Topicos en investigación clínica

¿SIGNIFICANCIA ESTADÍSTICA O CLÍNICA? ACLARANDO CONCEPTOS

María Alejandra Palacios-Ariza[1]

1. Médico y cirujano. MSc en Epidemiología, Esp. Educación para profesionales de la salud. Instructor Asistente Unidad de Investigación, Fundación Universitaria Sanitas, Bogotá D.C., Colombia.

RESUMEN

Es clave para el clínico que lee investigación realizar la diferenciación entre la significancia clínica y la significancia estadística. En esta revisión se indagará en primer lugar sobre qué es la significancia estadística para posteriormente aproximarnos a los conceptos que envuelven la significancia clínica. Se revisará qué opciones estadísticas existen para mejorar la interpretación de los resultados obtenidos en una investigación evitando caer sólo en una mala interpretación del "temido valor *p*". Por último, se revisarán dos ejemplos en los que se presentó una situación que puede ser común en investigación: significancia estadística sin significancia clínica.

Palabras Clave: Epidemiología; Probabilidad; Interpretación Estadística de Datos; Valor *p*.

DOI:

¿STATISTICAL OR CLINICAL SIGNIFICANCE? CLARIFYING CONCEPTS

ABSTRACT

It is key for the clinician reading research to recognize the difference between clinical and statistical significance. In this review we will first look into the meaning of statistical significance followed by a review of the concepts that give rise to clinical significance. We will review the statistical options available to improve the interpretation of research results so that we can avoid the misinterpretation of the "dreaded p-value". Lastly, we will review two real examples from the literature of a common situation in research studies: the finding of statistical significance without clinical significance.

Keywords: Epidemiology; Probability; Statistical Data Interpretation; P Value.

Recibido: 04/08/2022
Aceptado: 05/08/2022
Correo de correspondencia: mapalaciosar@unisanitas.edu.co

LA SIGNIFICANCIA ESTADÍSTICA

> *"La hipótesis nula, tomada literalmente (y esa es la única forma en que puede tomarse en la prueba de hipótesis formal), siempre es falsa en el mundo real... Si es falsa, incluso en un grado minúsculo, debe ser el caso que una gran muestra producirá un resultado significativo y conducirá a su rechazo. Entonces, si la hipótesis nula siempre es falsa, ¿cuál es el problema con rechazarla?"*
>
> Cohen J. Things I have Learned (So Far). American Psychologsit (Dic 1990).

La significancia estadística, controversial en el campo de la estadística, se refiere al rechazo de una hipótesis nula para adoptar la alterna bajo ciertos valores predefinidos del estadístico (1). Más formalmente hablaríamos de:

$$H_0 : \hat{\theta} = 0$$
$$H_a : \hat{\theta} \neq 0$$

En donde puede asumir la forma de una diferencia de medias, el logaritmo de un riesgo relativo, una razón de momios, entre otros. El umbral bajo el cual rechazamos la hipótesis nula con frecuencia se define en 0.05, a pesar de que no haya un soporte teórico para esta elección (2). Con esto, no se está simplemente haciendo la observación común de que cualquier umbral en particular es arbitrario. Más bien, se está señalando que incluso grandes cambios en los niveles de significación pueden corresponder a cambios pequeños y no significativos en las magnitudes de los estimadores subyacentes (3).

EL INFAME VALOR *P*

El valor *p* y su interpretación es un tema controversial, más que todo debido a las interpretaciones que este valor ha adquirido por fuera de la estadística como disciplina pura (4). Asumiendo que la hipótesis nula es cierta, el valor *p* es la probabilidad de observar un valor del estimador al menos tan extremo como el que se obtuvo en el estudio en cuestión. Nótese que no habla de la magnitud del estimador como tal. Claro está, para tamaños de muestra idénticos, un estimador con mayor magnitud tenderá a tener un valor *p* inferior. Sin embargo, es la alta dependencia del tamaño de muestra que hace difícil la interpretación del valor *p* y lo que ha llevado a que se malinterprete (5).

Es útil pensar en qué NO es el valor *p* de modo que se puedan evitar los errores en su interpretación. El valor *p* no es la probabilidad de que los resultados se deban al azar (6). El problema de las múltiples comparaciones es un buen ejemplo de por qué esta interpretación es incorrecta. Cada prueba de hipótesis que hacemos como parte de un análisis estadístico está tan sujeta al azar como el proceso de generación de datos y al realizar muchas pruebas sobre el mismo conjunto de datos corremos el riesgo de que, por azar, el valor *p* caiga por debajo de 0.05 (7). Los estudios de asociación genómica son un buen ejemplo de esto, puesto que se incluyen miles de polimorfismos dentro de modelos que buscan asociarlos con un fenotipo dado. Asumiendo que los polimorfismos ingresados no tengan ninguna asociación real con el fenotipo, de no realizar un ajuste del valor *p*, el número de falsos positivos sería enorme, siendo todos los positivos producto del azar (8)Linkage Disequilibrium (LD. Es por esto por lo que es una práctica usual en ese campo hacer ajustes al valor *p* usando métodos como el de Bonferroni o el False Discovery Rate.

El valor *p* tampoco es la probabilidad de que la hipótesis alterna sea verdadera. El valor *p* se evalúa como el complemento del área bajo la función de densidad de probabilidad, asumiendo que la nula es cierta, para el valor del estimador dado (1,6). Al no estar asociada con la distribución de probabilidad correspondiente a la hipótesis alterna, esta interpretación no es coherente.

El valor tampoco es evidencia que garantice que un modelo sea apropiado. El valor *p* no es una consideración, por ejemplo, de la presencia de confusión en la asociación entre una variable independiente y alguna de las variables incluidas en el modelo (6,9). Tampoco nos dice de forma inequívoca si nuestro

modelo se ajusta a los datos, puesto que el fenómeno del sobreajuste nos dará una confianza falsa.

Por último, el valor *p* tampoco es una medida de la magnitud del efecto, particularmente si tenemos en cuenta el tamaño de la muestra (5,6,9). Haciendo un ejercicio mental, en la *Figura 1* se muestra en los paneles A y B un ejemplo de *La Ronda de Noche* de Rembrandt. En el panel A se ve la obra original y en el B se ve la obra vandalizada. A la resolución que ofrece esta publicación, seguramente no pueden observar ninguna diferencia. Sin embargo, si se ven los recuadros C y D (acercamiento al 500%) podemos ver, sobre un segmento de la obra, que el recuadro C sí es diferente del recuadro D, el cual ha sido "vandalizado". La pregunta que debemos hacernos a continuación es si este vandalismo es significativo. Sin duda haciendo mayores y mayores acercamientos cualquier cambio microscópico podría interpretarse como vandalismo, pero este claramente no sería significativo. Es por este motivo que el valor *p* no es una medida del tamaño del efecto. Cuanto más grande sea la muestra (el acercamiento), mayor será la probabilidad de que se encuentre diferencias entre la obra en dos momentos en el tiempo, sin que esto sea un cambio importante. Otra noción que es importante recordar es que ningún valor *p* es "más significativo" que otro y que no permite comparar el efecto que tenga la modificación de una u otra variable sobre el desenlace (9).

FIGURA 1. *La Ronda de Noche* de Rembrandt. Las figuras A y C son la obra original. Las figuras B y D son la obra "vandalizada". Figura C y D tienen un acercamiento del 500%

LA SIGNIFICANCIA CLÍNICA

La significancia clínica es conceptualmente más difícil de capturar puesto que una misma patología representa distintas cosas para distintas personas. Puede definirse, por ejemplo, según la fuente: el paciente y su familia, el clínico, el salubrista (10). Para el paciente y su familia, la experiencia de servicio, la carga sintomática y calidad de vida tendrá un mayor peso dentro de su concepto de significancia clínica a la hora de valorar el impacto de una intervención (11). Para el clínico pueden existir medidas que no tengan mayor significado para el paciente y sus familias, como es el caso del número de lesiones hiperintensas en resonancia magnética en un paciente con esclerosis múltiple con sintomatología estable. Por último, para el salubrista existen impactos agregados de gran importancia que no serán significativos para el paciente o para el clínico, como lo son cambios sutiles en la incidencia de una enfermedad o de sus complicaciones que tienen grandes impactos sobre sistemas de salud, pero impactos discretos sobre pacientes individuales (12).

El concepto de significancia se ve modificado además por la severidad de la enfermedad. Al principio de un proceso de enfermedad es posible que el paciente quiera volver a su funcionalidad plena, no así para un paciente que ha acumulado discapacidad durante años y para el que incrementos menores en funcionalidad podrían ser muy significativos (10).

CÓMO DETERMINAR LA SIGNIFICANCIA CLÍNICA EN LA PRÁCTICA

En la literatura, la mejor forma de aproximarse a la significancia clínica es la verificación de los tamaños del efecto reportados. Se debe mirar en primer lugar la magnitud del estimador reportado y su significado con respecto a un riesgo dado. Una razón de momios de 2.00, por ejemplo, representa un incremento del 100% en el riesgo de un evento dado. Adicionalmente, una buena práctica es la del cálculo del número que es necesario tratar (NNT, por sus siglas en inglés) (13). Se debe ubicar la frecuencia final del evento de interés en el grupo control, restar a esa frecuencia la frecuencia en el grupo de intervención y tomar su inverso. Este valor da una idea de cuántos pacientes deben recibir la intervención para lograr o prevenir un desenlace dado.

Si lo que se pretende es determinar la significancia clínica en la planeación de un estudio o en el seguimiento de un paciente existen otras alternativas. Si se tiene acceso a personal con experiencia clínica se pueden usar procesos formales de elicitación. Si existen herramientas formales para medir estadios de la enfermedad bajo estudio, un cambio que logre desplazar la interpretación de la escala de un estadio a otro podría considerarse significativo. Existen también escalas con diferencias marcadas entre puntos, como es el caso de la escala de Rankin modificada para discapacidad, la cual fue diseñada para mostrar cambios significativos entre puntos consecutivos (14).

La literatura también puede ser una fuente valiosa a la hora de determinar la presencia de significancia clínica. Los entes reguladores pueden tener criterios predefinidos para la aprobación de una intervención dada, planteando de forma clara qué magnitud del efecto se considera suficiente para permitir la adopción de una intervención. Adicionalmente, para varias patologías existen estudios formales de determinación de la Diferencia Clínica Importante Mínima (MCID), que pueden servir en la planeación de estudios (15). Debe tenerse en cuenta que estos valores, frecuentemente fijos, no tienen en cuenta que las expectativas de los pacientes cambian a medida que progresa una enfermedad.

EJEMPLOS DE SIGNIFICANCIA ESTADÍSTICA SIN SIGNIFICANCIA CLÍNICA

Un estudio realizado por Westphal et al. en 2020 buscaba evaluar si el manejo previo con metformina confería un mejor desenlace tras trombólisis endovenosa en pacientes que sufrían un ataque cerebrovascular isquémico (ACVi) (16). Los investigadores tomaron datos de un registro colaborativo y realizaron control de sesgos con puntajes de propensión. Uno de los resultados presentados fue el del puntaje NIH Stroke

Scale Score NIHSS para severidad de un ACV. A pesar de las hipótesis, los investigadores encontraron una diferencia estadísticamente significativa entre los puntajes NIHSS de ingreso con un valor *p* claramente inferior a 0.05. Sin embargo, la diferencia absoluta entre los puntajes promedio fue de tan solo 1.3 puntos (en una escala que puede sumar hasta 42 puntos) en el NIHSS de ingreso de los pacientes (16). Los investigadores, correctamente, interpretan este hallazgo como estadística, pero no clínicamente significativo.

Otro ejemplo, tristemente célebre, proviene de un estudio del uso de erlotinib en pacientes con cáncer de páncreas avanzado. En este caso los investigadores llevaron a cabo un ensayo clínico aleatorizado en el que asignaron aleatoriamente pacientes a recibir manejo con gemcitabina vs. gemcitabina y erlotinib en terapia combinada (17). Al cabo del estudio los investigadores concluyen que el manejo con terapia combinada prolongó de forma significativa la supervivencia (HR: 0.82 IC95% 0.69-0.99). Sin embargo, al valorar la supervivencia de los sujetos en cada brazo se encuentra una supervivencia en el grupo de intervención de 6.24 meses y de 5.91 meses en el grupo control. Es decir, la diferencia en supervivencia fue de tan solo 10 días (17). Claramente, y a pesar de la significancia estadística, la terapia combinada que se evaluó en este estudio no provee una mejoría clínicamente significativa en la supervivencia de los pacientes.

REFERENCIAS

1. Cohen J. The earth is round (p < .05). American Psychologist. 1994;49(12):997–1003. https://doi.org/10.1037/0003-066X.49.12.997
2. Cohen J. Things I have learned (so far). American Psychologist. 1990;45(12):1304–12. https://doi.org/10.1037/0003-066X.45.12.1304
3. Gelman A, Stern H. The Difference Between "Significant" and "Not Significant" is not Itself Statistically Significant. null. 2006 Nov;60(4):328–31. https://doi.org/10.1198/000313006X152649
4. Cox DR. Statistical significance tests. Br J Clin Pharmacol. 1982 Sep;14(3):325–31. https://doi.org/10.1111/j.1365-2125.1982.tb01987.x
5. Goodman S. A dirty dozen: twelve p-value misconceptions. Semin Hematol. 2008 Jul;45(3):135–40. https://doi.org/10.1053/j.seminhematol.2008.04.003
6. Wasserstein RL, Lazar NA. The ASA Statement on p-Values: Context, Process, and Purpose. The American Statistician. 2016 Apr 2;70(2):129–33. https://doi.org/10.1080/00031305.2016.1154108
7. Curran-Everett D. Multiple comparisons: philosophies and illustrations. Am J Physiol Regul Integr Comp Physiol. 2000 Jul;279(1):R1-8. https://doi.org/10.1152/ajpregu.2000.279.1.R1
8. Sethuraman A, Gonzalez NM, Grenier CE, Kansagra KS, Mey KK, Nunez-Zavala SB, et al. Continued misuse of multiple testing correction methods in population genetics-A wake-up call? Mol Ecol Resour. 2019 Jan;19(1):23–6. https://doi.org/10.1111/1755-0998.12969
9. Sullivan GM, Feinn R. Using Effect Size—or Why the P Value Is Not Enough. J Grad Med Educ. 2012 Sep;4(3):279–82. https://doi.org/10.4300/JGME-D-12-00156.1
10. Kazdin AE. The meanings and measurement of clinical significance. J Consult Clin Psychol. 1999 Jun;67(3):332–9. https://doi.org/10.1037/0022-006X.67.3.332
11. Benson T. Measure what we want: a taxonomy of short generic person-reported outcome and experience measures (PROMs and PREMs). BMJ Open Qual. 2020 Mar;9(1):e000789. http://dx.doi.org/10.1136/bmjoq-2019-000789
12. Glass TA, Goodman SN, Hernán MA, Samet JM. Causal inference in public health. Annu Rev Public Health. 2013;34:61–75. https://doi.org/10.1146/annurev-publhealth-031811-124606
13. Vancak V, Goldberg Y, Levine SZ. Systematic analysis of the number needed to treat. Stat Methods Med Res. 2020 Sep;29(9):2393–410. https://doi.org/10.1177/0962280219890635
14. Quinn TJ, Dawson J, Walters M. Dr John Rankin; his life, legacy and the 50th anniversary of the Rankin Stroke Scale. Scott Med J. 2008 Feb;53(1):44–7. https://doi.org/10.1258/RSMSMJ.53.1.44

15. Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. Control Clin Trials. 1989 Dec;10(4):407–15. https://doi.org/10.1258/RSMSMJ.53.1.44
16. Westphal LP, Widmer R, Held U, Steigmiller K, Hametner C, Ringleb P, et al. Association of prestroke metformin use, stroke severity, and thrombolysis outcome. Neurology. 2020 Jul 28;95(4):e362–73. https://doi.org/10.1212/WNL.0000000000009951
17. Moore MJ, Goldstein D, Hamm J, Figer A, Hecht JR, Gallinger S, et al. Erlotinib Plus Gemcitabine Compared With Gemcitabine Alone in Patients With Advanced Pancreatic Cancer: A Phase III Trial of the National Cancer Institute of Canada Clinical Trials Group. JCO. 2007 May 20;25(15):1960–6. https://doi.org/10.1200/JCO.2006.07.9525