LAS PRUEBAS DE SIGNIFICACIÓN TIENEN UNA FUNCIÓN EN LA INVESTIGACIÓN EPIDEMIOLÓGICA: RESPUESTA A A. M. WALKER[1]

Joseph L. Fleiss[2]

Es indudable que tanto en epidemiología como en otras disciplinas se ha abusado de las pruebas de significación: las asociaciones o diferencias estadísticamente significativas se han considerado, erróneamente, equivalentes a asociaciones o diferencias importantes, y las asociaciones o diferencias estadísticamente no significativas se han considerado, erróneamente, iguales a cero. La inferencia apropiada que puede hacerse a partir de un resultado estadísticamente significativo es que se ha comprobado una asociación o diferencia distinta de cero; no tiene por qué ser necesariamente intensa, de tamaño considerable o importante; simplemente es distinta de cero. Igualmente, la conclusión apropiada deducida de un resultado estadísticamente no significativo es que los datos no han permitido comprobar la realidad del efecto investigado. Solo si el estudio tiene la potencia estadística adecuada será válida la conclusión de que no existe ningún efecto de importancia práctica. De lo contrario, a lo más que podemos llegar es al prudente "no demostrado".

En parte como reacción a los abusos debidos a malas interpretaciones, en el campo epidemiológico se está dando un movimiento que pretende erradicar de las publicaciones las pruebas de significación y los valores *P*. Un movimiento cuyo objetivo fuera mejorar la interpretación de las pruebas de significación contaría con la aprobación de la mayor parte de mis colegas estadísticos, pero lo que está ocurriendo es más preocupante. El autor de un trabajo enviado a una revista que publica artículos de temas epidemiológicos recibió de un editor la petición de que "*todas* (subrayado mío, J. F.) las referencias a pruebas estadísticas de hipótesis y significaciones estadísticas se eliminen del texto". Al autor de otro trabajo enviado a esa revista se le dijo que "estamos intentando desalentar el uso del concepto de 'significación estadística' que, a nuestro juicio, es obsoleto y puede confundir. Le pido por ello que elimine las referencias a valores *P* o 'significación estadística' ". El autor de un manuscrito enviado a otra de estas revistas fue informado por uno de sus editores de que "si no está de acuerdo con mis normas (relativas al carácter impropio de las pruebas de significación), siéntase libre para discutir ese extremo o, simplemente, desestime lo que para usted son mis puntos de vista equivocados y publique en otra revista. Sin embargo, como editor, sería inconcebible que aceptara trabajos que violan el principio científico que yo defiendo".

El insidioso mensaje que se está enviando a los que investigan en temas epidemiológicos es que las pruebas de significación no son válidas y no tienen lugar en la investigación, y que su uso en los artículos enviados a publicación producirá el rechazo o una demanda de revisión radical. Cuando se leen en el contexto del mo-

---

[1] Esta traducción del artículo "Significance tests have a role in epidemiological research: reactions to A. M. Walker" (*American Journal of Public Health*, 1986; 76(5): 559–560) se publica con autorización de la American Public Health Association y del autor. © American Public Health Association.

[2] Departamento de Bioestadística, Escuela de Salud Pública, Columbia University, Nueva York. Dirección postal: J. L. Fleiss, Head, Division of Biostatistics, Columbia University School of Public Health, 600 W. 168th Street, New York, NY 10032, Estados Unidos de América.

vimiento que he descrito, los argumentos aparentemente inocuos del Dr. Alexander Walker (*1*) sobre los valores *P* adquieren un significado diferente. El Dr. Walker cita la evaluación de los "términos de grado superior en los modelos estadísticos utilizados para analizar bases de datos muy ricas" como ejemplo de la utilidad de los valores *P* (de hecho, es su único ejemplo). La inferencia obvia es que las pruebas de significación posiblemente solo son útiles para investigar asuntos tangenciales y características poco importantes de los datos; por lo demás, están fuera de lugar.

Mi opinión es que, por el contrario, las pruebas de significación tienen una función válida en cada uno de los pasos importantes del análisis de los datos epidemiológicos. Los analistas que optan por utilizar estas pruebas e interpretan los valores *P* resultantes de manera prudente y apropiada no deben pedir excusas por hacerlo ni tolerar las demandas absurdas de reanalizar sus datos de manera contraria a lo que ellos piensan que es apropiado. Las siguientes son algunas de las aplicaciones válidas de las pruebas de significación en la investigación epidemiológica.

1. Refutación de resultados previos

Supongamos que una investigadora A encontró una asociación intensa entre los factores X e Y y que el investigador B piensa que la asociación es falaz. B no está interesado en determinar los límites del intervalo de confianza para la medida de la asociación entre X e Y, sino tan solo en demostrar que la asociación es nula. (La asociación entre consumo de café y cáncer de páncreas es un ejemplo.) Existen fórmulas para calcular el tamaño muestral y tablas (*2, 3*) que permiten a B diseñar un estudio con un número de sujetos suficiente para que un resultado estadísticamente no significativo sea aceptado como prueba en contra de la realidad (o, al menos, de la importancia práctica) de la asociación.

Sería deshonesto que B eliminara las referencias a las pruebas de significación en la descripción de los métodos del estudio o en la presentación de sus resultados, tanto si fue significativa la asociación que encontró entre X e Y como si no lo fue. Si la asociación no fue estadísticamente significativa, B habrá conseguido lo que quería. Si fue estadísticamente significativa, B está éticamente obligado a decirlo y, dado el contexto en el que el estudio está concebido, ello dará mayor credibilidad a la asociación.

2. Identificación de factores de confusión

Parece existir consenso respecto a qué condiciones biológicas y probabilísticas caracterizan, al menos en teoría, a una variable como factor de confusión (*3–5*). El que las condiciones biológicas resulten satisfechas o no, debe juzgarlo un conocedor de la materia, no un experto en estadística. Pero valorar las condiciones probabilísticas es un asunto puramente estadístico. Igual que en todos los problemas estadísticos importantes, hay varias soluciones razonables, no una sola solución correcta.

Una estrategia que mis colegas y yo aplicamos satisfactoriamente en un estudio prospectivo con muchos factores de confusión potenciales fue la siguiente. Primero, a partir del conocimiento de especialistas en la materia y publicaciones previas, identificamos 22 factores de confusión potenciales. Luego identificamos 15 de estos factores que se asociaban con la exposición, como mínimo a un nivel de sig-

nificación de 0,01. Finalmente, aplicamos el modelo de riesgo instantáneo proporcional de Cox[3] (*6*) y mediante la función de eliminación retrógrada de variables del programa BMDP2L (*7*) identificamos nueve de las 15 que eran relativamente independientes unas de otras y se asociaban con la variable de respuesta, como mínimo a un nivel de significación de 0,10 (o sea, 1/10). Solo entonces investigamos el efecto independiente del factor de riesgo hipotético sobre la respuesta.

Discutimos si ambos niveles de significación debían aflojarse o apretarse, pero nunca cuestionamos que fuera apropiado basarse en pruebas de significación para identificar qué subconjunto de los 22 potenciales factores de confusión deberíamos tener en cuenta en el análisis final. Nuestro ánimo era tomar decisiones —qué potenciales factores de confusión habrían de ser controlados y cuáles no— y las pruebas de significación proporcionaron reglas definidas *a priori*, tal como exige un proceso reproducible de toma de decisiones.

3. Análisis de subgrupo e interacciones

El Dr. Walker indica correctamente que las pruebas de heterogeneidad del efecto sobre subgrupos (es decir, de interacción) generalmente tienen escasa potencia. A mi parecer, eso es lo lógico, ya que está probado que las interacciones son notoriamente difíciles de reproducir. Una razón es que los errores aleatorios de clasificación al asignar los sujetos a los subgrupos pueden producir como artefacto una interacción aparente: una apariencia de asociación débil entre exposición y resultado en un subgrupo y asociación intensa en el otro (*8*). Traer a colación, como sugiere el Dr. Walker "observaciones relevantes ajenas al estudio" no es una salvaguardia adecuada contra los efectos espurios de subgrupo, tanto más cuando, con suficiente habilidad, es posible razonar la verosimilitud biológica de cualquier resultado imaginable.

Como fundamento para decidir si investigar o no las asociaciones dentro de los subgrupos, una prueba de significación formal para interacción proporciona un grado cuantificable de protección contra "cebos" que pueden ser solo artefactos. Yo recomiendo firmemente ese procedimiento.

4. Análisis de supervivencia por métodos no paramétricos

No es cierto que "(no existe) una observación epidemiológica importante que no (pueda) ser presentada claramente en unos pocos cuadros o tablas de datos numéricos en bruto y algunas estadísticas descriptivas sencillas". Los estudios en los que el resultado que se investiga es el tiempo de respuesta (período hasta la primera ocurrencia o la recidiva de la enfermedad, hasta la concepción, la muerte, etc.) constituyen un contraejemplo importante. Es la *función de supervivencia* estimada en conjunto, la serie de probabilidades de permanecer libre de enfermedad, o no embarazada, o vivo, la que constituye la estadística descriptiva "no tan simple". Cuando el investigador desea hacer ciertas suposiciones relativas a la forma matemática de la función de supervivencia (*6, 9, 10*), puede definirse una magnitud análoga al riesgo relativo (la probabilidad de que un miembro del grupo expuesto que durante las primeras T unidades de tiempo no ha experimentado el resultado investigado lo experimente durante el próximo intervalo momentáneo temporal, dividida por la probabilidad correspondiente para el grupo control) que es independiente de T. El riesgo relativo muestral puede

[3] *Cox's proportional hazard model* en inglés (*N. del t.*)

ser sometido a prueba de significación, pero la mayor parte de los analistas estarán de acuerdo en que un intervalo de confianza es igualmente informativo.

Sin embargo, supongamos que el investigador no desea hacer suposición alguna sobre la forma de la función de supervivencia y, en cambio, desea determinar si la función de supervivencia subyacente es en el grupo de expuestos la misma que en el grupo control. El análisis debe ser literalmente no paramétrico, ya que ni un solo parámetro ni un número finito de parámetros describe por completo las diferencias entre las dos funciones de supervivencia subyacentes. El procedimiento más popular para probar si las dos curvas difieren significativamente se denomina prueba del rango logarítmico[4] (*11*) o adaptación de Mantel (*12*) de la prueba de Mantel-Haenszel (*13*). Los epidemiólogos deberían mostrar tanto interés como los oncólogos o los cardiólogos en usar la prueba del rango logarítmico.

Intervalos de confianza múltiples

Pese a la descripción del Dr. Walker, el problema de las comparaciones múltiples implica los intervalos de confianza múltiples tanto como las pruebas de significación múltiples. De hecho, Scheffé (*14*) desarrolló su procedimiento clásico para comparaciones múltiples en un contexto de estimación de intervalos, no de pruebas de hipótesis. También Miller (*15*) subraya que la confianza que uno tiene en un intervalo dado que contiene la medida de asociación subyacente depende de si la asociación se especificó *a priori* como la de mayor interés científico o si fue señalada *a posteriori* durante una exploración *ad hoc* de todo el conjunto de datos.

El problema de cómo transmitir mejor la incertidumbre propia acerca de las asociaciones sugeridas por los datos aún no ha sido resuelto, pero un enfoque empírico usando técnicas bayesianas[5] parece prometedor (*16*). Lo que sugiere el Dr. Walker, presentar intervalos de confianza no ajustados "con un comentario apropiado", es ingenuo y, a mi entender, no se basa en teoría válida alguna de estimación de intervalos.

Conclusión

Algunos epidemiólogos creen que las pruebas de significación están muertas, y en algunas revistas han conseguido enterrarlas junto con los valores *P*. Hay que apoyar activamente a los investigadores que optan satisfactoriamente por estos procedimientos, ya que, pese a algunos, las pruebas de significación están vivitas y coleando.

Agradecimiento

Este trabajo fue financiado en parte por la beca DE 04068 del Instituto Nacional de Investigación Dental. Estoy muy agradecido por sus consejos y comentarios a mis colegas de la División de Bioestadística y Epidemiología de la Escuela

[4] *Log-rank test* en inglés (*N. del t.*).
[5] O sea, derivadas del teorema de Bayes (*N. del t.*).

de Salud Pública de la Columbia University, la División de Bioestadística de la Universidad de Iowa y el Centro Nacional de Estadísticas de Salud.

Referencias

1. Walker AM. Reporting the results of epidemiologic studies. *Am J Public Health*. 1986;76(5):556–558.
2. Fleiss JL. *Statistical methods for rates and proportions*. 2a ed. New York: Wiley; 1981.
3. Schlesselman JJ. *Case-control studies: design, conduct, analysis*. New York: Oxford University Press; 1982.
4. Breslow NE, Day NE. *Statistical methods in cancer research. Volume I, the analysis of case-control studies*. Lyon, France: IARC Scientific Pub. No. 32; 1980.
5. Kleinbaum DG, Kupper LL, Morgenstern H. *Epidemiologic research: principles and quantitative methods*. Belmont, CA: Lifetime Learning Publications; 1982.
6. Cox DR: Regression models and life tables. *J R Stat Soc*. (B) 1972;34:187–220.
7. Dixon WJ, Brown MB, Engelman L, et al. *BMDP statistical software*. Los Angeles: University of California Press; 1983.
8. Greenland S. The effect of misclassification in the presence of covariates. *Am J Epidemiol*. 1980;112: 564–569.
9. Glasser M. Exponential survival with covariance. *J Am Stat Assoc*. 1967;62:561–568.
10. Holford TR. Life tables with concomitant information. *Biometrics*. 1976;32:587–598.
11. Peto R, Peto J. Asymptotically efficient rank invariant test procedures. *J R Stat Soc*. (A). 1972;135: 185–206.
12. Mantel N. Evaluation of survival data and two new rank order statistics arising in consideration. *Cancer Chemother Rep*. 1966;50:163–170.
13. Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. *JNCI*. 1959;22:719–748.
14. Scheffé H. A method for judging all contrasts in the analysis of variance. *Biometrika*. 1953;40:87–104.
15. Miller RG. *Simultaneous statistical inference*. 2a ed. New York: Springer-Verlag; 1981.
16. Thomas DC, Siemiatycki J, Dewar R, et al. The problem of multiple inference in studies designed to generate hypotheses. *Am J Epidemiol*. 1985;122:1080–1095.