UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

Capítulo 9:

Prueba de significación y contraste de hipótesis

Erik Cobo, Jordi Cortés y José Antonio González

Laura Riba, Rosario Peláez, Marta Vilaró y Nerea Bielsa

Septiembre 2014

Departament d'Estadística i Investigació Operativa
UNIVERSITAT POLITÈCNICA DE CATALUNYA

equator network

UPC

ELSEVIER DOYMA

MEDICINA CLINICA

TRIALS

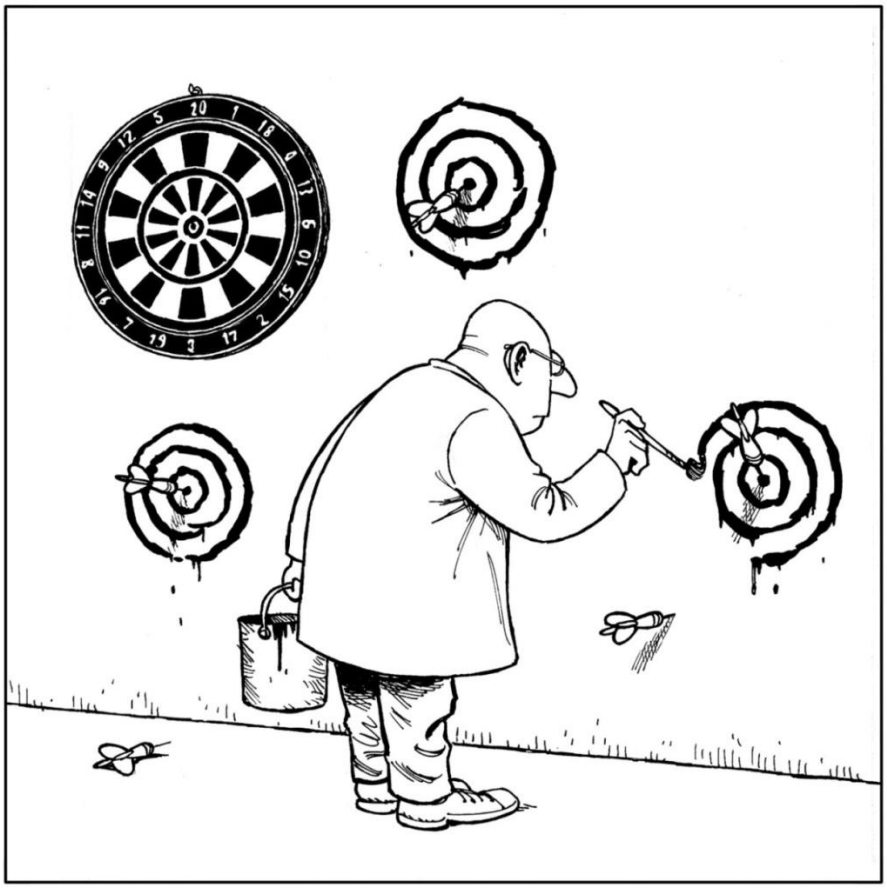
Prueba de significación y contraste de hipótesis

* Indica tema más avanzado que no es crucial para los ejercicios, aunque el lector debe recordar que aquí lo tiene —cuando lo necesite.

Presentación

Este capítulo formaliza la respuesta a dos preguntas diferentes pero relacionadas: "¿Qué sé?" (inferencia) y "¿Qué hago?" (decisión). Se define la prueba y el nivel p de significación en el entorno de la evidencia empírica o inferencia sobre conocimiento. Por su parte, los riesgos α y β y el contraste de hipótesis se enmarcan en la decisión entre dos acciones alternativas Finalmente distingue entre pruebas de diferencias y de equivalencia.

Al terminar este capítulo, el lector debe retener especialmente (1) la importancia de que las hipótesis sean independientes de los datos en que se contrastan; (2) que las reglas lógicas que gobiernan la adquisición de conocimiento y las que determinan la acción no son equivalentes; y (3) el papel del IC en las revistas científicas, y el del contraste de hipótesis en las agencias de decisión.



Contribuciones: (1) la versión original de 2013 descansa en el libro de Bioestadística para No estadísticos de Elsevier de EC, JAG y PM y en el material de la asignatura de PE de la FIB (UPC); fue editada por JC y revisada por RP y MV; (2) la de febrero de 2014 fue revisada por LR, JC, EC y MV para incorporar mejoras y sugerencias anónimas; y (3) la de septiembre de 2104 por NB y EC.

1. Objetivos de la inferencia estadística

A la vista de la información aportada por la muestra, las principales preguntas de la inferencia estadística son: (1) ¿qué valores del parámetro son creíbles?; (2) ¿se puede negar cierto valor del parámetro? Y, (3) a partir de ahora, ¿qué hago? La primera, mediante intervalos de confianza, se resolvió en el tema anterior; las 2 últimas se exponen en éste.

La pregunta que responden los intervalos de confianza (¿qué valores son creíbles?) engloba, de alguna manera, a la pregunta de la prueba de significación (¿se puede negar cierto valor?). Los intervalos de confianza aportan más información y son más fáciles de entender, asimilar y explicar. ¿Qué interés ofrece, entonces, poner a prueba una hipótesis? Pues quizás, que puede ser la auténtica pregunta de interés.

> **Ejemplo 1.1**: saber si un fármaco es más eficaz que otro puede reducirse a conocer si la diferencia de sus medias en la respuesta de interés es o no es exactamente el valor 0. Por tanto, poder negar el valor 0, implica haber demostrado que un producto es más eficaz que otro.

La pregunta sobre una hipótesis la aborda la inferencia estadística ("¿qué sé?") en la Prueba de Significación, PS, o valor de p.

La pregunta sobre la acción futura (¿qué hago?) la aborda la decisión estadística en el contraste de hipótesis, CH, acotando los riesgos alfa y beta de emprender acciones erróneas (tipo I y II).

2. Prueba de significación, PS

Se desea *poner a prueba* una hipótesis previa H confrontándola con los datos.

> **Ejemplo 2.1:** desde hace un tiempo, un residente se juega a cara y cruz las guardias que coinciden con las fiestas familiares. Su compañero lanza su moneda y... ¡siempre gana! Un día, el primero decide estudiar formalmente si la moneda está apañada. Así, el problema consiste en analizar si podemos descartarla hipótesis:
>
> H: $\pi = 0.5$ (moneda correcta)
>
> Donde π representa la probabilidad de cara que se desea negar.
>
> Lanzar n=100 veces la moneda y observar la proporción P de caras, proporcionará "*evidencia*" empírica. Suponga que observa P=0.63. Este resultado invita a creer que la moneda está "cargada": que no es cierto que π=0.5. En cambio, si el resultado fuera P=0.52,

se consideraría "compatible" con que la moneda no esté cargada. Cuanto más se aleje P de 0.5, más información en contra de H.

Hay que considerar la aleatoriedad del proceso. Es posible que una moneda perfecta, no cargada, genere una observación de 63 caras en 100 lanzamientos. Y, de forma recíproca, también es posible que una moneda con probabilidad de cara de 0.6 genere una muestra con un 50% de caras.

Nota: Se podría abordar el problema desde un punto de vista físico y, dando por bueno ('premisas') el conocimiento actual de esta ciencia, estudiar la composición de la moneda, su centro de gravedad, su circunferencia,... Ahora bien, sea cual sea su respuesta, siempre conviene estudiar qué dicen las observaciones, no sea que convenga revisar el modelo teórico.

La hipótesis H establece una condición sobre el parámetro poblacional que se desea negar . Esta información se "condensa" en un estadístico apropiado, que fluctúa aleatoriamente. Cuando H es correcta, la distribución es conocida, y el estadístico se localizará de forma previsible en una zona determinada por H. Y cuanto más lejos se aleje el estadístico de dicha zona, más credibilidad gana la posibilidad de que proceda de otra distribución con un parámetro distinto al de H.

Ejercicio 2.1

En una prueba de significación (elija una):

a) Se desea conocer el valor de cierto parámetro

b) Se construye una hipótesis sugerida por los datos

c) Se busca "evidencia" (pruebas) a favor de la hipótesis H que se desea demostrar que es cierta

d) Ninguna de las anteriores es correcta

Ejercicio 2.2

Vd. desea aportar evidencia de que un nuevo tratamiento es mejor que uno clásico. Escriba la hipótesis H:

a) H: el nuevo tratamiento no es mejor que el clásico

b) H: el nuevo tratamiento es mejor que el clásico

c) H: el rendimiento del nuevo tratamiento supera al clásico

d) Ninguna de las anteriores es correcta

Necesitamos un proceso que (1) permita "incorporar" la información muestral o "evidencia" empírica; y que (2) sea transparente, en el sentido de ser reproducible por otros investigadores.

Lectura: Nature, Lancet y BMJ han lamentado en 2014 la falta de reproducibilidad de los resultados de investigación y, por tanto, el despilfarro de recursos que implica

2.1. Valor p

Este método calcula el valor p (*p value)* o probabilidad de que se presente un valor del estadístico más alejado de H que el observado. Cuanto más pequeño es p, menos verosímil es H.

Recuerde

La verosimilitud de H disminuye si el valor p es pequeño.

El valor p (Figura 2.1) puede interpretarse como "cuán inverosímil es el resultado observado si H fuera cierta" o "hasta qué punto resultados como el observado (o más extremos) son probabilísticamente compatibles con H". Lo que suele interpretarse como que hay "suficiente evidencia o pruebas en contra de H" para negarla, lo que suele resumirse con un "el resultado es estadísticamente significativo".

Recuerde

Si p es muy pequeño, hay evidencias "estadísticamente significativas" en contra de H.

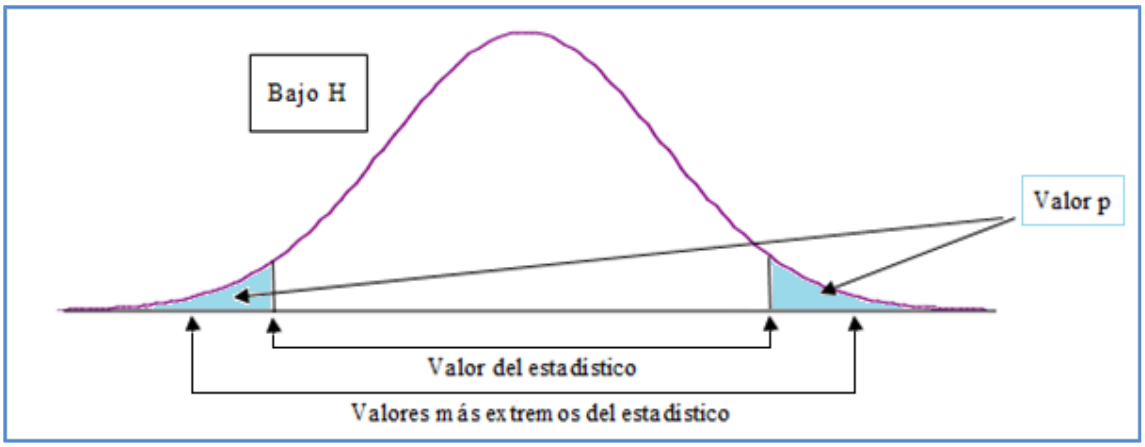


Figura 2.1 Distribución del estadístico si H es cierta. El valor p indica la probabilidad de observar valores del estadístico igual o más extremos que el observado, en el caso de que H sea cierta.

En función de si situamos los "valores más extremos" en 1 lado o en 2 hablaremos de pruebas uni o bilaterales. En las primeras, el valor de p es la probabilidad de obtener un valor o bien mayor, o bien menor, (dependiendo del problema) que el estadístico observado (probabilidad de una cola). En las

pruebas bilaterales, el valor de p es la probabilidad de obtener un valor más extremo del estadístico (se suman las probabilidades de ambas colas). Profundizaremos en este asunto en el punto 0.

2.2. Mecánica de la prueba de significación

La prueba de significación se basa en el siguiente proceso formal:

1) Antes de los datos
   a) Escoger una variable (*response, outcome, endpoint*) que valore el objetivo del estudio
   b) Fijar un diseño de recogida de datos y un estadístico que resuma los resultados de la variable
   c) Definir la hipótesis H que se desea rechazar
   d) Describir la distribución del estadístico bajo H y las premisas necesarias, escribiendo el plan de análisis estadístico.
   e) Acotar el valor de p que llevaría a rechazar H, usualmente p=0.05
2) Recoger, con calidad, los datos (realizar o el experimento o la observación "natural")
3) Una vez "cerrada" la base de datos:
   a) Calcular el valor p.
   b) Detallar el $IC_{95\%}$.

Nota: Se habla despectivamente de "p huérfana" cuando p no se acompaña de medidas del efecto y de su incertidumbre.

Recuerde

Primero el diseño (con la hipótesis y la variable); luego los datos; y al final la p con un $IC_{95\%}$.

A continuación exponemos este proceso para el caso de una probabilidad.

2.3. Prueba de significación de una probabilidad

Vamos a usar la distribución de la proporción P observada en una muestra para poner a prueba una hipótesis H sobre una probabilidad poblacional π.

Nota: Recuerde que P ~ N(π, π(1-π)/n). Note que, a diferencia de IC, ahora π viene dada por H.

Ejemplo 2.1 (cont): En el ejemplo anterior de la moneda, con n=100,

Variable: resultado cara o cruz

Estadístico: proporción **P** de caras

Hipótesis H: π = 0.5 (moneda correcta)

Si H es cierta: $P \sim N(\pi, \frac{\pi\ 1-\pi}{n}) = N(0.5, 0.05^2)$

Premisas: muestra grande $\pi \cdot n > 5$ y $(1-\pi) \cdot n > 5$

Límite de p=0.05

Caso a)

Con n=100 se observan 63 caras:

La proporción observada es: $P = \frac{63}{100} = 0.63 = 63\%$

El estadístico señal/ruido: $z = \frac{\pi_P - \pi}{\frac{\pi\ 1-\pi}{n}} = \frac{0.63-0.5}{\frac{0.5 \cdot 0.5}{100}} = \frac{0.13}{0.05} = \mathbf{2.6}$

```
# Cálculo del p-valor
# Unilateral:P(Z>2.6) cola sup.-> lower.tail=FALSE
> pnorm(q=2.6,lower.tail=FALSE)
[1] 0.004661188
# Bilateral: Multiplicando por 2 colas
> 2*pnorm(q=2.6,lower.tail=FALSE)
[1] 0.009322376
```

Por tanto, como el p valor (o probabilidad de observar un valor de **P** tan o más alejado de H) es p<0.01, se considera H poco verosímil y se rechaza que π valga 0.5 (Figura 2.2).

I) El intervalo de confianza es:

$$IC_{95\%}\ \pi\ :\ P \pm z_{\frac{\alpha}{2}} \cdot \sigma_P \approx 0.63 \pm 1.96 \cdot 0.05 \approx 0.63 \pm 0.10 = 0.53, 0.73$$

Creemos que la "auténtica" proporción de cara π se encuentra entre 53% y 73%.

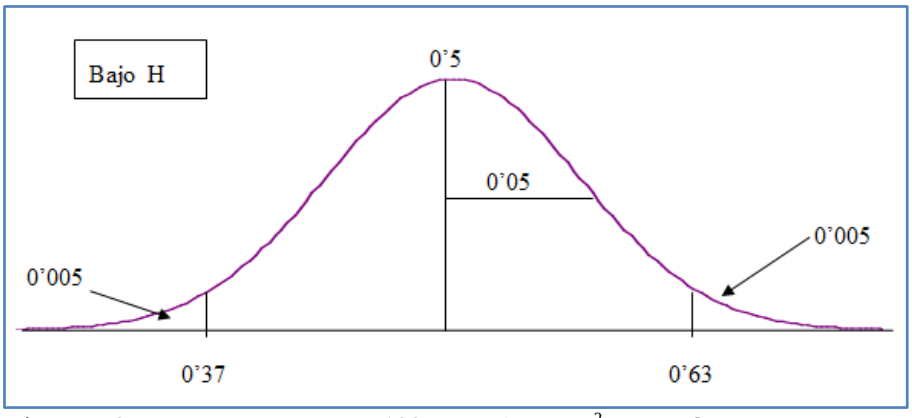


Figura 2.2 Bajo H: π=0.5 y con una muestra n=100, P ~ N(0.5,0.05^2). Si se observan 63 caras, P=0.63. Como P(P>0.63)≈ 0.005≈ P(P<0.37), el nivel de significación es p=2*0.005=0.01.

Caso b)

Con n=100 se observan 52 caras:

$$\mathbf{P} = \frac{52}{100} = 0.52 = 52\%$$

$$z = \frac{\pi_P - \pi}{\sqrt{\frac{\pi\ 1-\pi}{n}}} = \frac{0.52 - 0.5}{\sqrt{\frac{0.5 \cdot 0.5}{100}}} = \frac{0.02}{0.05} = \mathbf{0.4}$$

```
# Cálculo del p-valor
# P(Z>0.4) (cola superior, lower.tail=FALSE)
> pnorm(q=0.4,lower.tail=FALSE)
[1] 0.3445783
# Multiplicando por 2 (valor bilateral)
> 2*pnorm(q=0.4,lower.tail=FALSE)
[1] 0.6891565
```

Por tanto, valor p $= P(\mathbf{P} > 0.52) + P(\mathbf{P} < 0.48) \approx 0.69$.

Como p= 0.69 no es "pequeño", nada se opone a aceptar H (véase Figura 2.3).

VI) El intervalo de confianza es:

$$IC_{95\%}\ \pi : \mathrm{P} \pm z_{\frac{\alpha}{2}} \cdot \sigma_P \approx 0.52 \pm 1.96 \cdot 0.05 \approx 0.52 \pm 0.10 = 0.42, 0.62$$

Creemos que la "auténtica" proporción de cara π se encuentra entre 42% y 62%.

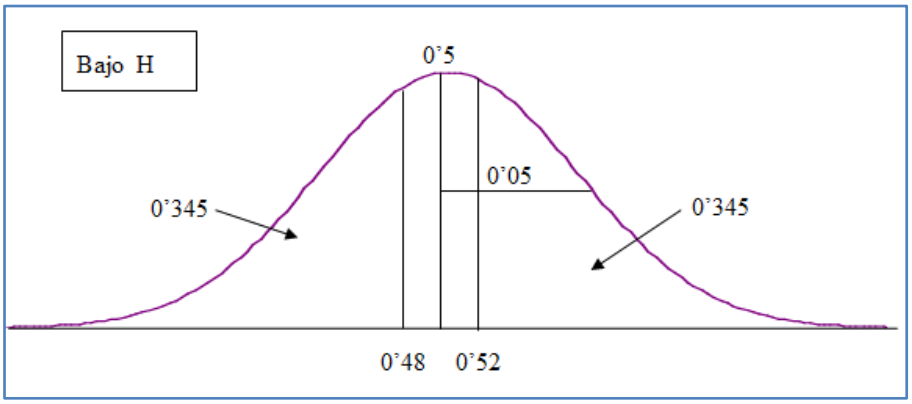


Figura 2.3 Si se observan 52 caras, P=0.52 y la probabilidad de observar 52 o más caras es de 0.345, que junto a su simétrica (observar 48 o menos caras) hace p=0.690.

A continuación se muestra cómo realizar esta prueba directamente con R:

```
Ejemplo 2.1. en R
Caso a): con n=100 se observan 63 caras:
> prop.test(x=63, n=100, p=0.5, conf.level=0.95, correct=FALSE)
1-sample proportions test without continuity correction
data:  63 out of 100, null probability 0.5
X-squared = 6.76, df = 1, p-value = 0.009322
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
  0.5322053 0.7181764
```

```
sample estimates:
  p
0.63
Caso b): con n=100 se observan 52 caras:
> prop.test(x=52, n=100, p=0.5, conf.level=0.95, correct=FALSE)
1-sample proportions test without continuity correction
data:  52 out of 100, null probability 0.5
X-squared = 0.16, df = 1, p-value = 0.6892
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
  0.4231658 0.6153545
sample estimates:
 p
0.52
```

Nota: El cálculo del p valor con R y según el método explciado cambia la distribución de referencia pero coinciden algebraicamente (la Ji cuadrado de 1 GdL es el cuadrado de una N(0,1)). En cambio los intervalos de confianza de R con la función *prop.test* usa el Wilson score method, que funciona bien incluso para tamaños de pocas decenas.

Ejercicio 2.3

Una serie de 400 pacientes con SIDA han recibido, en diferentes etapas de su seguimiento, dos tratamientos diferentes, A y B que son, a priori, potencialmente similares. Ahora, se les ha preguntado cuál prefieren y un 58% han optado por el A. Se desea saber si puede rechazarse que sean igualmente preferidos. [Escriba todos los pasos del proceso.]

Ejercicio 2.4

Repita el Ejercicio 2.3 asumiendo que las preferencias por A han sido 53%.

Muchos paquetes informáticos de estadística, al redondear, dan un nivel de significación con muchos ceros (por ejemplo, p=0.00000) que parecería indicar un resultado imposible, de probabilidad nula. Como ello no es así, cambie el último 0 por 1; y el '=' por '<': p<0.001.

Recuerde

No escriba p=0.000. En su lugar, ponga p<0.001.

Ejercicio 2.5

Los usuarios de una biblioteca llevan años protestando por las prestaciones del sistema de búsqueda instalado para realizar sus consultas. Los responsables de la biblioteca deciden valorar la posibilidad de cambiar el sistema. Durante el periodo de prueba, han realizado un experimento comparando ambos sistemas mediante una escala que mide la satisfacción de los usuarios.

Hacen la prueba anterior de preferencias y resumen sus resultados con la siguiente frase: *el nuevo sistema genera mayor satisfacción en los usuarios ($p<0.01$).* ¿Cuál o cuáles de las siguientes son ciertas?:

a) Se rechaza la hipótesis H de que la satisfacción sea igual en ambos grupos.
b) Suponiendo que ambos sistemas generen la misma satisfacción, la probabilidad de haber obtenido un resultado tan o más extremo que el observado es menor del 1%.
c) Creemos que el resultado observado refleja una diferencia poblacional del nivel de preferencias.
d) La proporción de casos más satisfechos con el sistema antiguo que con el nuevo es menor del 1%.
e) Cuando se dice que el nuevo sistema es mejor se tiene una probabilidad de error menor de 0.01.
f) La probabilidad de que el nuevo sistema sea mejor es 0.01.

2.4. Prueba de significación de una media ($\mu=\mu_H$)

La aplicación del mecanismo anterior a una variable continua en la que se desea contrastar una hipótesis sobre su media es muy similar.

Ejemplo 2.2: Se quiere *'testar'* en la respuesta Y si su media μ se corresponde con una cierta media μ_H especificada en la hipótesis H.

Para escribir que la media μ de la población origen de la muestra es una media μ_H pre-especificada, escribimos: $H: \mu = \mu_H$

Si, como es usual, la varianza poblacional σ^2 es desconocida, se recurre a su estimador muestral S^2 y a la distribución t de Student. Por lo tanto, bajo H:

$$t = \frac{y - \mu_H}{\frac{S}{\sqrt{n}}} \sim t_{n-1}$$

Y puede calcularse el nivel de significación p como:

$$\text{p} = P\ \ t_{n-1} > t$$

Ejemplo 2.3: ¿Recuerda el ejemplo para demostrar que las gasolineras estaban poniendo menos gasolina de la que cobraban? Se resolvió con un IC, pero ¿se puede demostrar que timan? En una muestra aleatoria de 100 servicios, con S=10cc, se debe tomar una decisión sobre si $\mu = 1000$, habiendo observado una media $y = 997$ cc.

Variable: contenido real en servicios de 1000cc

Estadístico: media y

H: μ_H = 1000cc

Regla para el rechazo de H: si p<0.05

Se usará el estadístico señal/ruido

$$t = \frac{y - \mu_H}{\frac{S}{\sqrt{n}}}$$

que bajo H tiene una distribución t-Student: $\boldsymbol{t} \sim t_{n-1=100-1=99}$ si la variable es normal (premisa).

Cálculo del estadístico:

$$\boldsymbol{t} = \frac{y - \mu_H}{\frac{S}{\sqrt{n}}} = \frac{997 - 1000}{\frac{10}{\sqrt{100}}} = -3$$

```
# Cálculo del p-valor
# Prob(t<-3) (cola inferior, lower.tail=TRUE)
> pt(q=-3, df=99, lower.tail=TRUE)
[1] 0.001707754
# Multiplicando por 2 (valor bilateral)
> 2*pt(q=-3, df=99, lower.tail=TRUE)
[1] 0.003415508
```

Decisión: como p=0.0034 < 0.05, rechazamos μ=1000cc con p=0.0034.

Conclusión práctica: Rechazamos que se esté dispensando la cantidad especificada.

I) Cálculo del intervalo de confianza:

$$IC_{95\%}\ \mu = y \pm t_{n-1;\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{n}} = 997 \pm 1.984 \cdot \frac{10}{\sqrt{100}} \approx 997 \pm 2 = 995,999$$

```
# t99,0.025
> qt(p=0.025,df=99,lower.tail=FALSE)
[1] 1.984217
```

La "auténtica" media μ de cantidad servida se encuentra entre 995 y 999 cc. Nos están timando, aunque a nivel individual, la cantidad es pequeña. La pequeña amplitud del $IC_{95\%}$ muestra que se dispone de mucha información.

Esta prueba completa con R es:

```
Ejemplo 2.3 en R
> install.packages('BSDA')
> library(BSDA)
> tsum.test(mean.x=997, s.x=10, n.x=100, mu=1000)
One-sample t-Test
data:  Summarized x
t = -3, df = 99, p-value = 0.003416
alternative hypothesis: true mean is not equal to 1000
95 percent confidence interval:
  995.0158 998.9842
sample estimates:
  mean of x
997
```

Ejemplo 2.4: En 9 voluntarios sanos se ha estudiado la diferencia D entre los tiempos de respuesta a un estímulo visual y auditivo, habiéndose observado, $\boldsymbol{d}$ = 6.71 y S=6.0. Asumiendo que D~N, ¿se puede aceptar que E(D)=μ=0, lo que implica que la respuesta a ambos estímulos es idéntica?

Solución:

Variable: diferencia entre el tiempo de respuesta a los estímulos visual y auditivo

Estadístico: media de las diferencias o d

Hipótesis que se quiere rechazar: H: E D $= \mu_H = 0$

Límite de p=0.05

Estadístico referencia: $\boldsymbol{t} = \frac{d-\mu_H}{S\ \overline{n}}$

Que bajo H se distribuye como: t ~$t_{n-1} = t_8$, si D normal (premisa).

Cálculo de p:

$$t = \frac{d - \mu_H}{\frac{S}{\sqrt{n}}} = \frac{6.71 - 0}{\frac{6}{\sqrt{9}}} = 3.355$$

```
# P = Prob [ (|t| > |3.355|)
> pt(q=3.355,df=8,lower.tail=FALSE)*2
[1] 0.01000575
```

Como p=0.01; H: $\mu_H = 0$ es poco verosímil. Conclusión práctica: ambos estímulos no tienen la misma respuesta (media).

I) Cálculo del intervalo de confianza:

$$IC_{95\%}\ D = d \pm t_{n-1;\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{n}} = 6.71 \pm 2.306 \cdot \frac{6}{\sqrt{9}} = 6.71 \pm 4.612 \approx 2.10, 11.32$$

```
# t8,0.025
> qt(p=0.025,df=8,lower.tail=FALSE)
[1] 2.306004
```

La "auténtica" diferencia entre la respuesta media a ambos estímulos se encuentra entre 2.10 y 11.32.

Prueba completa con R:

```
Ejemplo 2.4 en R
> install.packages('BSDA')
> library(BSDA)
> tsum.test(mean.x=6.71, s.x=6, n.x=9, mu=0)
One-sample t-Test
data:  Summarized x
t = 3.355, df = 8, p-value = 0.01001
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
  2.097992 11.322008  …
```

Ejercicio 2.6

La satisfacción de los usuarios se mide por una escala entre 0 y 100 con una distribución que se asemeja razonablemente a la Normal. El objetivo de un servicio sanitario es conseguir satisfacciones por encima de 70. En una muestra al azar de 16 usuarios se ha observado una media de 79 y una desviación típica de 12. ¿Se puede afirmar que la media poblacional está por encima de 70?

2.5. El estadístico "t" como cociente señal/ruido

El numerador de t representa la distancia entre el valor de la muestra y y el parámetro μ de la población. Y el denominador informa del error típico de y, ya que como μ es un parámetro de la población (forma parte de la pregunta), no tiene error aleatorio de muestreo.

Ejemplo 2.3 (cont.): En el ejemplo sobre el control de calidad en las gasolineras, si desea saber si el surtidor cumple con las especificaciones (μ), este numerador representa la señal que proporciona la muestra: cuánto se distancia de la media especificada en la hipótesis. Se ha observado un valor de -3. Por otro lado, la oscilación de $\boldsymbol{y}$ explicable por el muestreo aleatorio puede cuantificarse en $\boldsymbol{S_y} = \frac{s}{\sqrt{n}} = 1$. Y por tanto el cociente "señal/ruido" vale -3, indicando que la señal observada es negativa y 3 veces superior al error aleatorio.

Recuerde

Interprete el estadístico t como un cociente señal/ruido.

2.6. Prueba de significación de la comparación de dos medias

Para realizar una comparación de 2 medias, el estadístico a utilizar es:

Fórmula

El estadístico para comparar 2 medias es:

$$t = \frac{y_1 - y_2}{S \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Donde S es la desviación ponderada de las 2 muestras, *pooled*, raíz de:

$$S^2 = \frac{n_1 - 1 \cdot S_1^2 + n_2 - 1 \cdot S_2^2}{n_1 + n_2 - 2}$$

Ejemplo 2.5: Se realiza un Ensayo Clínico el que se quiere valorar la eficacia de un nuevo fármaco antidiabético. Para ello se asignan 18 pacientes al azar, con razón "1 a 2" a dos grupos: el de referencia, que recibirá el fármaco habitual, y el de la intervención, que recibirá el nuevo fármaco. A continuación se muestra la reducción en el nivel de glucosa (mg/dL) respecto el nivel inicial para cada individuo de cada uno de los grupos a los 3 meses del inicio del tratamiento:

G1 = grupo referencia: 13, 14, 10, 11, 14, 11 (mg/dL)

G2 = grupo experimental: 16, 11, 13, 12, 14, 12, 13, 13, 13, 12, 14, 15 (mg/dL)

La hipótesis nula es que no hay diferencias entre ambos fármacos en la reducción del nivel de glucosa.

Solución:

Variable: reducción de glicemia

Estadístico: diferencia de medias

Hipótesis que se quiere rechazar: H: $\mu_{G1} = \mu_{G2}$

Límite p=0.05

Estadístico de referencia $t = \frac{y_1 - y_2}{s \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$

Cuya distribución bajo H es: $t \sim t_{n_1+n_2-2} = t_{16}$

Premisas: las dos muestras provienen de una distribución normal, y sus varianzas son iguales.

Cálculos:

$$S = \sqrt{\frac{n_1 - 1 \cdot S_1^2 + n_2 - 1 \cdot S_2^2}{n_1 + n_2 - 2}} = \sqrt{\frac{6 - 1 \cdot 2.97 + 12 - 1 \cdot 1.97}{6 + 12 - 2}} = \sqrt{\frac{36.5}{16}} = 1.51$$

$$t = \frac{y_1 - y_2}{S \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{12.17 - 13.17}{1.51 \cdot \sqrt{\frac{1}{6} + \frac{1}{12}}} = -1.32$$

```
# P = Prob [(|t| < |-1.32|)]
> pt(q=1.32,df=16,lower.tail=FALSE)*2
[1] 0.2054096
```

Como p=0.20, no hay evidencia para rechazar H. No podemos afirmar que los fármacos sean diferentes en eficacia.

Nota: No hemos demostrado que tengan igual eficacia. Tan sólo no hemos logrado demostrar que sean diferentes. Tampoco hemos establecido que ambos sean eficaces: falta ver (1) si la reducción desde basal es significativa; y (2) cuál hubiera sido la evolución de otro grupo de referencia no tratado (que incluiría, entre otros, una posible calibración desigual de los aparatos).

I) Cálculo del intervalo de confianza:

$$IC_{95\%}\ \mu_1 - \mu_2 = y_1 - y_2 \pm t_{0.975,16} \cdot \frac{S}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = 12.17 - 13.17 \pm 2.12 \cdot \frac{1.51}{\sqrt{\frac{1}{6} + \frac{1}{12}}}$$

$$IC_{95\%}\ \mu_1 - \mu_2 = -1 \pm 1.60 = \mathbf{-2.60, 0.60}$$

PS con R:Ejemplo 2.5 en R

```
> g1 = c(13, 14, 10, 11, 14, 11) # grupo referencia
> g2 = c(16, 11, 13, 12, 14, 12, 13, 13, 13, 12, 14, 15) # grupo experimental
> t.test(g1, g2, alt="two.sided", var.equal=TRUE)

   Two Sample t-test

data:  g1 and g2
t = -1.3242, df = 16, p-value = 0.2041
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.6009321  0.6009321
sample estimates:
mean of x mean of y
 12.16667  13.16667
```

2.7. Valor p frente a IC

Digamos otra vez que las pruebas de significación (PS) y los intervalos de confianza (IC) son dos herramientas de inferencia: ambas permiten pasar de la muestra a la población. Mientras PS hace una pregunta concreta o "cerrada" sobre el valor del parámetro en la población (¿es $\pi = 0.5$?), el intervalo de confianza hace una pregunta "abierta", (¿cuál es el valor deπ?). Se podría argumentar que el intervalo es una herramienta positiva, que dice cuáles son los valores del parámetro compatibles con la muestra observada, mientras que la prueba de hipótesis es una herramienta negativa.

Ejemplo 2.6: Recuperemos el ejemplo de las 52 caras en 100 lanzamientos de una moneda. El intervalo de confianza del auténtico valor de la probabilidad de cara era:

$$IC_{95\%}\ \pi\ :\ P \pm z_{\frac{\alpha}{2}} \cdot \sigma_P \quad \approx \quad 0.52 \pm 1.96 \cdot 0.05 \quad \approx \quad 0.52 \pm 0.10 \quad = \quad 0.42, 0.62$$

Se cree, con una confianza del 95%, que esta moneda tiene una probabilidad de cara situada entre el 42% y el 62%. Este resultado coincide con el de la prueba de hipótesis que, con un p=0.69, no permite rechazar la H de π=0.5.

En el caso de observar 63 caras el IC es:

$$IC_{95\%}\ \pi\ :\ P \pm z_{\frac{\alpha}{2}} \cdot \sigma_P \quad \approx \quad 0.63 \pm 1.96 \cdot 0.05 \quad \approx \quad 0.63 \pm 0.10 \quad = \quad 0.53, 0.73$$

Por lo que ahora se cree, con una confianza del 95%, que esta probabilidad de cara, π, es alguno de los valores comprendidos entre el 53% y el 73%. Dado que excluye el valor 0.5, coincide con PS, que había rechazado H:π=0.5 con nivel de significación p=0.001.

Las conclusiones de IC y PS coinciden.

Definición

Un **intervalo de confianza** incluye el conjunto de valores del parámetro que, puestos en H, no pueden ser rechazados.

Nota técnica: en algunas situaciones, la estimación de la varianza del estimador no es la misma bajo los diferentes escenarios de IC y PS, por lo que no coincidirán plenamente. Por ejemplo, en el caso de π y P la amplitud de los intervalos suele diferir:

PS (P): aceptar si $P \in \pi_H \pm 1.96\sqrt{[\pi_H(1-\pi_H)/n]}$

IC (1-α): $\pi \in P \pm 1.96\sqrt{[P(1-P)/n]}$

En el modelo lineal (comparación medias, regresión,..) sí que coinciden.

Se puede utilizar $IC_{95\%}$ para hacer PS de H de interés, ya que valores del parámetro excluidos del IC generarían PS con valores de p<0.05.

Ejercicio 2.7

En 100 pacientes con SIDA el intervalo de confianza al 95% de la media μ del recuento de CD4 va de 375 a 500. Si se plantearan las dos pruebas de significación siguientes con α= 0.05:

(A) H_A: μ = 400

(B) H_B: μ = 350

Las conclusiones serían:

a) nada se opone a aceptar ambas H;

b) se rechazan ambas H;

c) nada se opone a aceptar H_A y se rechaza H_B;

d) se rechaza H_A y nada se opone a aceptar H_B.

IC ayuda a interpretar PS, ya que informa sobre los valores plausibles del parámetro.

Nota técnica: en el caso de rechazar una hipótesis H bilateral, p.e. π=0.5, la conclusión formal de la prueba de significación sería que se rechaza H sin decantarse hacia ninguno de los dos lados. Pero, a nivel

práctico, el intervalo de confianza permite conocer, no sólo el lado, sino también los valores razonables del parámetro.

Cuando no se rechaza H, IC distingue entre poca información (IC amplio) y efecto nulo o pequeño (IC estrecho).

Recuerde

Utilice siempre IC.

Lectura: las recomendaciones para los autores de revistas biomédicas anteponen el uso de IC al de PS: "Although P values may be provided in addition to confidence intervals, results should not be reported solely as P values" (Consort, item 17).

Cuando no es significativa, PS concluye: "nada se opone a aceptar la H". Pero ello puede ser, bien por falta de evidencia para establecer algo existente (¿muestra pequeña, diseño deficiente, análisis pobre,..?), o bien porque realmente no hay nada que ver.

Recuerde

En PS, ausencia de pruebas no es prueba de ausencia.

Lectura: siempre es frustrante no lograr demostrar el objetivo. Pero si la Ciencia no se lo permite, aún le quedan otros recursos. No se pierda esta página que recoge ejemplos sobre la retórica de los resultados negativos. Planteamientos unilaterales y bilaterales

Hasta el momento, hemos planteado pruebas **bilaterales o** de dos colas, como el ejemplo de la moneda, defectuosa tanto si salían caras de más o de menos. En consecuencia, el rechazo de H ha contemplado ambos lados (Figura 2.4).

$H: \pi = 0.5$

Figura 2.4 Las pruebas bilaterales miran la probabilidad en ambas colas.

Pero se pueden plantear también pruebas de una sola cola. En el ejemplo de la gasolinera, en el que se quería detectar si había timo, ¿qué se puede concluir si la media observada se situaba por encima de la media teórica? ¿Qué regalan gasolina? En esta situación, tiene más sentido una prueba **unilateral por la izquierda** (Figura 2.5)**:**

$H: \mu \geq 1000$

Figura 2.5 Las pruebas unilaterales por la izquierda miran la probabilidad en el lado izquierdo.

Y, de forma simétrica, si se estudia cómo aumenta la respuesta al aumentar la dosis, podría tener más sentido una prueba **unilateral por la derecha** (Figura 2.6):

$H: \mu \leq \mu_0$

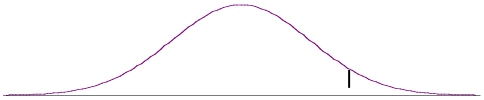


Figura 2.6 Las pruebas unilaterales por la derecha miran la probabilidad en el lado derecho.

Recuerde

En el caso de pruebas unilaterales, debe considerar sólo 1 cola.

Resaltemos dos aspectos relevantes:

(1) Al concentrar todo el nivel de significación en un lado, se hace algo mayor esa cola, por lo que una H uni o bilateral puede cambiar las conclusiones.

(2) El signo igual (acompañado, ahora, por el desigual) sigue figurando en H.

Recuerde

H es el punto de salida y debe establecerse antes de recoger los datos.

Ejercicio 2.8

Repita el Ejercicio 2.6 bajo un planteamiento unilateral.

Ejercicio 2.9

Se desea resolver la prueba H: $\mu \leq 0$ mediante un estadístico que sigue una distribución normal (0,1). El resultado de la prueba ofrece z=-2, por lo que se concluye (elija una):

a) que la media poblacional es 0

b) que la media poblacional es mayor que 0 (con un margen de error del 5%);

c) que la media poblacional es menor que 0 (con un margen de error del 5%);

d) hay una probabilidad del 95% de que la media poblacional sea 0;

e) nada se opone a aceptar la H de que la media es igual o inferior a 0.

3. Decisión: contraste de hipótesis, CH

CH es un instrumento para tomar una decisión manteniendo controlados los riesgos de error.

Definición

Un CH plantea elegir entre dos acciones alternativas.

Historieta: Tener 2 opciones es un tan sólo un dilema. El problema es no tener ninguna.

Ejemplo 3.1 (Prestado de un ejercicio de la profesora Monique Becue). Para clasificar cierto "garabato" como 8 o como B, un programa de reconocimiento de patrones mide la curvatura izquierda (Y) cuya distribución tiene una media de 12u si se trata de un "8", y una media superior si se trata de una "B". Se sabe que la distribución de Y es Normal y que σ=3u. Si se está dispuesto a aceptar que un 5% de "ochos" (8) sean reconocidos como "bes" (B), ¿a partir de qué valor se dirá que se trata de una "B"?

Variable: curvatura izquierda del garabato (Y)

Se usará el estadístico $\boldsymbol{z} = \frac{y-\mu_H}{\frac{\sigma}{\sqrt{n}}}$

H: μ = 12u (se trata de un 8)

Regla: Con p=0.05, se rechazará H si $\boldsymbol{z} > Z\alpha = 1.645$.

```
> qnorm(p=0.05,lower.tail=FALSE) # Z0.05
[1] 1.644854
```

Distribución del estadístico bajo H: $\boldsymbol{z} \sim N(0,1)$.

Premisas: dado que n=1, Y debe ser normal.

Cálculo del límite: $y = \mu_H + z_\alpha (\sigma/\sqrt{n}) = 12 + 1.645 * 3 = 16.935$

Si y > 16.935 se rechaza que sea un 8.

Suponga ahora que también conoce la distribución de las "B": N(21,3). Aceptando como límite de decisión y = 16.935, desea calcular la probabilidad de que una B sea reconocida como un 8 (Figura 3.1). Ahora dispone de dos situaciones hipotéticas, entre las que debe escoger.

H_0: μ = 12 (se trata de un 8)

H_1: μ = 21 (se trata de una B)

Prob[y ≤ 16.935 condicionado a Y~N(21,3)] = $P\left[Z \leq \frac{16.935 - 21}{3}\right] = P(Z \leq -1.355)$

```
> pnorm(q=-1.355,lower.tail=TRUE) # P ( Z ≤ -1.355 )
[1] 0.08770878
```

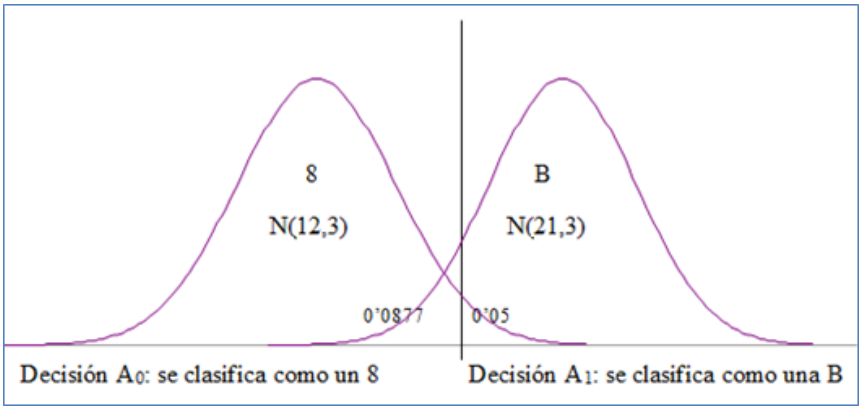


Figura 3.1 Si el valor supera el límite 16.935 clasifica el garabato como B (A_1) y en caso contrario como 8 (A_0).

Nota: En lo que sigue emplearemos A_0 y A_1 (acción 0, acción 1) para resaltar la acción que implica CH. Sea A_0 "conservadora" y A_1 "innovadora". Para tomar la acción A_1 hace falta rechazar H_0.

Nótese que se han identificado dos conclusiones erróneas y se han cuantificado los riesgos respectivos:

P (concluir B | realidad 8) = 0.05

P (concluir 8 | realidad B) ≅ 0.088

Un organismo científico, como la revista Medicina Clínica o la colaboración Cochrane, está interesado en lo que científicamente se sabe y, por tanto, en realizar intervalos de confianza o pruebas de significación. En cambio, un órgano ejecutivo, como una agencia reguladora del medicamento o un comité que elabora protocolos, debe proponer decisiones, acciones concretas.

Nota: ¿Cuál es el papel de las sociedades científicas? ¿Aportar un conocimiento que facilite una toma de decisión posterior por quién corresponda? ¿O elaborar consensos de guía de práctica clínica de uso posterior obligado? Esta pregunta nos supera. Como posibles usuarios, agradeceremos una guía consensuada de práctica clínica que se nos presente a modo de sugerencia. Sin lugar a duda, nuestros representantes, que deben asignar presupuestos a diferentes partidas, requerirán otra metodología.

Ejemplo 3.2: Fisher y Hill mantuvieron posiciones distintas en cuanto a la evidencia disponible sobre los efectos del tabaco. Sea cual sea esta evidencia, a un responsable de Salud Pública, lo que le concierne es, a la luz de dicha información, cuál debe ser su actuación. Greenland recuerda que un organismo de Salud Pública debe actuar y debe, por tanto, tomar decisiones: ante humo en un bosque, la acción pertinente es enviar bomberos, no científicos para averiguar si debajo del humo hay fuego.

A nivel personal, por ejemplo, un fumador debe valorar las consecuencias de los dos "errores" posibles: a) que decida seguir fumando, pero tenga razón Hill y él mismo sea de la proporción de casos que desarrollan el cáncer hacia los 50 años; o b) que decida no fumar, pero tenga razón Fisher y no se "ahorre" dicha enfermedad. Cada uno debe valorar qué consecuencias tiene cada posible situación.

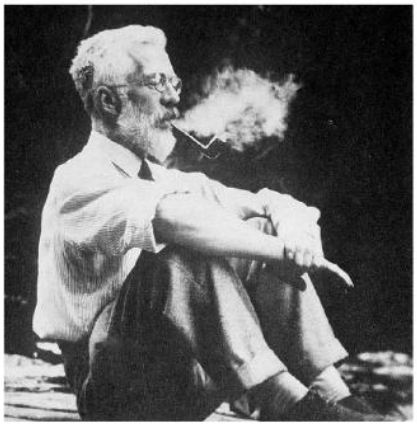


Figura 3.2. Ronald Fisher fumando en pipa

Recuerde

PS contesta "¿qué creo?" y CH, "¿qué hago?".

3.1. Límites de significación

El límite del nivel de significación p a partir del cual se rechaza H tiene un equivalente en la escala de los estadísticos, z o t. En la escala Z, los límites que corresponden a p=0.05 son -1.96 y +1.96. En la t de Student, dependerá de los grados de libertad.

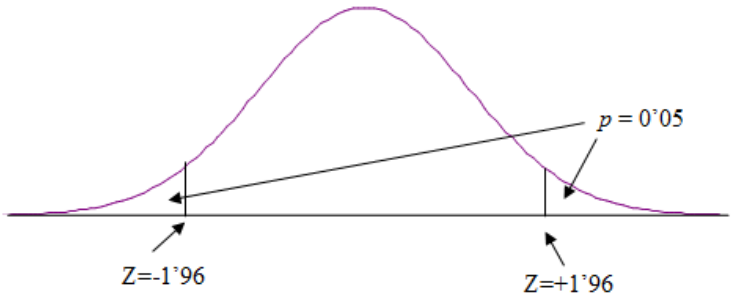


Figura 3.3 Es equivalente preguntarse si p<0.05 o si Z es mayor que 1.96 o menor que -1.96.

Ejercicio 3.1

En los ejercicios 2.3 y 2.4 comparó el valor de p con 0.05. ¿Cómo habría hecho la comparación con Z? ¿Y con t en el 2.6?

3.2. Errores tipo I y II. Riesgos α y β

En CH hay 2 tipos de errores.

Definición

El **error de primera especie o tipo I** consiste en decidir la acción alternativa (A_1) cuando era cierta H_0.

Tomar A_1 | es cierta H_0

Ejemplo 3.3: Concluir que es una B cuando en realidad es un 8 es un error tipo I.

Definición

El **error de segunda especie o tipo II** consiste en decidir la acción nula (A_0) cuando es cierta H_1.

Tomar A_0 | es cierta H_1

Ejemplo 3.3 (cont): Concluir que es un 8 cuando en realidad es una B es un error tipo II.

Definición

Las probabilidades correspondientes de cometer errores de primera y de segunda especie reciben el nombre de **riesgos α y β:**

$\alpha = P(\text{Decidir } A_1 \mid \text{es cierta } H_0)$

$\beta = P(\text{Decidir } A_0 \mid \text{es cierta } H_1)$

Ejemplo 3.4: Riesgo α = P(Decidir es una B | en realidad es un 8)

Riesgo β = P(Decidir es un 8 | en realidad es una B)

De esta manera, α representa la proporción de 8 que serán identificados como B y β su recíproco.

Nota: De aquí proviene el nombre de estadística frecuentista, ya que acota la frecuencia de errores.

Ejemplo 3.5: Un laboratorio farmacéutico propone a una agencia reguladora del medicamento un Ensayo Clínico para contrastar H_0 (misma eficacia que referencia) frente H_1 (eficacia mayor = Δ). Si rechaza H_0, acuerdan poner el fármaco en el mercado (A_1). El riesgo α sería la proporción de medicamentos como la referencia (H_0) que son finalmente puestos en el mercado (A_1). A su vez, el riesgo β es la proporción de medicamentos que no llegan al mercado (A_0) entre los que alcanzan el efecto Δ (H_1).

Ejercicio 3.2

Un proveedor entregaba un reactivo con un tiempo de reacción medio de 100 seg y desviación tipo de 10 seg. Ahora, ofrece uno mejor, con parámetros μ=50 seg y σ=5 seg y Vd. decide hacer un CH para guiar su actitud futura. Sean:

H_0: μ=100seg y σ=10seg (viejo); A_0: decidir usar el viejo;

H_1: μ=50seg y σ=5seg (nuevo); A_1: decidir usar el nuevo.

El riesgo α de cometer un error de primera especie es (cuál/cuáles son ciertas?):

a) la probabilidad de que el reactivo sea nuevo

b) decidir usar el nuevo (A_1) a pesar de ser como los viejos (H_0)

c) delante de reactivos con propiedades como los viejos, la probabilidad de decidir usar los nuevos

d) decidir usar el viejo (A_0) a pesar de ser de los nuevos (H_1)

e) con propiedades como los nuevos, la probabilidad de decidir usar los viejos

f) la proporción de reactivos como los viejos que serán aceptados como si fueran de los nuevos.

g) todas son falsas.

Ejercicio 3.3

En un contraste de hipótesis, si H_0 es cierta, es posible (elija una):

a) cometer dos errores, el de tipo I y el de tipo II

b) sólo se puede producir el de tipo I

c) sólo se puede producir el de tipo II

d) ninguno, ya que H_0 es cierta.

Definición

La **potencia** es 1-β o probabilidad de decidir A_1 cuando es cierta H_1:

$$Potencia = 1-\beta = P(Decidir\ A_1 \mid es\ cierta\ H_1)$$

Tipos de errores y riesgos		Decisión	
		A_0	A_1
Realidad	H_0	1-α	Tipo **I** (riesgo α)
	H_1	Tipo **II** (riesgo β)	Potencia = 1-β

Tabla 3.1 Resumen tipos de errores y riesgos

4. Use intervalos de confianza

Las guías de publicación aconsejan emplear siempre intervalos de confianza.

Recuerde

Use $IC_{95\%}$.

Si Vd. desea emplear P valores, lea los siguientes apartados (marcados con *) y estos 2 artículos sobre la distinción entre evidencia y decisión y 12 interpretaciones erróneas del P valor.

4.1. IC, PS y CH *

En IC, el nivel de confianza α se decide a priori. En CH también, y se opta por aquel diseño y estadístico que minimiza β, que también se establece a priori. Por tanto, en el entorno de IC y CH, lo único que tiene valor y debe, por tanto, ser reportado son los valores de α y β decididos a priori. En cambio, en PS, el nivel p es un resultado obtenido al final del experimento y el nivel de evidencia que aporta en contra de H sería diferente ante un valor de p=0.023 o de p<0.001, por lo que se recomienda reportar el valor de p exacto —hasta el decimal requerido.

Resumen

En IC debe informar del valor de α fijado a priori.

En PS debe reportar el valor exacto obtenido de p.

En CH se debe informar de los valores de α y β fijados a priori.

La misma concordancia en el cálculo que existe entre IC y PS, aplica también a CH. En cambio, los resultados de cada técnica deben interpretarse de acuerdo con sus objetivos.

Resumen

IC, PS y CH difieren en objetivos:

IC, estimar valores del parámetro

PS, aportar evidencia en contra de H

CH, decidir entre A_0 y A_1 minimizando los riesgos α y β

Pero coinciden en su mecánica:

IC (1-α): $\mu \in \ y \pm 1.96 \cdot \frac{\sigma}{\sqrt{n}}$

PS (y): aceptar H si $y \in \ \mu_H \pm 1.96 \cdot \frac{\sigma}{\sqrt{n}}$

CH (α,β): decidir A_0 si $y \in \ \mu_0 \pm 1.96 \cdot \frac{\sigma}{\sqrt{n}}$

4.2. Interpretación errónea de p y α *

En la **¡Error! No se encuentra el origen de la referencia.** los riesgos α y β representan probabilidades condicionadas a la fila, no a la columna. Es decir, proporcionan la probabilidad de una conclusión (acción) dada una H. Nótese que las filas representan valores del parámetro, que es, bajo el escenario definido, una constante; mientras que las columnas representan zonas en las que se sitúa el estadístico, que sí que es una variable aleatoria.

Así, en CH (H_0 frente a H_1) para tomar una decisión (A_0 frente a A_1), α y β representan la proporción o frecuencia de decisiones erróneas a largo plazo. En el Ejemplo 3.5, α es la proporción de fármacos iguales que el control (H_0) que a largo plazo son puestos en el mercado (A_1); y β, la de fármacos que superan el control en un valor Δ (H_1) que no son puestos en el mercado (A_0).

Nótese que en PS, p (y su máximo aceptado, 0.05) indica el nivel de evidencia en contra de H, mientras que en CH α y β indican la frecuencia de decisiones erróneas.

Recuerde

p en PS es medida de información empírica ("evidencia") en contra de H; mientras que α y β en CH cuantifican la frecuencia de decisiones erróneas.

La Tabla 4.1 expone términos que pueden emplearse para informar del resultado de PS o CH.

PRUEBA DE SIGNIFICACIÓN	**Si el valor de p es…**	
	Grande (p.ej. 0.634)	**Pequeño (p.ej. 0.0001)**
H es …	Verosímil	inverosímil
La diferencia…	es explicable por el azar del muestreo	no es explicable por el azar del muestreo
La diferencia…	no es estadísticamente significativa	sí es estadísticamente significativa
A nivel práctico …	no hemos logrado demostrar que la moneda está cargada	creemos que la moneda está cargada
CONTRASTE DE HIPÓTESIS	**Si el estadístico se sitúa en…**	
	Región de aceptación	**Región crítica**
Hipótesis…	Se acepta H_0	se rechaza H_0
Acción…	Se toma la acción A_0	Se toma la acción H_1

Tabla 4.1 La PS y CH en palabras

Recuerde

Ni el riesgo α ni el nivel de significación p pueden resumirse por "la probabilidad que tengo de haberme equivocado".

Ejercicio 4.1

¿Cuál o cuáles son correctas?

a) El nivel p es la probabilidad de equivocarse;
b) El nivel p es la probabilidad de equivocarse al rechazar H:
c) El nivel p es la probabilidad de equivocarse al aceptar H;
d) El nivel p es la probabilidad de observar el resultado actual (o más extremo) en caso de que fuera cierta H
e) El riesgo α es la probabilidad de equivocarse;
f) El riesgo α es la probabilidad de equivocarse al rechazar H:
g) El riesgo α es la probabilidad de equivocarse al aceptar H;
h) El riesgo α es la frecuencia esperada de ocasiones en las que siendo cierta H_O tomaremos la decisión (errónea) A_1.
i) El riesgo β es la probabilidad de equivocarse;
j) El riesgo β es la probabilidad de equivocarse al rechazar H:
k) El riesgo β es la probabilidad de equivocarse al aceptar H;
l) El riesgo β es la frecuencia esperada de ocasiones en las que siendo cierta H_1 tomaremos la decisión (errónea) A_0.

Ejemplo 4.1: La celebración final de carrera ha sido magnífica. A las 5 am los amigos se despiden, pero uno de ellos decide seguir la farra y le pide al taxista que le lleve a una buena partida de Póker. Tras pasar los controles típicos, que su amigo creía cosa de película, consigue entrar en un 5º piso de la calle Enrique Granados donde se sienta a una mesa y empieza a perder dinero. Sus rivales no paran de sacar magníficas jugadas. Tanto, que él calcula que, asumiendo que no hacen trampas, la probabilidad de esos resultados (o incluso mejores) es de tan sólo una entre cien. ¿Qué hace? Por supuesto, deja de jugar. El nivel de significación p=0.011 le permite rechazar la H de que no le hacen trampas.

Ejemplo 4.2: En la celebración de las Navidades, un joven investigador vuelve del hospital Mount Sinai para visitar a su familia. Y acaban jugando al Póker con idénticos resultados que el ejemplo anterior. A pesar de que este investigador calcula el mismo nivel de

significación anterior (asumiendo que no hacen trampas, esos resultados o mejores sólo ocurren 1 vez entre cien), sigue jugando confiado, ya que no se plantea la posibilidad alternativa, de que su familia le haga trampas. Por lo que dice, "caramba, qué mala suerte tengo hoy".

Lectura: La estadística Bayesiana lamenta que la solución de los dos ejemplos anteriores no tenga en cuenta toda la información contenida en el enunciado. Antes de empezar a jugar, el primer titulado ya podía sospechar que le harían trampas, pero no el segundo. Para poder calcular, a partir de los resultados muestrales, la probabilidad de que una hipótesis sea cierta, es preciso recurrir a una formalización del conocimiento científico previo: antes de los datos que actualmente se están analizando, ¿qué se sabía sobre este tema?, ¿qué se sabía sobre el valor del parámetro? Si se acepta representar el nivel de incertidumbre previa en forma de probabilidades sobre los diferentes valores del parámetro, ya se tienen los elementos necesarios para actualizar la información científica mediante el teorema de Bayes.

Ejercicio de Navegación

Referencias críticas sobre el abuso de las pruebas de significación, así como enlaces a paginas web aplicadas, y un "applet" muy instructivo, pueden encontrarse en:

http://www.stat.duke.edu/~berger/p-values.html

Recuerde

IC, PS y CH estudian la información aportada por los datos actuales, pero no la "suman" a la información previa.

4.3. Sólo el contraste de hipótesis permite "Aceptar H_0" *

PS no especifica H_1 y, por tanto, no tiene definida ninguna medida análoga al riesgo β. En consecuencia, PS no tiene argumento para defender H.

Nota técnica: el riesgo β puede delimitarse cuando el contraste de hipótesis tiene, como en el ejemplo del 8 y la B, la forma:

H_0: $\mu = \mu_0$

H_1: $\mu = \mu_1$

Pero si la prueba de significación es de la forma:

H: $\mu = \mu_H$

Entonces la definición de una medida análoga al riesgo β bajo todos los posibles $\mu \neq \mu_H$ s próximos a μ_H ese riesgo tiende hacia 1-α (Figura 4.1). Es decir, como PS sólo define H, este planteamiento "asimétrico" conduce a la conclusión asimétrica: si el valor de p es pequeño, se considera inverosímil. En cambio, si p es grande, "nada se opone a aceptar H".

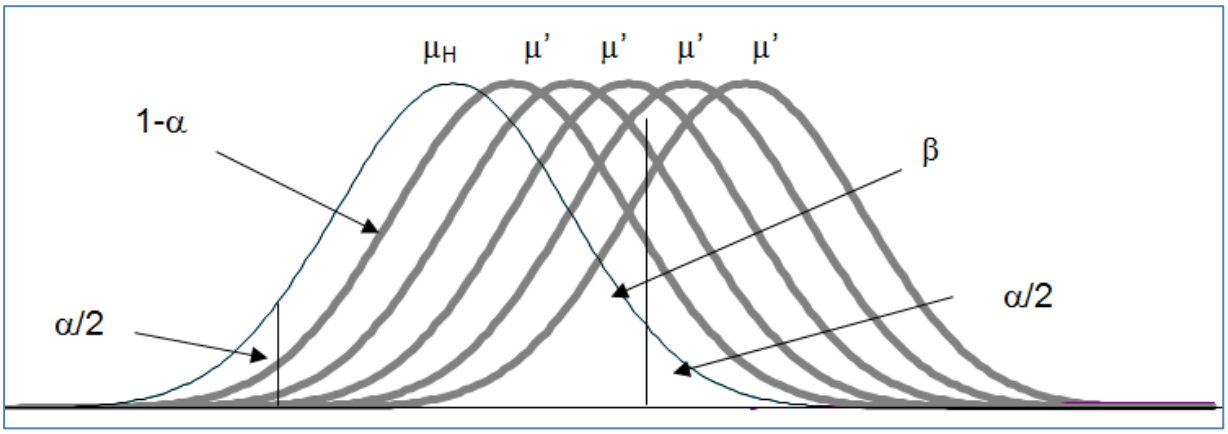


Figura 4.1 Si no hay hipótesis alternativa cerrada, el riesgo beta no está acotado

Recuerde

En PS, "ausencia de pruebas" no es "prueba de ausencia".

En PS diga "no se han detectado diferencias" en lugar de "no existen diferencias".

El CH, al tener acotados α y β, permite tomar ambas decisiones.

Ejercicio 4.2

El laboratorio Yotambién S.L., para demostrar que su genérico es tan eficaz como el de la compañía Losprimeros S.A., realiza un ensayo en el que compara ambos productos. Supóngase que obtiene un nivel de significación p=0.23, ¿puede concluir que ambos productos tienen la misma eficacia?

Conviene ir con mucho cuidado con las palabras que se utilizan para explicar las conclusiones de una prueba de significación. La Tabla 4.1 resume algunas de las más habituales. Nótese la asimetría de la conclusión a la que se llega en ambas regiones: mientras en la zona crítica se afirma que se rechaza H ("se ha demostrado la culpabilidad del acusado"), en la zona de aceptación no hay afirmaciones rotundas ("absuelto por falta de pruebas").

Ejercicio 4.3

¿Alguna(s) de las siguientes es falsa? :

a) En PS se buscan evidencias en contra de H

b) CH permite tomar ambas decisiones

c) Tanto p como α cuantifican áreas de las distribuciones de probabilidad, pero miden aspectos distintos.

d) En PS debe reportarse el valor exacto del nivel de significación p

e) En CH debe reportarse el valor previo α, usualmente, 0.05

f) En PS, si p>0.05, nada se opone a aceptar H

g) Una ventaja de CH es que permite decidir tanto A0 como A1

h) Una ventaja de CH es que cuantifica β

Ejercicio 4.4

PS es conservadora en el sentido de que se declara ..???.. H hasta que no haya clara evidencia en su contra:

a) ???=cierta

b) ???=falsa

c) PS no es conservadora

d) todas son incorrectas.

Ejercicio 4.5

En un estudio para comparar dos tratamientos, p=0.341. ¿Cuál/es son ciertas?

a) Nada se opone a aceptar H_0.

b) No existen diferencias

c) No se han detectado diferencias

d) La probabilidad de que sean diferentes es 0.341.

Lectura: Karl Popper ha contribuido a incorporar los avances estadísticos a la epistemología o metodología científica. De acuerdo con esta asimetría de la conclusión de una prueba de hipótesis, afirmó que lo único que se puede hacer con una teoría científica es ponerla a prueba y rechazarla en el caso de que encontremos pruebas en su contra, pero que nunca se podrá demostrar que sea cierta y constituya la última palabra de la ciencia en ese punto. Así, Popper dice que el criterio para establecer el status científico de una teoría es su refutabilidad o su testabilidad: "para ser colocados en el rango de científicos, los enunciados o sistemas de enunciados deben ser susceptibles de entrar en conflicto con observaciones posibles", lo que es conocido como problema de la demarcación. Así, una teoría científica es más fuerte cuando es más falseable, cuanto más fácilmente podría demostrarse su falsedad (caso de ser falsa).

4.4. Interpretación del CH *

Desde un punto de vista formal, disponer de dos hipótesis simples, cada una con un único valor, permite definir muchas propiedades interesantes para escoger el "mejor" estadístico. Los libros clásicos de estadística matemática exponen la teoría desarrollada por Pearson y Newman sobre el *contraste de* dos *hipótesis* simples. Fisher, se centraba en la inferencia sobre una hipótesis, por lo que sólo puede cuantificar p y sólo puede rechazar H en lo que él llamó PS.

Recuerde

PS es inferencia; si la p es pequeña, Fisher recomienda modificar nuestras *opiniones* sobre la veracidad de H.

CH es *decisión* que permite acotar los riesgos de tomar *acciones* erróneas.

Ejercicio 4.6

Las siguientes frases podrían figurar en la discusión de un artículo, ¿Cuáles son de inferencia y cuáles de decisión?

a) el riesgo es mayor en pacientes de tipo A.

b) el riesgo disminuye a la mitad si se adoptan las medidas X.

c) la obesidad abdominal es el componente de síndrome metabólico de mayor prevalencia en mujeres.

d) el valor predictivo de la escala de Z implica que debería utilizarse en el futuro para clasificar a este tipo de enfermos.

e) si hay dos o menos factores de riesgo presentes y la PAS ≥ 160 o la PAD ≥ 100 (siendo PAS < 180 y PAD < 110), conviene intentar cambios en el estilo de vida durante varios meses y luego, si se mantiene, tratamiento farmacológico.

Las acciones conllevan consecuencias. Y conviene tenerlas en cuenta. Ya expusimos que Greenland reclamó distinguir entre la Ciencia de la Epidemiología y la acción de fomentar la Salud Pública. Y en el capítulo de probabilidad y riesgo recordamos que la definición estadística de riesgo incluye la gravedad de las consecuencias.

Historieta: Los mismos datos en los Ejemplos 4.1 y 4.2 han llevado a decisiones diferentes: abandonar el garito de juego o seguir jugando con la familia. La diferencia es el grado previo de credibilidad de la hipótesis. Pero además, las consecuencias son diferentes, ya que seguir la partida familiar no conlleva pérdidas: incluso, en el caso de trampas, "el dinero se queda en casa".

Ejemplo 4.3: Gosset era un estadístico que trabajaba en la cervecera Guiness en su departamento de control de calidad, donde se planteaba la decisión de aceptar o rechazar una barrica de cerveza. Además de los riesgos α y β, debía considerar los costes por desechar una barrica correcta y por poner en el mercado una que no lo era.

Recuerde

El proceso de decisión, además de los riesgos de error debe valorar también sus consecuencias, el coste que se paga por cada decisión errónea y el premio que se obtiene con las decisiones correctas.

Ejemplo 4.4: Es bien conocido que aunque un tratamiento puede haber demostrado un cierto efecto positivo en una variable de interés, sus costes pueden aconsejar antes otra intervención sanitaria más eficiente, en el sentido de que una misma "inversión" origine un mayor "retorno", valorado en términos de salud.

Lectura: Aconsejar un producto químico o biológico de nueva creación no conlleva los mismos riesgos que aconsejar un hábito saludable que se ha practicado siempre.

Nota técnica: CH es el primer instrumento de la teoría de la decisión, que constituye toda una rama de la estadística y es ampliamente utilizada en otras disciplinas, como por ejemplo, la economía, donde los "costes" y los "premios" son fácilmente expresables en una única escala. El diagnóstico y el tratamiento son dos ejemplos de acciones médicas que podrían beneficiarse de las aportaciones de la teoría de la decisión.

Puede ser razonable esperar que el efecto de una intervención sea el mismo en diferentes condiciones (país, entorno de atención al paciente, raza,…). Incluso que lo sea la capacidad predictiva de un indicador valorada por su sensibilidad y especificidad. Pero no es en absoluto razonable esperar que las consecuencias de una decisión se valoren igual en diferentes entornos. Por ejemplo, el "valor" del mismo coste de un medicamento puede diferir de un país a otro.

Recuerde

El proceso de decisión implica una valoración de las consecuencias que tienen connotaciones *locales* y es más difícilmente extrapolable que la mera inferencia de conocimiento.

Ejercicio adicional

Encuentre un original científico reciente que, en su discusión, vaya de la interpretación de los resultados de inferencia a la decisión ulterior sin considerar formalmente el proceso de decisión, sus riesgos y sus consecuencias en el entorno en el que propone la acción.

En ocasiones resulta difícil discernir si el p valor reportado hace referencia a un objetivo de inferencia o decisión, pero esto no ocurre con el intervalo de confianza, donde queda claro el objetivo de inferencia.

5. Equivalencia

Hasta ahora hemos intentado establecer 'diferencias'.

Ejemplo 5.1: Por ejemplo, "el riesgo de sida es mayor en toxicómanos por vía parenteral", o bien "el nuevo tratamiento es mejor que el clásico".

Pero puede interesar establecer '*equivalencia'*.

Definición

Un tratamiento es **equivalente** a otro si la diferencia de sus efectos no alcanza un cierto valor ε que marca el límite de la irrelevancia.

Ejemplo 5.2: Se desea establecer que: $\varepsilon_1 < \mu_1-\mu_2 < \varepsilon_2$

ε_1 y ε_2 delimitan el intervalo de equivalencia.

Ejemplo 5.3: Se desea establecer, de forma simétrica, que: |□1-□2| < □̃

El concepto de equivalencia es más amplio que el de la estricta igualdad, pues incluye también los valores, cercanos a la igualdad, que no son relevantes desde el punto de vista práctico. La siguiente figura representa todos los posibles valores de la diferencia entre las dos medias de interés: $\mu_1-\mu_2$.

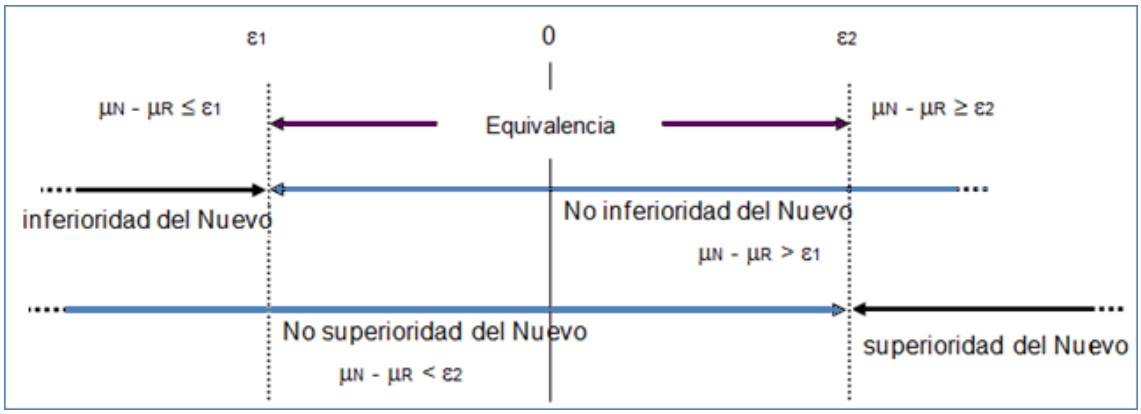


Figura 5.1 Definición de los conceptos de no superioridad, equivalencia y no inferioridad

Para establecer equivalencia se debe demostrar que las diferencias no alcanzan los dos límites especificados. Se puede hacer con un intervalo de confianza que deberá quedar comprendido entre estos límites, lo que equivale a realizar dos pruebas que deberán rechazar ambos límites.

Ejemplo 5.4: La figura 5.2 muestra 3 estudios en los que se concluiría equivalencia y 3 estudios en los que no.

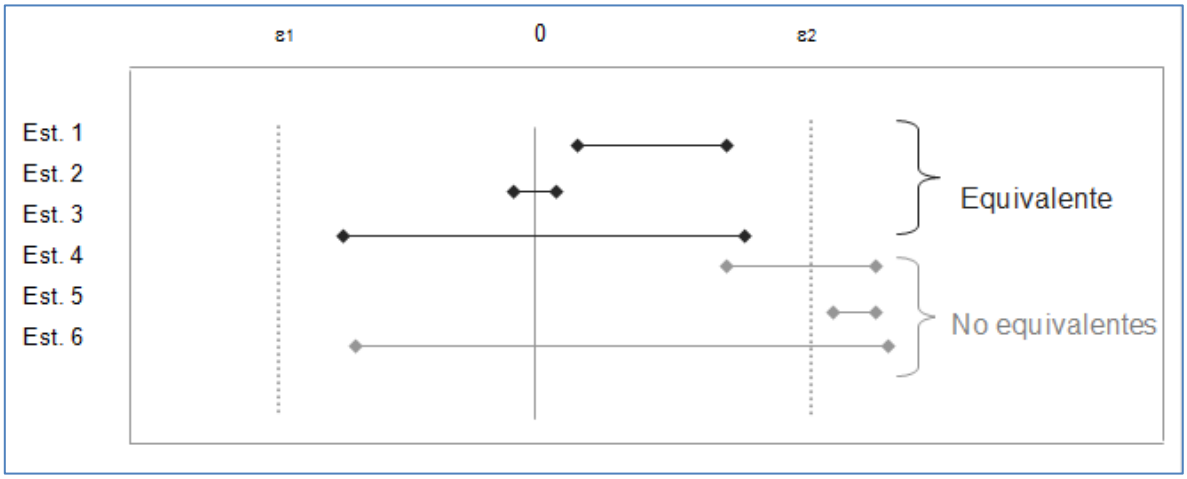


Figura 5.2 Los estudios 1 a 3, que dejan fuera ε_1 y ε_2, establecen equivalencia

Ejercicio 5.1

Si en el Ejemplo 5.4 de los datos de la figura 5.2 se hubiera hecho la PS para demostrar diferencias, ¿en qué estudios de los anteriores se concluiría que los tratamientos son diferentes? Razone posibles discordancias.

Recuerde

Ambos límites deben ser rechazados para poder establecer equivalencia.

Definición

Un tratamiento es **no inferior** a otro si éste no le supera en un cierto valor ε que hace relevantes las consecuencias.

Recuerde

Los planteamientos de "no inferioridad" y "no superioridad" sólo consideran un límite, sea ε_1 o ε_2.

Tanto la no inferioridad como la no superioridad se establecen mediante un constraste unilateral. Y la equivalencia puede establecerse mediante el uso simultáneo de ambos, por lo que el procedimiento que se utiliza recibe el nombre de Prueba Doblemente Unilateral (PDU) o two-one-sided test.

Ambos contrastes de la PDU se suelen realizar con riesgo $\alpha = 0.05$ y el riesgo α global de la PDU se mantiene en 0.05. Si el IC se calcula con una confianza $1-2\alpha$ (0.90, si $\alpha = 0.05$), coincidirán las conclusiones de la PDU con las del IC.

Nota: Aunque la estimación por intervalo se realiza con una confianza $1-2\alpha = 0.90$, el criterio de decisión basado en dicho IC tendrá un riesgo $\alpha=0.05$. Ello es así porque ambos límites de no equivalencia (que definen las dos H_0) no pueden ser simultáneamente ciertos; y, por tanto, sus riesgos α no necesitan ser sumados.

Nota: Si en lugar de dos CH se hubieran realizado dos PS con sus correspondientes niveles de significación p_1 y p_2, se acepta como nivel único de significación p, el mayor de los dos p_1, p_2 observados.

Ejemplo 5.5: Un nuevo (N) antiinflamatorio tiene una tolerabilidad superior a cierto producto clásico de referencia (R). Interesa poder demostrar que sus niveles de eficacia son parecidos. La eficacia se mide por la proporción de casos en los que desaparece el dolor a los 30'. Ambos fármacos serán equivalentes en eficacia si las proporciones de desaparición del dolor no difieren en más de un 8%. El intervalo de confianza (90%) de la diferencia de ambas proporciones va entre -6% y +3%. Dado que no alcanza los límites, se puede rechazar la no equivalencia ($\alpha=0.05$).

Ejercicio 5.2

Mediante un diseño en que todos los casos pasan por los dos tratamientos, se ha obtenido en n=20 casos el valor de la **D**iferencia entre ambas Presiones Arteriales Diastólicas (PAD) tras 3 meses con el tratamiento de **R**eferencia y 3 con el **N**uevo. Se ha establecido el límite de no equivalencia clínica de un hipotensor en ±10 mmHg. Los resultados han sido $X_D=3$ y $S_D=10$. Calcule el intervalo de confianza y decida si existe equivalencia.

Ejercicio 5.3

Decidir en el ejercicio anterior si existe equivalencia mediante el doble contraste de hipótesis unilateral. ¿Cómo cambiaría la presentación de resultados entre PS y CH?

Ejercicio 5.4 (mismo hipotensor, pero menos casos):

Repita el Ejercicio 5.2 y el Ejercicio 5.3 mediante IC y CH, asumiendo que los resultados han sido: n= 5; X_D=3 y S_D=10.

Ejemplo 5.6 (muy técnico): Para el establecimiento de equivalencia en biodisponibilidad (o bioequivalencia) se suele requerir que el cociente de los niveles en sangre entre R y N se encuentre entre 0.8 y 1.25. Es decir, que ni R puede estar al 80% de N (80%=cuatro quintos: 4/5), ni que N puede estar al 125% de R (125%=cinco cuartos: 5/4). En concreto, se pide que la media geométrica de dichos cocientes esté entre ambos valores o lo que es lo mismo, que la media aritmética de la diferencia entre ambos logaritmos se sitúe entre log(0.8)=-0.223 y log(1.25)=0.223. Así, se trabajará con la "diferencia de los logaritmos naturales", que se corresponde con el logaritmo de los cocientes que se desea mantener, en promedio, entre los dos valores requeridos. En un diseño de datos apareados los resultados han sido: n=12, X_{DL} =0.1 y S_{DL} =0.2.

Mediante $IC_{90\%}$: $\mu_{LR-LN} \in X_{LR-LN} \pm t_{\alpha/2} \cdot S_{LR-LN}/\sqrt{n}$

$$\mu_{LR-LN} \in 0.1 \pm t_{11,0.05} \cdot 0.2 / \sqrt{12} \rightarrow \mu_{LR-LN} \in \textbf{(-0.004, 0.204)}$$

```
# t11,0.05
> qt(p=0.05,df=11,lower.tail=FALSE)
[1] 1.795885
```

Mediante $PDU_{\alpha=0.05}$: H_{0A}: $\mu_{LR-LN} \leq -0.223$

H_{1A}: $\mu_{LR-LN} > -0.223$

$$t_1 = \frac{0.1 - (-0.223)}{0.2 \quad \overline{12}} = 5.597$$

```
> pt(q=5.597,df=11,lower.tail=FALSE)
[1] 8.05157e-05
```

H_{0B}: $\mu_{LR-LN} \geq 0.223$

H_{1B}: $\mu_{LR-LN} < 0.223$

$$t_2 = \frac{0.1 - 0.223}{\frac{0.2}{\sqrt{12}}} = -2.133$$

```
> pt(q=2.133,df=11,lower.tail=FALSE)
[1] 0.0281428
```

Por lo que tanto el IC como la PDU permiten concluir la equivalencia de ambos productos.

Lectura: Los planteamientos de equivalencia que se han resuelto en estos ejemplos hacen referencia a la equivalencia en media. Ello implica que un paciente tiene los mismos valores esperados bajo ambos productos en comparación y, por tanto, ambos preparados o productos son igualmente aconsejables para un nuevo paciente (equivalencia poblacional o prescribilidad). Para que dos preparados se puedan intercambiar en un paciente ya tratado (equivalencia individual o intercambiabilidad) es necesario, además, que no exista interacción entre el preparado y el paciente, es decir: que la diferencia (quizás nula) entre ambos preparados sea la misma para todos los pacientes. La demostración de esta condición ha sido exigida por algunos, resultando en una mayor dificultad para la salida al mercado de productos genéricos.

5.1. Sensibilidad de un estudio

Definición

Sensibilidad es la capacidad de un ensayo clínico concreto para distinguir entre un tratamiento eficaz y un tratamiento ineficaz o menos eficaz.

Es importante en cualquier ensayo pero tiene una implicación diferente en los ensayos que intentan demostrar diferencia entre tratamientos (de superioridad) que en los que intentan demostrar no-inferioridad.

En un ensayo de superioridad, si ésta se demuestra, queda también establecida su sensibilidad (se auto-valida). En cambio, un ensayo de equivalencia que alcanza el resultado deseado, o un ensayo de superioridad con un resultado negativo, siempre queda la duda de si: a) no ha demostrado diferencias porque no existen; o b) porque el estudio no hubiera sido capaz de establecerlas —por no ser "sensible" a ellas.

La sensibilidad se puede deducir a partir de: (1) Evidencia histórica de la sensibilidad a los efectos del tratamiento (ensayos pasados con un diseño similar lograron distinguir a los tratamientos efectivos); y (2) un apropiado diseño y desarrollo del ensayo, que no limitan su capacidad para distinguir entre tratamientos.

Recuerde

Un estudio de superioridad significativo permite *inducir* (aporta evidencia de) su sensibilidad. Un estudio de equivalencia requiere poder *deducir* su sensibilidad de su diseño y calidad de ejecución.

Varios factores pueden reducir la sensibilidad del ensayo: cambios en la población en estudio (criterios de selección), cambios en la dosis y pautas de tratamiento, cambios en las variables de eficacia y su momento de evaluación, periodos de lavado pre-inclusión, bajo cumplimiento con la medicación, baja respuesta de los pacientes a los tratamientos, uso de tratamientos concomitantes prohibidos, pacientes que tiendan a mejorar espontáneamente, criterios diagnósticos mal aplicados (pacientes sin la patología), evaluación sesgada debida al conocimiento de que todos los pacientes reciben algún tratamiento activo, etc.

Quien proponga un diseño de no-inferioridad, debe poder aportar evidencia histórica de la sensibilidad de diseños similares a los efectos del tratamiento del estudio. Así, el diseño debe ser similar a los ensayos previos respecto a: criterios de selección, variables, análisis, etc. Además, su ejecución debe de ser de alta calidad: reclutamiento, seguimiento, administración de la intervención, valoración, etc.

5.2 Margenes de equivalencia, no inferioridad y no superioridad

El margen ε se establece a priori a partir de criterios clínicos.

En cualquier caso, el margen ε de no-inferioridad siempre debe ser inferior al margen previo de eficacia Δ establecido respecto a un placebo, quizás la mitad o la tercera parte. Este hecho puede provocar un mayor tamaño muestral

Recuerde

ε debe ser menor a Δ

Referimos al lector interesado a la extensión de la Consort para equivalencia; a los documentos ICH E10 de elección del grupo control y E9 de análisis estadístico; y a la directriz de la EMA sobre estudios de equivalencia.

Soluciones a los ejercicios

2.1. La respuesta correcta es la d). La a) es incorrecta ya que "conocer el valor del parámetro" es el objetivo de la estimación por intervalo de confianza, no del contraste de hipótesis. La b) no es correcta ya que en la prueba de significación la hipótesis forma parte del enunciado del problema y debe ser siempre independiente de la obtención de los datos (lo que suele garantizarse especificándola previamente). La c) no es correcta ya que se buscan pruebas en contra de H, que se desea rechazar.

2.2. La respuesta correcta es la a) ya que debe situarse en H aquello que se desea rechazar para poder demostrar su complementario.

2.3. El proceso formal de decisión es el siguiente:

I) Variable: preferencia por A o por B

II) Estadístico: proporción P que prefieren A

III) Hipótesis que se desea rechazar: H: π = 0.5 (ambos fármacos tienen igual preferencia)

Se fija el límite del nivel de significación en p=0.05.

IV) Si H es cierta: $P \sim N(\pi,\ \pi(1-\pi)/n)\ =\ N(0.5, 0.025^2)$

Premisas: muestra grande $\pi \cdot n > 5$ y $(1-\pi) \cdot n > 5$

V) Cálculo del valor p:

$$z = \frac{\pi_p - \pi}{\frac{\pi\ 1-\pi}{n}} = \frac{0.58 - 0.5}{\frac{0.5 \cdot 0.5}{400}} = \frac{0.08}{0.025} = \mathbf{3.2}$$

```
# Prob(z>3.2) (cola superior, lower.tail=FALSE)
> pnorm(q=3.2,lower.tail=FALSE)
[1] 0.000687138
# Multiplicando por 2 (valor bilateral)
> pnorm(q=3.2,lower.tail=FALSE)*2
[1] 0.001374276
```

Por ello, puede rechazarse, con p=0.0014 que ambos tratamientos sean iguales: el tratamiento A es preferido al tratamiento B.

VI) Cálculo del intervalo de confianza:

$$IC_{95\%}\ \pi\ :\ P\ \pm\ z_{\frac{\alpha}{2}} \cdot \sigma_P\ \ \approx\ \ 0.58\ \pm 1.96 \cdot 0.025\ \ \approx\ \ 0.58\ \pm\ 0.05\ \ =\ \ 0.53, 0.63$$

La "auténtica" preferencia π por A se encuentra entre 53% y 63%. Al excluir 50%, IC permite la misma conclusión que PS.

2.4. En caso que P=53%, se tiene:

V) Cálculo del valor p:

$$z = \frac{\pi_p - \pi}{\frac{\pi\ 1-\pi}{n}} = \frac{0.53 - 0.5}{\frac{0.5 \cdot 0.5}{400}} = \frac{0.03}{0.025} = \mathbf{1.2}$$

```
# Prob(z>1.2) (cola superior, lower.tail=FALSE)
> pnorm(q=1.2,lower.tail=FALSE)
[1] 0.1150697
# Multiplicando por 2 (valor bilateral)
```

```
> pnorm(q=1.2,lower.tail=FALSE)*2
[1] 0.2301393
```

Se ha obtenido p=0.23. Nada se opone a aceptar que ambos tratamientos tienen la misma preferencia.

VI) Cálculo del intervalo de confianza:

$$IC_{95\%}\ \pi\ :\ P \pm z_{\frac{\alpha}{2}} \cdot \sigma_P \quad \approx \quad 0.53 \pm 1.96 \cdot 0.025 \quad \approx \quad 0.53 \pm 0.05 \quad = \quad 0.48, 0.58$$

La "auténtica" preferencia π por A se encuentra entre 48% y 58%. El intervalo contiene el valor 0.5, por lo que se llega a la misma conclusión con IC que con PS.

2.5. Las tres primeras son correctas, la cuarta no tiene sentido y las dos últimas son un error habitual de interpretación de p, que cuantifica la probabilidad de unos resultados condicionando a H, no la probabilidad de H condicionando a unos resultados. Más adelante insistimos en esta distinción.

2.6. El estadístico es:

$$\boldsymbol{t} = \frac{y - \mu_H}{\frac{S}{\sqrt{n}}} = \frac{79 - 70}{\frac{12}{\sqrt{16}}} = 3$$

```
# P = Prob(t>3)  con 15 grados de libertad
> pt(q=3,df=15,lower.tail=FALSE)*2
[1] 0.008972737
```

La probabilidad vale p <0.01 y, por tanto, se ha logrado demostrar que μ > 70.

Utilizando el t test de R se obtiene, además, el intervalo de confianza:

```
> library(BSDA)
> tsum.test(mean.x=79, s.x=12, n.x=16, mu=70)
…
t = 3, df = 15, p-value = 0.008973
alternative hypothesis: true mean is not equal to 70
95 percent confidence interval: 72.60565 85.39435
…
```

El intervalo excluye el valor 70, por lo que permite la misma conclusión que PS.

2.7. La correcta es la respuesta c).

2.8. Ahora cambia el nivel p de significación, que al dividirse por 2 da 0.0045, por lo que las conclusiones no cambian. Nótese que un planteamiento unilateral es más adecuado en este ejemplo.

2.9. Ejercicio difícil, ya que 2 es mayor que 1.96 y parece que podemos rechazar H, pero observe que -2 está a la izquierda de +1.96, por lo que se acepta H. Es correcta la respuesta e), ya que se trata de una prueba unilateral cuya H incluye el 0 y todos los valores negativos. Dado que el estadístico se sitúa en H, la única conclusión posible en una prueba de significación es "nada se opone a aceptar H".

3.1. En 2.3, al ser $z_3 = 3.2 > z_{\frac{\alpha}{2}} = 1.96$, se rechaza H.

En 2.4, al ser $z_4 = 1.2 < z_{\frac{\alpha}{2}} = 1.96$, nada se opone a aceptar H.

En 2.7, al ser $t_7 = 3.0 > t_{14,\frac{\alpha}{2}} = 2.145$, se rechaza H.

```
# t14,α/2
> qt(p=0.025,df=14,lower.tail=FALSE)
[1] 2.144787
```

3.2. Son correctas las respuestas c), expresada más formalmente en términos de probabilidad poblacional; y f), como frecuencia a largo plazo.

3.3. Es correcta la respuesta b) ya que H_0 es cierta.

4.1. Efectivamente, las correctas son las tres largas d), h) e I): ¡es peligroso abreviar!

4.2. No, los resultados de su experimento lo único que le dicen es que, asumiendo que los dos productos sean iguales, la probabilidad de obtener unos resultados como los suyos (o más extremos) no es muy pequeña. Por tanto, no puede demostrar que H sea falsa, lo que no equivale a haber demostrado que H sea cierta. Por ello, no puede afirmar que tengan la misma eficacia. Más adelante se estudia cómo puede demostrar equivalencia.

4.3. Todas son ciertas.

4.4. La correcta es la respuesta a).

4.5. Son correctas la a) y la c).

4.6. a) y c) son claramente inferencia, así como d) y e) decisión. b) hace inferencia sobre las consecuencias de una decisión.

5.1. Se rechazaría la H de estricta igualdad en los estudios 1, 4 y 5. Nótese que el estudio 1 tiene un IC, razonablemente estrecho, que le permite concluir tanto equivalencia (porque excluye ε_1 y ε_2) como diferencias (porque excluye 0); es decir, los dos tratamientos no son estrictamente iguales, pero sus diferencias no alcanzan el criterio de relevancia. El estudio 6, en cambio, tiene un IC tan amplio, aporta tan poca información, que no le permite ni rechazar la estricta igualdad ni el límite de relevancia clínica. Los restantes estudios no presentan estas paradojas: el 2 y el 3 no consiguen rechazar la estricta igualdad y sí que consiguen establecer equivalencia (aunque el 2 tiene un IC más estrecho que implica que se dispone de mucha información); y el 4 y el 5 consiguen rechazar la estricta igualdad y no consiguen establecer equivalencia. Nótese que el estudio 5 no incluye el margen de equivalencia ε_2, pero que se sitúa al lado de la no equivalencia (lo que coincide con el planteamiento unilateral).

5.2. $IC_{90\%}$:

$$\mu_D \in x_D \pm t_{\frac{\alpha}{2}} \cdot \frac{S_D}{\sqrt{n}} \rightarrow \mu_D \in 3 \pm t_{19;0.05} \cdot \frac{10}{\sqrt{20}} \rightarrow \mu_D \in 3 \pm 1.729 \cdot \frac{10}{\sqrt{20}} \rightarrow \boldsymbol{\mu_D \in -0.866, 6.866}$$

```
# t19,0.05
> qt(p=0.05,df=19,lower.tail=FALSE)
[1] 1.729133
```

Luego la media de las diferencias entre las presiones de ambos hipotensores está entre –0.866 (el de referencia consigue presiones más bajas en media: gana por 0.866 mmHg) y +6.866 (el nuevo consigue presiones más

bajas: gana por 6.866 mmHg). Luego la diferencia entre ambos fármacos está entre los límites –10 y +10: se ha establecido equivalencia.

5.3. PS : PDU $\quad H_A: \mu_D \leq -10 \quad t_1 = (3-(-10)) / (10/\sqrt{20}) = 5.814 \quad \rightarrow p < 0.001$

$H_B: \mu_D \geq 10 \quad t_2 = (3- 10) / (10/\sqrt{20}) = -3.130 \quad \rightarrow p \approx 0.003$

```
# p = Prob(t>5.814)  con 19 grados de libertad
> pt(q=5.814,df=19,lower.tail=FALSE)
[1] 6.677505e-06
# P = Prob(t<-3.130) con 19 grados de libertad
> pt(q=-3.130,df=19,lower.tail=TRUE)
[1] 0.002756741
```

La primera prueba permite afirmar que la media de las diferencias entre ambos hipotensores está por encima de –10. Y la segunda que está por debajo de +10. Por tanto, con un nivel de significación $P \cong 0.003$, se ha establecido que la media de las diferencias de ambos hipotensores está entre –10 y +10.

CH: $PDU_{\alpha=0.05}$: $\begin{cases} H_{0A}: \mu_D \leq -10 \\ H_{1A}: \mu_D > -10 \end{cases} \quad t_1 = (3-(-10)) / (10/\sqrt{20}) = 5.814$

$\begin{cases} H_{0B}: \mu_D \geq 10 \\ H_{1B}: \mu_D < 10 \end{cases} \quad t_2 = (3- 10) / (10/\sqrt{20}) = -3.130$

Se llega a la misma conclusión, pero ahora se dirá que, con riesgo $\alpha=0.05$, se autoriza el nuevo.

5.4. $IC_{90\%}$:

$$\mu_D \in x_D \pm t_{\frac{\alpha}{2}} \cdot \frac{S_D}{\sqrt{n}} \rightarrow \mu_D \in 3 \pm t_{4;0.05} \cdot \frac{10}{\sqrt{5}} \rightarrow \mu_D \in 3 \pm 2.132 \cdot \frac{10}{\sqrt{5}} \rightarrow \boldsymbol{\mu_D \in -6.534, 12.534}$$

```
# t4,0.05
> qt(p=0.05,df=4,lower.tail=FALSE)
[1] 2.131847
```

Ahora, el $IC_{90\%}$ sobrepasa el dintel superior que marca la no equivalencia y, por tanto, no se puede defender que haya equivalencia.

CH: $PDU_{\alpha=0.05}$: $\begin{cases} H_{0A}: \mu_D \leq -10 \\ H_{1A}: \mu_D > -10 \end{cases} \quad t_1 = (3-(-10)) / (10/\sqrt{5}) = 2.907 \quad \rightarrow p = 0.022 < 0.05 = \alpha$

$\begin{cases} H_{0B}: \mu_D \geq 10 \\ H_{1B}: \mu_D < 10 \end{cases} \quad t_2 = (3- 10) / (10/\sqrt{5}) = -1.565 \quad \rightarrow p \approx 0.096 > 0.05 = \alpha$

```
# p = Prob(t>2.907)  con 4 grados de libertad
> pt(q=2.907,df=4,lower.tail=FALSE)
[1] 0.02190478
# p = Prob(t<-1.565) con 4 grados de libertad
> pt(q=-1.565,df=4,lower.tail=TRUE)
[1] 0.0963175
```

Asimismo, aunque la primera prueba aún permite afirmar que la media de las diferencias está por encima de –10, la segunda no ha permitido establecer que esté por debajo de +10. Por tanto, no se ha podido demostrar que la media de las diferencias de ambos hipotensores esté entre –10 y +10. Por tanto, la acción debe ser no autorizar el nuevo (A_0).

Tabla salvadora

Tabla resumen de las pruebas de hipótesis vistas con el estadístico del test, su distribución (bajo H_0), sus premisas necesarias, el criterio de decisión y la función de R.

Parámetro	Hipótesis nula	Estadístico	Premisas	Distribución si H0 cierta	Criterio de decisión (riesgo α bilateral)	Función en R
$\boldsymbol{\pi}$	$H_o: \pi = \pi_H$	$z = \frac{P - \pi_H}{\sqrt{\frac{\pi_H\ 1 - \pi_H}{n}}}$	$\pi \cdot n > 5$ *y* $(1-\pi) \cdot n > 5$	$z \sim N\ 0{,}1$	*Rechazar* H_0 *si* $z\ >\ z_{\frac{\alpha}{2}}$	*prop.test*
$\boldsymbol{\mu}$	$H_o: \mu = \mu_H$	$t\ = \frac{y - \mu_H}{S\ \sqrt{n}}$	*Y ~ Normal*	$t \sim t_{n-1}$	*Rechazar* H_0 *si* $t\ >\ t_{n-1;\frac{\alpha}{2}}$	*t.test* *tsum.test* *(BSDA)*
$\boldsymbol{\mu}$	$H_o: \mu_1 = \mu_2$	$t = \frac{y_1 - y_2}{S \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ $S^2 = \frac{n_1 - 1\ \cdot S_1^2 + \ n_2 - 1\ \cdot S_2^2}{n_1 + n_2 - 2}$	$Y_1, Y_2 \sim Normal$ $\sigma_1 = \sigma_2$	$t \sim t_{n_1+n_2-2}$	*Rechazar* H_0 *si* $t\ >\ t_{n_1+n_2-2;\frac{\alpha}{2}}$	*t.test*