Revised standards for statistical evidence

Valen E. Johnson[1]

Department of Statistics, Texas A&M University, College Station, TX 77843-3143

Edited by Adrian E. Raftery, University of Washington, Seattle, WA, and approved October 9, 2013 (received for review July 18, 2013)

Recent advances in Bayesian hypothesis testing have led to the development of uniformly most powerful Bayesian tests, which represent an objective, default class of Bayesian hypothesis tests that have the same rejection regions as classical significance tests. Based on the correspondence between these two classes of tests, it is possible to equate the size of classical hypothesis tests with evidence thresholds in Bayesian tests, and to equate *P* values with Bayes factors. An examination of these connections suggest that recent concerns over the lack of reproducibility of scientific studies can be attributed largely to the conduct of significance tests at unjustifiably high levels of significance. To correct this problem, evidence thresholds required for the declaration of a significant finding should be increased to 25–50:1, and to 100–200:1 for the declaration of a highly significant finding. In terms of classical hypothesis tests, these evidence standards mandate the conduct of tests at the 0.005 or 0.001 level of significance.

Reproducibility of scientific research is critical to the scientific endeavor, so the apparent lack of reproducibility threatens the credibility of the scientific enterprise (e.g., refs. 1 and 2). Unfortunately, concern over the nonreproducibility of scientific studies has become so pervasive that a Web site, *Retraction Watch*, has been established to monitor the large number of retracted papers, and methodology for detecting flawed studies has developed nearly into a scientific discipline of its own (e.g., refs. 3–9).

Nonreproducibility in scientific studies can be attributed to a number of factors, including poor research designs, flawed statistical analyses, and scientific misconduct. The focus of this article, however, is the resolution of that component of the problem that can be attributed simply to the routine use of widely accepted statistical testing procedures.

Claims of novel research findings are generally based on the outcomes of statistical hypothesis tests, which are normally conducted under one of two statistical paradigms. Most commonly, hypothesis tests are performed under the classical, or frequentist, paradigm. In this approach, a "significant" finding is declared when the value of a test statistic exceeds a specified threshold. Values of the test statistic above this threshold define the test's rejection region. The significance level α of the test is defined to be the maximum probability that the test statistic falls into the rejection region when the null hypothesis—representing standard theory—is true. By long-standing convention (10), a value of $\alpha = 0.05$ defines a significant finding. The *P* value from a classical test is the maximum probability of observing a test statistic as extreme, or more extreme, than the value that was actually observed, given that the null hypothesis is true.

The second approach for performing hypothesis tests follows from the Bayesian paradigm and focuses on the calculation of the posterior odds that the alternative hypotheses is true, given the observed data and any available prior information (e.g., refs. 11 and 12). From Bayes theorem, the posterior odds in favor of the alternative hypothesis equals the prior odds assigned in favor of the alternative hypotheses, multiplied by the Bayes factor. In the case of simple null and alternative hypotheses, the Bayes factor represents the ratio of the sampling density of the data evaluated under the alternative hypothesis to the sampling density of the data evaluated under the null hypothesis. That is, it represents the relative probability assigned to the data by the two hypotheses. For composite hypotheses, the Bayes factor represents the ratio of the average value of the sampling density of the observed data under each of the two hypotheses, averaged with respect to the prior density specified on the unknown parameters under each hypothesis.

Paradoxically, the two approaches toward hypothesis testing often produce results that are seemingly incompatible (13–15). For instance, many statisticians have noted that *P* values of 0.05 may correspond to Bayes factors that only favor the alternative hypothesis by odds of 3 or 4–1 (13–15). This apparent discrepancy stems from the fact that the two paradigms for hypothesis testing are based on the calculation of different probabilities: *P* values and significance tests are based on calculating the probability of observing test statistics that are as extreme or more extreme than the test statistic actually observed, whereas Bayes factors represent the relative probability assigned to the observed data under each of the competing hypotheses. The latter comparison is perhaps more natural because it relates directly to the posterior probability that each hypothesis is true. However, defining a Bayes factor requires the specification of both a null hypothesis and an alternative hypothesis, and in many circumstances there is no objective mechanism for defining an alternative hypothesis. The definition of the alternative hypothesis therefore involves an element of subjectivity, and it is for this reason that scientists generally eschew the Bayesian approach toward hypothesis testing. Efforts to remove this hurdle continue, however, and recent studies of the use of Bayes factors in the social sciences include refs. 16–20.

Recently, Johnson (21) proposed a new method for specifying alternative hypotheses. When used to test simple null hypotheses in common testing scenarios, this method produces default Bayesian procedures that are uniformly most powerful in the sense that they maximize the probability that the Bayes factor in favor of the alternative hypothesis exceeds a specified threshold. A critical feature of these Bayesian tests is that their rejection regions can be matched exactly to the rejection regions of classical hypothesis tests. This correspondence is important because it provides a direct connection between significance levels, *P* values, and Bayes factors, thus making it possible to objectively

Significance

The lack of reproducibility of scientific research undermines public confidence in science and leads to the misuse of resources when researchers attempt to replicate and extend fallacious research findings. Using recent developments in Bayesian hypothesis testing, a root cause of nonreproducibility is traced to the conduct of significance tests at inappropriately high levels of significance. Modifications of common standards of evidence are proposed to reduce the rate of nonreproducibility of scientific research by a factor of 5 or greater.

Author contributions: V.E.J. designed research, performed research, contributed new reagents/analytic tools, analyzed data, and wrote the paper.

The author declares no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

[1]E-mail: vjohnson@stat.tamu.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1313476110/-/DCSupplemental.

STATISTICS

examine the strength of evidence provided against a null hypothesis as a function of a P value or significance level.

Results

Let $f(\mathbf{x}|\boldsymbol{\theta})$ denote the sampling density of the data $\mathbf{x}$ under both the null (H_0) and alternative (H_1) hypotheses. For $i = 0, 1$, let $\pi_i(\boldsymbol{\theta})$ denote the prior density assigned to the unknown parameter $\boldsymbol{\theta}$ belonging to Θ under hypothesis H_i, let $P(H_i)$ denote the prior probability assigned to hypothesis H_i, and let $m_i(\mathbf{x})$ denote the marginal density of the data under hypothesis H_i, i.e.,

$$m_i(\mathbf{x}) = \int_{\Theta} f(\mathbf{x}|\boldsymbol{\theta})\pi_i(\boldsymbol{\theta})d\boldsymbol{\theta}. \quad [1]$$

The Bayes factor in favor of the alternative hypothesis is defined as $BF_{10}(\mathbf{x}) = m_1(\mathbf{x})/m_0(\mathbf{x})$.

A condition of equipoise is said to apply if $p(H_0) = p(H_1) = 0.5$. It is assumed that no subjectivity is involved in the specification of the null hypothesis. Under these assumptions, a uniformly most powerful Bayesian test (UMPBT) for evidence threshold γ, denoted by UMPBT(γ), may be defined as follows (21).

Definition. A UMPBT for evidence threshold $\gamma > 0$ in favor of the alternative hypothesis H_1 against a fixed null hypothesis H_0 is a Bayesian hypothesis test in which the Bayes factor for the test satisfies the following inequality for any $\boldsymbol{\theta}_t \in \Theta$ and for all alternative hypotheses $H_{1'} : \boldsymbol{\theta} \sim \pi_{1'}(\boldsymbol{\theta})$:

$$\mathbf{P}_{\boldsymbol{\theta}_t}[BF_{10}(\mathbf{x}) > \gamma] \geq \mathbf{P}_{\boldsymbol{\theta}_t}[BF_{1'0}(\mathbf{x}) > \gamma]. \quad [2]$$

That is, the UMPBT(γ) is a Bayesian test in which the alternative hypothesis is specified so as to maximize the probability that the Bayes factor $BF_{10}(\mathbf{x})$ exceeds the evidence threshold γ for all possible values of the data generating parameter $\boldsymbol{\theta}_t$.

Under mild regularity conditions, Johnson (21) demonstrated that UMPBTs exist for testing the values of parameters in one-parameter exponential family models. Such tests include tests of a normal mean (with known variance) and a binomial proportion. In *SI Text*, UMPBTs are derived for tests of the difference of normal means, and for testing whether the noncentrality parameter of a χ^2 random variable on one degree of freedom is equal to 0. The form of alternative hypotheses, Bayes factors, rejection regions, and the relationship between evidence thresholds and sizes of equivalent frequentist tests are provided in Table S1.

The construction of UMPBTs is perhaps most easily illustrated in a z test for the mean μ of a random sample of normal observations with known variance σ^2. From Table S1, a one-sided UMPBT of the null hypothesis $H_0 : \mu = 0$ against alternatives that specify that $\mu > 0$ is obtained by specifying the alternative hypothesis to be

$$H_1 : \mu_1 = \sigma\sqrt{\frac{2\log(\gamma)}{n}}.$$

For $z = \sqrt{n}\bar{x}/\sigma$, the Bayes factor for this test is

$$BF_{10}(z) = \exp\left[z\sqrt{2\log(\gamma)} - \log(\gamma)\right].$$

By setting the evidence threshold $\gamma = 3.87$, the rejection region of the resulting test exactly matches the rejection region of a one-sided 5% significance test. That is, the Bayes factor for this test exceeds 3.87 whenever the sample mean of the data, $\bar{x}$, exceeds $1.645\sigma/\sqrt{n}$, the rejection region for a classical one-sided 5% test. If $\bar{x} = 1.645\sigma/\sqrt{n}$, then the UMPBT produces a Bayes factor that achieves the bounds described in ref. 13. Conversely if $\bar{x} = 0$, the Bayes factor in favor of the alternative hypothesis is $1/3.87 = 0.258$, which illustrates that UMPBTs—unlike P values—provide evidence in favor of both true null and true alternative hypotheses.

This example highlights several properties of UMPBTs. First, the prior densities that define one-sided UMPBT alternatives concentrate their mass on a single point in the parameter space. Second, the distance between the null parameter value and the alternative parameter value is typically $O(n^{-1/2})$, which means that UMPBTs share certain large sample properties with classical hypothesis tests. The implications of these properties are discussed further in *SI Text* and in ref. 21.

Unfortunately, UMPBTs do not exist for testing a normal mean or difference in means when the observational variance σ^2 is not known. However, if σ^2 is unknown and an inverse gamma prior distribution is imposed, then the probability that the Bayes factor exceeds the evidence threshold γ in a one-sample test can be expressed as

$$\mathbf{P}[BF_{10} > \gamma] = \mathbf{P}[a_n < \bar{x} < b_n], \quad [3]$$

and in a two-sample test as

$$\mathbf{P}[BF_{10} > \gamma] = \mathbf{P}[a_n < \bar{x}_2 - \bar{x}_1 < b_n]. \quad [4]$$

In these expressions, a_n and b_n are functions of the evidence threshold γ, the population means, and a statistic that is ancillary to both. Furthermore, $b_n \to \infty$ as the sample size n becomes large. For sufficiently large n, approximate, data-dependent UMPBTs can thus be obtained by determining the values of the population means that minimize a_n, because minimizing a_n maximizes the probability that the sample mean or difference in sample means will exceed a_n, regardless of the distribution of the sample means. The resulting approximate UMPBT tests are useful for examining the connection between Bayesian evidence thresholds and significance levels in classical t tests. Expressions for the values of the population means that minimize a_n for t tests are provided in Table S1.

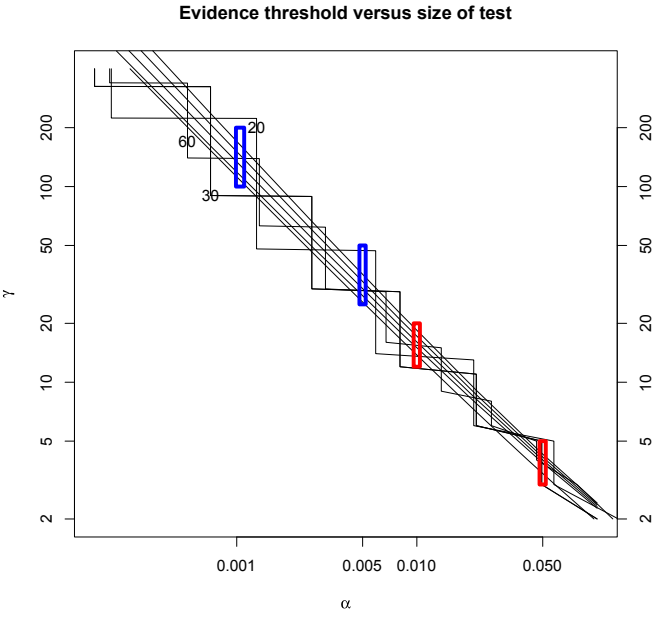

Fig. 1. Evidence thresholds and size of corresponding significance tests. The UMPBT and significance tests used to construct this plot have the same (z, χ^2, and binomial tests) or approximately the same (t tests) rejection regions. The smooth curves represent, from *Top* to *Bottom*, t tests based on 20, 30, and 60 degrees of freedom, the z test, and the χ^2 test on 1 degree of freedom. The discontinuous curves reflect the correspondence between tests of a binomial proportion based on 20, 30, or 60 observations when the null hypothesis is $p_0 = 0.5$.

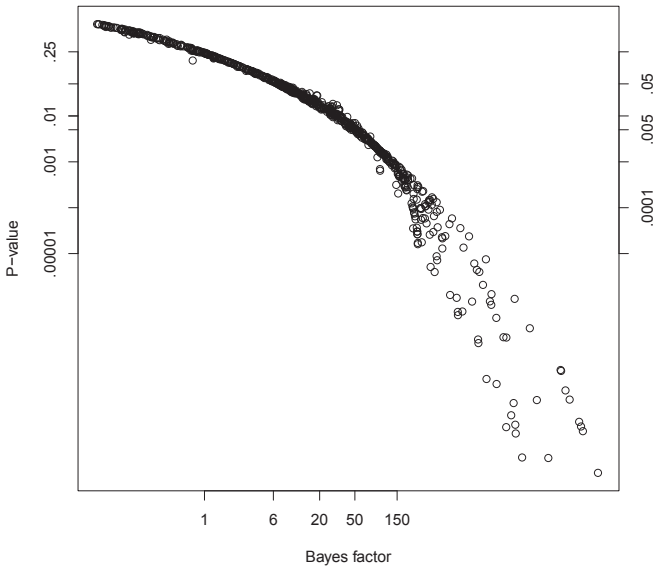

Fig. 2. *P* values versus UMPBT Bayes factors. This plot depicts approximate Bayes factors derived from 765 *t* statistics reported by Wetzels et al. (20). A breakdown of the curvilinear relationship between Bayes factors and *P* values occurs in the lower right portion of the plot, which corresponds to *t* statistics that produce Bayes factors that are near their maximum value.

Because UMBPTs can be used to define Bayesian tests that have the same rejection regions as classical significance tests, "a Bayesian using a UMPBT and a frequentist conducting a significance test will make identical decisions on the basis of the observed data. That is, a decision to reject the null hypothesis at a specified significance level occurs only when the Bayes factor in favor of the alternative hypothesis exceeds a specified evidence threshold" (21). The close connection between UMPBTs and significance tests thus provides insight into the amount of evidence required to reject a null hypothesis.

To illustrate this connection, curves of the values of the test sizes (α) and evidence thresholds (γ) that produce matching rejection regions for a variety of standard tests have been plotted in Fig. 1. Included among these are *z* tests, χ^2 tests, *t* tests, and tests of a binomial proportion.

The two red boxes in Fig. 1 highlight the correspondence between significance tests conducted at the 5% and 1% levels of significance and evidence thresholds. As this plot shows, the Bayesian evidence thresholds that correspond to these tests are quite modest. Evidence thresholds that correspond to 5% tests range between 3 and 5. This range of evidence falls at the lower end of the range that Jeffreys (11) calls "substantial evidence," or what Kass and Raftery (12) term "positive evidence." Evidence thresholds for 1% tests range between 12 and 20, which fall at the lower end of Jeffreys' "strong-evidence" category, or the upper end of Kass and Raftery's positive-evidence category. If equipoise applies, the posterior probabilities assigned to null hypotheses range from ~0.17 to 0.25 for null hypotheses that are rejected at the 0.05 level of significance, and from about 0.05 to 0.08 for nulls that are rejected at the 0.01 level of significance.

The two blue boxes in Fig. 1 depict the range of evidence thresholds that correspond to significance tests conducted at the 0.005 and 0.001 levels of significance. Bayes factors in the range of 25–50 are required to obtain tests that have rejection regions that correspond to 0.005 level tests, whereas Bayes factors between ~100 and 200 correspond to 0.001 level tests. In Jeffreys' scheme (11), Bayes factors in the range 25–50 are considered "strong" evidence in favor of the alternative, and Bayes factors in the range 100–200 are considered "decisive." Kass and Raftery (12) consider Bayes factors between 20 and 150 as "strong" evidence, and Bayes factors above 150 to be "very strong" evidence. Thus, according to standard scales of evidence, these levels of significance represent either strong, very strong, or decisive levels of evidence. If equipoise applies, then the corresponding posterior probabilities assigned to null hypotheses range from ~0.02 to 0.04 for null hypotheses that are rejected at the 0.005 level of significance, and from about 0.005 to 0.01 for null hypotheses that are rejected at the 0.001 level of significance.

The correspondence between significance levels and evidence thresholds summarized in Fig. 1 describes the theoretical connection between UMPBTs and their classical analogs. It is also informative to examine this connection in actual hypothesis tests. To this end, UMPBTs were used to reanalyze the 855 *t* tests reported in *Psychonomic Bulletin & Review* and *Journal of Experimental Psychology: Learning, Memory, and Cognition* in 2007 (20).

Because exact UMPBTs do not exist for *t* tests, the evidence thresholds obtained from the approximate UMPBTs described in *SI Text* were obtained by ignoring the upper bound on the rejection regions described in Eqs. **3** and **4**. From a practical perspective, this constraint is only important when the *t* statistic for a test is large, and in such cases the null hypothesis can be rejected with a high degree of confidence. To avoid this complication, *t* statistics larger than the value of the *t* statistic that maximizes the Bayes factor in favor of the alternative were excluded from this analysis. Also, because all tests reported by Wetzels et al. (20) were two-sided, the approximate two-sided UMPBTs described in ref. 21 were used in this analysis. The two-sided tests are obtained by defining the alternative hypothesis so that it assigns one-half probability to the two alternative hypotheses that represent the one-sided UMPBT(2γ) tests.

To compute the approximate UMPBTs for the *t* statistics reported in ref. 20, it was assumed that all tests were conducted at the 5% level of significance. The Bayes factors corresponding to the 765 *t* statistics that did not exceed the maximum value are plotted against their *P* values in Fig. 2.

Fig. 2 shows that there is a strong curvilinear relationship between the *P* values of the tests reported in ref. 20 and the Bayes factors obtained from the UMPBT tests. Furthermore, the relationship between the *P* values and Bayes factors is roughly

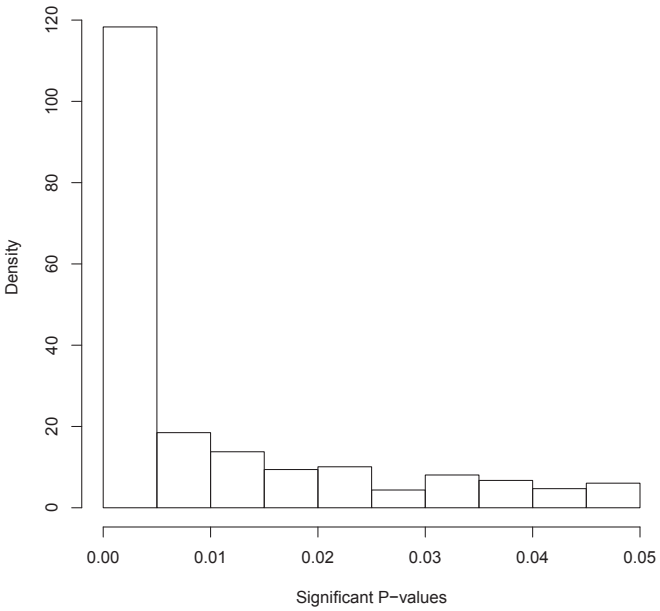

Fig. 3. Histogram of *P* values that were less than 0.05 and reported in ref. 20.

equivalent to the relationship observed with test size in Fig. 1. In this case, P values of 0.05 correspond to Bayes factors around 5, P values of 0.01 correspond to Bayes factors around 20, P values of 0.005 correspond to Bayes factors around 50, and P values of 0.001 correspond to Bayes factors around 150. As before, significant ($P = 0.05$) and highly significant ($P = 0.01$) P values seem to reflect only modest evidence in favor of the alternative hypotheses.

Discussion

The correspondence between P values and Bayes factors based on UMPBTs suggest that commonly used thresholds for statistical significance represent only moderate evidence against null hypotheses. Although it is difficult to assess the proportion of all tested null hypotheses that are actually true, if one assumes that this proportion is approximately one-half, then these results suggest that between 17% and 25% of marginally significant scientific findings are false. This range of false positives is consistent with nonreproducibility rates reported by others (e.g., ref. 5). If the proportion of true null hypotheses is greater than one-half, then the proportion of false positives reported in the scientific literature, and thus the proportion of scientific studies that would fail to replicate, is even higher.

In addition, this estimate of the nonreproducibility rate of scientific findings is based on the use of UMPBTs to establish the rejection regions of Bayesian tests. In general, the use of other default Bayesian methods to model effect sizes results in even higher assignments of posterior probability to rejected null hypotheses, and thus to even higher estimates of false-positive rates. This phenomenon is discussed further in *SI Text*, where Bayes factors obtained using several other default Bayesian procedures are compared with UMPBTs (see Fig. S1). These analyses suggest that the range 17–25% underestimates the actual proportion of marginally significant scientific findings that are false.

Finally, it is important to note that this high rate of nonreproducibility is not the result of scientific misconduct, publication bias, file drawer biases, or flawed statistical designs; it is simply the consequence of using evidence thresholds that do not represent sufficiently strong evidence in favor of hypothesized effects.

As final evidence of the severity of this effect, consider again the t statistics compiled by Wetzels et al. (20). Although the P values derived from these statistics cannot be considered a random sample from any meaningful population, it is nonetheless instructive to examine the distribution of the significant P values derived from these test statistics. A histogram estimate of this distribution is depicted in Fig. 3.

The P values displayed in Fig. 3 presumably arise from two types of experiments: experiments in which a true effect was present and the alternative hypothesis was true, and experiments in which there was no effect present and the null hypothesis was true. For the latter experiments, the nominal distribution of P values is uniformly distributed on the range (0.0, 0.05). The distribution of P values reported for true alternative hypotheses is, by assumption, skewed to the left. The P values displayed in this plot thus represent a mixture of a uniform distribution and some other distribution. Even without resorting to complicated statistical methods to fit this mixture, the appearance of this histogram suggests that many, if not most, of the P values falling above 0.01 are approximately uniformly distributed. That is, most of the significant P values that fell in the range (0.01–0.05) probably represent P values that were computed from data in which the null hypothesis of no effect was true.

These observations, along with the quantitative findings reported in *Results*, suggest a simple strategy for improving the replicability of scientific research. This strategy includes the following steps:

(*i*) Associate statistically significant test results with P values that are less than 0.005. Make 0.005 the default level of significance for setting evidence thresholds in UMPBTs.
(*ii*) Associate highly significant test results with P values that are less than 0.001.
(*iii*) When UMPBTs can be defined (or when other default Bayesian procedures are available), report the Bayes factor in favor of the alternative hypothesis and the default alternative hypothesis that was tested.

Of course, there are costs associated with raising the bar for statistical significance. To achieve 80% power in detecting a standardized effect size of 0.3 on a normal mean, for instance, decreasing the threshold for significance from 0.05 to 0.005 requires an increase in sample size from 69 to 130 in experimental designs. To obtain a highly significant result, the sample size of a design must be increased from 112 to 172.

These costs are offset, however, by the dramatic reduction in the number of scientific findings that will fail to replicate. In terms of evidence, these more stringent criteria will increase the odds that the data must favor the alternative hypothesis to obtain a significant finding from ~3–5:1 to ~25–50:1, and from ~12–15:1 to 100–200:1 to obtain a highly significant result. If one-half of scientifically tested (alternative) hypotheses are true, then these evidence standards will reduce the probability of rejecting a true null hypothesis based on a significant finding from ~20% to less than 4%, and from ~7% to less than 1% when based on a highly significant finding. The more stringent standards will thus reduce false-positive rates by a factor of 5 or more without requiring even a doubling of sample sizes.

Finally, reporting the Bayes factor and the alternative hypothesis that was tested will provide scientists with a mechanism for evaluating the posterior probability that each hypothesis is true. It will also allow scientists to evaluate the scientific importance of the alternative hypothesis that has been favored. Such reports are particularly important in large sample settings in which the default alternative hypothesis provided by the UMPBT may represent only a small deviation from the null hypothesis.

ACKNOWLEDGMENTS. I thank E.-J. Wagenmakers for helpful criticisms and the data used in Figs. 2 and 3. I also thank Suyu Liu, the referees and the editor for numerous suggestions that improved the article. This work was supported by National Cancer Institute Award R01 CA158113.

1. Zimmer C (April 16, 2012) A sharp rise in retractions prompts calls for reform. *NY Times*, Science Section.
2. Naik G (December 2, 2011) Scientists' elusive goal: Reproducing study results. *Wall Street Journal*, Health Section.
3. Begg CB, Mazumdar M (1994) Operating characteristics of a rank correlation test for publication bias. *Biometrics* 50(4):1088–1101.
4. Duval S, Tweedie R (2000) Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics* 56(2):455–463.
5. Ioannidis JP (2005) Contradicted and initially stronger effects in highly cited clinical research. *JAMA* 294(2):218–228.
6. Ioannidis JP, Trikalinos TA (2007) An exploratory test for an excess of significant findings. *Clin Trials* 4(3):245–253.
7. Miller J (2009) What is the probability of replicating a statistically significant effect? *Psychon Bull Rev* 16(4):617–640.
8. Francis G (2012) Evidence that publication bias contaminated studies relating social class and unethical behavior. *Proc Natl Acad Sci USA* 109(25):E1587, author reply E1588.
9. Simonsohn U, Nelson LD, Simmons JP (2013) P-curve: A key to the file drawer. *J Exp Psychol Gen*, in press.
10. Fisher RA (1926) *Statistical Methods for Research Workers* (Oliver and Boyd, Edinburgh).
11. Jeffreys H (1961) *Theory of Probability* (Oxford Univ Press, Oxford), 3rd Ed.
12. Kass RE, Raftery AE (1995) Bayes factors. *J Am Stat Assoc* 90(430):773–795.
13. Berger JO, Selke T (1987) Testing a point null hypothesis: The irreconcilability of p values and evidence. *J Am Stat Assoc* 82(397):112–122.
14. Berger JO, Delampady M (1987) Testing precise hypotheses. *Stat Sci* 2(3):317–335.
15. Edwards W, Lindman H, Savage LJ (1963) Bayesian statistical inference for psychological research. *Psychol Rev* 70(3):193–242.
16. Raftery AE (1995) Bayesian model selection in social research. *Sociol Methodol* 25:111–163.

17. Rouder JN, Speckman PL, Sun D, Morey RD, Iverson G (2009) Bayesian *t* tests for accepting and rejecting the null hypothesis. *Psychon Bull Rev* 16(2):225–237.
18. Wagenmakers E-J, Grünwald P (2006) A Bayesian perspective on hypothesis testing: A comment on Killeen (2005). *Psychol Sci* 17(7):641–642, author reply 643–644.
19. Wagenmakers E-J (2007) A practical solution to the pervasive problems of *p* values. *Psychon Bull Rev* 14(5):779–804.
20. Wetzels R, et al. (2011) Statistical evidence in experimental psychology: An empirical comparison using 855 *t* tests. *Perspect Psychol Sci* 6(3):291–298.
21. Johnson VE (2013) Uniformly most powerful Bayesian tests. *Ann Stat* 41(4):1716–1741.

Supporting Information

Johnson 10.1073/pnas.1313476110

SI Text

This supplement contains two sections. The first section presents a comparison of Bayes factors obtained using uniformly most powerful Bayesian tests (UMPBTs) to Bayes factors obtained using standard Cauchy priors (1–3), intrinsic priors (4), and Bayesian information criterion (BIC)-based approximations to Bayes factors (5–7), all in the context of z tests. In the second, several lemmas are presented that describe the UMPBT(γ) in common testing scenarios. Finally, a table summarizing the results of these lemmas is provided.

Comparison of Bayes Factors

In this section, Bayes factors generated from UMPBT alternatives are compared with Bayes factors obtained from other default Bayesian testing procedures. Each Bayesian testing procedure was used to test whether the mean μ of a random sample of n normal observations with known variance $\sigma^2 = 1$ was equal to 0. Several default procedures were tested. The first, due to Jeffreys (1), is based on the assumption that the prior density for μ under the alternative hypothesis is a standard Cauchy distribution. The extension of this test for unknown σ^2 leads to the Zellner–Siow prior for linear models (2) and testing procedures advocated for psychological tests in ref. 3. The second default method was obtained by assuming an intrinsic prior for μ under the alternative hypothesis (4). The third default method was based on converting the BIC criterion (5) into an approximate Bayes factor, as suggested in refs. 6 and 7.

The prior densities that define the alternative hypothesis in the comparison group are based on the specification of local alternative prior densities, which means that the order at which they accumulate evidence in favor of a true null hypothesis is only $O_p(n^{1/2})$ (8). This slow rate of convergence occurs because local alternative prior densities are not zero at the parameter value the defines a point null hypothesis. Data that support the null hypothesis thereby also provide some support to the alternative, making it difficult to distinguish between the two hypotheses when the null is true. In contrast, the evidence achieved by the UMPBTs in favor of true null hypotheses is bounded by a function of the evidence threshold γ. This means that only a finite amount of evidence can be obtained in favor of a true null hypothesis if γ is held constant as the sample size grows.

All tests were considered to be two-sided. The prior densities for μ under the alternative hypotheses in the approximate UMPBT(γ) two-sided tests were defined by placing one-half of the prior mass corresponding to each of the one-sided UMPBT (2γ)s on μ.

The Bayes factors in favor of the alternative hypotheses under each testing procedure can be expressed as follows.

Cauchy.

$$BF_{10}^{C}(\mathbf{x}) = \exp\left(\frac{n\bar{x}^2}{2}\right) \int_{-\infty}^{\infty} \frac{\exp\left[-n(\bar{x}-\mu)^2/2\right]}{\pi(1+\mu^2)}\, d\mu.$$

Intrinsic.

$$BF_{10}^{I}(\mathbf{x}) = \frac{1}{\sqrt{2n+1}} \exp\left[\frac{(n\bar{x})^2}{2n+1}\right].$$

[Note that the intrinsic prior in this setting is $\mu \sim N(0, 2)$.]

BIC.

$$BF_{10}^{B}(\mathbf{x}) = \exp\left\{0.5\left[n\bar{x}^2 - \log(n)\right]\right\}.$$

UMPBT.

$$BF_{10}^{U}(\mathbf{x}) = \exp\left(\frac{n\bar{x}^2}{2}\right) \left\{\frac{1}{2}\exp\left[-0.5n(\bar{x}-\mu_u)^2\right] + \frac{1}{2}\exp\left[-0.5n(\bar{x}+\mu_u)^2\right]\right\},$$

where

$$\mu_u = \sqrt{\frac{2\log(2\gamma)}{n}}.$$

To study the behavior of the Bayes factors obtained under each of the four procedures, the sample mean of the observed data was assumed to take one of the four values (0, 0.2, 0.4, 0.6). Note that the first value of 0 provides as much evidence in favor of the null hypothesis as can be obtained from the data. The remaining values represent standardized effect sizes of 0.2, 0.4, and 0.6, respectively, because the observational variance is 1. For each assumed value of the sample mean, the sample size was increased from 1 until either a sample size of 5,000 was reached or until the maximum of the Bayes factors exceeded 5,000. These maximum values were imposed to retain detail in the plots for values of the Bayes factors that are of practical interest. Finally, the evidence threshold γ for the UMPBT was determined by equating the rejection region for this test to the rejection region of a two-sided classical test of size 0.005. That is, γ was equal to $\exp(2.807^2/2)/2 = 25.7$.

The value of the Bayes factors obtained under these combinations of sample means and sample sizes is displayed in Fig. S1. This figure reveals a number of interesting features. Among these, this plot illustrates the consistency of the Bayes factors corresponding to the Cauchy, intrinsic, and BIC procedures. These procedures all produce Bayes factors that tend to 0 when $\bar{x} = 0$ and the sample size grows, even though this convergence is slow. In contrast, the UMPBT-based Bayes factor—based on a fixed evidence threshold γ—is constant and approximately equal to $1/2\gamma$ when $\bar{x} = 0$, independently of the sample size. In this respect, UMPBT tests with fixed evidence thresholds are similar to classical hypothesis tests: both maintain a constant "type I error" as the sample size is increased. Preliminary recommendations for increasing γ with sample size to achieve consistency are provided in ref. 9. Similarly, UMPBT-based Bayes factors eventually become smaller than the other three Bayes factors as n grows when γ is held constant, even though the UMPBT is consistent under a true alternative.

For sample sizes typically achieved in practice, the UMPBT-based Bayes factors appear to provide more useful summaries of the evidence in favor of either a true null or true alternative hypothesis than do the other Bayes factors. When $\bar{x} = 0$ for example, the Bayes factor in favor of the null hypothesis is ~50 for all values of n, whereas the other Bayes factors do not achieve this level of support for the null hypothesis until n is greater than ~1,250 (intrinsic), 1,700 (Cauchy), or 2,500 (BIC). For a standardized effect size of 0.2, none of the Bayes factors becomes much larger than 1 until sample sizes of about 50 are obtained, and then the UMPBT-based Bayes factors are larger than the

other Bayes factors for all sample sizes for which the Bayes factors are all less than 5,000. Similar comments apply to observed effect sizes of 0.4 and 0.6, except that smaller sample sizes are needed for all of the Bayes factors to exceed 1. As stated in the main article, these observations demonstrate that UMPBT-based Bayes factors produce more extreme Bayes factors than other default Bayesian procedures for sample sizes and effect sizes of practical interest. This means that the false-positive rates that would be estimated from the other procedures for marginally significant P values would be higher than 17–25%, the range suggested by the use of UMPBTs.

The relative performance of the various Bayes factors for small values of n is also interesting. For all values of $\bar{x}$ considered, the UMPBT-based Bayes factors obtained for $n<5$ suggest more support for the null hypotheses than do the other hypothesis tests. This fact can be attributed to the fact that the UMPBTs are obtained using nonlocal alternative priors on μ, whereas the other tests are based on local priors. As demonstrated in ref. 8, this means that UMPBTs are able to more quickly obtain evidence in support of the null hypothesis. For instance, when $\bar{x}=0.2$ and $n=1$, the UMPBT-based Bayes factor suggests strong support for the null hypothesis, whereas the other Bayes factors assume noncommittal values near 1.0.

When viewed from a scientific perspective, the evidence provided by UMPBTs in favor of the null hypothesis for small values of n and values of $|\bar{x}|\leq 0.6$ seems quite reasonable. Clearly, most scientists would not design an experiment to test whether a normal mean was equal to 0 with fewer than five observations. Unless, of course, μ was assumed to be large relative to σ under the alternative hypothesis. Under such an assumption, the observation of a sample mean less than 0.6σ provides strong evidence in favor of the null hypothesis.

Along similar lines, most classical statisticians regard the sample size n as fixed and ancillary when they conduct hypothesis tests. Under this assumption, UMPBTs violate the likelihood principle because the alternative hypothesis depends on n. In actual practice, however, the sample size selected by a researcher to test an effect size is generally highly informative about the magnitude of that effect size. For instance, few researchers would collect 100,000 observations to detect a standardized effect size of 0.4. A scientist who collects this many observations obviously hopes to detect a much subtler departure from the standard theory. It is also worth noting that sample size calculations themselves require the specification of an alternative hypothesis.

Because the value of the sample size selected for an experiment often reflects prior information regarding the magnitude of an effect size, it is the author's opinion that it is appropriate (and often desirable) to use the sample size chosen by an investigator to specify an alternative hypothesis.

Lemmas

The following lemmas describe the UMPBT(γ) for several common tests.

Lemma 1. *Suppose $X_1,\ldots,X_n$ are independent and identically distributed (iid) according to a normal distribution with mean μ and variance σ^2 (i.e., $N(\mu,\sigma^2)$). Then the one-sided UMPBT(γ) for testing $H_0:\mu=\mu_0$ against any alternative hypothesis that requires $\mu>\mu_0$ is obtained by taking $H_1:\mu=\mu_1$, where*

$$\mu_1=\mu_0+\sigma\sqrt{\frac{2\log(\gamma)}{n}}. \qquad \textbf{[S1]}$$

Similarly, the UMPBT(γ) one-sided test for testing $\mu<\mu_0$ is obtained by taking

$$\mu_1=\mu_0-\sigma\sqrt{\frac{2\log(\gamma)}{n}}.$$

Proof: Provided in ref. 9.

Lemma 2. *Suppose $X_{1,1},\ldots,X_{1,n_1}$ are iid $N\left(\mu-\frac{n_2}{n_1+n_2}\delta,\sigma^2\right)$, and $X_{2,1},\ldots,X_{2,n_2}$ are iid $N\left(\mu+\frac{n_1}{n_1+n_2}\delta,\sigma^2\right)$, where σ^2 is known and the prior distribution for μ is assumed to be uniform on the real line. The one-sided UMPBT(γ) for testing $H_0:\delta=0$ against alternatives that require $\delta>0$ is obtained by taking*

$$H_1:\delta=\sigma\sqrt{\frac{2(n_1+n_2)\log(\gamma)}{n_1n_2}}. \qquad \textbf{[S2]}$$

Proof. Consider first simple alternative hypotheses on $\delta>0$. Up to a constant factor that arises from the uniform distribution on μ, the marginal distribution of the data under the null hypothesis can be obtained by integrating out μ to obtain

$$m_0(\mathbf{x})=\left(2\pi\sigma^2\right)^{-(n_1+n_2-1)/2}(n_1+n_2)^{-1/2}\times\exp\left[-\frac{1}{2\sigma^2}\sum_{j=1}^{2}\sum_{i=1}^{n_j}\left(x_{j,i}-\bar{x}_j\right)^2\right]. \qquad \textbf{[S3]}$$

Similarly, the marginal distribution of the data under the alternative that $\mu_2-\mu_1=\delta$ can be obtained by integrating out μ to obtain

$$m_1(\mathbf{x})=m_0(\mathbf{x})\exp\left\{-\frac{1}{2\sigma^2}\left[\frac{n_1n_2}{n_1+n_2}\delta^2-\frac{2n_1n_2}{n_1+n_2}(\bar{x}_2-\bar{x}_1)\delta\right]\right\}. \qquad \textbf{[S4]}$$

It follows that

$$\mathbf{P}[\log(BF_{10})>log(\gamma)]=\mathbf{P}\left[\bar{x}_2-\bar{x}_1>\frac{(n_1+n_2)\sigma^2\log(\gamma)}{n_1n_2\delta}+\frac{\delta}{2}\right]. \qquad \textbf{[S5]}$$

Regardless of the distribution of $(\bar{x}_2-\bar{x}_1)$, this probability can be maximized by minimizing the right-hand side of the last inequality with respect to δ. The UMPBT value for δ is thus

$$\delta^*=\sigma\sqrt{\frac{2(n_1+n_2)\log(\gamma)}{n_1n_2}}. \qquad \textbf{[S6]}$$

Now consider composite alternative hypotheses, and let $BF_{10}(\delta)$ denote the value of the Bayes factor when evaluated at a particular value of δ and fixed $\mathbf{x}$. Define an indicator function s according to

$$s(\mathbf{x},\delta)=Ind(BF_{10}(\delta)>\gamma). \qquad \textbf{[S7]}$$

Then it follows from Eq. **S5** that

$$s(\mathbf{x},\delta)\leq s(\mathbf{x},\delta^*) \quad \text{for all } \mathbf{x}. \qquad \textbf{[S8]}$$

This implies that

$$\int_0^\infty s(\mathbf{x},\delta)\pi(\delta)\leq s(\mathbf{x},\delta^*) \qquad \textbf{[S9]}$$

for all probability densities $\pi(\delta)$. It follows that

$$\mathbf{P}_{\delta_t}(BF_{10} > \gamma) = \int_{\mathcal{X}} s(\mathbf{x}, \delta) f(\mathbf{x}|\delta_t) d\delta_t \quad \text{[S10]}$$

is maximized by a prior density that concentrates its mass δ^*. Here $f(\mathbf{x}|\delta_t)$ is the sampling density of $\mathbf{x}$ for $\delta = \delta_t$,

Lemma 3. *Suppose that X is distributed according to a χ^2 distribution on 1 degree of freedom and noncentrality parameter λ; that is, $X \sim \chi^2_1(\lambda)$. The UMPBT(γ) for testing $H_0 : \lambda = 0$ is obtained by taking $H_1 : \lambda = \lambda_1$, where λ_1 is the value of λ that minimizes*

$$\frac{1}{\sqrt{\lambda}} \log\left(e^{\lambda/2}\gamma + \sqrt{e^{\lambda}\gamma^2 - 1}\right). \quad \text{[S11]}$$

Proof. As in *Lemma 2*, consider first simple alternative hypotheses on $\lambda > 0$. By taking the ratio of a noncentral χ^2 density on 1 degree of freedom to the central χ^2 density on 1 degree of freedom, it follows that the Bayes factor in favor of the alternative can be expressed as

$$\sum_{i=0}^{\infty} \frac{e^{\lambda/2}\Gamma\left(\frac{1}{2}\right)\left(\frac{\lambda x}{2}\right)^i}{i! 2^i \Gamma\left(\frac{1}{2} + i\right)}. \quad \text{[S12]}$$

Noting that

$$\Gamma\left(\frac{1}{2} + i\right) = \frac{(2i)!\Gamma(1/2)}{4^i i!}, \quad \text{[S13]}$$

and that

$$\cosh\left(\sqrt{\lambda x}\right) = \sum_{i=0}^{\infty} \frac{(\lambda x)^i}{(2i)!}, \quad \text{[S14]}$$

it follows that

$$BF_{10}(\lambda) = e^{-\lambda/2}\cosh\left(\sqrt{\lambda x}\right). \quad \text{[S15]}$$

The probability that the Bayes factor exceeds the evidence threshold is given by

$$\begin{aligned}\mathbf{P}_{\lambda_t}[BF_{10} > \gamma] &= \mathbf{P}_{\lambda_t}\left[\cosh\left(\sqrt{\lambda x}\right) > e^{\lambda/2}\gamma\right] \\ &= \mathbf{P}_{\lambda_t}\left[\sqrt{x} > \lambda^{-1/2}\, \log\left(e^{\lambda/2}\gamma + \sqrt{e^{\lambda}\gamma^2 - 1}\right)\right].\end{aligned} \quad \text{[S16]}$$

Minimizing the right-hand side of the inequality maximizes the probability, regardless of the value of λ_t. The extension to composite hypotheses follows along from the same logic used in Eqs. **S7–S10**.

Lemma 4. *Suppose that X has a binomial distribution with success probability p and denominator n. The UMPBT(γ) for testing $H_0 : p = p_0$ against alternatives that require $p > p_0$ is obtained by taking $H_1 : p = p_1$, where p_1 is the value of p that minimizes*

$$\frac{\log(\gamma) - n[\log(1-p) - \log(1-p_0)]}{\log[p/(1-p)] - \log[p_0/(1-p_0)]}. \quad \text{[S17]}$$

The UMPBT(γ) for alternatives that require $p < p_0$ is obtained by taking p_1 to be the value of p that maximizes Eq. **S17**.

Proof. Provided in ref. 9.

Lemma 5. *Assume that the conditions of Lemma 1 apply, except that σ^2 is not known. Suppose that the prior distribution for σ^2 is an inverse gamma distribution with parameters α and λ, and define*

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad W = \sum_{i=1}^{n}(x_i - \bar{x})^2 + 2\lambda, \quad \text{and} \quad \gamma^* = \gamma^{\frac{2}{n+\alpha}}. \quad \text{[S18]}$$

Then the value of μ_1 that minimizes a_n in Eq. **S4** *is*

$$\mu_1 = \mu_0 + \sqrt{\frac{W(\gamma^* - 1)}{n}}. \quad \text{[S19]}$$

If a noninformative prior is assumed for σ^2 (i.e., $\alpha = \lambda = 0$), then the UMPBT(γ) alternative is obtained by taking

$$\mu_1 = \mu_0 + s\sqrt{(\gamma^* - 1)\frac{(n-1)}{n}},$$

where

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Proof. As in the previous proofs, consider first the case of simple alternative hypotheses. By integrating out the variance parameter, it follows that the Bayes factor in favor of the alternative hypothesis can be expressed as

$$BF_{10}(\mu_1) = \left[\frac{W + n(\bar{x} - \mu_0)^2}{W + n(\bar{x} - \mu_1)^2}\right]^{n/2+\alpha}. \quad \text{[S20]}$$

After some algebra, this expression leads to the following equation:

$$\mathbf{P}_{\mu_t}[BF_{10}(\mu_1) > \gamma] = \mathbf{P}_{\mu_t}[a_n < \bar{x} < b_n], \quad \text{[S21]}$$

where

$$a_n = \frac{\gamma^*\mu_1 - \mu_0}{\gamma^* - 1} - \sqrt{\frac{\gamma^*(\mu_1 - \mu_0)^2}{(\gamma^* - 1)^2} - \frac{W}{n}} \quad \text{[S22]}$$

and

$$b_n = \frac{\gamma^*\mu_1 - \mu_0}{\gamma^* - 1} + \sqrt{\frac{\gamma^*(\mu_1 - \mu_0)^2}{(\gamma^* - 1)^2} - \frac{W}{n}}. \quad \text{[S23]}$$

Minimizing a_n as a function of μ_1 leads to the stated result.

Lemma 6. *Assume that the conditions of Lemma 2 apply, except that the variance σ^2 is unknown. Suppose the prior distribution for σ^2 is an inverse gamma distribution with parameters α and λ, and define*

$$\bar{x}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} x_{j,i}, \quad W = \sum_{j=1}^{2}\sum_{i=1}^{n_j}\left(x_{j,i} - \bar{x}_j\right)^2 + 2\lambda, \quad \text{and} \quad \gamma^* = \gamma^{\frac{2}{n_1+n_2+2\alpha-1}}. \quad \text{[S24]}$$

Then the value of δ than minimizes a_n in Eq. **S5** *is*

$$\delta = \sqrt{\frac{W(\gamma^* - 1)(n_1 + n_2)}{n_1 n_2}}. \quad \text{[S25]}$$

Taking $\alpha = \lambda = 0$ *and*

$$s^2 = \frac{1}{n_1 + n_2 - 2} \sum_{j=1}^{2} \sum_{i=1}^{n_j} (x_{j,i} - \bar{x}_j)^2,$$

the UMPBT(γ) *alternative is defined by taking*

$$\delta = s \sqrt{\frac{(\gamma^* - 1)\nu(n_1 + n_2)}{n_1 n_2}}.$$

Proof. Similar to the proofs of *Lemmas 2* and *5*. Using the expressions for the marginal distributions obtained in the case of a known variance in *Lemma 2*, it can be shown that the Bayes factor takes the form of the ratio of t densities. Solving for the difference in means $\mu_2 - \mu_1$ leads to an inequality similar to Eq. **S21**, and the result follows.

A summary of the results of *Lemmas 1–6* appears in Table S1. Also provided in this table are expressions for the Bayes factors (expressed in terms of standard test statistics), rejection regions, and the relation between evidence threshold γ and the size of the corresponding classical test.

1. Jeffreys H (1961) *Theory of Probability* (Oxford Univ Press, Oxford), 3rd Ed.
2. Zellner A, et al., eds (1980) Posterior odds ratios for selected regression hypotheses. *Bayesian Statistics: Proceedings of the First International Meeting Held in Valencia (Spain)*, eds Bernardo JM, DeGroot MH, Lindley DV, Smith AFM (University Press, Valencia, Spain), pp 585–603.
3. Rouder JN, Speckman PL, Sun D, Morey RD, Iverson G (2009) Bayesian *t* tests for accepting and rejecting the null hypothesis. *Psychon Bull Rev* 16(2):225–237.
4. Berger JO, Pericchi LR (1996) The intrinsic Bayes factor for model selection and prediction. *J Am Stat Assoc* 91(433):109–122.
5. Schwarz G (1978) Estimating the dimension of a model. *Ann Stat* 6(2):461–464.
6. Kass RE, Raftery AE (1995) Bayes factors. *J Am Stat Assoc* 90(430):773–795.
7. Wagenmakers E-J (2007) A practical solution to the pervasive problems of *p* values. *Psychon Bull Rev* 14(5):779–804.
8. Johnson VE, Rossell D (2010) On the use of non-local prior densities in Bayesian hypothesis tests. *J Roy Stat Soc Ser B Method* 72(2):143–170.
9. Johnson VE (2013) Uniformly most powerful Bayesian tests. *Ann Stat* 41(4):1716–1741.

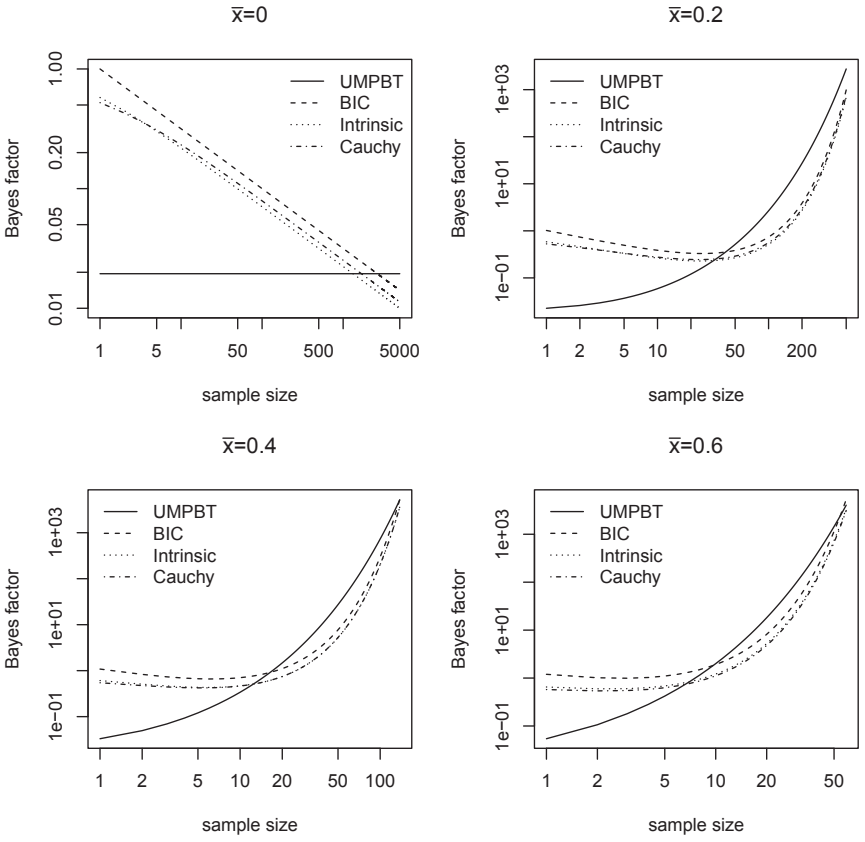

Fig. S1. Comparison of default Bayesian procedures for testing a null hypothesis that the mean of n $N(\mu,1)$ random variables is 0.

Table S1. Properties of UMPBTs in common testing situations

Test	Variables	H_1	Bayes factor	Reject region	$\gamma = f(\alpha)$
One-sample z	$z = \frac{\sqrt{n}(\bar{x}-\mu_0)}{\sigma}$	$\mu_1 = \mu_0 + \sigma\sqrt{\frac{2\log(\gamma)}{n}}$	$\exp\left[z\sqrt{2\log(\gamma)} - \log(\gamma)\right]$	$z > \sqrt{2\log(\gamma)}$	$\gamma = \exp\left(\frac{z_\alpha^2}{2}\right)$
Two-sample z	$z = \sqrt{\frac{n_1 n_2}{n_1+n_2}}\ \frac{\bar{x}_2-\bar{x}_1}{\sigma}$	$\delta = \sigma\sqrt{\frac{2(n_1+n_2)\log(\gamma)}{n_1 n_2}}$	$\exp\left[z\sqrt{2\log(\gamma)} - \log(\gamma)\right]$	$z > \sqrt{2\log(\gamma)}$	$\gamma = \exp\left(\frac{z_\alpha^2}{2}\right)$
One-sample t	$t = \frac{\sqrt{n}(\bar{x}-\mu_0)}{s}$				
	$s^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}$	$\mu_1 = \mu_0 + s\sqrt{\nu\gamma^*/n}$	$\left(\frac{\nu+t^2}{\nu+\left[t-\sqrt{\nu\gamma^*}\right]^2}\right)^m$	$t > \sqrt{\nu\gamma^*}$	$\gamma = \left(\frac{t_\alpha^2}{\nu}+1\right)^m$
	$\nu = n-1$				
	$\gamma^* = \gamma^{2/n} - 1$				
	$m = n/2$				
Two-sample t	$t = \sqrt{\frac{n_1 n_2}{n_1+n_2}}\ \frac{\bar{x}_2-\bar{x}_1}{s}$				
	$s^2 = \frac{\sum\sum(x_{i,j}-\bar{x}_j)^2}{n_1+n_2-2}$	$\delta = s\sqrt{\frac{2m\gamma^*\nu}{n_1 n_2}}$	$\left(\frac{\nu+t^2}{\nu+\left[t-\sqrt{\nu\gamma^*}\right]^2}\right)^m$	$t > \sqrt{\nu\gamma^*}$	$\gamma = \left(\frac{t_\alpha^2}{\nu}+1\right)^m$
	$\nu = n_1+n_2-2$				
	$\gamma^* = \gamma^{2/(n_1+n_2-1)} - 1$				
	$m = (n_1+n_2)/2$				
χ_1^2	x	$\lambda_1 = \arg\min_{\lambda} \frac{\log\left(a+\sqrt{a^2-1}\right)}{\sqrt{\lambda}}$	$\exp\left(-\frac{\lambda_1}{2}\right)\cosh\left(\sqrt{\lambda_1 x}\right)$		
	$a = \gamma e^{\lambda/2}$				
Proportion	(x,n)	$p_1 = \arg\min_{p} \frac{\log(\gamma) - n\Delta(p,p_0)}{\mathrm{logit}(p) - \mathrm{logit}(p_0)}$	$\left(\frac{p_1}{p_0}\right)^x \left(\frac{1-p_1}{1-p_0}\right)^{n-x}$		
p_0	$\Delta(p,p_0) = \log\left(\frac{1-p}{1-p_0}\right)$				

Note that the Bayes factors listed for the one- and two-sample t tests should only be used for $t < \sqrt{\nu\gamma^*} + \sqrt{\nu\gamma^* + 4\nu}$. Values for quantities in empty cells must be determined using numerical techniques.