En defensa de la hipótesis nula: un comentario acerca de la significación estadística y la aceptación de la hipótesis nula[1]

Armando H. Seuc[2]

En este trabajo se discuten dos errores comunes al analizar los resultados de una investigación: confundir la significación estadística con la significación práctica y aceptar la hipótesis nula cuando la potencia del estudio es baja. Estos errores generalmente se manifiestan al analizar los resultados, pero generalmente se cometen al diseñar el estudio. Un error frecuente asociado con los dos anteriores es considerar que el aumento "excesivo" del tamaño muestral es perjudicial porque aumenta la probabilidad de detectar como estadísticamente significativa una magnitud del efecto que no lo es desde el punto de vista práctico.

Son varios los errores que se cometen con frecuencia al diseñar, ejecutar y analizar los resultados de una investigación científica. Las razones por las que esto ocurre son numerosas y objeto de polémica, pero entre ellas merecen atención especial la inadecuada preparación de los investigadores desde el punto de vista metodológico y estadístico, la reticencia a incluir en el equipo de investigación a especialistas en estos temas y la falta de discusión de estos problemas entre los investigadores.

Dos errores frecuentes que se cometen al analizar los resultados de una investigación biomédica son confundir la significación estadística con la significación práctica (clínica, epidemiológica, etc.) y deducir conclusiones importantes a partir de la aceptación (o "no rechazo" como se solía decir en los textos de estadística) de la hipótesis nula (*1*) cuando la potencia del estudio es baja. Sin embargo, estos errores se originan en la fase de diseño de la investigación, cuando no se hace corresponder la significación estadística con la significación práctica y no se utiliza un tamaño muestral adecuado que permita la deducción de conclusiones importantes a partir de la aceptación de la hipótesis nula. El denominador común de estas insuficiencias es en mi opinión una concepción inadecuada acerca de la interrelación entre la población objeto de estudio y la muestra mediante la cual se estudia dicha población.

---

[1] Al final de este texto, en pág. 222 sigue un comentario de la Redacción a propósito de los temas aquí planteados.

[2] Laboratorio de Metodología de la Investigación, Instituto Nacional de Endocrinología, Zapata y D, Vedado, La Habana 4, Cuba

Se ha mencionado también un supuesto efecto pernicioso del aumento del tamaño muestral que incrementaría la probabilidad de detectar una diferencia (o en general una magnitud del efecto) estadísticamente significativa que no lo es desde el punto de vista práctico.

El debate de estos problemas mantiene su vigencia debido a que las causas que lo originan aún subsisten.

La significación estadística y la hipótesis nula

Confundir la significación estadística con la significación práctica es ciertamente un error muy común. Este error se manifiesta a menudo en la figura del investigador que busca afanosamente la significación estadística "hasta debajo de las piedras". Más que considerar la significación estadística y la significación práctica como equivalentes, este investigador realmente desconoce qué es la primera y su relación con la segunda. Su interés por las pruebas estadísticas se produce esencialmente porque "es lo que se acostumbra aplicar" a los resultados de una investigación, y porque piensa que ello le facilitará publicar sus resultados en revistas de prestigio. En consecuencia, no sabe qué hacer con los resultados de las pruebas estadísticas y generalmente los subutiliza o malinterpreta (2).

Otra manifestación frecuente de este error se da en la figura del investigador que "no cree en la estadística". Este considera que los resultados obtenidos en la muestra particular observada son no solo lo único que se tiene, sino además todo lo que se necesita. Este fenómeno se manifiesta a veces como una falsa modestia: el investigador autolimita el alcance del estudio a la muestra particular analizada, encubriendo con ello su desconfianza o desprecio por los métodos de la inferencia estadística. Para este investigador la significación estadística es un concepto vacío.

Las dos situaciones antes descritas, aunque extremas y aparentemente opuestas, tienen una base común: no comprender que *la muestra es un medio para estudiar la población*. El que casi nunca pueda incluirse en el estudio a toda la población de interés (por problemas de factibilidad) no debe llevar a trasladar el interés científico al estudio de una muestra particular.

La aceptación de la hipótesis nula *cuando la potencia del estudio es baja* es también otro error muy común. Se ha convertido en un mito la idea de que la hipótesis nula nunca puede ser aceptada, al parecer debido a una subvaloración crónica de la importancia del tamaño muestral para obtener conclusiones importantes y con un margen de error pequeño a partir de los resultados de una investigación científica.

Es en el diseño de una investigación donde se cometen los errores fundamentales, en particular los dos antes mencionados, y donde es conveniente y oportuno evitarlos. En el diseño de una investigación debe establecerse cuál es la magnitud del efecto importante desde el punto de vista práctico que, de existir, sería deseable no dejar de detectar (como estadísticamente significativo). De esta manera la significación estadística se hace corresponder con la significación práctica, con lo que desaparece la contraposición artificial que se ha establecido entre estos dos conceptos.

Es también en el diseño de una investigación donde puede y debe decidirse el tamaño muestral necesario para que un resultado "negativo", es decir un resultado que no permita aceptar la hipótesis alternativa, sea informativo, de forma que el no poder aceptar la hipótesis alternativa permita en estas circunstancias *aceptar* la hipótesis nula. Si el tamaño muestral garantiza una potencia adecuada, la pro-

babilidad de que al aceptar la hipótesis nula nos estemos equivocando será suficientemente pequeña y por lo tanto no habrá ningún problema en aceptarla cuando no se pueda aceptar la hipótesis alternativa. O al menos el problema será el mismo que existe cuando se acepta la hipótesis alternativa (¡al no poder aceptarse la hipótesis nula!).

Algunos autores han planteado (*1, 3*) que la hipótesis nula es un invento estadístico que puede crear confusión, en contraposición a la hipótesis alternativa que se corresponde con la hipótesis de investigación. Sin embargo, la aceptación de la hipótesis nula (si la potencia es alta) puede ser de interés no solo para evitar el despilfarro de recursos que implica insistir en una línea de trabajo sin perspectivas, sino también porque en determinadas situaciones la hipótesis de investigación establece que hay "equivalencia" entre los tratamientos y en estos casos la hipótesis de investigación *es* la hipótesis nula (*4, 5*).

Por otro lado, considerar la igualdad entre los tratamientos como hipótesis nula es esencialmente un problema de conveniencia para su tratamiento matemático, y puede no ser necesariamente lo más adecuado desde el punto de vista práctico. Una discusión más detallada de este tema requeriría demasiados argumentos matemáticos, razón por la que no se insistirá en él.

Un artículo publicado en esta revista (*1*) presentó algunos ejemplos que pretenden ilustrar los peligros que acarrea la aceptación de la hipótesis nula *en general*, razón por la cual parece que dicho artículo contribuye lamentablemente al mito de la imposibilidad de aceptación de la hipótesis nula. Sin embargo, una revisión detallada de tales ejemplos pone de relieve los peligros que acarrea *no diseñar correctamente una investigación*. Los errores de diseño que se evidencian en esos ejemplos (*1*, pp. 49 y 52) y que en general se presentan con frecuencia en las investigaciones biomédicas son: i) una inadecuada elección del marco muestral; ii) un tamaño muestral insuficiente; y iii) falta de comparabilidad en variables importantes de los grupos experimental y de control en un ensayo clínico.

Otro criterio bastante generalizado, erróneo en mi opinión, es que una muestra excesivamente grande es perjudicial para los propósitos científicos de la investigación, cuando en realidad, el aumento del tamaño muestral puede ser un inconveniente solo desde el punto de vista del costo, factibilidad o tiempo invertido en la investigación y no, como se alega, porque la prueba estadística produzca entonces con mayor frecuencia resultados estadísticamente significativos que no lo son desde el punto de vista práctico. El siguiente ejemplo ilustra este tema.

Ejemplo

Supongamos que queremos estimar la diferencia en la efectividad de dos tratamientos. Para medir la efectividad de cada tratamiento se determinará la media aritmética de una variable cuantitativa de distribución aproximadamente normal. Supongamos además que las hipótesis que se van a contrastar son

$$H_0: |\theta| = |\mu_1 - \mu_2| = 0;\ H_1 : |\theta| > \delta$$

donde δ es un valor prefijado que representa la magnitud de la diferencia a partir de la cual se considera que hay significación práctica. Valores de θ tales que $0 < |\theta| \leq \delta$ a efectos prácticos son considerados dentro de H_0: $|\theta| = 0$. Esta separación entre la hipótesis nula y la alternativa es la que facilita la toma de una decisión acertada. La regla o procedimiento óptimo para decidir a partir de los resultados de una muestra particular la aceptación o no aceptación de H_0 para la situación antes descrita con-

siste usualmente (*6–8*) en aceptar (no rechazar) H_0 si el intervalo de confianza al nivel $1 - \alpha$ para θ,

$$(\hat{\theta} - Z_{(1-\alpha/2)} \cdot EE(\hat{\theta}),\ \hat{\theta} + Z_{(1-\alpha/2)} \cdot EE(\hat{\theta})),$$

tiene intersección no vacía con el intervalo $(-\delta, \delta)$, siendo $\hat{\theta}$ la estimación de θ y $EE(\hat{\theta})$ el error estándar de dicha estimación.

Aumentar el tamaño muestral de n a n' puede y debe llevar a un aumento de la potencia del estudio mediante la reducción del error estándar de la estimación de θ, y la consiguiente reducción en la amplitud del intervalo de confianza; obsérvese que si mantenemos las hipótesis nula y alternativa y la regla de decisión antes propuestas, es decir, aceptar H_0 si el intervalo

$$(\hat{\theta}' - Z_{1-\alpha/2} \cdot EE\hat{\theta}'),\ \hat{\theta}' + Z_{1-\alpha/2} \cdot EE(\hat{\theta}')),$$

tiene intersección no vacía con $(-\delta, \delta)$, se mantiene la correspondencia entre la significación estadística y la significación práctica. Dicho de otra forma, el aumento del tamaño de la muestra *no conduce necesariamente a que aumente la probabilidad de detectar como estadísticamente significativa una diferencia que no lo es desde el punto de vista práctico.*

Supongamos por ejemplo que se comparan dos hipoglucemiantes orales para diabéticos. Cada hipoglucemiante se evalúa mediante la reducción de antes a después del tratamiento de la glucemia en ayunas. El investigador desea considerar significativa una diferencia entre los tratamientos de 10 mg/dL. Bajo el supuesto de que la desviación estándar en mediciones repetidas de la glucemia en ayunas en un mismo individuo es también 10 mg/dL, resulta necesaria entonces (*9*) una muestra de 22 sujetos en cada grupo para $\alpha = 0{,}05$ (dos colas) y $1 - \beta = 0{,}90$. En estas circunstancias la regla de decisión óptima establece la aceptación (no rechazo) de

$$H_0 : |\theta| = |\mu_1 - \mu_2| = 0 \text{ mg/dL}$$

y la no aceptación de

$$H_1 : |\theta| > 10 \text{ mg/dL}$$

si el intervalo

$$(\hat{\theta} - 2EE(\hat{\theta}),\ \hat{\theta} + 2EE(\hat{\theta}))$$

tiene intersección no vacía con el intervalo $(-10, 10)$. Si aumentáramos el tamaño muestral a 40 sujetos en cada grupo, ello nos podría servir para uno de los siguientes propósitos (*9*):

i) aumentar la potencia del estudio de 0,90 a 0,99, manteniendo $\alpha = 0{,}05$ (2 colas) y ambas hipótesis (H_0: $|\theta| = 0$ mg/dL; H_1: $|\theta| > 10$ mg/dL);

ii) bajar α de 0,05 a menos de 0,01 (2 colas), manteniendo $1 - \beta = 0{,}90$ y ambas hipótesis (H_0: $|\theta| = 0$ mg/dL; H_1: $|\theta| > 10$ mg/dL);

iii) aumentar la potencia de 0,90 a 0,95, bajar α de 0,05 a 0,01 (2 colas), manteniendo ambas hipótesis (H_0: $|\theta| = 0$ mg/dL; H_1: $|\theta| > 10$ mg/dL);

iv) mantener la potencia en 0,90 y α en 0,05 (2 colas), reduciendo la magnitud de la diferencia significativa a 7,5 mg/dL, es decir, ahora H_0': $|\theta| = 0$ mg/dL, H_1': $|\theta| > 7{,}5$ mg/dL (la regla de decisión deberá modificarse consecuentemente; se aceptará H_0' si

$$(\hat{\theta}' - 2EE(\hat{\theta}'),\ \hat{\theta}' + 2EE(\hat{\theta}'))$$

tiene intersección no vacía con $(-7{,}5, 7{,}5)$).

Obsérvese que el aumento en la probabilidad de detectar como estadísticamente significativa una diferencia que no lo es desde el punto de vista práctico (propósito iv), no es una consecuencia *inexorable* del aumento del tamaño muestral. Ese aumento puede emplearse provechosamente para los propósitos i) a iii).

Agradecimiento

Una versión inicial de este manuscrito fue sustancialmente mejorada mediante sugerencias de forma y de contenido realizadas por los revisores.

REFERENCIAS

1. Sheehan TJ. Publicaciones médicas: que los lectores sepan a qué atenerse. *Bol Oficina Sanit Panam* 1994; 116(1):47–53.
2. Mainland D. Medical statistics: thinking vs arithmetic. *J Chronic Dis* 1982; 35:413–417.
3. Salsburg D. The use of statistical methods in the analysis of clinical studies. *J Clin Epidemiol* 1993;46(1):17–27.
4. Farrington CP, Manning G. Test statistics and sample size formulae for comparative binomial trials with null hypothesis of non-zero risk difference or non-unity relative risk. *Stat Med* 1990;9; 1447–1454.
5. Cooper EC. Designs of clinical trials: active control (equivalence) trials. *J Acquir Immune Defic Syndr* 1990;3(Suppl 2):S77–S81.
6. Altman DG. *Practical statistics for medical research*. London: Chapman and Hall; 1991.
7. Hoel PG. *Introducción a la estadística matemática*, 2 ed. Barcelona: Ariel; 1976.
8. Armitage P. Inference and decision in clinical trials. *J Clin Epidemiol* 1989; 42(4):293–299.
9. Machin D, Campbell MJ. *Statistical tables for the design of clinical trials*. Oxford: Blackwell; 1987.

Este manuscrito fue *recibido* el 28 de junio de 1994 y fue *aceptado*, tras revisión, el 13 de abril de 1995.

COMENTARIO

1. Tal como indica A. H. Seuc, no se debe confundir la significación *estadística* con la significación *práctica* (sea clínica, epidemiológica, terapéutica, demográfica o del tipo que corresponda). La significación estadística depende entre otras cosas del tamaño muestral, mientras que la significación práctica no depende de ese tamaño sino de consideraciones *ad hoc* (biológicas, psicológicas, socioeconómicas, etc.) que se hacen normalmente al margen de los resultados del estudio. Incluso, una diferencia (sea o no estadísticamente significativa) entre dos proporciones o promedios poblacionales puede ser *prácticamente* significativa desde un punto de vista (por ejemplo, clínico), y no serlo desde otro (demográfico o epidemiológico).

Supongamos, por ejemplo, que un método A de prevención de una enfermedad infrecuente evita 34% de los casos potenciales de ese padecimiento, mientras que un método B solo previene 33% de los casos. Tales resultados se obtuvieron de un gran estudio en el que varios miles de personas se asignaron aleatoriamente a

uno u otro método. La diferencia de porcentajes de prevención de la enfermedad resultó estadísticamente muy significativa ($P < 0,01$). Independientemente de ello, puede considerarse que, *a efectos de prevención*, la diferencia entre 34% y 33% no es importante (o sea, no tiene "significación práctica", o "significación preventiva") y, por tanto, para optar entre los métodos A y B lo que hay que considerar no es esa pequeña diferencia en poder preventivo —aunque sea "estadísticamente significativa"— sino otros factores (costos, facilidad de aplicación, etc.).

2. Como indica A. H. Seuc, es incorrecto pensar que se puede aceptar la hipótesis nula (o sea, la hipótesis de que no hay diferencia) simplemente porque se obtenga un resultado que no permita rechazarla. Por desgracia, esa práctica equivocada es frecuente en los manuscritos que se reciben en esta revista. A menudo el tamaño muestral pequeño hace que la potencia ($1 - \beta$) de la prueba de hipótesis sea tan pequeña que la probabilidad β de error tipo II (aceptar una hipótesis nula que es falsa) es muy alta.

3. En cierta forma, el enfoque de estimación de intervalos de confianza permite resolver bastantes de los problemas planteados por las llamadas pruebas (o "contrastes", o "dócimas") de hipótesis (*1*). El intervalo de confianza es un conjunto de valores construido de tal forma que la mayor parte de las veces contendrá el verdadero valor poblacional. El valor estimado orienta sobre la magnitud más probable de ese parámetro según los datos obtenidos. Claro está que cuando los intervalos de confianza son muy amplios, es muy inseguro lo que podemos decir del verdadero valor poblacional. La amplitud del intervalo de confianza depende (igual que depende el valor *P*) del tamaño muestral, pero el cálculo de intervalos de confianza se presta menos que el procedimiento de pruebas de hipótesis a la práctica de ocultar al lector los datos obtenidos (o no prestarles atención) bajo comentarios tales como "$P > 0,05$" o "$P = 0,23$" o "la diferencia no fue estadísticamente significativa".

En los manuscritos que se reciben en esta revista se usan poco los intervalos de confianza. Cuando los hay, no pocas veces están calculados por un método inadecuado al caso. Por ejemplo, a veces se calculan intervalos de confianza simétricos para proporciones cercanas a 0% o a 100%, obteniéndose un intervalo que incluye valores absurdos, bien negativos, bien de más de 100%.

4. Un enfoque alternativo, el de estimación del carácter probatorio (*2, 3*), permite valorar en qué medida los datos de un estudio favorecen una hipótesis concreta *en detrimento* de otra hipótesis dada. Ese enfoque todavía se aplica muy poco, pero parece sólidamente fundamentado desde un punto de vista teórico y probablemente irá ganando terreno en el futuro.

5. En general, es importante rechazar "la idea de que en los datos hay pruebas y verdades absolutas que pueden ser reveladas mediante técnicas estadísticas" (*2*). Por más que se parta de la consideración de una realidad objetiva que se estudia mediante métodos experimentales u observacionales, la interpretación de los resultados siempre exige juicios que implican la subjetividad del investigador.

6. Toda esta temática es objeto de continua reflexión y polémica en las publicaciones científicas. El *Boletín de la Oficina Sanitaria Panamericana* ha publicado varios artículos en esa línea (*2–5*). Los lectores interesados pueden consultar esas referencias y otras publicaciones (*6–10*), que de seguro solo son una pequeña parte de lo mucho que se ha escrito respecto a estos aspectos de la inferencia estadística.[1]

La Redacción

[1] Véase la nota bibliográfica al final de las referencias.

REFERENCIAS

1. Gardner MJ, Altman DG. Intervalos de confianza y no valores *P*: estimación en vez de pruebas de hipótesis. *Bol Oficina Sanit Panam* 1993; 114(6):536–549 (ed. orig.: *Br Med J* 1986: 292:746–750).
2. Goodman SN, Royall R. Carácter probatorio e investigación científica. *Bol Oficina Sanit Panam* 1993;115(3):235–249 (ed. orig.: *Am J Public Health* 1988; 78(12):1568–1574).
3. Goodman SN. Valores *P*, pruebas de hipótesis y verosimilitud: las consecuencias para la epidemiología de un debate histórico ignorado. *Bol Oficina Sanit Panam* 1995;118(2):141–155 (ed. orig.: *Am J Epidemiol* 1993;137(5):485–496).
4. Walker AM. Cómo presentar los resultados de los estudios epidemiológicos. *Bol Oficina Sanit Panam* 1994;115(2):148–154 (ed. orig.: *Am J Public Health* 1986;76(5):556–558).
5. Fleiss JL. Las pruebas de significación tienen una función en la investigación epidemiológica: respuesta a A. M. Walker. *Bol Oficina Sanit Panam* 1993;115(2):155–159 (ed. orig.: *Am J Public Health* 1986:76(5):559–560).
6. Bailar JC III, Mosteller F, eds. *Medical uses of statistics* 2nd ed. Waltham, MA: NEJM; 1992.
7. Thompson WD. Statistical criteria in the interpretation of epidemiologic data. *Am J Public Health* 1987;77:191–194.
8. Poole C. Beyond the confidence interval. *Am J Public Health* 1987;77:195–199.
9. Susser M. Falsification, verification, and causal inference in epidemiology: reconsiderations in the light of Sir Karl Popper's philosophy. En: Susser M. *Epidemiology, health, and society*. New York: Oxford University Press; 1987:82–93.
10. Rothman K, ed. *Causal inference*. Chestnut Hill, MA: Epidemiology Resources; 1988.

NOTA BIBLIOGRÁFICA

Los artículos mencionados como referencias *1–5* fueron incluidos en el libro *Publicación científica: aspectos metodológicos, éticos y prácticos en ciencias de la salud* (Washington, DC: Organización Panamericana de la Salud; 1994; Publ. Científ. 550). El libro de Bailar y Mosteller *Medical uses of statistics* (ref. *6*) es una recopilación de artículos, muchos de ellos anteriormente publicados en *New England Journal of Medicine*. Un capítulo específico se dedica a los valores *P* y otro a la importancia de β, el error tipo II y el tamaño muestral en el diseño e interpretación de ensayos clínicos. La obra de Susser que se menciona como fuente de la referencia *9* es de interés general y contiene varios capítulos sobre temas metodológicos. El correspondiente a esa referencia está incluido también en el libro *Causal inference*, compilado por Rothman (ref. *10*), que presenta diversas perspectivas (algunas de ellas polémicas y enfrentadas) del tema de la causalidad en epidemiología. Indudablemente, ese tema tiene muchos puntos de contacto con la inferencia estadística y las pruebas de hipótesis.