Declaració de l'ASA sobre la significació estadística i els valors p

5 de febrer de 2016

Editat per Ronald L. Wasserstein, director executiu

en nom de la Junta Directiva de l'*American Statistical Association*[1]

Introducció

Els darrers anys, l'augment de la quantificació en la recerca científica i la proliferació de grans i complexos conjunts de dades han ampliat l'abast de les aplicacions dels mètodes estadístics. Aquest fet ha creat noves possibilitats per al progrés científic, però també ha portat certa preocupació sobre les conclusions obtingudes a partir de les dades. La validesa de les conclusions científiques i la possibilitat de reproduir-les no depèn únicament dels mètodes estadístics en sí mateixos. L'elecció apropiada de les tècniques, la correcció de les anàlisis efectuades i la interpretació adequada dels resultats estadístics juguen també un paper essencial a l'hora de garantir conclusions sòlides i per garantir que la incertesa que les envolta estigui ben representada.

El fonament de moltes conclusions científiques publicades és el concepte de "significació estadística", normalment avaluada mitjançant un índex anomenat valor p (*p-value*). Ara bé, tot i que el valor p pot ser una mesura estadística útil, sovint s'usa de forma incorrecta i també és mal interpretada. Això ha dut a què algunes revistes científiques dissuadeixin del seu ús i a què alguns científics i estadístics en recomanin el seu abandonament, basant-se en arguments que essencialment són els mateixos des de que els valors p van ser introduïts per primera vegada.

En aquest context, l'*American Statistical Association* (ASA) creu que la comunitat científica podria beneficiar-se d'una declaració formal que aclarís alguns principis que són generalment acceptats i que són implícits en la correcta utilització i interpretació dels valors p. Els aspectes considerats aquí no afecten només la recerca, sinó també el seu finançament, les pràctiques de les revistes, l'avenç professional, l'educació científica, les polítiques públiques, el periodisme i el dret. Aquesta declaració no pretén resoldre totes les qüestions relacionades amb les bones pràctiques estadístiques, ni tampoc resoldre les controvèrsies fonamentals. Més aviat presenta en termes no tècnics una selecció breu de principis que podrien millorar la pràctica i la interpretació de la ciència quantitativa, d'acord amb un consens ampli assolit a la comunitat estadística.

Què és el valor p?

Dit de forma planera, un valor p és la probabilitat, sota un model estadístic especificat, que un estadístic que sintetitza alguna característica de les dades (per exemple, la diferència de mitjanes en comparar dos grups) sigui igual o més extrem que el seu valor observat.

---

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

Principis

1. *Els valors p poden indicar fins a quin punt són incompatibles les dades amb un model estadístic especificat*

   Un valor p proporciona un mètode per resumir la incompatibilitat entre un conjunt de dades particular i un model proposat per a elles. El context més comú és que el model estigui construït sota un conjunt d'assumpcions, a més de l'anomenada "hipòtesi nul·la". Sovint la hipòtesi nul·la postula l'absència d'un efecte, com per exemple la manca de diferències entre dos grups, o l'absència de relació entre un factor i un resultat. Com més petit sigui el valor p major serà la incompatibilitat estadística de les dades amb la hipòtesis nul·la, si els supòsits fets per a calcular el valor p es mantenen. Aquesta incompatibilitat equival a posar en dubte la hipòtesi nul·la, o a proporcionar evidencia contra els supòsits considerats.

2. *Els valors p no mesuren la probabilitat que la hipòtesi estudiada sigui verdadera, o la probabilitat que les dades hagin estat produïdes només per l'atzar*

   Els investigadors sovint pretenen convertir el valor p en una afirmació sobre la certesa d'una hipòtesis nul·la, o sobre la probabilitat que l'atzar hagi generat les dades observades. El valor p no és ni una cosa ni l'altra. És una afirmació sobre les dades en relació amb una explicació hipotètica especificada, i no és una afirmació sobre l'explicació en sí mateixa.

3. *Les conclusions científiques i les decisions empresarials o polítiques no s'haurien de fonamentar únicament en el fet que el valor p sobrepassi un llindar específic*

   Les pràctiques que redueixen l'anàlisi de les dades o la inferència científica a l'aplicació mecànica de regles rígides per justificar afirmacions científiques (com, per exemple, "$p<0.05$") poden originar conclusions errònies, o una mala presa de decisions. Una conclusió no es transforma de cop i volta de "certa" per una banda a "falsa" per l'altra. Els investigadors han de considerar que a l'hora d'establir una inferència científica hi ha molts factors en joc que la contextualitzen —inclosos el disseny de l'estudi, la qualitat de les mesures, l'evidència externa sobre el fenomen en estudi i la validació dels supòsits subjacents sota l'anàlisi de les dades. Per consideracions d'ordre pràctic sovint cal prendre decisions binàries (del tipus "si-no"), però això no vol dir que els valors p considerats aïlladament puguin garantir la correcció o la incorrecció d'una decisió. L'ús generalitzat del concepte "significació estadística" (generalment interpretat com "$p\leq0.05$") per legitimar la reclamació d'un descobriment científic (o de la veritat que hi hagi implícita) produeix una distorsió considerable del procés científic.

4. *Realitzar una inferència apropiada requereix un informe complet i transparència*

   Els valors p i les anàlisis relacionades no es poden presentar selectivament. Realitzar diverses anàlisis de les dades i informar només d'aquelles que assoleixen un determinat nivell del valor p (normalment, aquells que passen un llindar de significació) fa que sigui impossible interpretar els valors p publicats. La tria selectiva de troballes prometedores, un costum també conegut amb termes com ara esgotament o tortura de les dades, cerca i fins i tot persecució de la significació, inferència selectiva o manipulació dels valors p,

produeixen un excés distorsionat de resultats estadísticament significatius a la literatura publicada, que haurien de ser vigorosament defugits. Calcular per costum múltiples contrastos estadístics presenta l'inconvenient que cada vegada que un investigador tria quines conclusions presenta d'acord amb uns resultats estadístics, la interpretació vàlida dels resultats queda molt compromesa si el lector no sap que hi ha hagut una elecció i quin ha estat el seu fonament. Els investigadors haurien d'explicar quin ha estat el nombre total d'hipòtesis explorades durant l'anàlisi, quines les decisions preses durant la recopilació de les dades, quines les anàlisis estadístiques dutes a terme, i també haurien de proporcionar tots els valors p calculats. No es poden obtenir conclusions científiques vàlides basades en els valors p i en els estadístics associats sense saber com a mínim quantes i quines anàlisis s'han dut a terme, i com es van seleccionar les que es van presentar (incloent els valors p).

5. *Un valor p, o la significació estadística, no mesura la mida d'un efecte o la importància d'un resultat*

   La significació estadística no és equivalent a la significació científica, humana o econòmica. Uns valors p menors que altres no impliquen necessàriament la presencia d'efectes més grans o més importants, ni valors p més grans impliquen manca d'importància, ni tan sols manca d'efecte. Qualsevol efecte, no importa com sigui de petit, pot produir un valor p petit si la mida mostral o la precisió de la mesura són prou grans, i els efectes grans poden produir valors p molt grans si la mida mostral és molt petita o les mesures són imprecises. De la mateixa manera, efectes estimats idèntics tindran diferents valors p si la precisió de l'estimació difereix.

6. *Per si mateix, un valor p no proporciona una bona mesura de l'evidència en relació amb un model o amb una hipòtesis*

   Els investigadors haurien d'admetre que un valor p descontextualitzat, aïllat d'altres evidències, proporciona una informació limitada. Per exemple, un valor p proper a 0.05 pres tot sol ofereix una evidència molt feble contra la hipòtesis nul·la. De la mateixa manera, un valor p relativament gran no implica cap evidència a favor de la hipòtesi nul·la, atès que moltes altres hipòtesis podrien ser tan consistents com ella amb les dades observades, o encara més. Per aquestes raons, l'anàlisi de dades no s'hauria d'aturar en el càlcul del valor p, si hi ha altres aproximacions apropiades i factibles.

Altres aproximacions

En vista dels mals usos freqüents i dels malentesos amb relació als valors p, alguns estadístics prefereixen complementar, o fins i tot substituir, els valors p per altres procediments. Hi ha mètodes que emfasitzen l'estimació pel damunt del mer posar a prova i contrastar, com ara els intervals de confiança, de credibilitat o de predicció. També es pot recórrer a mètodes bayesians, o a mesures alternatives de l'evidència, com per exemple la prova de la raó de versemblança o els factors de Bayes. I hi ha més possibilitats, com són els models de la teoria de presa de decisions, o la taxa de falsos descobriments. Encara que totes aquestes mesures i enfocaments es fonamenten en supòsits addicionals, podrien afrontar de manera més directa la mida d'un efecte (i la seva incertesa associada), o la comprovació de la validesa d'una hipòtesi.

Conclusió

Les bones pràctiques estadístiques, com a component essencial del bon quefer científic, emfasitzen els principis de dirigir i dur a terme un bon disseny dels estudis, de realitzar-los adequadament, d'aportar una varietat de resums numèrics i gràfics de les dades, d'entendre el fenomen que s'està estudiant, d'interpretar els resultats dins el seu context, de proporcionar una informació integra, i de comprendre de forma adequada, tant lògica com quantitativa, allò que signifiquen els resums de dades. Un índex únic no hauria de substituir el raonament científic.

Agraïment: la Junta Directiva de l'ASA agraeix a les següents persones haver compartit la seva experiència i els seus punts de vista durant la preparació d'aquest manifest. El text no reflecteix necessàriament el punt de vista de totes aquestes persones. De fet, algunes mantenen opinions contraries a les contingudes en la declaració (totalment o només en part). Ara bé, els agraïm enormement les seves contribucions. *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*.

Breu llista de referències sobre el valor p i la significació estadística,

per acompanyar la Declaració de l'ASA sobre la significació estadística i els valors p

La següent llista no és exhaustiva, però proporciona un bon punt de partida per a les persones que els agradaria explorar amb més deteniment les qüestions contingudes a la *Declaració de l'ASA sobre la significació estadística i els valors p*. Els articles apareixen en ordre alfabètic:

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 Medicin*e, 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 Scienc*es, 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