Declaración de la ASA sobre la significación estadística y los p-valores

5 de febrero de 2016

Editado por Ronald L. Wasserstein, director ejecutivo

en nombre de la Junta Directiva de la *American Statistical Association*[1]

Introducción

Durante los últimos años, el aumento de la cuantificación en la investigación científica y la proliferación de grandes y complejos conjuntos de datos han ampliado el alcance de las aplicaciones de los métodos estadísticos. Este hecho ha creado nuevas posibilidades para el progreso científico, pero también ha traído cierta preocupación sobre las conclusiones obtenidas a partir de los datos. La validez de las conclusiones científicas y la posibilidad de reproducirlas no depende únicamente de los métodos estadísticos en sí mismos. La elección apropiada de las técnicas, la corrección de los análisis efectuados y la adecuada interpretación de los resultados estadísticos juegan también un papel esencial para garantizar conclusiones sólidas y para asegurar que la incertidumbre que las rodea esté bien representada.

El fundamento de muchas conclusiones científicas publicadas es el concepto de "significación estadística", normalmente evaluada mediante un índice denominado p-valor. Ahora bien, a pesar de que el p-valor puede ser una medida estadística útil, a menudo se emplea de forma incorrecta y también se malinterpreta. Esto ha llevado a que algunas revistas científicas disuadan de su uso y a que algunos científicos y estadísticos recomienden su abandono, basándose en argumentos que esencialmente son los mismos desde que el p-valor se introdujo por primera vez.

En este contexto, la *American Statistical Association* (ASA) cree que la comunidad científica podría beneficiarse de una declaración formal que aclare algunos principios que son ampliamente aceptados y están implícitos en la correcta utilización e interpretación del p-valor. Los aspectos considerados aquí no sólo afectan a la investigación, sino también a su financiación, a las prácticas de las revistas, al progreso profesional, a la educación científica, a las políticas públicas, al periodismo y al derecho. Esta declaración no pretende resolver todas las cuestiones relacionadas con las buenas prácticas estadísticas, ni tampoco resolver las controversias fundamentales. Más bien presenta en términos no técnicos una breve selección de principios que podrían mejorar la práctica y la interpretación de la ciencia cuantitativa, de acuerdo con un consenso amplio alcanzado en la comunidad estadística.

¿Qué es el p-valor?

Expresado de forma sencilla, un p-valor es la probabilidad, bajo un modelo estadístico especificado, de que un estadístico que sintetiza alguna característica de los datos (por

[1] Ronald L. Wasserstein & Nicole A. Lazar (2016): *The ASA's statement on p-values: context, process, and purpose*. Reimpreso con el permiso de *The American Statistician*.

ejemplo, la diferencia de las medias al comparar dos grupos) sea igual o más extremo que su valor observado.

Principios

1. *Los p-valores pueden indicar hasta qué punto son incompatibles los datos con un modelo estadístico especificado*

   Un p-valor proporciona un método para resumir la incompatibilidad entre un conjunto particular de datos y un modelo propuesto para ellos. El contexto más común es que el modelo esté construido bajo un conjunto de premisas, además de la denominada "hipótesis nula". A menudo la hipótesis nula postula la ausencia de un efecto, como por ejemplo que no haya diferencias entre dos grupos, o la ausencia de relación entre un factor y un resultado. Cuanto más pequeño sea el p-valor mayor será la incompatibilidad estadística de los datos con la hipótesis nula, si los supuestos realizados para calcular el p-valor se mantienen. Esta incompatibilidad equivale a poner en entredicho la hipótesis nula, o a proporcionar evidencia contra los supuestos considerados.

2. *Los p-valores no miden la probabilidad de que la hipótesis estudiada sea verdadera, o la probabilidad de que los datos hayan sido producidas sólo por el azar*

   Los investigadores a menudo pretenden convertir el p-valor en una afirmación sobre la certeza de una hipótesis nula, o sobre la probabilidad de que el azar haya generado los datos observados. El p-valor no es ni una cosa ni la otra. Es una afirmación sobre los datos en relación con una explicación hipotética especificada, no una afirmación sobre la explicación en sí misma.

3. *Las conclusiones científicas y las decisiones empresariales o políticas no se deberían basar únicamente en el hecho de que el p-valor sobrepase un umbral específico*

   Las prácticas que reducen el análisis de los datos o la inferencia científica a la aplicación mecánica de reglas rígidas para justificar afirmaciones científicas (cómo, por ejemplo, "$p<0.05$") pueden originar conclusiones erróneas, o una mala toma de decisiones. Una conclusión no se transforma de repente de "cierta" por un lado a "falsa" por otro. Los investigadores deben considerar que para establecer una inferencia estadística hay muchos factores en juego que la contextualizan —incluidos el diseño del estudio, la calidad de las medidas, la evidencia externa sobre el fenómeno en estudio y la validación de los supuestos subyacentes bajo el análisis de los datos. Por consideraciones de orden práctico a menudo es necesario tomar decisiones binarias (del tipo "si-no"), pero esto no significa que los p-valores aisladamente considerados puedan garantizar la corrección o incorrección de una decisión. El uso generalizado del concepto "significación estadística" (generalmente interpretado como "$p\leq0.05$") para legitimar la reclamación de un descubrimiento científico (o de la verdad que esté implícita) produce a una distorsión considerable del proceso científico.

4. *Realizar una inferencia apropiada requiere un informe completo y transparencia*

   Los p-valores y los análisis relacionados no se deben presentar selectivamente. Realizar diversos análisis de los datos e informar sólo de aquellos que logran un determinado nivel del p-valor (normalmente, aquellos que pasan un umbral de significación) imposibilita la interpretación de los p-valores publicados. La selección de hallazgos prometedores, una

costumbre también conocida con términos como dragado de datos, búsqueda y hasta persecución de la significación, inferencia selectiva o manipulación de p-valores, produce un exceso distorsionado de resultados estadísticamente significativos en la literatura publicada, que deberían ser firmemente evitados. Calcular por costumbre múltiples contrastes estadísticos presenta el inconveniente que cada vez que un investigador elige qué conclusiones presenta en base a unos resultados estadísticos, la interpretación válida de los resultados queda severamente comprometida si el lector no sabe que ha habido una elección y cuál ha sido su fundamento. Los investigadores deberían explicar cuál ha sido el número total de hipótesis exploradas durante el estudio, cuáles han sido las decisiones en la recogida de los datos, qué análisis estadísticos se han llevado a cabo, y también debería proporcionar todos los p-valores calculados. No se pueden obtener conclusiones científicas válidas basadas en los p-valores y en los estadísticos asociados sin tener conocimiento al menos de qué análisis se han llevado a cabo, y de cómo se seleccionaron los que se presentaron (incluyendo los p-valores).

5. *Un p-valor, o la significación estadística, no mide el tamaño de un efecto o la importancia de un resultado*

   La significación estadística no es equivalente a la significación científica, humana o económica. Unos p-valores menores que otros no implican necesariamente la presencia de efectos mayores o más importantes, ni p-valores mayores implican carencia de importancia, ni siquiera la ausencia de efecto. Cualquier efecto, no importa cómo sea de pequeño, puede producir un p-valor pequeño si el tamaño de la muestra o la precisión de la medida son lo bastante altos, y los efectos grandes pueden producir p-valores muy grandes si el tamaño de la muestra es muy pequeño o las medidas son imprecisas. Del mismo modo, efectos estimados idénticos tendrán diferentes p-valores si la precisión de la estimación difiere.

6. *Por sí mismo, un p-valor no proporciona una buena medida de la evidencia en relación con un modelo o una hipótesis*

   Los investigadores deberían admitir que un p-valor descontextualizado, aislado de otras evidencias, proporciona una información limitada. Por ejemplo, un p-valor cercano a 0.05 tomado por sí solo ofrece una evidencia muy débil contra la hipótesis nula. Del mismo modo, un p-valor relativamente grande no implica ninguna evidencia a favor de la hipótesis nula, puesto que otras muchas hipótesis podrían ser tan consistentes como ella con los datos observados, o incluso más. Por estas razones, el análisis de datos no se debería detener en el cálculo del p-valor, si hay otras aproximaciones apropiadas y factibles.

<u>Otras aproximaciones</u>

En vista de los frecuentes malos usos y de los malentendidos relativos a los p-valores, algunos estadísticos prefieren complementar, o incluso sustituir, el p-valor por otros procedimientos. Hay métodos que enfatizan la estimación por encima del mero poner a prueba y contrastar, tales como los intervalos de confianza, de credibilidad o de predicción. También se puede recurrir a métodos bayesianos, o a medidas alternativas de la evidencia, como por ejemplo la prueba de la razón de verosimilitud o los factores de Bayes. Y hay más posibilidades, como son los modelos de la teoría de toma de decisiones, o la tasa de falsos descubrimientos. Aunque

todas estas medidas y enfoques se basan en supuestos adicionales, podrían abordar de forma más directa el tamaño de un efecto (y su incertidumbre asociada), o la comprobación de la validez de una hipótesis.

Conclusión

Las buenas prácticas estadísticas, como componente esencial del buen quehacer científico, enfatizan los principios de dirigir y llevar a cabo un buen diseño de los estudios y una realización adecuada, de aportar una variedad de resúmenes numéricos y gráficos de los datos, de entender el fenómeno que se está estudiando, de interpretar los resultados dentro de su contexto, de proporcionar una información íntegra, y de comprender de forma adecuada, tanto lógica como cuantitativa, aquello que signifiquen los resúmenes de datos. Un índice único no debería sustituir el razonamiento científico.

Agradecimiento: la Junta Directiva de la ASA agradece a las siguientes personas haber compartido su experiencia y sus puntos de vista durante la preparación de este manifiesto. El texto no refleja necesariamente el punto de vista de todas estas personas. De hecho, algunas mantienen opiniones contrarias a las contenidas en la declaración (totalmente o sólo en parte). Ahora bien, les agradecemos enormemente sus contribuciones. *Naomi Altman, Jim Berger, Yoav Benjamini, Don Berry, Brad Carlin, John Carlin, George Cobb, Marie Davidian, Steve Fienberg, Andrew Gelman, Steve Goodman, Sander Greenland, Guido Imbens, John Ioannidis, Valen Johnson, Michael Lavine, Michael Lew, Rod Little, Deborah Mayo, Chuck McCulloch, Michele Millar, Sally Morton, Regina Nuzzo, Hilary Parker, Kenneth Rothman, Don Rubin, Stephen Senn, Uri Simonsohn, Dalene Stangl, Philip Stark, Steve Ziliak*.

Breve lista de referencias sobre el p-valor y la significación estadística,

para acompañar la Declaración de la ASA sobre la significación estadística y los p-valores

La siguiente lista no es exhaustiva, pero proporciona un buen punto de partida para las personas que deseen explorar con mayor detenimiento las cuestiones contenidas en la *Declaración de la ASA sobre la significación estadística y los p-valores*. Los artículos aparecen en orden alfabético:

Altman D.G., Bland J.M. (1995), "Absence of evidence is not evidence of absence", British Medical Journal, 311:485

Altman, D.G., Machin, D., Bryant, T.N., and Gardner, M.J., eds. (2000), Statistics with Confidence, 2nd ed., London: BMJ Books

Berger, J.O., and Delampady, M. (1987), "Testing precise hypotheses," Statistical Science, 2,317–335

Berry, D. (2012), "Multiplicities in Cancer Research: Ubiquitous and Necessary Evils," Journal of the National Cancer Institute, 104, 1124–1132

Christensen, R. (2005), "Testing Fisher, Neyman, Pearson, and Bayes," The American Statistician, 59, 2, 121-126

Cox, D.R. (1982), "Statistical Significance Tests," British Journal of Clinical Pharmacology, 14, 325-331

Edwards, W., Lindman, H., and Savage, L.J. (1963), "Bayesian statistical inference for psychological research," Psychological Review, 70, 193–242

Gelman, A., and Loken, E. (2014), "The Statistical Crisis in Science [online]" American Scientist, 102. Available at http://www.americanscientist.org/issues/feature/2014/6/thestatistical-crisis-in-science

Gelman A, Stern HS. (2006), "The difference between 'significant' and 'not significant' is not itself statistically significant," The American Statistician, 60:328–331

Gigerenzer G (2004), "Mindless statistics," Journal of Socioeconomics, 33:567–606

Goodman, S.N. (1999a), "Toward Evidence-Based Medical Statistics 1: The P Value Fallacy,"Annals of Internal Medicine, 130, 995-1004

_____ (1999b), "Toward Evidence-Based Medical Statistics. 2: The Bayes Factor,"Annals of Internal Medicine, 130, 1005-1013

_____ (2008), "A Dirty Dozen: Twelve P-Value Misconceptions," Seminars in Hematology, 45, 135-140

Greenland, S. (2011), “Null misinterpretation in statistical testing and its impact on health risk assessment,” Preventive Medicine, 53, 225–228

_____ (2012). Nonsignificance plus high power does not imply support for the null over the alternative. Annals of Epidemiology, 22:364–368

Greenland, S., and Poole C (2011), “Problems in common interpretations of statistics in scientific articles, expert reports, and testimony,” Jurimetrics, 51, 113–129

Hoenig J.M., and Heisey D.M. (2001). The abuse of power: The pervasive fallacy of power calculations for data analysis. The American Statistician, 55:19–24

Ioannidis, J.P. (2005), “Contradicted and initially stronger effects in highly cited clinical research.” Journal of the American Medical Association, 294, 218-228

_____ (2008), “Why most discovered true associations are inflated (with discussion),” Epidemiology 19: 640-658

Johnson, V.E. (2013), “Revised standards for statistical evidence,” Proceedings of the National Academy of Sciences, 110(48), 19313–19317

_____ (2013), "Uniformly most powerful Bayesian tests,” Annals of Statistics, 41, 1716-1741

Lang, J., Rothman K.J., and Cann, C.I. (1998), “That confounded P-value. (Editorial),” Epidemiology, 9, 7-8

Lavine, M. (1999), “What is Bayesian Statistics and Why Everything Else is Wrong,” UMAP Journal, 20:2

Lew, M.J. (2012), “Bad statistical practice in pharmacology (and other basic biomedical disciplines): you probably don't know P,” British Journal of Pharmacology, 166:5, 1559-1567

Phillips, C.V. (2004), “Publication bias in situ,” BMC Medical Research Methodology, 4:20

Poole C. (1987), “Beyond the confidence interval,” American Journal of Public Health, 77, 195–199

Poole, C. (2001). Low P-values or narrow confidence intervals: Which are more durable? Epidemiology, 12, 291–294

Rothman, K.J. (1978), “A show of confidence (Editorial),” New England Journal of Medicine, 299, 1362-1363

_____ (1986), “Significance questing (Editorial),” Annals of Internal Medicine, 105, 445-447

_____ (2010), “Curbing type I and type II errors,” European Journal of Epidemiology, 25, 223-224

Rothman, K.J., Weiss, N.S., Robins, J., Neutra, R., and Stellman, S. (1992), “Amicus Curiae brief for the U. S. Supreme Court, Daubert v. Merrell Dow Pharmaceuticals, Petition for Writ of

Certiorari to the United States Court of Appeals for the Ninth Circuit," No. 92-102, October Term, 1992

Rozeboom, W.M. (1960), "The fallacy of the null-hypothesis significance test," Psychological Bulletin, 57:416–428

Schervish, M.J. (1996), "P Values: What They Are and What They Are Not," The American Statistician, 50:3, 203-206

Simmons, J.P., Nelson, L.D., and Simonsohn, U. (2011), "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant," Psychological Science, 22(11), 1359-1366

Stang, A., and Rothman, K.J. (2011), "That confounded P-value revisited," Journal of Clinical Epidemiology, 64(9), 1047-1048

Stang, A., Poole, C., and Kuss, O. (2010), "The ongoing tyranny of statistical significance testing in biomedical research," European Journal of Epidemiology, 25(4), 225-30

Sterne, J. A. C. (2002). "Teaching hypothesis tests – time for significant change?" Statistics in Medicine, 21, 985-994

Sterne, J. A. C. and G. D. Smith (2001). "Sifting the evidence – what's wrong with significance tests?" British Medical Journal, 322, 226-231

Ziliak, S.T. (2010), "The Validus Medicus and a New Gold Standard," The Lancet, 376, 9738, 324-325

Ziliak, S.T., and McCloskey, D.N. (2008), The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives, Ann Arbor: University of Michigan Press