Estimadores, Intervalos de Confianza y Pruebas de Hipótesis

Clase 6 — 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.

clazoq@uni.pe · linkedin.com/in/cristian-lazo-quispe · github.com/CristianLazoQuispe

¿Qué veremos hoy?

1. Motivación: ¿cuánta infraestructura necesita un proyecto de datos?
2. Evaluar estimadores: sesgo, varianza y MSE
3. Derivar estimadores: momentos y máxima verosimilitud
4. Intervalos de confianza
5. Pruebas de hipótesis
6. Errores y potencia
7. Manos a la obra

Contenido

Arquitectura mínima (1/3): publicar una app

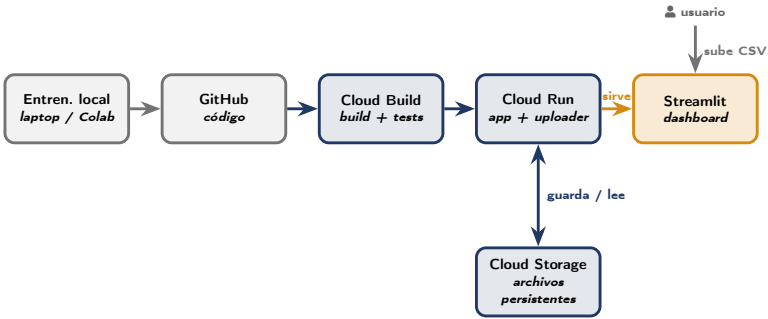
Un colegio que solo quiere analizar los datos de sus estudiantes *no* necesita el nivel 2 de la Clase 5. Lo mínimo: entrenas en local y despliegas con GitHub + Cloud Build + Cloud Run —sin base de datos.

■ tu entorno ■ servicios GCP (la nube) ■ tablero

push a GitHub ⇒ Cloud Build construye y prueba ⇒ Cloud Run sirve tu app. **Escala a cero**: si nadie la usa, no pagas.

Arquitectura mínima (2/3): subir y guardar datos (Cloud Storage)

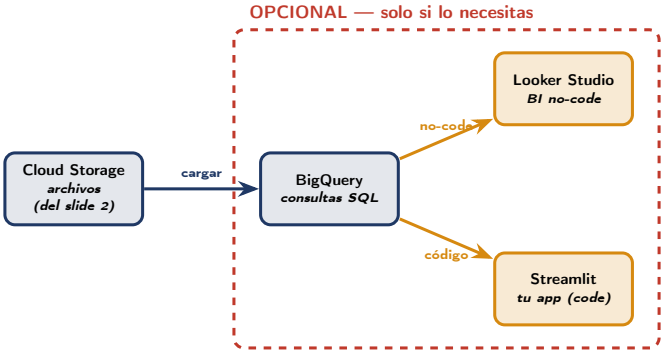
¿De dónde salen los datos? Los sube el usuario en la app (Streamlit). Cloud Run los guarda en Cloud Storage (persistente) y los lee de vuelta para mostrarlos. Así el dato entra por la subida, no aparece de la nada.



En vez de dejar el archivo en la **memoria** de Cloud Run (se borra al reiniciar o escalar a cero), Cloud Run lo guarda en **Cloud Storage** y de ahí lo lee para mostrarlo. Persistente y barato —**sin base de datos**.

Arquitectura mínima (3/3): consultas con BigQuery (opcional)

Opcional. Si acumulas muchos datos y necesitas consultas SQL o analítica, cargas de Cloud Storage a BigQuery y lo visualizas. Si no, con el slide anterior ya te basta.



Cuándo sí: tableros que cruzan *muchos* registros, históricos, consultas *ad-hoc* en SQL. **Cuándo no:** un archivo que subes y muestras —eso ya lo resuelve Cloud Storage. Marco de madurez: Google Cloud Architecture (clic).

¿Cuándo escalar? De la arquitectura a la estadística de hoy

Nivel 0–1 *(el colegio)*	Nivel 2 *(Clase 5)*
Entrenamiento **local**	Pipelines automáticos (Vertex AI)
Deploy simple: Cloud Build → Cloud Run	CI/CD completo
Reentrenas **a mano**	Reentrenamiento por *drift*
Pocos datos, 1 modelo	Muchos modelos, *feature store*

Escala **solo cuando el dolor lo justifique**: *YAGNI* (*"You Aren't Gonna Need It"*) —no construyas lo que aún no necesitas.

Y aquí entra la Clase 6

Ese tablero reporta **estimadores** (promedios, % de estudiantes en riesgo) con **intervalos de confianza**, y permite **pruebas de hipótesis**: ¿mejoró el rendimiento tras una intervención?

La estadística de hoy es el **motor** del tablero —y se entrega con infraestructura mínima. Las pruebas A/B enlazan con *causalidad* (clases siguientes).

Contenido

¿Qué hace "bueno" a un estimador?

La idea central

Un **estimador (estimator)** $\hat{\theta}$ es una *variable aleatoria*: tiene su propia distribución muestral. Lo juzgamos por características de esa distribución.

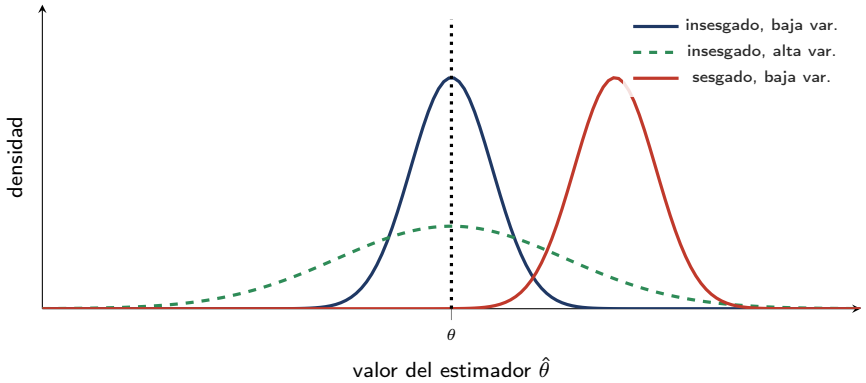
Definición 1 (Sesgo, varianza y MSE)

$$\text{Bias}(\hat{\theta}) = \mathbb{E}[\hat{\theta}] - \theta, \qquad \text{Var}(\hat{\theta}) = \mathbb{E}\big[(\hat{\theta} - \mathbb{E}[\hat{\theta}])^2\big],$$

$$\text{MSE}(\hat{\theta}) = \mathbb{E}\big[(\hat{\theta} - \theta)^2\big] = \underbrace{\text{Bias}(\hat{\theta})^2}_{\textit{centrado}} + \underbrace{\text{Var}(\hat{\theta})}_{\textit{precisión}}.$$

- **Insesgado (unbiased):** $\mathbb{E}[\hat{\theta}] = \theta$ para todo θ.
- **Consistente (consistent):** $\hat{\theta}$ colapsa a θ cuando $n \to \infty$.

Cuatro escenarios: centrado y precisión



Línea punteada: el parámetro verdadero θ. El **centrado** mide sesgo; el **ancho** mide varianza. El rojo está bien concentrado pero *desviado*.

Demo resuelto: ¿insesgado? y MSE

Problema

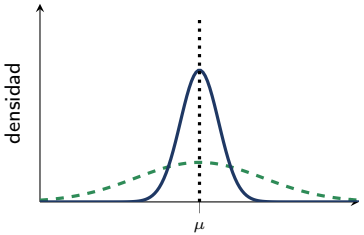
Sea $X_1, \ldots, X_n$ iid con media μ y varianza σ^2.
Considera dos estimadores de μ: $\hat{\mu}_1 = \bar{X}_n$ y $\hat{\mu}_2 = X_1$.
Se pide: ¿son insesgados? Compara sus MSE.

Solución

Ambos son insesgados: $\mathbb{E}[\bar{X}_n] = \mu$ y $\mathbb{E}[X_1] = \mu$, luego sesgo $= 0$. Como ambos están centrados, el MSE es solo varianza:

$$\mathrm{MSE}(\bar{X}_n) = \frac{\sigma^2}{n}, \qquad \mathrm{MSE}(X_1) = \sigma^2.$$

Para $n > 1$, $\bar{X}_n$ es **más eficiente**: menor varianza $\Rightarrow$ menor MSE.



Azul: $\bar{X}_n$ (angosto, σ^2/n).
Verde: X_1 (ancho, σ^2).

Mismo centro μ, distinta precisión.

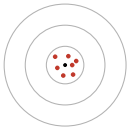
Intuición: el tiro al blanco

Sesgo vs. varianza, en una diana

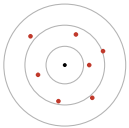
El centro es el verdadero θ; cada disparo es la estimación de una muestra.

- **Sesgo = apuntar mal:** la mira está desviada; los disparos se agrupan *lejos* del centro, de forma sistemática.
- **Varianza = pulso tembloroso:** la mano tiembla; los disparos se *dispersan* aunque apuntes al centro.

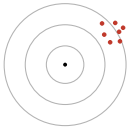
Ideal: buena puntería *y* pulso firme $\Rightarrow$ disparos agrupados *en* el centro (insesgado y baja varianza).



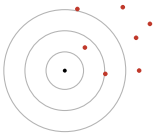
insesgado · baja var.



insesgado · alta var.



sesgado · baja var.



sesgado · alta var.

Contenido

Dos recetas para fabricar estimadores

Método de los momentos (Method of Moments, Pearson 1894)

Igualar momentos poblacionales a momentos muestrales y despejar el parámetro:

$$\mathbb{E}[X] = \frac{1}{n}\sum_i X_i, \quad \mathbb{E}[X^2] = \frac{1}{n}\sum_i X_i^2, \ \ldots$$

Con k parámetros se usan k *condiciones de momento*.

Máxima verosimilitud (Maximum Likelihood, Fisher)

Elegir el valor del parámetro que *más probablemente* generó los datos:

$$L(\theta \mid x) = \prod_{i=1}^{n} f(x_i \mid \theta), \qquad \hat{\theta}_{\text{MLE}} = \arg\max_{\theta} \ \log L(\theta \mid x).$$

Tomar log convierte el producto en suma (más fácil de derivar).

Demo resuelto: MLE de p (Bernoulli)

Problema

$X_1, \ldots, X_n$ iid Bernoulli(p) (éxito/fracaso). Observamos $\sum_i x_i = s$ éxitos. **Se pide:** el estimador de máxima verosimilitud de p.

Solución

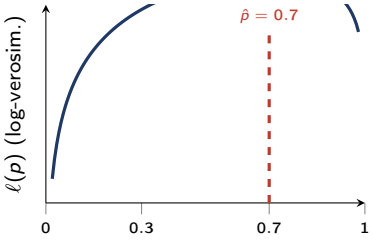
La log-verosimilitud es

$$\ell(p) = s \log p + (n - s) \log(1 - p).$$

Derivando e igualando a cero:

$$\ell'(p) = \frac{s}{p} - \frac{n - s}{1 - p} = 0 \implies \boxed{\hat{p} = \frac{s}{n} = \bar{X}_n.}$$

¡La proporción muestral! Coincide con el método de momentos ($\mathbb{E}[X] = p$).



Ejemplo $n = 10$, $s = 7$ éxitos.

El máximo de $\ell(p)$ está en $\hat{p} = s/n = 0.7$.

MLE en una frase (sin fórmulas)

La pregunta del MLE

*¿Qué valor del parámetro hace **más probables** los datos que ya observé?*

Ejemplo de la urna

Sacas 20 bolas y 6 son rojas. ¿Qué proporción de rojas p es la más *creíble*?

- Si $p = 0.9$: ver solo 6 rojas de 20 sería rarísimo.
- Si $p = 0.05$: también sería rarísimo.
- El valor que vuelve "lo más natural posible" el resultado es $\hat{p} = 6/20 = 0.3$. **Ese es el MLE.**

Ojo

El MLE no busca qué parámetro es "verdad" en abstracto, sino cuál *explica mejor* lo que el azar te entregó. Es bueno asintóticamente, pero puede ser **sesgado** con pocos datos.

Contenido

Intervalo de confianza para la media

Definición 2 (Intervalo de confianza (confidence interval))

Un IC al $1-\alpha$ *para* θ *es un par de funciones de la muestra* $[A, B]$ *tales que* $\mathbb{P}(A < \theta < B) = 1-\alpha$ *antes de observar los datos.*

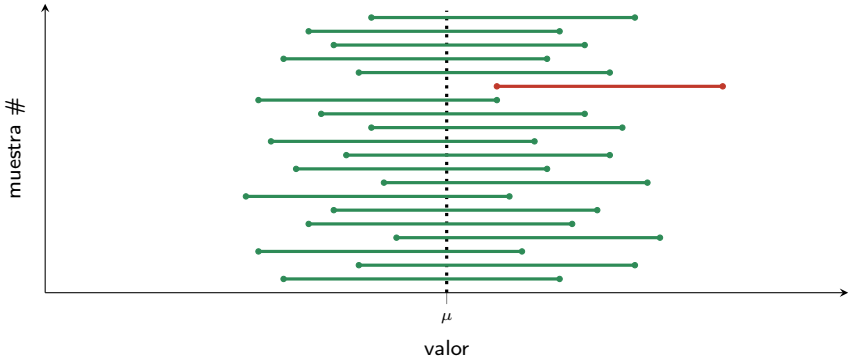
Dos casos clave (muestra normal, media μ)

- **Caso 1 — σ conocida:** $\bar{X}_n \pm z_{1-\alpha/2}\,\dfrac{\sigma}{\sqrt{n}}$, con $z_{0.975} = 1.96$ al 95 %.
- **Caso 2 — σ desconocida:** $\bar{X}_n \pm t_{n-1,\,1-\alpha/2}\,\dfrac{s}{\sqrt{n}}$ (la t tiene colas más anchas $\Rightarrow$ IC más ancho).

Cuidado de interpretación

Una vez con números, θ está dentro o no: la "confianza" es sobre el *procedimiento*, no sobre un intervalo concreto.

¿Qué significa “95 % de confianza”? (cobertura)



Cada barra es un IC del 95 % de una muestra distinta. La mayoría **cubren** μ; en promedio $\sim$1 de cada 20 **falla** (aquí, la muestra #15).

El "95 %": lo correcto vs. el error más común

✓ Correcto (interpretación frecuentista)

Si repitiéramos el muestreo muchas veces y construyéramos un IC cada vez, $\sim 95\,\%$ de esos intervalos contendrían a μ. La confianza está en el **procedimiento**, no en un intervalo particular.

× Incorrecto (el error más frecuente)

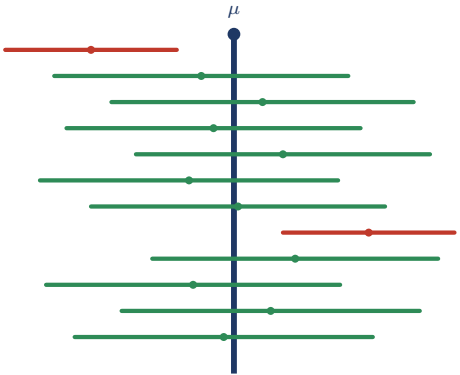
"Hay 95 % de probabilidad de que $\mu \in [a, b]$". Una vez calculado, $[a, b]$ son números fijos y μ es una constante: o está dentro o no —ya *no hay azar* (prob. $= 1$ ó 0). Lo aleatorio era el *intervalo*, no μ.

Ancho del intervalo

Más confianza ($90 \to 95 \to 99\,\%$) $\Rightarrow$ IC más **ancho**; más datos (n) $\Rightarrow$ IC más **angosto** (margen $\propto 1/\sqrt{n}$).

$\Rightarrow$ En la siguiente lámina, el "mapa de argollas" lo muestra de un vistazo.

El "95 %" de un vistazo: lanzar argollas a un poste fijo



— rodea el poste — falla

cada argolla = un IC de una muestra (centro = $\bar{X}$)

$\approx$ **95 de cada 100 rodean el poste μ**

✓ Correcto

Si lanzo **muchas** argollas (muchas muestras), 95 de cada 100 rodean el poste μ. La confianza está en el **procedimiento**.

× Incorrecto

"Esta argolla tiene 95 % de probabilidad de rodear el poste". **No**: una vez lanzada, ya lo rodeó o no. μ es **fija**; lo aleatorio era la *argolla*.

Tamaño de la argolla

más confianza → argolla **más grande** (IC más ancho);
más datos → argolla **más precisa** (IC más angosto).

¿Muchos intervalos o uno solo? (la duda más común)

En la práctica calculas UN solo IC

Tienes **una** muestra $\Rightarrow$ calculas **un** intervalo. Ese reportas y usas.

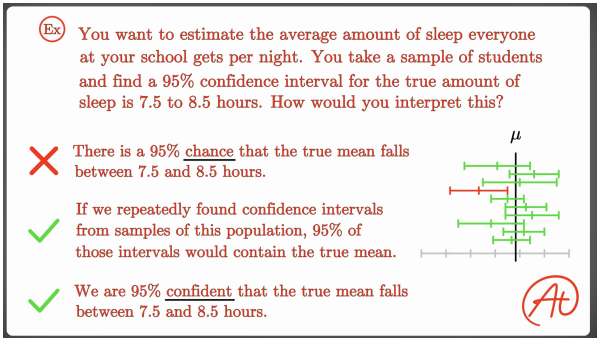
Entonces, ¿de dónde salen los "muchos"?

Son un **experimento mental**: *si* repitieras el estudio con muestras nuevas, cada una daría un IC *distinto*, y el 95 % de esos intervalos imaginarios contendrían a μ. Tú solo ves **el tuyo**; le pones "95 %" porque la **receta** acierta 95 de cada 100 veces.

En la analogía de las argollas

Lanzas **una** argolla (tu muestra $\rightarrow$ tu IC). El "95 %" es tu **puntería** (de imaginar muchos lanzamientos). Tu argolla ya cayó: confías en ella *porque tu técnica acierta* 95 *de* 100. Los "muchos" son hipotéticos; el "uno" es el tuyo.

Video recomendado: interpretar un intervalo de confianza



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

Interpreting Confidence Intervals

Ace Tutors — el error más común al concluir un IC.

- **Mal** (×): "hay 95 % de *probabilidad* de que μ caiga en el intervalo". μ es fijo; el intervalo es lo aleatorio.
- **Bien** (✓): de muchas muestras, el 95 % de los intervalos contendría a μ; estamos 95 % *confiados* en **el nuestro**.
- *Probabilidad* (antes de muestrear) vs. *confianza* (después): justo lo de esta sección.

youtube.com/watch?v=ftYdEm6pEkE

Demo resuelto: IC del 95 % para la media

Problema

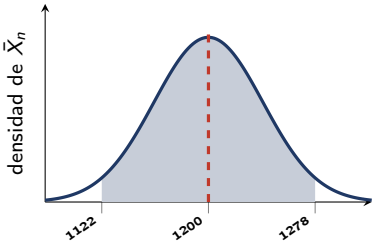
El ingreso mensual de una muestra de $n = 64$ trabajadores tiene media $\bar{X}_n = 1200$ soles. Se asume $\sigma = 320$ soles (conocida). **Se pide:** un IC del 95 % para μ.

Solución

Con $z_{0.975} = 1.96$ y error estándar $\sigma/\sqrt{n} = 320/8 = 40$:

$$\bar{X}_n \pm z_{0.975}\,\frac{\sigma}{\sqrt{n}} = 1200 \pm 1.96 \cdot 40 = 1200 \pm 78.4.$$

$$\boxed{[1121.6,\ 1278.4] \text{ soles.}}$$



Área sombreada = 95 % de la distribución de $\bar{X}_n$ (centro 1200).

Contenido

Anatomía de una prueba de hipótesis

Ingredientes

- **Hipótesis:** H_0 (nula, la que se pone a prueba) vs. H_A (alternativa).
- **Estadístico de prueba (test statistic):** resume la muestra, p. ej. Z.
- **Nivel de significancia α:** prob. de rechazar H_0 siendo cierta.
- **Región de rechazo (critical region):** valores del estadístico que llevan a rechazar H_0.
- ***p*-valor:** prob., *si H_0 es cierta*, de un estadístico tan o más extremo que el observado.

Regla de decisión

Rechazar H_0 si el estadístico cae en la región de rechazo $\iff$ si p-valor $< \alpha$.

Guía de decisión: ¿cómo identifico H_0 y H_1?

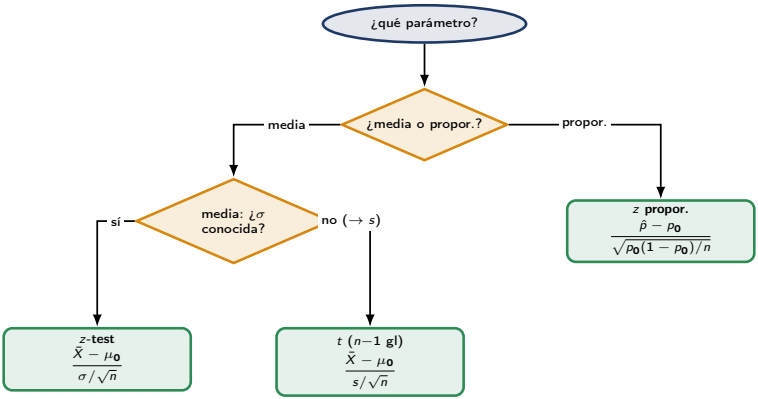
Regla de oro

- H_0 **(nula)** = **statu quo** ("no hay efecto"). *Siempre* con **igualdad**: $=$ (o $\leq/\geq$). **La igualdad vive en H_0.**
- H_1 **(alternativa)** = **lo que quiero detectar** (la sospecha). *Nunca* igualdad: usa $\neq$, $>$ o $<$.
- **Truco:** "¿qué quiero *demostrar*?" $\rightarrow H_1$. El signo sale de la pregunta: *sube* $\rightarrow >$, *baja* $\rightarrow <$, *difiere* $\rightarrow \neq$.

De la frase a las hipótesis

Pregunta	H_0	H_1	Colas
¿el programa *sube* el ingreso?	$\mu = \mu_0$	$\mu > \mu_0$	una
¿la media *difiere* de 100?	$\mu = 100$	$\mu \neq 100$	dos
¿la nueva tasa es *menor*?	$p = p_0$	$p < p_0$	una
¿la proporción *supera* 0.6?	$p = 0.6$	$p > 0.6$	una

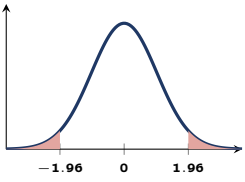
¿z-test, t-test u otro? (diagrama de flujo)



En palabras: media con σ conocida (o n grande) $\to z$; media con σ desconocida (usas s) $\to t$ con $n-1$ gl; proporción $\to z$. La t tiene **colas más gruesas** porque estimar σ añade ruido; al crecer n, $t_{n-1} \to \mathcal{N}(0,1)$.

¿Una cola o dos colas? Los tres casos (lo decide H_1)

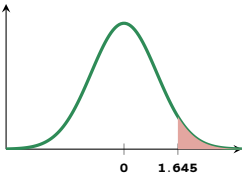
dos colas: $H_1 : \mu \neq \mu_0$



Rechazo en *ambos* extremos; ± 1.96 (2.5 % c/u).

Ej.: ¿el gasto promedio *difiere* de S/ 500?

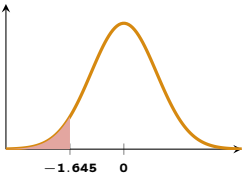
cola der.: $H_1 : \mu > \mu_0$



Rechazo en el extremo *derecho*; 1.645 (5 %).

Ej.: ¿el programa *aumenta* el ingreso?

cola izq.: $H_1 : \mu < \mu_0$



Rechazo en el extremo *izquierdo*; -1.645 (5 %).

Ej.: ¿la nueva máquina *reduce* el tiempo de espera?

Criterio y valores críticos (verificados)

H_1 con $\neq \Rightarrow$ **dos colas**, $z_{0.975} = 1.96$ (rechaza si $|z| > 1.96$). H_1 con $>$ o $< \Rightarrow$ **una cola**, $z_{0.95} = 1.645$ (todo el α en el extremo de la dirección de H_1).

La RECETA: checklist universal de 6 pasos

1. **Plantear** H_0 **y** H_1 (H_0 con "$=$"; H_1 con $>, <, \neq$). Esto fija las colas.
2. **Elegir test y nivel** α**:** media σ conocida $\to z$; σ desconocida $\to t$ ($n-1$ gl); proporción $\to z$.
3. **Calcular el estadístico:** $\dfrac{\text{estimación} - \text{valor nulo}}{\text{error estándar}}$.
4. **Hallar la región crítica** (valor crítico) *o* el p-valor.
5. **Decidir:** rechaza H_0 si cae en la zona de rechazo (o $p < \alpha$); si no, no rechazas.
6. **Interpretar en palabras** y en contexto: ¿hay efecto?, ¿de qué tamaño?, ¿qué error podría cometer?

Recuerda

"No rechazar" $\neq$ "probar H_0". Reporta también **tamaño del efecto** e **IC**, no solo el p-valor.

Demo A: z-test de DOS colas (σ conocida)

Problema

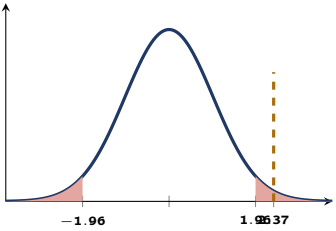
Bolsas de $\mu_0 = 500$ g, $\sigma = 8$ *conocida*. Muestra $n = 40$, $\bar{X} = 503$. ¿La media *difiere* de 500? ($\alpha = 0.05$.)

Solución (6 pasos)

$H_0 : \mu = 500$ vs $H_1 : \mu \neq 500$ (**dos colas**). z-test (σ conocida), críticos ± 1.96.

$$z = \frac{503 - 500}{8/\sqrt{40}} = \frac{3}{1.265} \approx 2.37.$$

$|z| = 2.37 > 1.96 \Rightarrow$ **rechazo** H_0;
$p = 2(1 - \Phi(2.37)) \approx 0.018$. La media *difiere* de 500 g.



z (bajo H_0)

$z_{\text{obs}} = 2.37$ cae en rechazo.

Demo B: z-test de UNA cola (σ conocida)

Problema

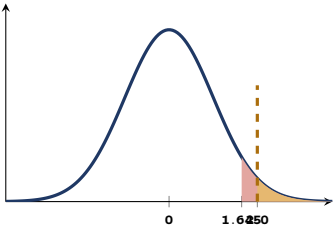
¿Una capacitación *aumenta* la productividad media?
$\mu_0 = 50$, $\sigma = 10$ *conocida*, $n = 64$, $\bar{X} = 52.5$.
($\alpha = 0.05$.)

Solución

$H_0 : \mu = 50$ vs $H_1 : \mu > 50$ (**una cola der.**). z-test, crítico 1.645.

$$z = \frac{52.5 - 50}{10/\sqrt{64}} = \frac{2.5}{1.25} = 2.0.$$

$z = 2.0 > 1.645 \Rightarrow$ **rechazo** H_0;
$p = 1 - \Phi(2.0) \approx 0.023$. El programa *subió* la productividad.



z (bajo H_0)

rojo: rechazo; ámbar: p-valor ≈ 0.023.

Demo C: t-test (σ desconocida, n chico)

Problema

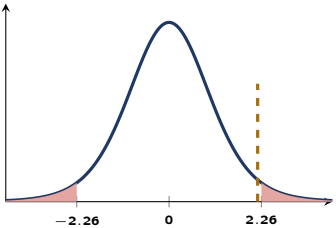
¿El rendimiento medio *difiere* de $\mu_0 = 30$? $n = 10$ parcelas, $\bar{X} = 32.4$, $s = 3.5$ (σ **desconocida**). ($\alpha = 0.05$.)

Solución

$H_0 : \mu = 30$ vs $H_1 : \mu \neq 30$ (**dos colas**). t-test, gl $= 9$, crítico $t_{9,0.975} \approx 2.262$.

$$t = \frac{32.4 - 30}{3.5/\sqrt{10}} = \frac{2.4}{1.107} \approx 2.17.$$

$|t| = 2.17 < 2.262 \Rightarrow$ **NO rechazo** (¡por poco!); $p \approx 0.058$. Con $n = 10$ la potencia es baja.



t (9 gl)

$t_{\text{obs}} = 2.17$ no entra a rechazo (línea ámbar).

Demo D: prueba de PROPORCIÓN (z)

Problema

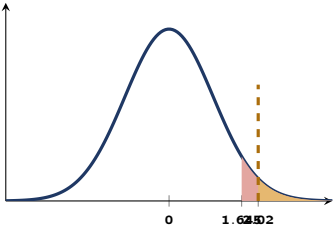
Antes renovaba el $p_0 = 60\,\%$. Tras una mejora, de $n = 200$ usuarios renuevan 134 ($\hat{p} = 0.67$). ¿La proporción *supera* 0.60? ($\alpha = 0.05$.)

Solución

$H_0 : p = 0.60$ vs $H_1 : p > 0.60$ (**una cola der.**). z de proporción, $\text{SE} = \sqrt{\frac{0.6 \cdot 0.4}{200}} = 0.0346$.

$$z = \frac{0.67 - 0.60}{0.0346} \approx 2.02.$$

$z = 2.02 > 1.645 \Rightarrow$ **rechazo** H_0; $p \approx 0.022$. La renovación *subió* sobre el 60 %.



z (bajo H_0)

rojo: rechazo; ámbar: p-valor ≈ 0.022.

Simulación: α como tasa de falsos positivos

Idea

$\alpha = \mathbb{P}(\text{rechazar } H_0 \mid H_0 \text{ cierta})$. Generamos muchas muestras *bajo* H_0 y contamos qué fracción rechaza: debe rondar α.

En R

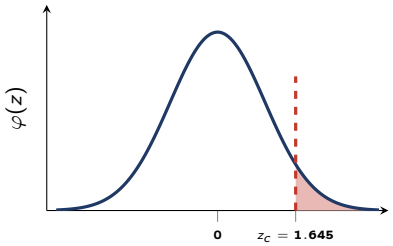
```
set.seed(2026); mu0 <- 100; sigma <- 15; n <- 30
alpha <- 0.05; B <- 20000
rechaza <- replicate(B, {
  x <- rnorm(n, mean = mu0, sd = sigma)  # H0 CIERTA
  t.test(x, mu = mu0)$p.value < alpha    # TRUE = falso positivo
})
mean(rechaza)   # ~ 0.05  ==> coincide con alpha
```

Lectura

Aunque H_0 *es* cierta, $\sim 5\,\%$ de los experimentos la rechaza por azar: **eso es** α. Bajar α a 0.01 baja los falsos positivos (pero también la potencia para efectos reales).

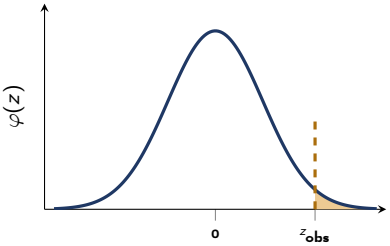
Región de rechazo y p-valor (prueba Z de cola derecha)

región de rechazo ($\alpha = 0.05$)



Área roja $= \alpha = 0.05$ a la derecha del valor crítico $z_c = 1.645$.

p-valor del estadístico observado



Área ámbar $= p$-valor (cola más allá de z_{obs}). Si $p < \alpha$, se rechaza H_0.

Demo resuelto: prueba Z de una media

Problema

Una política busca elevar el gasto medio en salud por encima de $\mu_0 = 500$ soles. Muestra: $n = 100$, $\bar{X}_n = 512$, $\sigma = 60$ conocida. Con $\alpha = 0.05$ y $H_0 : \mu = 500$ vs. $H_A : \mu > 500$, ¿se rechaza H_0?

Solución

Estadístico de prueba bajo H_0:

$$z_{\text{obs}} = \frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}} = \frac{512 - 500}{60/\sqrt{100}} = \frac{12}{6} = 2.0.$$

Como es cola derecha al 5 %, el valor crítico es $z_c = 1.645$. Dado que $z_{\text{obs}} = 2.0 > 1.645$, **se rechaza** H_0. Equivalente vía p-valor: $p = 1 - \Phi(2.0) \approx 0.023 < 0.05$.

Errores comunes al interpretar pruebas de hipótesis

1. El p-valor NO es $\mathbb{P}(H_0$ cierta)

Es $\mathbb{P}$(datos así de extremos $\mid H_0$), *no* $\mathbb{P}(H_0 \mid$ datos). Confundir $\mathbb{P}(A \mid B)$ con $\mathbb{P}(B \mid A)$ es la *falacia del fiscal*.

2. "No rechazar" $\neq$ "aceptar H_0"

Ausencia de evidencia $\neq$ evidencia de ausencia. Quizá el efecto existe pero la muestra fue chica.

3. Significancia $\neq$ importancia

Con n enorme, un efecto *minúsculo* (sin relevancia práctica) puede salir "significativo". Reporta **tamaño del efecto** e **IC**, no solo p.

4. "p-hacking" / muchas hipótesis

Probar 20 hipótesis al 5 % con *todas* falsas $\Rightarrow$ esperas ≈ 1 "significativa" por azar. Remedio: preinscribir y corregir (α/m, Bonferroni).

Metáfora

Si 1000 personas tiran una moneda justa 5 veces, *alguien* sacará 5 caras ($p \approx 0.03$) sin tener una moneda mágica. Eso es buscar el "afortunado" y llamarlo descubrimiento.

Contenido

Tipo I, tipo II y potencia

La tabla 2×2 de decisiones

	H_0 verdadera	H_0 falsa
No rechazar H_0	decisión correcta	error tipo II (β)
Rechazar H_0	error tipo I (α)	decisión correcta (potencia)

Definición 3 (Significancia, operación y potencia)

$$\alpha = \mathbb{P}(\textit{rechazar } H_0 \mid H_0), \qquad \beta = \mathbb{P}(\textit{no rechazar } H_0 \mid H_A),$$

$$\textit{potencia} = 1 - \beta = \mathbb{P}(\textit{rechazar } H_0 \mid H_A).$$

Hay un **intercambio**: bajar α (mover el umbral) sube β; subir n encoge ambas distribuciones y mejora las dos.

Casos cotidianos: ¿qué es cada error?

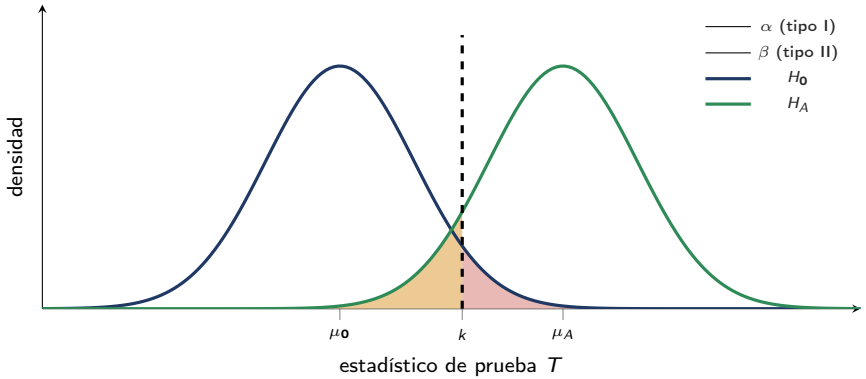
H_0 = "no pasa nada / statu quo".

Contexto	H_0	**Tipo I (α):** rechazar H_0 cierta	**Tipo II (β):** no rechazar H_0 falsa
Justicia penal	el acusado es inocente	condenar a un *inocente*	liberar a un *culpable*
Tamizaje médico	la persona está sana	alarmar a un *sano*	no detectar al *enfermo*
Programa social (RCT)	el programa *no* sirve	escalar algo inútil (gasto)	descartar algo que *sí* servía
Filtro de spam	el correo es legítimo	bueno → spam	dejar pasar spam

¿Cuál es peor? Depende del costo

En justicia, condenar a un inocente (tipo I) es lo más grave ⇒ "inocente hasta que se pruebe lo contrario" fija H_0 = inocente y exige evidencia fuerte (α chico).

Las dos densidades: α y β sombreados



Umbral k. Cola **roja** de $H_0 = \alpha$; zona **ámbar** de H_A a la izquierda de $k = \beta$; la potencia es el resto de H_A $(1 - \beta)$.

Demo resuelto: calcular α, β y potencia

Problema

X_i iid $N(\mu, \sigma^2)$ con $\sigma^2 = 4$, $n = 25$, de modo que $T = \bar{X}_n \sim N(\mu,\, 4/25)$, es decir desviación 0.4. Probamos $H_0 : \mu = 0$ vs. $H_A : \mu = 1$ con umbral $k = 0.66$ (rechazar si $T > k$). **Se pide:** α, β y la potencia.

Solución

$$\alpha = \mathbb{P}(T > 0.66 \mid \mu = 0) = 1 - \Phi\Big(\tfrac{0.66-0}{0.4}\Big) = 1 - \Phi(1.65) \approx 0.05.$$

$$\beta = \mathbb{P}(T < 0.66 \mid \mu = 1) = \Phi\Big(\tfrac{0.66-1}{0.4}\Big) = \Phi(-0.85) \approx 0.20.$$

Por tanto la **potencia** es $1 - \beta \approx 0.80$. Subir n estrecharía ambas densidades y reduciría β manteniendo α.

Demo accesible: control de calidad (falsa alarma vs. efecto perdido)

Problema

Una máquina debe hacer tornillos de $\mu_0 = 20$ mm; al desgastarse, los hace más largos. Cada hora: muestra $n = 36$, $\sigma = 0.6$. Test $H_0 : \mu = 20$ vs. $H_A : \mu > 20$; si rechazo, **paro la línea** para recalibrar.

¿Qué es cada error?

- **Tipo I (falsa alarma):** parar una máquina que estaba *bien* (producción detenida en vano).
- **Tipo II (efecto perdido):** *no* parar una máquina ya descalibrada (lote defectuoso).

Números

SE $= 0.6/\sqrt{36} = 0.1$. Con $\alpha = 0.05$ (cola derecha):

$$k = 20 + 1.645(0.1) = 20.16.$$

Si descalibrada da $\mu_1 = 20.30$:

$$\beta = \Phi\Big(\tfrac{20.16-20.30}{0.1}\Big) = \Phi(-1.40) \approx 0.08.$$

$$\textbf{potencia} = 1 - \beta \approx 0.92.$$

Si la descalibración fuera *sutil* ($\mu_1 = 20.10$), la potencia cae a ≈ 0.26: los problemas pequeños son los más fáciles de "perder".

Contenido

¿Y ahora?

- Revisa el **PDF de teoría**: sesgo–varianza, distribuciones $\chi^2/t/F$, derivación de IC y de pruebas "desde cero".
- Practica con el **PDF de ejercicios** (3 niveles: Basic · Core · Challenge).
- Verifica con el **PDF de soluciones** paso a paso.
- En **R**: `qnorm`, `qt`, `t.test`, `prop.test`, `power.t.test` y `replicate(...)` para *ver* la cobertura de los intervalos y la potencia.

¡A practicar!