90e

COMUNICACIÓN BIOMÉDICA

Publicaciones médicas: que los lectores sepan a qué atenerse[1]

T. Joseph Sheehan[2]

Los médicos han de estar atentos a posibles trampas en las publicaciones; dos terceras partes de los estudios que aparecen en las mejores revistas médicas contienen conclusiones injustificadas. Hay dos errores muy comunes: el primero es confundir la significación estadística con la significación médica; el segundo, deducir conclusiones sustanciales de la aceptación de la hipótesis nula. En el lenguaje coloquial, significación indica importancia y significado; por el contrario, la significación estadística especifica la probabilidad de que el efecto observado pudiera haberse producido por variación aleatoria. La hipótesis nula es la hipótesis de ausencia de efecto experimental o de correlación. Puede aceptarse o rechazarse. Aceptar la hipótesis nula no demuestra que sea cierta; o sea, que no existe un efecto experimental o una correlación.

Como el ejercicio físico, la literatura médica puede ser dañina o beneficiosa. El ejercicio físico parece beneficiar a quienes lo practican con regularidad y prudencia. La lectura es una fuente importante de conocimiento, pero como han señalado Schor y Karten (*1*), dos terceras partes de los estudios que aparecen en las revistas médicas más exigentes tienen fallas de diseño e interpretación lo suficientemente graves para invalidar sus conclusiones.

Este artículo pretende revisar y mostrar a los lectores de publicaciones médicas dos errores comunes de interpretación. El primero se refiere a la hipótesis nula; el segundo, a lo que quiere decir la significación estadística. En la primera sección se revisan los conceptos principales. A continuación se presentan dos ejemplos tomados de publicaciones médicas. Los lectores familiarizados con los temas desde el punto de vista conceptual pueden pasar directamente a los ejemplos.

LA HIPÓTESIS NULA Y LA SIGNIFICACIÓN ESTADÍSTICA

Una hipótesis nula no es lo mismo que una hipótesis de investigación y significación estadística no es lo mismo que significación médica. Una hipótesis de

---

[1] Esta traducción del artículo "The medical literature—Let the reader beware" (*Archives of Internal Medicine* 1980; 140:472–474, © American Medical Association 1980/90) se publica con autorización de la American Medical Association y del autor.

[2] Departamento de Investigación en Educación para la Salud, Centro de Salud de la Universidad de Connecticut, Farmington. Las separatas de la edición original en inglés deben pedirse a: Dr. T. Joseph Sheehan, Department of Research in Health Education, University of Connecticut Health Center, Farmington, CT 06032, Estados Unidos de América.

investigación es una propuesta para explicar los hallazgos de la investigación. Por ejemplo, una hipótesis de investigación respecto a la relación entre tabaquismo y cáncer podría predecir una relación positiva entre ambos. Por el contrario, la hipótesis nula es un invento estadístico y puede crear confusión. La hipótesis nula es la hipótesis de que no existe efecto experimental o que no hay correlación;[3] afirma que nada sucedió o que no se hallará ningún efecto.

Los siguientes enunciados son ejemplos de hipótesis nulas:

- ☐ No hay diferencia en cuanto a resultados finales entre el grupo bajo tratamiento y el grupo que recibió placebo.
- ☐ No hay diferencia de efectividad entre el analgésico A y el B.
- ☐ No hay correlación entre ejercicio y pérdida de peso.

Generalmente se considera pedante formular la hipótesis nula en términos tan explícitos. Sin embargo, es importante que el lector pueda determinar la cuestión central sometida a estudio y expresarla en forma de hipótesis nula. Luego será más fácil comprender si el investigador ha lidiado con ella correctamente.

A menudo la ciencia va en contra de la intuición; la hipótesis nula es un buen ejemplo de un proceder que parece retrógrado. El investigador diseña su estudio de tal manera que, si su hipótesis de investigación es cierta, es bastante posible que sus datos lleven al rechazo de la hipótesis nula, a favor de la hipótesis de investigación. Intenta así probar que la hipótesis nula no es cierta. Por ejemplo, dada la hipótesis nula de que "no hay asociación entre fumar cigarrillos y padecer cáncer de pulmón", el investigador intenta diseñar su estudio de manera que pueda rechazar la hipótesis nula. Si tras aplicar la prueba estadística adecuada sus datos le permiten rechazar la hipótesis nula, se dice que sus resultados son estadísticamente significativos al nivel de 0,05 ($P < 0,05$), o al nivel de 0,01 ($P < 0,01$). Esto significa que habría menos de 5% o 1% de posibilidades, respectivamente, de encontrar una asociación de esta magnitud si operase el azar exclusivamente, o sea, si realmente en la población no hubiera asociación entre fumar cigarrillos y padecer cáncer de pulmón. Siendo tan baja la probabilidad de que la asociación se produzca por azar, el investigador está dispuesto a argüir que hay una asociación no casual entre fumar cigarrillos y cáncer de pulmón, y ha "demostrado" su hipótesis de investigación, en la medida en que esta puede demostrarse. Puede estar equivocado, pero especifica en qué medida es verosímil que lo esté.

Por el contrario, si el investigador no logra rechazar la hipótesis nula, intuitivamente podría concluirse que la hipótesis nula es verdadera y concluir así justificadamente que no existe relación entre fumar cigarrillos y cáncer de pulmón. Pero la intuición a veces es engañosa y este es un error habitual al que quizá no presten atención los lectores no experimentados. Una cosa es obtener una conclusión negativa (que no hay efecto o que no hay relación) y otra es no llegar a ninguna conclusión (el estudio no fue concluyente). La conclusión de "no hay un efecto" nos dice que el tratamiento no funcionó, o que fumar cigarrillos no produce daño. Esto no es lo mismo que decir que en el estudio no se alcanzó ninguna conclusión. Cuando la prueba estadística muestra que el azar podría haber generado los resultados observados, el investigador ha de ser cauteloso al considerar otras explicaciones. Que no se obser-

[3] En este artículo el autor usa el término correlación para referirse a dos conceptos que no son del todo idénticos. En algunos contextos (este por ejemplo), el término alude al cambio simultáneo de dos variables; en otros, el término es sinónimo de "coeficiente de correlación", valor entre −1 y 1 que cuantifica el grado de variación simultánea y que habitualmente se representa por *r*. (*N. del t.*)

vara nada no prueba que no sucediera nada. De manera que, si los datos no muestran un relación estadísticamente significativa entre fumar cigarrillos y cáncer de pulmón, es incorrecto concluir que no existe tal asociación. Es correcto decir que los datos de esta investigación no mostraron una relación, pero un estudio de mayor sensibilidad podría mostrarla.

Hay muchas razones para tomar con cautela las conclusiones surgidas de hipótesis nulas aceptadas. Una es que la significación estadística depende en gran medida del tamaño muestral. Supongamos que nos interesa la relación entre talla y peso en una población de corredores de fondo y que tomamos una muestra de nueve atletas de las listas de equipos de atletismo de las universidades locales. Medimos la talla y el peso de cada sujeto y calculamos la correlación entre talla y peso. Suponiendo que el estereotipo de corredor delgado sea cierto, esperaríamos obtener una baja correlación entre talla y peso. Ahora supongamos que por casualidad nuestra muestra resultó integrada por lanzadores de peso y obtuvimos una correlación de 0,55 entre talla y peso. (Elegir la muestra al azar no garantiza que la muestra elegida represente fielmente a la población. Lo que sí garantiza es que cada sujeto tiene igual posibilidad de ser elegido. Que la muestra resulte integrada exclusivamente por lanzadores de peso es así un ejemplo traído por los pelos; es posible, pero no es muy probable.) La teoría estadística nos dice que, en una muestra de nueve sujetos, una correlación debe ser de al menos 0,58 para que sea estadísticamente significativa al nivel de 0,05. Como 0,55 es menor que 0,58, no podríamos rechazar la hipótesis nula (o sea, la hipótesis de que la correlación en la población de la que fue tomada la muestra es cero). Como las correlaciones varían en magnitud entre -1,0 y +1,0, una correlación de 0,55 es bastante grande en valor absoluto, así que obviamente, no podríamos concluir que la correlación poblacional es cero. Tampoco podríamos concluir que hay una fuerte relación positiva, porque el resultado negativo de la prueba de significación estadística nos alerta de la posibilidad de que una correlación hasta de 0,55 pueda obtenerse al azar al tomar nueve pares de observaciones de una población en la que la verdadera correlación es cero.

En nuestro ejemplo, presuntamente sabíamos que la correlación entre talla y peso de los corredores es cercana a cero, de manera que los resultados de la prueba estadística no nos llevaron a una conclusión errónea. Estábamos en lo cierto al no rechazar la hipótesis nula. Sin embargo, si la correlación muestral hubiese sido ligeramente mayor, digamos de 0,59, habríamos rechazado la hipótesis nula erróneamente. Como sabemos que la verdadera correlación entre talla y peso de los corredores es cero, habríamos cometido un error, un error de tipo I, rechazar la hipótesis nula cuando es verdadera.

Por otro lado, si nos interesa la correlación entre talla y peso en una población de hombres adultos, que cabe suponer es de alrededor de 0,55, y por azar nuestra muestra resulta integrada por corredores delgados, podríamos errar en la otra dirección. Por ejemplo, es posible que obtuviéramos una correlación de tan solo 0,12, en cuyo caso aceptaríamos la hipótesis nula erróneamente. El error de aceptar la hipótesis nula cuando es falsa se llama error de tipo II. El ejemplo nos muestra lo engañoso que puede ser deducir conclusiones sustanciales cuando se acepta la hipótesis nula. Es tentador concluir que no existe relación entre talla y peso. A menudo se sacan conclusiones como esta en casos en los que se acepta la hipótesis nula. El riesgo de cometer un error de tipo II, riesgo que es muy grande en casos de muestras pequeñas, hace que estas conclusiones raramente estén justificadas.

Aumentar el tamaño muestral protege contra los errores de tipo II porque la desviación estándar del estadístico de la prueba disminuye proporcional-

mente a la raíz cuadrada del tamaño muestral. Por ejemplo, como hemos dicho, con una muestra de nueve observaciones, la correlación observada debe ser por lo menos de 0,58 para ser estadísticamente significativa a un nivel de $P < 0,05$; con un tamaño muestral de 26, la correlación observada debe ser por lo menos 0,33 para ser significativa; con un tamaño muestral de 100, solo debe ser 0,17. Además, la posibilidad de cometer un error de tipo II a menudo puede predecirse al planificar el estudio. Diseñar un estudio y calcular el tamaño muestral a efectos de disminuir el error de tipo II no parece ser una práctica común en la investigación clínica, como señaló una revisión reciente de más de 300 ensayos clínicos controlados (2). En más de dos terceras partes de 71 estudios que pasaban otros criterios de selección, era muy alta la probabilidad de cometer un error de tipo II, y en todos los casos esto se podía haber sabido antes de realizar el estudio. Además, en casi todos los casos en que se aceptó la hipótesis nula se concluyó que el tratamiento no tenía efecto, a pesar de que en algunos casos se hallaron diferencias de hasta 25%, lo que representa una mejora clínica significativa.

Finalmente, aun cuando se rechaza la hipótesis nula, hay que ser cuidadoso y no confundir la significación estadística con la significación médica. De la misma manera que una muestra demasiado pequeña puede hacer que un investigador acepte la hipótesis nula cuando realmente es falsa, diferencias muy pequeñas o correlaciones basadas en muestras muy grandes pueden resultar estadísticamente significativas a pesar de ser triviales por su magnitud desde el punto de vista médico. Lo correcto es que el clínico especifique antes de realizar el estudio qué magnitud del efecto del tratamiento o qué correlación consideraría clínicamente relevante. Entonces es posible especificar un tamaño muestral que no sea demasiado pequeño ni demasiado grande para detectar la correlación o el cambio clínicamente relevante. En este caso no importa si se confunde la significación estadística con la significación médica, porque ambas se refieren a lo mismo.

Ejemplo 1: Significación estadística frente a significación médica

Recientemente, los resultados más importantes de un estudio sobre consumo de alcohol y niveles de lipoproteínas (3) fueron resumidos por un destacado cardiólogo en una conocida revista destinada al público femenino, además de aparecer en numerosas revistas de actualidad y en publicaciones médicas. En todos estos sitios se reiteraron las engañosas interpretaciones contenidas en el artículo original. Veamos primero los datos y luego las interpretaciones.[4]

Los investigadores midieron el consumo de alcohol y las concentraciones hemáticas de lípidos en cinco poblaciones distintas y luego calcularon la correlación para cada grupo. Las cinco correlaciones entre consumo de alcohol y triglicéridos fueron 0,04, 0,07, 0,12, 0,11 y 0,13. Las cinco correlaciones entre consumo de alcohol y colesterol de lipoproteínas de baja densidad (colesterol LDL) fueron $-0,04$, $-0,35$, $-0,09$, $-0,24$ y $-0,23$. Las cinco correlaciones entre consumo de alcohol y colesterol de lipoproteínas de alta densidad (colesterol HDL) fueron 0,28, 0,19, 0,30, 0,28 y 0,25. Los autores ofrecen la siguiente interpretación de estas correlaciones:

[4] Lo que aquí se pretende no es avergonzar a los autores; pero sin ejemplos específicos e importantes sería imposible ilustrar aspectos que el lector sin formación estadística puede comprender y que son muy comunes en las publicaciones médicas. Partiendo de esa idea, agradezco los acertados comentarios de un revisor anónimo. (*N. del autor*)

> Frente a trabajos anteriores, el cuadro que se deduce de los resultados del Estudio Cooperativo de Fenotipaje de Lipoproteínas (Cooperative Lipoprotein Phenotyping Study) es bastante claro. El consumo de alcohol se encuentra *moderadamente* asociado con incrementos de los triglicéridos, de *moderada a fuertemente* asociado con disminución del colesterol LDL, y *fuertemente* asociado con incrementos del colesterol HDL [subrayados añadidos].

¿Cuál es la base para afirmar que estas correlaciones son "moderadas" o "fuertes"? Y, antes de responder a esa pregunta, ¿son las correlaciones observadas estadísticamente significativas? No consta de ninguna manera que se hayan hecho pruebas de significación estadística para estas correlaciones; o sea, que no parece que se haya sometido a prueba la hipótesis nula según la cual el verdadero valor de las correlaciones en la población sería cero. La cuestión de la significación estadística resulta una pista falsa en este estudio, ya que la falta de pruebas de significación no es una omisión seria, por el tamaño muestral extremadamente grande. Con muestras tan grandes, incluso correlaciones de 0,11 pueden ser estadísticamente significativas. Pero lo que importa es la significación médica, no la significación estadística. Incluso habiendo una asociación no debida al azar entre consumo de alcohol y niveles de lípidos en sangre, ¿es correcto caracterizar las correlaciones observadas como "moderadas", "de moderadas a fuertes" y "fuertes"? La significación estadística indica simplemente que es pequeña (menor que 0,05 o que 0,01) la probabilidad de que la correlación observada provenga de una población en la que la verdadera correlación es cero. La significación estadística es muy distinta de la significación médica, salvo cuando las correlaciones son de gran magnitud.

¿Cómo se juzga si una correlación es grande? No hay un estándar absoluto, porque el juicio depende de las variables que se estén correlacionando, pero una posibilidad es elevar la correlación al cuadrado. El cuadrado de la correlación es una estimación de cuánto varían en conjunto los fenómenos asociados. Por ejemplo, la correlación de 0,5 a menudo observada entre el cociente intelectual (CI) y las calificaciones escolares indica que 25% de la variación de las calificaciones escolares se asocia con el CI. De manera que se podría afirmar que hay una relación "moderada" entre CI y calificaciones. Por supuesto que no hay reglas fijas ni seguras, pero en vista de que otro 75% de la variación del rendimiento escolar no puede explicarse por el CI, parece razonable concluir que la asociación es "modesta".

En el estudio de las lipoproteínas, la máxima correlación en la categoría de asociación "fuerte" es 0,30, lo cual da cuenta de 9% de la variación de las concentraciones de HDL. La máxima correlación en la categoría "de moderada a fuerte" es 0,35, lo que da cuenta de 12% de la variación simultánea. En cuanto a la categoría "moderada", la máxima correlación es 0,13, lo que da cuenta de menos de 2% de la variación en común. Siendo tan escasa la variación compartida entre estas variables, es difícil justificar los calificativos de "fuerte" y "de moderada a fuerte". Es posible que todas las correlaciones sean estadísticamente significativas, pero no parecen tener significación médica.

Quizá los autores hicieron pruebas de significación estadística, pero luego confundieron significación estadística con significación médica. De no ser así, es inexplicable que utilicen calificativos como "moderada" o "fuerte" al referirse a correlaciones de tan escasa magnitud. Afirmar que las correlaciones son estadísticamente significativas no nos dice nada de su significado o importancia práctica o biológica; simplemente significa que hay una baja probabilidad de que las correlaciones observadas hubieran ocurrido por casualidad si la verdadera correlación en la población de la que proceden las observaciones fuese cero.

Ejemplo 2: La hipótesis nula

El siguiente ejemplo muestra por qué es peligroso sacar conclusiones de un estudio en el que no se ha podido rechazar la hipótesis nula.

Hace mucho tiempo que los clínicos utilizan el diámetro torácico anteroposterior (AP) para diagnosticar el enfisema. Sin embargo, los autores del estudio de nuestro segundo ejemplo no pudieron encontrar estudios empíricos que justificaran esta costumbre (4). Por ello, midieron el diámetro AP del tórax en un grupo de pacientes enfisematosos hospitalizados y compararon los resultados con los de otros dos grupos de pacientes hospitalizados por otras enfermedades y de profesionales sanitarios sin síntomas respiratorios. El diámetro AP medio fue de 20,2 cm para los sujetos sanos y de 23,0 cm para los pacientes enfisematosos. Basándose en la hipótesis nula de inexistencia de diferencia entre las medias, realizaron una prueba estadística no especificada. No encontraron una diferencia estadísticamente significativa en diámetro AP y por lo tanto no pudieron rechazar la hipótesis nula. Concluyeron que no había diferencia entre los grupos y arguyeron que el diámetro AP quedaba así desacreditado como signo diagnóstico útil de enfisema. Los autores sostuvieron que habían "echado abajo la vieja y repetida cantinela según la cual un diámetro AP aumentado es signo común y útil de enfisema" (4). Y agregaron: "Esperamos que el rechazo de esta creencia errónea lleve en el futuro a su eliminación de los tratados de diagnóstico, medicina y enfermedades pulmonares". He oído que se han incluido preguntas basadas en esas conclusiones en los exámenes de acreditación de la especialidad de medicina interna.

¿Hay algún error en todo esto? Este caso es claramente uno de esos en los que se deducen conclusiones de una hipótesis nula aceptada. Una hipótesis nula explícita sería que "no existe diferencia en diámetro AP medio entre sujetos normales y pacientes con enfisema". Los autores no pudieron rechazar la hipótesis nula basándose en su prueba estadística no especificada y concluyeron así que la hipótesis nula era verdadera, o sea, que no existen diferencias de diámetro AP entre sujetos normales y pacientes enfisematosos. ¿Podría haber otras razones por las cuales no pudo rechazarse la hipótesis nula?

Veamos nuevamente los grupos que se compararon. La edad media en el grupo de sujetos sanos es de 30,9 años, en contraste con una media de 56 años en el grupo de pacientes enfisematosos. El grupo de sanos tiene un peso medio de 74,5 kg y una altura media de 179,2 cm, mientras que el grupo de pacientes enfisematosos tiene un peso medio de 62,1 kg y una altura media de 174,4 cm.

¿Qué significa tener un grupo control más joven, más alto y más pesado? Es lógico que sujetos más altos y más pesados tengan un diámetro AP proporcionalmente mayor que sujetos más bajos y menos pesados, aun siendo similares en otros aspectos. ¿Cómo afecta esto a la hipótesis nula? Los tórax más grandes que esperaríamos encontrar en el grupo control podrían reducir cualquier diferencia entre el grupo control y los pacientes. Pero esta misma reducción genera un sesgo a favor de la aceptación de la hipótesis nula. Sin embargo, los datos muestran que a pesar de este sesgo, el diámetro AP de los sujetos normales sigue siendo menor que el de los pacientes enfisematosos.

Parece razonable esperar que la diferencia fuese incluso mayor si el grupo control fuese comparable. El hecho de que el grupo control no sea comparable es una buena razón para no sacar conclusiones a partir de la aceptación de la hipótesis nula en este caso.

Hay una ironía final en este estudio del diámetro AP. No solo se dedujeron conclusiones a partir de la aceptación de la hipótesis nula, sino que, además, si se aplica una prueba *t* a los datos publicados, se rechaza la hipótesis nula. La discusión que se ha presentado cuestiona las conclusiones del estudio, pero los resultados de la prueba *t* contradicen dichas conclusiones por completo. Cuando se compara el diámetro AP medio de 20,2 cm del grupo de sujetos normales con el diámetro AP medio de 23,0 cm del grupo de pacientes enfisematosos, teniendo en cuenta las desviaciones estándar y los tamaños muestrales indicados en el artículo, la diferencia es estadísticamente significativa ($P < 0{,}001$), lo cual indica que hay indicios más que suficientes para resucitar el mito de los clínicos, e incluso para reintroducirlo en los textos de medicina.

COMENTARIO

Este artículo se ha circunscrito a la discusión de dos puntos: la significación estadística frente a la significación práctica y la deducción de conclusiones a partir de la aceptación de la hipótesis nula. Hemos dado solo un ejemplo de cada caso, pero el lector debe comprender que estos problemas son muy comunes y muy importantes. Ya dijimos anteriormente que en un artículo reciente en *The New England Journal of Medicine* (2) en el que se revisaron 71 ensayos clínicos aleatorizados con resultados "negativos", se señaló claramente que, al menos en dos tercios de esos estudios, podían predecirse los resultados negativos antes de realizar el estudio, ya que era muy alta la probabilidad de no rechazar la hipótesis nula (cometer un error de tipo II). El artículo también indica que "en la mayor parte de los estudios, la ausencia de una diferencia significativa al nivel de 5% fue interpretada como indicativa de que no existe diferencia clínicamente significativa". Esto, a pesar de que en muchos de estos ensayos las diferencias entre el grupo bajo tratamiento y el grupo control podrían haber sido de hasta 25%.

REFERENCIAS

1. Schor S, Karten I. Statistical evaluation of medical journal manuscripts. *JAMA* 1966;195:1123–1128.
2. Freiman JA, Chalmers TC, Smith H Jr, et al. The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial: survey of 71 "negative" trials. *N Engl J Med* 1978;299:600–694.
3. Castelli WP, Gordon T, Hjortland M, et al. Alcohol and blood lipids: The Cooperative Lipoprotein Phenotyping Study. *Lancet* 1977;2:153–155.
4. Kilbum KH, Asmurdsson T. Anteroposterior chest diameter in emphysema: from maxim to measurement. *Arch Intern Med* 1969;123:379–382.