Redefine statistical significance

We propose to change the default *P*-value threshold for statistical significance from 0.05 to 0.005 for claims of new discoveries.

Daniel J. Benjamin, James O. Berger, Magnus Johannesson, Brian A. Nosek, E.-J. Wagenmakers, Richard Berk, Kenneth A. Bollen, Björn Brembs, Lawrence Brown, Colin Camerer, David Cesarini, Christopher D. Chambers, Merlise Clyde, Thomas D. Cook, Paul De Boeck, Zoltan Dienes, Anna Dreber, Kenny Easwaran, Charles Efferson, Ernst Fehr, Fiona Fidler, Andy P. Field, Malcolm Forster, Edward I. George, Richard Gonzalez, Steven Goodman, Edwin Green, Donald P. Green, Anthony Greenwald, Jarrod D. Hadfield, Larry V. Hedges, Leonhard Held, Teck Hua Ho, Herbert Hoijtink, Daniel J. Hruschka, Kosuke Imai, Guido Imbens, John P. A. Ioannidis, Minjeong Jeon, James Holland Jones, Michael Kirchler, David Laibson, John List, Roderick Little, Arthur Lupia, Edouard Machery, Scott E. Maxwell, Michael McCarthy, Don Moore, Stephen L. Morgan, Marcus Munafó, Shinichi Nakagawa, Brendan Nyhan, Timothy H. Parker, Luis Pericchi, Marco Perugini, Jeff Rouder, Judith Rousseau, Victoria Savalei, Felix D. Schönbrodt, Thomas Sellke, Betsy Sinclair, Dustin Tingley, Trisha Van Zandt, Simine Vazire, Duncan J. Watts, Christopher Winship, Robert L. Wolpert, Yu Xie, Cristobal Young, Jonathan Zinman and Valen E. Johnson

The lack of reproducibility of scientific studies has caused growing concern over the credibility of claims of new discoveries based on 'statistically significant' findings. There has been much progress toward documenting and addressing several causes of this lack of reproducibility (for example, multiple testing, *P*-hacking, publication bias and under-powered studies). However, we believe that a leading cause of non-reproducibility has not yet been adequately addressed: statistical standards of evidence for claiming new discoveries in many fields of science are simply too low. Associating statistically significant findings with $P < 0.05$ results in a high rate of false positives even in the absence of other experimental, procedural and reporting problems.

For fields where the threshold for defining statistical significance for new discoveries is $P < 0.05$, we propose a change to $P < 0.005$. This simple step would immediately improve the reproducibility of scientific research in many fields. Results that would currently be called significant but do not meet the new threshold should instead be called suggestive. While statisticians have known the relative weakness of using $P \approx 0.05$ as a threshold for discovery and the proposal to lower it to 0.005 is not new[1,2], a critical mass of researchers now endorse this change.

We restrict our recommendation to claims of discovery of new effects. We do not address the appropriate threshold for confirmatory or contradictory replications of existing claims. We also do not advocate changes to discovery thresholds in fields that have already adopted more stringent standards (for example, genomics and high-energy physics research; see the 'Potential objections' section below).

We also restrict our recommendation to studies that conduct null hypothesis significance tests. We have diverse views about how best to improve reproducibility, and many of us believe that other ways of summarizing the data, such as Bayes factors or other posterior summaries based on clearly articulated model assumptions, are preferable to *P* values. However, changing the *P* value threshold is simple, aligns with the training undertaken by many researchers, and might quickly achieve broad acceptance.

Strength of evidence from *P* values

In testing a point null hypothesis H_0 against an alternative hypothesis H_1 based on data x_{obs}, the *P* value is defined as the probability, calculated under the null hypothesis, that a test statistic is as extreme or more extreme than its observed value. The null hypothesis is typically rejected — and the finding is declared statistically significant — if the *P* value falls below the (current) type I error threshold $\alpha = 0.05$.

From a Bayesian perspective, a more direct measure of the strength of evidence for H_1 relative to H_0 is the ratio of their probabilities. By Bayes' rule, this ratio may be written as:

$$\frac{\Pr(H_1 \mid x_{obs})}{\Pr(H_0 \mid x_{obs})} = \frac{f(x_{obs} \mid H_1)}{f(x_{obs} \mid H_0)} \times \frac{\Pr(H_1)}{\Pr(H_0)} \equiv \mathrm{BF} \times (\text{prior odds}) \quad (1)$$

where BF is the Bayes factor that represents the evidence from the data, and the prior odds can be informed by researchers' beliefs, scientific consensus, and validated evidence from similar research questions in the same field. Multiple-hypothesis testing, *P*-hacking and publication bias all reduce the credibility of evidence. Some of these practices reduce the prior odds of H_1 relative to H_0 by changing the population of hypothesis tests that are reported. Prediction markets[3] and analyses of replication results[4] both suggest that for psychology experiments, the prior odds of H_1 relative to H_0 may be only about 1:10. A similar number has been suggested in cancer clinical trials, and the number is likely to be much lower in preclinical biomedical research[5].

There is no unique mapping between the *P* value and the Bayes factor, since the Bayes factor depends on H_1. However, the connection between the two quantities can be evaluated for particular test statistics under certain classes of plausible alternatives (Fig. 1).

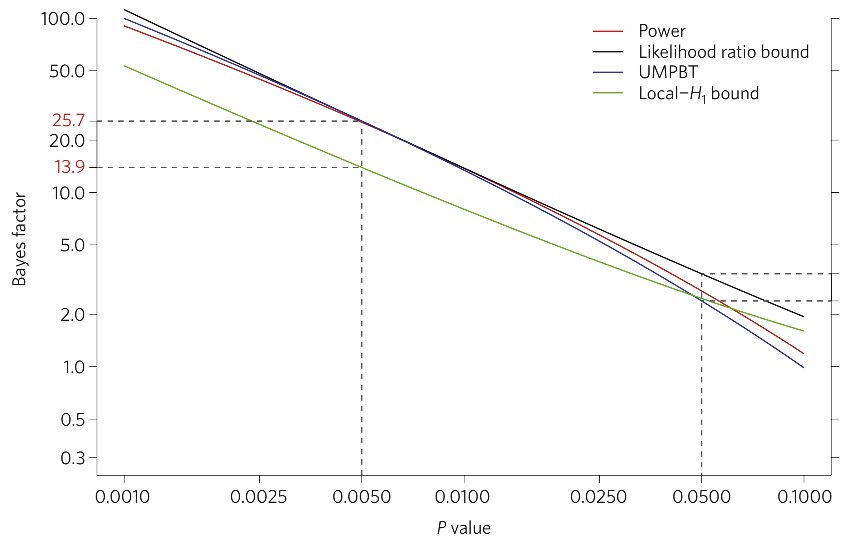

Fig. 1 | Relationship between the *P* value and the Bayes factor. The Bayes factor (BF) is defined as $\frac{f(x_{\text{obs}} \mid H_1)}{f(x_{\text{obs}} \mid H_0)}$. The figure assumes that observations are independent and identically distributed (i.i.d.) according to $x \sim N(\mu,\sigma^2)$, where the mean μ is unknown and the variance σ^2 is known. The *P* value is from a two-sided *z*-test (or equivalently a one-sided χ_1^2-test) of the null hypothesis H_0: $\mu = 0$. Power (red curve): BF obtained by defining H_1 as putting ½ probability on $\mu = \pm m$ for the value of *m* that gives 75% power for the test of size $\alpha = 0.05$. This H_1 represents an effect size typical of that which is implicitly assumed by researchers during experimental design. Likelihood ratio bound (black curve): BF obtained by defining H_1 as putting ½ probability on $\mu = \pm\hat{x}$, where $\hat{x}$ is approximately equal to the mean of the observations. These BFs are upper bounds among the class of all H_1 terms that are symmetric around the null, but they are improper because the data are used to define H_1. UMPBT (blue curve): BF obtained by defining H_1 according to the uniformly most powerful Bayesian test[2] that places ½ probability on $\mu = \pm w$, where *w* is the alternative hypothesis that corresponds to a one-sided test of size 0.0025. This curve is indistinguishable from the 'Power' curve that would be obtained if the power used in its definition was 80% rather than 75%. Local-H_1 bound (green curve): $\text{BF} = \frac{1}{-ep\ln p}$, where *p* is the *P* value, is a large-sample upper bound on the BF from among all unimodal alternative hypotheses that have a mode at the null and satisfy certain regularity conditions[15]. The red numbers on the *y* axis indicate the range of Bayes factors that are obtained for *P* values of 0.005 or 0.05. For more details, see the Supplementary Information.

A two-sided *P* value of 0.05 corresponds to Bayes factors in favour of H_1 that range from about 2.5 to 3.4 under reasonable assumptions about H_1 (Fig. 1). This is weak evidence from at least three perspectives. First, conventional Bayes factor categorizations[6] characterize this range as 'weak' or 'very weak'. Second, we suspect many scientists would guess that $P \approx 0.05$ implies stronger support for H_1 than a Bayes factor of 2.5 to 3.4. Third, using equation (1) and prior odds of 1:10, a *P* value of 0.05 corresponds to at least 3:1 odds (that is, the reciprocal of the product $\frac{1}{10} \times 3.4$) in favour of the null hypothesis!

Why 0.005

The choice of any particular threshold is arbitrary and involves a trade-off between type I and type II errors. We propose 0.005 for two reasons. First, a two-sided *P* value of 0.005 corresponds to Bayes factors between approximately 14 and 26 in favour of H_1. This range represents 'substantial' to 'strong' evidence according to conventional Bayes factor classifications[6].

Second, in many fields the $P < 0.005$ standard would reduce the false positive rate to levels we judge to be reasonable. If we let ϕ denote the proportion of null hypotheses that are true, $1 - \beta$ the power of tests in rejecting false null hypotheses, and α the type I error/significance threshold, then as the population of tested hypotheses becomes large, the false positive rate (that is, the proportion of true null effects among the total number of statistically significant findings) can be approximated by:

$$\text{False positive rate} \approx \frac{\alpha\phi}{\alpha\phi + (1-\beta)(1-\phi)} \qquad (2)$$

For different levels of the prior odds that there is a true effect, $\frac{1-\phi}{\phi}$, and for significance thresholds $\alpha = 0.05$ and $\alpha = 0.005$, Fig. 2 shows the false positive rate as a function of power $1-\beta$.

In many studies, statistical power is low[7]. Figure 2 demonstrates that low statistical power and $\alpha = 0.05$ combine to produce high false positive rates.

For many, the calculations illustrated by Fig. 2 may be unsettling. For example, the false positive rate is greater than 33% with prior odds of 1:10 and a *P* value threshold of 0.05, regardless of the level of statistical power. Reducing the threshold to 0.005 would reduce this minimum false positive rate to 5%. Similar reductions in false positive rates would occur over a wide range of statistical powers.

Empirical evidence from recent replication projects in psychology and experimental economics provide insights into the prior odds in favour of H_1. In both projects, the rate of replication (that is, significance at $P < 0.05$ in the replication in a consistent direction) was roughly double for initial studies with $P < 0.005$ relative to initial studies with $0.005 < P < 0.05$: 50% versus 24% for psychology[8], and 85% versus 44% for experimental economics[9]. Although based on relatively small samples of studies (93 in psychology, and 16 in experimental economics, after excluding initial studies with $P > 0.05$), these numbers are suggestive of the potential gains in reproducibility that would accrue from the new threshold of $P < 0.005$ in these fields. In biomedical research, 96% of a sample of recent papers claim statistically significant results with the $P < 0.05$ threshold[10]. However, replication rates were very low[5] for these studies, suggesting a potential for gains by adopting this new standard in these fields as well.

Potential objections

We now address the most compelling arguments against adopting this higher standard of evidence.

The false negative rate would become unacceptably high. Evidence that does not reach the new significance threshold should be treated as suggestive, and where possible further evidence should be accumulated; indeed, the combined results from several studies may be compelling even if any particular study is not. Failing to reject the null hypothesis does not mean accepting the null hypothesis. Moreover, the false negative rate will not increase if sample sizes are increased so that statistical power is held constant.

For a wide range of common statistical tests, transitioning from a *P* value threshold of $\alpha = 0.05$ to $\alpha = 0.005$ while maintaining 80% power would require an increase in sample sizes of about 70%. Such an increase means that fewer studies can be conducted using

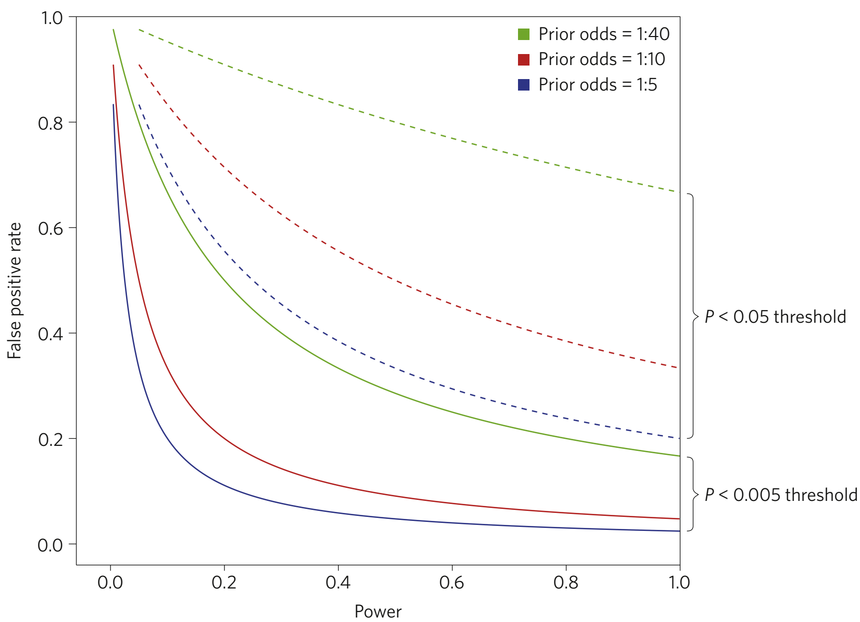

Fig. 2 | Relationship between the *P* value threshold, power, and the false positive rate. Calculated according to equation (2), with prior odds defined as $\frac{1-\phi}{\phi} = \frac{\Pr(H_1)}{\Pr(H_0)}$. For more details, see the Supplementary Information.

current experimental designs and budgets. But Fig. 2 shows the benefit: false positive rates would typically fall by factors greater than two. Hence, considerable resources would be saved by not performing future studies based on false premises. Increasing sample sizes is also desirable because studies with small sample sizes tend to yield inflated effect size estimates[11], and publication and other biases may be more likely in an environment of small studies[12]. We believe that efficiency gains would far outweigh losses.

The proposal does not address multiple-hypothesis testing, *P*-hacking, publication bias, low power, or other biases (for example, confounding, selective reporting, and measurement error), which are arguably the bigger problems. We agree. Reducing the *P* value threshold complements — but does not substitute for — solutions to these other problems, which include good study design, ex ante power calculations, pre-registration of planned analyses, replications, and transparent reporting of procedures and all statistical analyses conducted.

The appropriate threshold for statistical significance should be different for different research communities. We agree that the significance threshold selected for claiming a new discovery should depend on the prior odds that the null hypothesis is true, the number of hypotheses tested, the study design, the relative cost of type I versus type II errors, and other factors that vary by research topic. For exploratory research with very low prior odds (well outside the range in Fig. 2), even lower significance thresholds than 0.005 are needed. Recognition of this issue led the genetics research community to move to a 'genome-wide significance threshold' of 5×10^{-8} over a decade ago. And in high-energy physics, the tradition has long been to define significance by a '5-sigma' rule (roughly a *P* value threshold of 3×10^{-7}). We are essentially suggesting a move from a 2-sigma rule to a 3-sigma rule.

Our recommendation applies to disciplines with prior odds broadly in the range depicted in Fig. 2, where use of $P < 0.05$ as a default is widespread. Within those disciplines, it is helpful for consumers of research to have a consistent benchmark. We feel the default should be shifted.

Changing the significance threshold is a distraction from the real solution, which is to replace null hypothesis significance testing (and bright-line thresholds) with more focus on effect sizes and confidence intervals, treating the *P* value as a continuous measure, and/or a Bayesian method. Many of us agree that there are better approaches to statistical analyses than null hypothesis significance testing, but as yet there is no consensus regarding the appropriate choice of replacement. For example, a recent statement by the American Statistical Association addressed numerous issues regarding the misinterpretation and misuse of *P* values (as well as the related concept of statistical significance), but failed to make explicit policy recommendations to address these shortcomings[13]. Even after the significance threshold is changed, many of us will continue to advocate for alternatives to null hypothesis significance testing.

Concluding remarks

Ronald Fisher understood that the choice of 0.05 was arbitrary when he introduced it[14]. Since then, theory and empirical evidence have demonstrated that a lower threshold is needed. A much larger pool of scientists are now asking a much larger number of questions, possibly with much lower prior odds of success.

For research communities that continue to rely on null hypothesis significance testing, reducing the *P* value threshold for claims of new discoveries to 0.005 is an actionable step that will immediately improve reproducibility. We emphasize that this proposal is about standards of evidence, not standards for policy action nor standards for publication. Results that do not reach the threshold for statistical significance (whatever it is) can still be important and merit publication in leading journals if they address important research questions with rigorous methods. This proposal should not be used to reject publications of novel findings with $0.005 < P < 0.05$ properly labelled as suggestive evidence. We should reward quality and transparency of research as we impose these more stringent standards, and we should monitor how researchers' behaviours are affected by this change. Otherwise, science runs the risk that the more demanding threshold for statistical significance will be met to the detriment of quality and transparency.

Journals can help transition to the new statistical significance threshold. Authors and readers can themselves take the initiative by describing and interpreting results more appropriately in light of the new proposed definition of statistical significance. The new significance threshold will help researchers and readers to understand and communicate evidence more accurately. ❐

Daniel J. Benjamin[1]*, James O. Berger[2], Magnus Johannesson[3]*, Brian A. Nosek[4,5], E.-J. Wagenmakers[6], Richard Berk[7,10], Kenneth A. Bollen[8], Björn Brembs[9], Lawrence Brown[10], Colin Camerer[11], David Cesarini[12,13], Christopher D. Chambers[14], Merlise Clyde[2], Thomas D. Cook[15,16], Paul De Boeck[17], Zoltan Dienes[18], Anna Dreber[3], Kenny Easwaran[19], Charles Efferson[20], Ernst Fehr[21], Fiona Fidler[22], Andy P. Field[18], Malcolm Forster[23], Edward I. George[10], Richard Gonzalez[24], Steven Goodman[25], Edwin Green[26], Donald P. Green[27], Anthony G. Greenwald[28], Jarrod D. Hadfield[29], Larry V. Hedges[30], Leonhard Held[31], Teck Hua Ho[32], Herbert Hoijtink[33], Daniel J. Hruschka[34], Kosuke Imai[35], Guido Imbens[36], John P. A. Ioannidis[37], Minjeong Jeon[38], James Holland Jones[39,40], Michael Kirchler[41], David Laibson[42], John List[43], Roderick Little[44], Arthur Lupia[45], Edouard Machery[46], Scott E. Maxwell[47], Michael McCarthy[48], Don A. Moore[49], Stephen L. Morgan[50], Marcus Munafó[51,52], Shinichi Nakagawa[53], Brendan Nyhan[54], Timothy H. Parker[55], Luis Pericchi[56], Marco Perugini[57], Jeff Rouder[58], Judith Rousseau[59], Victoria Savalei[60], Felix D. Schönbrodt[61], Thomas Sellke[62], Betsy Sinclair[63], Dustin Tingley[64], Trisha Van Zandt[65], Simine Vazire[66], Duncan J. Watts[67], Christopher Winship[68], Robert L. Wolpert[2], Yu Xie[69], Cristobal Young[70], Jonathan Zinman[71] and Valen E. Johnson[72]*

[1]Center for Economic and Social Research and Department of Economics, University of Southern California, Los Angeles, CA 90089-3332, USA. [2]Department of Statistical Science, Duke University, Durham, NC 27708-0251, USA. [3]Department of Economics, Stockholm School of Economics, Stockholm SE-113 83, Sweden. [4]University of Virginia, Charlottesville, VA 22908, USA. [5]Center for Open Science, Charlottesville, VA 22903, USA. [6]Department of Psychology, University of Amsterdam, Amsterdam 1018 VZ, The Netherlands. [7]School of Arts and Sciences and Department of Criminology, University of Pennsylvania, Philadelphia, PA 19104-6286, USA. [8]Department of Psychology and Neuroscience, Department of Sociology, University of North Carolina Chapel Hill, Chapel Hill, NC 27599-3270, USA. [9]Institute of Zoology — Neurogenetics, Universität Regensburg, Universitätsstrasse 31, 93040 Regensburg, Germany. [10]Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104, USA. [11]Division of the Humanities and Social Sciences, California Institute of Technology, Pasadena, CA 91125, USA. [12]Department of Economics, New York University, New York, NY 10012, USA. [13]The Research Institute of Industrial Economics (IFN), Stockholm SE-102 15, Sweden. [14]Cardiff University Brain Research Imaging Centre (CUBRIC), Cardiff CF24 4HQ, UK. [15]Northwestern University, Evanston, IL 60208, USA. [16]Mathematica Policy Research, Washington, DC 20002-4221, USA. [17]Department of Psychology, Quantitative Program, Ohio State University, Columbus, OH 43210, USA. [18]School of Psychology, University of Sussex, Brighton BN1 9QH, UK. [19]Department of Philosophy, Texas A&M University, College Station, TX 77843-4237, USA. [20]Department of Psychology, Royal Holloway University of London, Egham Surrey TW20 0EX, UK. [21]Department of Economics, University of Zurich, 8006 Zurich, Switzerland. [22]School of BioSciences and School of Historical & Philosophical Studies, University of Melbourne, Parkville, VIC 3010, Australia. [23]Department of Philosophy, University of Wisconsin — Madison, Madison, WI 53706, USA. [24]Department of Psychology, University of Michigan, Ann Arbor, MI 48109-1043, USA. [25]Stanford University, General Medical Disciplines, Stanford, CA 94305, USA. [26]Department of Ecology, Evolution and Natural Resources SEBS, Rutgers University, New Brunswick, NJ 08901-8551, USA. [27]Department of Political Science, Columbia University in the City of New York, New York, NY 10027, USA. [28]Department of Psychology, University of Washington, Seattle, WA 98195-1525, USA. [29]Institute of Evolutionary Biology School of Biological Sciences, The University of Edinburgh, Edinburgh EH9 3JT, UK. [30]Weinberg College of Arts & Sciences Department of Statistics, Northwestern University, Evanston, IL 60208, USA. [31]Epidemiology, Biostatistics and Prevention Institute (EBPI), University of Zurich, 8001 Zurich, Switzerland. [32]National University of Singapore, Singapore 119077, Singapore. [33]Department of Methods and Statistics, Universiteit Utrecht, Utrecht 3584 CH, The Netherlands. [34]School of Human Evolution and Social Change, Arizona State University, Tempe, AZ 85287-2402, USA. [35]Department of Politics and Center for Statistics and Machine Learning, Princeton University, Princeton, NJ 08544, USA. [36]Stanford University, Stanford, CA 94305-5015, USA. [37]Departments of Medicine, of Health Research and Policy, of Biomedical Data Science, and of Statistics and Meta-Research Innovation Center at Stanford (METRICS), Stanford University, Stanford, CA 94305, USA. [38]Advanced Quantitative Methods, Social Research Methodology, Department of Education, Graduate School of Education & Information Studies, University of California, Los Angeles, CA 90095-1521, USA. [39]Department of Life Sciences, Imperial College London, Ascot SL5 7PY, UK. [40]Department of Earth System Science, Stanford, CA 94305-4216, USA. [41]Department of Banking and Finance, University of Innsbruck and University of Gothenburg, Innsbruck A-6020, Austria. [42]Department of Economics, Harvard University, Cambridge, MA 02138, USA. [43]Department of Economics, University of Chicago, Chicago, IL 60637, USA. [44]Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109-2029, USA. [45]Department of Political Science, University of Michigan, Ann Arbor, MI 48109-1045, USA. [46]Department of History and Philosophy of Science, University of Pittsburgh, Pittsburgh, PA 15260, USA. [47]Department of Psychology, University of Notre Dame, Notre Dame, IN 46556, USA. [48]School of BioSciences, University of Melbourne, Parkville, VIC 3010, Australia. [49]Haas School of Business, University of California at Berkeley, Berkeley, CA 94720-1900A, USA. [50]Johns Hopkins University, Baltimore, MD 21218, USA. [51]MRC Integrative Epidemiology Unit, University of Bristol, Bristol BS8 1TU, UK. [52]UK Centre for Tobacco and Alcohol Studies, School of Experimental Psychology, University of Bristol, Bristol BS8 1TU, UK. [53]Evolution & Ecology Research Centre and School of Biological, Earth and Environmental Sciences, University of New South Wales, Sydney, NSW 2052, Australia. [54]Department of Government, Dartmouth College, Hanover, NH 03755, USA. [55]Department of Biology, Whitman College, Walla Walla, WA 99362, USA. [56]Department of Mathematics, University of Puerto Rico, Rio Piedras Campus, San Juan, PR 00936-8377, Puerto Rico. [57]Department of Psychology, University of Milan-Bicocca, Milan 20126, Italy. [58]Department of Cognitive Sciences, University of California, Irvine, CA 92617, USA. [59]Université Paris Dauphine, 75016, Paris, France. [60]Department of Psychology, The University of British Columbia, Vancouver V6T 1Z4 BC, Canada. [61]Department Psychology, Ludwig-Maximilians-University Munich, Leopoldstraße 13, 80802 Munich, Germany. [62]Department of Statistics, Purdue University, West Lafayette, IN 47907-2067, USA. [63]Department of Political Science, Washington University in St. Louis, St. Louis, MO 63130-4899, USA. [64]Government Department, Harvard University, Cambridge, MA 02138, USA. [65]Department of Psychology, Ohio State University, Columbus, OH 43210, USA. [66]Department of Psychology, University of California, Davis, CA 95616, USA. [67]Microsoft Research, 641 Avenue of the Americas, 7th Floor, New York, NY 10011, USA. [68]Department of Sociology, Harvard University, Cambridge, MA 02138, USA. [69]Department of Sociology, Princeton University, Princeton, NJ 08544, USA. [70]Department of Sociology, Stanford University, Stanford, CA 94305-2047, USA. [71]Department of Economics, Dartmouth College, Hanover, NH 03755-3514, USA. [72]Department of Statistics, Texas A&M University, College Station, TX 77843, USA.

**e-mail: daniel.benjamin@gmail.com; magnus.johannesson@hhs.se; vejohnson@exchange.tamu.edu*

Published online: 1 September 2017
DOI: 10.1038/s41562-017-0189-z

References

1. Greenwald, A. G. et al. *Psychophysiology* **33**, 175–183 (1996).
2. Johnson, V. E. *Proc. Natl Acad. Sci. USA* **110**, 19313–19317 (2013).
3. Dreber, A. et al. *Proc. Natl Acad. Sci. USA* **112**, 15343–15347 (2015).
4. Johnson, V. E. et al. *J. Am. Stat. Assoc.* **112**, 1–10 (2016).
5. Begley, C. G. & Ioannidis, J. P. A. *Circ. Res.* **116**, 116–126 (2015).
6. Kass, R. E. & Raftery, A. E. *J. Am. Stat. Assoc.* **90**, 773–795 (1995).
7. Szucs, D. & Ioannidis, J. P. A. *PLoS Biol.* **15**, e2000797 (2017).
8. Open Science Collaboration. *Science* **349**, aac4716 (2015).
9. Camerer, C. F. et al. *Science* **351**, 1433–1436 (2016).
10. Chavalarias, D. et al. *JAMA* **315**, 1141–1148 (2016).
11. Gelman, A. & Carlin, J. *Perspect. Psychol. Sci.* **9**, 641–651 (2014).
12. Fanelli, D., Costas, R. & Ioannidis, J. P. A. *Proc. Natl Acad. Sci. USA* **114**, 3714–3719 (2017).
13. Wasserstein, R. L. & Lazar, N. A. *Am. Stat.* **70**, 129–133 (2016).
14. Fisher, R. A. *Statistical Methods for Research Workers* (Oliver & Boyd, Edinburgh, 1925).

15. Sellke, T., Bayarri, M. J. & Berger, J. O. *Am. Stat.* **55**, 62–71 (2001).

Acknowledgements
We thank D. L. Lormand, R. Royer and A. T. Nguyen Viet for excellent research assistance.

Competing interests
One of the 72 authors, Christopher Chambers, is a member of the Advisory Board of *Nature Human Behaviour*. Christopher Chambers was not a corresponding author and did not communicate with the editors regarding the publication of this article. The other authors declare no competing interests.

Additional information
Supplementary information is available for this paper at doi:10.1038/s41562-017-0189-z.

Supplementary Information:

Supplementary Text

R code used to generate Figures 1 and 2

Supplementary Information:

Supplementary Text

<u>Figure 1</u>

All four curves in Figure 1 describe the relationship between (i) a *P*-value based on a two-sided normal test and (ii) a Bayes factor or a bound on a Bayes factor. The *P*-values are based on a two-sided test that the mean μ of an independent and identically distributed sample of normally distributed random variables is 0. The variance of the observations is known. Without loss of generality, we assume that the variance is 1 and the sample size is also 1. The curves in the figure differ according to the alternative hypotheses that they assume for calculating (ii).

Because these curves involve two-sided tests, all alternative hypotheses are restricted to be symmetric around 0. That is, the density assumed for the value of μ under the alternative hypothesis is always assumed to satisfy $f(\mu) = f(-\mu)$.

The curve labeled "Power" corresponds to defining the alternative hypothesis so that power is 75% in a two-sided 5% test. This is achieved by assuming that μ under the alternative hypothesis is equal to $\pm(z_{0.025} + z_{0.75}) = \pm 2.63$. That is, the alternative hypothesis places ½ its prior mass on 2.63 and ½ its mass on -2.63.

The curve labeled UMPBT corresponds to the uniformly most powerful Bayesian test (*2*) that corresponds to a classical, two-sided test of size $\alpha = 0.005$. The alternative hypothesis for this Bayesian test places ½ mass at 2.81 and ½ mass at -2.81. The null hypothesis for this test is rejected if the Bayes factor exceeds 25.7. Note that this curve is nearly identical to the "Power" curve if that curve had been defined using 80% power, rather than 75% power. The Power curve for 80% power would place ½ its mass at ± 2.80.

The Likelihood Ratio Bound curve represents an approximate upper bound on the Bayes factor obtained by defining the alternative hypothesis as putting ½ its mass on $\pm\bar{x}$, where $\bar{x}$ is the observed sample mean. Over the range of *P*-values displayed in the figure, this alternative hypothesis very closely approximates the maximum Bayes factor that can be attained from among the set of alternative hypotheses constrained to be of the form $0.5 \times [f(\mu) + f(-\mu)]$ for some density function *f*.

The Local-H_1 curve is described fully in the figure caption. A fuller explanation and discussion of this bound can be found in ref. *15*.

<u>Equation 2 and Figure 2</u>

This equation defines the large-sample relationship between the false positive rate, power $1 - \beta$, type I error rate α, and the probability that the null hypothesis is true when a large number of independent experiments have been conducted. More specifically, suppose that n independent hypothesis tests are conducted, and suppose that in each test the probability that the null hypothesis is true is ϕ. If the null hypothesis is true, assume that the probability that it is falsely rejected (i.e., a false positive occurs) is α. For the test $j = 1, \ldots, n$, define the random variable $X_j = 1$ if the null hypothesis is true *and* the null hypothesis is rejected, and $X_j = 0$ if either the alternative hypothesis is true or the null hypothesis is not rejected. Note that the X_j are independent Bernoulli random variables with $\Pr(X_j = 1) = \alpha\phi$. Also for test *j*, define another random variable $Y_j = 1$ if the alternative hypothesis is true *and* the null hypothesis is rejected, and 0 otherwise. It follows that the Y_j are independent Bernoulli random variables with $\Pr(Y_j = 1) = (1 - \phi)(1 - \beta)$. Note that Y_j is independent of Y_k for $j \neq k$, but Y_j is not independent of X_j. For the *n* experiments, the false positive rate can then be written as:

$$FPR = \frac{\sum_{j=1}^{n} X_j}{\sum_{j=1}^{n} X_j + \sum_{j=1}^{n} Y_j} = \frac{\sum_{j=1}^{n} X_j/n}{\sum_{j=1}^{n} X_j/n + \sum_{j=1}^{n} Y_j/n}.$$

By the strong law of large numbers, $\sum_{j=1}^{n} X_j/n$ converges almost surely to $\alpha\phi$, and $\sum_{j=1}^{n} Y_j/n$ converges almost surely to $(1 - \phi)(1 - \beta)$. Application of the continuous mapping theorem yields

$$FPR \xrightarrow{\text{a.s.}} \frac{\alpha\phi}{\alpha\phi + (1 - \phi)(1 - \beta)}.$$

Figure 2 illustrates this relationship for various values of α and prior odds for the alternative, $\frac{1-\phi}{\phi}$.

R code used to generate Figure 1:

```
type1=.005
type1Power=0.05
type2=0.25
p=1-c(9000:9990)/10000
xbar = qnorm(1-p/2)

# alternative based on 80% POWER IN 5% TEST
muPower = qnorm(1-type2)+qnorm(1-type1Power/2)
bfPow = 0.5*(dnorm(xbar,muPower,1)+dnorm(xbar,-muPower,1))/dnorm(xbar,0,1)

muUMPBT = qnorm(0.9975)
bfUMPBT = 0.5*(dnorm(xbar,muUMPBT,1)+dnorm(xbar,-
muUMPBT,1))/dnorm(xbar,0,1)

# two-sided "LR" bound
bfLR = 0.5/exp(-0.5*xbar^2)

bfLocal = -1/(2.71*p*log(p))

#coordinates for dashed lines
data = data.frame(p,bfLocal,bfLR,bfPow,bfUMPBT)
U_005 = max(data$bfLR[data$p=="0.005"])
L_005 = min(data$bfLocal[data$p=="0.005"])
U_05 = max(data$bfLR[data$p=="0.05"])
L_05 = min(data$bfUMPBT[data$p=="0.05"])

# Local bound; no need for two-sided adjustment


#plot margins
par(mai=c(0.8,0.8,.1,0.4))
par(mgp=c(2,1,0))

matplot(p,cbind(bfLR,-1/(2.71*p*log(p))),type='n',log='xy',
        xlab=expression(paste(italic(P) ,"-value")),
        ylab="Bayes Factor",
        ylim = c(0.3,100),
        bty="n",xaxt="n",yaxt="n")
lines(p,bfPow,col="red",lwd=2.5)
lines(p,bfLR,col="black",lwd=2.5)
lines(p,bfUMPBT,col="blue",lwd=2.5)
lines(p,bfLocal,col="green",lwd=2.5)
legend(0.015,100,c(expression(paste("Power")),"Likelihood Ratio
Bound","UMPBT",expression(paste("Local-",italic(H)[1],"
Bound"))),lty=c(1,1,1,1),
       lwd=c(2.5,2.5,2.5,2.5),col=c("red","black","blue","green"), cex =
0.8)
#text(0.062,65, "\u03B1", font =3, cex = 0.9)

#customizing axes
#x axis
axis(side=1,at=c(-2,0.001,0.0025,0.005,0.010,0.025,0.050,0.100,0.14),
```

```
     labels =
c("","0.0010","0.0025","0.0050","0.0100","0.0250","0.0500","0.1000",""),lw
d=1,
     tck = -0.01, padj = -1.1, cex.axis = .8)
#y axis on the left - main
axis(side=2,at=c(-0.2, 0.3,0.5,1,2,5,10,20,50,100),labels =
c("","0.3","0.5","1.0","2.0","5.0","10.0","20.0","50.0","100.0"),lwd=1,las
= 1,
     tck = -0.01, hadj = 0.6, cex.axis = .8)
#y axis on the left - secondary (red labels)
axis(side=2,at=c(L_005,U_005),labels = c(13.9,25.7),lwd=1,las= 1,
     tck = -0.01, hadj = 0.6, cex.axis = .6,col.axis="red")
#y axis on the right - main
axis(side=4,at=c(-0.2, 0.3,0.5,1,2,5,10,20,50,100),labels =
c("","0.3","0.5","1.0","2.0","5.0","10.0","20.0","50.0","100.0"),lwd=1,las
= 1,
     tck = -0.01, hadj = 0.4, cex.axis = .8)
#y axis on the right - secondary (red labels)
axis(side=4,at=c(L_05,U_05),labels = c(2.4,3.4),lwd=1,las= 1,
     tck = -0.01, hadj = 0.4, cex.axis = .6,col.axis="red")

###dashed lines
segments(x0 = 0.000011, y0= U_005, x1 = 0.005, y1 = U_005, col = "gray40",
lty = 2)
segments(x0 = 0.000011, y0= L_005, x1 = 0.005, y1 = L_005, col = "gray40",
lty = 2)
segments(x0 = 0.005, y0= 0.00000001, x1 = 0.005, y1 = U_005, col =
"gray40", lty = 2)

segments(x0 = 0.05, y0= U_05, x1 = 0.14, y1 = U_05, col = "gray40", lty =
2)
segments(x0 = 0.05, y0= L_05, x1 = 0.14, y1 = L_05, col = "gray40", lty =
2)
segments(x0 = 0.05, y0= 0.00000001, x1 = 0.05, y1 = U_05, col = "gray40",
lty = 2)
```

R code used to generate Figure 2:

```
pow1=c(5:999)/1000   # power range for 0.005 tests
pow2=c(50:999)/1000  # power range for 0.05 tests
alpha=0.005 # test size
pi0=5/6  # prior probability
N=10^6  # doesn't matter


#graph margins
par(mai=c(0.8,0.8,0.1,0.1))
par(mgp=c(2,1,0))


plot(pow1,alpha*N*pi0/(alpha*N*pi0+pow1*(1-pi0)*N),type='n',ylim = c(0,1),
xlim = c(0,1.5),
     xlab='Power                                      ',
     ylab='False positive rate', bty="n", xaxt="n", yaxt="n")
#grid lines
segments(x0 = -0.058, y0 = 0, x1 = 1, y1 = 0,lty=1,col = "gray92")
segments(x0 = -0.058, y0 = 0.2, x1 = 1, y1 = 0.2,lty=1,col = "gray92")
segments(x0 = -0.058, y0 = 0.4, x1 = 1, y1 = 0.4,lty=1,col = "gray92")
segments(x0 = -0.058, y0 = 0.6, x1 = 1, y1 = 0.6,lty=1,col = "gray92")
segments(x0 = -0.058, y0 = 0.8, x1 = 1, y1 = 0.8,lty=1,col = "gray92")
segments(x0 = -0.058, y0 = 1, x1 = 1, y1 = 1,lty=1,col = "gray92")


lines(pow1,alpha*N*pi0/(alpha*N*pi0+pow1*(1-
pi0)*N),lty=1,col="blue",lwd=2)
odd_1_5_1 = alpha*N*pi0/(alpha*N*pi0+pow1[995]*(1-pi0)*N)
alpha=0.05
pi0=5/6
lines(pow2,alpha*N*pi0/(alpha*N*pi0+pow2*(1-
pi0)*N),lty=2,col="blue",lwd=2)
odd_1_5_2 = alpha*N*pi0/(alpha*N*pi0+pow2[950]*(1-pi0)*N)


alpha=0.05
pi0=10/11
lines(pow2,alpha*N*pi0/(alpha*N*pi0+pow2*(1-pi0)*N),lty=2,col="red",lwd=2)
odd_1_10_2 = alpha*N*pi0/(alpha*N*pi0+pow2[950]*(1-pi0)*N)
alpha=0.005
pi0=10/11
lines(pow1,alpha*N*pi0/(alpha*N*pi0+pow1*(1-pi0)*N),lty=1,col="red",lwd=2)
odd_1_10_1 = alpha*N*pi0/(alpha*N*pi0+pow1[995]*(1-pi0)*N)

alpha=0.05
pi0=40/41
lines(pow2,alpha*N*pi0/(alpha*N*pi0+pow2*(1-
pi0)*N),lty=2,col="green",lwd=2)
odd_1_40_2 = alpha*N*pi0/(alpha*N*pi0+pow2[950]*(1-pi0)*N)
alpha=0.005
pi0=40/41
```

```
lines(pow1,alpha*N*pi0/(alpha*N*pi0+pow1*(1-
pi0)*N),lty=1,col="green",lwd=2)
odd_1_40_1 = alpha*N*pi0/(alpha*N*pi0+pow1[995]*(1-pi0)*N)




#customizing axes
axis(side=2,at=c(-0.5,0,0.2,0.4,0.6,0.8,1.0),labels =
c("","0.0","0.2","0.4","0.6","0.8","1.0"),
     lwd=1,las= 1,tck = -0.01, hadj = 0.4, cex.axis = .8)
axis(side=1,at=c(-0.5,0,0.2,0.4,0.6,0.8,1.0),labels =
c("","0.0","0.2","0.4","0.6","0.8","1.0"),
     lwd=1,las= 1, tck = -0.01, padj = -1.1, cex.axis = .8)



legend(1.05,1,c("Prior odds = 1:40","Prior odds = 1:10","Prior odds =
1:5"),pch=c(15,15,15),
       col=c("green","red","blue"), cex = 1)


###############  Use these commands to add brackets in Figure 2

library(pBrackets)


#add text and brackets
text(1.11,(odd_1_5_2+odd_1_40_2)/2, expression(paste(italic(P)," < 0.05
threshold")), cex = 0.9,adj=0)
text(1.11,(odd_1_5_1+odd_1_40_1)/2, expression(paste(italic(P)," < 0.005
threshold")), cex = 0.9,adj=0)
brackets(1.03, odd_1_40_1, 1.03, odd_1_5_1, h = NULL, ticks = 0.5,
curvature = 0.7, type = 1,
        col = 1, lwd = 1, lty = 1, xpd = FALSE)
brackets(1.03, odd_1_40_2, 1.03, odd_1_5_2, h = NULL, ticks = 0.5,
curvature = 0.7, type = 1,
         col = 1, lwd = 1, lty = 1, xpd = FALSE)
```