Trabajo práctico 10: Pruebas de hipótesis: la lógica

AL TERMINAR ESTA GUÍA

- Plantear la hipótesis nula y la alternativa a partir de una pregunta, y decidir si la prueba es unilateral o bilateral.
- Resolver una prueba t para una media y una prueba para una proporción, con la región crítica y con el p-valor.
- Calcular la probabilidad de un error de tipo II y la potencia, y ver cómo cambian con α y con el tamaño de la muestra.
- Interpretar correctamente un p-valor y distinguir la significación estadística de la relevancia práctica.

PARA EL DOCENTE

El ejercicio 1 abre el encuentro: conviene discutir en voz alta las situaciones del arsénico y del juicio, que muestran que elegir H_0 decide qué error se controla. Los ejercicios 2 a 4 son para el práctico, con las tablas de la t y de la normal; el 2 y el 3 son los centrales y el 4 se puede dejar como tarea.

El 5 necesita computadora y se complementa con los simuladores del p-valor y de los errores. El 6 retoma las lecturas erróneas del p-valor, que son las que aparecen en los parciales y en los diarios.

EJERCICIO 1 Plantear las hipótesis

A mano

Para cada situación, planteá la hipótesis nula y la alternativa, en símbolos y en palabras. Decidí si la prueba es unilateral o bilateral, y describí qué sería, en ese contexto, cometer un error de tipo I y uno de tipo II.

a) Un fabricante de lámparas LED dice en la caja que duran en promedio 25.000 horas. Una asociación de consumidores sospecha que duran menos.

> RESOLUCIÓN
>
> $H_0 : \mu = 25\,000 \qquad H_1 : \mu < 25\,000$
>
> Unilateral a izquierda: a la asociación solo le interesa detectar que duran menos; si duraran más, no habría reclamo. Error de tipo I: concluir que las lámparas duran menos de lo anunciado cuando en realidad cumplen, y acusar al fabricante sin razón. Error de tipo II: no detectar que duran menos cuando de verdad es así, y que los consumidores sigan pagando por una duración que no reciben.

> La afirmación del fabricante va en H_0 porque es la que se pone a prueba: se le da el beneficio de la duda, y la carga de la prueba la tiene quien sostiene lo contrario.

b) Durante años, el peso medio al nacer en una maternidad fue de 3.300 g. Después de un programa de controles prenatales, se quiere saber si cambió.

> RESOLUCIÓN
>
> $H_0 : \mu = 3300 \qquad H_1 : \mu \neq 3300$
>
> Bilateral: la pregunta es si cambió, en cualquier dirección. Aunque se espere que el programa aumente el peso, un cambio hacia abajo también sería importante, y una prueba unilateral no lo podría detectar. Error de tipo I: concluir que el peso medio cambió cuando sigue igual. Error de tipo II: no detectar un cambio que sí ocurrió.
>
> La dirección de la prueba se decide antes de ver los datos, a partir de la pregunta. Elegirla después, mirando para qué lado salió la diferencia, es hacer trampa: duplica en la práctica el nivel de significación.

c) Con el método de enseñanza habitual, aprueba el 60 % de los estudiantes de una materia. Una docente prueba un método nuevo y quiere saber si con él aprueban más.

> RESOLUCIÓN
>
> $H_0 : p = 0{,}60 \qquad H_1 : p > 0{,}60$
>
> Unilateral a derecha. Error de tipo I: adoptar el método nuevo creyendo que es mejor cuando no lo es (con el costo de capacitar y cambiar materiales para nada). Error de tipo II: descartar un método que de verdad mejora la aprobación.
>
> H_0 también se puede escribir $p \leq 0{,}60$: la cuenta se hace con el valor del borde, 0,60, porque es el caso de H_0 más difícil de distinguir de H_1. Por eso la igualdad va siempre en la hipótesis nula: hace falta un valor concreto para calcular la distribución del estadístico.

d) Un municipio mide el arsénico del agua de un pozo y lo clausura si la concentración media supera el límite de 0,01 mg/L.

> RESOLUCIÓN
>
> $H_0 : \mu = 0{,}01 \qquad H_1 : \mu > 0{,}01$
>
> Unilateral a derecha. Error de tipo I: clausurar un pozo que cumple el límite (un costo económico). Error de tipo II: dejar abierto un pozo contaminado (un costo de salud). Acá el error de tipo II es el más grave, y conviene usar un nivel de significación más alto o una muestra grande, para que la prueba tenga potencia.
>
> Para discutir: hay quien sostiene que en salud pública la carga de la prueba debería invertirse, con H_0: $\mu \geq 0{,}01$ (el agua está contaminada hasta que se demuestre lo contrario). Así, el error que se controla con α es justamente el de declarar segura el agua cuando no lo es. Qué va en H_0 no es una cuestión técnica: decide qué error se controla.

e) En un juicio penal, la hipótesis nula es que la persona acusada es inocente.

RESOLUCIÓN

H_0: la persona es inocente. H_1: es culpable. Error de tipo I: condenar a una persona inocente. Error de tipo II: absolver a una persona culpable. El sistema penal fija un umbral alto («más allá de toda duda razonable») porque considera mucho más grave el primer error: es un α chico.

La sentencia es «culpable» o «no culpable», nunca «inocente»: no encontrar pruebas suficientes no demuestra la inocencia. Con las pruebas de hipótesis pasa lo mismo: se rechaza H_0 o no se rechaza, pero no se «acepta». Un resultado no significativo puede deberse a que H_0 es cierta o a que la muestra fue chica.

EJERCICIO 2 ¿Las bolsas pesan lo que dicen?

A mano

Una cooperativa envasa nueces en bolsas que dicen «500 g». Una inspección de defensa del consumidor pesó 16 bolsas elegidas al azar de un lote y obtuvo, en gramos:

Bolsas 1 a 8	500	503	503	498	504	500	507	490

Bolsas 9 a 16	498	488	494	501	481	485	497	489

Se puede suponer que el peso de las bolsas tiene distribución aproximadamente normal. La inspección usa un nivel de significación del 5 %.

a) Planteá las hipótesis. ¿Por qué la prueba es unilateral?

RESOLUCIÓN

$H_0 : \mu = 500 \qquad H_1 : \mu < 500$

A la inspección le interesa detectar si las bolsas traen menos de lo que dicen, que es lo que perjudica al consumidor. Si trajeran más, no habría infracción. La pregunta se fija antes de pesar: no se elige la dirección mirando los datos.

b) Calculá la media y el desvío de la muestra, y el estadístico de prueba.

RESOLUCIÓN

$$\bar{x} = \frac{7938}{16} = 496{,}125 \qquad s = 7{,}518 \qquad EE = \frac{7{,}518}{\sqrt{16}} = 1{,}8794$$

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{496{,}125 - 500}{1{,}8794} = -2{,}062$$

El estadístico mide a cuántos errores estándar está la media muestral del valor de H_0. Si H_0 es cierta, tiene distribución t con n − 1 = 15 grados de libertad.

c) Con α = 0,05, ¿cuál es la región crítica? ¿Qué decisión se toma?

RESOLUCIÓN

$$\text{Se rechaza } H_0 \text{ si } t < -t_{15;\,0{,}95} = -1{,}753 \qquad t = -2{,}062 < -1{,}753$$

El estadístico cae en la región crítica: se rechaza H_0. Toda el área de α va a la cola izquierda, porque solo los valores muy bajos de $\bar{x}$ son evidencia a favor de H_1.

d) Acotá el p-valor con la tabla de la t. Si tenés una calculadora o un software, calculalo exactamente.

RESOLUCIÓN

En la fila de 15 grados de libertad, $|t| = 2{,}062$ queda entre 1,753 (columna 0,05) y 2,131 (columna 0,025). Entonces $0{,}025 < p < 0{,}05$. El valor exacto es $p = 0{,}0285$.

El p-valor es la probabilidad de obtener una media de 496,125 g o menos si en realidad el peso medio fuera 500 g. Como es menor que 0,05, se llega a la misma decisión que con la región crítica: los dos caminos son equivalentes.

e) Si la inspección hubiera planteado una prueba bilateral ($H_1: \mu \neq 500$), ¿la decisión sería la misma?

RESOLUCIÓN

$$|t| = 2{,}062 < t_{15;\,0{,}975} = 2{,}131 \qquad p_{\text{bilateral}} = 2 \cdot 0{,}0285 = 0{,}0570$$

No: con la prueba bilateral no se rechaza H_0. El 5 % se reparte entre las dos colas, el valor crítico se aleja y el p-valor se duplica. Por eso la dirección se decide antes de ver los datos: quien elige la prueba unilateral después de ver para qué lado salió la diferencia está usando, en los hechos, un nivel del 10 %.

f) Redactá la conclusión para el informe de la inspección. ¿Qué error se podría estar cometiendo?

RESOLUCIÓN

«Con un nivel de significación del 5 %, los datos dan evidencia de que el peso medio de las bolsas del lote es menor que 500 g ($t = -2{,}06$, gl = 15, $p = 0{,}028$). El peso medio estimado es de 496,1 g.» Como se rechazó H_0, el único error posible es el de tipo I: que las bolsas cumplan y la muestra haya salido baja por azar. La probabilidad de que pase eso, si H_0 es cierta, es $\alpha = 0{,}05$.

Además, que sea significativo no dice si importa: 3,9 g menos en una bolsa de 500 es el 0,8 %. Si es una diferencia relevante lo deciden las normas de envasado, no la prueba.

EJERCICIO 3 La cobertura de vacunación

A mano

Un informe oficial sostiene que el 90 % de los chicos que ingresan a primer grado en una provincia tiene completo el esquema de vacunación. Un equipo de investigación sospecha que la cobertura real es menor. Revisó las libretas de 250 chicos de primer grado elegidos al azar y encontró que 215 tenían el esquema completo.

a) Planteá las hipótesis y verificá que se pueda usar la aproximación normal.

RESOLUCIÓN

$$H_0 : p = 0{,}90 \qquad H_1 : p < 0{,}90$$

Se usa el valor de H_0 para verificar las condiciones: $n \cdot p_0 = 225$ y $n \cdot (1 - p_0) = 25$, los dos por encima de 10.

b) Calculá la proporción muestral y el estadístico de prueba. ¿Por qué el error estándar se calcula con 0,90 y no con la proporción de la muestra?

> RESOLUCIÓN
>
> $$\hat{p} = \frac{215}{250} = 0{,}8600 \qquad z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0\,(1-p_0)}{n}}} = \frac{0{,}8600 - 0{,}90}{\sqrt{\frac{0{,}9 \cdot 0{,}1}{250}}} = \frac{-0{,}0400}{0{,}0190} = -2{,}108$$
>
> La prueba calcula qué pasaría si H_0 fuera cierta, y si H_0 es cierta la proporción es 0,90: con ella se obtiene el error estándar. En el intervalo de confianza, en cambio, no hay un valor supuesto y se usa $\hat{p}$ (daría 0,0219).

c) Calculá el p-valor con la tabla de la normal y explicá qué significa en este contexto.

> RESOLUCIÓN
>
> $$p = P(Z < -2{,}11) = 0{,}0174$$
>
> Si la cobertura fuera realmente del 90 %, la probabilidad de que una muestra de 250 chicos diera una cobertura de 86,0 % o menos sería de 0,0174. Es la probabilidad de un resultado así de extremo suponiendo H_0 cierta, no la probabilidad de que H_0 sea cierta.

d) ¿Qué se decide con α = 0,05? ¿Y con α = 0,01?

> RESOLUCIÓN
>
> Con α = 0,05 se rechaza H_0, porque 0,0174 < 0,05: hay evidencia de que la cobertura es menor que el 90 %. Con α = 0,01 no se rechaza, porque 0,0174 > 0,01. Lo mismo con los valores críticos: z = −2,108 es menor que −1,645 pero no que −2,326.
>
> Por eso conviene informar el p-valor y no solo la decisión: quien lee puede ver que la evidencia es moderada y aplicar su propio criterio. El nivel se fija antes de mirar los datos; cambiarlo después para que dé significativo no es válido.

e) Redactá la conclusión. ¿Demuestra el estudio que el informe oficial está mal?

> RESOLUCIÓN
>
> «Con un nivel de significación del 5 %, los datos indican que la cobertura de vacunación completa en primer grado es menor que el 90 % informado (cobertura en la muestra: 86,0 %; z = −2,11; p = 0,017).»
>
> No lo demuestra: hay un riesgo de error de tipo I, y la evidencia no es muy fuerte. Tampoco dice por qué la diferencia: puede ser una cobertura real más baja o una diferencia en cómo se cuentan los esquemas completos (el informe oficial puede usar registros y el estudio, libretas). La prueba responde a una pregunta estadística; la interpretación necesita conocer cómo se produjo cada dato.

EJERCICIO 4 Los dos errores y la potencia

A mano

En una población de adultos, la presión arterial sistólica media es de 120 mmHg, con un desvío de 18 mmHg. Se sospecha que en quienes trabajan de noche es más alta. Se va a medir la presión de 25 trabajadores nocturnos elegidos al azar y a plantear H_0: μ = 120 contra H_1: μ >

120, con α = 0,05. Para simplificar, se supone que el desvío es conocido y el mismo en los dos grupos, y se usa la normal.

a) Describí con palabras, en este contexto, el error de tipo I y el de tipo II. ¿Cuál de los dos controla el investigador al fijar α?

> RESOLUCIÓN
>
> Tipo I: concluir que los trabajadores nocturnos tienen la presión más alta cuando en realidad no es así. Tipo II: no detectar que la tienen más alta cuando sí la tienen. Al fijar α = 0,05 se controla el primero: si H_0 es cierta, la prueba se equivoca el 5 % de las veces. La probabilidad del segundo, β, depende de cuánto más alta sea la presión de verdad, del tamaño de la muestra y del desvío.

b) ¿A partir de qué media muestral se rechaza H_0?

> RESOLUCIÓN
>
> $$EE = \frac{\sigma}{\sqrt{n}} = \frac{18}{\sqrt{25}} = 3{,}6 \qquad \bar{x}_c = 120 + 1{,}645 \cdot 3{,}6 = 125{,}922$$
>
> Se rechaza H_0 si la media de los 25 trabajadores supera 125,922 mmHg. El valor crítico se calcula suponiendo H_0 cierta: deja un 5 % de probabilidad a la derecha cuando μ = 120.

c) Si en realidad la presión media de los trabajadores nocturnos fuera 125 mmHg, ¿cuál sería la probabilidad de cometer un error de tipo II? ¿Y la potencia de la prueba?

> RESOLUCIÓN
>
> $$\beta = P(\bar{X} < 125{,}922 \mid \mu = 125) = P\left(Z < \frac{125{,}922 - 125}{3{,}6}\right) = P(Z < 0{,}26) = 0{,}6026$$
>
> $$\text{Potencia} = 1 - \beta = 0{,}3974$$
>
> Ahora la distribución de x̄ está centrada en 125, no en 120: el valor crítico sigue siendo el mismo, pero la probabilidad se calcula con la distribución verdadera. Con esta muestra, una diferencia real de 5 mmHg pasaría inadvertida en el 60 % de los estudios.

d) Repetí el cálculo de β con α = 0,01. ¿Qué pasa con el error de tipo II?

> RESOLUCIÓN
>
> $$\bar{x}_c = 120 + 2{,}326 \cdot 3{,}6 = 128{,}374 \qquad \beta = P\left(Z < \frac{128{,}374 - 125}{3{,}6}\right) = P(Z < 0{,}94) = 0{,}8264$$
>
> β sube de 0,6026 a 0,8264. Con la misma muestra, achicar un error agranda el otro: exigir más evidencia para rechazar hace que se rechace menos, también cuando H_0 es falsa.

e) Volvé a α = 0,05 y calculá β si se midieran 100 trabajadores en lugar de 25.

> RESOLUCIÓN
>
> $$EE = \frac{18}{\sqrt{100}} = 1{,}80 \qquad \bar{x}_c = 120 + 1{,}645 \cdot 1{,}80 = 122{,}961 \qquad \beta = P(Z < -1{,}13) = 0{,}1292$$

β baja a 0,1292, y la potencia sube a 0,8708. La única manera de bajar los dos errores a la vez es aumentar la muestra. Por eso, antes de un estudio se calcula el tamaño necesario para tener una potencia razonable (en general, del 80 %) para detectar la diferencia más chica que importaría en la práctica.

EJERCICIO 5 Pruebas con la EPH, y un p-valor simulado

Con software

Se usa la base «Ocupación e ingresos en Gran Catamarca», extracto de la EPH del primer trimestre de 2026 (estadistica.ar/datos/eph-catamarca-2026t1). Se puede usar R, Python, Excel, InfoStat o SPSS. En R, t.test(x, mu = ...) hace la prueba t para una media y prop.test(k, n, p = ...) la de una proporción.

Como en la semana anterior, se trata a la base como una muestra aleatoria simple de personas, sin el ponderador ni el diseño muestral de la EPH.

a) Alguien afirma que el ingreso medio de la ocupación principal en Gran Catamarca es de al menos un millón de pesos. Con las personas ocupadas que declararon ingreso (ESTADO = 1 y P21 > 0), probá H_0: μ = 1.000.000 contra H_1: μ < 1.000.000.

RESOLUCIÓN

$$n = 594 \qquad \bar{x} = 881.992 \qquad s = 935.780 \qquad t = \frac{881.992 - 1\,000\,000}{935.780/\sqrt{594}} = -3{,}073$$

El p-valor es 0,0011: se rechaza H_0 con cualquier nivel habitual. Los datos indican que el ingreso medio es menor que un millón. En R: x ← eph$P21[eph$ESTADO == 1 & eph$P21 > 0]; t.test(x, mu = 1e6, alternative = "less"). La asimetría de los ingresos no es un problema con 594 casos.

b) Supongamos que la tasa de desocupación de referencia es del 6 %. Probá si la de Gran Catamarca es distinta, con una prueba bilateral y la aproximación normal.

RESOLUCIÓN

$$\hat{p} = \frac{26}{657} = 0{,}0396 \qquad z = \frac{0{,}0396 - 0{,}06}{\sqrt{0{,}06 \cdot 0{,}94/657}} = -2{,}205 \qquad p = 2\,P(Z < -2{,}205) = 0{,}0275$$

Con α = 0,05 se rechaza H_0: la tasa sería distinta del 6 % (más baja). Pero el p-valor no está lejos de 0,05, y esta cuenta ignora el diseño muestral, que agranda el error estándar: con el diseño de la EPH, la conclusión podría cambiar. prop.test(26, 657, p = 0.06) da un p-valor algo mayor, porque aplica la corrección por continuidad.

c) Simulá 4.000 muestras de 20 datos de una normal con media 50 y desvío 10, y a cada una aplicale la prueba t bilateral de H_0: μ = 50. ¿En qué proporción de las muestras el p-valor es menor que 0,05? ¿Y menor que 0,10? ¿Y que 0,50?

RESOLUCIÓN

En R: pv ← replicate(4000, t.test(rnorm(20, 50, 10), mu = 50)$p.value); mean(pv < 0.05); hist(pv).

En una corrida, el p-valor fue menor que 0,05 en el 5,2 % de las muestras, menor que 0,10 en el 9,3 % y menor que 0,50 en el 50,2 %. Cuando H_0 es cierta, el p-valor tiene distribución uniforme entre 0 y 1: cualquier valor es igual de probable, y el histograma sale plano. Por eso, con α = 0,05, la prueba rechaza

> por error el 5 % de las veces: eso es exactamente el error de tipo I.

d) Repetí la simulación con muestras de una normal con media 55 (H_0 es falsa). ¿En qué proporción se rechaza H_0 con $\alpha = 0{,}05$?

> RESOLUCIÓN
>
> En una corrida se rechazó H_0 en el 57,2 % de las muestras: esa es la potencia de la prueba para una diferencia de medio desvío con 20 datos. En el resto de los casos se cometió un error de tipo II, aunque la diferencia existía. Ahora el histograma de los p-valores se amontona cerca de 0.

e) Separá los p-valores de la simulación con H_0 cierta en grupos de 20, como si fueran 20 investigaciones de algo que no tiene efecto. ¿En qué proporción de los grupos aparece al menos un resultado «significativo»? Compará con la cuenta $1 - 0{,}95^{20}$.

> RESOLUCIÓN
>
> $$1 - 0{,}95^{20} = 0{,}6415$$
>
> En la simulación, el 65,0 % de los grupos tuvo al menos un p-valor menor que 0,05. Si se hacen veinte pruebas sobre cosas sin efecto, lo más probable es que alguna dé significativa por azar. Publicar solo esa, o probar muchas variables hasta que una dé, es lo que se llama p-hacking, y explica parte de los resultados que después no se replican.

EJERCICIO 6 ¿Qué dice un p-valor?

Para discutir

Un ensayo comparó un medicamento contra un placebo para bajar la presión arterial, con H_0: el medicamento no tiene efecto. El resultado fue $p = 0{,}03$. Decidí si cada afirmación es correcta y, si no lo es, corregila.

a) «Hay un 3 % de probabilidad de que la hipótesis nula sea cierta.»

> RESOLUCIÓN
>
> Incorrecta: es la confusión más común. El p-valor se calcula suponiendo que H_0 es cierta: es la probabilidad de obtener un resultado tan extremo como el observado, o más, si el medicamento no tuviera efecto. No es la probabilidad de H_0 dados los datos; para eso haría falta una probabilidad previa de que el medicamento funcione, como en el teorema de Bayes.

b) «Si el medicamento no tuviera ningún efecto, una diferencia como la observada, o mayor, aparecería en alrededor del 3 % de los ensayos.»

> RESOLUCIÓN
>
> Correcta: es la definición. Un p-valor chico indica que los datos serían raros si H_0 fuera cierta, y por eso se los toma como evidencia en contra de H_0.

c) «Como $p < 0{,}05$, el medicamento baja mucho la presión.»

> RESOLUCIÓN
>
> Incorrecta. El p-valor no mide el tamaño del efecto. Con una muestra muy grande, una baja de 1 mmHg, sin importancia clínica, puede dar un p-valor chiquísimo. Para saber cuánto baja hay que mirar la estimación y su intervalo de confianza, y para saber si importa, el criterio clínico.

d) «Otro ensayo, con menos pacientes, dio p = 0,08. Eso demuestra que el medicamento no funciona.»

> RESOLUCIÓN
>
> Incorrecta. No rechazar H_0 no prueba que sea cierta: puede faltar potencia. Un p-valor de 0,08 dice que la evidencia no alcanzó el umbral, no que haya evidencia de ausencia de efecto. Además, los dos resultados no se contradicen: un ensayo chico puede no detectar un efecto que existe.

e) «Si se repite el ensayo, hay un 97 % de probabilidad de volver a obtener un resultado significativo.»

> RESOLUCIÓN
>
> Incorrecta. La probabilidad de que un nuevo ensayo dé significativo es su potencia, que depende del efecto verdadero y del tamaño de la muestra, no de 1 − p. Aun si el efecto verdadero fuera exactamente el observado, la probabilidad de que un ensayo igual vuelva a dar $p < 0,05$ rondaría el 58 %: poco más que tirar una moneda.

f) «Los investigadores habían fijado $\alpha = 0,05$. Si hubiera dado p = 0,07, podrían haber usado $\alpha = 0,10$ y concluir igual que el medicamento funciona.»

> RESOLUCIÓN
>
> Incorrecta. El nivel de significación se fija antes de ver los datos. Elegirlo después, para que el resultado dé significativo, invalida la prueba: el error de tipo I deja de ser el α declarado. Lo mismo vale para cambiar la hipótesis, sacar pacientes o probar varias variables hasta que alguna dé.

g) «Si se fija $\alpha = 0,05$ y se aplica la prueba a muchos medicamentos sin efecto, a la larga se va a concluir por error que funciona uno de cada veinte.»

> RESOLUCIÓN
>
> Correcta: es el sentido del nivel de significación. Es una propiedad del procedimiento a la larga, igual que el nivel de confianza de un intervalo.
