Contraste de hipótesis: el valor *p*

V. Abraira

Unidad de Bioestadística Clínica. Hospital Ramón y Cajal. Madrid.

El contraste de hipótesis es la técnica estadística más frecuentemente usada en la literatura clínica[1] y epidemiológica[2], sin embargo es reconocido, que también con mucha frecuencia, es mal entendida e interpretada erróneamente[3]. La confusión es en parte debida a que la manera actual de usar los contrastes de hipótesis es una combinación de dos metodologías originalmente enfrentadas entre sí, y que contienen elementos incompatibles[4,5].

La idea inicial de las pruebas de significación se debe a Fisher[6]. Supóngase que se quiere evaluar si un tratamiento suministrado después de un episodio de ictus mejora el pronóstico de los pacientes. Para ello[7], una serie de 6.105 pacientes que han sufrido un ictus, se distribuyen aleatoriamente en dos grupos, a los pacientes de uno de los grupos se les administra el tratamiento propuesto, mientras que a los del otro grupo se les suministra placebo y se sigue a todos ellos durante 4 años; la evaluación de la eficacia del tratamiento se obtiene de la comparación de las proporciones de recurrencia del ictus entre los grupos tratados (10,1%) y placebo (13,8%). Si el tratamiento no fuera eficaz ambas proporciones serían iguales, aunque no necesariamente exactamente iguales[8]. La idea de Fisher consiste en realizar la comparación calculando una probabilidad, el "famoso" valor *p* o nivel de significación: la probabilidad de encontrar una diferencia en las proporciones de recurrencia de ictus como la que se ha encontrado o mayor en la hipótesis, llamada hipótesis nula, de que el tratamiento no tenga efecto y usar este valor *p* como un índice de la fuerza probatoria de los datos contra la hipótesis nula, cuanto menor sea *p*, mayor será la carga de la prueba en contra de la hipótesis nula; propone, además, el valor de 0,05 como punto de corte "conveniente", aunque argumenta enfáticamente que la interpretación última la debe hacer el investigador. Es obvio que en la actualidad ese punto de corte se usa como una regla mucho más rígida. En nuestro ejemplo, los autores señalan que $p<0{,}0001$, por lo tanto los resultados del estudio aportan una gran fuerza probatoria contra la hipótesis de que el tratamiento no afecta al pronóstico, y los autores concluyen que el tratamiento debería ser rutinariamente considerado en los pacientes con historia de ictus[7].

Con posterioridad a Fisher, y como reacción a la subjetividad inherente a la interpretación del valor *p*, Neyman y Pearson proponen los denominados contrastes de hipótesis[4] en los que se reemplaza el subjetivo concepto de fuerza probatoria por un procedimiento para decidir entre dos hipótesis, la hipótesis nula (en nuestro ejemplo el tratamiento no es eficaz) y la hipótesis alternativa (el tratamiento sí es eficaz). Se fijan a priori unas tasas aceptables para los dos tipos de error que se pueden cometer (tabla 1), se calcula el valor *p* (de la misma manera que en la propuesta de Fisher, aunque en la actualidad existe una gran variedad de procedimientos, los llamados tests estadísticos, para calcular ese valor en distintas situaciones experimentales y para distintos parámetros y funciones de parámetros) y se usa para tomar una decisión: si *p* es menor que la tasa aceptada de error tipo I se rechaza la hipótesis nula a favor de la alternativa, de lo contrario no se rechaza la hipótesis nula. El procedimiento garantiza a la larga una frecuencia preestablecida de decisiones correctas, pero no dice nada sobre la verdad o falsedad de cada hipótesis concreta.

En la actualidad las dos concepciones se usan mezcladas de un modo que seguramente disgustaría a los creadores de ambas y se quiere ver en el valor *p* tanto un índice de la fuerza probatoria como una tasa de error en la decisión: se dice, por ejemplo, que la diferencia es significativa al nivel *p*, pero también que se acepta o rechaza la hipótesis nula con el nivel α. Esta mezcla ha dado lugar, por ejemplo, a distintos estilos de comunicar los

Tabla 1. Tipos de errores en un contraste de hipótesis

	La "verdad"	
Resultado del experimento	H_0 cierta	H_0 falsa H_1 cierta
H_0 rechazada	Error tipo I (α)	Decisión correcta
H_0 no rechazada	Decisión correcta	Error tipo II (β)

H_0: hipótesis nula; H_1: hipótesis alternativa; error tipo I: rechazar la hipótesis nula siendo cierta; error tipo II: aceptar la hipótesis nula siendo falsa; α: probabilidad error tipo I; β: probabilidad error tipo II.

Correspondencia:
Dr. V. Abraira.
Unidad de Bioestadística Clínica.
Hospital Ramón y Cajal. Crta. Colmenar km 9,100.
28034 Madrid.
Correo electrónico: victor.abraira@hrc.es

resultados que pueden incluso coexistir en un mismo artículo, a veces los investigadores dan el valor exacto de *p*, a veces sólo comunican que es menor que un cierto punto de corte, a veces el punto de corte es el mismo en todo el artículo (por ejemplo el ubicuo 0,05) pero a veces se usan puntos de corte diferentes. En el artículo de nuestro ejemplo[7], se puede ver p=0,7, p<0,01, p<0,001 y también p>0,1.

La interpretación errónea más frecuente[3] en el uso de los contrastes consiste en interpretar el valor *p* como la probabilidad de que la hipótesis nula sea cierta y que, por tanto, un resultado "significativo" significa que es muy improbable que la hipótesis nula sea cierta. Para interpretarlo correctamente, hay que darse cuenta que el valor *p* es la probabilidad de unos resultados dada la hipótesis nula, que es distinta de la probabilidad de la hipótesis nula dados los resultados, es decir, son probabilidades que están en la misma relación que, por ejemplo, en el caso más familiar a los clínicos de las pruebas diagnósticas, la sensibilidad (probabilidad de un resultado positivo de la prueba en los enfermos) y el valor predictivo positivo (probabilidad de estar enfermo en los individuos con resultado positivo)[9].

La crítica más importante que recibe el valor de *p* como índice de la fuerza probatoria es que no depende sólo del tamaño del efecto observado, sino, y sobre todo, del tamaño muestral[4,5]. Así, en el ejemplo del ictus se obtuvo p<0,0001 para la diferencia entre las proporciones 10,1% y 13,8% encontradas en 6.105 pacientes (3.051 en el grupo del tratamiento activo y 3.054 en el del placebo); si el experimento se hubiera hecho con 300 en cada grupo, para las mismas proporciones se hubiera encontrado p=0,164; en el otro extremo, si se hubiera hecho con 30.000 en cada grupo, y se hubieran encontrado las proporciones 10,0% y 10,5%, el valor *p* hubiera sido 0,042, es decir una diferencia significativa para unas proporciones cuya diferencia desde el punto de vista clínico sería irrelevante.

Con la otra interpretación, la crítica más importante a los contrastes de hipótesis como forma de tomar decisiones es que éstas se toman sin tener en cuenta ninguna información ajena al experimento, el formalismo de los contrastes de hipótesis no contempla la información proveniente de otros estudios, se asume que los investigadores y los lectores son vírgenes respecto a las hipótesis en juego, asunción que parece bastante irreal y que ha dado lugar a un pernicioso estilo del apartado "Discusión" de los artículos, en el que rara vez se discuten los resultados del estudio en el contexto de una revisión sistemática actualizada de artículos anteriores[10].

Las reacciones ante estos frecuentes errores, malas interpretaciones y limitaciones van básicamente en dos sentidos: uno recomendar limitaciones en su uso[2] y proponer utilizar en su lugar los intervalos de confianza[8]; la revista *Epidemiology* es una abanderada de esta posición, aunque últimamente la ha suavizado un poco[11], pero también se ha señalado que ambas aproximaciones comparten la misma base teórica y por lo tanto los mismos problemas[12]. La otra propuesta supone una alternativa radicalmente distinta y aunque hasta ahora su uso es muy limitado, es probable que en un futuro próximo asistamos a su despegue: se trata de los métodos bayesianos[4,5] a cuyo fundamento se dedicará una nota más adelante.

Puntos clave

- El contraste de hipótesis es la técnica estadística más frecuentemente usada en la literatura clínica, pero con mucha frecuencia es mal entendida e interpretada.
- Se basa en poner a prueba una hipótesis de no diferencia (hipótesis nula) calculando la probabilidad de encontrar una diferencia como la que realmente se ha encontrado o mayor, en el supuesto de que la hipótesis nula sea cierta.
- Esa probabilidad se usa como un índice de la fuerza probatoria de los datos contra la hipótesis nula, aunque también como instrumento para tomar una decisión garantizando a la larga unas tasas de error preestablecidas.
- A pesar de su uso prácticamente ubicuo en la literatura médica, esta doble interpretación no está exenta de contradicciones y está en el origen de los errores y malas interpretaciones.

BIBLIOGRAFÍA

1. Guyatt G, Jaeschke R, Heddle N, Cook D, Shannon H, Wermuth L. Basic statistics for clinicians: 1. Hypothesis testing. CMAJ 1995;152:27-32.
2. Poole C. Low P-values or narrow confidence intervals: which are more durable? Epidemiology 2001;12:291-4.
3. Sterne JAC, Smith GD. Sifting the evidence - what's wrong with significance tests? Br Med J 2001;322:226-31.
4. Silva LC, Muñoz A. Debate sobre métodos frecuentistas vs bayesianos. Gac Sanit 2000;14:482-94.
5. Goodman SN. Toward evidence-based medical statistics. 1: The P value fallacy. Ann Intern Med 1999;130:995-1004.
6. Fisher RA. Statistical Methods, Experimental Design and Scientific Inference (Re-issue). Oxford: Oxford University Press; 1995.
7. PROGRESS Collaborative Group. Randomised trial of a perindopril-based blood-pressure-lowering regimen among 6105 individuals with previous stroke or transient ischaemic attack. Lancet 2001;358:1033-41.
8. Abraira V. Estimación: intervalos de confianza. SEMERGEN 2002;28:84-5.
9. Abraira V. Índices de rendimiento de las pruebas diagnósticas. SEMERGEN 2002;28:193-4.
10. Clarke M, Chalmers I. Discussion Sections in Reports of Controlled Trials Published in General Medical Journals: Islands in Search of Continents? JAMA 1998;280:280-2.
11. The Editors. The value of *p*. Epidemiology 2001;12:286.
12. Feinstein AR. P-values and confidence intervals: two sides of the same unsatisfactory coin. J Clin Epidemiol 1998; 51:355-60.