MÈTODE Science Studies Journal (2014). Universitat de València.
DOI: 17.7203/metode.83.3913
Article rebut: 29/07/2014, acceptat: 13/10/2014.

LA FALTA DE REPRODUCTIBILITAT DE LA INVESTIGACIÓ

L'ESTADÍSTICA COM A LEGITIMACIÓ DEL RESULTAT

SCOTT D. GODDARD i VALEN E. JOHNSON

La investigació científica es legitima mitjançant la replicació dels seus resultats, però els esforços per replicar afirmacions enganyoses exhaureixen el finançament. Ens centrarem en una d'aquestes errades: els resultats de proves estadístiques que ofereixen falsos positius a causa de l'atzar. Els mètodes estadístics clàssics confien en un p-valor per a ponderar les proves enfront d'una hipòtesi nul·la, però les proves d'hipòtesis bayesianes ofereixen resultats més fàcils de comprendre, sempre que hom puga especificar distribucions a priori per a la hipòtesi alternativa. Descriurem noves proves, les UMPBT, tests bayesians que ofereixen una especificació per defecte de les alternatives a priori, i mostrarem que aquests tests també maximitzen la potència estadística.

Paraules clau: evidència estadística, test d'hipòtesi, anàlisi bayesià, tests bayesians uniformement més potents.

Poques persones racionals acceptarien els resultats d'una investigació científica si els intents posteriors de validar aquests resultats han fracassat. Llavors, què li passaria al bon nom de la ciència si es descobrira que les troballes de molts estudis prestigiosos no són replicables? Potser anem camí de descobrir-ho. Per una casualitat extensament divulgada, dues firmes farmacèutiques van anunciar recentment que només havien pogut reproduir per complet els resultats revisats i publicats d'una petita fracció d'estudis: entre un 20 i un 25% en el cas d'una de les empreses (Prinz *et al*., 2011) i un 11% en el cas de l'altra (Begley i Ellis, 2012). La majoria d'aquests estudis provaven l'eficàcia de tractaments contra el càncer, un camp en què se sap que l'índex de fracàs de les proves clíniques és alt. Però aquests resultats no són únics en absolut. Els investigadors d'altres camps científics han observat l'escassetat de resultats experimentals reproduïbles (vegeu Hirschhorn *et al*., 2002, per exemple).

Ens fem eco del sentiment expressat en un altre article: «Quan es fan afirmacions aparentment inversemblants amb mètodes convencionals, és un moment ideal per a reexaminar els dits mètodes.» (Rouder i Morey, 2011). Podríem començar un examen d'aquesta mena amb els mètodes estadístics convencionals. Encara que no s'ha difós molt fora de la literatura estadística, hi ha una creixent quantitat de proves que suggereixen que els tests d'hipòtesis clàssiques, tal com s'usen normalment, tendeixen a exagerar la solidesa de les tendències estadístiques (Edwards *et al*. 1963; Berger i Sellke, 1987; Johnson, 2013*a*, 2013*b*). Com a conseqüència, les mateixes pràctiques que usen els científics per a analitzar les seues dades són al seu torn causa de la falta de reproductibilitat de la recerca científica.

> «QUÈ LI PASSARIA AL BON NOM DE LA CIÈNCIA SI ES DESCOBRIRA QUE LES TROBALLES DE MOLTS ESTUDIS PRESTIGIOSOS NO SÓN REPLICABLES?»

ARRIBAR A CONCLUSIONS ERRÒNIES

El problema associat a les proves clàssiques es pot il·lustrar amb un exemple senzill. Imaginem que sabem que la malaltia *W* mata 2 de cada 3 pacients que la contrauen. Suposem que un fàrmac experimental (*A*) promet millorar la taxa de supervivència. Si els inves-


Edu Bayer/SINC

tigadors realitzen un estudi clínic, administrant *A* a 16 pacients, i 9 dels quals sobreviuen, com podem concloure si el fàrmac és eficaç o no? Si no és eficaç, es pot esperar que al voltant d'un terç dels 16 (posem-ne 5) pacients sobrevisqueren. 9 pacients són «aproximadament» 5 pacients? O es diferencia prou de 5 com per justificar l'afirmació que els resultats de la prova són «significatius», és a dir, que el fàrmac *A* és efectiu?

El mètode convencional per a respondre aquesta pregunta és realitzar un test d'hipòtesis unilaterals, en el qual contrastem una hipòtesi nul·la davant la seua hipòtesi alternativa. Diguem que *p* indica la taxa de supervivència de la població després del tractament amb el fàrmac *A*, siga la que siga. La hipòtesi nul·la (H_0) indica que *p* és menor o igual a 1/3, la qual cosa significa que el medicament no és eficaç. La hipòtesi alternativa (H_1) afirma que *p* és major que 1/3, la qual cosa significa que *A* hi ajuda, en certa manera.

En la pràctica estadística estàndard, la hipòtesi nul·la es rebutja en favor de la hipòtesi alternativa si el p-valor de l'experiment és menor que 0,05, on el p-valor es defineix com la probabilitat (si H_0 és certa) d'arreplegar dades almenys tan extremes com les observades. Per tant, 0,05 (el que es coneix com «grandària» de la prova) és un llindar que divideix els p-valors que rebutgen H_0 d'aquells que no ho fan. En la prova del fàrmac, 9 de cada 16 pacients van sobreviure a la malaltia després del tractament amb *A*. El p-valor, la probabilitat d'observar 9 o més supervivents d'entre 16 pacients, si *p* és 1/3, pot calcular-se simplement usant teoria de probabilitat. Resulta ser lleugerament menor que 0,05. Així, en una prova amb grandària 0,05 podem rebutjar la hipòtesi nul·la i concloure que el fàrmac és eficaç.

El problema en aquest cas, respecte als falsos descobriments i la falta de reproductibilitat, és que és més probable del que sembla que hàgem arribat a una conclusió incorrecta. Encara que alguns opinen el contrari, un p-valor de 0,05 no significa que la probabilitat que la hipòtesi nul·la siga vertadera és 0,05 (una interessant discussió al respecte es pot trobar en Sellke *et al*., 2001). De fet, si suposem que el nou medicament tenia la mateixa probabilitat de ser eficaç com de no ser-ho, llavors la probabilitat a favor de la hipòtesi nul·la és com a mínim de 0,15. Una xifra preocupantment alta tenint en compte que acabem de rebutjar-la! Aquest és el principal problema dels tests d'hipòtesis clàssiques: el p-valor, en comparació amb un llindar de 0,05, pot ser prou petit per rebutjar la hipòtesi nul·la (és a dir, que el medicament no és eficaç), però així i tot pot tenir una probabilitat relativament alta de ser cert. Que els científics (que disciplines científiques senceres, de fet) continuen utilitzant un llindar tan alt, mentre que rares vegades informen de la probabilitat que la hipòtesi nul·la siga vertadera, obre una bretxa en la defensa del rigor estadístic que permet que tot d'afirmacions errònies es colen en l'àmbit sagrat de les dades científiques.

«ENTRE UN 17% I UN 25% DE TOTS ELS DESCOBRIMENTS SIGNIFICATIUS DE DUES REVISTES DE PSICOLOGIA EN 2007 EREN, EN REALITAT, FALSOS DESCOBRIMENTS»

En realitat, en el millor dels casos la probabilitat és 0,15. Calcular la probabilitat a favor de la hipòtesi nul·la no és un còmput clàssic, sinó més aviat un de bayesià. Els càlculs bayesians requereixen supòsits addicionals, més enllà dels realitzats en els mètodes clàssics. D'aquests supòsits se'n diu supòsits «a priori» perquè es fan abans d'arreplegar les dades, com una idea preconcebuda que l'investigador aporta a la investigació. Al contrari, dels resultats que es deriven de l'anàlisi de dades se'n diu «a posteriori»; el valor de 0,15 és una probabilitat a posteriori a favor de la hipòtesi nul·la. Calcular-lo requereix realitzar dos supòsits anteriors. En primer lloc, hem d'especificar la probabilitat a priori que la hipòtesi nul·la siga vertadera, o la nostra confiança en H_0, abans de reclutar un sol pacient. En segon lloc, hem de suposar un valor per a *p* sota la hipòtesi alternativa, ja que si el fàrmac és eficaç, serà òbviament superior a un terç.

Quant al primer supòsit, d'ara en avant, hem simplificat l'exposició en suposar que la probabilitat a priori de H_0 (i també H_1) és de 0,5. En absència de qualsevol informació prèvia sobre el nou fàrmac, aquesta ben bé podria ser una suposició raonable.

La ciència es basa en la reproductibilitat dels seus resultats. Recentment, dues firmes farmacèutiques han anunciat que només havien pogut reproduir els resultats d'un percentatge no superior al 25% dels seus estudis publicats en revistes revisades per parells. En la imatge, pacient durant la realització d'un assaig clínic.

El principal problema dels tests d'hipòtesis clàssiques és que el valor *p* pot ser prou petit per a rebutjar la hipòtesi nul·la però que continue havent-hi una probabilitat relativament alta de ser cert. Això permet que es colen tot tipus d'afirmacions errònies com a dades científiques comprovades, com l'aval científic a un fàrmac que no és realment eficient.

L'ÚS DE MÈTODES ESTADÍSTICS INADEQUATS POT DONAR LLOC FÀCILMENT A RESULTATS PERILLOSOS I INEFICIENTS »

Més preocupant, però, és la qüestió de quin valor hauríem de prendre per a *p*, en el cas que la hipòtesi alternativa siga certa (en l'exemple, que el medicament és efectiu). Els diferents supòsits sobre aquesta probabilitat porten a diferents probabilitats a posteriori a favor de la hipòtesi contrària, H_0, i per tant a conclusions diferents. Calculem el valor de 0,15 suposant que si *p* no és 1/3, llavors és 9/16. Per descomptat, podríem haver escollit qualsevol valor entre 0 i 1, però una vegada que es va realitzar l'assaig i 9 pacients van sobreviure, una suposició a priori que $p=9/16$ resulta ser, de totes les suposicions a priori que podríem haver fet, la més hostil respecte a H_0 (i no obstant això, recordem que la probabilitat a posteriori resultant de la hipòtesi nul·la H_0, 0,15, va ser decebedora perquè no era prou hostil). Si en canvi haguérem assumit algun altre valor a priori de *p,* la probabilitat a posteriori seria fins i tot major que 0,15. Per exemple, per a una suposició a priori que *p* és bé 0,3618 o bé 0,75, la probabilitat a posteriori de la hipòtesi nul·la s'eleva a 0,39.

QUAN L'ESTADÍSTICA AVALA AFIRMACIONS ANTICIENTÍFIQUES

Hi ha tres punts clau en aquest exemple. En primer lloc, és evident que les probabilitats a posteriori ben sovint no transmeten una acusació tan decidida contra la hipòtesi nul·la com els p-valors clàssics. En segon lloc, les probabilitats a posteriori depenen en gran manera dels supòsits a priori realitzats per al paràmetre d'interès sota la hipòtesi alternativa, de manera que els supòsits previs afecten subjectivament el resultat de l'anàlisi. En tercer lloc, l'ús de mètodes estadístics inadequats pot donar lloc fàcilment a resultats perillosos i ineficients. Tal seria el cas d'un fàrmac ineficaç contra el càncer que rebera suport científic.

Els punts primer i segon s'il·lustren en una investigació molt mediàtica sobre la percepció extrasensorial. Bem (2011) informa dels resultats de nou experiments que tracten de provar l'existència de la percepció extrasensorial, en els quals la hipòtesi nul·la suposa que no hi ha tal cosa i la hipòtesi alternativa suposa que sí. L'autor analitza les dades de cada experiment calculant p-valors clàssics, i vuit dels nou experiments ofereixen p-valors inferiors a 0,05. N'hi hagué vuit, de resultats significatius a favor de l'existència de la percepció extrasensorial.

Wagenmakers *et al*. (2011) van criticar Bem per, entre altres coses, confiar en els dits p-valors, per la seua coneguda tendència a exagerar el pes de les proves contra la hipòtesi nul·la, i van oferir una reanàlisi de les dades utilitzant mètodes bayesians. Van concloure que les probabilitats a posteriori a favor de la hipòtesi que re-

presentava la no existència de la percepció extrasensorial oscil·laven entre 0,15 i 0,88 en els nou experiments, per la qual cosa «les dades de Bem no protegeixen la hipòtesi de la precognició». En resposta, Bem *et al*. (2011) van assenyalar que els resultats de Wagenmakers *et al*. eren molt sensibles als supòsits a priori realitzats en la grandària de l'efecte sota la hipòtesi alternativa. Sostenien també que aquells supòsits consideraven grandàries d'efecte a pesos elevats que no es troben normalment en experiments psicològics. Finalment, es va reanalitzar les dades utilitzant els mateixos mètodes bayesians, però amb supòsits a priori «basats en el coneixement» sota la hipòtesi alternativa que donaven més pes a grandàries d'efecte menors, i es va descobrir que les probabilitats a posteriori a favor de la hipòtesi nul·la (la no existència de percepció extrasensorial) oscil·laven entre 0,09 i 0,67, la majoria de les quals per sota de 0,3.

TESTS BAYESIANS UNIFORMEMENT MÉS POTENTS

L'acalorat debat sobre els mètodes utilitzats per Bem (2011) –que comprèn molts més articles que els que apareixen citats ací– subratlla la naturalesa poc fiable dels p-valors i la polèmica entorn dels mètodes de càlcul de probabilitats a posteriori. Depenent de la seua opinió sobre la percepció extrasensorial, també pot demostrar com una confiança inadequada en els tests d'hipòtesis clàssiques pot premiar amb l'aprovació dels revisors una afirmació enganyosa i anticientífica.

Recentment, hem proposat un nou acostament a la resolució del segon d'aquests problemes, el d'establir supòsits a priori per a p. La idea bàsica de la nostra proposta és que, en primer lloc, experts rellevants en la investigació haurien d'establir un llindar d'evidència per a la probabilitat a posteriori a favor de la hipòtesi nul·la, anàleg en certa manera al límit establit per als p-valors. Després d'això, però abans d'arreplegar les dades, s'hauria de permetre als investigadors –que normalment esperen rebutjar H_0 quan arriben els resultats– que realitzen supòsits a priori sota la hipòtesi alternativa que maximitzen les possibilitats de rebutjar la hipòtesi nul·la. Això es pot fer d'una manera relativament senzilla en gran part dels tests. Les proves resultants es denominen tests bayesians uniformement més potents (UMPBT), i podem il·lustrar-ne l'ús amb la nostra hipotètica prova clínica de fàrmacs.

Suposem que el mecenes de la prova clínica exigeix, per exemple, que la hipòtesi nul·la es rebutge únicament si la probabilitat a posteriori cau per sota de 0,05. Per a un investigador que vulga declarar l'èxit del nou fàrmac, la pregunta rellevant a l'hora d'establir p sota la hipòtesi alternativa és: «Quin valor suposat per a p maximitzarà la possibilitat que la probabilitat a posteriori a favor de H_0 *siga* inferior a 0,05?»

Utilitzant la metodologia de Johnson (2013*a*), el supòsit a priori més favorable que pot realitzar l'investigador és $p=0{,}63$. Aquest valor maximitza la possibilitat que la probabilitat a posteriori de la hipòtesi nul·la siga inferior a 0,05, sense importar el valor real de p. Des de la perspectiva de l'investigador, és l'elecció òptima d'entre tots els possibles supòsits sota la hipòtesi alternativa, i si permetem que els investigadors ho trien, s'elimina la subjectivitat en seleccionar la hipòtesi alternativa.

D'altra banda, usar una probabilitat a priori per a la supervivència de $p=0{,}63$ i un llindar de 0,05 sota la hipòtesi alternativa implica que la hipòtesi nul·la només es rebutjarà si 11 o més pacients sobreviuen després de rebre el medicament. Aquest és el mateix criteri que s'usaria per a rebutjar la hipòtesi nul·la en un test clàssic

CERN

La probabilitat de l'existència del bosó de Higgs està entre el 0,999963 i 0,999977, la qual cosa és una prova molt clara però no tant com la que s'argumentava en l'informe original. En la imatge, els físics François Englert i Peter Higgs, durant l'anunci del descobriment en el CERN.

«UNA APROVACIÓ INADEQUADA EN ELS TESTS D'HIPÒTESIS CLÀSSIQUES POT PREMIAR AMB L'APROVACIÓ DELS REVISORS UNA AFIRMACIÓ ENGANYOSA I ANTICIENTÍFICA»

de grandària 0,004. Per tant, el requisit que la probabilitat a posteriori per a la hipòtesi nul·la siga inferior a un llindar baix (0,05) perquè resulte significatiu implica que el p-valor haurà de ser inferior a un llindar molt baix (0,004). En aquest cas, els tests bayesians uniformement més potents ofereixen una manera objectiva de realitzar supòsits a priori sota H_1 i al mateix temps limiten l'excessiva permissivitat dels tests d'hipòtesis clàssiques.

«ELS MÈTODES D'ANÀLISI BAYESIANS SÓN UNA FORMA SIMPLE I POTENT DE REDUIR LA FALTA DE REPRODUCTIBILITAT EN LA CIÈNCIA MODERNA»

A més, com que els tests bayesians uniformement més potents es poden utilitzar per a establir supòsits a priori objectius sota la hipòtesi alternativa, són útils per a revisar publicacions en què es van utilitzar originàriament p-valors clàssics i calcular les probabilitats a posteriori. Usant aquests mètodes, Johnson (2013*a*) sosté que la probabilitat a posteriori a favor de l'existència del bosó de Higgs pot estar prop d'entre 0,999963 i 0,999977 –encara una prova molt clara, però potser no tant com s'argumentava en l'informe original amb un p-valor de 3×10^{-7}. En un altre article (Johnson 2013*b*), va utilitzar aquests tests per a valorar que entre un 17 % i un 25 % de tots els descobriments significatius de dues revistes de psicologia en 2007 eren, en realitat, falsos descobriments. Finalment, tornant a l'estudi sobre la percepció extrasensorial, podem usar una versió aproximada dels tests bayesians uniformement més potents per a establir les probabilitats a posteriori per a la hipòtesi nul·la entre 0,12 i 0,39, quan s'utilitza 0,05 com a llindar de significació.

CONCLUSIÓ

En resum, volem remarcar que els llindars que s'utilitzen actualment en els tests clàssics de significació estadística són responsables de gran part de la falta de reproductibilitat dels estudis científics que s'ha observat en la premsa popular i en revistes especialitzades. Entre els milers d'afirmacions que apareixen publicades cada any, una gran part dels estudis marginalment significatius al nivell 0,05 són, de fet, falses troballes. No obstant això, els mètodes d'anàlisi bayesians que calculen la probabilitat a posteriori a favor de la hipòtesi nul·la pal·lien la falta de fiabilitat dels p-valors, i quan els supòsits a priori sota la hipòtesi alternativa es realitzen mitjançant tests bayesians uniformement més potents, la probabilitat resultant a posteriori és objectiva i equivalent a un test clàssic, però segueix estàndards de proves més alts. Considerem que aquests mètodes bayesians de contrast són una forma simple i potent de reduir la falta de reproductibilitat en la ciència moderna.

REFERÈNCIES

Begley, C. i L. Ellis, 2012. «Drug Development: Raise Standards for Preclinical Cancer Research». *Nature*, 483(7391): 531-533. DOI: <10.1038/483531a>.

Bem, D., 2011. «Feeling the Future: Experimental Evidence for Anomalous Retroactive Influences on Cognition and Effect». *Journal Personality and Social Psychology*, 100(3): 407-425. DOI: <10.1037/a0021524>.

Bem, D.; Utts, J. i W. Johnson, 2011. «Must Psychologists Change the Way they Analyze Their Data?». *Journal Personality and Social Psychology*, 101(4): 716-719. DOI: <10.1037/a0024777>.

Berger, J. i T. Sellke, 1987. «Testing a Point Null Hypothesis: Irreconcilability of p-values and Evidence». *Journal of the American Statistical Association*, 82(397): 112-122. DOI: <10.2307/2289131>.

Edwards, W.; Lindman, H. i L. Savage, 1963. «Bayesian Statistical Inference for Psychological Research». *Psychological Review*, 70(3): 193-242. DOI: <10.1037/h0044139>.

Hirschhorn, J.; Lohmueller, K.; Byrne, E. i K. Hirschhorn, 2002. «A Comprehensive Review of Genetic Association Studies». *Genetics in Medicine*, 4(2): 45-61. DOI: <10.1097/00125817-200203000-00002>.

Johnson, V. E., 2013*a*. «Uniformly Most Powerful Bayesian Test». *The Annals of Statistics*, 41(1): 1716-1741. DOI: <10.1214/13-AOS1123>.

Johnson, V. E., 2013*b*. «Revised Standards for Statistical Evidence». *PNAS*, 110(48): 19313-19317. DOI: <10.1073/pnas.1313476110>.

Prinz, F.; Schlange, T. i K. Asadullah, 2011. «Believe It or Not: How Much Can We Rely on Published Data on Potential Drug Targets?». *Nature Reviews Drug Discovery*, 10(9): 712. DOI: <10.1038/nrd3439-c1>.

Rouder, J. i R. Morey, 2011. «A Bayes Factor Meta-analysis of Bem's ESP Claim». *Psychonomic Bulleton and Review*, 18(4): 682-689. DOI: <10.3758/s13423-011-0088-7>.

Sellke, T.; Bayarri, M. i J. Berger, 2001. «Calibration of p-values for Testing Precise Null Hypotheses». *The American Statistician*, 55(1): 62-71. DOI: <10.1198/000313001300339950>.

Wagenmakers, E.; Wetzels, R.; Borsboom, D. i H. van der Maas, 2011. «Why Psychologists Must Change the Way they Analyze Their Data: the Case of Psi: Comment on Bem (2011)». *Journal of Personality and Social Psychology*, 100(3): 426-432. DOI: <10.1037/a0022790>.

ABSTRACT

The Lack of Reproducibility in Research. How Statistics Can Endorse Results.

Scientific research is validated by reproduction of the results, but efforts to reproduce spurious claims drain resources. We focus on one cause of such failure: false positive statistical test results caused by random variability. Classical statistical methods rely on p-values to measure the evidence against null hypotheses, but Bayesian hypothesis testing produces more easily understood results, provided one can specify prior distributions under the alternative hypothesis. We describe new tests, UMPBTs, which are Bayesian tests that provide default specification of alternative priors, and show that these tests also maximize statistical power.

Keywords: statistical evidence, hypothesis test, Bayesian analysis, uniformly most powerful Bayesian tests.

Scott D. Goddard. Estudiant de doctorat del departament d'Estadística. Universitat de Texas (EUA).

Valen E. Johnson. Cap del departament d'Estadística. Universitat de Texas (EUA).