanza:

$$\mathbb{E}[\hat{\theta}] = \mathbb{E}[2\bar{X}_n] = 2\mathbb{E}[\bar{X}_n] = 2 \cdot \frac{\theta}{2} = \theta.$$

Como $\mathbb{E}[\hat{\theta}] = \theta$, el sesgo es 0: $\hat{\theta} = 2\bar{X}_n$ es **insesgado**.

Matiz

$\hat{\theta} = \max\{X_1, \dots, X_n\}$ es *sesgado* ($\leq \theta$): insesgadez no es lo único que importa.

Contenido

- 1 Motivación: distribuciones en la práctica (MLOps)
- 2 Distribuciones discretas especiales
- 3 Distribuciones continuas especiales
- 4 Sumas de variables independientes
- 5 La media muestral $\bar{X}_n$
- 6 Teorema del Límite Central (TLC)
- 7 Introducción a la estimación
- 8 Simular en R**
- 9 Manos a la obra

En R: muestrear una distribución y verla

Idea

Las funciones `r*` muestrean; `d*` dan la pmf/pdf teórica. Muestreamos mucho y comparamos el histograma con la curva teórica: **deben coincidir**.

Código (córrelo en RStudio)

```
set.seed(2026)
# Poisson(3): teorica vs simulada
y <- rpois(100000, lambda = 3)
barplot(table(y)/length(y), col = "lightgreen")
points(0:12 + 0.7, dpois(0:12, 3), pch = 19, col = "red")

# Exponencial(1.5): histograma vs pdf
e <- rexp(100000, rate = 1.5)
hist(e, breaks = 60, freq = FALSE, col = "lightblue")
curve(dexp(x, rate = 1.5), add = TRUE, lwd = 2, col = "red")
```

En R: estimar $\mathbb{E}[X]$ y $\text{Var}(X)$ por simulación

Verificar fórmulas a mano

Con muchas réplicas, `mean()` y `var()` se acercan a los valores teóricos (ley de los grandes números).

Código

```
set.seed(2026); N <- 200000
```

```
y <- rpois(N, 3)           # E[X]=Var=3  
c(mean(y), var(y))        # ~ 3 , ~ 3
```

```
e <- rexp(N, rate = 1.5)   # E=1/1.5, Var=1/1.5^2  
c(mean(e), var(e))        # ~ 0.667 , ~ 0.444
```

```
b <- rbinom(N, 10, 0.4)    # E=np=4, Var=npq=2.4  
c(mean(b), var(b))        # ~ 4 , ~ 2.4
```

En R: ver el TLC con distribuciones madre distintas

El experimento clave

Tomamos 3 distribuciones muy distintas (uniforme, exponencial, Bernoulli) y promediamos n datos muchas veces. Al crecer n , el histograma de $\bar{X}_n$ se **vuelve normal**.

Código

```
set.seed(2026); B <- 20000
tlc <- function(rngen, n, mu, sigma, tit) {
  m <- replicate(B, mean(rngen(n)))
  hist(m, breaks = 40, freq = FALSE, main = paste(tit, n))
  curve(dnorm(x, mu, sigma/sqrt(n)), add = TRUE, col = "blue", lwd = 2)
}
par(mfrow = c(3,3))
for (n in c(1,5,30)) tlc(function(n) runif(n),      n, 0.5, 1/sqrt(12), "Unif")
for (n in c(1,5,30)) tlc(function(n) rexp(n,1),     n, 1, 1, "Exp")
for (n in c(1,5,30)) tlc(function(n) rbinom(n,1,.3),n, .3, sqrt(.21), "Bern")
```

Contenido

- 1 Motivación: distribuciones en la práctica (MLOps)
- 2 Distribuciones discretas especiales
- 3 Distribuciones continuas especiales
- 4 Sumas de variables independientes
- 5 La media muestral $\bar{X}_n$
- 6 Teorema del Límite Central (TLC)
- 7 Introducción a la estimación
- 8 Simular en R
- 9 Manos a la obra

¿Y ahora?

- Revisa el **PDF de teoría** para el detalle (más distribuciones, demostraciones del TLC y propiedades de estimadores).
- Practica con el **PDF de ejercicios** (3 niveles: Basic · Core · Challenge).
- Verifica con el **PDF de soluciones** paso a paso.
- En **R**: `dbinom`, `dpois`, `pnorm`, `qnorm`, `rnorm` y simular promedios con `replicate(..., mean(...))` para ver el TLC.

¡A practicar!