ESTADISTICA.AR

Pruebas de hipótesis: la lógica

Estadística · Semana 10 · Unidad 6: Pruebas de hipótesis

1 / 27

NOTAS · 1

Pensada para 90 a 120 minutos. Es una clase de ideas más que de cuentas: si se entiende la lógica, todas las pruebas que vienen (dos grupos, chi cuadrado, ANOVA, regresión) son la misma con otra fórmula.

PREGUNTA PARA LA CLASE

Una señora dice que distingue si en la taza se sirvió primero la leche o primero el té. Se le dan 8 tazas, 4 de cada tipo, en orden al azar, y acierta todas. ¿Le creen?

2 / 27

NOTAS · 2

Es el ejemplo con el que Fisher abrió su libro sobre diseño de experimentos (1935). Si la señora adivina, hay 70 maneras de elegir cuáles 4 tazas son «leche primero», y solo una es la correcta: acertar todas tiene probabilidad 1/70 = 0,0143. Es tan improbable por casualidad que es razonable creerle.

Preguntar después: ¿y si hubiera acertado 6 de las 8 (3 de cada tipo)? Acertar 3 o más de las 4 por casualidad tiene probabilidad 17/70 = 24,3 %: no alcanza para creerle. Toda la clase es formalizar este razonamiento.



NOTAS · 3

La semana que viene se aplica la misma lógica para comparar dos grupos y para la tabla de contingencia.



SEMANA 10 · LA LÓGICA DE UNA PRUEBA

La primera prueba de hipótesis (1710)

John Arbuthnot contó los bautismos de Londres de 1629 a 1710: en los 82 años hubo más varones que mujeres.

82

años seguidos con más varones

1/2

la probabilidad de un año así, si los dos sexos fueran igual de probables

(1/2)^82

la de 82 años seguidos: un número con 24 ceros después de la coma

Fuente: Arbuthnot (1710), Philosophical Transactions of the Royal Society 27: 186-190.

5 / 27

NOTAS · 5

Arbuthnot concluyó que no podía ser casualidad. Su explicación fue teológica, pero el razonamiento es el de una prueba de hipótesis moderna: suponer que es azar, calcular qué tan improbable es lo observado, y rechazar el azar si lo es mucho.

Hoy se sabe que nacen unos 105 varones por cada 100 mujeres en casi todas las poblaciones.

SEMANA 10 · LA LÓGICA DE UNA PRUEBA

La dama que prueba el té

Si la señora adivina, cuántas de las 4 tazas con la leche primero identifica bien.

Aciertos	Maneras de elegir	Probabilidad si adivina
0	1	1/70 = 0,014
1	16	16/70 = 0,229
2	36	36/70 = 0,514
3	16	16/70 = 0,229
4	1	1/70 = 0,014
Total	70	1

Fuente: Fisher (1935), The Design of Experiments.

6 / 27

NOTAS · 6

Las maneras salen de combinar: para acertar k, elegir k de las 4 tazas correctas y 4 − k de las 4 incorrectas. Es la combinatoria de la semana 5.

Con esta tabla se decide de antemano: si acierta las 4 se le cree, porque por azar pasa menos de 2 veces en 100. Con 3 no alcanza.

SEMANA 10 · LA LÓGICA DE UNA PRUEBA

La lógica, en tres pasos

Suponer que es casualidad

Que la señora adivina, que los sexos son igual de probables: la hipótesis nula.

Calcular qué tan raro es

La probabilidad de un resultado como el observado, o más extremo, si fuera casualidad.

Decidir

Si es muy improbable, se rechaza la casualidad. Si no, no hay evidencia contra ella.

7 / 27

NOTAS · 7

Es una prueba por el absurdo, pero con probabilidades: no se llega a una contradicción, sino a algo muy improbable.

2

Hipótesis y errores

Las dos formas de equivocarse

8 / 27

SEMANA 10 · HIPÓTESIS Y ERRORES

Las dos hipótesis

Hipótesis nula, H_0

La de «no hay efecto»: igualdad, ausencia de diferencia, casualidad.

Es la que se pone a prueba.

Hipótesis alternativa, H_1

Lo que se quiere mostrar: que hay un efecto, una diferencia.

Se acepta solo si los datos la apoyan con fuerza.

Las dos hablan del parámetro de la población (μ, p), nunca de la muestra: la muestra ya se conoce.

9 / 27

NOTAS · 9

La asimetría importa: H_0 se mantiene hasta que haya evidencia en contra, como la inocencia en un juicio. Por eso el resultado es «rechazar H_0» o «no rechazar H_0», y nunca «aceptar H_0».

Las hipótesis se plantean antes de mirar los datos. Plantearlas después de verlos es hacer trampa sin querer.

SEMANA 10 · HIPÓTESIS Y ERRORES

Dos formas de equivocarse

	Se rechaza H_0	No se rechaza H_0
H_0 es cierta	Error de tipo I (probabilidad α)	Decisión correcta
H_0 es falsa	Decisión correcta (potencia, 1 − β)	Error de tipo II (probabilidad β)

Nunca se sabe en qué fila se está: si se supiera, no haría falta la prueba.

10 / 27

NOTAS · 10

Una forma de recordarlo con la alarma de humo de la cocina: el error de tipo I es que suene sin fuego (una falsa alarma); el de tipo II, que no suene cuando hay fuego.

SEMANA 10 · HIPÓTESIS Y ERRORES

α, β y la potencia

Nivel de significación, α

La probabilidad de error de tipo I que se acepta. Se fija **antes**: 0,05 es lo habitual.

Con la alarma: qué tan seguido se tolera que suene sin fuego.

β

La probabilidad de error de tipo II. Depende de cuánto se aleje la realidad de H_0 y del tamaño de muestra.

No se fija: se calcula para un valor concreto de H_1.

La potencia, 1 − β

La probabilidad de detectar un efecto que existe.

Crece con n y con el tamaño del efecto.

11 / 27

NOTAS · 11

Con la misma muestra, bajar α sube β: una alarma menos sensible suena menos sin fuego, pero también se le escapan más incendios. La única forma de bajar los dos es más datos.

El práctico tiene el cálculo de β y de la potencia, y el simulador de errores de tipo I y II muestra cómo se mueven.

3

El p-valor y la región crítica

Dos formas de decidir lo mismo

12 / 27

SEMANA 10 · EL P-VALOR Y LA REGIÓN CRÍTICA

La moneda de Kerrich

John Kerrich tiró una moneda 10.000 veces mientras estaba preso en un campo de internación durante la Segunda Guerra Mundial.

5.067

caras en 10.000 tiradas: $\hat{p}$ = 0,5067

0,005

el error estándar si la moneda es equilibrada: $\sqrt{(0{,}5 \cdot 0{,}5 / 10.000)}$

z = 1,34

(0,5067 − 0,5) / 0,005: a cuántos errores estándar de 0,5 quedó

Fuente: Kerrich (1946), An Experimental Introduction to the Theory of Probability.

13 / 27

NOTAS · 13

H_0: p = 0,5 (la moneda es equilibrada). H_1: p ≠ 0,5. Salieron 67 caras más de las esperadas: ¿es mucho o es azar? z responde en errores estándar.

El error estándar se calcula con el 0,5 de H_0, no con $\hat{p}$: la cuenta se hace suponiendo que H_0 es cierta.

SEMANA 10 · EL P-VALOR Y LA REGIÓN CRÍTICA

El p-valor

Si la moneda es equilibrada, ¿qué probabilidad hay de quedar tan lejos de 0,5 como Kerrich, o más? $P(|Z| \geq 1{,}34) = 2 \cdot 0{,}0901 = 0{,}1802$



Las dos colas, porque H_1 es p ≠ 0,5: un desvío hacia las cecas sería tan sospechoso como hacia las caras.

14 / 27

NOTAS · 14

El p-valor es 0,1802: si la moneda fuera equilibrada, en el 18 % de los experimentos como el de Kerrich se vería un desvío como ese o mayor. No es raro: no hay evidencia contra la moneda.

Cuanto más chico el p-valor, más difícil es explicar los datos con H_0. La regla: si p ≤ α, se rechaza H_0.

SEMANA 10 · EL P-VALOR Y LA REGIÓN CRÍTICA

La región crítica

Con α = 0,05 y una prueba bilateral, se rechaza H_0 si z cae más allá de ±1,96. El z de Kerrich, 1,34, queda adentro.



Es la otra forma de decidir: comparar el estadístico con el valor crítico, en lugar del p-valor con α. Llevan siempre a la misma decisión.

15 / 27

NOTAS · 15

z = 1,34 queda adentro: no se rechaza H_0. Equivale a decir que el p-valor, 0,1802, es mayor que 0,05.

El p-valor dice más que la región crítica: no solo si se rechaza, sino con cuánto margen. Por eso hoy se informa el p-valor.

SEMANA 10 · EL P-VALOR Y LA REGIÓN CRÍTICA

Bilateral o unilateral

Bilateral: $H_1: p \neq 0{,}5$

- Interesa un desvío en cualquier dirección.
- Región crítica: $|z| > 1{,}96$.
- Kerrich: p-valor = 0,1802.

Unilateral: $H_1: p > 0{,}5$

- Solo interesa una dirección, y se decide **antes** de ver los datos.
- Región crítica: $z > 1{,}645$.
- Kerrich: p-valor = 0,0901.

La unilateral tiene más potencia en su dirección y ninguna en la otra.

16 / 27

NOTAS · 16

Elegir la unilateral después de ver hacia dónde salieron los datos es hacer trampa: duplica en los hechos la probabilidad de error de tipo I.

Las unilaterales tienen sentido cuando solo una dirección importa: un control de calidad que solo se preocupa si el contenido es menor que lo declarado.

PREGUNTA PARA LA CLASE

Con la moneda de Kerrich no se rechazó Ho. ¿Quedó demostrado que la moneda es equilibrada?

17 / 27

NOTAS · 17

No. No rechazar H_0 quiere decir que los datos son compatibles con ella, no que sea cierta. El intervalo del 95 % para p va de 0,497 a 0,517: los datos también son compatibles con una moneda que salga cara el 51 % de las veces.

«La ausencia de evidencia no es evidencia de ausencia.» Por eso no se dice «se acepta H_0».

4

Pruebas para una proporción y una media

Las dos pruebas de hoy

18 / 27

SEMANA 10 · PRUEBAS PARA UNA PROPORCIÓN Y UNA MEDIA

La prueba para una proporción

Hipótesis

H_0: $p = p_0$, contra $p \neq p_0$, $p > p_0$ o $p < p_0$.

Estadístico

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

Con p_0: la cuenta supone H_0 cierta.

Condición

$np_0 \geq 5$ y $n(1 - p_0) \geq 5$.

Para usar la normal.

19 / 27

NOTAS · 19

Es lo que se hizo con Kerrich, con $p_0 = 0{,}5$. A diferencia del intervalo de confianza, que usa $\hat{p}$ en el error estándar, la prueba usa p_0.

SEMANA 10 · PRUEBAS PARA UNA PROPORCIÓN Y UNA MEDIA

La prueba t para una media

Los datos

La muestra de 30 hogares de la semana 9: $\bar{x} = 1{,}400$, $s = 1{,}003$, EE = 0,183.

Las hipótesis

Un informe anterior decía 1,8 ocupados por hogar. H_0: $\mu = 1{,}8$; H_1: $\mu \neq 1{,}8$.

El estadístico

$t = (\bar{x} - \mu_0) / EE = (1{,}400 - 1{,}8) / 0{,}183 =$ **−2,19**

Con 29 grados de libertad.

20 / 27

NOTAS · 20

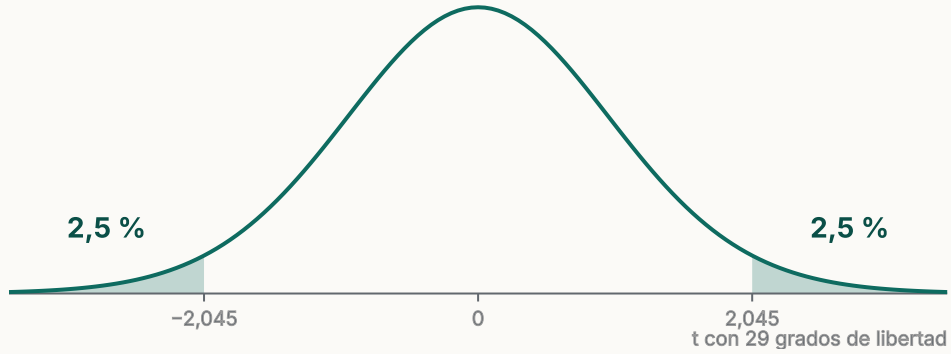
El informe de 1,8 es un supuesto para el ejemplo. Lo que no es supuesto es la muestra: es la misma con la que se armó el intervalo de la semana pasada.

Las condiciones son las del intervalo t: muestra aleatoria y datos normales o n razonable.

SEMANA 10 · PRUEBAS PARA UNA PROPORCIÓN Y UNA MEDIA

La decisión

Valor crítico de la t con 29 grados de libertad: ±2,045. El t de la muestra, −2,19, cae en la región crítica.



Con la calculadora, p-valor = 0,0371: menor que 0,05. Se rechaza H_0.

21 / 27

NOTAS · 21

Conclusión para un informe: «la cantidad media de ocupados por hogar es distinta de 1,8 (t = −2,19, 29 gl, p = 0,037); la muestra sugiere que es menor».

Con la tabla de la t solo se puede acotar el p-valor: está entre 0,02 y 0,05, porque 2,19 queda entre los valores críticos 2,045 y 2,462.

SEMANA 10 · PRUEBAS PARA UNA PROPORCIÓN Y UNA MEDIA

La prueba y el intervalo dicen lo mismo

El intervalo del 95 % de la semana 9: de 1,026 a 1,774 ocupados por hogar.

Ho: μ = 1,8

1,8 queda afuera del intervalo: se rechaza. t = −2,19, p = 0,037.

Decisión correcta: μ no es 1,8.

Ho: μ = 1,6

1,6 queda adentro: no se rechaza. t = −1,09, p = 0,283.

Error de tipo II: μ = 1,471, no 1,6.

22 / 27

NOTAS · 22

Una prueba bilateral con α = 0,05 rechaza exactamente los valores de μ_0 que quedan afuera del intervalo del 95 %. El intervalo es la lista de todos los valores que la prueba no rechazaría.

Como en esta población se conoce μ = 1,471, se ve algo que en la realidad nunca se ve: la segunda prueba se equivoca. Con 30 hogares, la diferencia entre 1,471 y 1,6 es demasiado chica para detectarla: falta potencia.

5

Leer bien un resultado

Significativo no quiere decir importante

23 / 27

SEMANA 10 · LEER BIEN UN RESULTADO

Los pasos de una prueba

- Plantear H_0 y H_1, y elegir α, **antes** de ver los datos.
- Verificar las condiciones: muestra aleatoria, tamaño, forma de los datos.
- Calcular el estadístico de prueba.
- Calcular el p-valor, o comparar con el valor crítico.
- Decidir y redactar la conclusión **en el contexto del problema**, con el estadístico y el p-valor.

24 / 27

NOTAS · 24

El último paso es el que más se olvida. «Se rechaza H_0» no le dice nada a quien leyó el informe; «los hogares tienen en promedio menos de 1,8 ocupados» sí.

SEMANA 10 · LEER BIEN UN RESULTADO

Significativo no es lo mismo que importante

Una moneda que sale cara el 50,1 % de las veces, tirada un millón de veces.

$z = 2{,}00$

$(0{,}501 - 0{,}5) / \sqrt{(0{,}25 / 1.000.000)}$

$p = 0{,}0456$

«estadísticamente significativo» con $\alpha = 0{,}05$

Con muestras enormes, cualquier diferencia, por mínima que sea, da un p-valor chico.

25 / 27

NOTAS · 25

Para cualquier uso práctico esa moneda es equilibrada. «Significativo» solo quiere decir «difícil de explicar por azar»; si la diferencia importa lo decide el tamaño del efecto, que se ve en el intervalo de confianza.

Lo contrario también pasa: con muestras chicas, un efecto grande puede dar $p > 0{,}05$. Por eso conviene informar siempre el intervalo junto con el p-valor.

SEMANA 10 · LEER BIEN UN RESULTADO

Qué es y qué no es el p-valor

Es

- La probabilidad de datos como los observados, o más extremos, **si H_0 fuera cierta**.
- Una medida de qué tan compatibles son los datos con H_0.

No es

- La probabilidad de que H_0 sea cierta.
- La probabilidad de que el resultado sea casualidad.
- Una medida del tamaño o la importancia del efecto.

26 / 27

NOTAS · 26

La primera confusión es la más extendida, también entre investigadores: el p-valor se calcula suponiendo H_0, así que no puede dar la probabilidad de H_0. Es el mismo error de dar vuelta una condicional de la semana 5: $P(\text{datos} \mid H_0)$ no es $P(H_0 \mid \text{datos})$.

El práctico tiene un ejercicio para discutir frases sobre el p-valor.

SEMANA 10 · PRUEBAS DE HIPÓTESIS: LA LÓGICA

Para llevarse

- Una prueba supone H_0 y pregunta qué tan raros serían los datos si fuera cierta.
- Hay dos errores: rechazar una H_0 cierta (α) y no rechazar una falsa (β).
- El p-valor es la probabilidad de un resultado así o más extremo si H_0 fuera cierta; se rechaza H_0 si $p \leq \alpha$.
- No rechazar H_0 no la demuestra; rechazarla no dice que el efecto sea importante.
- La prueba bilateral y el intervalo de confianza llevan a la misma decisión.

Guía de trabajo práctico 10
estadistica.ar/docentes/trabajos-practicos/semana-10

Simulador del p-valor
estadistica.ar/herramientas/sim-p-valor

Simulador de errores de tipo I y II
estadistica.ar/herramientas/sim-errores-i-ii

Calculadora de prueba t
estadistica.ar/herramientas/calc-prueba-t

27 / 27

NOTAS · 27

Para el práctico: las tablas de la normal y de la t, y calculadora.