MÈTODE Science Studies Journal (2014). Universitat de València.
DOI: 10.7203/metode.83.3913
Artículo recibido: 29/07/2014, aceptado: 13/10/2014.

LA FALTA DE REPRODUCIBILIDAD DE LA INVESTIGACIÓN

LA ESTADÍSTICA COMO LEGITIMACIÓN DEL RESULTADO

SCOTT D. GODDARD Y VALEN E. JOHNSON

La investigación científica se legitima mediante la replicación de sus resultados, pero los esfuerzos por replicar afirmaciones engañosas agotan la financiación. Nos centraremos en una de las causas de esos errores: los resultados de pruebas estadísticas que ofrecen falsos positivos debido al azar. Los métodos estadísticos clásicos confían en los p-valores para ponderar las pruebas frente a una hipótesis nula, pero las pruebas de hipótesis bayesianas ofrecen resultados más fáciles de comprender, siempre que uno pueda especificar distribuciones a priori para la hipótesis alternativa. Describiremos nuevas pruebas, los UMPBT, test bayesianos que ofrecen una especificación por defecto de las alternativas a priori, y mostraremos que estos test también maximizan la potencia estadística.

Palabras clave: evidencia estadística, test de hipótesis, análisis bayesiano, test bayesianos uniformemente más potentes.

Pocas personas racionales aceptarían los resultados de una investigación científica si los intentos posteriores de validar esos resultados han fracasado. Entonces, ¿qué le pasaría al buen nombre de la ciencia si se descubriese que los hallazgos de muchos estudios prestigiosos no fueran replicables? Quizás estemos en camino de descubrirlo. Por una casualidad extensamente divulgada, dos firmas farmacéuticas anunciaron recientemente que solo habían podido reproducir por completo los resultados revisados y publicados de una pequeña fracción de sus estudios: entre un 20 y un 25 % en el caso de una de las empresas (Prinz *et al*., 2011) y un 11 % en el caso de la otra (Begley y Ellis, 2012). La mayoría de estos estudios probaban la eficacia de tratamientos contra el cáncer, un campo en el que se sabe que el índice de fracaso de las pruebas clínicas es alto. Pero estos resultados no son únicos en absoluto. Los investigadores de otros campos científicos han observado la escasez de resultados experimentales reproducibles (véase Hirschhorn *et al*., 2002, por ejemplo).

Nos hacemos eco del sentimiento expresado en otro artículo: «Cuando se hacen afirmaciones aparentemente inverosímiles con métodos convencionales, es un momento ideal para reexaminar dichos métodos.» (Rouder y Morey, 2011). Podríamos comenzar un examen de este tipo con los métodos estadísticos convencionales. Aunque no se ha difundido mucho fuera de la literatura estadística, existe una creciente cantidad de pruebas que sugieren que los test de hipótesis clásicos, tal y como se usan normalmente, tienden a exagerar la solidez de las tendencias estadísticas (Edwards *et al*., 1963; Berger y Sellke, 1987; Johnson, 2013*a*, 2013*b*). Como consecuencia, las propias prácticas que usan los científicos para analizar sus datos son a su vez causa de la falta de reproductibilidad de la investigación científica.

«¿QUÉ LE PASARÍA AL BUEN NOMBRE DE LA CIENCIA SI SE DESCUBRIESE QUE LOS HALLAZGOS DE MUCHOS ESTUDIOS RESPETADOS NO SON REPLICABLES?»

LLEGAR A CONCLUSIONES ERRÓNEAS

El problema asociado a las pruebas clásicas se puede ilustrar con un ejemplo sencillo. Imaginemos que sabemos que la enfermedad *W* mata a 2 de cada 3 pacientes que la contraen. Supongamos que un fármaco experimental (*A*) promete mejorar la tasa de supervivencia. Si los investigadores realizan un estudio clínico, adminis-


Edu Bayer/SINC

trando *A* a 16 pacientes, y 9 de ellos sobreviven, ¿cómo podemos concluir si el fármaco es eficaz o no? Si no es eficaz, sería de esperar que alrededor de un tercio de los 16 (pongamos que 5) pacientes sobrevivieran. ¿9 pacientes son «aproximadamente» 5 pacientes? ¿O se diferencia lo suficiente de 5 como para justificar la afirmación de que los resultados de la prueba son «significativos», es decir, que el fármaco *A* es efectivo?

El método convencional para responder a esta pregunta es realizar un test de hipótesis unilateral, en el que contrastamos una hipótesis nula frente a su hipótesis alternativa. Digamos que *p* indica la tasa de supervivencia de la población después del tratamiento con el fármaco *A*, cualquiera que sea. La hipótesis nula (H_0) indica que *p* es menor o igual a 1/3, lo que significa que el medicamento no es eficaz. La hipótesis alternativa (H_1) implica que *p* es mayor que 1/3, lo que significa que *A* ayuda, en cierta medida.

«ENTRE UN 17% Y UN 25% DE TODOS LOS DESCUBRIMIENTOS SIGNIFICATIVOS DE DOS REVISTAS DE PSICOLOGÍA EN 2007 ERAN, EN REALIDAD, FALSOS DESCUBRIMIENTOS»

En la práctica estadística estándar, la hipótesis nula se rechaza en favor de la hipótesis alternativa si el p-valor del experimento es menor que 0,05, donde el p-valor se define como la probabilidad (si H_0 es cierta) de recoger datos al menos tan extremos como los observados. Por lo tanto, 0,05 (lo que se conoce como el «tamaño» de la prueba) es un umbral que divide los p-valores que rechazan H_0 de aquellos que no lo hacen. En la prueba del fármaco, 9 de cada 16 pacientes sobrevivieron a la enfermedad tras el tratamiento con *A*. El p-valor, la probabilidad de observar 9 o más supervivientes de entre 16 pacientes, si *p* es 1/3, puede calcularse simplemente usando teoría de probabilidad. Resulta ser ligeramente menor que 0,05. Así, en una prueba con tamaño 0,05 podemos rechazar la hipótesis nula y concluir que el fármaco es eficaz.

El problema en este caso, con respecto a los falsos descubrimientos y la falta de reproducibilidad, es que es más probable de lo que parece que hayamos llegado a una conclusión incorrecta. Aunque algunos opinan lo contrario, un p-valor de 0,05 no significa que la probabilidad de que la hipótesis nula sea verdadera es 0,05 (una interesante discusión al respecto se puede encontrar en Sellke *et al*., 2001). De hecho, si suponemos que el nuevo medicamento tenía la misma probabilidad de ser eficaz como de no serlo, entonces la probabilidad a favor de la hipótesis nula es de al menos 0,15. ¡Una cifra preocupantemente alta teniendo en cuenta que acabamos de rechazarla! Este es el principal problema de los test de hipótesis clásicos: el p-valor, en comparación con un umbral de 0,05, puede ser lo suficientemente pequeño como para rechazar la hipótesis nula (el medicamento no es eficaz), pero aun así puede tener una probabilidad relativamente alta de ser cierto. Que los científicos (que disciplinas científicas enteras, de hecho) continúen utilizando un umbral tan alto, mientras que rara vez informan de la probabilidad de que la hipótesis nula sea verdadera, crea una brecha en la defensa del rigor estadístico que permite que todo tipo de afirmaciones erróneas se cuelen en el ámbito sagrado de los datos científicos.

En realidad, en el mejor de los casos la probabilidad es 0,15. Calcular la probabilidad a favor de la hipótesis nula no es un cómputo clásico, sino más bien uno bayesiano. Los cálculos bayesianos requieren supuestos adicionales, más allá de los realizados en los métodos clásicos. A estos supuestos se les llama supuestos «a priori» porque se hacen antes de recoger los datos, como una idea preconcebida que el investigador aporta a la investigación. Por el contrario, a los resultados que se derivan del análisis de datos se les llama «a posteriori»; el valor de 0,15 es una probabilidad a posteriori a favor de la hipótesis nula. Calcularlo requiere realizar dos supuestos anteriores. En primer lugar, debemos especificar la probabilidad a priori de que la hipótesis nula sea verdadera, o nuestra confianza en H_0, antes de reclutar a un solo paciente. En segundo lugar, debemos suponer un valor para *p* bajo la hipótesis alternativa, ya que si el fármaco es eficaz, será obviamente superior a un tercio.

En cuanto al primer supuesto, de aquí en adelante, hemos simplificado la exposición al suponer que la probabilidad a priori de H_0 (y también H_1) es de 0,5. En ausencia de cualquier información previa acerca del nuevo fármaco, esta bien podría ser una suposición razonable.

La ciencia se basa en la reproducibilidad de sus resultados. Recientemente, dos firmas farmacéuticas han anunciado que solo habían podido reproducir los resultados de un porcentaje no superior al 25% de sus estudios publicados en revistas revisadas por pares. En la imagen, pacientes durante la realización de un ensayo clínico.

El principal problema de los test de hipótesis clásicos es que el valor *p* puede ser lo suficientemente pequeño como para rechazar la hipótesis nula pero que siga habiendo una probabilidad relativamente alta de ser cierto. Esto permite que se cuelen todo tipo de afirmaciones erróneas como datos científicos comprobados, como el respaldo científico a un fármaco que no es realmente eficiente.

«EL USO DE MÉTODOS ESTADÍSTICOS INADECUADOS PUEDE DAR LUGAR FÁCILMENTE A RESULTADOS PELIGROSOS E INEFICIENTES»

Más preocupante, sin embargo, es la cuestión de qué valor deberíamos tomar para *p*, en el supuesto de que la hipótesis alternativa es cierta (en el ejemplo, que el medicamento es efectivo). Los diferentes supuestos sobre esta probabilidad llevan a diferentes probabilidades a posteriori a favor de la hipótesis contraria, $H_{0,}$ y por lo tanto a conclusiones diferentes. Calculamos el valor de 0,15 suponiendo que si p no es 1/3, entonces es 9/16. Por supuesto, podríamos haber elegido cualquier valor entre 0 y 1, pero una vez que se realizó el ensayo y 9 pacientes sobrevivieron, una suposición a priori de que $p=9/16$ resulta ser, de todas las suposiciones a priori que podríamos haber hecho, la más hostil con respecto a H_0 (y sin embargo, recordemos que la probabilidad a posteriori resultante de la hipótesis nula H_0, 0,15, fue decepcionante porque no era lo suficientemente hostil). Si en cambio hubiésemos asumido algún otro valor a priori de *p,* la probabilidad a posteriori sería incluso mayor que 0,15. Por ejemplo, para una suposición a priori de que p es bien 0,3618 o bien 0,75, la probabilidad a posteriori de la hipótesis nula se eleva a 0,39.

CUANDO LA ESTADÍSTICA APOYA AFIRMACIONES ANTICIENTÍFICAS

Hay tres puntos clave en este ejemplo. En primer lugar, es evidente que las probabilidades a posteriori con frecuencia no transmiten una acusación tan decidida contra la hipótesis nula como los p-valores clásicos. En segundo lugar, las probabilidades a posteriori dependen en gran medida de los supuestos a priori realizados para el parámetro de interés bajo la hipótesis alternativa, de modo que los supuestos previos afectan subjetivamente al resultado del análisis. En tercer lugar, el uso de métodos estadísticos inadecuados puede dar lugar fácilmente a resultados peligrosos y antieconómincos. Tal sería el caso de un fármaco ineficaz contra el cáncer que recibiera respaldo científico.

Los puntos primero y segundo se ilustran en una investigación muy mediática sobre la percepción extrasensorial. Bem (2011) informa de los resultados de nueve experimentos que tratan de probar la existencia de la percepción extrasensorial, en los que la hipótesis nula supone que no existe tal cosa y la hipótesis alternativa supone que sí. El autor analiza los datos de cada experimento calculando p-valores clásicos, y ocho de los nueve experimentos ofrecen p-valores inferiores a 0,05. Hubo ocho resultados significativos a favor de la existencia de la percepción extrasensorial.

Wagenmakers *et al*. (2011) criticaron a Bem por, entre otras cosas, confiar en dichos p-valores, por su conocida tendencia a exagerar el peso de las pruebas contra la hipótesis nula, y ofrecieron un reanálisis de los

datos utilizando métodos bayesianos. Concluyeron que las probabilidades a posteriori a favor de la hipótesis que suponía la no existencia de la percepción extrasensorial oscilaban entre 0,15 y 0,88 en los nueve experimentos, por lo que «los datos de Bem no respaldan la hipótesis de la precognición». En respuesta, Bem *et al*. (2011) señalaron que los resultados de Wagenmakers *et al*. eran muy sensibles a los supuestos a priori realizados en el tamaño del efecto bajo la hipótesis alternativa. Sostenían también que aquellos supuestos consideraban tamaños de efecto con pesos elevados que no se encuentran normalmente en experimentos psicológicos. Finalmente, se reanalizaron los datos utilizando los mismos métodos bayesianos, pero con supuestos a priori «basados en el conocimiento» bajo la hipótesis alternativa que daban más peso a tamaños de efecto menores, y se descubrió que las probabilidades a posteriori a favor de la hipótesis nula (la no existencia de percepción extrasensorial) oscilaban entre 0,09 y 0,67, la mayoría de ellas por debajo de 0,3.

CERN

La probabilidad de la existencia del bosón de Higgs está entre el 0,999963 y 0,999977, lo que es una prueba muy clara pero no tanto como la que se argumentaba en el informe original. En la imagen, los físicos François Englert y Peter Higgs, durante el anuncio del descubrimiento en el CERN.

TEST BAYESIANOS UNIFORMEMENTE MÁS POTENTES

El acalorado debate acerca de los métodos utilizados por Bem (2011) –que comprende muchos más artículos que los que aparecen citados aquí– subraya la naturaleza poco fiable de los p-valores y la polémica en torno a los métodos de cálculo de probabilidades a posteriori. Dependiendo de su opinión sobre la percepción extrasensorial, también puede demostrar cómo una confianza inadecuada en los test de hipótesis clásicos puede premiar con la aprobación de los revisores una afirmación engañosa y anticientífica.

«LOS MÉTODOS DE ANÁLISIS BAYESIANOS SON UNA FORMA SIMPLE Y POTENTE DE REDUCIR LA FALTA DE REPRODUCIBILIDAD EN LA CIENCIA MODERNA»

Recientemente, hemos propuesto un nuevo acercamiento a la resolución del segundo de estos problemas, el de establecer supuestos a priori para *p*. La idea básica de nuestra propuesta es que, en primer lugar, expertos relevantes en la investigación deberían establecer un umbral de evidencia para la probabilidad a posteriori a favor de la hipótesis nula, análogo en cierta medida al límite establecido para los p-valores. Después de eso, pero antes de recoger los datos, se debería permitir a los investigadores –que normalmente esperan rechazar H_0 cuando llegan los resultados– que realicen supuestos a priori bajo la hipótesis alternativa que maximicen las posibilidades de rechazar la hipótesis nula. Esto se puede hacer de una manera relativamente sencilla en gran parte de los test. Las pruebas resultantes se denominan test bayesianos uniformemente más potentes (UMPBT), y podemos ilustrar su uso con nuestra hipotética prueba clínica de fármacos.

Supongamos que el mecenas de la prueba clínica exige, por ejemplo, que la hipótesis nula se rechace únicamente si la probabilidad a posteriori cae por debajo de 0,05. Para un investigador que quiera declarar el éxito del nuevo fármaco, la pregunta relevante a la hora de establecer *p* bajo la hipótesis alternativa es: «¿Qué valor supuesto para *p* maximizará la posibilidad de que la probabilidad a posteriori a favor de H_0 sea inferior a 0,05?»

Utilizando la metodología de Johnson (2013*a*), el supuesto a priori más favorable que puede realizar el investigador es $p=0{,}63$. Este valor maximiza la posibilidad de que la probabilidad a posteriori de la hipótesis nula sea inferior a 0,05, sin importar el valor real de *p*. Desde la perspectiva del investigador, es la elección óptima de entre todos los posibles supuestos bajo la hipótesis alternativa, y si permitimos que los investigadores lo elijan, se elimina la subjetividad al seleccionar la hipótesis alternativa.

Por otra parte, usar una probabilidad a priori para la supervivencia de $p=0{,}63$ y un umbral de 0,05 bajo

la hipótesis alternativa implica que la hipótesis nula solo se rechazará si 11 o más pacientes sobreviven tras recibir el medicamento. Este es el mismo criterio que se usaría para rechazar la hipótesis nula en un test clásico de tamaño 0,004. Por lo tanto, el requisito de que la probabilidad a posteriori para la hipótesis nula sea inferior a un umbral bajo (0,05) para que resulte significativo implica que el p-valor deberá ser inferior a un umbral muy bajo (0,004). En este caso, los test bayesianos uniformemente más potentes ofrecen una forma objetiva de realizar supuestos a priori bajo H_1 y al mismo tiempo limitan la excesiva permisividad de los test de hipótesis clásicos.

Además, como los test bayesianos uniformemente más potentes se pueden utilizar para establecer supuestos a priori objetivos bajo la hipótesis alternativa, son útiles para revisar publicaciones en las que se utilizaron originalmente p-valores clásicos y calcular las probabilidades a posteriori. Usando estos métodos, Johnson (2013*a*) sostiene que la probabilidad a posteriori a favor de la existencia del bosón de Higgs puede estar cerca de entre 0,999963 y 0,999977 –todavía una prueba muy clara, pero quizás no tanto como se argumentaba en el informe original con un p-valor de 3×10^{-7}. En otro artículo (Johnson 2013*b*), utilizó estos test para valorar que entre un 17 % y un 25 % de todos los descubrimientos significativos de dos revistas de psicología en 2007 eran, en realidad, falsos descubrimientos. Por último, volviendo al estudio sobre la percepción extrasensorial, podemos usar una versión aproximada de los test bayesianos uniformemente más potentes para establecer las probabilidades a posteriori para la hipótesis nula entre 0,12 y 0,39, cuando se utiliza 0,05 como umbral de significación.

CONCLUSIÓN

En resumen, queremos hacer hincapié en que los umbrales que se utilizan actualmente en los test clásicos de significación estadística son responsables de gran parte de la falta de reproducibilidad de los estudios científicos que se ha observado en la prensa popular y en revistas especializadas. Entre las miles de afirmaciones que aparecen publicadas cada año, una gran parte de los estudios marginalmente significativos al nivel 0,05 son, de hecho, falsos hallazgos. Sin embargo, los métodos de análisis bayesianos que calculan la probabilidad a posteriori a favor de la hipótesis nula palian la falta de fiabilidad de los p-valores, y cuando los supuestos a priori bajo la hipótesis alternativa se realizan mediante test bayesianos uniformemente más potentes, la probabilidad a posteriori resultante es objetiva y equivalente a un test clásico, pero sigue estándares de pruebas más altos. Consideramos que estos métodos bayesianos de contraste son una forma simple y potente de reducir la falta de reproducibilidad en la ciencia moderna.

REFERENCIAS

Begley, C. y 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. y 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. y 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. y 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. y 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. y 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. y 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. y 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. y 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. Estudiante de doctorado del departamento de Estadística. Universidad de Texas (EE UU).
Valen E. Johnson. Jefe del departamento de Estadística. Universidad de Texas (EE UU).