14.310x – Data Analysis for Social Scientists

MicroMasters en Statistics and Data Science (SDS) – MITx

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

Clase 7

Pruebas de Hipótesis: Guía Completa

Teoría

Módulos oficiales del MicroMaster:
M6 (extensión) – Confidence Intervals and Hypothesis Testing, ampliado con χ^2, F y pruebas de dos muestras

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 6 abrimos las pruebas de hipótesis: planteamos H_0/H_1, vimos el test z y el test t, una y dos colas, el p-valor, y una primera mirada al error Tipo I/Tipo II y la potencia. No alcanzó el tiempo para cerrar bien varias preguntas que **sí o sí** aparecen en la práctica: ¿cómo se calcula el p-valor paso a paso y qué significa exactamente? ¿por qué normalizamos con $\sigma/\sqrt{n}$ o $s/\sqrt{n}$? ¿cuándo uso una cola, dos colas, o cómo elijo la dirección? ¿la región de rechazo y el p-valor son "dos maneras distintas" de decidir, o la misma? ¿qué hago si tengo *dos* muestras en vez de una? ¿qué son las pruebas χ^2 y F?

Esta clase es un **repaso profundo y autocontenido**: recorremos *todo* el terreno de pruebas de hipótesis desde cero, con más detalle del que dio tiempo en la Clase 6, y agregamos tres familias de pruebas que aún no habíamos visto: pruebas de **dos muestras** (medias y proporciones, pareadas e independientes), la prueba χ^2 (bondad de ajuste e independencia) y la prueba F (comparación de varianzas). Cada concepto se explica en **dos capas**: primero la idea intuitiva, en lenguaje llano; después la versión formal con fórmulas. Al final hay una **guía maestra de decisión** y un **catálogo de referencia rápida** con decenas de escenarios "¿qué prueba uso aquí?".

Si ya viste la Clase 6, vas a reconocer varias ideas — eso es intencional. Aquí las vemos con más ejemplos, más despacio, y conectándolas con las pruebas nuevas (χ^2, F, dos muestras). Si te saltás la Clase 6, este documento alcanza por sí solo.

2. Repaso relámpago: H_0, H_1 y el vocabulario básico

La igualdad vive en H_0

La **hipótesis nula** H_0 es el "statu quo" (nada cambió, no hay diferencia) y **siempre** lleva el signo "=" (o $\leq/\geq$ en una cola). La **hipótesis alternativa** H_1 es lo que el investigador sospecha o quiere *detectar*, y nunca lleva "=′": usa $\neq$, $>$ o $<$. El signo de H_1 sale de la pregunta: ¿sube? $\rightarrow >$; ¿baja? $\rightarrow <$; ¿cambia/difiere? $\rightarrow \neq$.

Pregunta / sospecha	H_0	H_1	**Colas**
¿La tasa de defectos *bajó* tras el cambio de proveedor?	$p = p_0$	$p < p_0$	una (izq.)
¿El puntaje promedio de un colegio *superó* la meta regional?	$\mu = \mu_0$	$\mu > \mu_0$	una (der.)
¿La presión arterial media *cambió* con el nuevo fármaco?	$\mu = \mu_0$	$\mu \neq \mu_0$	dos
¿La participación electoral *difiere* entre dos distritos?	$p_1 = p_2$	$p_1 \neq p_2$	dos
¿La campaña de marketing *aumentó* la conversión?	$p_1 = p_2$	$p_2 > p_1$	una (der.)
¿El nivel educativo está *asociado* con el empleo formal?	independencia	asociación	— (única)
¿Dos líneas de producción tienen la *misma variabilidad*?	$\sigma_1^2 = \sigma_2^2$	$\sigma_1^2 \neq \sigma_2^2$	dos

3. Anatomía del estadístico de prueba: ¿por qué dividimos entre $\sigma/\sqrt{n}$?

Esta es la pregunta que más quedó en el aire: **¿por qué normalizamos, y por qué justo así?** Vamos a construirlo en tres pasos.

Construcción paso a paso

1. **La distancia cruda no es comparable.** Si $\bar{X} = 520$ y $\mu_0 = 500$, la distancia $\bar{X} - \mu_0 = 20$ no dice nada por sí sola: 20 puntos es enorme en un examen sobre 30, pero insignificante en uno sobre 1000. Necesitamos una *regla* para medir esa distancia.
2. **La regla es la dispersión de la media muestral.** Por el TLC (visto en la Clase 5), $\bar{X}$ tiene su propia distribución muestral con desviación estándar (el **error estándar**, *standard error*): $\text{SE} = \sigma/\sqrt{n}$ si σ es conocida (viene de $\text{Var}(\bar{X}) = \sigma^2/n$).
3. **Dividir entre el SE convierte "distancia cruda" en "número de errores estándar".** Ese cociente es *el mismo truco* del z-score que ya conoces para estandarizar una observación individual — solo que ahora lo aplicamos a la distribución muestral del estimador, no al dato individual:

$$z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}.$$

Casi nunca conocemos σ (la desviación estándar *poblacional*); en la práctica la estimamos con s (la desviación *muestral*). La anatomía es idéntica —distancia sobre error estándar— pero ahora la "regla" (s) *también* es una estimación, con su propio ruido. Eso agrega

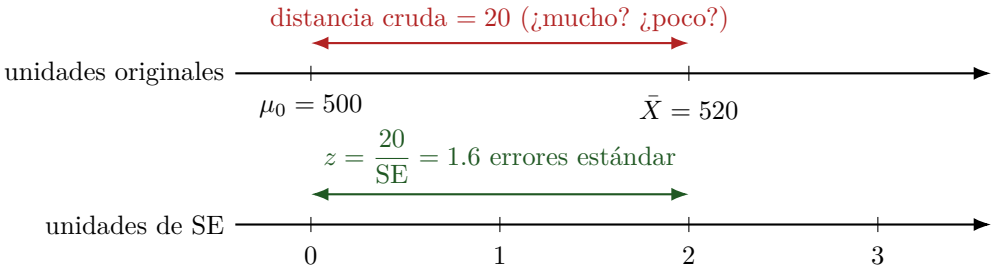

Figura 1: La "regla del metro": la misma distancia (arriba, en unidades originales) se re-expresa abajo en *unidades de error estándar*. Solo después de esta conversión la distancia es comparable entre problemas distintos.

incertidumbre extra, y por eso el estadístico ya no sigue una Normal exacta sino una t-Student, con colas más pesadas (§5).

4. El test z (media, σ conocida)

4.1. ¿Qué es el test z y cuándo se usa?

Definición 4.1 (Test z para una media)**.** El **test** z pone a prueba una afirmación sobre la **media poblacional** μ (p. ej. "el promedio es 500") usando una muestra, cuando la **desviación estándar poblacional** σ **es conocida** (o cuando n es grande, $n \gtrsim 30$, y usamos s como aproximación razonable de σ). El estadístico, ya construido en la Sección 3, es

$$z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}},$$

que bajo H_0 sigue una $\mathcal{N}(0,1)$ (Normal estándar).

Usálo cuando: (1) tu pregunta es sobre una **media** (no una proporción ni una varianza); (2) tienes **una sola muestra** (no comparas dos grupos); (3) conoces σ de antemano (poco común, pero pasa con instrumentos calibrados, procesos industriales muy estudiados, o cuando n es grande). Si σ es desconocida y n es chico, usa el test t (Sección 5) en su lugar.

4.2. El valor crítico: qué es y cómo se lee

Definición 4.2 (Valor crítico)**.** El **valor crítico** es el punto (o los puntos) que separan la "zona normal" de la **región de rechazo**: los valores de z tan extremos que, si aparecieran, dudaríamos de H_0. Se leen de la tabla de la Normal estándar (o de una calculadora/software) según el α elegido y el número de colas.

α	Crítico (dos colas) $z_{1-\alpha/2}$	Crítico (una cola) $z_{1-\alpha}$
0.10	1.645	1.282
0.05	1.960	1.645
0.01	2.576	2.326

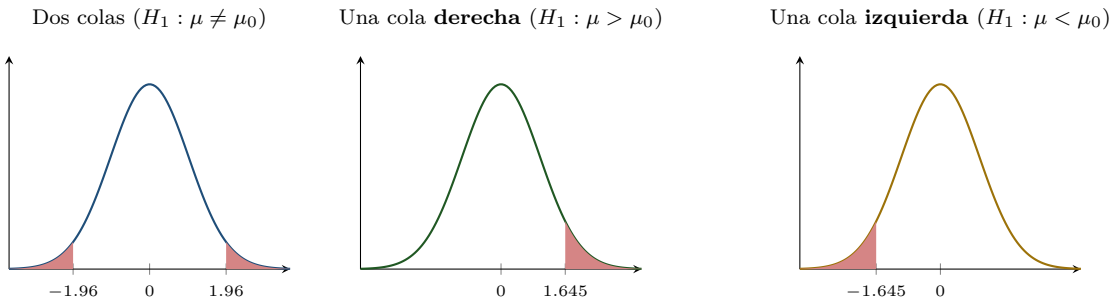

Figura 2: Las tres regiones de rechazo posibles (rojo) para $\alpha = 0.05$. El área roja **siempre** suma α: en dos colas se reparte $\alpha/2$ a cada extremo; en una cola, todo α va al lado que marca H_1.

4.3. El p-valor: qué es y cómo se calcula

Definición 4.3 (p-valor, otra vez, con fórmula explícita)**.** El p-valor es $\mathbb{P}(\text{tan o más extremo que } z_{\text{obs}} \mid H_0)$. Con Φ la CDF de la Normal estándar:

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

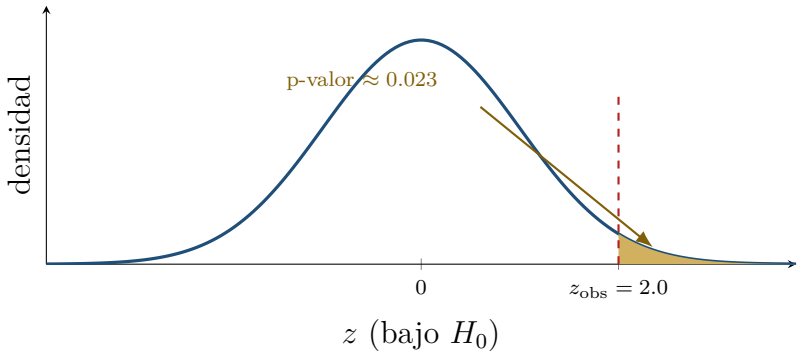

Figura 3: El p-valor (una cola derecha) es el área bajo la curva a la derecha de $z_{\text{obs}} = 2.0$ (el dato de la barra de chocolate, Demo 2). Como $0.023 < 0.05 = \alpha$, rechazamos H_0.

El p-valor responde: *"si H_0 fuera cierta, ¿qué tan raro sería obtener un resultado como el mío (o más extremo)?"* Un p-valor chico ($< \alpha$) dice "muy raro bajo H_0" $\Rightarrow$ dudamos de H_0. **No** es la probabilidad de que H_0 sea cierta —esa es una pregunta distinta que el p-valor no responde (ver la trampa #1 en §16).

Qué es realmente $\Phi(z_{\text{obs}})$. Φ es la CDF de la Normal estándar, así que $\Phi(z_{\text{obs}}) = \mathbb{P}(Z \leq z_{\text{obs}} \mid H_0 \text{ cierta})$: una probabilidad **sobre el estadístico** Z, *asumiendo* que H_0 es cierta. El p-valor (sea $\Phi(z_{\text{obs}})$, $1 - \Phi(z_{\text{obs}})$, o $2(1 - \Phi(|z_{\text{obs}}|))$ según la dirección) 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 $\mathbb{P}(H_0 \mid$ datos$)$.** Esa probabilidad "al revés" exigiría el teorema de Bayes:

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

Para calcularla hace falta 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. Por eso $\Phi(z_{\text{obs}})$ y el p-valor son afirmaciones sobre los *datos dado* H_0, mientras que $\mathbb{P}(H_0 \mid \text{datos})$ es una pregunta distinta que requiere información que el p-valor no tiene.

4.4. ¿Son lo mismo? Valor crítico y p-valor

Dos caminos, la misma decisión

	Camino A — valor crítico	**Camino B — p-valor**
Qué calculas	el estadístico z_{obs} (o t, etc.)	la probabilidad de un resultado tan o más extremo bajo H_0
Con qué comparas	un valor crítico fijo (p. ej. 1.96)	directamente contra α
Regla de decisión	rechaza si $\lvert z_{\text{obs}}\rvert > z_{\text{crít}}$	rechaza si p-valor $< \alpha$
Quién lo usa en la práctica	tablas impresas de estadística	software (R, Python) casi siempre
Ventaja	rápido si solo tienes tablas	cuantifica *qué tan* extremo, no solo sí/no

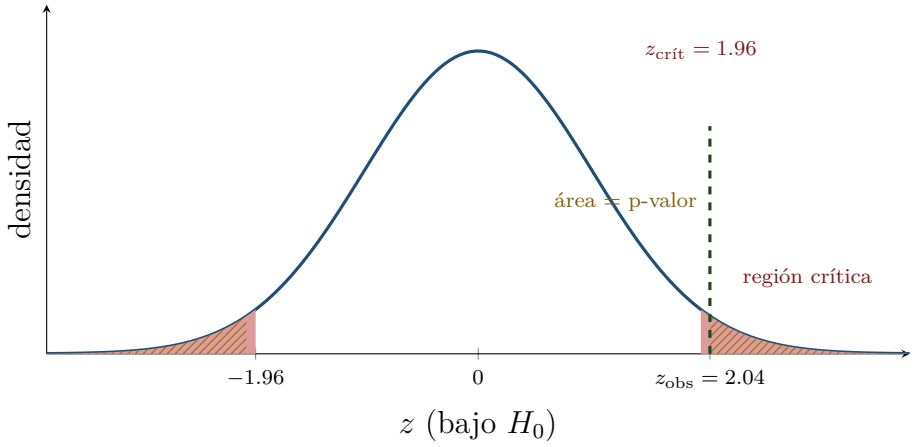

Figura 4: Dos caminos, un mismo resultado. Camino A: como $z_{\text{obs}} = 2.04$ cae *dentro* de la región crítica (rojo, $|z| > 1.96$), rechazamos. Camino B: el p-valor (rayado ambar, $\mathbb{P}(|Z| \geq 2.04) \approx 0.041$) es menor que $\alpha = 0.05$, también rechazamos. $|z_{\text{obs}}| > z_{\text{crít}} \iff$ p-valor $< \alpha$: es la **misma** regla vista desde dos ángulos.

Por qué son equivalentes

La función de distribución acumulada (CDF) es **monótona creciente**: a mayor $|z_{\text{obs}}|$, menor el área a su derecha (el p-valor). Por eso "el estadístico superó el crítico" y "el p-valor es menor que α" son la *misma* afirmación, dicha de dos maneras. No hay "dos versiones de $\alpha = 0.05$": hay un solo umbral, mirado desde el eje del estadístico (Camino A) o desde el eje de probabilidad (Camino B).

Con cualquiera de los dos caminos basta para decidir. En la práctica, casi siempre usarás el p-valor porque el software te lo da directo; entender la equivalencia con el valor crítico es lo que te permite *interpretar* lo que el software calculó.

4.5. ¿Una cola o dos? La guía de decisión

Cómo elegir

Mira el signo de H_1, que a su vez sale de la pregunta de investigación:

- ¿"*cambió*", "*es distinto*", "*difiere*"? $\rightarrow$ $H_1 : \mu \neq \mu_0$, **dos colas**.
- ¿"*aumentó*", "*superó*", "*es mayor*"? $\rightarrow$ $H_1 : \mu > \mu_0$, **una cola derecha**.
- ¿"*bajó*", "*es menor*", "*se redujo*"? $\rightarrow$ $H_1 : \mu < \mu_0$, **una cola izquierda**.

Sólo existen estas tres formas de H_1: $\neq$ (dos colas), $>$ (una cola derecha), $<$ (una cola izquierda). **No existe** una "media cola" ni una cuarta opción. La dirección se fija **antes** de ver los datos, a partir de la pregunta de investigación — elegirla *después* de ver los datos (para que "convenga" rechazar) es una forma de p-hacking (ver §16).

Receta universal, 6 pasos

1. Plantea H_0 y H_1.
2. Fija α (usualmente 0.05).
3. Calcula el estadístico: $z = \dfrac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$.
4. Obtén el valor crítico (o el p-valor).
5. Compara y decide: rechaza H_0 si z cae en la región de rechazo (o si p-valor $< \alpha$).
6. Interpreta en palabras, en el contexto del problema.

4.6. Tres ejemplos resueltos, uno por dirección

Ejemplo 4.1 (Demo 1 — dos colas)**.** Un call center afirma que el tiempo medio de atención es $\mu_0 = 8$ minutos, con $\sigma = 2.4$ (conocida, de un historial largo). En $n = 49$ llamadas, $\bar{X} = 8.7$. ¿Cambió el tiempo medio? ($\alpha = 0.05$.)

Paso 1–2. $H_0 : \mu = 8$ vs. $H_1 : \mu \neq 8$ (dos colas), $\alpha = 0.05$.

Paso 3. $z = \dfrac{8.7 - 8}{2.4/\sqrt{49}} = \dfrac{0.7}{0.3429} \approx 2.04$.

Paso 4. Crítico (dos colas): ± 1.96.

Paso 5. $|z| = 2.04 > 1.96 \Rightarrow$ **rechazamos** H_0. El p-valor es $2(1 - \Phi(2.04)) \approx 0.041 < 0.05$.

Paso 6. Hay evidencia de que el tiempo medio de atención cambió.

Ejemplo 4.2 (Demo 2 — una cola derecha)**.** Una barra de chocolate dice pesar en promedio $\mu_0 = 50$ g, con $\sigma = 10$ g (conocida). Un inspector sospecha que en realidad pesan *más* (¿nos están regalando chocolate?). Pesa $n = 25$ barras: $\bar{X} = 54$ g. ($\alpha = 0.05$.)

Paso 1–2. $H_0 : \mu = 50$ vs. $H_1 : \mu > 50$ (una cola derecha).

Paso 3. $z = \dfrac{54 - 50}{10/\sqrt{25}} = \dfrac{4}{2} = 2.0$.

Paso 4. Crítico (una cola, der.): 1.645.

Paso 5. $z = 2.0 > 1.645 \Rightarrow$ **rechazamos** H_0; p-valor $= 1 - \Phi(2.0) \approx 0.023 < 0.05$.

Paso 6. Sí, en promedio las barras pesan más de 50 g.

Ejemplo 4.3 (Demo 3 — una cola izquierda, reducción)**.** Una fábrica de tornillos tenía $\mu_0 = 30$ defectuosos por lote (de $\sigma = 6$, conocida). Tras un cambio de máquina, en $n = 36$ lotes se observa $\bar{X} = 27.5$. ¿Bajó el número medio de defectos? ($\alpha = 0.05$.)

Paso 1–2. $H_0 : \mu = 30$ vs. $H_1 : \mu < 30$ (una cola izquierda).

Paso 3. $z = \dfrac{27.5 - 30}{6/\sqrt{36}} = \dfrac{-2.5}{1} = -2.5.$
Paso 4. Crítico (una cola, izq.): -1.645.
Paso 5. $z = -2.5 < -1.645 \Rightarrow$ **rechazamos** H_0; p-valor $= \Phi(-2.5) \approx 0.0062 < 0.05$.
Paso 6. Sí, el cambio de máquina redujo los defectos.

5. El test t (media, σ desconocida)

5.1. ¿Qué es el test t y cuándo se usa?

Definición 5.1 (Test t para una media)**.** El **test** t responde exactamente la misma pregunta que el test z (¿la media poblacional es μ_0?) pero para el caso —mucho más común en la práctica— en que **no conocemos** σ. La estimamos con la desviación *muestral* s, y usamos

$$t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}},$$

que bajo H_0 sigue una distribución t de Student con $\text{gl} = n - 1$ grados de libertad (no una Normal exacta).

Es el caso "por defecto" para una media: úsalo cada vez que tengas una sola muestra, tu pregunta sea sobre una media, y **no** conozcas σ de antemano (lo normal: casi nunca conocemos la desviación estándar *poblacional*, solo podemos estimarla de los datos).

5.2. ¿Por qué la curva t es distinta? Colas más pesadas

De dónde sale la diferencia

Al estimar σ con s, agregamos una fuente extra de variabilidad (la propia incertidumbre de s como estimador). El estadístico t "sabe" esto y compensa con colas **más gruesas** (más probabilidad lejos de 0) que la Normal: así, para rechazar H_0 con la misma confianza, se exige un valor *un poco más extremo* que con z. Cuando n crece, s se vuelve una estimación muy precisa de σ y $t_{n-1} \to \mathcal{N}(0,1)$: con muestras grandes, t y z casi coinciden.

5.3. El valor crítico t: depende de los grados de libertad

Cómo leer una tabla t

A diferencia de z (un solo número por α), el crítico t depende **también** de $\text{gl} = n - 1$: a menos datos, crítico más grande (colas más gruesas). Ejemplos para $\alpha = 0.05$, dos colas:

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

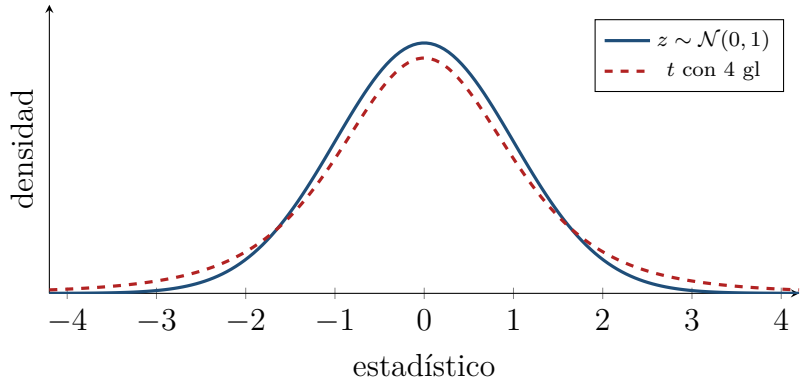

Figura 5: Con pocos grados de libertad (aquí, 4), la curva t (roja, punteada) tiene un pico más bajo y colas visiblemente más gruesas que la Normal (azul): valores "extremos" son más probables, así que el valor crítico debe ser mayor para mantener el mismo α.

Nota que a medida que gl crece, el crítico se acerca a 1.96 (el de z) — consistente con la Figura 5.

5.4. El p-valor con t

Se calcula con la **misma lógica** que con z, y de nuevo son tres casos según la dirección de H_1 — sólo que ahora usamos $F_t(\cdot\,; \text{gl})$, la CDF de la t de Student con $\text{gl} = n - 1$, en vez de Φ:

$$\text{p-valor} = \begin{cases} 2\,(1 - F_t(|t_{\text{obs}}|;\, \text{gl})) & \text{dos colas} \\ 1 - F_t(t_{\text{obs}};\, \text{gl}) & \text{una cola derecha} \\ F_t(t_{\text{obs}};\, \text{gl}) & \text{una cola izquierda} \end{cases}$$

Software como R o Python calcula F_t automáticamente conociendo gl (en R, la función es `pt(t_obs, df = gl)`).

Sí: la equivalencia valor-crítico $\Longleftrightarrow$ p-valor de la Sección 4.4 no depende de qué distribución de referencia se use — solo depende de que la CDF sea monótona creciente, lo cual es cierto tanto para $\mathcal{N}(0,1)$ como para t_{n-1}. Rechazar porque $|t_{\text{obs}}| > t_{\text{crít}}$ y rechazar porque p-valor $< \alpha$ siguen siendo la **misma** decisión, vista desde los mismos dos ángulos de siempre.

El mismo argumento de la página anterior aplica sin cambios: 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". El argumento de Bayes tampoco cambia: $\mathbb{P}(H_0 \mid \text{datos}) = \mathbb{P}(\text{datos} \mid H_0)\mathbb{P}(H_0)/\mathbb{P}(\text{datos})$ sigue exigiendo el prior $\mathbb{P}(H_0)$, que el test t tampoco usa ni conoce, sin importar cuántos grados de libertad tenga. 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.

5.5. Tres ejemplos resueltos

Ejemplo 5.1 (Demo 1 — dos colas, n chico)**.** Un programa de nutrición afirma que el peso medio al nacer no cambia respecto a $\mu_0 = 3.20$ kg. En $n = 16$ nacimientos, $\bar{X} = 3.32$ kg y $s = 0.28$ kg. ($\alpha = 0.05$.)

Paso 1–2. $H_0 : \mu = 3.20$ vs. $H_1 : \mu \neq 3.20$, gl$= n - 1 = 15$.

Paso 3. $t = \dfrac{3.32 - 3.20}{0.28/\sqrt{16}} = \dfrac{0.12}{0.07} \approx 1.71$.

Paso 4. Crítico $t_{15,0.975} \approx 2.131$.

Paso 5. $|t| = 1.71 < 2.131 \Rightarrow$ **NO rechazamos**; p-valor $\approx 0.107 > 0.05$.

Paso 6. No hay evidencia suficiente de cambio en el peso medio.

Ejemplo 5.2 (Demo 2 — una cola)**.** Un taller mecánico afirma reparar autos en $\mu_0 = 5.0$ horas en promedio. Una muestra de $n = 20$ reparaciones da $\bar{X} = 5.6$ h, $s = 1.2$ h. ¿Se demora *más* de lo que dice? ($\alpha = 0.05$.)

Paso 1–2. $H_0 : \mu = 5.0$ vs. $H_1 : \mu > 5.0$, gl$= 19$.

Paso 3. $t = \dfrac{5.6 - 5.0}{1.2/\sqrt{20}} \approx 2.24$.

Paso 4. Crítico $t_{19,0.95} \approx 1.729$.

Paso 5. $t = 2.24 > 1.729 \Rightarrow$ **rechazamos**; p-valor $\approx 0.019 < 0.05$.

Paso 6. Sí, en promedio se demora más de lo anunciado.

Ejemplo 5.3 (Demo 3 — ejemplo extendido, paso a paso con datos)**.** El gasto mensual en alimentos de un distrito tenía media histórica $\mu_0 = 800$ soles. Se encuesta a $n = 16$ hogares:

$$812,\ 845,\ 798,\ 860,\ 830,\ 875,\ 805,\ 850,\ 822,\ 840,\ 865,\ 795,\ 835,\ 858,\ 810,\ 848.$$

Paso 1–2. $H_0 : \mu = 800$ vs. $H_1 : \mu \neq 800$, $\alpha = 0.05$, gl$= 15$.

Cálculo de $\bar{X}$ y s. $\bar{X} = \frac{1}{16}\sum x_i = 834.25$; $s = \sqrt{\frac{1}{15}\sum (x_i - \bar{X})^2} \approx 25.01$.

Paso 3. SE $= s/\sqrt{n} = 25.01/4 = 6.25$; $t = \dfrac{834.25 - 800}{6.25} \approx 5.48$.

Paso 4. Crítico $t_{15,0.975} \approx 2.131$.

Paso 5. $|t| = 5.48 \gg 2.131 \Rightarrow$ **rechazamos** H_0 con contundencia; p-valor ≈ 0.00006.

Paso 6. El gasto medio actual es claramente mayor al histórico.

Verificación en R

```
x <- c(812,845,798,860,830,875,805,850,822,840,865,795,835,858,810,848)
t.test(x, mu = 800)
# t = 5.478, df = 15, p-value = 6.36e-05
```

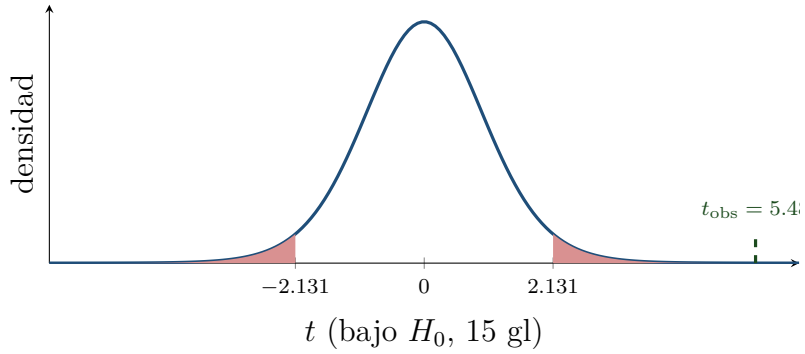

Figura 6: Región de rechazo (rojo) para el Demo 3, t_{15} con críticos ± 2.131. El estadístico observado $t_{\text{obs}} = 5.48$ cae muy adentro de la región de rechazo (la curva ya casi toca el eje ahí — por eso el p-valor es tan chico, ≈ 0.00006).

6. Test z de una proporción

6.1. ¿Qué es y cuándo se usa?

Definición 6.1 (Test z para una proporción)**.** Pone a prueba una afirmación sobre una **proporción** poblacional p (p. ej. "el 60 % aprueba"). Es la misma anatomía de siempre (distancia sobre error estándar), solo que ahora la "regla" (el SE) se calcula con p_0 bajo H_0, **no** con $\hat{p}$:

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

Bajo H_0, este estadístico sigue (aproximadamente, por el TLC aplicado a una proporción) una $\mathcal{N}(0,1)$ — exactamente como el z de una media.

Cuando tu pregunta es sobre una **proporción** (no una media), con **una sola muestra**. No hay "versión t" de este test: como el SE bajo H_0 ya usa el valor hipotético p_0 (no una desviación estimada aparte), no hay una fuente extra de incertidumbre que justifique colas más pesadas.

6.2. El valor crítico (las mismas reglas de siempre, aunque sea redundante repetirlo)

Como z sigue siendo aproximadamente $\mathcal{N}(0,1)$, el valor crítico se lee de la **misma** tabla Normal que en el test z de una media (Sección 3 en adelante):

α	Crítico (dos colas) $z_{1-\alpha/2}$	Crítico (una cola) $z_{1-\alpha}$
0.10	1.645	1.282
0.05	1.960	1.645
0.01	2.576	2.326

6.3. El p-valor: las mismas tres fórmulas, aplicadas a proporciones

De nuevo, exactamente las mismas tres fórmulas del test z de una media (solo que ahora z se calculó con $\hat{p}$ y p_0 en vez de $\bar{X}$ y μ_0):

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

Y la guía de una/dos colas es la misma de siempre: ¿"cambió"? dos colas; ¿"aumentó/superó"? una cola derecha; ¿"bajó"? una cola izquierda.

6.4. Dos ejemplos resueltos

Ejemplo 6.1 (Dos colas)**.** Históricamente, la mitad de los clientes elige el plan básico ($p_0 = 0.50$). En $n = 400$ clientes nuevos, 188 lo eligen ($\hat{p} = 0.47$). ¿Cambió la preferencia? ($\alpha = 0.05$.)

SE $= \sqrt{0.5 \cdot 0.5/400} = 0.025$; $z = \dfrac{0.47 - 0.50}{0.025} = -1.20$. Como $|z| = 1.20 < 1.96$, **no rechazamos**; p-valor ≈ 0.230.

Ejemplo 6.2 (Una cola)**.** Un programa de salud tenía $p_0 = 15\,\%$ de adherencia. Tras una campaña, de $n = 250$ pacientes, 52 se adhieren ($\hat{p} = 0.208$). ¿Mejoró la adherencia? ($\alpha = 0.05$.)

SE $= \sqrt{0.15 \cdot 0.85/250} = 0.02258$; $z = \dfrac{0.208 - 0.15}{0.02258} \approx 2.57$. Como $2.57 > 1.645$, **rechazamos** H_0; p-valor ≈ 0.005.

7. Tabla resumen: ejemplo, colas, estadístico, crítico y p-valor

Una tabla de consulta con los ejemplos ya vistos, columna por columna: el escenario con sus datos, cuántas colas (y de qué lado), la fórmula del estadístico, el valor crítico, y la fórmula **completa** del p-valor — **sin abreviar**, aunque se repita la misma forma varias veces, para que quede clarísimo caso por caso.

Escenario (datos)	**Cola(s)**	**Estadístico**	**Crítico** (α=0.05)	**Fórmula del p-valor**
Call center: $\mu_0 = 8$, $\sigma = 2.4$, $n = 49$, $\bar{X} = 8.7$. ¿Cambió?	dos colas	$z = \dfrac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} = 2.04$	± 1.96	$2\big(1 - \Phi(\lvert z_{\text{obs}}\rvert)\big) \approx 0.041$
Chocolate: $\mu_0 = 50$, $\sigma = 10$, $n = 25$, $\bar{X} = 54$. ¿Aumentó?	1 cola **derecha**	$z = \dfrac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} = 2.0$	1.645	$1 - \Phi(z_{\text{obs}}) \approx 0.023$
Tornillos: $\mu_0 = 30$, $\sigma = 6$, $n = 36$, $\bar{X} = 27.5$. ¿Bajó?	1 cola **izquierda**	$z = \dfrac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} = -2.5$	-1.645	$\Phi(z_{\text{obs}}) \approx 0.0062$
Peso al nacer: $\mu_0 = 3.20$, $n = 16$, $\bar{X} = 3.32$, $s = 0.28$. ¿Cambió?	dos colas, gl= 15	$t = \dfrac{\bar{X} - \mu_0}{s/\sqrt{n}} = 1.71$	± 2.131	$2\big(1 - F_t(\lvert t_{\text{obs}}\rvert; 15)\big) \approx 0.107$
Taller mecánico: $\mu_0 = 5.0$, $n = 20$, $\bar{X} = 5.6$, $s = 1.2$. ¿Se demora más?	1 cola **derecha**, gl= 19	$t = \dfrac{\bar{X} - \mu_0}{s/\sqrt{n}} = 2.24$	1.729	$1 - F_t(t_{\text{obs}}; 19) \approx 0.019$
Plan básico: $p_0 = 0.50$, $n = 400$, $\hat{p} = 0.47$. ¿Cambió?	dos colas	$z = \dfrac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} = -1.20$	± 1.96	$2\big(1 - \Phi(\lvert z_{\text{obs}}\rvert)\big) \approx 0.230$
Adherencia: $p_0 = 0.15$, $n = 250$, $\hat{p} = 0.208$. ¿Mejoró?	1 cola **derecha**	$z = \dfrac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} = 2.57$	1.645	$1 - \Phi(z_{\text{obs}}) \approx 0.005$

Cuadro 1: Cada fila es un ejemplo ya resuelto en este documento, con su fórmula **completa** de estadístico, crítico y p-valor (nada abreviado como "igual que arriba"). Φ es la CDF de $\mathcal{N}(0,1)$; $F_t(\cdot; \text{gl})$ es la CDF de la t de Student con esos grados de libertad. Nota: χ^2 y F no tienen "colas" en este mismo sentido — χ^2 es siempre una cola derecha por construcción (ver §8 en adelante), y F depende de la convención del software.

8. Error Tipo I y Tipo II — dos capas

8.1. La idea, sin fórmulas

Un detector de humo puede fallar de dos maneras. **Tipo I**: suena la alarma sin que haya incendio (falsa alarma, te despierta a las 3 a.m. por nada). **Tipo II**: hay un incendio real y la alarma *no* suena (el peor error). En pruebas de hipótesis pasa lo mismo: **Tipo I** es rechazar H_0 cuando en realidad era cierta ("falsa alarma" estadística); **Tipo II** es no rechazar H_0 cuando en realidad era falsa ("no detectamos el incendio real").

	H_0 **cierta (no hay fuego)**	H_0 **falsa (HAY fuego)**
No rechazo (no suena)	✓ Correcto	× **Tipo II** (β)
Rechazo (suena)	× **Tipo I** (α)	✓ Correcto

Figura 7: Los cuatro escenarios posibles, en lenguaje cotidiano (analogía del detector de humo): dos aciertos y dos tipos de error.

8.2. La versión formal

Definición 8.1 (Tabla 2×2 de errores)**.**

	H_0 cierta	H_0 falsa
No rechazo H_0	correcto $(1-\alpha)$	**Tipo II** (β)
Rechazo H_0	**Tipo I** (α)	correcto, **potencia** $(1-\beta)$

$\alpha = \mathbb{P}(\text{rechazar} \mid H_0 \text{ cierta})$; $\beta = \mathbb{P}(\text{no rechazar} \mid H_0 \text{ falsa})$; **potencia** $= 1-\beta = \mathbb{P}(\text{rechazar} \mid H_0 \text{ falsa})$: la probabilidad de detectar un efecto real cuando existe. Este marco (fijar α y minimizar β) se conoce como el enfoque de **Neyman–Pearson**.

Cuatro contextos, ahora en números

Contexto	Tipo I (α)	Tipo II (β)	Ejemplo num.
Tamizaje médico	falso positivo (sano marcado enfermo)	falso negativo (enfermo marcado sano)	test con $\alpha = 0.05$, $\beta = 0.20$
Control de calidad	rechazar un lote bueno	aceptar un lote defectuoso	$\alpha = 0.05$, potencia $= 0.81$ (Demo §9)
Programa social (RCT)	creer que hay efecto sin haberlo	no detectar un efecto real	n chico $\Rightarrow$ β alto
Sistema judicial	condenar a un inocente	absolver a un culpable	estándar penal: α muy chico

La Figura 7 y la tabla formal de arriba describen **exactamente lo mismo**: el "Tipo I" del detector de humo *es* α; el "Tipo II" *es* β. La única diferencia es que ahora podemos *calcularlos* con números (§9).

9. Potencia y tamaño de muestra — derivación completa

A diferencia de solo memorizar una fórmula, vamos a **derivar** de dónde sale el tamaño de muestra necesario para una potencia objetivo.

Planteamiento

Sea $H_0 : \mu = \mu_0$ vs. $H_1 : \mu = \mu_1 > \mu_0$ (una cola derecha), con $\Delta = \mu_1 - \mu_0 > 0$ el **tamaño del efecto**. Rechazamos si $\bar{X} > k$, donde el valor crítico k se fija para que $\mathbb{P}(\bar{X} > k \mid \mu_0) = \alpha$:

$$k = \mu_0 + z_{1-\alpha}\,\frac{\sigma}{\sqrt{n}}.$$

La **potencia** es la probabilidad de rechazar *si* μ_1 *es la verdad*:

$$1 - \beta = \mathbb{P}(\bar{X} > k \mid \mu_1) = \mathbb{P}\left(Z > \frac{k - \mu_1}{\sigma/\sqrt{n}}\right) = \Phi\left(\frac{\Delta\sqrt{n}}{\sigma} - z_{1-\alpha}\right).$$

Para una potencia *objetivo* $1 - \beta$, igualamos $\dfrac{\Delta\sqrt{n}}{\sigma} - z_{1-\alpha} = z_{1-\beta}$ y despejamos n:

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

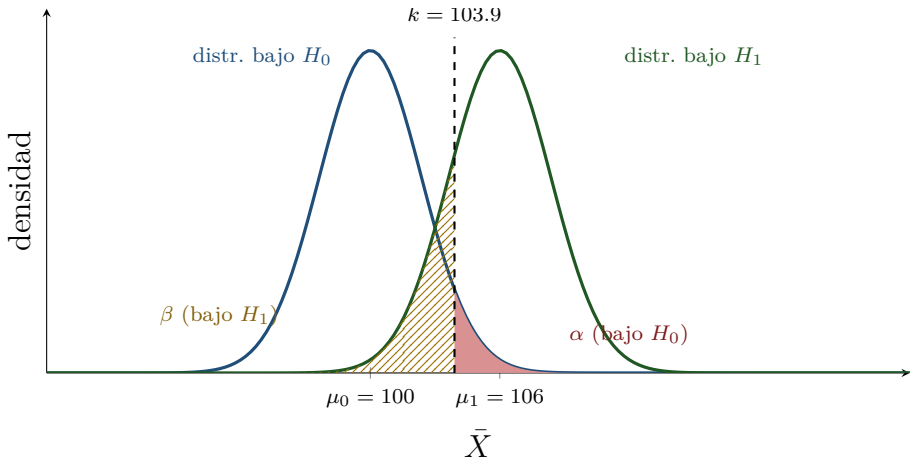

Figura 8: Dos mundos posibles: si H_0 es cierta (azul) o si H_1 es cierta (verde). α (rojo) es el área de falsa alarma bajo H_0; β (ambar rayado) es el área de no detección bajo H_1. La potencia es el área verde *restante* a la derecha de k.

Ejemplo 9.1 ((a) Potencia lograda, dado n)**.** Un programa de refuerzo escolar quiere subir el puntaje medio de $\mu_0 = 100$ a $\mu_1 = 106$ ($\Delta = 6$), con $\sigma = 15$ y $n = 40$ estudiantes, $\alpha = 0.05$ (una cola). $k = 100 + 1.645 \cdot \frac{15}{\sqrt{40}} \approx 103.90$. La potencia lograda es

$$1 - \beta = \Phi\left(\frac{6\sqrt{40}}{15} - 1.645\right) = \Phi(2.530 - 1.645) = \Phi(0.885) \approx 0.812.$$

Con $n = 40$, hay $\approx 81\,\%$ de probabilidad de detectar el efecto si de verdad existe.

Ejemplo 9.2 ((b) Tamaño de muestra para potencia objetivo)**.** Un RCT quiere detectar un efecto $\Delta = 5$ (sobre una variable con $\sigma = 15$) con potencia objetivo $1 - \beta = 0.80$ y $\alpha = 0.05$ (una cola). Con $z_{1-\alpha} = 1.645$ y $z_{1-\beta} = z_{0.80} = 0.842$:

$$n = \frac{15^2\,(1.645 + 0.842)^2}{5^2} = \frac{225 \cdot 6.183}{25} \approx 55.6 \;\Rightarrow\; n = 56 \text{ (redondeando hacia arriba).}$$

Tres palancas, todas visibles en la fórmula de n: (1) aumentar n (más datos, pero cuesta); (2) buscar un efecto Δ más grande (a veces no es decisión nuestra); (3) reducir σ (mejor diseño, menos ruido). La superficie de la Figura 9 muestra el mismo mensaje: para detectar efectos chicos hace falta mucho más n.

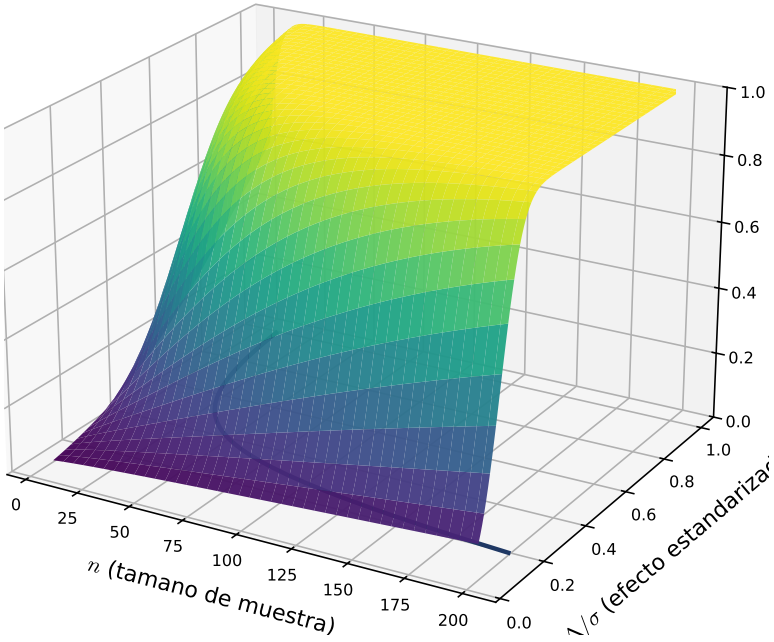

Figura 9: Potencia $1-\beta$ como superficie sobre n (tamaño de muestra) y $d=\Delta/\sigma$ (efecto estandarizado), test z de dos colas con $\alpha=0.05$. La curva marcada es el umbral de planeamiento habitual, potencia $=0.80$: para un efecto pequeño (d chico) se necesita un n mucho mayor para alcanzarla.

10. Pruebas de dos muestras — medias

¿Pareado o independiente? La pregunta que decide todo

Pareado (paired): los mismos sujetos (u objetos) se miden dos veces (antes/después, izquierda/derecha). **Independiente (independent):** dos grupos de sujetos *distintos y no relacionados*. Esta pregunta se responde **antes** de elegir fórmula, mirando el *diseño* del estudio — no los números.

10.1. Independientes, varianzas conocidas (caso puente)

$$z=\frac{(\bar{X}_1-\bar{X}_2)-0}{\sqrt{\sigma_1^2/n_1+\sigma_2^2/n_2}}.$$

Ejemplo 10.1. Dos sucursales: $\bar{X}_1=1450$, $\bar{X}_2=1390$ (ventas diarias, soles), $\sigma_1=180$, $\sigma_2=200$ (conocidas de datos históricos), $n_1=40$, $n_2=45$. SE $=\sqrt{180^2/40+200^2/45}\approx 41.22$; $z=\dfrac{1450-1390}{41.22}\approx 1.46$. Como $1.46<1.96$, **no rechazamos** (p-valor ≈ 0.145).

10.2. Independientes, varianzas desconocidas asumidas iguales (t pooled)

$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}, \qquad t = \frac{\bar{X}_1 - \bar{X}_2}{s_p\sqrt{1/n_1 + 1/n_2}}, \qquad \text{gl} = n_1 + n_2 - 2.$$

Ejemplo 10.2 (Salarios en dos regiones)**.** $n_1 = n_2 = 15$; $\bar{X}_1 = 2850$, $\bar{X}_2 = 2540$ soles; $s_1 = 380$, $s_2 = 400$. $s_p^2 = 152{,}200$, $s_p \approx 390.1$; $\text{SE} = 390.1\sqrt{2/15} \approx 142.5$; $t = \dfrac{2850-2540}{142.5} \approx 2.18$, gl= 28, crítico $t_{28,0.975} \approx 2.048$. Como $2.18 > 2.048$, **rechazamos** H_0 (p-valor ≈ 0.038): los salarios promedio difieren entre regiones.

10.3. Independientes, varianzas desconocidas y distintas (Welch)

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}},$$

con grados de libertad dados por la fórmula de Welch–Satterthwaite (no la derivamos aquí; así es como `R` calcula `t.test(var.equal=FALSE)` por defecto).

Ejemplo 10.3. $n_1 = 10$, $n_2 = 24$; $\bar{X}_1 = 48.5$, $\bar{X}_2 = 44.0$; $s_1 = 12.0$ (grupo chico y disperso), $s_2 = 4.5$ (grupo grande y estable) — varianzas muy distintas, usar pooled sería engañoso. SE ≈ 3.90; $t \approx 1.15$; gl (Welch)≈ 10.1; crítico ≈ 2.23. Como $1.15 < 2.23$, **no rechazamos** (p-valor ≈ 0.28).

10.4. Pareado: se reduce a un test t de una muestra sobre diferencias

$$d_i = X_i - Y_i, \qquad t = \frac{\bar{d}}{s_d/\sqrt{n}}, \qquad \text{gl} = n - 1.$$

Ejemplo 10.4 (Los mismos números, dos diseños distintos)**.** Un tutor mide el puntaje de 10 alumnos **antes y después** de un curso de refuerzo (mismos alumnos, medidos dos veces):

Antes	58	62	55	70	60	65	52	68	59	63
Después	64	66	61	73	62	70	55	74	63	68

Análisis correcto (pareado), porque son los MISMOS alumnos: diferencias d_i = después−antes; $\bar{d} = 4.4$, $s_d \approx 1.43$; $t = \dfrac{4.4}{1.43/\sqrt{10}} \approx 9.73$, gl= 9, crítico ≈ 2.262. $9.73 \gg 2.262 \Rightarrow$ **rechazamos con contundencia** (p-valor ≈ 0.0000045): el curso sí subió el puntaje.
Si (por error) tratamos los mismos 20 **números como dos grupos independientes** (ignorando quién es quién): $\bar{X}_{\text{antes}} = 61.2$, $\bar{X}_{\text{después}} = 65.6$, $s_1 \approx 5.59$, $s_2 \approx 5.83$; pooled $t \approx 1.72$, gl= 18, crítico ≈ 2.101. $1.72 < 2.101 \Rightarrow$ **no rechazaríamos** (p-valor ≈ 0.102).

Con exactamente los mismos 20 números, el análisis pareado detecta el efecto con claridad y el análisis (mal) independiente no. La razón: al parear, *restamos* la variabilidad entre alumnos (unos parten más alto que otros) y nos quedamos solo con el cambio individual — mucho menos ruido. **El diseño del estudio decide la prueba, no los números.**

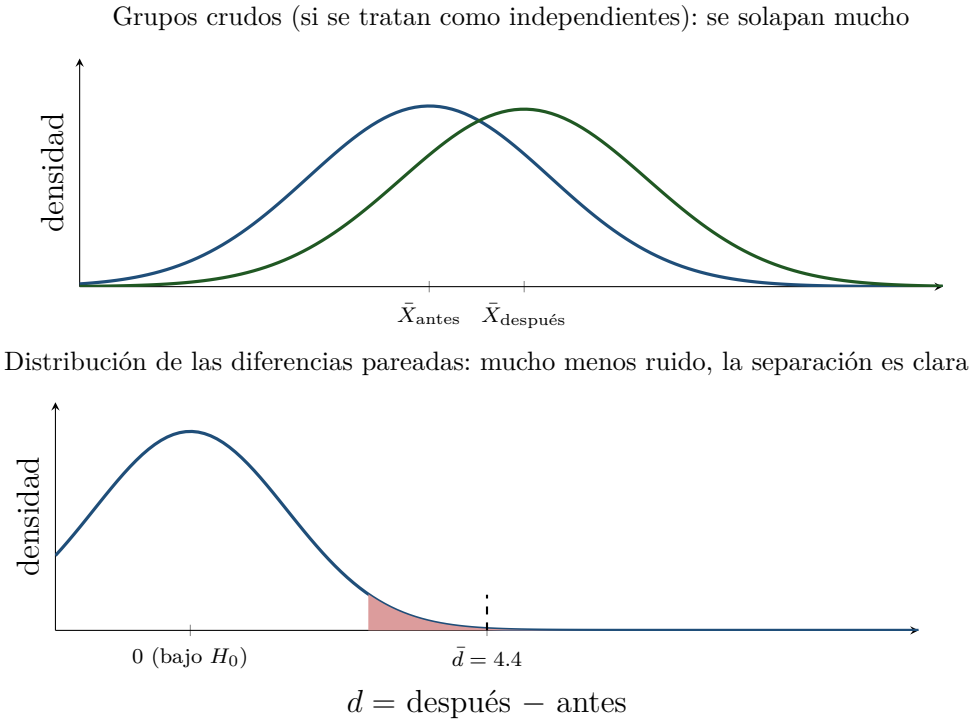

Figura 10: Arriba: los grupos crudos (antes/después) se solapan bastante si se ven como si fueran independientes. Abajo: al mirar las *diferencias individuales*, la variabilidad entre alumnos desaparece y $\bar{d} = 4.4$ queda claramente dentro de la región de rechazo (rojo).

11. Pruebas de dos muestras — proporciones

$$\hat{p}_{\text{pool}} = \frac{x_1 + x_2}{n_1 + n_2}, \qquad z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_{\text{pool}}(1 - \hat{p}_{\text{pool}})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}.$$

Ejemplo 11.1 (Participación electoral entre distritos)**.** Distrito A: 312 de 600 votan ($\hat{p}_1 = 0.520$). Distrito B: 285 de 580 ($\hat{p}_2 = 0.491$). $\hat{p}_{\text{pool}} \approx 0.506$; SE ≈ 0.0291; $z \approx 0.98$. Como $0.98 < 1.96$, **no rechazamos** (p-valor ≈ 0.326): no hay evidencia de que la participación difiera.

Ejemplo 11.2 (Test A/B de conversión)**.** Versión A de una página: 58 de 500 compran ($\hat{p}_1 = 0.116$). Versión B: 82 de 500 ($\hat{p}_2 = 0.164$). ¿B convierte *más*? ($\alpha = 0.05$, una cola.)

$\hat{p}_{\text{pool}} = 0.14$; SE ≈ 0.0219; $z = \dfrac{0.164 - 0.116}{0.0219} \approx 2.19 > 1.645$. **Rechazamos** H_0 (p-valor ≈ 0.014): la versión B convierte significativamente más.

12. La prueba chi-cuadrado (χ^2)

Definición 12.1 (Estadístico χ^2)**.**

$$\chi^2 = \sum_{\text{categorías}} \frac{(O - E)^2}{E},$$

donde O son las frecuencias **observadas** y E las **esperadas** bajo H_0. Siempre $\chi^2 \geq 0$; su distribución depende de los grados de libertad y es **asimétrica a la derecha** (muy distinta de la z/t, simétricas).

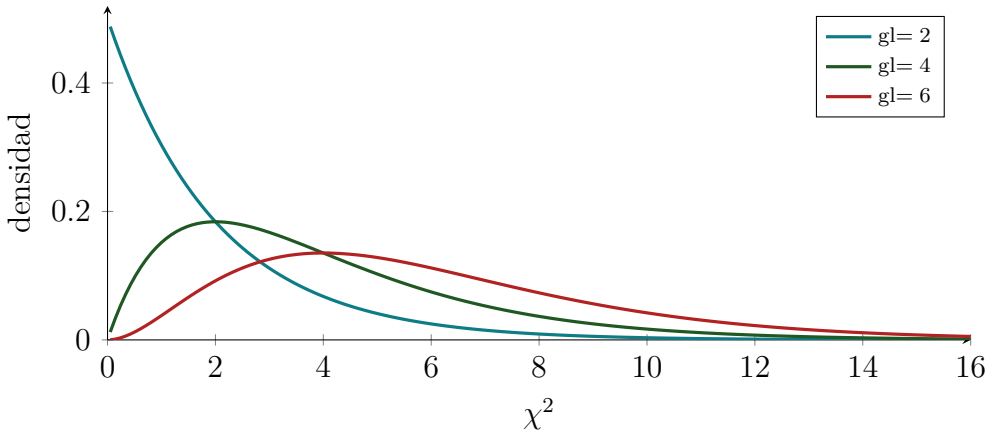

Figura 11: La familia χ^2 según los grados de libertad: siempre positiva y sesgada a la derecha; con más gl, el pico se desplaza a la derecha y la curva se aplana.

En el ejercicio Reto+ de razón de verosimilitud de la Clase 6 apareció $z^2 \sim \chi^2_1$ (teorema de Wilks). Allí era un subproducto de una derivación; aquí χ^2 es la **herramienta central** para dos preguntas nuevas: bondad de ajuste e independencia.

12.1. Bondad de ajuste (goodness-of-fit)

Compara una distribución de categorías **observada** contra una **esperada** (hipótesis o histórico). gl $= k - 1$ ($k =$ núm. de categorías).

Ejemplo 12.1 (Encuesta de preferencias)**.** Históricamente, 4 categorías de opinión se repartían 25 % cada una. En una encuesta reciente ($n = 500$): observados 120, 95, 140, 145.

Esperados bajo H_0: 125 cada uno.

$$\chi^2 = \frac{(120-125)^2}{125} + \frac{(95-125)^2}{125} + \frac{(140-125)^2}{125} + \frac{(145-125)^2}{125} \approx 12.40,$$

gl= 3, crítico $\chi^2_{3,0.95} \approx 7.81$. Como $12.40 > 7.81$, **rechazamos** H_0 (p-valor ≈ 0.006): la distribución de opinión cambió.

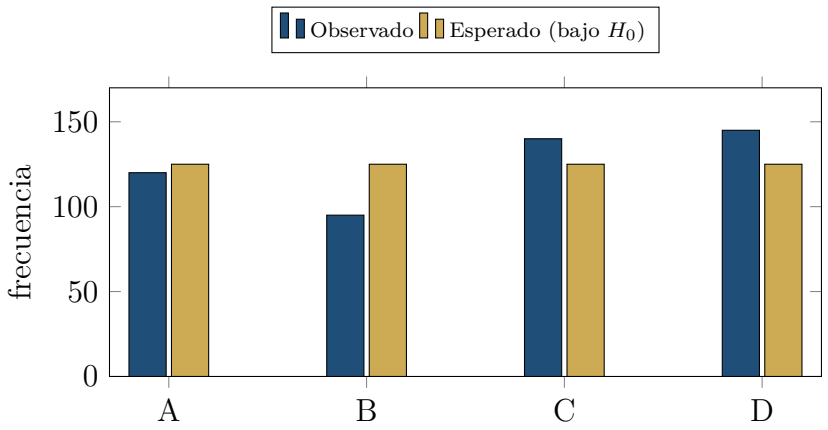

Figura 12: Frecuencias observadas vs. esperadas de las 4 categorías de opinión. Las diferencias parecen chicas a simple vista, pero acumuladas dan $\chi^2 = 12.40$, suficiente para rechazar.

Ejemplo 12.2 (¿El dado está cargado?)**.** Se lanza un dado 60 veces: $8, 12, 10, 9, 7, 14$ (caras 1–6). Esperado si es justo: 10 cada una. $\chi^2 = \sum(O-10)^2/10 \approx 3.40$, gl= 5, crítico $\chi^2_{5,0.95} \approx 11.07$. Como $3.40 < 11.07$, **no rechazamos**: no hay evidencia de que el dado esté cargado (p-valor ≈ 0.64).

12.2. Independencia (tablas de contingencia)

Prueba si dos variables categóricas están **asociadas**. Los conteos esperados de cada celda son $E_{ij} = \dfrac{(\text{total fila } i)(\text{total columna } j)}{\text{total general}}$; $\text{gl} = (r-1)(c-1)$.

Ejemplo 12.3 (Nivel educativo y empleo formal)**.**

	Empleo formal	Empleo informal	Total
Educ. superior	90	60	150
Educ. básica	40	110	150
Total	130	170	300

Esperados bajo independencia: todas las celdas = 65 o 85 (proporcional a los totales). $\chi^2 \approx 33.94$, gl= $(2-1)(2-1) = 1$, crítico $\chi^2_{1,0.95} \approx 3.84$. Como $33.94 \gg 3.84$, **rechazamos**: nivel educativo y empleo formal **sí** están asociados (p-valor $\approx 5.7 \times 10^{-9}$).

Ejemplo 12.4 (Género y participación en un programa)**.** 45 de 100 mujeres participan; 70 de 100 hombres participan. $\chi^2 \approx 12.79$, gl= 1, crítico ≈ 3.84. $12.79 > 3.84 \Rightarrow$ **rechazamos** (p-valor ≈ 0.0003): la participación **sí** depende del género en esta muestra.

El test χ^2 es una *aproximación*, válida cuando los conteos esperados E no son muy chicos (regla práctica: todos $E \geq 5$). Con celdas esperadas menores a 5, el p-valor puede ser poco confiable; en ese caso se usa una alternativa exacta (p. ej. el test exacto de Fisher).

13. La prueba F (cumpliendo lo prometido en la Clase 6)

Al cerrar la Clase 6 quedó pendiente: "revisa distribuciones $\chi^2/t/F$". Ya vimos χ^2; ahora cerramos con F, la prueba **para comparar dos varianzas**.

Definición 13.1 (Estadístico F)**.**

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

Como χ^2, F es siempre positiva y sesgada a la derecha, pero depende de **dos** parámetros (los gl del numerador y del denominador), no de uno solo.

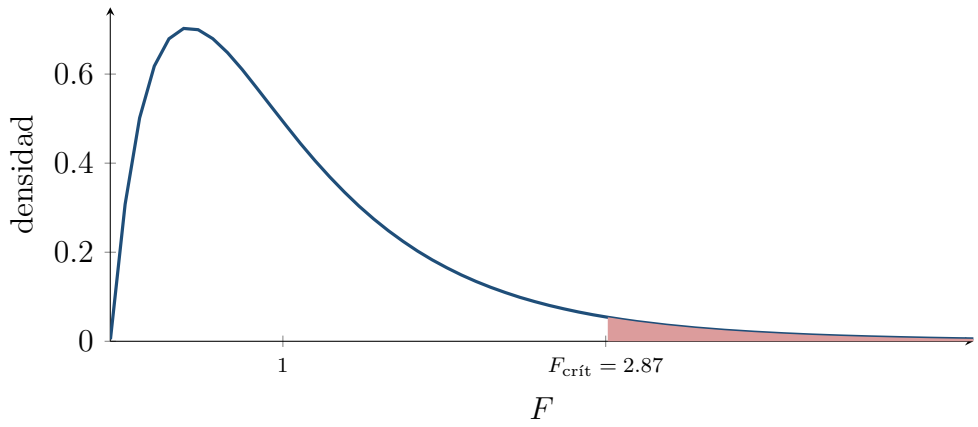

Figura 13: Densidad $F(4, 20)$: siempre positiva, sesgada a la derecha, con dos parámetros (gl del numerador y del denominador). En rojo, la región de rechazo al 5 % ($F > 2.87$).

Ejemplo 13.1 (Variabilidad de dos líneas de producción)**.** Línea 1: $n_1 = 5$, $s_1 = 4.8$. Línea 2: $n_2 = 21$, $s_2 = 2.6$. ¿Son igual de variables? ($\alpha = 0.05$.) $F = \dfrac{4.8^2}{2.6^2} \approx 3.41$, gl$= (4, 20)$, crítico $F_{4,20,0.95} \approx 2.87$. Como $3.41 > 2.87$, **rechazamos** H_0 (p-valor ≈ 0.028): la línea 1 es significativamente más variable.

La prueba F reaparecerá más adelante en el curso como la prueba de significancia **global** de un modelo de regresión (¿el modelo en conjunto explica algo?) — misma lógica (varianza explicada vs. varianza residual), otro contexto.

14. Guía maestra de decisión: ¿qué test uso?

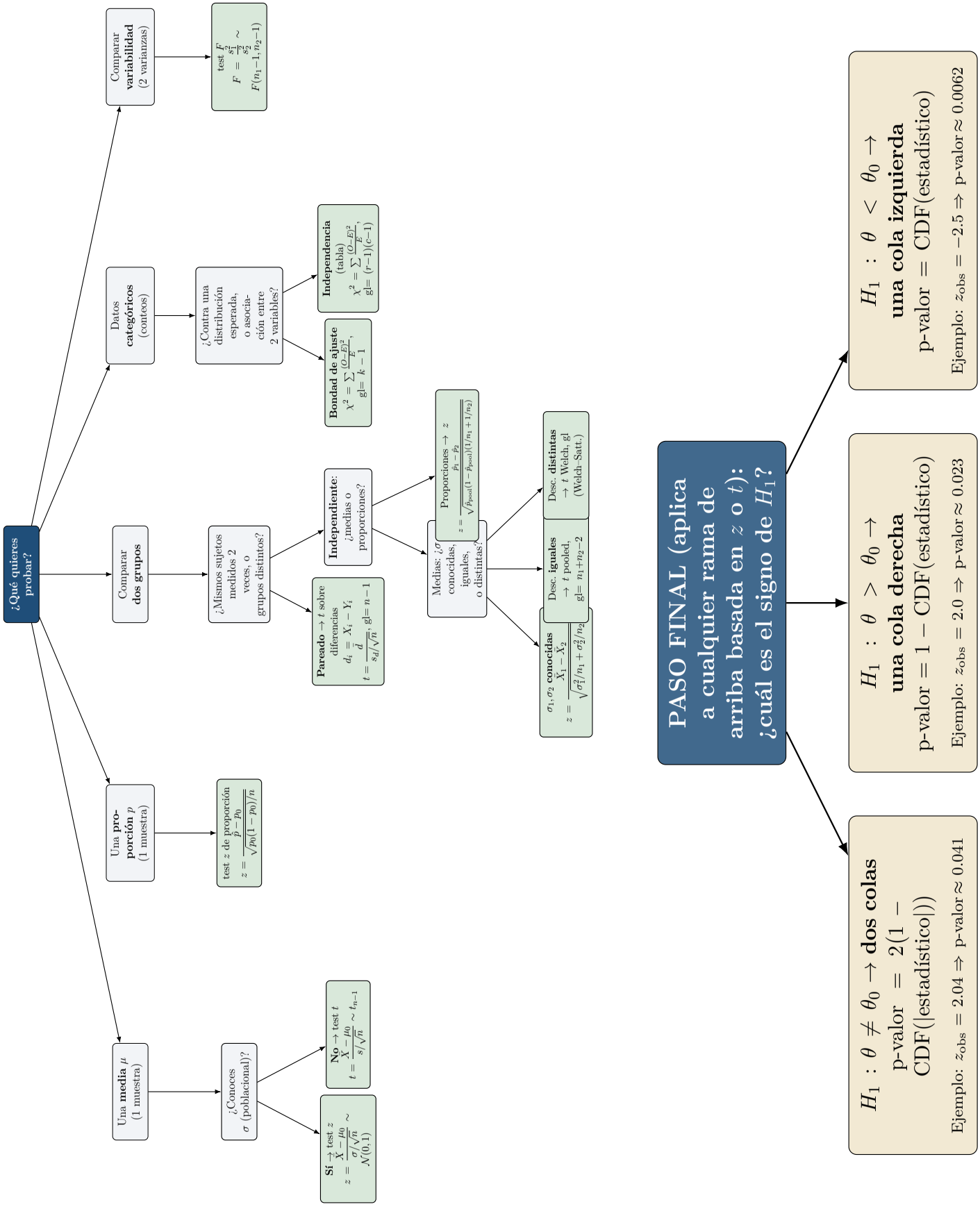
¿Qué quieres probar?
Una media μ (1 muestra)
¿Conoces σ (poblacional)?
Si → test z $z = \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$
No → test t $t = \frac{\bar{X}-\mu_0}{s/\sqrt{n}} \sim t_{n-1}$
Una proporción p (1 muestra)
test z de proporción $z = \frac{\hat{p}-p_0}{\sqrt{p_0(1-p_0)/n}}$
Comparar dos grupos
¿Mismos sujetos medidos 2 veces, o grupos distintos?
Pareado → t sobre diferencias $d_i = X_i - Y_i$, $t = \frac{\bar{d}}{s_d/\sqrt{n}}$, gl= $n-1$
Independiente: ¿medias o proporciones?
Proporciones → z $z = \frac{\hat{p}_1-\hat{p}_2}{\sqrt{\hat{p}_{\text{pool}}(1-\hat{p}_{\text{pool}})(1/n_1+1/n_2)}}$
Medias: ¿σ conocidas, iguales, o distintas?
σ_1, σ_2 conocidas $z = \frac{\bar{X}_1-\bar{X}_2}{\sqrt{\sigma_1^2/n_1+\sigma_2^2/n_2}}$
Desc. iguales → t pooled, gl= n_1+n_2-2
Desc. distintas → t Welch, gl (Welch-Satt.)
Datos categóricos (conteos)
¿Contra una distribución esperada, o asociación entre 2 variables?
Bondad de ajuste $\chi^2 = \sum \frac{(O-E)^2}{E}$, gl= $k-1$
Independencia (tabla) $\chi^2 = \sum \frac{(O-E)^2}{E}$, gl= $(r-1)(c-1)$
Comparar variabilidad (2 varianzas)
test F $F = \frac{s_1^2}{s_2^2} \sim F(n_1-1, n_2-1)$
PASO FINAL (aplica a cualquier rama de arriba basada en z o t): ¿cuál es el signo de H_1?
$H_1 : \theta \neq \theta_0$ → **dos colas** p-valor = $2(1 -$ CDF(|estadístico|)) Ejemplo: $z_{\text{obs}} = 2.04 \Rightarrow$ p-valor ≈ 0.041
$H_1 : \theta > \theta_0 \rightarrow$ **una cola derecha** p-valor = 1 − CDF(estadístico) Ejemplo: $z_{\text{obs}} = 2.0 \Rightarrow$ p-valor ≈ 0.023
$H_1 : \theta < \theta_0 \rightarrow$ **una cola izquierda** p-valor = CDF(estadístico) Ejemplo: $z_{\text{obs}} = -2.5 \Rightarrow$ p-valor ≈ 0.0062

Tabla matriz — identificación

Test	# muestras	Parámetro	Varianzas	Pareado/Indep.
z media	1	media	σ conocida	—
t media	1	media	σ desconocida	—
z proporción	1	proporción	—	—
z dif. medias	2	media	σ_1, σ_2 conocidas	independiente
t pooled	2	media	desconocidas, iguales	independiente
t Welch	2	media	desconocidas, distintas	independiente
t pareado	2 (mismo suj.)	media de dif.	desconocida	**pareado**
z dif. proporciones	2	proporción	—	independiente
χ^2 bondad de ajuste	1	distr. categórica	—	—
χ^2 independencia	1 tabla	asociación	—	—
F	2	varianza	—	independiente

Tabla matriz — mecánica (mismo orden de filas)

Test	Estadístico	Distribución	gl
z media	$(\bar{X}-\mu_0)/(\sigma/\sqrt{n})$	$\mathcal{N}(0,1)$	—
t media	$(\bar{X}-\mu_0)/(s/\sqrt{n})$	t	$n-1$
z proporción	$(\hat{p}-p_0)/\sqrt{p_0(1-p_0)/n}$	$\mathcal{N}(0,1)$	—
z dif. medias	$(\bar{X}_1-\bar{X}_2)/\sqrt{\sigma_1^2/n_1+\sigma_2^2/n_2}$	$\mathcal{N}(0,1)$	—
t pooled	$(\bar{X}_1-\bar{X}_2)/(s_p\sqrt{1/n_1+1/n_2})$	t	n_1+n_2-2
t Welch	$(\bar{X}_1-\bar{X}_2)/\sqrt{s_1^2/n_1+s_2^2/n_2}$	t	Welch–Satterthwaite
t pareado	$\bar{d}/(s_d/\sqrt{n})$	t	$n-1$
z dif. proporciones	$(\hat{p}_1-\hat{p}_2)/\sqrt{\hat{p}_{\text{pool}}(1-\hat{p}_{\text{pool}})(1/n_1+1/n_2)}$	$\mathcal{N}(0,1)$	—
χ^2 bondad de ajuste	$\sum(O-E)^2/E$	χ^2	$k-1$
χ^2 independencia	$\sum(O-E)^2/E$	χ^2	$(r-1)(c-1)$
F	s_1^2/s_2^2	F	(n_1-1, n_2-1)

15. Catálogo rápido de referencia (escenarios)

Esto **no son ejercicios**: es tu chuleta de consulta rápida antes de un examen. Cada fila es un escenario en una frase, con el test/cola/estadístico que le corresponde — sin resolverlo.

Test z de una media.

1	¿El tiempo medio de espera *supera* 10 min? (σ conocida)	1 (der.)	$z = (\bar{X} - 10)/(\sigma/\sqrt{n})$
2	¿El peso medio de un producto *difiere* de 500 g? (σ conocida)	2	$z = (\bar{X} - 500)/(\sigma/\sqrt{n})$
3	¿La temperatura media *bajó* de 36.5°? (σ conocida)	1 (izq.)	$z = (\bar{X} - 36.5)/(\sigma/\sqrt{n})$
4	¿El ingreso medio mensual *cambió* de S/2000? (σ conocida)	2	igual que #2
5	¿La velocidad media de descarga *superó* 50 Mbps?	1 (der.)	igual que #1

Test t de una media.

6	¿El gasto medio de hogares *difiere* de S/800? (n chico, σ desconocida)	2	$t = (\bar{X} - 800)/(s/\sqrt{n})$, gl= $n-1$
7	¿El rendimiento medio de un cultivo *superó* 30 qq/ha?	1 (der.)	$t = (\bar{X} - 30)/(s/\sqrt{n})$
8	¿La duración media de una batería *bajó* de 8 h?	1 (izq.)	igual que #7 con signo
9	¿El tiempo medio de reparación *difiere* de 5 h?	2	igual que #6

Test z de una proporción.

10	¿La tasa de aprobación *superó* el 70 % histórico?	1 (der.)	$z = (\hat{p} - 0.70)/\sqrt{0.70 \cdot 0.30/n}$
11	¿La proporción de defectuosos *cambió* de 5 %?	2	similar, dos colas
12	¿La adherencia a un tratamiento *mejoró* de 15 %?	1 (der.)	igual que #10

Dos medias independientes.

13	¿Salarios promedio en Lima vs. provincias son distintos? (grupos distintos)	2	t pooled o Welch según varianzas
14	¿Dos máquinas producen piezas de largo medio distinto?	2	igual que #13
15	¿Un grupo tratamiento tiene ingreso medio mayor que control (RCT)?	1 (der.)	igual, una cola

Dos medias pareadas.

16	¿El puntaje *mejoró* tras el curso, mismos alumnos antes/después?	1 (der.)	$t = \bar{d}/(s_d/\sqrt{n})$
17	¿La presión arterial *cambió* en los mismos pacientes con un fármaco?	2	igual que #16
18	¿El mismo grupo de obreros produce distinto de mañana vs. tarde?	2	igual que #16

Dos proporciones.

19	¿La tasa de aprobación *difiere* entre dos secciones?	2	z con $\hat{p}_{\text{pool}}$
20	¿La versión B de una app convierte *más* que la A?	1 (der.)	igual que #19
21	¿La participación electoral *difiere* entre dos distritos?	2	igual que #19

Chi-cuadrado, bondad de ajuste.

22	¿La distribución de preferencia partidaria observada coincide con la histórica?	— (1 cola por construcción)	gl= $k-1$
23	¿Un dado (o moneda) está cargado?	—	igual que #22
24	¿Las ventas por día de la semana siguen la distribución esperada?	—	igual que #22

Chi-cuadrado, independencia.

25	¿Nivel educativo está *asociado* con empleo formal?	—	gl= $(r-1)(c-1)$
26	¿Género está asociado con participación en un programa?	—	igual que #25
27	¿Región de residencia está asociada con preferencia de partido?	—	igual que #25

Prueba F.

28	¿La variabilidad de notas es distinta entre dos profesores?	2 (o 1, según convención)	$F = s_1^2/s_2^2$
29	¿Dos líneas de producción tienen igual variabilidad?	igual	igual que #28

16. Errores comunes al interpretar pruebas de hipótesis

1. **p-valor $\neq \mathbb{P}(H_0 \mid \textbf{datos})$.** El p-valor es $\mathbb{P}(\text{datos} \mid H_0)$, no al revés (falacia del fiscal).
2. **No rechazar $\neq$ probar H_0.** Solo dice que no hay evidencia suficiente; quizá la muestra fue chica (potencia baja).
3. **Significancia estadística $\neq$ relevancia práctica.** Un efecto diminuto puede ser "significativo" con n enorme.
4. **p-hacking / pruebas múltiples.** Probar muchas hipótesis a la vez infla los falsos positivos (con m pruebas independientes, $\mathbb{P}(\geq 1 \text{ falso positivo}) = 1-(1-\alpha)^m$); corrección de Bonferroni: usar α/m.
5. **Confundir pareado con independiente.** Como vimos en §10, tratar datos parea-

dos como independientes (o viceversa) puede cambiar por completo la conclusión.

6. **Chi-cuadrado con conteos esperados chicos.** Con $E < 5$ en varias celdas, el p-valor de χ^2 deja de ser confiable (§12.2).

17. En R: recorrido de todas las pruebas

Todas las familias, en un solo lugar

```
# 1) z/t de una media -- R no tiene z.test base; con sigma conocida se calcula
    a mano,
#    pero t.test() cubre el caso mas comun (sigma desconocida):
x <- c(812,845,798,860,830,875,805,850,822,840,865,795,835,858,810,848)
t.test(x, mu = 800)                        # una muestra

# 2) t pareado
antes    <- c(58,62,55,70,60,65,52,68,59,63)
despues  <- c(64,66,61,73,62,70,55,74,63,68)
t.test(despues, antes, paired = TRUE)     # pareado

# 3) t independiente (pooled vs Welch)
t.test(despues, antes, paired = FALSE, var.equal = TRUE)   # pooled
t.test(despues, antes, paired = FALSE, var.equal = FALSE)  # Welch (default)

# 4) proporciones (una y dos muestras)
prop.test(x = 52, n = 250, p = 0.15, alternative = "greater")
prop.test(x = c(58, 82), n = c(500, 500))

# 5) chi-cuadrado (bondad de ajuste e independencia)
chisq.test(c(120,95,140,145), p = rep(0.25, 4))
tabla <- matrix(c(90,60,40,110), nrow = 2, byrow = TRUE)
chisq.test(tabla)

# 6) F (comparacion de varianzas)
var.test(x1, x2)   # razon de varianzas, con su IC y p-valor
```

18. Resumen de la clase

Resumen de la Clase 7

- Normalizamos ($z = \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}$, etc.) para convertir una distancia cruda en "número de errores estándar", comparable entre problemas.
- z (media, σ conocida), t (media, σ desconocida), z (proporción): mismo esqueleto,

distinta "regla".

- Valor crítico y p-valor son **el mismo criterio**, visto desde dos ángulos: $|z_{\text{obs}}| > z_{\text{crít}} \iff$ p-valor $< \alpha$.
- Tipo I (α, falsa alarma) y Tipo II (β, no detectar) se entienden igual de simple (detector de humo) que formal (tabla 2×2).
- La potencia $1 - \beta$ sube con n, con un efecto Δ mayor, o con menos ruido σ.
- **Pareado vs. independiente** lo decide el *diseño*, no los números — y cambia radicalmente el resultado.
- χ^2 (categorías: bondad de ajuste e independencia) y F (varianzas) completan la caja de herramientas junto a z/t.

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.

- Este documento se solapa intencionalmente con la sección de pruebas de hipótesis de la Clase 6 (Confidence Intervals and Hypothesis Testing): aquí se profundiza y se agregan χ^2, F y pruebas de dos muestras.
- Orloff, J. & Kamrin, J. *18.05 Introduction to Probability and Statistics*, MIT OpenCourseWare — capítulos de χ^2, F, t de dos muestras (pooled/Welch/pareado) y potencia/tamaño de muestra.
- Diez, D., Çetinkaya-Rundel, M. & Barr, C. *OpenIntro Statistics* — §6.2 (diferencia de dos proporciones).

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`.