Artículo de revisión
Ginecol Obstet Mex. 2020; 88 (8): 536-541.

Las falacias de la p y significación estadística

The fallacies of P and statical significance.

José Niz-Ramos

Resumen

ANTECEDENTES: El valor de p es el método más empleado para estimar la significación estadística de cualquier hallazgo; sin embargo, en los últimos años se ha intensificado su debate al respecto, debido a la baja credibilidad y reproducibilidad de diversos estudios.

OBJETIVO: Describir el estado actual del concepto del valor de p y la significación estadística (prueba de significación de la hipótesis nula [por sus siglas en inglés: *Null Hypothesis Significance Testing*: NHST]), especificar los problemas más importantes y puntualizar las soluciones propuestas para la mejor utilización de los conceptos.

METODOLOGÍA: Se llevó a cabo la búsqueda bibliográfica en MEDLINE y Google Scholar, con los términos: "NHST", "Statistical significance; P value" en idioma inglés y español, de 2018-2019, limitándose a la selección de artículos publicados entre 2005 y 2019, mediante la revisión de tipo narrativo con búsqueda manual; sobre todo estudios de metodología.

RESULTADOS: La búsqueda global reportó 1411 artículos: 875 de *PubMed* y 536 de *Google Scholar*. Se excluyeron 817 por duplicación, 155 sin acceso completo y 414 ensayos clínicos (sin metodología estadística); los 25 restantes fueron el motivo del análisis.

CONCLUSIONES: El concepto del valor de p no es simple, tiene varias falacias y malas interpretaciones que deben considerarse para evitarlas en lo posible. Se recomienda no usar el término "estadísticamente significativo" o "significativo", sustituir el umbral de 0.05 por 0.005, informar valores de p precisos y con IC95%, riesgo relativo, razón de momios, tamaño del efecto o potencia y métodos bayesianos.

PALABRAS CLAVE: Valor de p; MEDLINE; reproducibilidad; significación estadística; riesgo relativo; razón de momios; métodos bayesianos.

Abstract

BACKGROUND: The P value is the most widely used method of estimating the statistical significance of any finding, however, in recent years the debate over the P value has been increasingly intensified due to the low credibility and reproducibility of many studies.

OBJECTIVE: To describe the current state of the concept of the value of P and the statistical significance (Null Hypothesis Significance Testing (NHST), specify the most important problems and point out the solutions proposed in the literature for their best use.

METHODOLOGY: Search in MEDLINE and Google Scholar, with the terms: "NHST", "Statistical significance; P value "in English and Spanish, carried out from 2018-2019, limited to articles published from 2005 to 2019, and a narrative-type review with manual search. Articles on methodology were preferably selected.

RESULTS: The global search yielded 1411 articles, 875 from PubMed and 536 from Google Scholar. 817 were excluded by duplication, 155 without full access, 414 from clinical trials, without statistical methodology. The 25 selected articles were the reason for the analysis.

CONCLUSIONS: The concept of the value of P is not simple, and it has several fallacies and misinterpretations that must be taken into account to avoid them as much as possible. Recommendations: Do not use "statistically significant" or "significant", replace the threshold of 0.05 with 0.005, report accurate P values with 95% CI, relative risk, odds ratio, effect size or power, and Bayesian methods.

KEYWORDS: *P* value; MEDLINE; Reproducibility; Statistical significance; Relative Risk; Odds Ratio; Bayesian Methods.

Ginecoobstetra, coeditor de Ginecología y Obstetricia de México

Recibido: mayo 2020
Aceptado: julio 2020

Correspondencia
José Niz-Ramos
nizjose@gmail.com

Este artículo debe citarse como
Niz-Ramos J. Las falacias de la p y la significación estadística. Ginecol Obstet Mex. 2020; 88 (8): 536-541.
https://doi.org/10.24245/gom.v88i8.4534

ANTECEDENTES

El valor de p (*P value* en inglés) es el método más empleado para estimar la significación estadística en una prueba de hipótesis y en la mayor parte de las investigaciones científicas, desde la t de Student y la χ^2 hasta el análisis de regresión; sin embargo, se ha suscitado una discusión debido a su mal uso, por ignorancia o conflicto de intereses. Ronald Fisher propuso los límites entre la significación y la falta de significación basándola en la probabilidad (P), y estableciendo arbitrariamente su límite con el valor de p = 0.05; donde p significa la probabilidad de obtener un hallazgo de interés por casualidad.[1] También se ha criticado por qué no consideran la importancia y relevancia del efecto observado.[2]

El objetivo de este estudio fue: describir el estado actual del concepto del valor de p y la significación estadística (NHST: *Null Hypothesis Significance Testing,* por sus siglas en inglés), especificar los problemas más importantes y puntualizar las soluciones propuestas para una mejor utilización de los conceptos.

METODOLOGÍA

Revisión narrativa de la bibliografía, llevada a cabo entre 2018 y 2019, en la que se seleccionaron artículos publicados de 2005 a 2019 en PubMed y Google Académico, que incluyeran las palabras clave: "NHST", "P value" y "Statistical significance", utilizando OR como término booleano, mediante la revisión narrativa con análisis manual (desplegando artículos relacionados y citas).

Se tomaron en cuenta artículos de revistas y *blogs* que analizan la metodología de NHST y el valor de p, mediante artículos originales, revisiones, editoriales, cartas al editor y artículos de opinión.

RESULTADOS

La búsqueda global arrojó 1411 artículos: 875 de PubMed y 536 de Google Scholar. Se excluyeron 817 por duplicación, 155 sin acceso completo y 414 ensayos clínicos (sin metodología estadística). Los 25 artículos restantes fueron el motivo de análisis de la revisión. **Figura 1**

DISCUSIÓN

El valor de p es la probabilidad de observar un parámetro hipotético (por ejemplo: una razón de posibilidades), tan extremo como el observado debido solo al azar y que varía entre 0-1. Se interpreta de tres maneras: *1)* $p \leq 0.05$: indica fuerte evidencia contra la hipótesis nula ([H0] podría ser rechazada), *2)* > 0.05: sugiere débil evidencia contra la H0 (podría fallar el rechazo de hipótesis nula) y *3)* valores de p cercanos al límite son marginales.[3] Se realiza a través de la NHST (*Null Hypothesis Significance Testing*), mediante pruebas de inferencia estadística (t de Student, ANOVA, χ^2, correlación de Pearson, etc.).

La definición es clara y precisa, pero las interpretaciones incorrectas siguen siendo abundantes y repetidas, por ejemplo, Nuzzo[4] señala que 89% de los estudios publicados en 2011 informaron el valor p sin proporcionar ningún modelo de estimación, tamaño del efecto o potencia estadística, y otras publicaciones indican el mal empleo de dichos valores.[5-7]

Los valores de p siempre se han criticado, algunos autores señalan que son como el vestido nuevo del emperador[4] (con innegables inconvenientes) o como los mosquitos (incómodos y difíciles de ahuyentar),[7] incluso se ha comparado a la falacia del valor de p con la fábula del "*zorro de Esopo*", por ser un índice generalizado, incomprendido, mal interpretado y calculado.[3]

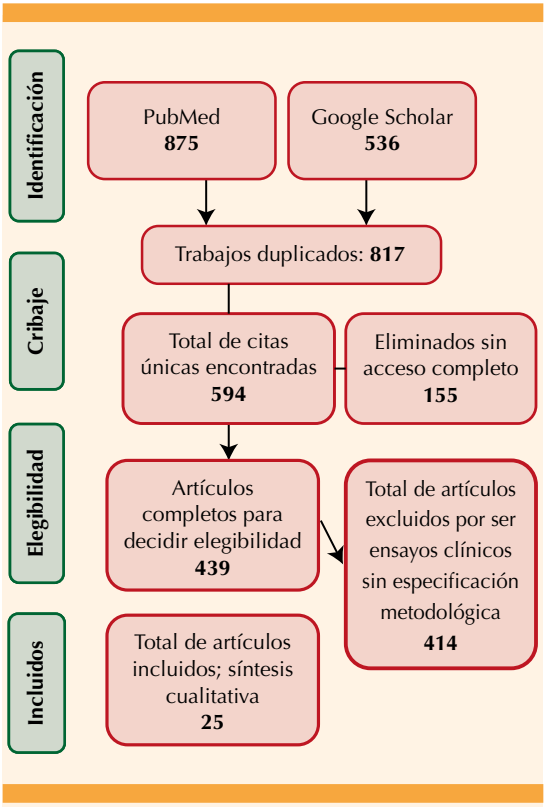

Figura 1. Diagrama de los resultados de la búsqueda.

En la actualizad se utiliza un híbrido, que considera las pruebas de significación estadística de Neyman-Pearson (hipótesis alternativa, error tipo I y II, y potencia), y se informa el valor de p aceptado o rechazado (Fisher), según sea el caso. Esto ha creado confusión, perpetuada por libros y revistas médicas, y ha permitido que los resultados salgan a la luz publicados con el término: "estadísticamente significativo", "p <0.05" o "p >0.05".[8]

Prestigiosos autores critican el mal uso de los valores de p en investigaciones biomédicas y en la ciencia en general.[4-9] Incluso la revista *Basic and Applied Social Psychology* prohibió en 20015 el uso de pruebas de hipótesis:[5]

Existen varias interpretaciones erróneas, también llamadas falacias, del valor de p;[3,4,10] un autor señala 12[11] otro 25,[12] pero las cuatro más mencionadas son:

1. La *"falacia de la probabilidad inversa"* es la falsa creencia que el valor de p se refiere a la probabilidad de tener una hipótesis nula (H_0) verdadera.
2. La *"falacia de las probabilidades contra el azar"* señala que el valor de p es la probabilidad de obtener el resultado por azar.

Ambas están relacionadas al confundir la probabilidad del resultado, asumiendo que la hipótesis nula es cierta. Las pruebas de significación estadística no ofrecen información de la probabilidad de la hipótesis nula.

3. La *"falacia del tamaño del efecto"* vincula la significación estadística con la magnitud del efecto. Así, los valores pequeños de p se interpretan como efectos grandes. Sin embargo, no informan la magnitud de un efecto.
4. La *"falacia de la significación clínica o práctica"* asocia el valor de p con la importancia clínica de un hallazgo. No obstante, el resultado estadísticamente significativo no indica que sea importante desde el punto de vista clínico.

También se ha indicado que algunos resultados en investigaciones son probablemente falsos y su reproductibilidad es mínima, por ejemplo, Ioannidis[13] señaló que de 49 estudios de investigación clínica originales, muy citados en tres revistas de alto impacto, 56% no pudieron reproducirse.

También existen autores a favor de las interpretaciones:[14,15] señalan que el valor de p es un instrumento y su falla depende de quién lo usa, o que los valores son "el patrón de referencia" de la validez estadística. La aparición de la llamada crisis de replicación, encabezada por

https://doi.org/10.24245/gom.v88i8.4534

Nuzzo (2014),[4] indica que el valor de p no es tan confiable ni tan objetivo como suponen la mayoría de los científicos. Por su parte, Baker[16] describió, en una encuesta con 1576 expertos de reproducibilidad de la investigación, que más de 70% de los autores no puede reproducir los experimentos del otro, y más de 50% no pude reproducir ni sus propios experimentos.

Ante esta situación, la Asociación Estadounidense de Estadística (ASA), en 2016,[6] tomó cartas en el asunto publicando una serie de principios:

1. Los valores de p pueden indicar hasta qué punto son incompatibles los datos con un modelo estadístico específico.
2. Los valores de p no miden la probabilidad de que la hipótesis estudiada sea verdadera o que los datos se originaron el azar.
3. Las conclusiones científicas y las decisiones empresariales o políticas no deberían basarse solo en el hecho de que el valor de p sobrepase un umbral específico.
4. Realizar una inferencia apropiada requiere un informe completo y transparente.
5. El valor de p o la significación estadística no miden el tamaño de un efecto ni la importancia de un resultado.
6. Por sí mismo, el valor de p no proporciona ninguna evidencia en relación con algún modelo o hipótesis.

Aunque estos principios fueron descritos anteriormente, representaron un parteaguas para que la asociación señalara los defectos y aunque no proporcionaron sugerencias para mejorar estas condiciones, repercutieron para que diversos investigadores se interesaran en el tema, y en octubre de 2017, la misma ASA favoreció el Simposio de Inferencia Estadística, una reunión de dos días que sentó las bases para la publicación de un número especial de la prestigiosa revista *The American Statistician*,[17] que para la fecha de consulta (30 de abril de 2020) tenía 170,045 vistas y 278 citas (5.3 citas por semana).

Enseguida se citan las principales sugerencias para disminuir los errores del valor de p y la significación estadística que propusieron los 43 autores en ese número[7] y otros habían postulado anteriormente:[18,19-21]

1. Sustituir el umbral de p = 0.05 por p = 0.005 y referir los valores entre 0.05 y 0.005 como sugerentes.
2. Asesorarse de expertos en estadística para interpretar los resultados de una investigación científica.
3. Reiterar la importancia clínica del estudio y proporcionar enunciados claros y explícitos de la(s) pregunta(s) de investigación y la(s) hipótesis(s) que se comprobarán.
4. Detallar la metodología del análisis estadístico (justificación del tamaño de la muestra y razones del empleo de métodos estadísticos). Si utilizan NHST y valores de p deberán justificar su aplicación.
5. Incitar a los revisores y consejos editoriales de las revistas para no permitir el uso de la frase "estadísticamente significativo" o "significativo".
6. Recalcar que los resúmenes contengan resultados con valores numéricos (tasas, porcentajes, proporciones) de los efectos demostrados.
7. Informar valores de p precisos (no menores de 0.05 o 0.01), incluso exactos, por ejemplo 0.002, utilizando índices de evidencia adicionales: IC95%, riesgo relativo, razón de momios (odds ratio), tamaño del efecto o potencia, y métodos bayesianos.

Posteriormente, otros autores sugirieron recomendaciones semejantes.[22,23] Algunas revistas cambiaron sus directrices para los informes estadísticos. *New England Journal of Medicine*[24] publicó en julio de 2019 que deben reemplazarse los valores de p con estimaciones de efectos o asociación con IC95%. Y la revista *Pediatric Anesthesia*[25] sugirió reportar los valores de p con IC95% y eliminar el concepto de "significación estadística".

CONCLUSIONES

El concepto del valor de p no es simple, tiene varias falacias y malas interpretaciones que deben tomarse en cuenta para evitarlas en lo posible. Además, cualquier declaración asociada con el valor de p debe considerarse con precaución.

Recomendaciones

- No usar los términos: "estadísticamente significativo" o "significativo".
- Sustituir el umbral de significación estadística de 0.05 por 0.005, y referirse a los valores p = 0.05 y p = 0.005 como sugerentes.
- Informar valores de p precisos (no menores de 0.05 o 0.01) o exactos.
- Utilizar en conjunto con el valor de p las pruebas que incluyen IC95%, riesgo relativo, razón de momios (odds ratio), tamaño del efecto o potencia, y métodos bayesianos.

REFERENCIAS

1. Gigerenzer, G.,et al. Surrogate science: The idol of a universal method for scientific inference. Journal of Management, 2015;41:421-440. doi: 10.1177/0149206314547522
2. Molina Arias M. ¿Qué significa realmente el valor de p?. Rev Pediatr Aten Primaria. 2017 Dic; 19(76): 377-381. Disponible en: http://scielo.isciii.es/scielo.php?script=sci_arttext&pid=S1139-76322017000500014&lng=es
3. Bertolaccini L, et al. Are the fallacies of the P value finally ended? J Thorac Dis. 2016 Jun;8(6):1067-8. DOI:10.21037/jtd.2016.04.48
4. Nuzzo R. Scientific method: statistical errors. Nature. 2014;506:150-2. http://ns.leg.ufpr.br/lib/exe/fetch.php/disciplinas:ce008:506150a.pdf
5. Trafimow, D, et al. Editorial, Basic and Applied Social Psychology, 2015;37:1, 1-2, DOI: 10.1080/01973533.2015.1012991
6. Wasserstein,R. et al. The ASA Statement on p-Values: Context, Process, and Purpose, The American Statistician, 2016;70:2, 129-133, https://doi.org/10.1080/00031305.2016.1154108
7. Lambdin C. Significance tests as sorcery: Science is empirical significance tests are not. Theory & Psychology, 2012; 22(1):67–90.http://psychology.okstate.edu/faculty/jgrice/psyc5314/SignificanceSorceryLambdin2012.pdf
8. Jiroutek MJ, Turner JR. Buying a significant result:Do we need to reconsider the role of the P value. Clin Hypertens. 2017;19:919–921. https://onlinelibrary.wiley.com/doi/epdf/10.1111/jch.13021
9. Badenes-Ribera L, et al. Errores de interpretación de los valores p entre psicólogos profesionales españoles. International Journal of Developmental and Educational Psychology, 2017;2:551-559. https://www.redalyc.org/pdf/3498/349853220053.pdf
10. Kühberger A. et all. The significance fallacy in inferential statistics. BMC Res Notes. 2015;17;8:84. https://doi.org/10.1186/s13104-015-1020-4.
11. Goodman S. A. Dirty Dozen: Twelve P-Value Misconceptions. Semin Hematol 2008;45:135-140. doi:10.1053/j.seminhematol.2008.04.003
12. Greenland S, et al, Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. Eur J Epidemiol 2016;31:337–350. DOI 10.1007/s10654-016-0149-3
13. Ioannidis JP. Contradicted and Initially Stronger Effects in Highly Cited Clinical Research. JAMA. 2005;294(2):218-228. doi:10.1001/jama.294.2.218
14. Palmer, A., et al. Recommendations for the use of statistics in clinical and health psychology. Clínica y Salud, 2013;24:47-54. http:// dx.doi.org/10.5093/cl2013a6
15. Lytsy P, P in the right place: Revisiting the evidential value of P-values. J Evid Based Med. 2018 Nov;11(4):288-291. doi: 10.1111/jebm.12319.
16. Baker, M.1,500 scientists lift the lid on reproducibility. Nature, 2016;533(7604), 452–454. doi:10.1038/533452ª
17. Wasserstein, R L et al. Moving to a World Beyond "p < 0.05" The American Statistician, 2019;73:sup1, 1-19, DOI: 10.1080/00031305.2019.1583913
18. Molina Arias, M. El significado de los intervalos de confianza. Pediatría Atención Primaria, 2013;15(57), 91-94. https://dx.doi.org/10.4321/S1139-76322013000100016

https://doi.org/10.24245/gom.v88i8.4534

19. Morey, R.D., et al. The fallacy of placing confidence in confidence intervals. Psychon Bull Rev 2016;23, 103–123. https://doi.org/10.3758/s13423-015-0947-8.

20. Esarey, J. Lowering the threshold of statistical significance to p< 0.005 to encourage enriched theories of politics. The Political Methodologist, 2017, https://thepolitical-methodologist.com/2017/08/07/in-support-of-enriched-theories-of-politics-a-case-for-lowering-the-threshold-of-statistical-significance-to-p-0-005/

21. Benjamin, D. J., et al. Redefine Statistical Significance, Nature Human Behaviour, 2018,2, 6–10. https://doi.org/10.1038/s41562-017-0189-z

22. Hurlbert, S., et al. "Coup de Grâce for a Tough Old Bull: 'Statistically Significant' Expires," The American Statistician, 2019;73. https://doi.org/10.1080/00031305.2018.1543616

23. Blakeley B. et al. Abandon Statistical Significance, The American Statistician, 2019;73: sup1, 235-245, DOI: 10.1080 / 00031305.2018.1527253

24. Harrington D. New Guidelines for Statistical Reporting in the Journal. N Engl J Med 2019; 381:285-286 DOI: 10.1056/NEJMe1906559

25. Davidson, A. Embracing uncertainty: The days of statistical significance are numbered. Pediatr Anaesth, 2019;29: 978-980. doi:10.1111/pan.13721

CITACIÓN ACTUAL

De acuerdo con las principales bases de datos y repositorios internacionales, la nueva forma de citación para publicaciones periódicas, digitales (revistas en línea), libros o cualquier tipo de referencia que incluya número doi (por sus siglas en inglés: Digital Object Identifier) será de la siguiente forma:

REFERENCIAS

1. Katarina V, Gordana T. Oxidative stress and neuroinflammation should be both considered in the occurrence of fatigue and depression in multiple sclerosis. Acta Neurol Belg. 2018;34(7):663-9. doi: 10.1007/s13760-018-1015-8.
2. Yang M, et al. A comparative study of three different forecasting methods for trial of labor after cesarean section. J Obstet Gynaecol Res. 2017;25(11):239-42. doi: https://doi.org/10.1016/j.gyobfe.2015.04.015.