Pruebas de Hipótesis: Guía Completa

Clase 7 — 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: lo que quedó pendiente
2. Anatomía del estadístico
3. El test z, paso a paso
4. El test t, la misma secuencia
5. Error Tipo I/II en dos capas
6. Potencia y tamaño de muestra
7. Pruebas de dos muestras
8. Prueba chi-cuadrado
9. Prueba F
10. Guía maestra de decisión
11. Cierre

Contenido

Un repaso profundo, autocontenido

¿Por qué esta clase?

En la Clase 6 quedaron dudas reales: ¿cómo se calcula el p-valor y qué significa? ¿por qué normalizamos con $\sigma/\sqrt{n}$? ¿una cola, dos colas, "media cola"? ¿el valor crítico y el p-valor son dos reglas distintas?

Esta clase re-recorre todo desde cero

z, t, proporciones, **dos muestras**, χ^2, F, potencia derivada, y una guía maestra de decisión. Redundante a propósito con la Clase 6 — para que quede *completamente* claro.

Contenido

Antes de empezar: ¿qué es "mucho" o "poco"?

Piensa

La altura promedio de los estudiantes es $\mu_0 = 1.65$ m. Mido una muestra y obtengo $\bar{X} = 1.68$ m. **Son las mismas unidades** (ambas alturas, en metros) — ¿esa diferencia de 3 cm es mucha o poca?

Antes de empezar: ¿qué es "mucho" o "poco"?

Piensa

La altura promedio de los estudiantes es $\mu_0 = 1.65$ m. Mido una muestra y obtengo $\bar{X} = 1.68$ m. **Son las mismas unidades** (ambas alturas, en metros) — ¿esa diferencia de 3 cm es mucha o poca?

Pista

Todavía no se puede saber. Falta un dato clave: ¿cuánto varía *normalmente* la media de una muestra, solo por azar? Sin eso, "3 cm" no significa nada por sí solo. Vamos a verlo.

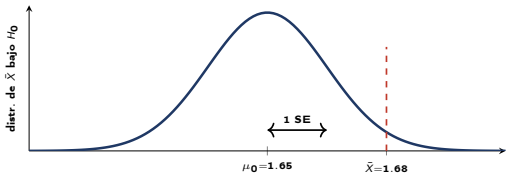
¿Por qué dividimos entre $\sigma/\sqrt{n}$?

No es un problema de unidades

$\bar{X}$ y μ_0 **ya están en las mismas unidades**. El problema real: la distancia cruda $\bar{X} - \mu_0$ no dice si es "mucha" o "poca" sin saber cuánto **varía** $\bar{X}$ **de muestra en muestra**, solo por azar.

La regla: el error estándar

Por el TLC, $\bar{X}$ tiene su propia dispersión: $\mathrm{SE} = \sigma/\sqrt{n}$ — cuánto se espera que "baile" la media muestral aunque H_0 sea cierta.



Dividir = comparar contra ese vaivén normal

$z = \dfrac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$: no mide metros ni ninguna otra unidad — mide **cuántos "SEs de sorpresa"** es tu resultado.

Demo: alturas

$\mu_0 = 1.65$ m, $\bar{X} = 1.68$ m, con $n = 100$ estudiantes $\Rightarrow \mathrm{SE} = 0.015$ m.

$$z = \frac{1.68 - 1.65}{0.015} = 2.0$$

Interpretación

Los 3 cm equivalen a **2 errores estándar** de distancia — bastante más que el vaivén normal esperado por puro azar: empieza a ser sospechoso para H_0.

σ vs. s

Si σ es desconocida, la estimamos con s — y esa estimación extra de incertidumbre es justo lo que hace que t tenga colas más pesadas.

Contenido

Antes de empezar: ¿el mismo dato, dos conclusiones?

Piensa

¿Puede el **mismo** resultado de una muestra llevarte a conclusiones **distintas** según cómo planteaste tu pregunta de investigación (¿cambió? ¿subió? ¿bajó?)?

Antes de empezar: ¿el mismo dato, dos conclusiones?

Piensa

¿Puede el **mismo** resultado de una muestra llevarte a conclusiones **distintas** según cómo planteaste tu pregunta de investigación (¿cambió? ¿subió? ¿bajó?)?

Pista

Sí — la dirección de H_1 cambia la regla de decisión, aunque el dato sea idéntico. Vamos a verlo con números.

¿Qué es una prueba de hipótesis?

La pregunta de fondo

Dada una muestra, ¿hay **suficiente evidencia** para dudar de una afirmación sobre la población?
Ejemplo: "el promedio es 500" ¿lo sostenemos, o los datos lo contradicen?

Vocabulario mínimo

H_0 = la afirmación que ponemos a prueba (el "statu quo"). H_1 = lo contrario, lo que sospechamos.
Rechazar H_0 = los datos son demasiado raros para sostenerla.

Todavía sin fórmulas

Antes de cualquier cálculo, la idea es simple: ¿mis datos serían muy raros *si* H_0 fuera cierta? Si sí, dudamos de H_0.

El test z: ¿qué es y cuándo se usa?

Definición

Pone a prueba una afirmación sobre la **media** μ, cuando **conocemos** σ (o n es grande). Estadístico (visto en Anatomía): $z = \dfrac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$ bajo H_0.

Con z_{obs} ya calculado, ¿qué hago?

Como $z \sim \mathcal{N}(0,1)$ bajo H_0, puedo convertir z_{obs} en una probabilidad con la CDF normal Φ:

$$\Phi(z_{\text{obs}}) = \mathbb{P}(Z \leq z_{\text{obs}} \mid H_0)$$

Esta $\Phi(z_{\text{obs}})$ es la pieza que usaremos más adelante tanto para la región de rechazo como para el p-valor.

¿Cuándo lo uso?

(1) tu pregunta es sobre una **media**; (2) **una sola muestra**; (3) σ conocida. Si σ es desconocida, viene el test t (más adelante).

Ejemplo paso a paso: crítico y p-valor juntos

Datos (retomamos la demo de alturas)

$\mu_0 = 1.65$ m, σ conocida, $n = 100$ estudiantes, $\bar{X} = 1.68$ m, $\mathrm{SE} = \sigma/\sqrt{n} = 0.015$ m. $H_0 : \mu = 1.65$ vs. $H_1 : \mu \neq 1.65$ (dos colas), $\alpha = 0.05$.

Paso 1: el estadístico

$$z_{\text{obs}} = \frac{\bar{X} - \mu_0}{\mathrm{SE}} = \frac{1.68 - 1.65}{0.015} = 2.0$$

Camino A: valor crítico

Dos colas, $\alpha = 0.05 \Rightarrow z_{\text{crít}} = \pm 1.96$ (tabla Normal).
Comparo:

$$|z_{\text{obs}}| = 2.0 > 1.96$$

$\Rightarrow$ **rechazo** H_0.

Camino B: p-valor

Busco $\Phi(2.0) = 0.9772$ en la tabla Normal. Dos colas:

$$\text{p-valor} = 2\big(1 - \Phi(2.0)\big) = 2(1 - 0.9772) = 0.0456$$

Comparo: $0.0456 < 0.05 = \alpha \Rightarrow$ **rechazo** H_0.

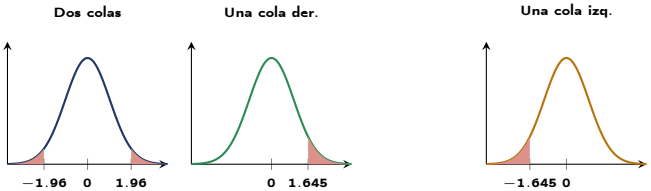
Misma decisión, dos caminos

Los 3 cm de diferencia sí son estadísticamente significativos: la altura promedio real **no** es 1.65 m.

La región de rechazo: ¿qué es y por qué esa zona?

Definición

Los valores de z tan extremos que, si aparecieran, dudaríamos de H_0. El **valor crítico** es la frontera; se lee de la tabla Normal según α y el número de colas.

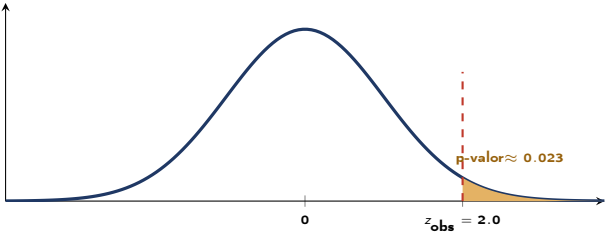


¿Por qué esa zona y no otra?

El área roja **siempre** suma α: así, la probabilidad de rechazar H_0 *siendo cierta* (falsa alarma) queda controlada en α por diseño.

El p-valor: ¿qué es, cómo se calcula, para qué sirve?

$$\text{p-valor} = \begin{cases} 2(1-\Phi(|z_{\text{obs}}|)) & \text{dos colas} \\ 1-\Phi(z_{\text{obs}}) & \text{una cola derecha} \\ \Phi(z_{\text{obs}}) & \text{una cola izquierda} \end{cases}$$



Para qué sirve

Responde: *"si H_0 fuera cierta, ¿qué tan raro sería mi dato (o peor)?"* Chico $\Rightarrow$ raro bajo H_0 $\Rightarrow$ dudamos de H_0.

¿El p-valor es $\mathbb{P}(H_0 \mid \text{datos})$? La falacia del fiscal

Qué es realmente $\Phi(z_{\text{obs}})$

$\Phi(z_{\text{obs}}) = \mathbb{P}(Z \le z_{\text{obs}} \mid H_0$ **cierta**): una probabilidad **sobre el estadístico**, *asumiendo* H_0. El p-valor (sea Φ, $1-\Phi$, o $2(1-\Phi)$) es $\mathbb{P}(\text{datos tan o más extremos} \mid H_0)$ — **siempre** condicionado en "H_0 cierta", nunca al revés.

¿Por qué **no** puede ser $\mathbb{P}(H_0 \mid \text{datos})$?

Esa probabilidad "al revés" exigiría el teorema de Bayes:

$$\mathbb{P}(H_0 \mid \text{datos}) = \frac{\mathbb{P}(\text{datos} \mid H_0)\,\mathbb{P}(H_0)}{\mathbb{P}(\text{datos})}$$

Necesita el **prior** $\mathbb{P}(H_0)$ (qué tan probable era H_0 *antes* de ver los datos) — algo que el cálculo del p-valor **jamás** usa ni conoce.

Conclusión

$\Phi(z_{\text{obs}})$, $1-\Phi(z_{\text{obs}})$ y el p-valor son afirmaciones sobre los **datos dado** H_0. $\mathbb{P}(H_0 \mid \text{datos})$ es una pregunta distinta, que el p-valor **no responde**.

¿Son lo mismo? Valor crítico vs. p-valor

Camino A

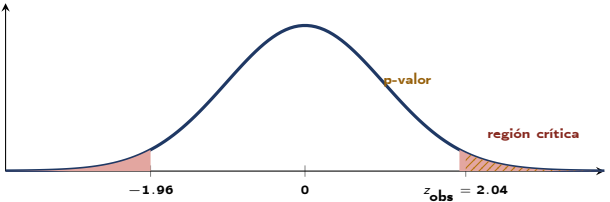
Calcula z_{obs}, compara contra un crítico fijo (p. ej. 1.96). Rechaza si $|z_{\text{obs}}| > z_{\text{crít}}$.

Camino B

Calcula la probabilidad de algo tan extremo bajo H_0 (p-valor). Rechaza si p-valor $< \alpha$.

Son equivalentes

$|z_{\text{obs}}| > z_{\text{crít}} \iff$ p-valor $< \alpha$ (monotonía de la CDF). No hay "dos versiones de 0.05": un umbral, dos ángulos. Con cualquiera de los dos basta; el software casi siempre da el p-valor.



¿Una cola o dos? Guía de decisión

Mira el signo de H_1

¿"cambió", "es distinto"? $\rightarrow \neq$ (**dos colas**). ¿"aumentó", "superó"? $\rightarrow >$ (**una cola derecha**). ¿"bajó", "se redujo"? $\rightarrow <$ (**una cola izquierda**).

Ninguna otra opción

No existe una "media cola". La dirección se fija **antes** de ver los datos — elegirla después (para que "convenga") es p-hacking.

Receta universal, 6 pasos

(1) H_0/H_1 (2) fija α (3) calcula z (4) valor crítico o p-valor (5) decide (6) interpreta en palabras.

Ejemplo 1 — dos colas

Call center

$\mu_0 = 8$ min, $\sigma = 2.4$, $n = 49$, $\bar{X} = 8.7$. $H_0 : \mu = 8$ vs. $H_1 : \mu \neq 8$.

$$z_{\text{obs}} = \frac{8.7 - 8}{2.4/\sqrt{49}} \approx 2.04$$

Crítico

Dos colas, $\alpha = 0.05 \Rightarrow z_{\text{crít}} = \pm 1.96$: $|2.04| > 1.96 \Rightarrow$ **rechazamos** H_0.

P-valor, paso a paso

Busco $\Phi(2.04) = 0.9793$ en la tabla Normal. Dos colas:

$$\text{p-valor} = 2\big(1 - \Phi(2.04)\big) = 2(1 - 0.9793) = 2(0.0207) \approx 0.041$$

Comparo: $0.041 < 0.05 = \alpha \Rightarrow$ **rechazamos** H_0 (misma decisión que con el crítico).

Interpretación

El tiempo medio de atención sí cambió.

Ejemplo 2 — una cola derecha

Barra de chocolate

$\mu_0 = 50$ g, $\sigma = 10$ g, $n = 25$, $\bar{X} = 54$ g. $H_0 : \mu = 50$ vs. $H_1 : \mu > 50$.

$$z = \frac{54 - 50}{10/\sqrt{25}} = 2.0$$

Decisión e interpretación

Crítico 1.645: $2.0 > 1.645 \Rightarrow$ **rechazamos** H_0 (p-valor ≈ 0.023). En promedio, las barras pesan más de 50 g.

Ejemplo 3 — una cola izquierda

Fábrica de tornillos

$\mu_0 = 30$ defectos, $\sigma = 6$, $n = 36$, $\bar{X} = 27.5$. $H_0 : \mu = 30$ vs. $H_1 : \mu < 30$.

$$z = \frac{27.5 - 30}{6/\sqrt{36}} = -2.5$$

Decisión e interpretación

Crítico -1.645: $-2.5 < -1.645 \Rightarrow$ **rechazamos** H_0 (p-valor ≈ 0.0062). El cambio de máquina sí redujo los defectos.

Contenido

Antes de empezar: ¿confiar más o menos?

Piensa

Si no conozco σ y tengo que **estimarla** con mis propios datos (s), ¿debería exigir un resultado **más** o **menos** extremo para rechazar H_0, comparado con si conociera σ exactamente?

Antes de empezar: ¿confiar más o menos?

Piensa

Si no conozco σ y tengo que **estimarla** con mis propios datos (s), ¿debería exigir un resultado **más** o **menos** extremo para rechazar H_0, comparado con si conociera σ exactamente?

Pista

Más extremo — estimar σ agrega incertidumbre, así que el "listón" sube un poco. Vamos a ver cuánto, y por qué.

El test t: ¿qué es y cuándo se usa?

Definición

Responde la misma pregunta que z (¿la media es μ_0?) pero cuando **no conocemos** σ. Se estima con s:

$$t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} \sim t_{n-1}.$$

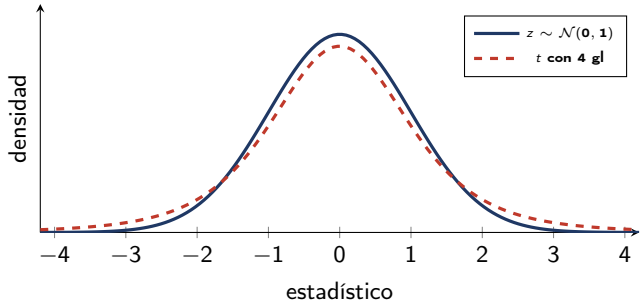
¿Cuándo lo uso?

Es el caso "por defecto" para una media: casi nunca conocemos σ *poblacional*, solo podemos estimarla de los datos.

¿Por qué cambia la curva? Colas más pesadas

De dónde sale la diferencia

Estimar σ con s agrega incertidumbre extra $\Rightarrow$ colas más gruesas que la Normal. Cuando n crece, $t_{n-1} \to \mathcal{N}(0,1)$.



El valor crítico t: depende de los grados de libertad

gl$= n-1$	5	10	15	30
$t_{\text{gl},0.975}$	2.571	2.228	2.131	2.042

A más gl, más cerca de z

Con menos datos, crítico más grande (colas más gruesas); a medida que gl crece, se acerca a 1.96.

¿Son lo mismo aquí también?

Sí: la equivalencia valor-crítico $\Longleftrightarrow$ p-valor vista para z aplica igual con t — solo cambia la CDF de referencia (t_{n-1} en vez de Φ), pero la CDF sigue siendo monótona creciente.

¿El p-valor de t es $\mathbb{P}(H_0 \mid \text{datos})$? Tampoco

Mismo argumento que con z

El p-valor con t es $\mathbb{P}(\text{datos tan o más extremos} \mid H_0)$, usando la CDF de t_{n-1} en vez de Φ — sigue condicionado en "H_0 cierta". Cambiar z por t **no** cambia la dirección de la condición.

El argumento de Bayes no cambia

$\mathbb{P}(H_0 \mid \text{datos}) = \dfrac{\mathbb{P}(\text{datos} \mid H_0)\,\mathbb{P}(H_0)}{\mathbb{P}(\text{datos})}$ sigue exigiendo el prior $\mathbb{P}(H_0)$ — el test t tampoco lo usa ni lo conoce, sin importar cuántos grados de libertad tenga.

Conclusión

Ni con z ni con t el p-valor es $\mathbb{P}(H_0 \mid \text{datos})$. Es $\mathbb{P}(\text{datos} \mid H_0)$ — siempre esa dirección.

Ejemplo 1 — dos colas, n chico

Peso al nacer

$\mu_0 = 3.20$ kg, $n = 16$, $\bar{X} = 3.32$, $s = 0.28$. $H_0 : \mu = 3.20$ vs. $H_1 : \mu \neq 3.20$, gl$= 15$.

$$t = \frac{3.32 - 3.20}{0.28/\sqrt{16}} \approx 1.71$$

Decisión e interpretación

Crítico $t_{15,0.975} \approx 2.131$: $1.71 < 2.131 \Rightarrow$ **NO rechazamos** (p-valor ≈ 0.107). No hay evidencia suficiente de cambio.

Ejemplo 2 — una cola

Taller mecánico

$\mu_0 = 5.0$ h, $n = 20$, $\bar{X} = 5.6$, $s = 1.2$. $H_0 : \mu = 5.0$ vs. $H_1 : \mu > 5.0$, gl$= 19$.

$$t = \frac{5.6 - 5.0}{1.2/\sqrt{20}} \approx 2.24$$

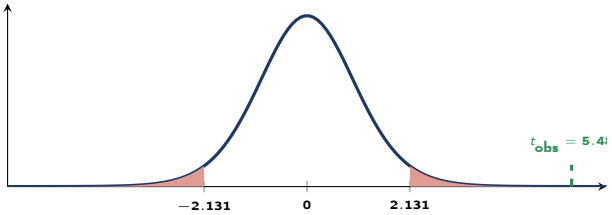
Decisión e interpretación

Crítico $t_{19,0.95} \approx 1.729$: $2.24 > 1.729 \Rightarrow$ **rechazamos** (p-valor ≈ 0.019). En promedio, se demora más de lo anunciado.

Ejemplo extendido, con R

Gasto mensual ($n = 16$ hogares)

$\mu_0 = 800$, $\bar{X} = 834.25$, $s = 25.01 \Rightarrow t \approx 5.48 \gg 2.131$: **rechazamos** con contundencia (p-valor ≈ 0.00006).



t (bajo H_0, 15 gl)

En R

`t.test(x, mu = 800)` → `t = 5.478, df = 15, p = 6.36e-05`

Test z de una proporción

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}}$$

Demo

$p_0 = 15\,\%$, $n = 250$, $\hat{p} = 0.208 \Rightarrow z \approx 2.57 > 1.645$: **rechazamos** (p-valor ≈ 0.005): mejoró la adherencia.

Contenido

Antes de empezar: ¿qué tipo de error es?

Piensa

Una prueba de embarazo da **positivo**, pero la persona **no** está embarazada. ¿Es error Tipo I o Tipo II?

Antes de empezar: ¿qué tipo de error es?

Piensa

Una prueba de embarazo da **positivo**, pero la persona **no** está embarazada. ¿Es error Tipo I o Tipo II?

Pista

Tipo I: H_0 =“no embarazada” era cierta, y la rechazamos incorrectamente (falsa alarma). Si el test diera **negativo** estando embarazada, sería **Tipo II**: no detectar algo real. Vamos a formalizarlo.

La idea, sin fórmulas: el detector de humo

La idea, sin fórmulas

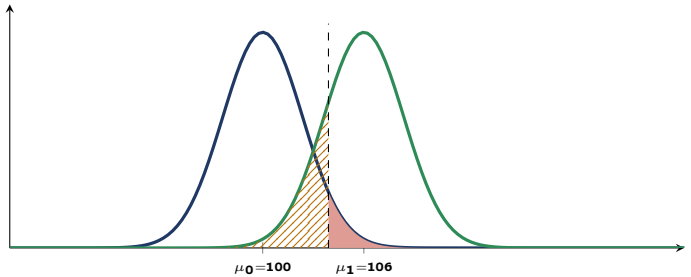
Tipo I: suena la alarma sin incendio (falsa alarma). **Tipo II**: hay incendio y no suena (el peor error).

	H_0 cierta	H_0 falsa
No rechazo	✓ Correcto	Tipo II (β)
Rechazo	Tipo I (α)	✓ Correcto

La versión formal: potencia

Formal

$\alpha = \mathbb{P}(\text{rechazar} \mid H_0)$; $\beta = \mathbb{P}(\text{no rechazar} \mid H_1)$; potencia $= 1 - \beta$.



Con $n = 40$, $\Delta = 6$, $\sigma = 15$: potencia ≈ 0.81.

Contenido

Antes de empezar: ¿sube o baja la potencia?

Piensa

Si duplico n, ¿qué le pasa a la potencia? ¿Y si el efecto Δ que quiero detectar es más chico?

Antes de empezar: ¿sube o baja la potencia?

Piensa

Si duplico n, ¿qué le pasa a la potencia? ¿Y si el efecto Δ que quiero detectar es más chico?

Pista

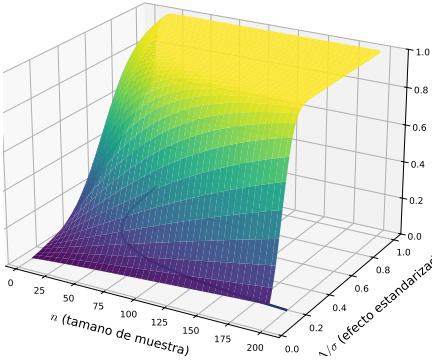
Más $n \Rightarrow$ SE más chico $\Rightarrow$ **potencia sube**. Δ más chico $\Rightarrow$ más difícil de distinguir de $H_0 \Rightarrow$ **potencia baja**. Vamos a derivarlo.

Derivación: ¿cuánto n necesito?

$$n = \frac{\sigma^2 (z_{1-\alpha} + z_{1-\beta})^2}{\Delta^2}$$

Demo: planeamiento de un RCT

$\Delta = 5$, $\sigma = 15$, potencia objetivo 0.80, $\alpha = 0.05$ (una cola): $n = \dfrac{225(1.645 + 0.842)^2}{25} \approx 55.6 \Rightarrow n = 56$.



Contenido

Antes de empezar: ¿pareado o independiente?

Piensa

Mido la presión arterial de 20 pacientes **antes y después** de un fármaco (mismos pacientes). ¿Pareado o independiente? ¿Por qué importaría la diferencia?

Antes de empezar: ¿pareado o independiente?

Piensa

Mido la presión arterial de 20 pacientes **antes y después** de un fármaco (mismos pacientes). ¿Pareado o independiente? ¿Por qué importaría la diferencia?

Pista

Pareado (mismos sujetos, dos mediciones). Importa porque pareado *resta* la variabilidad entre pacientes y detecta el efecto con muchísima más potencia que tratarlos como independientes. Vamos a verlo con un demo que sorprende.

¿Pareado o independiente? La pregunta que decide todo

Regla, ANTES de la fórmula

Pareado: mismos sujetos, medidos dos veces. **Independiente**: dos grupos distintos, no relacionados.

Los mismos 20 números, dos veredictos

Antes/después de un curso, mismos alumnos: pareado $t \approx 17.1$ (**rechaza con contundencia**); tratados (mal) como independientes: $t \approx 0.78$ (**no rechaza**). **El diseño decide la prueba.**

Independientes: pooled vs. Welch; pareado

$$t_{\text{pooled}} = \frac{\bar{X}_1 - \bar{X}_2}{s_p\sqrt{1/n_1 + 1/n_2}} \ (\text{gl} = n_1{+}n_2{-}2), \qquad t_{\text{pareado}} = \frac{\bar{d}}{s_d/\sqrt{n}} \ (\text{gl} = n{-}1)$$

Demo: salarios Lima vs. provincias

$n_1 = n_2 = 15$, $\bar{X}_1 = 2850$, $\bar{X}_2 = 2540$, varianzas similares $\rightarrow$ pooled: $t \approx 2.18 > 2.048$: **rechazamos** (p-valor ≈ 0.038).

Welch

Si las varianzas son muy distintas, usar pooled subestima el SE (falsa confianza). R: `t.test(var.equal=FALSE)`.

Dos proporciones

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_{\text{pool}}(1-\hat{p}_{\text{pool}})(1/n_1 + 1/n_2)}}$$

Demo: test A/B

A: 58/500, B: 82/500 $\Rightarrow z \approx 2.19 > 1.645$: **B convierte más** (p-valor ≈ 0.014).

Contenido

Antes de empezar: ¿dos versiones de χ^2?

Piensa

Vamos a ver dos preguntas distintas, ambas resueltas con χ^2: (a) ¿el dado es justo? (b) ¿género y preferencia de partido están asociados? ¿En qué se parecen y en qué difieren?

Antes de empezar: ¿dos versiones de χ^2?

Piensa

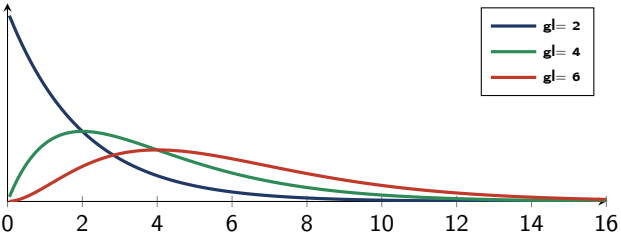
Vamos a ver dos preguntas distintas, ambas resueltas con χ^2: (a) ¿el dado es justo? (b) ¿género y preferencia de partido están asociados? ¿En qué se parecen y en qué difieren?

Pista

(a) es **bondad de ajuste**: 1 variable vs. una distribución esperada. (b) es **independencia**: 2 variables, ¿asociadas? Misma fórmula, pregunta distinta.

χ^2: bondad de ajuste e independencia

$$\chi^2 = \sum \frac{(O-E)^2}{E}, \qquad \text{gl} = k-1 \text{ (GOF) o } (r-1)(c-1) \text{ (indep.)}$$



Demo: educación $\times$ empleo formal (tabla 2×2)

$\chi^2 \approx 33.94$, gl$= 1$, crítico $\approx 3.84 \Rightarrow$ **sí** están asociados.

Contenido

Antes de empezar: ¿quién varía más?

Piensa

¿Cómo compararías si dos líneas de producción son igual de **consistentes** — no cuál produce más, sino cuál **varía** más?

Antes de empezar: ¿quién varía más?

Piensa

¿Cómo compararías si dos líneas de producción son igual de **consistentes** — no cuál produce más, sino cuál **varía** más?

Pista

Comparando sus **varianzas** — eso es exactamente el F-test.

F: comparar dos varianzas (¡lo prometido en Clase 6!)

$$F = \frac{s_1^2}{s_2^2} \sim F(n_1 - 1, n_2 - 1) \text{ bajo } H_0 : \sigma_1^2 = \sigma_2^2$$

Demo: dos líneas de producción

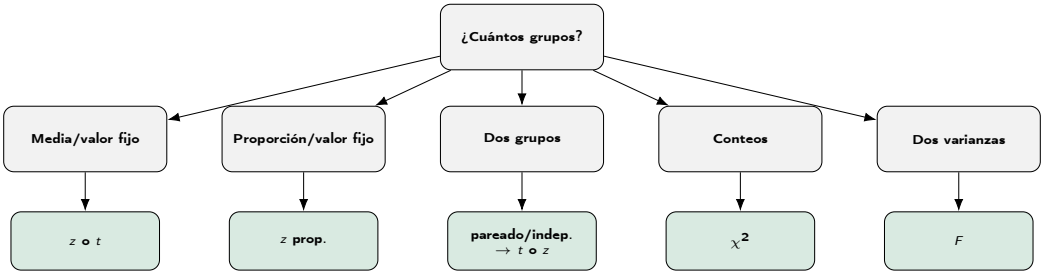
$n_1 = 5, s_1 = 4.8$; $n_2 = 21, s_2 = 2.6 \Rightarrow F \approx 3.41 > 2.87$: **rechazamos** (p-valor ≈ 0.028): línea 1 más variable.

Adelanto

F reaparece como la prueba de significancia **global** de una regresión (varianza explicada vs. residual).

Contenido

¿Qué test uso?



Colas: último paso, siempre separado

¿Signo de H_1? $\neq\rightarrow$ dos colas; $>/<\rightarrow$ una cola.

Catálogo rápido (extracto)

#	Escenario	Cola(s)	Test
1	¿Espera media *supera* 10 min? (σ conocida)	1 (der.)	z media
9	¿Gasto medio *difiere* de S/800? (n chico)	2	t media
16	¿Puntaje *mejoró*, mismos alumnos?	1 (der.)	t pareado
38	¿Preferencia observada coincide con la histórica?	—	χ^2 GOF

Tabla completa (48 escenarios en 8 familias) en la teoría, §14.

Contenido

Resumen

- Normalizar = distancia $\rightarrow$ número de errores estándar.
- Valor crítico y p-valor: **misma** regla, dos ángulos.
- Tipo I/II: detector de humo = tabla 2×2 formal.
- Potencia sube con n, Δ, o menos ruido σ.
- **Pareado vs. independiente** lo decide el diseño, no los números.
- χ^2 (categorías) y F (varianzas) completan la caja de herramientas.

¿Y ahora?

Con z, t, χ^2, F y dos muestras dominados, seguimos con **Causalidad y RCT** (Clase 8) — donde la comparación de medias entre grupos tratamiento y control es, precisamente, un test de dos muestras.