Diferencias estadísticamente significativas vs. relevancia clínica

Statistically significant differences vs. clinical relevance

MAURICIO BARRERA VALENCIA[1]

Forma de citar: Barrera M. Diferencias estadísticamente significativas vs. relevancia clínica: Rev CES Med 2008; 22 (1): 89-96

RESUMEN

En este artículo se hace una revisión de los peligros que conlleva el uso del término significación estadística y la importancia de analizar la magnitud de las diferencias que se encuentran al final de los estudios de investigación. Para ello, se hace una presentación del concepto de significación estadística, los errores tipo I *y tipo* II *y del concepto de relevancia clínica. Asimismo, se discute el uso de otro tipo de medidas como son los intervalos de confianza. Finalmente se presentan, a manera de conclusión, dos ideas básicas: la primera tiene que ver con la importancia de identificar la prueba estadística que mejor se ajuste al estudio para rechazar o aceptar la hipótesis nula y la necesidad de establecer si la magnitud de las diferencias obtenidas tienen alguna importancia desde el punto de vista clínico.*

PALABRAS CLAVE

Significación estadística
Valor p
Intervalos de confianza
Investigación

[1] MSc. Profesor Universidad de Antioquia. E mail: maobarrera@une.net.co

Recibido: 22 noviembre / 2007. Revisado: 24 febrero / 2007. Aceptado: 7 abril / 2007

SUMMARY

This article reviews the potential hazards of using the term 'statistical significance' as well as the importance of analyzing size effects of differences found at the research reports articles. Thus, this article presents a review of concepts like statistical significance, type I and type II errors, and clinical relevance. Similarly, a discussion regarding other statistical measures, such as confidence intervals, is presented. At last, two ideas are presented as main conclusions of this analysis: the first deals with the importance of identifying the best statistical tests to either accept or reject the null hypothesis in a research study. The second idea highlights the need of clarifying the clinical relevance of differences' size effect.

KEY WORDS

Statistical significance
p value
Confidence intervals
Research

Las pruebas de significación, tal y como usualmente se usan en la actualidad, surgieron de la fusión de dos métodos: uno desarrollado por Fisher en los años veinte, que permitía valorar el grado de incompatibilidad de los datos con una hipótesis y el otro formulado por Neyman y Pearson, en los años treinta, que se basaba en la elección entre dos hipótesis. A mediados de los años cuarenta, los dos métodos se unen, en un intento por conciliar estas perspectivas originalmente contrapuestas y dan lugar a la prueba estadística de hipótesis que actualmente se conoce. El método toma de Fisher su valor ***p*** para ser usado como un índice que mide la fuerza de la evidencia y de Neyman y Pearson, el propósito de adoptar una decisión consistente en rechazar la hipótesis nula si el valor de ***p*** es pequeño (normalmente, cuando $p<0.05$) y en no rechazar la hipótesis nula, si el valor de *p* es más grande (1).

Desde entonces, y particularmente en los últimos treinta años, el uso de la prueba de hipótesis en las revistas de ciencias de la salud se ha incrementado enormemente, buscando dar un soporte más objetivo a las conclusiones que sacan los investigadores de sus estudios. Sin embargo, una consecuencia desafortunada de este auge ha sido el énfasis que se le ha otorgado al resultado en términos del valor obtenido, olvidando el significado clínico que se desprende de este resultado. Así, en esta prueba, los datos se examinan con relación a una hipótesis estadística "nula", práctica que ha llevado a la creencia errónea de que el objetivo de los estudios debe ser obtener una significación estadística, cuando en realidad el objetivo de la mayor parte de las investigaciones en ciencias de la salud, es determinar la magnitud de algún factor, evento o relación objeto de estudio (2).

Al respecto Fleiss afirma: "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, también erróneamente, iguales a cero". Y más adelante concluye: "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" (3).

Un aspecto, que hace más complejo el tema es el hecho de que, al estar hablando en términos de probabilidades (de hecho por esa razón se emplean las herramientas de la estadística) las posibilidades de encontrar en un estudio una diferencia "estadísticamente significativa" se aumentan a medida que aumentan el número de comparaciones.

Para comprender más claramente lo expuesto hasta aquí, vale la pena revisar los conceptos de nivel de significancia estadística y los errores tipo I y II.

En general las pruebas de significación estadística, se aplican usualmente en estudios analíticos que buscan identificar asociaciones causales entre la exposición a factores de riesgo y la presencia de eventos mórbidos. Se debe advertir claramente que el establecimiento de una asociación se fundamenta en la detección de una diferencia, y que la obtención de esta diferencia entre los grupos comparados puede ser, en principio, el resultado del simple azar, y no como la expresión de una diferencia real existente entre los grupos comparados (4).

Ahora bien, dado que es virtualmente imposible obtener resultados con un 100% de certeza, el investigador de forma a priori, define un valor de certeza con el cual aspira obtener resultados confiables de su estudio. Este valor seleccionado es lo que se conoce como nivel de significancia y hace referencia a la probabilidad de obtener en un estudio un valor tan extremo como el realmente observado si la hipótesis nula fuera cierta (5). Para clarificar mejor este punto, es conveniente explicar el concepto de probabilidad, mediante un ejemplo dado por Wiersman (citado por Hernández y col):

"La probabilidad de que un evento ocurra oscilará entre 0 y 1, donde 0 significa la imposibilidad de ocurrencia y 1 la certeza de que ocurra el fenómeno. Al lanzar al aire una moneda, la probabilidad de que salga sello es de 0,5 y la probabilidad de que salga cara es de 0,5. Si se lanza un dado, la probabilidad de obtener cualquiera de sus lados es de 1/6= 0.1667. En ambos casos la suma de las posibilidades siempre es igual a 1" (6).

Con base en este concepto de probabilidad, el investigador elige de manera arbitraria qué tanto error está dispuesto a aceptar en los resultados de su estudio. Esa elección le permite al investigador decidir si sus resultados son válidos o no, aún sabiendo que siempre existirá la posibilidad de equivocarse en su decisión. En otras palabras, siempre existe la posibilidad de realizar una investigación cuyos resultados señalan una diferencia, cuando realmente dicha diferencia no existe. Por tradición se han aceptado generalmente los valores de 0,05 o 0,01 (lo que conlleva a tener una seguridad de acierto del 95 o 99% respectivamente). Este es el valor que se denomina valor "***p***" y su interpretación puede ser expresada en términos generales del siguiente modo "un resultado que es significativo al nivel de 0.05 puede ocurrir por azar no más de 5 veces en 100 ensayos" (7).

Sin embargo, así como es posible confirmar la hipótesis alterna con base en los resultados obtenidos siendo esta falsa, también es posible rechazar dicha hipótesis (es decir aceptar la hipótesis nula) cuando en realidad era verdadera. A esta particularidad es a lo que se le denomina error tipo I y error tipo II, respectivamente (para una presentación más detallada ver tabla 1) (8, 9).

Tabla 1. RESULTADOS POSIBLES DE UN ESTUDIO

RESULTADO DE UN ESTUDIO	VERDAD EN LA POBLACIÓN	
	La Hipótesis Nula es Correcta	**La Hipótesis Alterna es Correcta**
Se confirma Hipótesis Nula	*Verdadero negativo*	Error tipo II o falso negativo
Se confirma Hipótesis Alterna	Error tipo I o falso positivo	*Verdadero positivo*

Para evitar alguno de estos errores, los investigadores concentran sus esfuerzos en seleccionar adecuadamente la muestra, controlar aquellas variables que puedan afectar los resultados de sus mediciones, emplear instrumentos válidos y confiables para medir la variable de interés y seleccionar un diseño de la investigación que pueda dar respuesta a las preguntas de investigación que se formulan. Es por estas razones que el concepto de nivel de significancia estadística es tan importante para el estudio, pues de alguna manera brinda la confianza necesaria al investigador de que sus resultados no son producto del azar.

Desafortunadamente, esta forma de proceder no es infalible y puede dar lugar a conclusiones erróneas. La objeción más importante a este método proviene precisamente de la naturaleza de los valores ***p***, ya que el rechazo o la aceptación de una hipótesis resulta ser en la mayoría de los casos, fruto del tamaño de la muestra. Al respecto, Silva y Benavides sostienen:

...El rechazo o la aceptación de una hipótesis resulta ser, simplemente, un reflejo del tamaño de la muestra: si esta es suficientemente grande, siempre se rechazará la hipótesis nula. Esto nos coloca en una aparente paradoja: cuando operamos con una parte muy pequeña de la realidad (una muestra muy pequeña), entonces no podemos obtener, conclusión alguna, como es lógico e intuitivo, lo cual conduce a que muchos investigadores, cuyos resultados no alcanzan la esperada significación estadística, proclaman que con un tamaño de muestra mayor casi seguramente lo hubieran logrado. Lamentablemente y esto es lo realmente grave, tienen razón; pero eso significa que tampoco se puede sacar nada en claro cuando se trabaja con una muestra muy grande, puesto que en tal caso el rechazo de la hipótesis nula queda virtualmente asegurado (10).

Por esta razón, autores como Gardner y Altman, (2) afirman que pequeñas diferencias sin interés real pueden ser estadísticamente significativas cuando el tamaño de la muestra es grande, mientras que efectos clínicamente importantes pueden no ser estadísticamente significativos solo porque el número de sujetos estudiados fue escaso.

Otro aspecto, tiene que ver con el juicio que hace el investigador de sus resultados. Aún aceptando que al comparar los valores de las medidas, estas arrojen una diferencia estadísticamente significativa, pocas veces el investigador se preocupa por evaluar el tamaño de dicha diferencia. Dicho de otra forma, el investigador concentra todos sus esfuerzos en evitar resultados productos del azar, pero olvida analizar si dichos resultados tienen alguna relevancia clínica.

De acuerdo a Fernández y Díaz (11) y Coolican (12), la relevancia clínica de un fenómeno va más allá de cálculos aritméticos y está determinada por el juicio clínico. La relevancia depende de la magnitud de la diferencia, la gravedad del problema a investigar, la vulnerabilidad, la morbimortalidad generada por el mismo, su coste y por su frecuencia entre otros elementos.

En este sentido Sarria y Silva plantean que "A menudo se olvida que el análisis estadístico es solo un elemento más que ha de sumarse al arsenal de conocimientos científicos e información aportada por estudios anteriores para configurar una conclusión. En consecuencia, se comenten muchos errores, tales como convertir en una conclusión algo que no pasa de ser un resultado. En este contexto, resulta bastante frecuente el uso incorrecto de la palabra significativo (o sus derivados) para referirse a un resultado importante" (13). En otras palabras: a pesar de que el resultado de la prueba de hipótesis puede arrojar como resultado el rechazo de la hipótesis nula, al indicar que existe una diferencia en los parámetros poblacionales de interés, esta conclusión no siempre representa el mismo significado en la práctica. En algunas

ocasiones se puede encontrar que una diferencia estadísticamente significativa no representa una diferencia de magnitud relevante de acuerdo a la naturaleza del problema que se ha definido para el estudio. Así, la magnitud de la diferencia clínicamente significativa la establece el investigador basándose en múltiples factores como la gravedad del problema que se va a investigar, morbimortalidad asociada con el fenómeno, los costos que conlleva la implementación de nuevos tratamientos o la presentación de efectos secundarios (14, 15).

Para ilustrar lo expuesto hasta aquí, supóngase un estudio hipotético en el cual se pretende establecer los efectos ocasionados luego de un accidente cerebro vascular (ACV) de la arterial cerebral anterior, sobre la capacidad de inhibición de respuestas automáticas de los pacientes, luego de dos semanas del evento. Para ello, se toma un grupo de pacientes con ACV (n=20) de la arteria cerebral anterior (grupo 1), otro grupo de pacientes con ACV (n=18) en la arteria cerebral media o posterior (grupo 2) y un grupo control (n=22) sin antecedentes de daño cerebral (grupo 3). Los grupos se equiparan en las variables de edad, sexo, nivel educativo y estrato socioeconómico. Para la evaluación se incluye la prueba del Stroop y los resultados se resumen en la tabla 2.

Tabla 2. DATOS DE EJEMPLO DE DOS GRUPOS CON ACV Y UN GRUPO CONTROL EN LA PRUEBA DEL STROOP

Variable	**Grupo 1** **Me** ***Ds***	**Grupo 2** **Me** ***Ds***	**Grupo 3** **Me** ***Ds***	**Valor** ***p**<0.05*
Errores ensayo 1	0.42 0.8	0.53 0.5	0.21 0.3	0.62
Tiempo ensayo 1	18.3 0.76	19.5 1.02	15.6 0.56	0.41
Errores ensayo 2	0.62 0.45	1.14 1.26	0.23 0.35	0.26
Tiempo ensayo 2	21.5 2.11	22.7 1.14	19.1 0.98	0.51
Errores fase de conflicto	8.5 3.56	7.9 2.28	7.7 1.18	**0.03**
Tiempo fase de conflicto	32.5 5.65	33.2 4.48	31.9 3.15	**0.02**

* Prueba de Kruskall Wallis

Los valores de ***p*** que aparecen en negrita, informan de un resultado significativo. Sin embargo, al observar las medias y las desviaciones estándar, los valores señalan diferencias de menos de un error entre los grupos y de menos de un segundo en el tiempo empleado por los tres grupos. La pregunta que a continuación deberá plantearse el investigador es: ¿Clínicamente estas diferencias en el número de errores y en el tiempo son significativas o representan una diferencia real? Como puede verse, para responder a este interrogante no basta con establecer la magnitud estadística del evento; es necesario apelar a la experiencia del investigador con el instrumento

de medida y a los resultados obtenidos por otros estudios similares, entre otros aspectos, para poder, ahí si, determinar si el valor obtenido merece ser tenido en cuenta (para una revisión de algunos ejemplos adicionales en relación con el tema consultar a Sarria y Silva) (13).

Todo esto, ha originado una seria controversia alrededor de la conveniencia o no de indicar la relevancia de los resultados en términos del valor ***p***. (16, 17) e incluso se ha planteado el uso de otras medidas que puedan ser más útiles para interpretar los resultados de un estudio como son la estimación y los intervalos de confianza (18). En esta línea revistas como Lancet, British Journal of Medicine o New England Journal of Medicine han sugerido el uso de medidas diversas al valor de *p* (19) y la American Psychological Association recomienda el uso de medidas como los intervalos de confianza, que den cuenta de los efectos de tamaño de las diferencias encontradas (20), dejando en libertad al autor de una publicación de usar o no el valor ***p*** en sus análisis (21).

No obstante, es necesario señalar que existen diseños experimentales que justifican plenamente el uso del valor ***p*** y por tanto hay autores como Fleiss (3) que critican fuertemente la posición extrema de abolir por completo su uso.

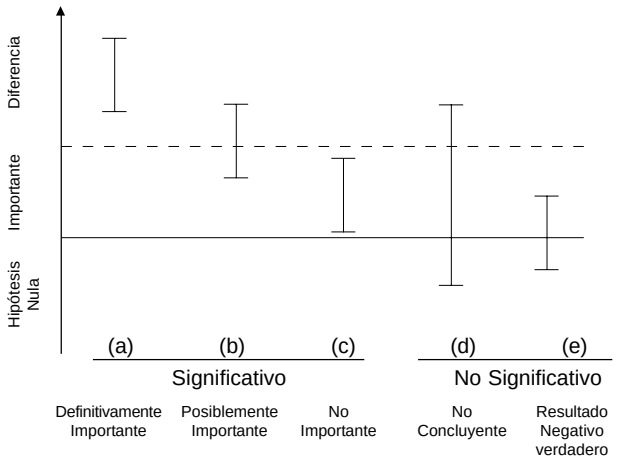
A manera de conclusión se plantean dos ideas que pueden contribuir a la mejor comprensión de los resultados que se obtienen en un estudio: en primer lugar, está la importancia de seleccionar con la misma dedicación y empeño con la que se plantea aspectos como el tamaño muestral o el diseño de la investigación, la prueba que se ajuste de mejor forma a la naturaleza de los datos para aceptar o rechazar la hipótesis nula.

En segundo lugar, se hace necesario, una vez obtenidos los resultados, no solamente identificar si las diferencias son producto o no del azar, sino además establecer si la magnitud de dichas diferencias posee alguna importancia desde el punto de vista clínico. Solo así será factible disminuir la posibilidad de encontrar asociaciones "***estadísticamente posibles pero conceptualmente estériles***" (9).

Una sugerencia útil consiste en utilizar además del valor ***p*** otro tipo de medida como los intervalos de confianza (22) los cuales pueden facilitar la distinción entre significación estadística y significación clínica. En la figura 1 se ilustran cinco interpretaciones posibles de una prueba de significación en términos del intervalo de confianza de una diferencia entre dos grupos que pudiera eventualmente ser útil para el investigador en el análisis de sus datos:

Figura 1. PRUEBA DE SIGNIFICANCIA

Intervalos de confianza que muestran las cinco interpretaciones posibles en términos de significación estadística e importancia clínica (a) La diferencia es significativa y con seguridad, suficientemente grande para tener importancia clínica; (b) La diferencia es significativa, pero no está claro si es suficientemente grande para ser clínicamente importante; (c) La diferencia es significativa, pero demasiado pequeña para ser importante; (d) La diferencia no es estadísticamente significativa pero puede ser suficientemente grande para ser importante; (e) La diferencia no es significativa ni tampoco lo bastante grande para ser clínicamente importante. (22)



REFERENCIAS

1. Benavidez, A y Silva, L.C. Contra la sumisión estadística: un apunte sobre las pruebas de significación Metas. 2000 27: 35-40.

2. Gardner, M.J. y Altman, D. G.) Confidence intervals rather than P values: estimation rather than hypothesis testing British Medical Journal. 1986; 292: 746-750.

3. Fleiss, J.L. Las pruebas de significación tienen una función en la investigación epidemiológica: Respuesta a M. Wlaker Boletín Sanitario de Panamá. 1993; 2: 115.

4. Londoño, J.L. Metodología de la Investigación Epidemiológica (3ª ed) Manual Moderno Bogotá, Colombia 2004.

5. Smith, Peter G & Morrow, Richard H. Ensayos de campo de intervenciones en salud en países en desarrollo. 2° edición. OPS, 1998 256-257.

6. Hernández, R. Fernández, C. y Baptista, P. Metodología de la Investigación 3ª ed. Mc Graw Hill Mexico D.F. Mexico 2003.

7. Kerlinger F.N. y Lee, HB Investigación del Comportamiento: Métodos de Investigación en Ciencias Sociales 4ª ed. Mc Graw Hill México 2001.

8. Pérez, A. Gómez, C. Sánchez, R. Dennis, R. Ruiz, A. Selección de la muestra y factores determinantes para el cálculo de su tamaño En Investigación Clínica: Epidemiología Clínica Aplicada Ruiz, A. Gómez, C y Londoño D. Editores Centro Editorial Javeriano Bogotá Colombia 2001 p. 412 -443.

9. Siegel, S. Estadística no Paramétrica México: Trillas; 1974

10. Silva. L.C. y Benavides, A. Apuntes sobre subjetividad y estadística en la investigación en salud Revista Cubana de Salud Pública. 2003; 29 (2):170 – 173.

11. Fernández, P. y Díaz, P. Significancia estadística y relevancia clínica Cadena de Atención Primaria. 2001; 8: 191-195.

12. Coolican, H. Métodos de Investigación en Psicología 3ª ed. Mexico: Manual Moderno 2005.

13. Sarria, M y Silva, L.C. Las Pruebas de Significación estadística en tres revistas biomédicas: una revisión crítica Revista Panamericana de Salud Pública. 2004; 15 (5): 300 – 305.

14. Gil, J.F. Una Mirada al Valor de *p* en Investigación Revista Colombiana de Psiquiatría. 2005; 34 (3): 414 – 424.

15. Schatz, P. Jay, K.A. McComb, J. McLaughlin, J.R. Misuse of Statistical Test in Archives of Clinical Neuropsychology publications Archives of Clinical Neuropsychology. 2005; 20: 1053 – 1059.

16. Wilhelmus, K.R. Beyond the P: posible insignificance of de nonsignificant P value Journal of cataract Refract Surgeon. 2004; 30:2425-2426.

17. Wilhelmus, K.R. Beyond the P: precluding a puddle of p values Journal of cataract Refract Surgeon. 2004; 30: 2207-2208.

18. Castañeda, J y Gil. J.F. Una mirada a los intervalos de confianza en investigación Revista Colombiana de Psiquiatría. 2004; 33:(2) 193 – 201.

19. Fidler, F. Cumming, G. Burgman, M y Neil, T. Statistical reform in medicine, psychology and ecology The Journal of socio-Economics. 2004; 33: 615 – 630.

20. Cumming, G. Fidler, F, Leonrard, M. Kalinowski, P. Chistiansen, A. Kleinig, A. Lo, J. Mcnenamin, N. y Wilson, S. Statistical Reform in Psychology Psychological Science. 2007;18 (3): 230 – 232.

21. American Psychological Association Manual de Estilo de Publicaciones 2ª ed. Manual Moderno. 2002. Mexico.

22. Armitage, P y Berry, G. Estadística para la Investigación Biomédica 3ª Ed. Harcourt Brace 1997. Madrid España.