INFERENCIA ESTADÍSTICA: PRUEBAS DE HIPÓTESIS

Jorge Dagnino S.[1]

- En la inferencia estadística existen dos aproximaciones complementarias: pruebas de hipótesis y estimación.
- La pruebas de hipótesis evalúan la probabilidad asociada a la hipótesis nula (H_0) de que no hay efecto o diferencia.
- El valor de p obtenido refleja la probabilidad de rechazar la H_0 siendo esta verdadera; en ningún caso prueba que la hipótesis alternativa, de que si hay efecto o diferencia, sea verdadera.
- Error tipo I (α) es un falso negativo: rechazar H_0 cuando esta es verdadera.
- Error tipo II (β) es un falso negativo, aceptar H_0 la cuando esta es falsa.
- La potencia de un experimento o test describe la probabilidad de detectar una diferencia verdadera de una determinada magnitud.

Los datos obtenidos en el transcurso de una investigación médica frecuentemente son usados para comparar el efecto de diferentes maniobras o situaciones: tratados *versus* no tratados, casos versus controles o normales versus enfermos. Aunque existen numerosas alternativas para hacer esas comparaciones - que dependerán fundamentalmente del diseño experimental, del tipo de datos y escalas de medición, y del número de muestras - y a pesar de la enorme variedad en el tipo de problemas médicos que pueden plantearse o de las soluciones estadísticas que se aplique, existen básicamente dos aproximaciones: **pruebas de hipótesis** y **estimación**. Vale la pena discutir algunos puntos de estas alternativas antes de elaborar sobre las pruebas estadísticas propiamente tales.

HIPÓTESIS DE TRABAJO E HIPÓTESIS NULA

Una hipótesis es una proposición que puede o no ser verdadera pero que se adopta provisionalmente hasta recabar información que sugiera lo contrario. Si hay inconsistencia, se rechaza la hipótesis. Las pruebas de hipótesis se usan precisamente para evaluar el grado de esa inconsistencia.

Se puede describir formalmente los pasos a seguir:

1) Formular la hipótesis y su alternativa. Normalmente la hipótesis de trabajo (por ejemplo, tal tratamiento es mejor que el control o tal procedimiento tiene menos morbilidad) es contrastada con una hipótesis estadística que supone que **no** existe tal efecto o tal diferencia. La razón para hacer esto es que se puede calcular de antemano la distribución de probabilidades asociadas con tal situación. Esta hipótesis se conoce con el nombre de hipótesis nula que se abrevia como H_0 (Nullus: Nula, falto de valor y fuerza para obligar o tener efecto). La expresión matemática es H_0: $\mu 1 = \mu 2$. La hipótesis alternativa es que el efecto **sí** existe, que es distinto de cero, y que en algunos casos se puede especificar el signo de esa diferencia. Normalmente corresponde a la hipótesis de trabajo, se abrevia como H_1 y tiene tres alternativas: $\mu 1 \neq \mu 2$, $\mu 1 < \mu 2$ o bien $\mu 1 > \mu 2$.
2) Elegir la prueba estadística apropiada de acuerdo al diseño experimental, el tipo de datos y el número de grupos que se comparan. La cifra que resulta de usar la prueba (aplicar la o las fórmulas) en los datos recolectados se conoce como el estadístico del test en cuestión: z; estadístico t o de Student, la r de Pearson, F del análisis de varianza, χ^2. La distribución del estadístico puede ser calculada de antemano cuando la H_0 es verdadera y por lo tanto conocer los valores

[1] Profesor Titular, División de Anestesiología, Pontificia Universidad Católica de Chile.

que delimitarán distintas porciones del área bajo la curva de esa distribución; éstas se conocen como distribuciones de muestreo. Vale la pena decir aquí, y lo reiteraremos luego, que las pruebas de hipótesis en ningún caso **prueban** la veracidad de la hipótesis alternativa o de trabajo, limitándose a decir que no hay suficiente evidencia para rechazar la hipótesis nula basándose en un nivel preestablecido de probabilidades.

3) Elegir el nivel de significación α de la prueba, el límite para rechazar H_0. En general, se acepta α = 0,01 ó 0,05, cifras que implican un 1%, o un 5% respectivamente, de posibilidades de equivocarse cuando se rechaza H_0, de decir que hay una diferencia cuando en realidad no la hay. Este es el llamado error tipo I.
4) Calcular el valor de P. Esta es la probabilidad de obtener los resultados observados u otros más extremos si la H_0 es verdadera, cifra que es determinada por el área de la distribución que queda más allá del valor calculado.
5) Si p es menor que α, rechazar H_0 y aceptar la alternativa; en caso contrario, se acepta la hipótesis nula. El conjunto de valores que resultarían en el rechazo de H_0 - calculados conociendo la prueba usada, α y el número de observaciones - se conoce con el nombre de región crítica (Figura 1). Este punto puede refrasearse así: se rechaza la H_0 si el estadístico cae en la región crítica. En los apéndices de los textos de estadística aparecen tablas con la distribución de estos estadísticos, dando el valor de p y donde el tamaño de la muestra se considera en los grados de libertad.

Como se evalúa el estadístico calculando la probabilidad de observar el valor encontrado u otro más extremo, el valor de P constituye la cola de la distribución. Este concepto es importante pues permite entender qué significa un test de una cola o de dos colas. Si la hipótesis de trabajo implica que existe una diferencia, sin especificar la dirección de esa diferencia ($\mu 1 \neq \mu 2$) debe usarse una prueba de dos colas. Si se es capaz de especificar de antemano el signo de ella ($\mu 1 < \mu 2$ o bien $\mu 1 > \mu 2$), se puede y se debe usar una prueba de una cola. El punto es importante pues el área crítica es mayor en este último caso lo que equivale a decir que se puede rechazar con un valor menor del estadístico. Como veremos luego, esto equivale a aumentar la potencia de la prueba en cuestión.

EL VALOR DE P

El valor de P es tan ubicuo y tan buscado en la literatura médica que ha llegado a alcanzar poderes mágicos. Feinstein, en su estilo claro e irónico, comenta sobre cómo un acto de juicio científico crítico se convirtió en un templo numérico arbitrario, asombrándose que la comunidad científica haya aceptado tan livianamente una guía que no es ni confiable ni efectiva. La P no es confiable porque depende absolutamente del tamaño de la muestra y esto la hace especialmente peligrosa en los trabajos epidemiológicos. No importa cuan trivial o incluso estúpida sea una hipótesis o cuan pequeña o inconsecuente sea la diferencia que se analiza, los resultados pueden ser estadísticamente significativos si se estudia un número suficiente de casos. La p no es efectiva como guía de la decisión científica, pues el valor de P no tiene que ver con la magnitud de la diferencia que se estudia y, por cierto, la interpretación de significado no debe leerse a partir del valor de P. Es intuitivamente obvio que es un error interpretar un trabajo en forma muy diferente si P resulta 0,048 o bien 0,052.

Es necesario tener claro que se rechaza la hipótesis nula porque es poco probable que sea verdadera con los datos obtenidos. En ningún caso se está probando la hipótesis alternativa de que sí hay efecto. Sólo se dice que la probabilidad de que el efecto observado no exista realmente es tan baja que se acepta que lo más probable es que efectivamente exista. A la inversa, el rechazo de la hipótesis nula jamás im-

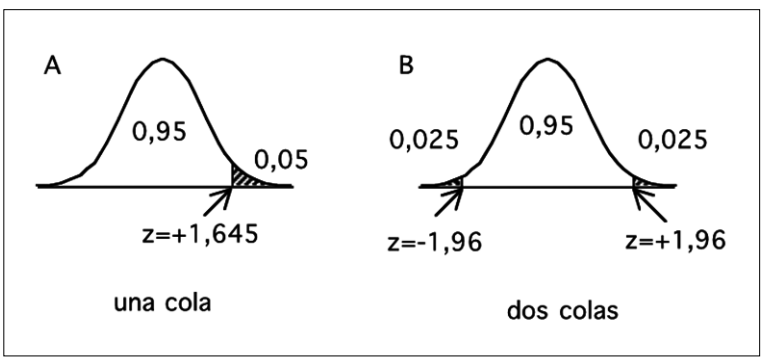

Figura 1. Distribución de un estadístico, en este caso z, y decisión entre la hipótesis nula o su alternativa. En blanco, área de aceptación y sombreada área crítica de rechazo de la H_0. En A, se destaca el área crítica en un extremo o cola de la distribución correspondiente a p = 0,05; en B, las áreas se distribuyen por mitades en los dos extremos o colas. Es aparente que se necesita un valor menor del estadístico para rechazar H_0 cuando se trata sólo de un extremo.

plica la comprobación de la igualdad. La distinción podría parecer sutil semánticamente pero es muy importante en la interpretación de los resultados.

Estas pruebas también se denominan de significancia pues los valores de P por sobre o debajo del límite se denominan como significativos, altamente significativos o no significativos. Desgraciadamente, su connotación estadística ha sido trasladada a la clínica, creando mucha confusión entre lo que es estadísticamente significativo y lo que es clínicamente significativo. En cualquier caso, es indispensable tener claro que lo que debe interesar es probar estadísticamente sólo aquellas diferencias que pueden ser clínicamente importantes. En este sentido sería mejor y científicamente correcto que se estableciese, ya desde la formulación de la hipótesis, la magnitud de la diferencia que justificaría usar un test de hipótesis y no al revés, en que se aplica éste y después se intenta justificar la importancia de la diferencia. La situación empeora pues con frecuencia ni siquiera se hace la justificación y autores y lectores se convencen que la p hace la verdad.

Otro problema, derivado de la dependencia de P para declarar un estudio significativo o no significativo, es lo que se denomina sesgo de publicación, por el que con mayor frecuencia son publicados o enviados a publicación los trabajos con una P significativa comparados con los negativos. Ello es especialmente importante a la hora de revisiones de la literatura y particularmente cuando se usan los hoy tan en boga metanálisis. El aceptar los resultados de éstos sin otras consideraciones podría estar generando otro tipo de mitos o errores.

Por otro lado, una P grande se asemeja mucho al fracaso y se habla de resultados negativos cuando en realidad podrían ser muy positivos para los enfermos pues podrían ahorrarse tratamientos costosos y riesgosos que no sirven. También se ha criticado el uso o mal uso de la expresión "fracaso en alcanzar significación estadística" como si eso fuese un pecado o demostración de ineficiencia.

CONCLUSIONES Y ERRORES DE UN TEST DE HIPÓTESIS

Frente a dos posibilidades reales, no hay diferencias (H_0) o bien sí las hay (H_1), las pruebas de hipótesis pueden dar dos resultados: rechazar o aceptar H_0. En estas circunstancias, en forma análoga a lo que sucede con los exámenes de laboratorio diagnósticos, las alternativas son cuatro. Dos no constituyen más que la coincidencia entre la realidad y el resultado de las pruebas:

1) Se rechaza H_0 cuando ésta es falsa, una diferencia verdadera es declarada estadísticamente significativa. Es un verdadero positivo.
2) Se acepta H_0 cuando ésta es verdadera, no hay una diferencia estadísticamente significativa y en realidad no la hay. Un verdadero negativo.

Las otras alternativas implican una incongruencia entre la realidad y los resultados y, por lo tanto, constituyen errores.

3) Se rechaza H_0 cuando ésta es verdadera, concluyendo que hay una diferencia que en realidad no existe, un falso positivo. Se ha cometido un error que se denomina de tipo I (α). La probabilidad de que ocurra este tipo de error es la que se controla al establecer α y normalmente no va más allá del 5%. Sin embargo, inadvertidamente puede ser mayor cuando no se cumplen los requisitos necesarios para aplicar la prueba de hipótesis elegida: usar un test paramétrico cuando en realidad se debió usar uno no paramétrico, una prueba de una cola en vez de una de dos colas o comparaciones múltiples con tests diseñados para comparar sólo dos medias o medianas.
4) Se acepta H_0 cuando en realidad es falsa, un falso negativo, concluyendo que no hay diferencia cuando en realidad existe. Este es el error tipo II (β), que la mayoría de las veces se debe a un tamaño insuficiente de la muestra. La probabilidad de cometer un error tipo II es β cuyo valor depende de la magnitud del efecto de interés y del tamaño de la muestra. Sin embargo, es más frecuente hablar de la potencia de la prueba para detectar un efecto de un tamaño determinado.

Estos dos errores deben ser considerados al evaluar el resultado de un trabajo de investigación que haya empleado pruebas de hipótesis, considerando la posibilidad de un error I cuando los resultados son significativos y de un error tipo II cuando son no significativos. La Tabla 1 resume la relación entre los resultados de una prueba de hipótesis y la realidad.

POTENCIA

La potencia de un test es igual a 1- β y describe la probabilidad de detectar una diferencia verdadera de una magnitud determinada. Mientras mayor es la potencia de un test menor es la probabilidad de tener un falso negativo. Frente a cada resultado no significativo debe considerarse la posibilidad de un error tipo II y la potencia de la prueba usada para evaluar la diferencia. En otras palabras, una prueba tiene una elevada potencia cuando tiene una escasa probabilidad de rechazar la hipótesis nula cuando

Tabla 1. Relación entre los resultados de una prueba de hipótesis y la realidad

		SITUACIÓN VERDADERA	
		Hay diferencia (H_1)	No hay diferencia (H_0)
PRUEBA	Significativa (Rechaza H_0)	Verdadero (+) Conclusión correcta Potencia 1-β	Falso (+) Error I o error β
	No significativa (Acepta H_0)	Falso (-) Error II o error β	Verdadero (-) Conclusión correcta

ésta es verdadera, pero una gran posibilidad de rechazarla cuando ésta es falsa.

Para calcular la potencia de un test es necesario conocer la magnitud de la diferencia, la variabilidad y el tamaño de la muestra. En la planificación de un trabajo se debe estimar el tamaño de la muestra usando para ello una estimación de la diferencia esperada y clínicamente significativa, la variabilidad esperada y decidir cuál es el error β que se declarará aceptable (usualmente 20% lo que en muchas circunstancias es excesivo). Las cifras de diferencia y variabilidad se pueden obtener de trabajos en áreas similares en la literatura o bien desarrollar un piloto que permita hacer la estimación.

Para aumentar la potencia de un test hay diversas alternativas:

1) Como existe una relación inversa entre α y β, se puede aumentar la potencia elevando el nivel de significación. Es la peor alternativa pues implica aumentar la posibilidad de un error tipo I.
2) Considerar sólo diferencias mayores concentrándose sólo en aquellas médica o biológicamente relevantes, algo que debería estar siempre presente desde el inicio.
3) Aumentar el tamaño de la(s) muestra(s), lo que aumenta el costo y la duración de la investigación.
4) Disminuir la variabilidad con mediciones más precisas o escogiendo muestras más homogéneas.
5) Hacer la mayor cantidad de presunciones válidas posibles (por ejemplo, el test de 1 cola es más potente que el de 2 colas).
6) Usar la prueba de mayor poder aplicable al caso en cuestión (por ejemplo, un test paramétrico tiene mayor poder que uno no paramétrico).

Existe un caso frecuente de error tipo II que se da al evaluar la seguridad de un tratamiento o procedimiento en término de los efectos adversos o la morbimortalidad. Cuando éstos tienen una baja incidencia, es frecuente que en una serie relativamente pequeña no aparezcan casos del evento, hecho que muchas veces es interpretado como hallazgo negativo y la droga o el procedimiento son catalogados como seguros. Debe quedar muy claro el hecho que un numerador de cero no significa que no exista riesgo y no impide que se puedan hacer estimaciones sobre la magnitud de ese riesgo. De hecho, los principios de estadística inferencial que se aplican a numeradores distintos de cero también se aplican cuando este es cero. Por ser un problema de estimación, será analizado en el artículo correspondiente.

REFERENCIAS

1. Altman DG. Practical Statistics for Medical Research. London: Chapman & Hall, 1991.
2. Bland JM, Altman DG. One and two sided tests of significance. BMJ 1994; 309: 248.
3. Bland M. An Introduction to Medical Statistics. 3rd Ed, Oxford: OUP, 2006.
4. Browner WS, Newman TB. Are all significant P values created equal? The analogy between diagnostic tests and clinical research. JAMA 1987; 257: 2459-2463.
5. Dawson-Saunders B, Trapp RG. Bioestadística Médica. México D.F: Manual Moderno, 1993.
6. Freiman JA, Chalmers TC, Smith H Jr, Kuebler RR. The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial. Survey of 71 "negative" trials. NEJM 1978; 299: 690-694.
7. Glantz SA. Primer of Biostatistics. 3a edición, New York: McGraw-Hill, 1992.
8. Goodman SN. Toward Evidence-Based Medical Statistics. 1. The P Value Fallacy. Ann Intern Med 1999; 130: 995-1004.
9. Holland BK. A Classroom Demonstration of Hypothesis Testing. Teaching Statistics 2007; 29: 71-73.
10. Portney LG, Watkins MP. Foundations of Clinical Research. Applications to practice. 2nd ed. Upper Saddle River: Prentice-Hall, 2000.
11. Siegel S. Non parametric statistics for the behavioral sciences. New York: McGraw-Hill, 1956.
12. Sterne JA, Smith GD. Sifting the evidence-what's wrong with significance tests?. BMJ 2001; 322: 226-231.

Correspondencia a:
Dr. Jorge Dagnino S.
jdagnino@med.puc.cl