Editorial

Los límites de las pruebas de significación estadística y los valores p

Luis Carlos Silva Ayçaguer
Investigador Titular.
Centro Nacional de Información de Ciencias Médicas. La Habana (Cuba).

Entre los recursos inferenciales más extendidos en la investigación empírica contemporánea se hallan las *pruebas de significación estadísticas* (PSE), que datan de la tercera década del siglo precedente. Lamentablemente, en pocas áreas de la estadística más que en ésta se han arraigado las recetas, presentes tanto en los centenares de libros introductorios sobre estadística inferencial que –con pocas diferencias entre sí– se fueron publicando desde entonces, como en los cursos clásicos de estadística que ofrecen las universidades y otros servicios formativos. Unos y otros –salvo excepciones– suelen concentrarse acríticamente en la comunicación de una secuencia mecánica de códigos operativos que desembocan en el cómputo de los llamados «valores p» y, por su conducto, en reglas para establecer veredictos con los que no pocos investigadores se consideran relevados de la obligación de pensar. El problema no sería grave si se tratara de un procedimiento sólido y carente de insuficiencias relevantes. Pero son notables las endebleces metodológicas de las PSE; y enconada la controversia a que han dado lugar. Lo cierto es que cada día se torna más difícil soslayar la necesidad de suplirlas o, como mínimo, complementarlas con recursos más racionales.

Las PSE han sido tan castigadas en el plano conceptual que hoy son contados los estadísticos profesionales que se empeñan en defenderlas explícitamente. Dichas pruebas conservan una tenaz presencia en la literatura científica, acaso debida a la comodidad que supone dejarse llevar por la inercia y usar procedimientos que proporcionan «soluciones» inmediatas, capaces de generar conclusiones «científicas» por sí mismos[1]. Sin embargo, la acumulación de observaciones críticas aparecidas desde su creación conforma hoy un voluminoso prontuario.

Aunque muchos lo ignoran, las PSE nacen y se desarrollan en un contexto altamente conflictivo. Las contribuciones de Ronald Fisher (inventor de la p y las «pruebas de significación») y más tarde de Jerzy Neyman y Egon Pearson (creadores de las «pruebas de hipótesis») estuvieron animadas por agrias impugnaciones mutuas, tanto en el plano conceptual como práctico. Los pormenores de tales controversias pueden consultarse en los libros de historia[2].

Como una forzada conciliación de estas perspectivas originalmente divergentes, se desembocó hacia 1940 en la actual PSE. En esencia, ésta consiste en enjuiciar la validez de una hipótesis nula H_o mediante la comparación del valor p obtenido con cierto umbral preestablecido (usualmente, si $p < 0,05$, se rechaza H_o en favor de una propuesta novedosa).

Y es precisamente ese ritual el que ha producido el debate más violento, crecientemente matizado por declaraciones tan cáusticas como que «las PSE constituyen con toda seguridad el más idiota proceder jamás institucionalizado en el entrenamiento maquinal de los estudiantes de ciencia» y tan estridentes como que «las PSE no deberían siquiera existir, pues entrañan una bancarrota intelectual y son profundamente inconsistentes, tanto desde una perspectiva lógica como práctica»[3]. ¿Cómo explicarse aseveraciones de ese calibre sobre un procedimiento tan universalmente aplicado?

Las razones que subyacen son de diversa índole y complejidad variable, pero las esenciales resultan relativamente fáciles de comprender. Para comenzar, cabe precisar qué es y qué no es la p. La definición formal es la siguiente: asumiendo que la hipótesis nula H_o es válida, si el mismo experimento se repitiera infinitas veces, la frecuencia con la cual teóricamente obtendríamos un valor al menos tan alejado de lo que ella anuncia como el que objetivamente se obtuvo, es igual al valor p.

Esta definición es tan enrevesada que no sorprende que muy pocos sean realmente capaces de captarla adecuadamente. La mayoría de los usuarios, e incluso profesores de estadística e investigadores consagrados –así lo demuestran nítidamente investigaciones recientes[4]– están convencidos de que p mide la probabilidad de que H_o sea verdadera; un valor, digamos, $p = 0,03$ significaría que dicha hipótesis tiene una probabilidad de sólo un 3% de ser cierta.

Es una interpretación «comprensible», ya que el investigador evalúa hipótesis; de modo que nada más natural que el deseo de cuantificar el grado de validez que ellas merezcan. Lamentablemente, es una interpretación equivocada. La falacia se

aprecia con toda claridad si se tiene en cuenta que los valores p se calculan bajo el supuesto de que la hipótesis nula es verdadera; es imposible, por tanto, que pueda ofrecer una medida directa de la probabilidad de que ella sea, en efecto, verdadera.

El rasgo más claramente cuestionable del valor p estriba en que éste puede resultar tan pequeño como se desee en virtualmente cualquier situación práctica; basta con tomar una muestra suficientemente grande. Una PSE sólo permite que nos pronunciemos dicotómicamente acerca de si H_o puede considerarse falsa o no. Supongamos que se valora si dos hormigas son o no diferentes entre sí. Si se llegara a la conclusión de que no podemos aseverarlo, sólo hay una explicación: que las hormigas no se han examinado con suficiente detalle. Salvo contadísimas excepciones, usar una PSE para evaluar si dos tratamientos son o no iguales, es completamente análogo a emplear determinado instrumento con el único propósito de determinar si son o no iguales dos hormigas.

Basta que haya una minúscula diferencia, por intrascendente que sea (incluso el más mínimo sesgo), para que la diferencia cuya nulidad es proclamada por H_o pueda ser declarada «significativa»; sólo se trata de que se cuente con suficientes medios como para tomar una muestra adecuadamente grande.

Ésta es una imputación medular, pues nos dice que la decisión queda en manos de un elemento externo a la realidad que se examina, de modo que la respuesta a la pregunta formulada depende a la postre de los recursos disponibles. Hay algunos especialistas que defienden las PSE, pero no conozco un solo artículo, un solo libro, un solo profesor que consiga refutar o siquiera matizar esta gravísima objeción. No se me ocurre respaldo mayor para esa objeción que el clamoroso silencio que produce.

Ante la «no significación», es frecuente que se culpe al escaso tamaño de muestra. Algo patético, pues lo peor de la afirmación «no se ha encontrado significación, pero con una muestra mayor muy verosímilmente podríamos haberla hallado» es que *siempre* es verdadera. Es tan ridícula y estéril como afirmar: «no hemos podido corroborar que estas dos hormigas son diferentes; pero probablemente, con una lupa de mayor potencia, hubiéramos podido hallar suficientes indicios como para afirmarlo». Por otra parte, tiene tanta validez como otra afirmación que nunca se hace cuando el resultado es coherente con lo que desean los investigadores: «si hubiéramos trabajado con una muestra menor, muy posiblemente no hubiéramos encontrado significación». Este incuestionable doble rasero bastaría por sí mismo para señalar que algo no anda bien con las PSE.

Las PSE constituyen un proceso de decisión sobre una hipótesis (que se ha de rechazar o no); su tarea se reduce a dar elementos para pronunciarse dicotómicamente (significativo o no significativo; rechazo o no rechazo). Pero el pensamiento científico no discurre así realmente. La noción de encarar una hipótesis como si fuera una invitación al cine, que aceptamos o rechazamos, poco tiene que ver con el avance del conocimiento científico tal y como se ha verificado históricamente. Estamos, por tanto, ante otra de las graves insuficiencias que las PSE padecen en términos epistemológicos: conceptualmente, están concebidas para examinar resultados aislados. Para algunos autores[5] se trata directamente de «la más seria de sus deficiencias». Sólo la replicación y los procesos permanentes de autocorrección permiten ir forjando el consenso acerca de lo que provisionalmente consideramos como cierto. En el caso de las ciencias médicas y la salud pública, podría decirse que tal conocimiento se ha conseguido más bien *a pesar* del esquematismo propio de las PSE.

Es natural que algunos lectores, ante «noticias» como ésta, se pregunten si existen y cuáles serían las alternativas, los sucedáneos a las PSE. Puedo afirmar que existen, y que bien podría ser motivo para futuras y más pausadas contribuciones.

BIBLIOGRAFÍA

1. Sarria M, Silva LC. Las pruebas de significación estadística en tres revistas biomédicas: una revisión crítica. Revista Panamericana de Salud Pública. 2004;15:300-6.
2. Almenara J, Silva LC, Benavides A, García C, González JL. Historia de la bioestadística: la génesis, la normalidad y la crisis. Cádiz: Quórum S.A.; 2003.
3. Gill J. Grappling with Fisher's legacy in social science hypothesis testing. Journal de la Société Française de Statistique; 2004. Disponible en: psbla-de.ucdavis.edu/papers/denis.pdf en enero de 2008
4. Gigerenzer G, Krauss S, Vitouch O. The null ritual: what you always wanted to know about significance testing but were afraid to ask. En: David Kaplan, editor. The Handbook of Methodology for the Social Sciences [cap. 21]; 2004.
5. Howard GS, Maxwell SE, Fleming KJ. The proof of the pudding: an illustration of the relative strengths of null hypothesis, meta-analysis, and Bayesian analysis. Psychological Methods. 2000;5:315-32.