14.310x – Data Analysis for Social Scientists

MicroMasters en Statistics and Data Science (SDS) – MITx

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

Clase 6

Estimadores, Intervalos de Confianza y Pruebas de Hipótesis

Teoría

Módulos oficiales del MicroMaster:
M6 – Assessing and Deriving Estimators, Confidence Intervals and Hypothesis Testing

B.Sc. Cristian Lazo Quispe
clazoq@uni.pe
linkedin.com/in/cristian-lazo-quispe · github.com/CristianLazoQuispe

Material del *Advanced Program in Data Science & Global Skills*, diseñado y desarrollado por **BREIT (Brescia Institute of Technology)**, iniciativa de **Aporta** —plataforma de innovación e impacto social del **Grupo Breca**— en alianza con el **MIT IDSS** (Institute for Data, Systems, and Society).

11 de agosto de 2026

Índice

1. De qué trata esta clase

En la clase anterior dimos el salto de la *probabilidad* a la *estadística*: vimos que un **estimador (estimator)** $\hat{\theta}$ es una función de la muestra —por tanto, una variable aleatoria— que usamos para adivinar un **parámetro (parameter)** θ desconocido de la población. Ahora vamos a tres preguntas centrales de toda la inferencia estadística:

- **¿Cómo sé si un estimador es bueno?** Criterios para *evaluar* estimadores: sesgo, varianza, error cuadrático medio, consistencia y eficiencia.
- **¿Cómo construyo un estimador desde cero?** Dos métodos clásicos para *derivar* estimadores: el método de los momentos y la máxima verosimilitud (MLE).
- **¿Qué tan confiable es mi estimación?** Dos formas de cuantificar la incertidumbre: los **intervalos de confianza** y las **pruebas de hipótesis** (con su análisis de errores y *potencia*).

Un estimador es una *receta* para adivinar la verdad de la población usando datos. Esta clase trata de tres habilidades del estadístico aplicado: (1) *juzgar* recetas (¿se inclina hacia un lado?, ¿cuánto varía?), (2) *inventar* recetas razonables, y (3) *reportar* no solo un número sino también cuánta confianza merece. En ciencias sociales, donde casi nunca medimos a toda la población, estas tres habilidades son el corazón del oficio.

2. Evaluar estimadores: sesgo, varianza, MSE, consistencia y eficiencia

Un estimador $\hat{\theta}$ es una variable aleatoria, así que tiene una **distribución muestral (sampling distribution)**. Todos los criterios para juzgarlo se basan en características de esa distribución: ¿está centrada en el valor correcto? ¿qué tan dispersa es?

2.1. Sesgo (bias)

Definición 2.1 (Sesgo y estimador insesgado)**.** El **sesgo (bias)** de un estimador $\hat{\theta}$ de θ es

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

$\hat{\theta}$ es **insesgado (unbiased)** si $\mathbb{E}[\hat{\theta}] = \theta$ para todo valor posible del parámetro $\theta \in \Theta$: en promedio, da en el blanco.

El sesgo mide un error *sistemático*: si repitiéramos el muestreo muchas veces, ¿el promedio de nuestras estimaciones coincidiría con la verdad, o se inclinaría siempre hacia arriba o hacia abajo? Un estimador insesgado no se inclina; no garantiza acertar en *una* muestra,

sino no equivocarse *en promedio*.

Ejemplo 2.1 (La media y la varianza muestral)**.** Si $X_1, \ldots, X_n$ son iid con media μ y varianza σ^2:

- La **media muestral** $\bar{X} = \frac{1}{n}\sum_i X_i$ es *insesgada* para μ, pues $\mathbb{E}[\bar{X}] = \mu$.
- La **varianza muestral** con divisor $n-1$,

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2,$$

es *insesgada* para σ^2: $\mathbb{E}[s^2] = \sigma^2$. (El divisor $n-1$ en vez de n corrige el sesgo que aparecería por usar $\bar{X}$ en lugar del μ verdadero.)

2.2. Varianza y eficiencia (efficiency)

Definición 2.2 (Eficiencia entre estimadores insesgados)**.** Dados *dos* estimadores insesgados $\hat{\theta}_1$ y $\hat{\theta}_2$ de θ, decimos que $\hat{\theta}_1$ es **más eficiente (more efficient)** que $\hat{\theta}_2$ si, para un mismo tamaño de muestra,

$$\mathrm{Var}(\hat{\theta}_1) < \mathrm{Var}(\hat{\theta}_2).$$

Entre estimadores insesgados, *menor varianza = mejor*.

Si dos recetas aciertan en promedio (ambas insesgadas), preferimos la que *varía menos*: nos da estimaciones más parecidas entre muestras, es decir, más precisas y confiables. La eficiencia compara esa precisión.

Piensa en un estimador como un tirador que dispara a una diana, donde el *centro* es el verdadero θ y cada disparo es la estimación de una muestra distinta:

- **Sesgo = apuntar mal.** Si la mira está desviada, los disparos se agrupan lejos del centro de forma *sistemática*, no importa cuántas veces dispares. Eso es un estimador sesgado.
- **Varianza = pulso tembloroso.** Si te tiembla la mano, los disparos se dispersan por toda la diana aunque apuntes bien al centro. Eso es alta varianza.

El *ideal* es buena puntería (insesgado) *y* pulso firme (baja varianza): disparos agrupados *en* el centro. Un estimador puede fallar por cualquiera de los dos males —o por ambos— y el MSE (más adelante) los suma para darnos una sola nota.

2.3. Error cuadrático medio (MSE)

A veces conviene aceptar un *poco* de sesgo a cambio de *mucha* menos varianza. El criterio que combina ambos en una sola cifra es el error cuadrático medio.

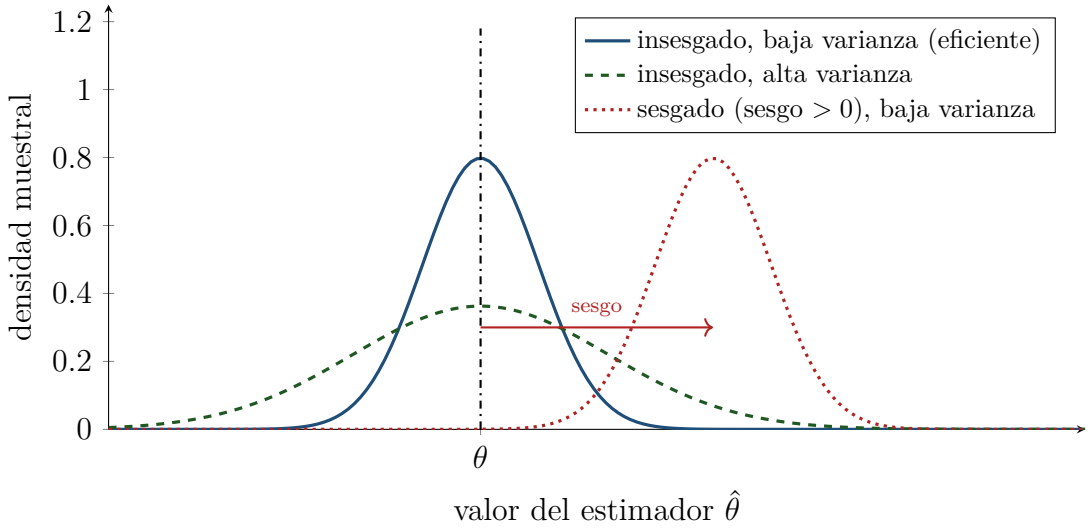


Figura 1: Distribuciones muestrales de tres estimadores de un mismo θ (línea negra). El **azul** está centrado en θ (insesgado) y es angosto (baja varianza): el mejor. El **verde** también está centrado en θ (insesgado) pero es ancho (alta varianza): menos eficiente. El **rojo** es angosto pero está *corrido* a la derecha (sesgado): da estimaciones precisas pero sistemáticamente equivocadas.

Definición 2.3 (Error cuadrático medio (Mean Squared Error))**.** El **MSE** de $\hat{\theta}$ es el promedio del error al cuadrado:

$$\mathrm{MSE}(\hat{\theta}) = \mathbb{E}\Big[(\hat{\theta} - \theta)^2\Big].$$

Descomposición sesgo–varianza

El MSE se separa en dos piezas:

$$\boxed{\mathrm{MSE}(\hat{\theta}) = \Big[\mathrm{Bias}(\hat{\theta})\Big]^2 + \mathrm{Var}(\hat{\theta})}$$

es decir, **MSE** = **sesgo**^{2} + **varianza**. Para un estimador insesgado el sesgo es 0, así que $\mathrm{MSE}(\hat{\theta}) = \mathrm{Var}(\hat{\theta})$.

Deducción. Sumando y restando $\mathbb{E}[\hat{\theta}]$ dentro del cuadrado:

$$\mathbb{E}\Big[(\hat{\theta} - \theta)^2\Big] = \mathbb{E}\Big[(\hat{\theta} - \mathbb{E}[\hat{\theta}] + \mathbb{E}[\hat{\theta}] - \theta)^2\Big] = \underbrace{\mathbb{E}\Big[(\hat{\theta} - \mathbb{E}[\hat{\theta}])^2\Big]}_{\mathrm{Var}(\hat{\theta})} + \underbrace{(\mathbb{E}[\hat{\theta}] - \theta)^2}_{\mathrm{sesgo}^2},$$

porque el doble producto $2(\mathbb{E}[\hat{\theta}] - \theta)\,\mathbb{E}[\hat{\theta} - \mathbb{E}[\hat{\theta}]] = 0$.

El MSE penaliza los dos males a la vez: estar *descentrado* (sesgo) y estar *disperso* (varianza). Esto formaliza el famoso **compromiso sesgo–varianza (bias–variance trade-**

off): a veces un estimador ligeramente sesgado pero muy estable tiene *menor* MSE que uno insesgado pero muy ruidoso. En la Figura 1, el rojo (sesgado, baja varianza) podría tener menor MSE que el verde (insesgado, alta varianza) si el sesgo es pequeño comparado con la diferencia de varianzas.

Ejemplo 2.2 (Comparar dos estimadores por MSE)**.** Para estimar θ, tenemos dos recetas:

- $\hat{\theta}_A$: insesgado, con $\text{Var}(\hat{\theta}_A) = 4 \Rightarrow \text{MSE}_A = 0^2 + 4 = 4$.
- $\hat{\theta}_B$: sesgo $= 1$, con $\text{Var}(\hat{\theta}_B) = 2 \Rightarrow \text{MSE}_B = 1^2 + 2 = 3$.

Aunque $\hat{\theta}_B$ es *sesgado*, tiene *menor* MSE ($3 < 4$): en promedio se equivoca menos. Si nos importa el error total (y no solo no inclinarnos), $\hat{\theta}_B$ es preferible. Este es el trade-off en acción.

2.4. Consistencia (consistency)

Definición 2.4 (Estimador consistente)**.** $\hat{\theta}_n$ (basado en n datos) es **consistente (consistent)** para θ si converge en probabilidad al parámetro al crecer la muestra:

$$\hat{\theta}_n \xrightarrow[n\to\infty]{p} \theta, \qquad \text{es decir,} \quad \mathbb{P}\Big(|\hat{\theta}_n - \theta| > \varepsilon\Big) \to 0 \ \text{ para todo } \varepsilon > 0.$$

Consistencia es una promesa *asintótica*: con cada vez más datos, la distribución del estimador se "colapsa" en un punto, justo en θ. Una condición suficiente cómoda: si $\hat{\theta}_n$ es insesgado (o su sesgo $\to 0$) y $\text{Var}(\hat{\theta}_n) \to 0$, entonces $\hat{\theta}_n$ es consistente. La media muestral lo cumple: $\mathbb{E}[\bar{X}] = \mu$ y $\text{Var}(\bar{X}) = \sigma^2/n \to 0$ (es la *Ley de los Grandes Números*).

Son criterios *distintos*. Un estimador puede ser insesgado pero inconsistente (p. ej. usar solo X_1 siempre: $\mathbb{E}[X_1] = \mu$ pero su varianza no baja con n). Puede ser sesgado pero consistente (p. ej. un sesgo que se desvanece con n). "Bueno" depende de qué te importa: no inclinarte (insesgado), precisión (eficiente), o garantía con muchos datos (consistente). El MSE intenta resumirlos.

3. Derivar estimadores: momentos y máxima verosimilitud

Ya sabemos juzgar un estimador; ¿pero cómo *conseguimos* uno? Hay dos marcos clásicos: el método de los momentos y la máxima verosimilitud.

3.1. Método de los momentos (Method of Moments)

Definición 3.1 (Momentos poblacionales y muestrales)**.** El k-ésimo **momento poblacional** es $\mathbb{E}[X^k]$ (una función del parámetro). El k-ésimo **momento muestral** es $\frac{1}{n}\sum_{i=1}^{n} X_i^k$ (calculable con los datos).

Receta del método de los momentos

Para estimar un parámetro, *iguala* el primer momento poblacional al primer momento muestral y *despeja* el parámetro. Si hay k parámetros, usa k momentos (k ecuaciones, k incógnitas). Cada igualdad es una *condición de momento (moment condition)*. Idea de Karl Pearson (1894).

Ejemplo 3.1 (Momentos para λ de una Poisson)**.** Sea $X_1, \ldots, X_n$ iid Pois(λ) (p. ej. número de protestas por distrito en un mes). El primer momento poblacional es $\mathbb{E}[X] = \lambda$. Igualando al primer momento muestral $\bar{X}$:

$$\lambda = \bar{X} \implies \boxed{\hat{\lambda}_{\text{MoM}} = \bar{X}}$$

El estimador por momentos de la tasa de Poisson es, simplemente, el promedio de los conteos. Intuitivo: la media de una Poisson *es* λ.

3.2. Máxima verosimilitud (Maximum Likelihood, MLE)

Definición 3.2 (Función de verosimilitud y MLE)**.** Dada una muestra iid con densidad/pmf $f(x \mid \theta)$, la **función de verosimilitud (likelihood)** es la densidad conjunta vista *como función del parámetro*:

$$L(\theta \mid x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i \mid \theta).$$

El **estimador de máxima verosimilitud (MLE)** $\hat{\theta}_{\text{MLE}}$ es el valor de θ que *maximiza* L (equivalentemente, la log-verosimilitud $\ell(\theta) = \log L(\theta)$, porque el log es creciente y convierte el producto en suma).

El MLE responde: "¿qué valor del parámetro hace que los datos que *de hecho* observamos sean lo más probable posible?". Entre todas las distribuciones de una familia, elegimos la que mejor "explica" la muestra. Por eso es tan usado: tiene buenas propiedades (es consistente y, bajo condiciones de regularidad, asintóticamente normal y eficiente).

Imagina que sacas de una urna 20 bolas y 6 salen rojas. ¿Qué proporción de rojas p es la más *creíble* en esa urna? Si dijeras $p = 0.9$, ver solo 6 rojas de 20 sería rarísimo; si dijeras $p = 0.05$, también. El valor que vuelve "lo más natural posible" el resultado observado es $p = 6/20 = 0.3$. Eso *es* el MLE: **el valor del parámetro que hace más probables**

los datos que ya viste. No buscamos qué parámetro es "verdad" en abstracto, sino cuál *explica mejor* lo que el azar nos entregó.

Ejemplo 3.2 (MLE de p en una Bernoulli)**.** Sea $X_1, \ldots, X_n$ iid Bernoulli(p) (p. ej. $X_i = 1$ si el hogar i tiene acceso a internet). Con $f(x \mid p) = p^x(1-p)^{1-x}$, la verosimilitud es

$$L(p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{s}(1-p)^{n-s}, \qquad s = \textstyle\sum_i x_i \text{ (número de éxitos).}$$

Tomamos log y derivamos:

$$\ell(p) = s\ln p + (n-s)\ln(1-p), \qquad \ell'(p) = \frac{s}{p} - \frac{n-s}{1-p} = 0.$$

Despejando: $s(1-p) = (n-s)p \Rightarrow s = np \Rightarrow$

$$\boxed{\hat{p}_{\text{MLE}} = \frac{s}{n} = \bar{X}}$$

El MLE de p es la proporción muestral de éxitos. (Aquí coincide con el método de los momentos, porque $\mathbb{E}[X] = p$.)

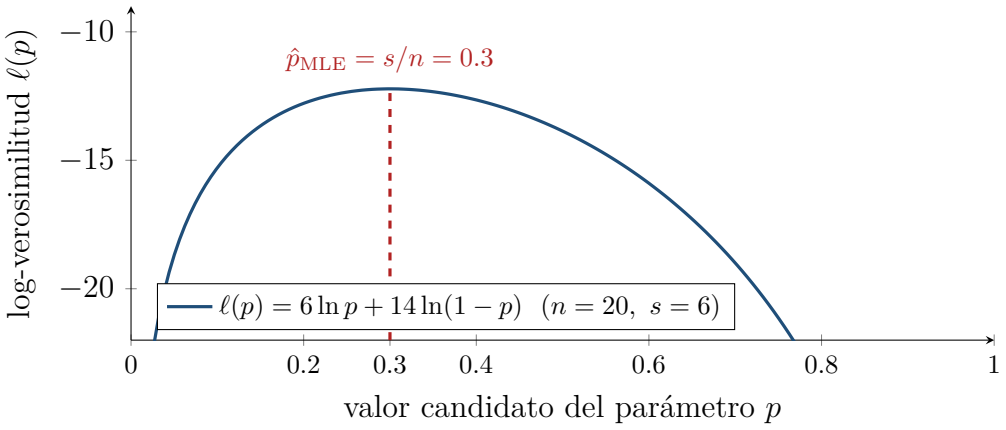


Figura 2: Log-verosimilitud de una muestra Bernoulli con $n = 20$ ensayos y $s = 6$ éxitos. La curva se maximiza exactamente en $\hat{p} = s/n = 6/20 = 0.3$ (línea roja): el valor de p que hace "más probable" haber observado 6 éxitos en 20 intentos.

Cuando el modelo tiene *dos* parámetros —por ejemplo una Normal con media μ y desviación σ desconocidas— la log-verosimilitud $\log L(\mu, \sigma)$ ya no es una curva sino una **superficie** (una "loma") sobre el plano de parámetros. El MLE $(\hat{\mu}, \hat{\sigma})$ es simplemente la **cima** de esa loma: el punto del espacio de parámetros que hace más creíbles los datos observados (figura 3). Con k parámetros viviríamos en una loma de k dimensiones; la idea —buscar la cima— es la misma.

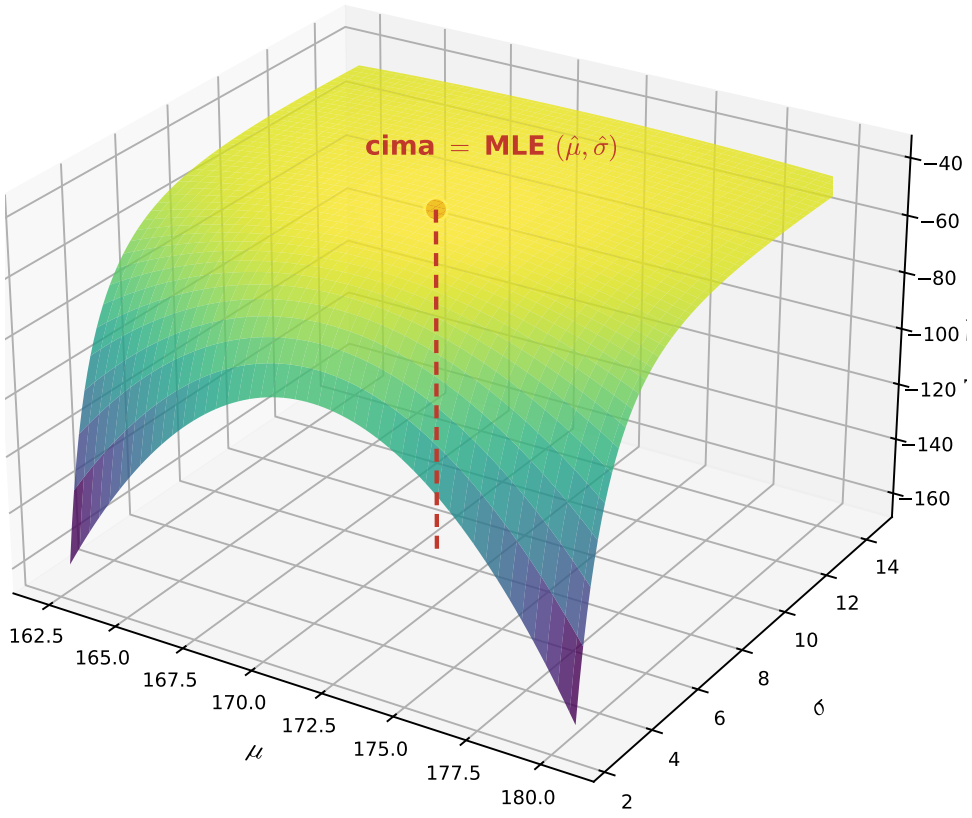


Figura 3: Log-verosimilitud $\log L(\mu, \sigma)$ de una muestra Normal, como superficie sobre el espacio de parámetros (μ, σ). El **MLE** es la **cima** (punto rojo): $\hat{\mu} = \bar{X}$ y $\hat{\sigma} = \sqrt{\frac{1}{n}\sum_i (X_i - \bar{X})^2}$. Maximizar la verosimilitud es, literalmente, "subir hasta lo más alto" de esta loma.

La máxima verosimilitud tiene grandes propiedades *asintóticas*, pero no es mágica: el MLE puede ser **sesgado** en muestras finitas. Ejemplo clásico: con $X_i \sim \text{Unif}[0, \theta]$, el MLE de θ es el máximo muestral $\hat{\theta} = X_{(n)} = \text{máx}\{X_1, \ldots, X_n\}$, que *siempre* es $\leq \theta$ y por tanto *subestima* θ (es sesgado hacia abajo, aunque consistente).

4. Intervalos de confianza (confidence intervals)

Reportar solo un número ($\hat{\theta} = 0.42$) esconde cuán precisos somos. El **error estándar (standard error, SE)** —la desviación estándar del estimador— ya da una idea: para $\bar{X}$, $\text{SE}(\bar{X}) = \sigma/\sqrt{n}$. Pero a menudo es más útil presentar esa información como un *intervalo*.

Definición 4.1 (Intervalo de confianza)**.** Un **intervalo de confianza (confidence interval, CI)** de nivel $1 - \alpha$ para θ es un par de funciones de la muestra $A(X_1, \ldots, X_n)$ y $B(X_1, \ldots, X_n)$ tales que

$$\mathbb{P}\Big(A < \theta < B\Big) = 1 - \alpha.$$

El número $1-\alpha$ es el **nivel de confianza**; lo común es $1-\alpha = 0.95$ (95 por ciento). Tras evaluar en los datos, $[A, B]$ es un par de números concretos.

CI para la media: σ conocida (Normal)

Si muestreamos de una $\mathcal{N}(\mu, \sigma^2)$ con σ *conocida* (o n grande, por el TLC), el CI de nivel $1-\alpha$ para μ es

$$\boxed{\bar{X} \pm z_{1-\alpha/2}\,\frac{\sigma}{\sqrt{n}}}$$

donde $z_{1-\alpha/2}$ es el cuantil de la normal estándar. Para 95 %: $z_{0.975} = 1.96$. Es decir, **estimación ± margen de error**, con margen $= z \cdot \text{SE}$.

CI para la media: σ desconocida (t de Student)

Si σ es *desconocida* (lo habitual), la estimamos con s y usamos la distribución **t de Student** con $n-1$ grados de libertad, porque

$$\frac{\bar{X}-\mu}{s/\sqrt{n}} \sim t_{n-1}.$$

El CI de nivel $1-\alpha$ para μ es

$$\boxed{\bar{X} \pm t_{n-1,\,1-\alpha/2}\,\frac{s}{\sqrt{n}}}$$

La t tiene **colas más gruesas** que la normal, así que $t_{n-1,\,1-\alpha/2} > z_{1-\alpha/2}$: el intervalo es *más ancho*. Esto "castiga" nuestra mayor incertidumbre por no conocer σ. Cuando $n \to \infty$, $t_{n-1} \to \mathcal{N}(0,1)$ y ambos intervalos coinciden.

Ejemplo 4.1 (Ingreso medio con σ conocida)**.** Una muestra de $n = 100$ hogares da un ingreso promedio $\bar{X} = 2000$ soles. Suponemos $\sigma = 600$ soles *conocida*. Un CI al 95 % para el ingreso medio poblacional μ:

$$\bar{X} \pm 1.96\,\frac{\sigma}{\sqrt{n}} = 2000 \pm 1.96 \cdot \frac{600}{\sqrt{100}} = 2000 \pm 1.96 \cdot 60 = 2000 \pm 117.6,$$

es decir, $[1882.4,\ 2117.6]$ soles.

Interpretación en palabras (correcta): "Con un 95 % de confianza, el ingreso medio poblacional está entre 1882.4 y 2117.6 soles". Es decir, la receta que produjo este intervalo atrapa al verdadero μ el 95 % de las veces; en *este* caso confiamos (sin certeza) en que μ quedó adentro. *Evita* decir "hay 95 % de probabilidad de que $\mu \in [1882.4, 2117.6]$".

Ejemplo 4.2 (Misma muestra, σ desconocida (t))**.** Si en cambio *no* conocemos σ y la muestra da $s = 600$ soles, usamos la t con $n-1 = 99$ gl. El cuantil $t_{99,\,0.975} \approx 1.984$ (apenas mayor que 1.96):

$$\bar{X} \pm t_{99,\,0.975}\,\frac{s}{\sqrt{n}} = 2000 \pm 1.984 \cdot 60 = 2000 \pm 119.0 \ \Rightarrow\ [1881.0,\ 2119.0].$$

El intervalo t es ligeramente más ancho que el normal, reflejando la incertidumbre extra de estimar σ. Con $n = 99$ la diferencia es pequeña; con n chico sería notoria.

Interpretación (correcta): "Con 95 % de confianza, el ingreso medio poblacional está entre 1881.0 y 2119.0 soles". El intervalo t es un poco más ancho porque pagó el "impuesto" de no conocer σ y haberla estimado con los mismos datos.

4.1. ¿Qué significa REALMENTE el "95 %"? (lo más importante)

Esta es, con diferencia, la idea que más se malinterpreta en toda la estadística. Vále la pena ir despacio.

Interpretación frecuentista correcta del nivel de confianza

El "95 %" describe el **procedimiento**, no un intervalo concreto:

> *Si repitiéramos el muestreo muchísimas veces —sacando cada vez una muestra nueva y construyendo un CI con la misma receta— alrededor del* 95 % *de* esos *intervalos contendrían al verdadero* μ.

La aleatoriedad vive en *cómo cae la muestra* (y por tanto en dónde caen los extremos A y B del intervalo), no en μ. El parámetro μ es un número **fijo** de la población, que no conocemos pero no se mueve. Lo que "acierta el 95 % de las veces" es la receta de construir intervalos.

Piensa en el juego de **lanzar argollas** a un **poste** clavado en el suelo (o el **tiro al sapo**: la boca del sapo no se mueve de su sitio). El poste es μ: está *fijo*, solo que no sabemos exactamente dónde. Cada vez que tomas una muestra es como **lanzar una argolla**: cae en una posición un poco distinta cada vez, porque cada muestra sale distinta. Esa argolla es tu intervalo de confianza.
"Confianza del 95 %" quiere decir que, si lanzaras muchas argollas, 95 **de cada** 100 **rodearían el poste**. Pero una argolla *ya lanzada* o rodeó el poste o no lo hizo: es un hecho consumado, ya no hay azar. Por eso **no** se dice "esta argolla tiene 95 % de probabilidad de rodear el poste". El 95 % describe tu *puntería* (el procedimiento de muestreo), no una tirada concreta. En la figura de cobertura de abajo, cada línea horizontal es una argolla (el IC de una muestra): casi todas cruzan la línea vertical μ, y solo unas pocas la dejan fuera.

En tu estudio real calculas UN solo intervalo, a partir de *tu* única muestra. Ese es el que reportas y usas. Entonces, ¿de dónde salen los "muchos intervalos" del 95 %?
Son un **experimento mental**: *si* repitieras el estudio con muestras nuevas (o si muchos investigadores lo hicieran), cada muestra daría un intervalo *distinto*, porque cada muestra sale distinta. De *todos* esos intervalos imaginarios, el 95 % contendría a μ. Tú solo ves

el tuyo, pero como tu intervalo salió de una *receta* que acierta el 95 % de las veces, le pones el sello "95 % de confianza".
En la analogía: lanzas **una** argolla (tu muestra → tu IC). El "95 %" es lo que sabes de tu *puntería* (de imaginar muchos lanzamientos). Tu argolla ya cayó y ya rodeó el poste o no; confías en ella *porque tu técnica acierta* 95 *de cada* 100, no porque "esta" tirada tenga azar. Los "muchos" son hipotéticos; el "uno" es el tuyo.

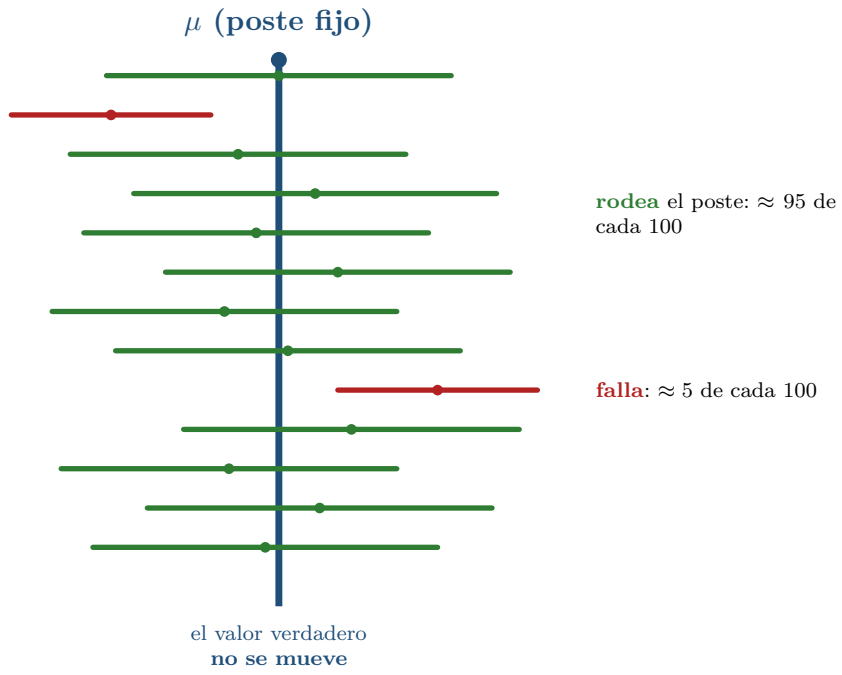


Figura 4: El "95 %" como *tiro al sapo*. El **poste** es μ: está clavado y no se mueve. Cada **argolla** (segmento horizontal, con su centro $= \bar{X}$) es el IC de una muestra distinta; lo que cambia entre muestras es *dónde cae la argolla*, no el poste. La gran mayoría **rodean** μ (verde) y solo ~ 1 de cada 20 **falla** (rojo). Por eso el "95 %" describe la *puntería del procedimiento*, no una argolla ya lanzada: una vez en el suelo, ya rodeó el poste o no.

Decir "hay 95 % de probabilidad de que μ esté en $[1882.4,\ 2117.6]$" es **incorrecto**, aunque suene natural y lo diga mucha gente (incluso en artículos publicados). ¿Por qué está mal?

- Una vez calculado, $[1882.4,\ 2117.6]$ son **números fijos** y μ es una **constante** (desconocida, pero constante). La afirmación "$1882.4 < \mu < 2117.6$" es o bien *verdadera* o bien *falsa*: su "probabilidad" es 1 o 0, no 0.95. Ya no hay azar que medir.
- Lo aleatorio era *el intervalo* (dependía de la muestra), no μ. La probabilidad 0.95 se refiere a la *familia* de intervalos que la receta produce *antes* de ver los datos, no a este intervalo ya cerrado.

Forma correcta de decirlo: "con 95 % de confianza, μ está entre 1882.4 y 2117.6 soles",

entendiendo "confianza" como "el procedimiento que me dio este intervalo acierta el 95 % de las veces". (La frase con "probabilidad" sí es válida en el enfoque *bayesiano* con un *intervalo de credibilidad*, que es otra cosa y se basa en supuestos distintos.)

El semiancho (margen de error) es $z_{1-\alpha/2}\,\sigma/\sqrt{n}$. De ahí se leen dos intuiciones útiles:

- **Más confianza ⇒ más ancho.** Pasar de 90 % a 95 % a 99 % sube z ($1.645 \to 1.96 \to 2.576$), así que el intervalo se ensancha. Es lógico: para estar *más seguro* de atrapar a μ, debo lanzar una "herradura más grande". La certeza total (100 %) exigiría un intervalo infinito —inútil.
- **Más datos ⇒ más angosto.** El $\sqrt{n}$ en el denominador hace que cuadruplicar n reduzca el margen a la mitad. Más información = más precisión = intervalo más fino. Hay un *trade-off*: precisión (intervalo angosto) vs. confianza (intervalo ancho); n es la palanca que compra ambas a la vez.

"CI del 95 %" **no** significa "hay 95 % de probabilidad de que μ esté en *este* intervalo". Una vez calculado, $[1882.4,\ 2117.6]$ son números fijos y μ es una constante: o está dentro o no, no hay azar. Lo correcto: *si repitiéramos el muestreo muchas veces y construyéramos un CI cada vez, alrededor del* 95 % *de esos intervalos contendrían a* μ. La confianza está en el *procedimiento*, no en un intervalo particular.

5. Pruebas de hipótesis (hypothesis testing)

Muchas preguntas de ciencias sociales son de tipo *sí/no*: ¿el programa social aumentó el ingreso? ¿la nueva política cambió las horas trabajadas? La **prueba de hipótesis** responde: dada una muestra, ¿hay *suficiente evidencia* para contradecir cierta afirmación sobre la población?

Definición 5.1 (Hipótesis nula y alternativa)**.** La **hipótesis nula (null hypothesis)** H_0 es la afirmación que se pone a prueba (típicamente "no hay efecto", $\theta = \theta_0$). La **hipótesis alternativa (alternative hypothesis)** H_1 (o H_A) es lo contrario. Si H_1 es $\theta \neq \theta_0$ la prueba es de **dos colas (two-sided)**; si es $\theta > \theta_0$ (o $<$), es de **una cola (one-sided)**.

Definición 5.2 (Estadístico, nivel, región de rechazo)**.** Un **estadístico de prueba (test statistic)** T resume la muestra (p. ej. $T = \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}$). El **nivel de significancia (significance level)** α es la probabilidad máxima que aceptamos de rechazar H_0 siendo cierta (común: $\alpha = 0.05$). La **región de rechazo (critical / rejection region)** es el conjunto de valores de T para los que rechazamos H_0; sus bordes son los **valores críticos**.

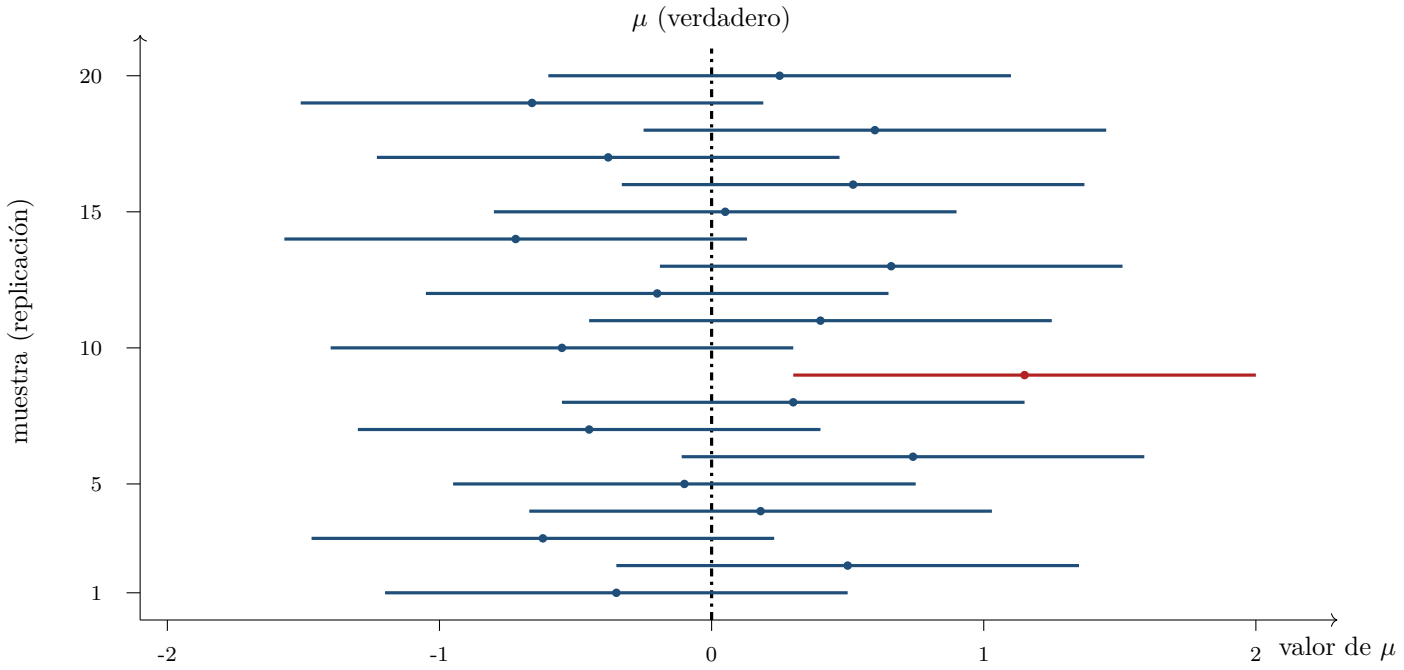


Figura 5: Veinte intervalos de confianza al 95 %, uno por muestra. El punto es la estimación $\bar{X}$ de cada muestra; la barra, su intervalo. La línea vertical es el μ *verdadero*. La mayoría (azul) **cubre** μ; uno (rojo, muestra 9) **no** lo cubre. A largo plazo, alrededor de 1 de cada 20 (5 %) fallará: eso es lo que significa "95 % de confianza".

Receta de una prueba de hipótesis

1. Plantea H_0 y H_1.
2. Fija α (p. ej. 0.05).
3. Calcula el estadístico T con los datos.
4. Compara T con el/los valor(es) crítico(s) (o calcula el p-valor).
5. **Rechaza** H_0 si T cae en la región de rechazo (o si p-valor $< \alpha$); si no, **no rechazas**.

Para un test z de dos colas al 5 %, los valores críticos son ± 1.96: rechazas si $|T| > 1.96$. Para una cola (derecha) al 5 %, el crítico es 1.645.

Definición 5.3 (p-valor (p-value))**.** El **p-valor** es la probabilidad, *suponiendo* H_0 *cierta*, de observar un estadístico tan o más extremo que el obtenido. Regla de decisión:

$$\boxed{\text{p-valor} < \alpha \implies \text{rechazar } H_0}$$

Cuanto menor el p-valor, más fuerte la evidencia contra H_0.

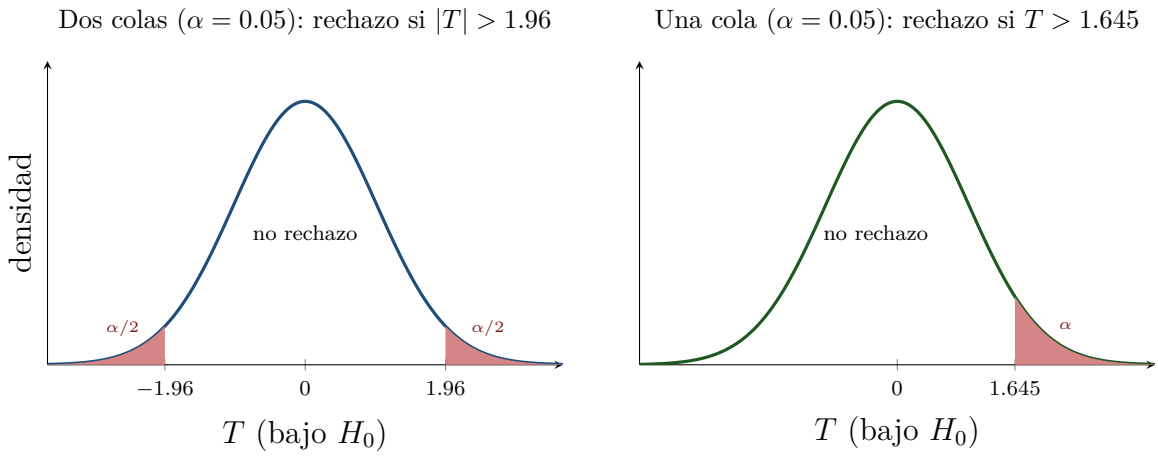


Figura 6: Regiones de rechazo (en rojo) para un test z al 5 %. **Izquierda:** dos colas, con $\alpha/2 = 2.5\,\%$ en cada extremo y valores críticos ± 1.96. **Derecha:** una cola (derecha), con todo el $\alpha = 5\,\%$ en un extremo y valor crítico 1.645. Si el estadístico observado cae en zona roja, rechazamos H_0.

El p-valor mide qué tan "sorprendentes" son los datos *si la nula fuera cierta*. Un p-valor de 0.002 dice: "datos como estos ocurrirían apenas 2 de cada 1000 veces si H_0 fuera verdad" —demasiado raro, dudamos de H_0. La regla $\text{p} < \alpha$ es *equivalente* a "T cayó en la región de rechazo": son dos formas de la misma decisión.

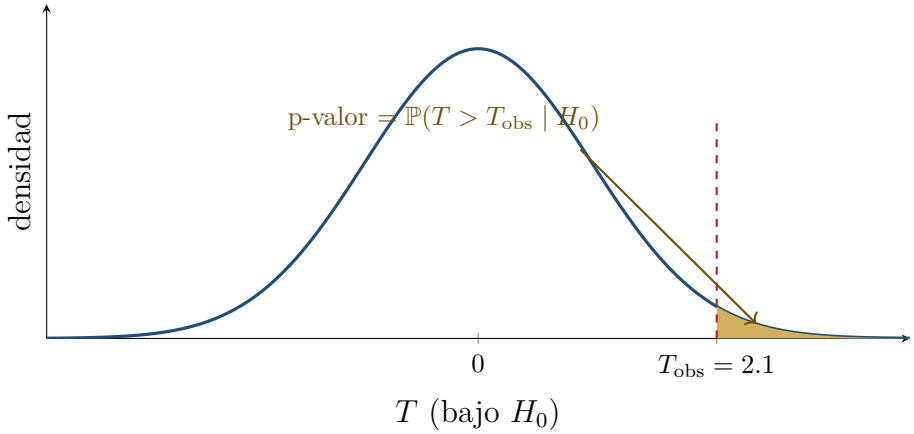


Figura 7: El **p-valor** (una cola) es el área bajo la densidad de H_0 a la derecha del estadístico observado $T_{\text{obs}} = 2.1$. Aquí vale ≈ 0.018. Como $0.018 < 0.05 = \alpha$, rechazamos H_0. (Para un test de dos colas se sumarían ambas colas, dando ≈ 0.036.)

Ejemplo 5.1 (Test sobre un puntaje promedio)**.** Un programa educativo afirma no cambiar el puntaje medio, históricamente $\mu_0 = 500$ con $\sigma = 100$ (conocida). Tomamos $n = 64$ alumnos y obtenemos $\bar{X} = 520$. Probamos $H_0 : \mu = 500$ vs $H_1 : \mu \neq 500$ al $\alpha = 0.05$. El

estadístico:

$$T = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{520 - 500}{100/\sqrt{64}} = \frac{20}{12.5} = 1.6.$$

Como $|T| = 1.6 < 1.96$, *no* rechazamos H_0. El p-valor de dos colas es $2\,(1 - \Phi(1.6)) = 2(0.0548) \approx 0.110 > 0.05$: la evidencia no basta para concluir que el programa cambió el puntaje.

No rechazar H_0 **no** demuestra que H_0 sea cierta; solo dice que *no hay evidencia suficiente* para descartarla (quizá la muestra fue chica). Y un resultado "significativo" ($p < 0.05$) no implica que el efecto sea *grande* o *importante*: la significancia estadística no es lo mismo que la relevancia práctica.

6. Guía de decisión: cómo plantear y resolver CUALQUIER prueba

Esta sección es un *manual de campo*. Responde, una por una, las dudas que más aparecen al empezar con pruebas de hipótesis: cómo escribir H_0 y H_1, qué test usar (z, t, proporción), cuántas colas, y un *checklist* universal con muchos demos resueltos.

6.1. Duda 1: ¿Cómo identifico H_0 y H_1?

Regla de oro para escribir las hipótesis

- H_0 **(nula) = el "statu quo"**: "no hay efecto", "no hay diferencia", "nada cambió". *Siempre* lleva la **igualdad**: se escribe con $=$ (o con $\leq$ / $\geq$ en pruebas de una cola). La igualdad **vive en** H_0.
- H_1 **(alternativa) = lo que se quiere *detectar***: la sospecha, la afirmación del investigador, el efecto que buscamos probar. *Nunca* lleva la igualdad: usa $\neq$, $>$ o $<$.
- **Trucos rápidos:** (1) "¿qué quiero *demostrar*?" $\rightarrow$ eso es H_1. (2) El signo de H_1 sale de la *pregunta*: "¿*sube*?" $\rightarrow >$; "¿*baja*?" $\rightarrow <$; "¿*difiere / cambia*?" $\rightarrow \neq$. (3) H_0 es lo *contrario* de H_1 y se queda con el "=".

De la frase de investigación a H_0 y H_1

Pregunta / sospecha	H_0 **(statu quo)**	H_1 **(lo que detecto)**	**Colas**
¿El programa *sube* el ingreso medio?	$\mu = \mu_0$	$\mu > \mu_0$	una (der.)
¿La nueva máquina *reduce* el tiempo medio?	$\mu = \mu_0$	$\mu < \mu_0$	una (izq.)
¿La media *difiere* de 100?	$\mu = 100$	$\mu \neq 100$	dos
¿La proporción de aprobados *supera* $p_0 = 0.6$?	$p = 0.6$	$p > 0.6$	una (der.)
¿La nueva tasa de fallas es *menor* que p_0?	$p = p_0$	$p < p_0$	una (izq.)
¿El nuevo método *cambia* la proporción?	$p = p_0$	$p \neq p_0$	dos
¿El fármaco *tiene* algún efecto (cualquiera)?	$\mu = \mu_0$	$\mu \neq \mu_0$	dos

1. **Poner el "=" en H_1.** Mal: "$H_1 : \mu = 520$". La igualdad va en H_0; H_1 describe la *desviación* ($>$, $<$ o $\neq$).
2. **Decidir las colas *después* de ver los datos.** La dirección de H_1 se fija *antes* de mirar la muestra, según la pregunta; si no, inflas los falsos positivos.
3. **Confundir lo que quieres con H_0.** Lo que quieres *probar* casi siempre es H_1. H_0 es "aburrido" (no pasa nada) y es lo que intentas *derribar* con evidencia.

6.2. Duda 2: ¿z-test, t-test u otro? (diagrama de flujo)

Qué test usar, en una regla

- **Media μ, con σ poblacional CONOCIDA** (o n grande por TLC) $\rightarrow$ **z-test**: $z = \dfrac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$.
- **Media μ, con σ DESCONOCIDA** (usas la muestral s) $\rightarrow$ **t-test de Student** con $\text{gl} = n - 1$: $t = \dfrac{\bar{X} - \mu_0}{s/\sqrt{n}}$. Para n grande, $t \approx z$.
- **Proporción p $\rightarrow$ z-test** (aprox. normal): $z = \dfrac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$ (usa p_0 en el SE bajo H_0).

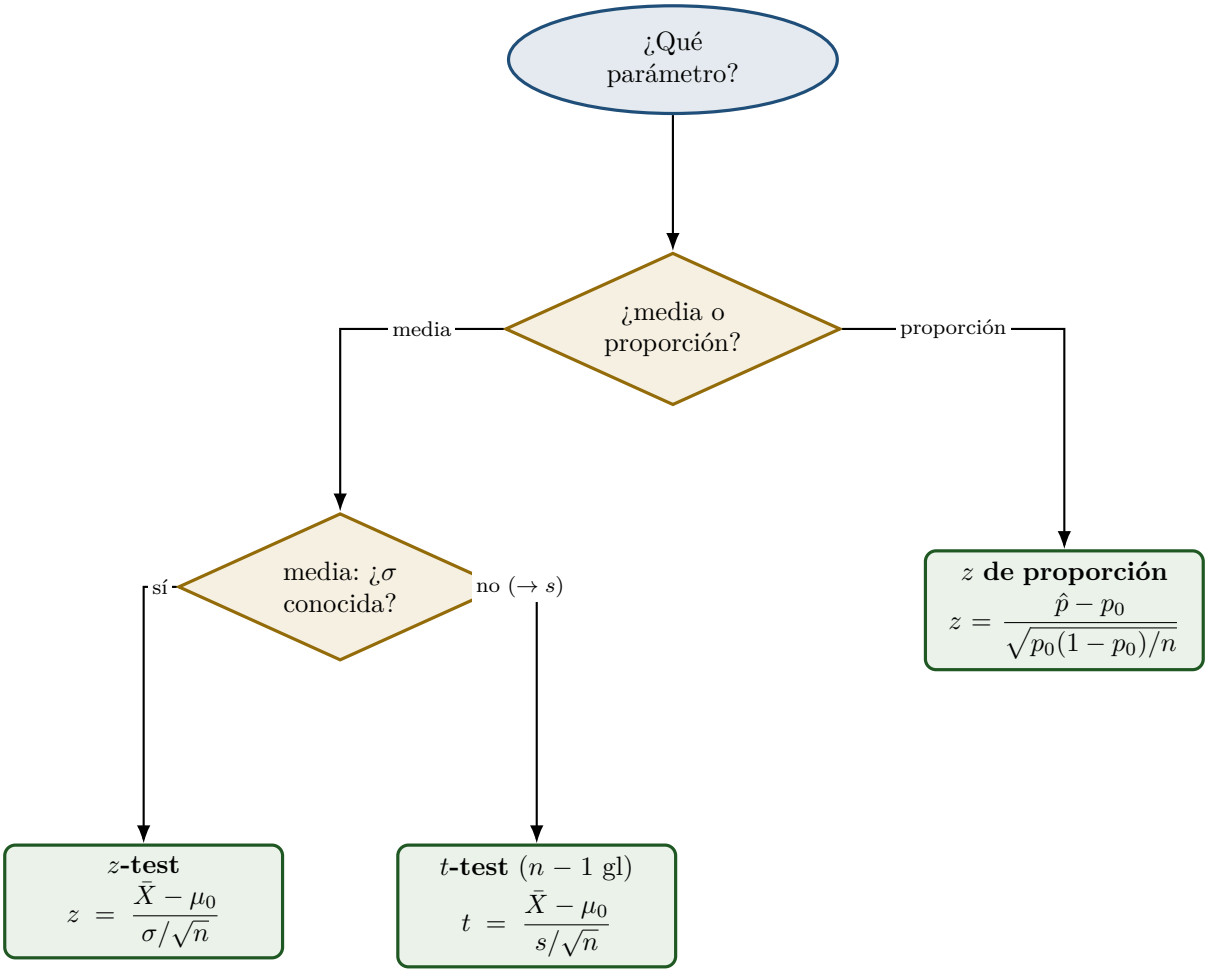


Figura 8: **Diagrama de flujo para elegir el test.** Primero pregunta qué parámetro estimas. Si es una *media*: ¿conoces σ? Sí $\rightarrow$ z-test; no $\rightarrow$ t-test con $n-1$ grados de libertad (usando s). Si es una *proporción*: z-test con el error estándar bajo H_0. (Para n grande, $t \approx z$.)

Al no conocer σ la estimamos con s, que *también* es aleatorio: añade una incertidumbre extra. La t de Student tiene *colas más gruesas* que la normal para reflejar ese "ruido adicional", así que sus valores críticos son un poco más grandes (test más conservador). Cuando n crece, s se vuelve muy preciso, las colas se adelgazan y $t_{n-1} \rightarrow \mathcal{N}(0,1)$: por eso con n grande "$t \approx z$".

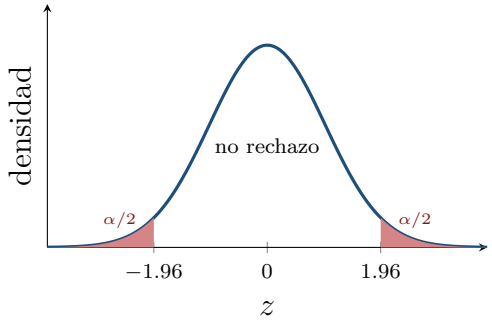
6.3. Duda 3: ¿una cola o dos colas?

El número de colas lo decide H_1

- H_1 **con $\neq$ ("difiere / cambia") → dos colas**: el α se reparte en ambos extremos ($\alpha/2$ cada uno). Crítico al 5 %: $z_{0.975} = 1.96$ (rechazas si $|z| > 1.96$).
- H_1 **direccional ($>$ o $<$) → una cola**: todo el α va a un solo extremo. Crítico al 5 %: $z_{0.95} = 1.645$ (rechazas si $z > 1.645$ para ">", o si $z < -1.645$ para "<").

Verifica los números: $z_{0.975} = 1.96$ (dos colas, deja 2.5 % en cada extremo); $z_{0.95} = 1.645$ (una cola, deja 5 % en un extremo).

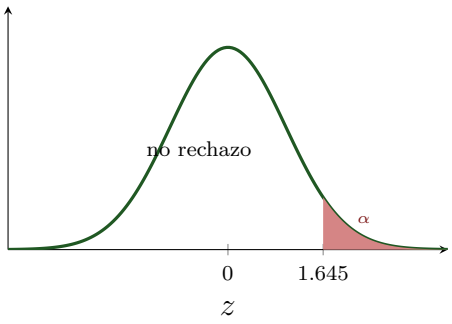
Dos colas: $H_1 : \mu \neq \mu_0$



Rechazo en *ambos* extremos; críticos ± 1.96 ($\alpha/2 = 2.5$ % c/u).

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

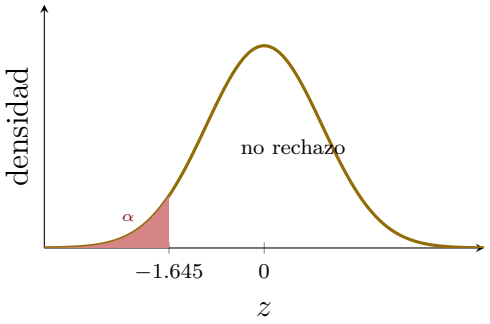
Una cola derecha: $H_1 : \mu > \mu_0$



Rechazo en el extremo *derecho*; crítico 1.645 (todo el $\alpha = 5$ %).

Ejemplo: ¿el programa *aumenta* el ingreso?

Una cola izquierda: $H_1 : \mu < \mu_0$



Rechazo en el extremo *izquierdo*; crítico -1.645 (todo el $\alpha = 5$ %).

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

Figura 9: **Los tres casos de colas** (test z al $\alpha = 0.05$). *Arriba izq.:* dos colas, $\alpha/2 = 2.5$ % en cada extremo, críticos ± 1.96. *Arriba der.:* una cola derecha, todo el 5 % en el extremo derecho, crítico 1.645. *Abajo:* una cola izquierda, todo el 5 % en el extremo izquierdo, crítico -1.645. Las pruebas de una cola tienen *más potencia* en la dirección esperada, pero la dirección debe fijarse *antes* de ver los datos.

6.4. Duda 4: la RECETA paso a paso (checklist universal)

Checklist de 6 pasos para CUALQUIER prueba de hipótesis

1. **Plantear H_0 y H_1.** H_0 con el "=" (statu quo); H_1 con $>$, $<$ o $\neq$ (lo que quiero detectar). Esto ya fija las *colas*.
2. **Elegir el test y el nivel α.** Usa la Figura 8: media con σ conocida $\to z$; media con σ desconocida $\to t$ ($n-1$ gl); proporción $\to z$. Fija α (común 0.05).
3. **Calcular el estadístico** con los datos: $\dfrac{\text{estimación} - \text{valor nulo}}{\text{error estándar}}$.
4. **Hallar la región crítica** (valor(es) crítico(s)) *o* el **p-valor**.
5. **Decidir:** rechaza H_0 si el estadístico cae en la zona de rechazo (o si p-valor $< \alpha$); si no, *no* rechazas.
6. **Interpretar en palabras** y en el contexto: ¿hay evidencia del efecto?, ¿de qué tamaño?, ¿qué error podría estar cometiendo?

6.5. Duda 5: muchos demos, uno por tipo de prueba

A continuación, un demo por cada caso, todos con la misma estructura: *enunciado* $\to$ H_0/H_1 $\to$ *test/colas* $\to$ *cálculo* $\to$ *decisión* $\to$ *interpretación*, con su gráfico de región de rechazo o p-valor.

Ejemplo 6.1 (Demo A — z-test de DOS colas (σ conocida))**. Enunciado.** Una máquina envasadora debe llenar bolsas de $\mu_0 = 500$ g. Se sabe por histórico que $\sigma = 8$ g (*conocida*). Una muestra de $n = 40$ bolsas da $\bar{X} = 503$ g. ¿La media *difiere* de 500 g? ($\alpha = 0.05$.)

Paso 1 (H_0/H_1). $H_0 : \mu = 500$ vs $H_1 : \mu \neq 500$ ("difiere" $\to$ **dos colas**).
Paso 2 (test). Media con σ *conocida* $\to$ **z-test**, $\alpha = 0.05$, críticos ± 1.96.
Paso 3 (estadístico). $z = \dfrac{503 - 500}{8/\sqrt{40}} = \dfrac{3}{1.2649} \approx 2.37$.
Paso 4–5 (decisión). $|z| = 2.37 > 1.96 \Rightarrow$ **rechazamos** H_0. p-valor $= 2\,(1 - \Phi(2.37)) \approx 2(0.0089) = 0.0178 < 0.05$.
Paso 6 (interpretación). Hay evidencia ($\alpha = 5\,\%$) de que el llenado medio *difiere* de 500 g (parece sobrellenar ~ 3 g). Conviene recalibrar.

Ejemplo 6.2 (Demo B — z-test de UNA cola (σ conocida))**. Enunciado.** Un programa de capacitación busca *aumentar* la productividad media, históricamente $\mu_0 = 50$ unidades/día con $\sigma = 10$ (*conocida*). Tras el programa, $n = 64$ trabajadores dan $\bar{X} = 52.5$. ¿El programa *aumentó* la media? ($\alpha = 0.05$.)

Paso 1. $H_0 : \mu = 50$ vs $H_1 : \mu > 50$ ("aumenta" $\to$ **una cola derecha**).
Paso 2. σ conocida $\to$ **z-test**; crítico de una cola $z_{0.95} = 1.645$.
Paso 3. $z = \dfrac{52.5 - 50}{10/\sqrt{64}} = \dfrac{2.5}{1.25} = 2.0$.

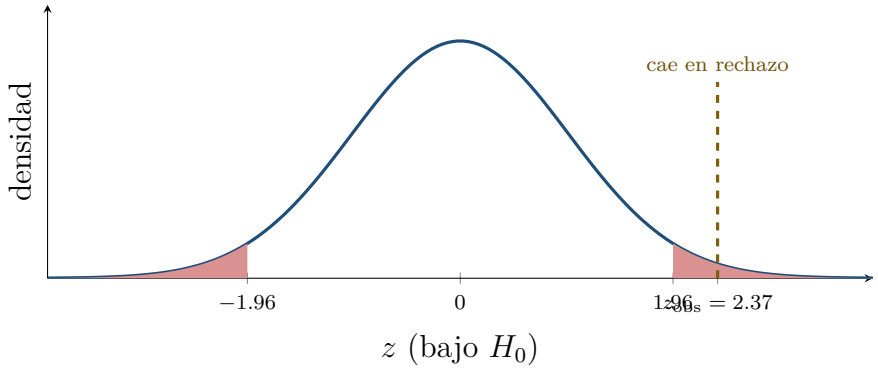


Figura 10: Demo A (dos colas). El estadístico $z_{\text{obs}} = 2.37$ cae en la cola derecha de rechazo (rojo), más allá de 1.96: rechazamos H_0.

Paso 4–5. $z = 2.0 > 1.645 \Rightarrow$ **rechazamos** H_0. p-valor $= 1 - \Phi(2.0) \approx 0.023 < 0.05$.
Paso 6. Hay evidencia de que el programa *subió* la productividad media. Reportar también el tamaño (+2.5 unidades) y su IC.

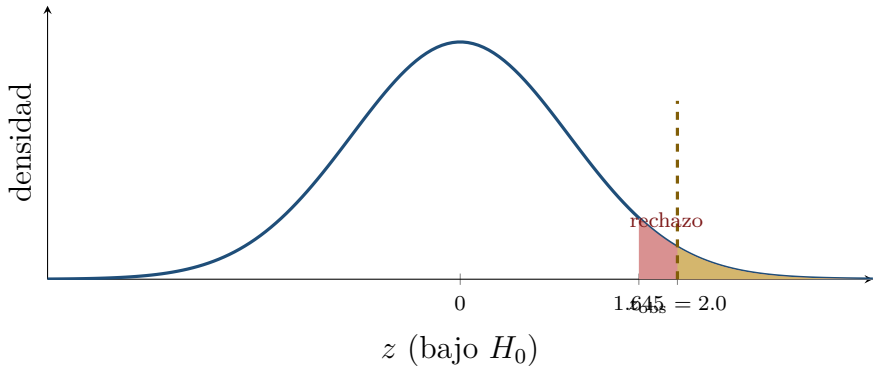


Figura 11: Demo B (una cola derecha). Región de rechazo $z > 1.645$ (rojo); el p-valor es el área a la derecha de $z_{\text{obs}} = 2.0$ (ámbar) ≈ 0.023.

Ejemplo 6.3 (Demo C — t-test (σ desconocida, n chico))**. Enunciado.** Probamos si el rendimiento medio de un cultivo *difiere* de $\mu_0 = 30$ qq/ha. Tomamos $n = 10$ parcelas y obtenemos $\bar{X} = 32.4$ y desviación muestral $s = 3.5$ (σ *desconocida*). ($\alpha = 0.05$.)

Paso 1. $H_0 : \mu = 30$ vs $H_1 : \mu \neq 30$ (**dos colas**).
Paso 2. Media con σ *desconocida* $\rightarrow$ **t-test** con gl $= n - 1 = 9$. Crítico $t_{9,\,0.975} \approx 2.262$.
Paso 3. $t = \dfrac{32.4 - 30}{3.5/\sqrt{10}} = \dfrac{2.4}{1.1068} \approx 2.17$.
Paso 4–5. $|t| = 2.17 < 2.262 \Rightarrow$ **NO rechazamos** H_0 (¡por poco!). El p-valor de dos colas es $\approx 0.058 > 0.05$.
Paso 6. No hay evidencia *suficiente* (al 5 %) de que el rendimiento difiera de 30. Ojo: con $n = 10$ la potencia es baja; un n mayor podría detectar el efecto. Nota cómo el crítico de la t (2.262) es *mayor* que el de la normal (1.96): la t "castiga" no conocer σ.

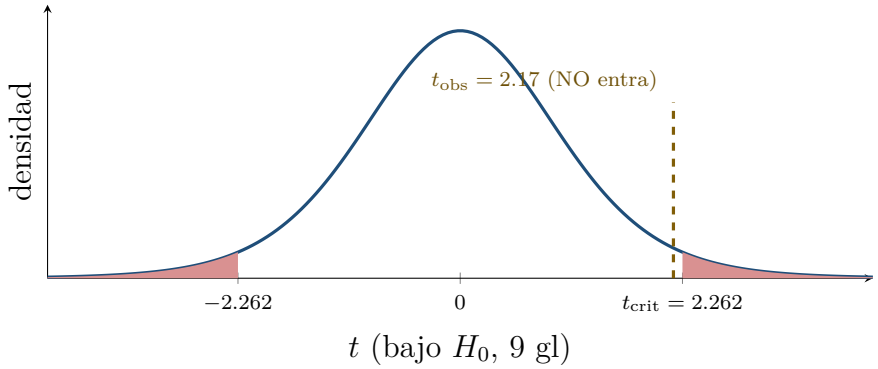


Figura 12: Demo C (t con 9 gl, dos colas). El estadístico $t_{\text{obs}} = 2.17$ se queda *justo a la izquierda* del crítico 2.262: no rechazamos. Las colas de la t son más gruesas que las de la normal.

Ejemplo 6.4 (Demo D — prueba de PROPORCIÓN (z))**. Enunciado.** Históricamente, el $p_0 = 60\,\%$ de los usuarios renueva su suscripción. Tras una mejora, en $n = 200$ usuarios renuevan 134 ($\hat{p} = 134/200 = 0.67$). ¿La proporción *supera* 0.60? ($\alpha = 0.05$.)

Paso 1. $H_0 : p = 0.60$ vs $H_1 : p > 0.60$ (**una cola derecha**).
Paso 2. Proporción → z**-test**; SE bajo H_0: $\sqrt{p_0(1-p_0)/n} = \sqrt{0.6 \cdot 0.4/200} = \sqrt{0.0012} = 0.03464$. Crítico 1.645.
Paso 3. $z = \dfrac{0.67 - 0.60}{0.03464} = \dfrac{0.07}{0.03464} \approx 2.02$.
Paso 4–5. $z = 2.02 > 1.645 \Rightarrow$ **rechazamos** H_0. p-valor $= 1 - \Phi(2.02) \approx 0.0217 < 0.05$.
Paso 6. Hay evidencia de que la tasa de renovación *subió* por encima del 60 % tras la mejora.

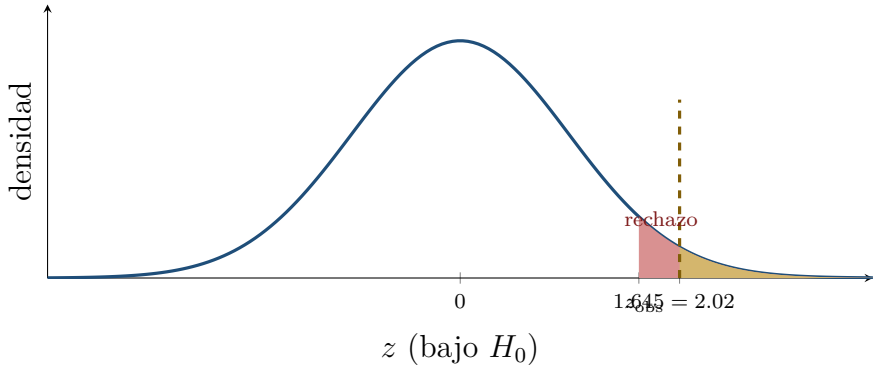


Figura 13: Demo D (proporción, una cola). $z_{\text{obs}} = 2.02$ cae en la región de rechazo ($z > 1.645$); el p-valor (ámbar) ≈ 0.022.

6.6. Duda 6: simular el nivel α como tasa de falsos positivos

El nivel α es la probabilidad de rechazar H_0 *cuando* H_0 *es cierta*. Eso se "ve" simulando: generamos muchas muestras *bajo* H_0 y contamos qué fracción rechaza. Debe rondar α.

Simulacion: alpha como tasa de falsos positivos a la larga

```
set.seed(2026)
mu0   <- 100     # H0 verdadera: la media poblacional ES 100
sigma <- 15
n     <- 30
alpha <- 0.05
B     <- 20000   # numero de experimentos (todos bajo H0)

# En cada experimento: muestrear BAJO H0 y hacer un t.test de H0: mu = 100.
# Contamos cuantas veces rechazamos (falso positivo, porque H0 es cierta).
rechaza <- replicate(B, {
  x <- rnorm(n, mean = mu0, sd = sigma)   # datos generados con H0 CIERTA
  t.test(x, mu = mu0)$p.value < alpha     # TRUE = rechazo = falso positivo
})

mean(rechaza)    # ~ 0.05  ==> la tasa de falsos positivos coincide con alpha

# Cambia alpha a 0.01 o 0.10 y la tasa simulada seguira a alpha.
# Un unico t.test (para ver un p-valor concreto bajo H0):
x <- rnorm(n, mean = mu0, sd = sigma)
t.test(x, mu = mu0)            # casi siempre NO rechaza (H0 es cierta)
```

La salida `mean(rechaza)` cae cerca de 0.05: aunque H_0 *es* cierta, en torno al 5 % de los experimentos la rechaza *por puro azar*. **Eso es exactamente α:** la tasa de falsos positivos que aceptamos a la larga. Si bajas `alpha` a 0.01, la tasa simulada bajará a ~ 0.01 (menos falsos positivos, pero a costa de menos potencia para detectar efectos reales).

6.7. Errores comunes al interpretar pruebas de hipótesis

Las pruebas de hipótesis son fáciles de *ejecutar* y fáciles de *malinterpretar*. Estos cuatro malentendidos son los más costosos en la práctica.

1) El p-valor NO es la probabilidad de que H_0 sea verdadera.
El p-valor es $\mathbb{P}(\text{datos tan o más extremos} \mid H_0 \text{ cierta})$, no $\mathbb{P}(H_0 \text{ cierta} \mid \text{datos})$. ¡Son cosas distintas! El p-valor *supone* que H_0 es cierta y mide qué tan raros serían los datos bajo ese supuesto. No te dice la probabilidad de la hipótesis (eso requeriría el enfoque bayesiano y una probabilidad *a priori*). Confundir $\mathbb{P}(A \mid B)$ con $\mathbb{P}(B \mid A)$ es la *falacia del fiscal* aplicada a la estadística.

2) "No rechazar H_0" $\neq$ "aceptar / probar H_0".
Ausencia de evidencia no es evidencia de ausencia. Si no encontraste un efecto, puede ser que no exista *o* que tu muestra fue demasiado pequeña para detectarlo. Reportar "no hay evidencia de efecto" es correcto; "probamos que no hay efecto" es casi siempre falso.

3) Significancia estadística ≠ importancia práctica.
Con un n enorme, hasta una diferencia minúscula y sin relevancia (p. ej. subir el ingreso en 0.50 soles) puede salir "significativa" ($p < 0.05$). Al revés, un efecto grande puede no ser significativo si n es chico. *Siempre* reporta el **tamaño del efecto** y su **intervalo de confianza**, no solo el p-valor.

4) Mirar muchas hipótesis / "p-hacking" infla los falsos positivos.
Si pruebas 20 hipótesis al $\alpha = 0.05$ *aunque todas sean falsas (no haya efecto)*, esperas $20 \times 0.05 = 1$ "descubrimiento" significativo *por puro azar*. Probar muchísimas variables y reportar solo las que "salieron", o seguir recolectando datos hasta que p baje de 0.05, fabrica falsos positivos. Remedios: preinscribir las hipótesis y corregir por comparaciones múltiples (p. ej. Bonferroni: usar α/m con m tests).

Si tiras una moneda justa 5 veces, sacar "5 caras" es raro (probabilidad $1/32 \approx 0.03$). Pero si *mil personas* lo intentan, que *alguien* saque 5 caras es casi seguro —y esa persona no tiene una moneda mágica. Buscar entre muchas hipótesis hasta hallar una "significativa" es exactamente eso: confundir el afortunado de turno con un descubrimiento real. Por eso importa decidir *qué* vamos a probar *antes* de mirar los datos.

7. Errores, potencia y el rol de n

Toda decisión puede equivocarse de dos maneras.

Definición 7.1 (Errores tipo I y II; potencia)**.**

	H_0 cierta	H_0 falsa
No rechazar H_0	correcto	Error tipo II (β)
Rechazar H_0	Error tipo I (α)	correcto (potencia)

El **error tipo I** (falso positivo) tiene probabilidad α (el nivel de significancia). El **error tipo II** (falso negativo) tiene probabilidad β. La **potencia (power)** es

$$\boxed{\text{potencia} = 1 - \beta = \mathbb{P}(\text{rechazar } H_0 \mid H_0 \text{ falsa})}$$

la probabilidad de detectar un efecto que *sí* existe.

α es la tasa de "falsas alarmas" (gritar "¡hay efecto!" cuando no lo hay); β es la de "efectos perdidos" (no detectar uno real). Hay un *trade-off*: bajar α (ser más exigente para rechazar) sube β y baja la potencia. Elegimos α pequeño cuando el falso positivo es el error más grave (p. ej. aprobar un tratamiento que no funciona).

Cuatro casos para fijar la idea

H_0 es siempre el "no pasa nada / statu quo". Mira qué significa cada error en contextos reales:

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

¿Cuál es peor? Depende del costo de cada error. En justicia, condenar a un inocente (tipo I) se considera lo más grave: por eso se asume "inocente hasta que se pruebe lo contrario" (eso fija H_0 = inocente) y se exige evidencia fuerte (α pequeño). En salud pública, a veces el tipo II (no detectar) es peor y se acepta un α mayor.

Cómo subir la potencia

Para un α fijo, la potencia $1-\beta$ aumenta cuando:

- el **efecto verdadero** es mayor (más distancia entre H_0 y la realidad);
- la **varianza** σ^2 es menor (datos menos ruidosos);
- el **tamaño de muestra** n es mayor.

Al crecer n, las distribuciones de T bajo H_0 y bajo H_1 se vuelven más angostas y se separan: *se puede mejorar α y β a la vez*. Por eso el **cálculo de potencia (power calculation)** —elegir n para garantizar cierta potencia— es clave al diseñar experimentos sociales (ensayos aleatorizados).

Ejemplo 7.1 (Calcular α y β con números)**.** Sea $T = \bar{X}$ con X_i iid $\mathcal{N}(\mu, \sigma^2)$, $\sigma^2 = 4$, $n = 25$, así que $\text{SE} = \sigma/\sqrt{n} = 2/5 = 0.4$. Probamos $H_0 : \mu = 0$ vs $H_1 : \mu = 1$, rechazando si $\bar{X} > k$. Bajo H_0, $\bar{X} \sim \mathcal{N}(0, 0.4^2)$; bajo H_1, $\bar{X} \sim \mathcal{N}(1, 0.4^2)$. Para $\alpha = 0.05$ el crítico es $k = 1.645 \cdot 0.4 = 0.658$. Entonces

$$\alpha = \mathbb{P}(\bar{X} > 0.658 \mid \mu = 0) = 1 - \Phi\Big(\tfrac{0.658-0}{0.4}\Big) = 1 - \Phi(1.645) = 0.05,$$

$$\beta = \mathbb{P}(\bar{X} < 0.658 \mid \mu = 1) = \Phi\Big(\tfrac{0.658-1}{0.4}\Big) = \Phi(-0.855) \approx 0.196.$$

La potencia es $1 - \beta \approx 0.804$. Si subiéramos n, SE caería, β bajaría y la potencia crecería.

Ejemplo 7.2 (Demo aplicada: ¿funciona un programa de tutorías?)**.** Un municipio prueba un programa de tutorías y mide el cambio promedio de puntaje en $n = 100$ estudiantes; se

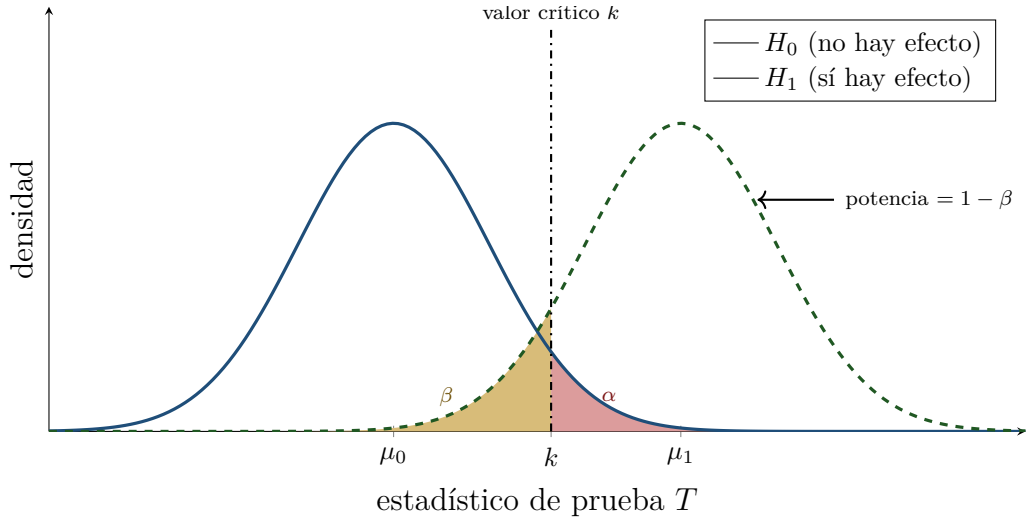


Figura 14: Densidades del estadístico bajo H_0 (azul, centrada en μ_0) y bajo H_1 (verde, centrada en μ_1). El valor crítico k separa rechazo (a la derecha) de no rechazo. El área roja a la derecha de k bajo H_0 es α (error tipo I); el área ámbar a la izquierda de k bajo H_1 es β (error tipo II). La potencia es el resto bajo H_1 a la derecha de k. Al aumentar n, ambas campanas se angostan y se separan: α y β bajan juntos.

sabe que $\sigma = 15$ puntos. Plantea $H_0 : \mu = 0$ (sin efecto) vs $H_1 : \mu > 0$ (mejora). Observa $\bar{X} = 3.2$ puntos.

- **¿Qué sería cada error aquí?** Tipo I: concluir que el programa funciona y *escalarlo* (gastar presupuesto) cuando en realidad no sirve. Tipo II: *archivar* el programa cuando sí ayudaba a los estudiantes.
- **Decisión:** $\text{SE} = \sigma/\sqrt{n} = 15/10 = 1.5$; $z = \dfrac{\bar{X} - 0}{\text{SE}} = \dfrac{3.2}{1.5} = 2.13$. Como $z = 2.13 > 1.645$ (crítico de una cola con $\alpha = 0.05$), **rechazamos** H_0: hay evidencia de un efecto positivo. El p-valor es $1 - \Phi(2.13) \approx 0.017 < 0.05$.

Importante: rechazar H_0 *no* prueba que el efecto sea grande ni que valga la pena; solo dice que es poco probable verlo si el programa no sirviera. Hay que mirar también el *tamaño* del efecto (3.2 puntos) y su intervalo de confianza.

Ejemplo 7.3 (Demo accesible: control de calidad en una fábrica de tornillos)**.** Una máquina debe producir tornillos de longitud media $\mu_0 = 20$ mm. Por desgaste, puede "descalibrarse" y empezar a producir tornillos más largos. Cada hora se toma una muestra de $n = 36$ tornillos (se sabe $\sigma = 0.6$ mm) y se prueba $H_0 : \mu = 20$ (máquina OK) vs $H_1 : \mu > 20$ (descalibrada). Si rechazamos, *se detiene la línea* para recalibrar.

¿Qué es cada error aquí?

- **Error tipo I (falsa alarma):** parar una máquina que *estaba bien*. Costo: producción detenida y técnico movilizado en vano.
- **Error tipo II (efecto perdido):** *no* parar una máquina que ya estaba descalibrada. Costo: un lote entero de tornillos defectuosos.

Números. El error estándar es $\text{SE} = \sigma/\sqrt{n} = 0.6/6 = 0.1$ mm. Fijamos $\alpha = 0.05$ (test de una cola), así que rechazamos si $\bar{X} > k$ con $k = \mu_0 + 1.645 \cdot \text{SE} = 20 + 1.645(0.1) = 20.16$ mm. Entonces

$$\alpha = \mathbb{P}(\bar{X} > 20.16 \mid \mu = 20) = 1 - \Phi\left(\tfrac{20.16-20}{0.1}\right) = 1 - \Phi(1.645) = 0.05.$$

Supongamos que, descalibrada, la máquina produce $\mu_1 = 20.30$ mm. La probabilidad de *no detectarlo* (tipo II) es

$$\beta = \mathbb{P}(\bar{X} < 20.16 \mid \mu = 20.30) = \Phi\left(\tfrac{20.16-20.30}{0.1}\right) = \Phi(-1.40) \approx 0.081.$$

La **potencia** de detectar esa descalibración es $1 - \beta \approx 0.92$: un buen sistema de control. Si la descalibración fuera más sutil ($\mu_1 = 20.10$, apenas por encima), la potencia caería mucho ($1 - \beta \approx 0.26$): los problemas *pequeños* son los más fáciles de "perder". La intuición "**falsa alarma vs. efecto perdido**" es exactamente el trade-off α vs. β que vimos en la Figura 14.

Dualidad CI ↔ test

Hay un puente directo: un test de dos colas al nivel α **rechaza** $H_0 : \mu = \mu_0$ *si y solo si* μ_0 **cae fuera** del CI de nivel $1 - \alpha$. Por eso, mirar si un intervalo de confianza contiene (o no) el valor nulo equivale a hacer la prueba.

8. En R: simular la cobertura de un intervalo de confianza

La interpretación frecuentista del CI se "ve" muy bien por simulación: generamos muchas muestras de una población con μ *conocido*, construimos un CI en cada una, y contamos qué porcentaje captura el verdadero μ. Debería rondar el 95 por ciento.

Simulacion de cobertura de un IC al 95% (t de Student)

```
set.seed(2026)
mu      <- 2000      # media verdadera de la poblacion (la "verdad")
sigma   <- 600       # desviacion verdadera
n       <- 100       # tamano de cada muestra
B       <- 10000     # numero de muestras (replicaciones)

# Para cada replicacion: sacar n datos, construir el IC al 95% (caso t),
# y verificar si el intervalo CUBRE mu.
cubre <- replicate(B, {
  x  <- rnorm(n, mean = mu, sd = sigma)
  ic <- t.test(x, conf.level = 0.95)$conf.int   # IC con s (sigma desconocida)
  (ic[1] <= mu) & (mu <= ic[2])                 # TRUE si cubre mu
})
```

```
mean(cubre)          # ~ 0.95  (proporcion de intervalos que cubren mu)

# Tambien se puede a mano (sin t.test):
cubre2 <- replicate(B, {
  x    <- rnorm(n, mean = mu, sd = sigma)
  xbar <- mean(x); s <- sd(x)
  tcrit <- qt(0.975, df = n - 1)                # cuantil t_{n-1, 0.975}
  margen <- tcrit * s / sqrt(n)
  (mu >= xbar - margen) & (mu <= xbar + margen)
})
mean(cubre2)         # ~ 0.95 tambien

# Una sola prueba de hipotesis sobre H0: mu = 2000
x <- rnorm(n, mean = mu, sd = sigma)
t.test(x, mu = 2000)$p.value   # p-valor: casi siempre > 0.05 (H0 es cierta)
```

La proporción `mean(cubre)` sale cerca de 0.95: de las 10 000 muestras, alrededor del 95 por ciento produce un intervalo que *sí* contiene al verdadero $\mu = 2000$, y cerca del 5 por ciento *falla* —exactamente como en la Figura 5. Prueba a cambiar `conf.level` a 0.90 o 0.99 y observa cómo cambia la cobertura (y el ancho de los intervalos).

Resumen de la clase

Lo esencial de la Clase 6

- **Evaluar estimadores:** sesgo $= \mathbb{E}[\hat{\theta}] - \theta$ (insesgado si $= 0$); varianza (entre insesgados, menor = más eficiente); **MSE = sesgo^2 + varianza**; consistencia ($\hat{\theta}_n \to \theta$ en probabilidad).
- **Compromiso sesgo–varianza:** a veces un estimador con poco sesgo y mucha menos varianza tiene *menor* MSE que uno insesgado y ruidoso.
- **Derivar estimadores:** método de los *momentos* (iguala momentos poblacionales y muestrales; p. ej. $\hat{\lambda} = \bar{X}$ en Poisson); *máxima verosimilitud* (maximiza $L(\theta) = \prod f(x_i \mid \theta)$; p. ej. $\hat{p} = \bar{X}$ en Bernoulli). El MLE es bueno asintóticamente pero puede ser sesgado en muestras finitas.
- **Intervalos de confianza:** $\bar{X} \pm z_{1-\alpha/2}\,\sigma/\sqrt{n}$ (σ conocida) o $\bar{X} \pm t_{n-1,\,1-\alpha/2}\,s/\sqrt{n}$ (σ desconocida); $z_{0.975} = 1.96$. La t es más ancha (colas gruesas). Interpretación: el 95 % se refiere al *procedimiento*, no a un intervalo dado.
- **Pruebas de hipótesis:** H_0 vs H_1; estadístico T; nivel α; región de rechazo (dos colas ± 1.96, una cola 1.645 al 5 %); **rechaza si p-valor** $< \alpha$.
- **Errores y potencia:** tipo I (α, falso positivo), tipo II (β, falso negativo), **potencia** $= 1 - \beta$. Más n, menos ruido o mayor efecto $\Rightarrow$ más potencia. Dualidad CI$\leftrightarrow$test.

Fuentes y atribución

Material de repaso **basado en MIT 14.310x – Data Analysis for Social Scientists** (Esther Duflo & Sara Fisher Ellison), MIT OpenCourseWare, licencia **CC BY-NC-SA 4.0** (uso no comercial, con atribución, compartir bajo la misma licencia), y en las fuentes de la bibliografía. Los enunciados se basan en el material del curso y en diversas fuentes abiertas; cuando un problema se adapta de una fuente específica, se cita en el enunciado.

Bibliografía.

- Duflo, E. & Ellison, S. F. *14.310x: Data Analysis for Social Scientists*. MIT OpenCourseWare (slides, lecture notes y problem sets). ocw.mit.edu.
- Larsen, R. J. & Marx, M. L. *An Introduction to Mathematical Statistics and Its Applications*, 6.ª ed., Pearson, 2018.
- DeGroot, M. H. & Schervish, M. J. *Probability and Statistics*, 4.ª ed., Pearson, 2012.
- Blitzstein, J. & Hwang, J. *Introduction to Probability* (Harvard Stat 110); W. Chen, *Probability Cheatsheet*.
- Diez, D., Çetinkaya-Rundel, M. & Barr, C. *OpenIntro Statistics*. openintro.org.
- Materiales abiertos de la comunidad del curso en GitHub: `gevorg-hunanyan/probability`; `hanmochen/Probability-Theory-2020`; `mynameisjanus/14310xDataAnalysis`.