Endocrinol Diabetes Nutr. 2018;65(9):477–478

Endocrinología, Diabetes y Nutrición

www.elsevier.es/endo

EDITORIAL

¿Sería conveniente reducir el valor p considerado significativo?

Should we reduce the p value considered significant?

Ignasi Gich Saladich

Servicio de Epidemiología y Salud Pública, Hospital de la Santa Creu i Sant Pau, Barcelona, España

Disponible en Internet el 27 de septiembre de 2018

Para ser honesto con el lector, debo indicar que una parte importante de mi trabajo se basa en la utilización del contraste de hipótesis y la obtención de valores p, por tanto este texto no podrá tener una objetividad absoluta.

Pongamos que tenemos un grupo de sujetos sobre los que deseamos evaluar si una dieta consigue modificar sus valores de hemoglobina glicosilada (HbA1c), para estudiar el efecto obtendríamos el valor medio antes de iniciar la dieta y el valor medio posterior. El procedimiento más usual, sin entrar en análisis más complicados, sería comparar los 2 valores medios (con sus desviaciones típicas), por ejemplo: valor inicial del 6,8%, valor final 5,9%, la comparación (en este caso un test de «t» de medidas repetidas) nos da un valor p = 0,026.

La pregunta: ¿podemos confirmar que la dieta ha sido beneficiosa? Clínicamente es evidente la reducción del valor de HbA1c, así mismo, estadísticamente el resultado ha sido significativo; dicho de otra forma, tenemos evidencia de que la repetición del estudio con otro grupo de casos, similares a los de nuestro estudio, empleando el mismo diseño y con un número parecido de casos, arrojará valores similares.

El punto de corte para el rechazo de la igualdad se denomina nivel de significación y suele fijarse, de forma arbitraria, en un 5%. En el ejemplo 0,026 es inferior a 0,05, lo que nos permite decir que la diferencia no ha sido exclusiva de nuestros sujetos (en estos la diferencia es indiscutible), sino que parece razonable asumir que en la población de la que hemos extraído los casos se produciría dicha disminución.

Una expresión más formal, que habitualmente encontramos en la sección de métodos de un artículo científico, es decir que el valor del alfa se ha fijado en 0,05. Cuanto menor es el valor de p, mayor será la probabilidad de haber rechazado la hipótesis nula (que es la igualdad) y, por tanto, mayor será la probabilidad de que exista una diferencia real entre lo que se esté comparando.

Como ya he mencionado, el valor es arbitrario y todas las guías explicitan que, de forma razonada, puede modificarse, usualmente rebajándolo.

Es importante resaltar que la magnitud del efecto (en el ejemplo disminución del 6,8% al 5,9%) es el dato más importante y que siempre hay que discutir, no podemos ceñirnos a las consideraciones estadísticas sin matizar la relevancia clínica de los resultados.

Ahora puedo recordar la definición más formal del valor p: la probabilidad empírica de cometer un error de tipo I. Esto es, rechazar la igualdad y aceptar la existencia de la diferencia en nuestra comparación, aun y cuando no deberíamos. El principal inconveniente de esta forma de proceder es que no es posible, a partir de dicho valor p, conocer la magnitud del efecto y por ende de su relevancia clínica, que es en última instancia nuestro verdadero objetivo, como ya he comentado.

Así mismo la interpretación errónea de que el valor p indica la probabilidad de la que la H_0 (la igualdad) sea cierta, conlleva conclusiones incorrectas.

Abundando en interpretaciones falsas, en un artículo, cuyo último firmante fue el recientemente fallecido Dr.

Correo electrónico: igichs@santpau.cat

https://doi.org/10.1016/j.endinu.2018.09.001

Altman (Grenland, 2016)[1], se detallan multitud de errores y malas interpretaciones; de todas ellas resaltaré la declaración de que un resultado significativo implica un resultado clínicamente relevante. La magnitud del efecto no se puede interpretar a partir del valor de p.

Por todo ello, repetidamente a lo largo de los años, se ha postulado la necesidad de rebajar el nivel de significación, como apunta el Ioannidis (JAMA, 2018)[2]; en cualquier caso, la disminución del valor p no corrige del todo el problema. Una solución ha sido no considerar el valor p como significativo o no, sino facilitar la magnitud del efecto con su intervalo de confianza, que permite una evaluación de la precisión de dicha magnitud, por tanto, una visión claramente más clínica.

Recientemente se ha publicado una iniciativa, firmada por una gran grupo de expertos metodólogos (Benjamin et al.[3], 2018), sugiriendo la reducción del nivel de significación usualmente empleado de 0,05 a 0,005. La idea subyacente es mejorar la reproducibilidad de la investigación, de lo que se desprende que los autores consideran que los investigadores no son del todo rigurosos y que los resultados no son sólidos para las conclusiones de los artículos. En este sentido, Ioannidis se preguntaba por qué la mayoría de los resultados de las investigaciones publicadas, son falsos (Ioannidis, 2005)[4].

Un listado, no exhaustivo, de los problemas que expliquen la falta de validez, de las conclusiones, considerando solo el valor de p, incluiría: eliminación de valores, análisis de múltiples variables, comparaciones múltiples, comparaciones no previstas, análisis de subgrupos, aumentar la muestra hasta conseguir resultados significativos, etc., en definitiva, la tendencia a «torturar los datos», o simplificar la interpretación de los resultados de un estudio de acuerdo solo con el valor p.

La disminución del nivel de significación no evita el problema, pero presumiblemente lo convierte en menos prevalente y es una solución sencilla y fácil de implementar. En un amplio rango de las pruebas estadísticas usualmente empleadas, fijar el valor del alfa en 0,005, con la potencia usual del 80%, requeriría incrementos del tamaño de la muestra alrededor de un 70% (Benjamin et al., 2018[3]). Una ventaja, no despreciable, es el efecto sobre futuras investigaciones, puesto que los estudios con pocos casos tienden a exagerar las estimas de los tamaños del efecto, por tanto, los valores que un investigador extraiga de un estudio con un gran tamaño de muestra serán más robustos.

Otra forma de abordar el problema es utilizando la aproximación bayesiana, lo que implica añadir a nuestro análisis información de estudios previos (que se denomina «prior»), que puede mejorar la reproducibilidad de las conclusiones, como muestra Nuzzo (2014)[5], aproximación más compleja y que adolece del problema de la validez del «prior» empleado de los valores que se toman como referencia.

Como ya he indicado, la reducción del valor p puede entenderse como una distracción a la solución real, tal como la Asociación de Estadísticos Americana (Wasserstein y Lazar, 2016)[6] explicitó en su artículo, concluyendo que la buena práctica estadística, como componente esencial de la correcta práctica científica, enfatiza los principios del buen diseño del estudio y su realización, una variedad de resúmenes numéricos y gráficos de datos, la comprensión del fenómeno en estudio, la interpretación de los resultados en contexto, un informe completo y la adecuada comprensión lógica y cuantitativa de qué significan los datos.

Por todo ello, parece más razonable la elección de un buen diseño, una clara explicitación de los análisis previstos, de los detalles concretos utilizados para el cálculo del tamaño de la muestra, indicando claramente todas las exclusiones (en caso de haberlas) y todos los cálculos llevados a cabo, así como las variables evaluadas, en resumen, mejorar la calidad y la transparencia de las investigaciones, basadas en protocolos científicos previos, con un plan de análisis estadístico detallado.

Ningún índice en particular debe sustituir al razonamiento científico.

Agradecimientos

Al Sr. Ivan Solà Arnau, investigador del centro Cochrane Iberoamericano de Barcelona, y a la Dra. Carmen Fajardo, por sus indicaciones.

Bibliografía

1. Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN, et al. Statistical tests, P values, confidence intervals, and power: A guide to misinterpretations. Eur J Epidemiol. 2016;31(4):337–50.
2. Ioannidis JPA. The proposal to lower p value thresholds to.005. JAMA. 2018;319(14):1429–30.
3. Benjamin DJ, Berger JO, Johannesson M, Nosek BA, Wagenmakers EJ, Berk R, et al. Redefine statistical significance. Nat Hum Behav. 2018;2:6–10.
4. Ioannidis JPA. Why most published research findings are false. PLoS Med. 2005;2(8):e124.
5. Nuzzo R. Scientific method: Statistical errors. Nature. 2014;506(7487):150–2.
6. Wasserstein RL, Lazar NA. The ASA's Statement on p-values: Context, process, and purpose. Am Stat. 2016;70(2):129–33.