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.

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.

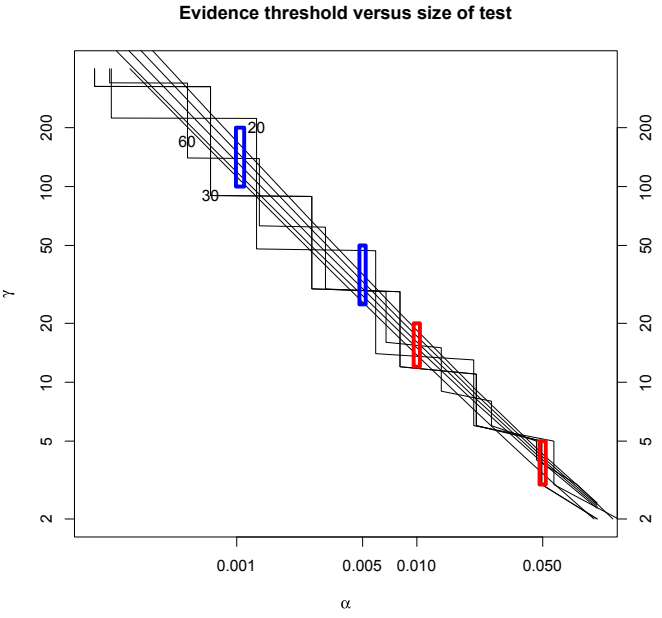

Fig. 1. Evidence thresholds and size of corresponding significance tests. The UMPBT and significance tests used to construct this plot have the same (z, χ^2, and binomial tests) or approximately the same (t tests) rejection regions. The smooth curves represent, from *Top* to *Bottom*, t tests based on 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 $p_0 = 0.5$.

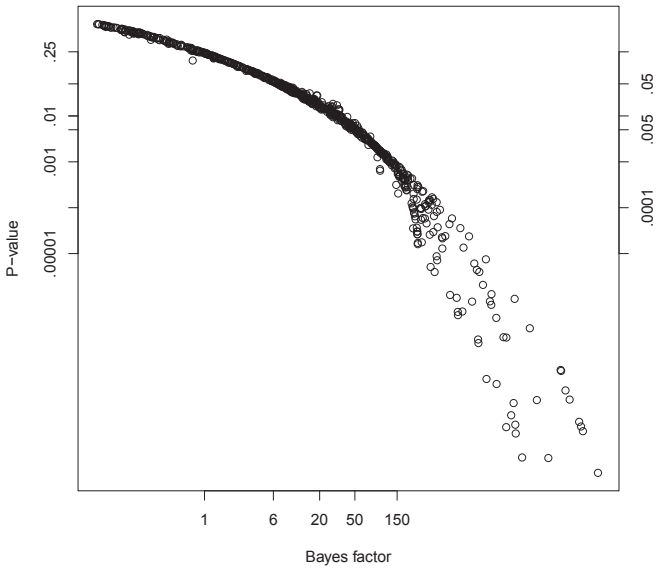

Fig. 2. *P* values versus UMPBT Bayes factors. This plot depicts approximate Bayes factors derived from 765 *t* statistics reported by Wetzels et al. (20). A breakdown of the curvilinear relationship between Bayes factors and *P* values occurs in the lower right portion of the plot, which corresponds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

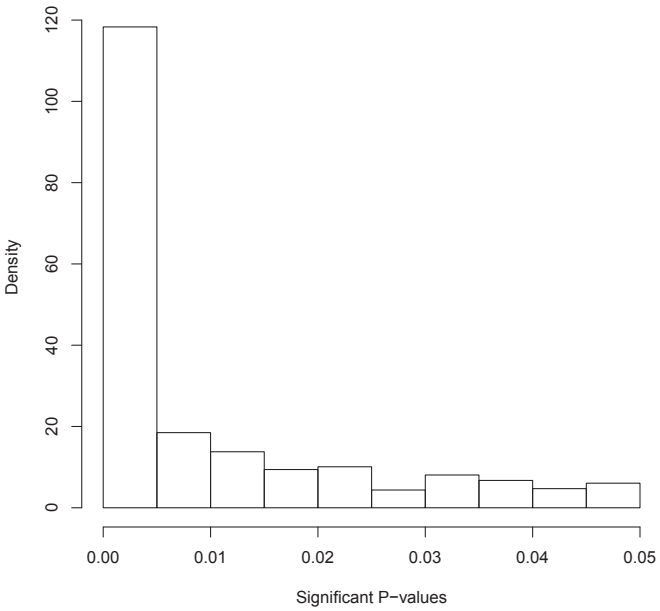

Fig. 3. Histogram of *P* values that were less than 0.05 and reported in ref. 20.

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.