CONTRASTE DE HIPÓTESIS

Esquema general: hipótesis nula y alternativa, errores tipo I y II, estadístico de contraste y región crítica. Riesgos alfa y beta. Tipos de contrastes. Prueba para la proporción de una población. Prueba para la media de una población. Potencia de un contraste. Grado de significación (P-valor).

INTRODUCCIÓN

"Hypothesis testing provides an objective frame work for making decisions using probabilistic methods, rather than relying on subjective impressions" – B Rosner, Fundamentals of Biostatistics, Brooks/Cole 2006 (6ªed)

A. Definiciones (Diccionario de la Lengua Española – R.A.E.)[1]

- Contrastar. Del lat. tardío contrastāre 'oponerse'.
  3. tr. Comprobar la exactitud o autenticidad de algo.
- Hipótesis. Del gr. ὑπόθεσις hypóthesis.
  1. f. Suposición de algo posible o imposible para sacar de ello una consecuencia.

  → Hipótesis de trabajo. Hipótesis que se establece provisionalmente como base de una investigación que puede confirmar o negar la validez de aquella.
- Refutar. Del lat. refutāre.
  1. tr. Contradecir o impugnar con argumentos o razones lo que otros dicen.

B. La Medicina Basada en la Evidencia (EBM)[2] aboga para que la decisión de aplicar un tratamiento a un paciente se apoye en los beneficios contrastados objetivamente. La evaluación de la evidencia científica tiene que hacerse mediante la aplicación de pruebas imparciales y fiables. El uso de fármacos como la talidomida o la aplicación de la terapia de reemplazo hormonal son ejemplos de que lo nuevo no es siempre mejor y que los efectos deseados no siempre se materializan[3]. Esta noticia de prensa es otro ejemplo reciente:

> Varapalo al controvertido fármaco de 40.000 euros contra el alzhéimer
> Un grupo de expertos pone en duda la eficacia del aducanumab, un carísimo tratamiento que podría ralentizar ligeramente la enfermedad
> EL PAIS | M. Ansede 06 NOV 2020 - 23:43 CET

[1] <http://dle.rae.es/>; acceso 19/10/2017
[2] Sackett DL et al. Evidence based medicine: what it is and what it isn't. BMJ 1996;312:71.
[3] Evans I, Thornton H, Chalmers I and Glasziou P (2011). Testing Treatments, 2nd Edition; London: Pinter and Martin.

C. El método científico como combinación de razonamiento y observación hace uso de la lógica inductiva para extraer conclusiones de resultados experimentales. En particular se aplica la regla de inferencia lógica *modus tollendo tollens.* Por ejemplo[4]:

Si P→Q	Si el agua hierve, entonces soltará vapor.
noQ:	No suelta vapor.
⇒ noP	Por lo tanto, no está hirviendo el agua.

D. En el contexto de la investigación médica la inferencia estadística aporta la tecnología para evaluar las conclusiones en términos de probabilidad[5].

E. La INFERENCIA ESTADÍSTICA se define como el conjunto de métodos que permiten mediante inducción llegar a conclusiones sobre propiedades de una población a partir las observaciones recogidas en una muestra[6]. Las preguntas que se intentan responder parten de dos planteamietos:

- ESTIMACIÓN: Determinar un valor desconocido
  Ejemplo: *¿Cuánto vale el peso medio en la población de estudio?*
  Respuesta: 67k ± 8.6k. Con una confianza del 95% se afirma que el peso medio puede ser cualquier valor entre 58.4 y 75.6
- PRUEBA de HIPÓTESIS: Tomar una decisión
  Ejemplo: *El peso medio poblacional vale 54k, ¿Falso o no?*
  Respuesta: P-valor = 0,0034 < 0,05. Es falso, porque los datos no son compatibles con 54k

ESQUEMA DE UNA PRUEBA DE HIPÓTESIS

A. Un CONTRASTE o prueba de hipótesis es un conjunto de reglas para tomar una decisión acerca de una hipótesis, falsa o no falsa, en base a una probabilidad.

B. Las etapas a seguir son:

→ Plantear una hipótesis y utilizarla como premisa
→Deducir de lo anterior una situación esperable
→Usar lo observado en los datos como prueba
→Cuantificar la discrepancia entre lo observado y lo esperado: Obs vs Esp
→Tomar una decisión a favor o en contra de la hipótesis

C. El punto de partida es plantear la HIPÓTESIS sobre la que se tomará la decisión. Será una afirmación acerca de una característica de la población de estudio. Generalmente una hipótesis estadística hace referencia a un parámetro.
Ejemplos:

H: "La altura media de la población de jóvenes españoles es de 174 cm"

[4] <https://es.wikipedia.org/wiki/Modus_tollendo_tollens>; acceso 27/10/2017
[5] Wassertheil-Smoller S. (1990). Ver en bibliografía.
[6] Spiegel MR (2011) Statistics. Shaum's easy outlines. McGrawHill. 2ª ed

H: "La proporción de curaciones usando T en una población de pacientes es del 40%"
H: "La prevalencia de caries en niños de ciudad es igual a la de niños rurales"

D. La DECISIÓN sobre la hipótesis sólo puede ser de dos tipos: se rechaza la hipótesis nula o no se rechaza. Se basa en evaluar los datos de la muestra seleccionada de la población de estudio. Si la información de los datos es compatible con lo planteado en la hipótesis, entonces ésta no se puede rechazar. Sin embargo, la hipótesis quedará refutada, si la información apunta a otra idea.

E. La hipótesis sobre la que se decide ha de ser una hipótesis NULA, en el sentido de que no hay cambio o efecto respecto a una situación aceptada. Un aforismo del derecho dice: «lo normal se entiende que está probado, lo anormal se prueba»[7] . En el mundo académico se prueba lo anormal refutando la hipótesis nula. Por eso, si se pone en duda el comportamiento de una moneda, la hipótesis nula o Ho ha de afirmar que "la moneda no está trucada":

Ho: "La moneda no está trucada" o P(cara) = 0.5

F. Acompañando a la hipótesis nula también se define una hipótesis ALTERNATIVA, que se simboliza por Ha. Será la afirmación que se adoptará en caso de rechazar la nula. En el ejemplo de la moneda sería declarar "la moneda está trucada".

Ho: "La moneda no está trucada" o P(cara) = 0.5
Ha: "La moneda está trucada" o P(cara) $\neq 0.5$

Generalmente es de tipo compuesto, pues será un conjunto de posibilidades no incluidas en la hipótesis nula. Por ejemplo:

Ha: "La altura media de la población de jóvenes españoles es < ó > 174 cm"

Se dice que un contraste es BILATERAL cuando la hipótesis alternativa incluye cualquier sentido, como en el ejemplo anterior, mientras que se dice que UNILATERAL si la afirmación es en sentido único. Por ejemplo:

Ha: "La altura media de la población de jóvenes españoles es > 174 cm"

G. La información se extraerá de los datos aplicando el ESTADÍSTICO adecuado a cada caso. Ejemplos:

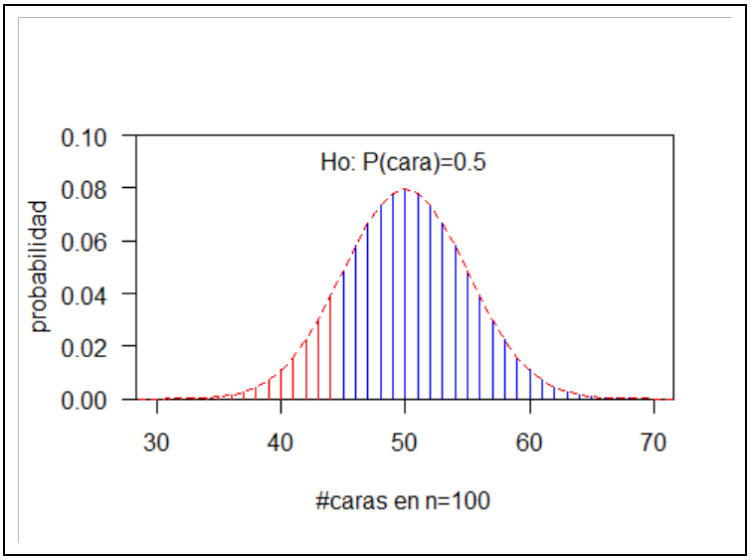
- La media aritmética de una muestra de 56 jóvenes es de 173,5 cm
- La frecuencia relativa de curaciones en la muestra de 120 pacientes es de 0,45

H. Puesto que los datos de la muestra se obtienen por un proceso aleatorio, las estimaciones estarán afectadas de incertidumbre, es decir, error de muestreo. Es por ello que, aun siendo verdadera la hipótesis, la realidad observada no se ajustará perfectamente a lo esperado. En el ejemplo de la moneda sospechosa los datos se obtendrán de repetir varias veces el lanzamiento de la moneda y difícilmente lo observado será mitad caras y mitad cruces por muy perfecta que sea la moneda. Lo esperado será 50 caras en 100 lanzamientos, pero 45 caras también sería aceptable.

I. La discrepancia entre lo esperado y lo observado se medirá en términos de PROBABILIDAD. Los datos se considerarán incompatibles con la hipótesis sólo si la

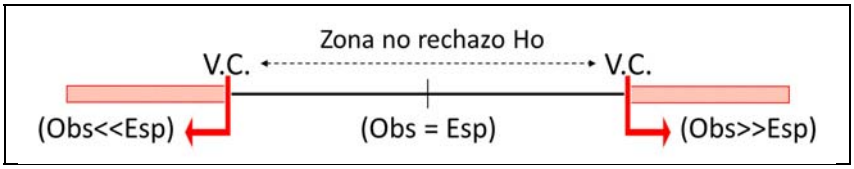
[7] <https://es.wikipedia.org/wiki/Onus_probandi>; acceso: 27/10/2017

probabilidad es baja. Si se asume que la moneda no está trucada, entonces aplicamos p(cara)=0.5 para deducir que lo esperado en 100 lanzamientos son 50 caras. Observar 28 caras o menos en 100 lanzamientos tiene una probabilidad 0,00000629, lo cual es demasiado baja para seguir apoyando la idea de que no hay truco. Sin embargo, la probabilidad de 44 caras o menos en 100 lanzamientos es de 0,13562651, por lo que no se podría rechazar la hipótesis inicial.



J. El cálculo de la probabilidad se hace utilizando la DISTRIBUCIÓN MUESTRAL del estadístico asumiendo que la hipótesis nula es cierta. Esto significa que los parámetros de la distribución quedan marcados por lo declarado en la hipótesis nula y que la probabilidad calculada estará condicionada a ello.

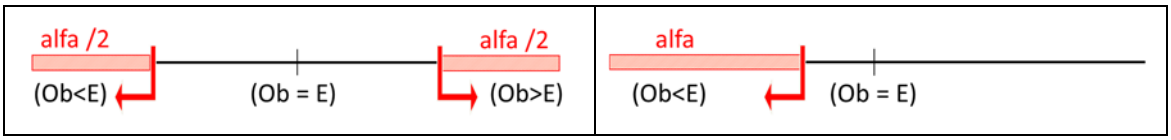
K. La regla de decisión se construirá fijando a priori un nivel de probabilidad. Si la probabilidad calculada es menor, entonces se rechazará la hipótesis nula. Habitualmente la probabilidad se expresa por sus correspondientes percentiles que se establecen como VALORES CRÍTICOS (V.C.). Su función es marcar lo máximo que puede diferir lo observado en la muestra con lo esperado bajo la hipótesis nula.



En un contraste bilateral habrá dos valores críticos, uno para detectar que lo observado es menor que lo esperado y otro para detectar la situación contraria. Siguiendo con el ejemplo de la moneda significaría delimitar cuántas caras por debajo o por encima de 50 es razonable esperar de una moneda no trucada. En contraste unilateral sólo habrá un valor crítico que se situará en el lado que indique la hipótesis alternativa.

L. Se denomina REGIÓN CRÍTICA al conjunto de valores que están más allá de los valores críticos. Corresponde con los resultados muy poco probables en caso de que la hipótesis nula fuera cierta. El tamaño de la región crítica viene impuesto por el nivel de probabilidad prefijado, que se denomina NIVEL DE SIGNIFICACIÓN o ALFA. Lógicamente la región crítica ocupa las colas de la distribución muestral del estadístico

usado en el contraste. Si es de tipo bilateral, el valor de alfa se reparte entre los dos extremos, pero si es unilateral entonces se concentra en el lado que indique la hipótesis alternativa.



M. La decisión consiste en seleccionar una opción de dos que son contrapuestas. Cualquiera que sea la decisión, la opción elegida puede ser correcta o no, ya que se basa en probabilidad. Así, por ejemplo, si la moneda es perfecta, el resultado de 28 caras en 100 lanzamientos es posible, aunque con una probabilidad muy pequeña. Se distinguen dos tipos de ERROR:

- Error de tipo I. Rechazar una hipótesis nula que es verdadera
- Error de tipo II: No rechazar una hipótesis nula que no es verdadera

Estos dos errores no pueden ocurrir simultáneamente, pues están condicionados a decisiones distintas. La situación general se puede ilustrar con un diagrama de árbol de cuatro ramas, pero normalmente se presenta en forma de tabla como la siguiente.

DECISIÓN	Ho Verdadera	Ho No-Verdadera
$\Rightarrow$ Se rechaza Ho	Error de tipo I α = Pr(Error I)	OK
$\Rightarrow$ No se rechaza Ho	OK	Error tipo II β = Pr(Error II)

N. A pesar de que siempre existirá el RIESGO de tomar la decisión errónea, el contraste se ha de construir para que esta probabilidad sea mínima. Se trata de que el error ocurra el menor número de veces posibles. Hay que distinguir dos tipos de riesgo:

- Riesgo alfa. Probabilidad del error tipo I.

$$\alpha = Pr(error\ I) = Pr(r\ H_0 | H_0\ verdadera)$$

- Riesgo beta. Probabilidad del error de tipo II.

$$\beta = Pr(error\ II) = Pr(no - r\ H_0 | H_0\ no - verdadera)$$

El riesgo alfa se fija antes del contraste y determina el tamaño de la región crítica, por eso también se denomina nivel de significación. Si se adopta un alfa grande, los valores críticos se acercarán al centro reduciendo la zona de no rechazo. En consecuencia, se favorece el rechazo de la hipótesis. Por el mismo razonamiento, un valor pequeño de alfa provocará que los valores críticos se desplacen hacia los extremos y dificultará el rechazo de la hipótesis. Esto posibilita el error de tipo II y aumenta el riesgo beta. Los riesgos alfa y beta están relacionados, aunque no son complementarios.

TIPOS DE CONTRASTES

A. Los distintos contrastes de hipótesis suelen clasificarse en tres grandes tipos:

- CONFORMIDAD ("one-sample inference"[8]). La hipótesis hace una afirmación sobre una característica de una población y se contrasta con una muestra de datos obtenida de ella.

  H: "La altura media de la población de jóvenes españoles es de 174 cm"
- HOMOGENEIDAD ("two-sample inference"). La hipótesis hace una afirmación comparando una característica entre dos o más poblaciones y se contrasta comparando sendas muestras obtenidas de ellas.

  H: "La prevalencia de caries en niños de ciudad es igual a la de niños rurales"
- INDEPENDENCIA. La hipótesis afirma que dos variables estudiadas en una misma población no están asociadas entre ellas y se contrasta con una muestra obtenida de ella.

  H: "Las notas de Anatomía son independientes de las notas de Bioquímica en la población de alumnos de primero"

B. Para resolver un contraste de hipótesis hay cuatro pasos básicos a seguir:

1) Definir las hipótesis.

   Esto implica entender bien el tipo de datos para identificar el parámetro en cuestión, el número de poblaciones y muestras estudiadas y si el contraste ha de ser unilateral o bilateral. Al final se formularán las hipótesis nula y alternativa correspondientes.

2) Estadístico de contraste.

   Para cuantificar la diferencia entre lo establecido en la hipótesis y lo observado en la muestra se escogerá el estadístico adecuado al tipo de datos a analizar. Para asignar probabilidad a la diferencia encontrada se usará la distribución muestral del estadístico, según lo indicado en la hipótesis nula, siempre y cuando se cumplan las condiciones de aplicación. La región crítica se situará en las colas de la distribución y su tamaño dependerá de alfa. Con esto quedará constituida la regla de decisión que se aplicará más tarde.

3) Cálculos.

   La ejecución de los cálculos para obtener el estadístico de contraste podrá hacerse a mano o con programas informáticos. Para valorar el resultado se consultarán las tablas para encontrar los valores críticos, o de forma alternativa se usará el P-valor calculado por el programa informático.

4) Decisión y Conclusión.

   En primer lugar, hay que decidir si se rechaza o no la hipótesis nula en coherencia con la regla de decisión anterior. Además, habrá que considerar el posible error asociado. Después se concluye con las consecuencias que se derivan de la decisión, que tienen más que ver con el contexto que con la estadística.

[8] Rosner B. (2011). Ver bibliografía.

CONFORMIDAD DE UNA PROPORCIÓN

A. Objetivo. Comprobar si los datos aportan o no evidencia para responder a:

¿Está usted conforme en afirmar que la proporción poblacional vale ….?

B. La hipótesis nula de una conformidad de proporción tiene la forma:

Ho : π=número p.ej. Ho: π=0,4

C. El parámetro proporción poblacional (π) se interpreta como una probabilidad de éxito en experiencias con resultados dicotómicos

Ejemplos:

E= *lanzar al aire una moneda y observar el resultado*

X= {cara: 1, cruz: 0}

π = Pr(cara)

E= *seleccionar al azar un individuo de la población y observar el resultado*

X= {diabético: 1, no-diabético: 0}

π = Pr(diabético)

D. Se supone que los datos se obtienen seleccionando n individuos de la población por muestreo aleatorio simple. En cada uno de ellos se mide la presencia o ausencia del atributo que se define como éxito. Las observaciones formarán una colección de ceros y unos.

E. El mejor estimador de la proporción poblacional es la frecuencia relativa de éxitos en la muestra (F.R.), también conocida como proporción muestral:

$$FR = \frac{\#exitos}{n} = \frac{FA}{n} \quad \rightarrow \quad P(\acute{e}xito) = \pi$$

F. La distribución muestral de FR depende principalmente del numerador, frecuencia absoluta, ya que el denominador es el tamaño de la muestra y es una constante.

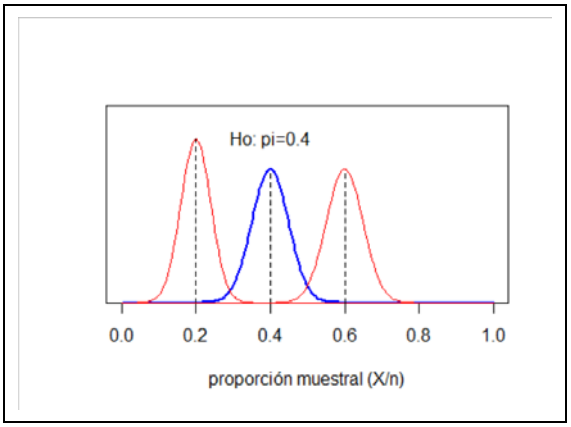
$$FA = \#exitos \ \sim \ Binomial(n, \pi)$$

En muestras grandes el modelo Binomial se puede aproximar por una Normal.

$$FA = \#exitos \ \xrightarrow{n\ grande} \ \sim \ Normal\left(n\pi, \sqrt{n\pi(1-\pi)}\right)$$

Por tanto, la distribución de FR en muestras grandes, y probabilidad de éxito no extrema, es:

$$FR = \frac{FA}{n} \ \xrightarrow{n\ grande} \ \sim \ Normal\left(\pi, \sqrt{[\pi(1-\pi)/n]}\right)$$



G. La distribución muestral de FR bajo Ho, es decir, asumiendo cierto lo que en ella se afirma, aparece al sustituir el parámetro por lo declarado en la hipótesis:

$$H_0: \pi = 0{,}4 \qquad FR \xrightarrow{n\ grande} \sim Normal\left(0{,}4, \sqrt{[0{,}4 * 0{,}6/n]}\right)$$

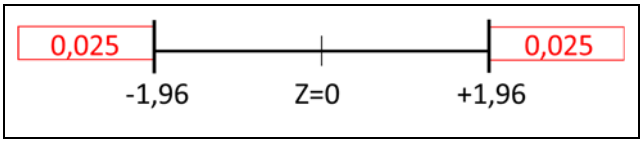
H. Para trabajar con probabilidades de una Normal lo más cómodo es tipificar el estadístico para convertirlo en una Zeta:

$$FR \implies \frac{FR - \pi}{\sqrt{[\pi(1-\pi)/n]}} \longrightarrow H_o: \pi = 0{,}4 \rightarrow \frac{FR - 0{,}4}{\sqrt{[0{,}4(1-0{,}4)/n]}} = z \sim N(0{,}1)$$

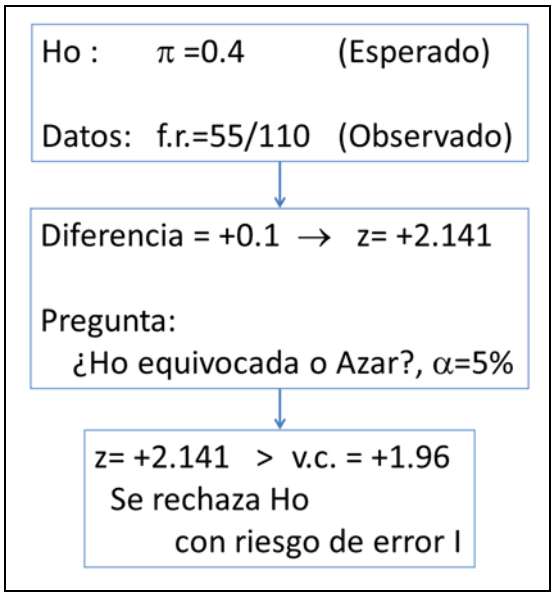
I. La decisión sobre Ho se basará en comparar la frecuencia relativa observada en la muestra (fr) con lo establecido en la hipótesis (π=0,4). No hay que entenderlo como una comparación matemática, porque hay que tener en cuenta la incertidumbre de muestreo:

Si fr se acerca a 0,4	→	z estará en el centro	→ No se rechaza Ho
Si fr se aleja de 0,4	→	z estará en los extremos	→ Sí se rechaza Ho

J. En torno al centro de la distribución, z=0, se sitúa la zona de no rechazo y más allá de los valores críticos está la zona de rechazo. La región crítica estará en los extremos, porque contiene valores poco probables siendo cierta la hipótesis nula. El tamaño de la región crítica vendrá determinado por el riesgo alfa que implica rechazar la hipótesis nula. En un contraste bilateral la región crítica se reparte en las dos colas y los valores críticos de la z son -1,96 y +1,96 para alfa=5%.



K. Esquema de resolución con ejemplo



CONFORMIDAD DE UNA MEDIA

A. Objetivo. Comprobar si los datos aportan o no evidencia para responder a:
¿Está usted conforme en afirmar que la media poblacional vale ….?

B. La hipótesis nula de una conformidad de media tiene la forma:
Ho : μ=número p.ej. Ho: μ=174

C. El parámetro media poblacional (μ) se refiere al valor medio de una variable continua medida en una población determinada.
Ejemplo:
E= *seleccionar al azar una persona de una población de jóvenes varones*
X= altura (cm)
μ= valor medio de altura en la población

D. Se supone que los datos se obtienen seleccionando n individuos de la población por muestreo aleatorio simple. En cada uno de ellos se mide la característica X. Las observaciones formarán una colección de números.

E. El mejor estimador de la media poblacional es la media aritmética de la muestra:

$$\bar{X}_n = \frac{1}{n}\sum x_i \;\rightarrow\; \mu$$

F. La distribución muestral de $\bar{X}$ será Normal si:
- la variable X original es Normal (teorema de la adición)
- la variable X original no es Normal, pero el tamaño de muestra es muy grande (teorema del límite central)

El centro de la distribución está en μ y el error típico depende de n:

$$\bar{X} \sim Normal\left(\mu, \sigma/\sqrt{n}\right)$$

G. La distribución muestral de $\bar{X}$ bajo Ho, es decir, asumiendo cierto lo que en ella se afirma, aparece al sustituir el parámetro por lo declarado en la hipótesis:

$$H_0: \mu = 174 \qquad \bar{X} \sim Normal\left(174, \sigma/\sqrt{n}\right)$$

H. Para trabajar con probabilidades de una Normal lo más cómodo es transformar el estadístico. Si se conoce el valor de la variancia poblacional (σ), la transformación es una tipificación que genera una Zeta:

$$\bar{X} \implies \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \longrightarrow H_o: \mu = 174 \rightarrow \frac{\bar{X}-174}{\sigma/\sqrt{n}} = z \sim N(0,1)$$

Generalmente sigma es desconocida y habrá que estimarla a partir de la muestra. La transformación de $\bar{X}$ con S da lugar al estadístico t que sigue una distribución t de Student con n-1 grados de libertad:

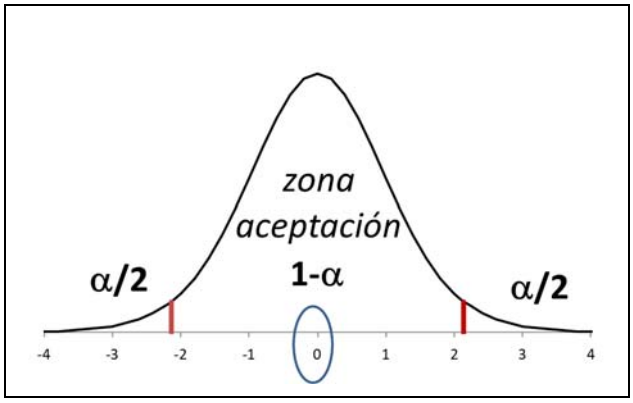
$$\bar{X} \implies \frac{\bar{X}-\mu}{S/\sqrt{n}} \longrightarrow H_o: \mu = 174 \rightarrow \frac{\bar{X}-174}{S/\sqrt{n}} = t \sim t - Student(gl = n-1)$$

I. La decisión sobre Ho se basará en comparar la media muestral ($\bar{x}$) con lo establecido en la hipótesis (μ=174). No hay que entenderlo como una comparación matemática, porque hay que tener en cuenta la incertidumbre de muestreo:

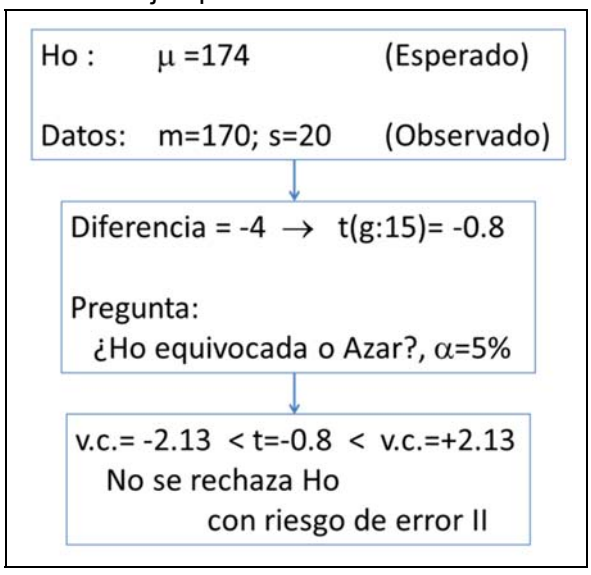
Si $\bar{x}$ se acerca a 174 $\rightarrow$ t estará en el centro $\rightarrow$ No se rechaza Ho

Si $\bar{x}$ se aleja de 174 $\rightarrow$ t estará en los extremos $\rightarrow$ Sí se rechaza Ho

J. En torno al centro de la distribución, t=0, se sitúa la zona de no rechazo y más allá de los valores críticos está la zona de rechazo. La región crítica estará en los extremos, porque contiene valores poco probables siendo cierta la hipótesis nula. El tamaño de la región crítica vendrá determinado por el riesgo alfa que implica rechazar la hipótesis nula. En un contraste bilateral la región crítica se reparte en las dos colas y los valores críticos de la t dependerán de los grados de libertad. Para n=16, es decir 15 grados de libertad, los valores críticos son -2,131 y +2,131 para alfa=5%.



K. Esquema de resolución con ejemplo



POTENCIA

A. La potencia de una prueba cuantifica la capacidad para detectar la falsedad de una hipótesis nula. Se calcula como probabilidad condicionada:

$$Potencia = P(rechazar\ Ho \mid Ho\ es\ falsa)$$

B. La potencia es complementaria del riesgo beta, que es la probabilidad de cometer error de tipo II:

$$Potencia = 1 - \beta = 1 - P(NO\ rechazar\ Ho \mid Ho\ es\ falsa)$$

Por tanto, poca potencia significa mucha posibilidad de error.

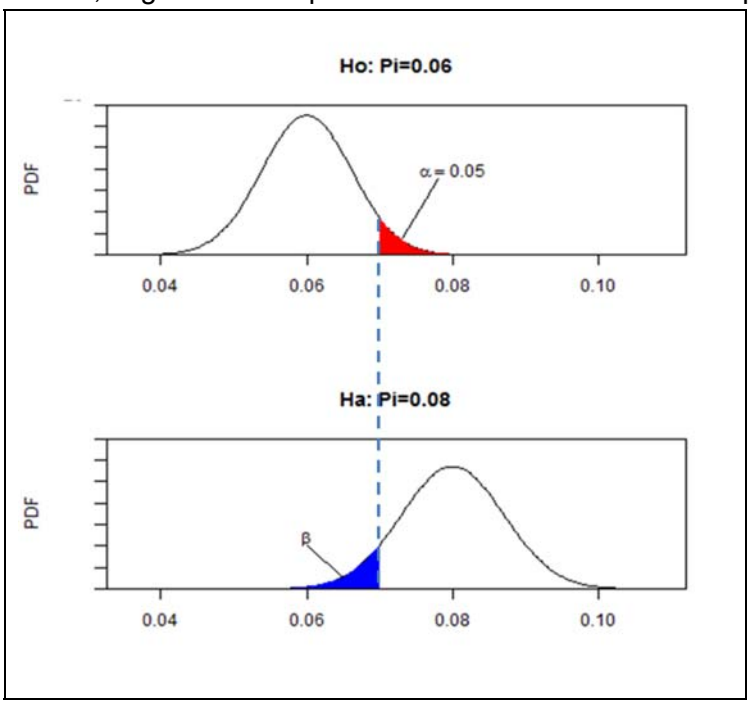
C. Afirmar que Ho es falsa puede significar muchas opciones distintas. Para hacer cálculos de potencia la hipótesis alternativa se define de forma simple fijando un valor alternativo al de la hipótesis nula. Ejemplo:

$$H_0: \pi = 0{,}6 \quad vs \quad H_a: \pi = 0{,}8$$

A la diferencia entre ambas se le denomina delta o tamaño del efecto.

$$Delta = \pi(H_a) - \pi(H_0) = 0{,}8 - 0{,}6 = 0{,}2$$

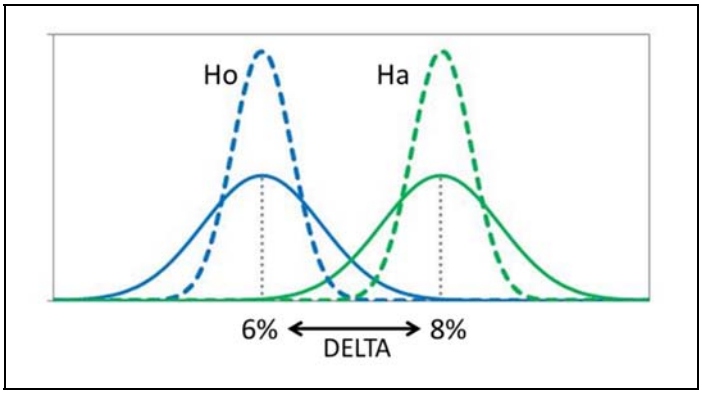
D. Puesto que el riesgo alfa y el riesgo beta están relacionados, cambiar alfa afectará a la potencia. En concreto aumentar alfa, disminuirá beta y por tanto aumentará la potencia. No obstante, la ganancia en potencia es a costa del error de tipo I.



E. Otra manera de aumentar la potencia es definiendo una hipótesis alternativa más alejada de la nula. A mayor delta, menos se solapan las dos curvas y mayor será la potencia. Sin embargo, esto afecta a la resolución del contraste pues limitará la posibilidad de detectar cambios pequeños del parámetro.

F. Sin modificar alfa ni delta, también se puede aumentar la potencia aumentando el tamaño de la muestra. Esto es así porque disminuye el error típico de la distribución muestral, con lo cual las curvas se concentran y hace que sea más fácil discriminar

entre la hipótesis nula y la alternativa. La contrapartida es el gasto que implica trabajar con una muestra más numerosa.



P-VALOR o GRADO DE SIGNIFICACIÓN

A. "Informally, a p-value is the probability under a specified statistical model that a statistical summary of the data (e.g. the sample mean difference between two compared groups) would be equal to or more extreme than its observed value"[9]

B. Cálculo del p-valor en un contraste de conformidad unilateral del parámetro media poblacional.

   Ejemplo:

   $$H_o\colon \mu = 120 \qquad vs \qquad H_a\colon \mu > 120$$

   El modelo estadístico especificado en este problema quedará definido por el uso de la media aritmética como estadístico para resumir los datos. Si se reúnen las condiciones de aplicación para suponer que $\bar{X}$ tiene distribución normal y además se usa el valor de la hipótesis nula para centrarla, entonces el modelo será:

   $$\bar{X} \sim Normal\left(\mu_{Ho}, \sigma/\sqrt{n}\right) = Normal\left(120, \sigma/\sqrt{n}\right)$$

   El valor observado del estadístico será la media aritmética calculada con los datos de la muestra, p.ej. n=9

   $$\bar{X}\{(x_1, x_2, \cdots, x_9)\} = 132$$

   Por tanto, el P-valor corresponde a la probabilidad de que el resultado sea 132 o más extremo bajo una campana de Gauss centrada en 120, para lo cual hay que integrar el modelo desde 132 hasta más infinito:

   $$P - valor(132) = P(\bar{X} \geq 132)$$

   El cálculo de la probabilidad suele hacerse tipificando previamente la $\bar{X}$ para pasar a una variable Zeta, suponiendo sigma conocida, p.ej. σ =12

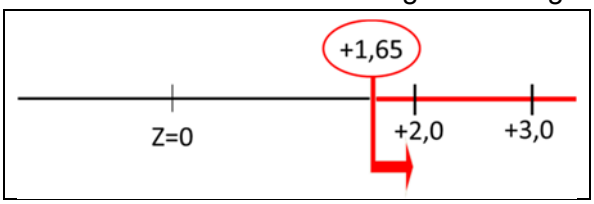
   $$P - valor(132) = P(\bar{X} \geq 132) = P\left(Z \geq \frac{132 - 120}{4} = +3\right) = 0.00135$$

[9] Wasserstein RL, Lazar NA. The ASA's Statement on p-Values: Context, Process, and Purpose. The American Statistician, 70:2, 129-133.

C. El P-valor se usa como criterio para decidir sobre la hipótesis nula de forma equivalente al uso de los valores críticos. Si el P-valor es menor que el riesgo alfa fijado, entonces se rechaza Ho porque eso significa que el resultado del estadístico ha traspasado los valores críticos y ha caído dentro de la región crítica. En el ejemplo anterior la zeta crítica asociada a un alfa del 5% es +1.645 y consecuentemente el P-valor de un resultado z=+3 es menor que alfa.

P-valor < alfa $\Rightarrow$ resultado en la R.C. $\Rightarrow$ Se rechaza Ho

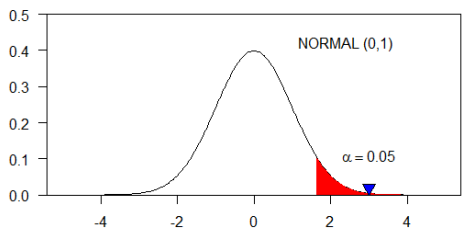
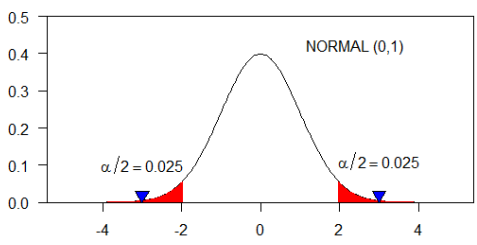
D. En estadística se dice que "el resultado es significativo" cuando los datos aportan evidencia para rechazar la hipótesis nula. En el ejemplo anterior el resultado $\bar{x} = 132$ ó z=+3 es significativo, pero también lo hubiera sido un resultado de $\bar{x} = 128$ ó z=+2. El P-valor pone de manifiesto la diferencia entre ambos casos pues refleja a qué profundidad de la región crítica ha caído el resultado. El resultado z=+3 tiene un P-valor menor y por ello se califica de más significativo. Esta interpretación justifica que al P-valor también se lo denomine "grado de significación".



E. Interpretar lo que es "más extremo" requiere tener en cuenta el sentido de la desigualdad indicado en la hipótesis alternativa. En el ejemplo anterior se ha de entender que más extremo es por encima del resultado, porque en la hipótesis alternativa se apunta hacia más infinito. Sin embargo, en un contraste bilateral la hipótesis alternativa apunta en ambos sentidos, porque no importa si el resultado se desvía por la derecha de cero o por la izquierda. El cálculo del P-valor ha de tener en cuenta ambas posibilidades:

$$|(Obs - Esp)| = |132 - 120| = |12|$$

$$P - valor(|12|) = P(Z \leq -3) + P(Z \geq +3) = 2 * (0.00135)$$

Unilateral (alfa=5%) vc=+1.645 P-valor(+3)=0.00135	Bilateral (alfa=5%) vc=-1.96; +1.96 P-valor(|3|)=0.0027
	

F. Nota histórica: Los primeros usos de la probabilidad para valorar si unos datos observados apoyan o contradicen una hipótesis aparecen ya en el siglo XVIII[10]. La introducción formal del concepto de P-valor ocurre a principios del siglo XX. Se debe a K. Pearson quien lo aplicó en sus trabajos para desarrollar la prueba ji-cuadrado. Después fue R. Fisher quien popularizó su uso a todas las pruebas de significación. A esto contribuyó que lo incluyera en su libro "Statistical Methods for Research Workers" (1925) que es una obra clásica de la estadística. También se debe a R. Fisher la convención de comparar el P-valor con un riesgo alfa del 5%, lo que significa asumir una decisión errónea de tipo I de cada 20 intentos.

G. El uso de ordenadores para realizar los cálculos estadísticos ha facilitado que todos los análisis vayan acompañados de los P-valores correspondientes. Por otro lado, las revistas científicas han promovido durante muchos años que los resultados se valoraran en función del grado de significación. En consecuencia, se ha producido una focalización en el uso del P-valor que ha tenido efectos contraproducentes. Son muchos los estudios que señalan el mal uso y la mala interpretación de este concepto. En 1999 S. Goodman[11] acuñó el término "p-value fallacy" para denunciar esta influencia en la interpretación de los datos de investigación médica, y en 2008[12] publicó una lista de doce interpretaciones incorrectas. La primera de ellas, por ser la más ubicua y perniciosa, es:
Misconception #1: If P.05, the null hypothesis has only a 5% chance of being true.
Entrado ya el siglo XXI, la Asociación Americana de Estadística (ASA en inglés) acogió un debate sobre esta cuestión que acabó en una declaración que se publicó bajo el nombre "The ASA's Statement on p-Values: context, process and purpose"[13]. La conclusión se reproduce completa a continuación:
"Good statistical practice, as an essential component of good scientific practice, emphasizes principles of good study design and conduct, a variety of numerical and graphical summaries of data, understanding of the phenomenon under study, interpretation of results in context, complete reporting and proper logical and quantitative understanding of what data summaries mean. No single index should substitute for scientific reasoning".

BIBLIOGRAFÍA RECOMENDADA

- Bland M. An Introduction to medical statistics. Oxford University Press , 2015, 4th ed. (*en linea en CRAI-UB*)
- Forthofer RN, Lee ES, Hernandez M. Biostatistics : a guide to design, analysis, and discovery. Elsevier Academic Press, 2007, 2nd ed (*en linea en CRAI-UB*)

[10] Stigler SM. The History of Statistics: The Measurement of Uncertainty before 1900. The Belknap Press of Harvard University Press. 1986
[11] Goodman S. Toward evidence-based medical statistics. 1. The P value fallacy. Annals of Internal Medicine. 1999 Jun; 130(12): 995-1004.
[12] Goodman S. A Dirty Dozen: Twelve P-Value Misconceptions. Seminars in hematology. 2008 Jul; 45(3): 135-40.
[13] Wasserstein RL, Lazar NA. The ASA's Statement on p-Values: Context, Process, and Purpose. The American Statistician, 70:2, 129-133.

- Johnson RA, Bhattacharyya GK. Statistics: principles and methods. Hoboken, N.J: Wiley; cop. 2010, 6ª ed.
- Rosner BA. Fundamentals of biostatistics. 7th ed. Pacific Grove, Calif: Brooks/Cole, Cengage Learning; 2011.
- Spiegel, MR. Statistics. Shaum's easy outlines. McGrawHill; 2011, 2ªed
- Wassertheil-Smoller S. Biostatistics and Epidemiology: a primer for health professionals. 2rd ed. New York. Springer-Verlag, 1990.

GLOSARIO

Conclusión clínica o práctica
Conclusión estadística
Confianza
Distribución del estadístico de la prueba
Distribución muestral: exacta vs asintótica
Error de tipo I
Error de tipo II
Estadístico de la prueba
Falsos resultados de la prueba de hipótesis
Grado de significación = P-valor
Hipótesis alternativa (Ha)
Hipótesis nula (Ho)
Hipótesis simple vs h. compuesta
Incertidumbre
Magnitud del efecto (delta)
Nivel de significación = riesgo alfa
P-valor = grado de significación
Potencia
Prueba o contraste de hipótesis
Prueba bilateral vs unilateral
Región crítica (o rechazo) vs r. de aceptación
Resultado de la prueba: rechazo vs no-rechazo
Riesgo alfa = nivel de significación
Riesgo beta
Significación estadística
Valor crítico (de tablas) = límite r. rechazo