Los valores *P* y los intervalos de confianza: ¿en qué confiar?

María Luisa Clark[1]

Las pruebas de hipótesis, que también se conocen por pruebas de significación estadística o pruebas de la hipótesis de nulidad (o de la hipótesis nula), forman parte fundamental del material que se imparte en los cursos de estadística universitaria. Una vez que aprende a manejarlas, el estudiante o investigador no vacila en aplicarlas libremente, pese a que hoy en día la mayoría de los estadísticos desaconsejan su uso debido a las graves deficiencias de estas pruebas y a su dudosa utilidad en comparación con otros métodos de análisis inferencial.

Los primeros argumentos en contra del uso de pruebas de hipótesis aparecieron durante la primera mitad del siglo pasado con carácter esporádico (1, 2), pero cuando en 1986 *British Medical Journal* dio a conocer su postura al respecto, el debate en torno a ellas cobró un ímpetu que ha quedado evidenciado en los centenares de artículos posteriores que contrastan las ventajas de los intervalos de confianza con las carencias del valor *P* (3–6). Lo cierto es que estas últimas, raras veces abordadas en la sala de clase, son considerables desde el punto de vista de los fines que persigue un investigador. Hoy por hoy se reconoce que los intervalos de confianza aventajan a las pruebas de hipótesis como instrumento analítico para muchos tipos de investigación, entre ellos los estudios observacionales y experimentales relacionados con las ciencias médicas y sociales, con el resultado de que la mayoría de las revistas biomédicas alientan a sus autores a proporcionar intervalos de confianza en lugar de valores *P*.

Fáciles de calcular con los paquetes estadísticos modernos, los valores *P* ejercen un poderoso atractivo sobre el investigador por la exigua reflexión que exigen y la falsa sensación de seguridad que confieren. Un solo número encierra la clave que determina si los resultados de un estudio han de sumarse a las pruebas a favor o en contra de una hipótesis, y el investigador que obtiene resultados significativos suele sentirse satisfecho de haber logrado su meta, sin darse cuenta de que no ha conseguido mejorar en modo alguno su comprensión del fenómeno que estudia. Para entender a fondo esta afirmación, conviene examinar qué es un valor *P*.

Comencemos con un ejemplo elemental. Un investigador sospecha que las personas expuestas a un factor determinado están en mayor riesgo de contraer cierta enfermedad que las personas que no lo están y se propone demostrarlo matemáticamente. Su hipótesis de trabajo es que hay una diferencia entre el grupo expuesto y el grupo sin exposición (grupo testigo) en lo referente a la frecuencia de la enfermedad en cuestión, pero para poder demostrarlo valiéndose de una prueba de hipótesis, tiene que empezar por plantear la hipótesis contraria (hipótesis nula o de nulidad), es decir, que no hay diferencia alguna entre los grupos comparados en lo referente a la frecuencia de la enfermedad de interés. La finalidad es tener bases numéricas para descartar la hipótesis de nulidad y poder dar por verdadera la hipótesis alterna, confirmándose así que, muy verosímilmente, la frecuencia de la enfermedad de interés en los grupos comparados sí es distinta.

Una vez planteada la hipótesis de nulidad, es preciso que el investigador determine el margen de equivocación que está dispuesto a tolerar y fije el llamado valor de significación o valor alfa (α), para luego calcular el valor *P*. En términos sencillos, este último valor no es otra cosa que la probabilidad de observar la diferencia encontrada entre los grupos o una más extrema si es correcta la hipótesis de nulidad.

Si el valor *P* es menor del valor α fijado por el investigador (0,05 la mayor parte de las veces, o en ocasiones 0,01 ó 0,10), se descarta que los resultados observados puedan atribuirse a mero azar si en realidad no hay una diferen-

[1] Jefa de Redacción, *Revista Panamericana de Salud Pública*.

cia, o, dicho de otro modo, la incompatibilidad entre los datos observados y la hipótesis de nulidad se considera lo suficientemente grande como para poder descartar esta hipótesis. En cambio, si el valor *P* es α o mayor, se considera que no hay suficientes indicios para descartar la hipótesis de nulidad.

Varios errores fundamentales se asocian con el uso de pruebas de hipótesis cuando no se entiende qué es lo que permiten calcular. Uno de ellos consiste en pensar que un valor *P* de algún modo cuantifica la magnitud de la diferencia encontrada entre dos grupos sometidos a comparación. En parte ello se debe a una confusión semántica, ya que en el lenguaje común y corriente "significativo" equivale a "notable", "destacado" o "importante." Pero un valor *P*, como hemos visto, no refleja en absoluto la magnitud de la diferencia que el investigador encuentra, sino la probabilidad de haber observado esa diferencia si en realidad no hay ninguna. Sobre la base de un resultado estadísticamente significativo se puede concluir que un medicamento supera a otro en eficacia, pero no si es tanta su superioridad que se justifique exponer al paciente a sus efectos secundarios, por ejemplo. Para fines prácticos, lo que interesa es conocer la magnitud de la diferencia, para lo cual el valor *P* carece por completo de utilidad.

Otra idea equivocada es que un valor *P* mayor de α confirma que la hipótesis de nulidad es correcta, o que representa la probabilidad de que lo sea. Cabe aclarar que la estadística, basada por entero en probabilidades y frecuencias, no cuenta con ninguna herramienta que sirva para confirmar la hipótesis de nulidad. Siempre hay una posibilidad, por remota que sea, de que a la luz de los datos una hipótesis de nulidad parezca verdadera aun siendo falsa, por obra del azar (error tipo II). El no poder rechazar la hipótesis de nulidad no equivale a poder confirmarla, y la diferencia entre una cosa y otra influye decisivamente sobre las conclusiones que pueden derivarse de un estudio (7).

Mientras más grande una muestra, menor es la influencia del azar sobre los resultados y menor la probabilidad de cometer errores de interpretación, y esto nos lleva a otro disparate muy frecuente, que es el de afirmar, frente a un resultado sin significación estadística, que el carácter reducido de la muestra explica la falta de significación. El problema con esta afirmación es, precisamente, que es cierta para cualquier resultado y, por ende, completamente vacía. En otras palabras, cuando una muestra es lo suficientemente grande, cualquier resultado puede cobrar significación estadística, razón de por sí suficiente para dudar del valor de estas pruebas.

Otra práctica muy difundida es el uso de valores de α fijados por convención, sin atención a los antecedentes del problema particular que se examina y a las consecuencias prácticas de tomar decisiones basadas en el uso de un nivel de significación u otro. El investigador consciente busca información que tenga utilidad en la práctica y su elección de α debe contemplar los fines de su estudio y el uso que tendrán sus resultados a la luz de los conocimientos derivados de estudios anteriores. Los valores de α convencionales son arbitrarios y carecen de respaldo teórico. Si $P = 0{,}045$, por lo general concluimos que los resultados tienen significación; en cambio, concluimos que no la tienen cuando $P = 0{,}050$. ¡Dos conclusiones radicalmente distintas derivan de una diferencia de 0,005 en el valor de *P*! Sin embargo, los resultados de metaanálisis se apoyan en buena medida sobre estos cimientos endebles (1).

Para muchos especialistas las pruebas de hipótesis encierran un error conceptual que reduce aun más su fiabilidad. Al tener que partir de la premisa de que la hipótesis de nulidad es verdadera, toda prueba de hipótesis se basa en una suposición que raras veces se cumple en la práctica. Hay incluso quienes argumentan que, en principio, la mayoría de las hipótesis de nulidad son falsas, ya que en la vida real dos grupos nunca son idénticos con respecto a una característica determinada (6, 8). La presunta falsedad *a priori* de toda hipótesis de nulidad menoscaba, como es de suponer, la integridad de este tipo de prueba.

Basta con lo señalado hasta ahora para entender por qué las pruebas de hipótesis gozan de tan poca popularidad entre los expertos, pero si hubiese que escoger el defecto que más descalifica su uso, este sería, sin duda, el primero que hemos señalado: la poca información que aportan a la luz de lo que el investigador necesita saber. Los valores *P* no miden la magnitud del efecto observado, como tampoco su precisión; es decir, no dan ninguna idea de cuán confiable o fuerte es el efecto detectado en un estudio, ni permiten saber cuánto variarían los resultados si el estudio se repitiese con distintas muestras. No aportan información que lleve a acumular conocimientos útiles en términos prácticos ni a formular nuevos postulados teóricos que marquen el rumbo de futuras investigaciones. Dicho de otro modo, la significación estadística de un resultado no es ningún indicio de su "significación" clínica, aunque a menudo las dos cosas se confunden (1). La interpretación de un resultado a la luz de un valor *P* es una práctica mecánica e irreflexiva cuya persistencia es difícil de comprender si se considera que los intervalos de confianza, que le revelan al investigador el margen de error de sus resultados y la magnitud del efecto que observa, fomentan la actividad analítica imprescindible para la evolución del conocimiento científico. Veamos ahora qué es un intervalo de confianza y en qué aspectos aventaja al valor *P*.

Un intervalo de confianza es un recorrido de valores, basados en una muestra tomada de una población, en el que cabe esperar que se encuentre el verdadero valor de un parámetro poblacional con cierto grado de confianza. La distribución de un parámetro fisiológico en la población sirve de fundamento teórico para calcular estos intervalos. Recordemos que en una distribución normal o gausiana, cerca de 68% de los valores se encuentran en el intervalo abarcado por la media ± 1 desviación estándar (DE); más de 95% de los valores, en el intervalo abarcado por la media ± 2 DE; y más de 99% de los valores, en el intervalo abarcado por la media ± 3 DE. Sobre esta base, un intervalo de confianza de 95%, que es el que se busca con mayor frecuencia, se obtiene sumándole y restándole a la media el error estándar multiplicado por 1,96. Si quisiese calcularse un intervalo de confianza de 99%, el error estándar se multiplicaría por 2,58. ¿Qué indica, entonces, un intervalo de confianza de 95%? Que si el investigador repitiese su estudio en las mismas condiciones pero con distintas muestras aleatorias, noventa y cinco de cada cien veces obtendría intervalos que contendrían el verdadero parámetro poblacional y cinco veces obtendría intervalos que no lo contendrían. En otras palabras, se puede tener gran confianza en que el intervalo resultante abarca el valor verdadero, pues dicho intervalo se ha obtenido por un método que casi siempre acierta. Esto no equivale a decir que hay una probabilidad de 95% de que el verdadero valor se encuentre dentro del intervalo, error de interpretación que es bastante común. La confianza deriva de la aplicación de un método respaldado por lo que se sabe acerca de la distribución poblacional de determinado parámetro fisiológico. El investigador que lo aplica contará con información valiosa —la magnitud y precisión del efecto observado— que no puede conseguir mediante un valor *P*.

Un intervalo de confianza posee la ventaja de que se puede calcular para cualquier valor. Si se desea determinar si es verdadera la diferencia observada entre dos grupos, se calcula el intervalo de confianza de 95% de la diferencia entre sus respectivas medias. Si el intervalo abarca el valor cero, no se puede descartar que no haya una diferencia; si no lo abarca, la probabilidad de que se esté observando una diferencia que en realidad no existe se considera remota. La misma lógica se aplica al calcular el intervalo de confianza de una razón de posibilidades o de un riesgo relativo, solo que en estos casos el valor 1 es el que indica la ausencia de una diferencia porque se trata de una proporción. No obstante, si los intervalos de confianza solo se usaran de esta manera, entonces no se diferenciarían en nada de las pruebas de significación. Podemos ver, nuevamente, que no solo ofrecen mucha más información que los valores *P*, sino que abarcan a las pruebas de significación y hasta podrían usarse como sucedáneos de ellas.

La precisión de los resultados guarda relación con el tamaño muestral y con la variabilidad de los datos, de tal manera que cuanto más grande la muestra, más se acercarán los resultados al verdadero valor poblacional y más estrecho será el intervalo de confianza. Asimismo, mientras más grande sea la desviación estándar de los datos, menos precisos serán los resultados y más ancho el intervalo de confianza. Un investigador no puede controlar la desviación estándar, pero puede manipular el tamaño muestral para mejorar la precisión y utilidad de sus resultados. Si lo juzga necesario, puede efectuar estudios sucesivos con muestras cada vez mayores para llegar a conclusiones clínicas con mayor certidumbre.

Esta larga disquisición nos trae al interesante estudio de Sarria Castro y Silva Ayçaguer (9), "Las pruebas de hipótesis en tres revistas biomédicas: una revisión crítica", que se publica en este número de la *Revista Panamericana de Salud Pública/Pan American Journal of Public Health* (RPSP/PAJPH). Atentos al uso irreflexivo de pruebas de hipótesis, los autores comparan tres revistas de salud pública, incluida la RPSP/PAJPH, desde el punto de vista de la frecuencia con que los artículos publicados en ellas dan valores *P* o intervalos de confianza. Sobre la base de los trabajos publicados en las tres revistas de 1996 a 2000, los autores concluyen que se sigue abusando de las pruebas de hipótesis e incurriendo en disparates hoy en día inadmisibles. También reparan en otros errores que ya hemos mencionado: el uso de la palabra "significativo" en un sentido ajeno al concepto estadístico de significación; la atribución de la falta de significación estadística a una muestra demasiado pequeña; el uso automático de niveles α convencionales (0,05, 0,01), sin tener en cuenta los antecedentes del problema estudiado, entre otros. Por tales motivos consideramos oportuno reiterar lo explicitado en la "Información para los autores" publicadas en cada número de enero de la RPSP/PAJPH y en www.paho.org, al efecto de sustituir o complementar los valores *P* con intervalos de confianza. También instamos a nuestros autores a estar atentos a los errores comunes que suelen asociarse con el uso de pruebas de hipótesis y la presentación de sus resultados. Vaya también el agradecimiento de la redacción por la contribución de los doctores Sarria Castro y Silva Ayçaguer a mejorar los estándares editoriales de nuestra revista. Los lectores que deseen adentrarse en el tema aquí tratado encontrarán muy útiles las listas de referencias que se proporcionan en las siguientes direcciones de Internet: http://www.cnr.colostate.edu/~anderson/thompson1.html y http://www.cnr.colostate.edu/~anderson/nester.html

REFERENCIAS

1. Berkson J. Some difficulties of interpretation encountered in the application of the chi-square test. J Am Stat Assoc. 1938;33:526–36.
2. Berkson J. Tests of significance considered as evidence. J Am Stat Assoc. 1942;37:325–35.
3. Gardner MJ. Altman DG. Confidence intervals rather than P values: estimation rather than hypothesis testing. Br Med J. 1986;292:746–50.
4. Carver RP. The case against statistical significance testing. Harvard Educ Rev. 1978;48:378–99.
5. Cohen J. The earth is round (p<.05). Am Psicol. 1994;49:997–1003.
6. Anderson DR, Burnham KP, Thompson WL. Null hypothesis testing: problems, prevalence, and an alternative. J Wildlife Management. 2000;64(4):912–23.
7. Parkhurst DF. Interpreting failure to reject a null hypothesis. Bull Ecol Soc Am. 1985; 66:301–2.
8. Johnson DH. Statistical sirens: the allure of nonparametrics. Ecology. 1995;76:1998–2000.
9. Sarria Castro M, Silva Ayçaguer LC. Las pruebas de significación estadística en tres revistas biomédicas en lengua española: una revisión crítica. Rev Panam Salud Publica. 2004; 15(5):300–6.