UNIFORMLY MOST POWERFUL BAYESIAN TESTS

BY VALEN E. JOHNSON[1]

Texas A&M University

Uniformly most powerful tests are statistical hypothesis tests that provide the greatest power against a fixed null hypothesis among all tests of a given size. In this article, the notion of uniformly most powerful tests is extended to the Bayesian setting by defining uniformly most powerful Bayesian tests to be tests that maximize the probability that the Bayes factor, in favor of the alternative hypothesis, exceeds a specified threshold. Like their classical counterpart, uniformly most powerful Bayesian tests are most easily defined in one-parameter exponential family models, although extensions outside of this class are possible. The connection between uniformly most powerful tests and uniformly most powerful Bayesian tests can be used to provide an approximate calibration between *p*-values and Bayes factors. Finally, issues regarding the strong dependence of resulting Bayes factors and *p*-values on sample size are discussed.

1. Introduction. Uniformly most powerful tests (UMPTs) were proposed by Neyman and Pearson in a series of articles published nearly a century ago [e.g., Neyman and Pearson (1928, 1933); see Lehmann and Romano (2005) for a comprehensive review of the subsequent literature]. They are defined as statistical hypothesis tests that provide the greatest power among all tests of a given size. The goal of this article is to extend the classical notion of UMPTs to the Bayesian paradigm through the definition of uniformly most powerful Bayesian tests (UMPBTs) as tests that maximize the probability that the Bayes factor against a fixed null hypothesis exceeds a specified threshold. This extension is important from several perspectives.

From a classical perspective, the outcome of a hypothesis test is a decision either to reject the null hypothesis or not to reject the null hypothesis. This approach to hypothesis testing is closely related to Popper's theory of critical rationalism, in which scientific theories are never accepted as being true, but instead are only subjected to increasingly severe tests [e.g., Mayo and Spanos (2006), Popper (1959)]. Many scientists and philosophers, notably Bayesians, find this approach unsatisfactory for at least two reasons [e.g., Howson and Urbach (2005), Jeffreys (1939)].

TABLE 1
Probabilities of a random variable under competing hypotheses

X	1	2	3
Null hypothesis	0.99	0.008	0.001
Alternative hypothesis	0.01	0.001	0.989

First, a decision not to reject the null hypothesis provides little quantitative information regarding the truth of the null hypothesis. Second, the rejection of a null hypothesis may occur even when evidence from the data strongly support its validity. The following two examples—one contrived and one less so—illustrate these concerns.

The first example involves a test for the distribution of a random variable X that can take values 1, 2 or 3; cf. Berger and Wolpert (1984). The probability of each outcome under two competing statistical hypotheses is provided in Table 1. From this table, it follows that a most powerful test can be defined by rejecting the null hypothesis when $X = 2$ or 3. Both error probabilities of this test are equal to 0.01.

Despite the test's favorable operating characteristics, the rejection of the null hypothesis for $X = 2$ seems misleading: $X = 2$ is 8 times more likely to be observed under the null hypothesis than it is under the alternative. If both hypotheses were assigned equal odds a priori, the null hypothesis is rejected at the 1% level of significance even though the posterior probability that it is true is 0.89. As discussed further in Section 2.1, such clashes between significance tests and Bayesian posterior probabilities can occur in variety of situations and can be particularly troubling in large sample settings.

The second example represents a stylized version of an early phase clinical trial. Suppose that a standard treatment for a disease is known to be successful in 25% of patients, and that an experimental treatment is concocted by supplementing the standard treatment with the addition of a new drug. If the supplemental agent has no effect on efficacy, then the success rate of the experimental treatment is assumed to remain equal to 25% (the null hypothesis). A single arm clinical trial is used to test this hypothesis. The trial is based on a one-sided binomial test at the 5% significance level. Thirty patients are enrolled in the trial.

If y denotes the number of patients who respond to the experimental treatment, then the critical region for the test is $y \geq 12$. To examine the properties of this test, suppose first that $y = 12$, so that the null hypothesis is rejected at the 5% level. In this case, the *minimum* likelihood ratio in favor of the null hypothesis is obtained by setting the success rate under the alternative hypothesis to $12/30 = 0.40$ (in which case the power of the test is 0.57). That is, if the new treatment's success rate were defined a priori to be 0.4, then the likelihood ratio in favor of the null

hypothesis would be

$$(1) \qquad L_{\min} = \frac{0.25^{12}0.75^{18}}{0.4^{12}0.6^{18}} = 0.197.$$

For any other alternative hypothesis, the likelihood ratio in favor of the null hypothesis would be larger than 0.197 [e.g., Edwards, Lindman and Savage (1963)]. If equal odds are assigned to the null and alternative hypothesis, then the posterior probability of the null hypothesis is *at least* 16.5%. In this case, the null hypothesis is rejected at the 5% level of significance even though the data support it. And, of course, the posterior probability of the null hypothesis would be substantially higher if one accounted for the fact that a vast majority of early phase clinical trials fail.

Conversely, suppose now that the trial data provide clear support of the null hypothesis, with only 7 successes observed during the trial. In this case, the null hypothesis is not rejected at the 5% level, but this fact conveys little information regarding the relative support that the null hypothesis received. If the alternative hypothesis asserts, as before, that the success rate of the new treatment is 0.4, then the likelihood ratio in favor of the null hypothesis is 6.31; that is, the data favor the null hypothesis with approximately 6:1 odds. If equal prior odds are assumed between the two hypotheses, then the posterior probability of the null hypothesis is 0.863. Under the assumption of clinical equipoise, the prior odds assigned to the two hypotheses are assumed to be equal, which means the only controversial aspect of reporting such odds is the specification of the alternative hypothesis.

For frequentists, the most important aspect of the methodology reported in this article may be that it provides a connection between frequentist and Bayesian testing procedures. In one-parameter exponential family models with monotone likelihood ratios, for example, it is possible to define a UMPBT with the same rejection region as a UMPT. This means that a Bayesian using a UMPBT and a frequentist conducting a significance test will make identical decisions on the basis of the observed data, which suggests that either interpretation of the test may be invoked. 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 level. This fact provides a remedy to the two primary deficiencies of classical significance tests—their inability to quantify evidence in favor of the null hypothesis when the null hypothesis is not rejected, and their tendency to exaggerate evidence against the null when it is. Having determined the corresponding UMPBT, Bayes factors can be used to provide a simple summary of the evidence in favor of each hypothesis.

For Bayesians, UMPBTs represent a new objective Bayesian test, at least when objective Bayesian methods are interpreted in the broad sense. As Berger (2006) notes, "there is no unanimity as to the definition of objective Bayesian analysis..." and "many Bayesians object to the label 'objective Bayes,'" preferring

other labels such as "noninformative, reference, default, conventional and non-subjective." Within this context, UMPBTs provide a new form of default, nonsubjective Bayesian tests in which the alternative hypothesis is determined so as to maximize the probability that a Bayes factor exceeds a specified threshold. This threshold can be specified either by a default value—say 10 or 100—or, as indicated in the preceding discussion, determined so as to produce a Bayesian test that has the same rejection region as a classical UMPT. In the latter case, UMPBTs provide an objective Bayesian testing procedure that can be used to translate the results of classical significance tests into Bayes factors and posterior model probabilities. By so doing, UMPBTs may prove instrumental in convincing scientists that commonly-used levels of statistical significance do not provide "significant" evidence against rejected null hypotheses.

Subjective Bayesian methods have long provided scientists with a formal mechanism for assessing the probability that a standard theory is true. Unfortunately, subjective Bayesian testing procedures have not been—and will likely never be—generally accepted by the scientific community. In most testing problems, the range of scientific opinion regarding the magnitude of violations from a standard theory is simply too large to make the report of a single, subjective Bayes factor worthwhile. Furthermore, scientific journals have demonstrated an unwillingness to replace the report of a single *p*-value with a range of subjectively determined Bayes factors or posterior model probabilities.

Given this reality, subjective Bayesians may find UMPBTs useful for communicating the results of Bayesian tests to non-Bayesians, even when a UMPBT is only one of several Bayesian tests that are reported. By reducing the controversy regarding the specification of prior densities on parameter values under individual hypotheses, UMPBTs can also be used to focus attention on the specification of prior probabilities on the hypotheses themselves. In the clinical trial example described above, for example, the value of the success probability specified under the alternative hypothesis may be less important in modeling posterior model probabilities than incorporating information regarding the outcomes of previous trials on related supplements. Such would be the case if numerous previous trials of similar agents had failed to provide evidence of increased treatment efficacy.

UMPBTs possess certain favorable properties not shared by other objective Bayesian methods. For instance, most objective Bayesian tests implicitly define local alternative prior densities on model parameters under the alternative hypothesis [e.g., Berger and Pericchi (1996), Jeffreys (1939), O'Hagan (1995)]. As demonstrated in Johnson and Rossell (2010), however, the use of local alternative priors makes it difficult to accumulate evidence in favor of a true null hypothesis. This means that many objective Bayesian methods are only marginally better than classical significance tests in summarizing evidence in favor of the null hypothesis. For small to moderate sample sizes, UMPBTs produce alternative hypotheses that correspond to nonlocal alternative prior densities, which means that they are able

to provide more balanced summaries of evidence collected in favor of true null and true alternative hypotheses.

UMPBTs also possess certain unfavorable properties. Like many objective Bayesian methods, UMPBTs can violate the likelihood principle, and their behavior in large sample settings can lead to inconsistency if evidence thresholds are held constant. And the alternative hypotheses generated by UMPBTs are neither vague nor noninformative. Further comments and discussion regarding these issues are provided below.

In order to define UMPBTs, it useful to first review basic properties of Bayesian hypothesis tests. In contrast to classical statistical hypothesis tests, Bayesian hypothesis tests are based on comparisons of the posterior probabilities assigned to competing hypotheses. In parametric tests, competing hypotheses are characterized by the prior densities that they impose on the parameters that define a sampling density shared by both hypotheses. Such tests comprise the focus of this article. Specifically, it is assumed throughout that the posterior odds between two hypotheses H_1 and H_0 can be expressed as

$$(2) \qquad \frac{\mathbf{P}(H_1 \mid \mathbf{x})}{\mathbf{P}(H_0 \mid \mathbf{x})} = \frac{m_1(\mathbf{x})}{m_0(\mathbf{x})} \times \frac{p(H_1)}{p(H_0)},$$

where $\mathrm{BF}_{10}(\mathbf{x}) = m_1(\mathbf{x})/m_0(\mathbf{x})$ is the Bayes factor between hypotheses H_1 and H_0,

$$(3) \qquad m_i(\mathbf{x}) = \int_\Theta f(\mathbf{x} \mid \boldsymbol{\theta}) \pi_i(\boldsymbol{\theta} \mid H_i) \, d\boldsymbol{\theta}$$

is the marginal density of the data under hypothesis H_i, $f(\mathbf{x} \mid \boldsymbol{\theta})$ is the sampling density of data $\mathbf{x}$ given $\boldsymbol{\theta}$, $\pi_i(\boldsymbol{\theta} \mid H_i)$ is the prior density on $\boldsymbol{\theta}$ under H_i and $p(H_i)$ is the prior probability assigned to hypothesis H_i, for $i = 0, 1$. The marginal prior density for $\boldsymbol{\theta}$ is thus

$$\pi(\theta) = \pi_0(\theta \mid H_0) P(H_0) + \pi_1(\theta \mid H_1) p(H_1).$$

When there is no possibility of confusion, $\pi_i(\theta \mid H_i)$ will be denoted more simply by $\pi_i(\theta)$. The parameter space is denoted by Θ and the sample space by $\mathcal{X}$. The logarithm of the Bayes factor is called the *weight of evidence.* All densities are assumed to be defined with respect to an appropriate underlying measure (e.g., Lebesgue or counting measure).

Finally, assume that one hypothesis—the null hypothesis H_0—is fixed on the basis of scientific considerations, and that the difficulty in constructing a Bayesian hypothesis test arises from the requirement to specify an alternative hypothesis. This assumption mirrors the situation encountered in classical hypothesis tests in which the null hypothesis is known, but no alternative hypothesis is defined. In the clinical trial example, for instance, the null hypothesis corresponds to the assumption that the success probability of the new treatment equals that of the standard treatment, but there is no obvious value (or prior probability density) that should

be assigned to the treatment's success probability under the alternative hypothesis that it is better than the standard of care.

With these assumptions and definitions in place, it is worthwhile to review a property of Bayes factors that pertains when the prior density defining an alternative hypothesis is misspecified. Let $\pi_t(\boldsymbol{\theta} \mid H_1) = \pi_t(\boldsymbol{\theta})$ denote the "true" prior density on $\boldsymbol{\theta}$ under the assumption that the alternative hypothesis is true, and let $m_t(\mathbf{x})$ denote the resulting marginal density of the data. In general $\pi_t(\boldsymbol{\theta})$ is not known, but it is still possible to compare the properties of the weight of evidence that would be obtained by using the true prior density under the alternative hypothesis to those that would be obtained using some other prior density. From a frequentist perspective, π_t might represent a point mass concentrated on the true, but unknown, data generating parameter. From a Bayesian perspective, π_t might represent a summary of existing knowledge regarding $\boldsymbol{\theta}$ before an experiment is conducted. Because π_t is not available, suppose that $\pi_1(\boldsymbol{\theta} \mid H_1) = \pi_1(\boldsymbol{\theta})$ is instead used to represent the prior density, again under the assumption that the alternative hypothesis is true. Then it follows from Gibbs's inequality that

$$\int_{\mathcal{X}} m_t(\mathbf{x}) \log\left[\frac{m_t(\mathbf{x})}{m_0(\mathbf{x})}\right] d\mathbf{x} - \int_{\mathcal{X}} m_t(\mathbf{x}) \log\left[\frac{m_1(\mathbf{x})}{m_0(\mathbf{x})}\right] d\mathbf{x}$$
$$= \int_{\mathcal{X}} m_t(\mathbf{x}) \log\left[\frac{m_t(\mathbf{x})}{m_1(\mathbf{x})}\right] d\mathbf{x}$$
$$\geq 0.$$

That is,

$$(4) \qquad \int_{\mathcal{X}} m_t(\mathbf{x}) \log\left[\frac{m_t(\mathbf{x})}{m_0(\mathbf{x})}\right] d\mathbf{x} \geq \int_{\mathcal{X}} m_t(\mathbf{x}) \log\left[\frac{m_1(\mathbf{x})}{m_0(\mathbf{x})}\right] d\mathbf{x},$$

which means that the expected weight of evidence in favor of the alternative hypothesis is always decreased when $\pi_1(\boldsymbol{\theta})$ differs from $\pi_t(\boldsymbol{\theta})$ (on a set with measure greater than 0). In general, the UMPBTs described below will thus decrease the average weight of evidence obtained in favor of a true alternative hypothesis. *In other words, the weight of evidence reported from a UMPBT will tend to underestimate the actual weight of evidence provided by an experiment in favor of a true alternative hypothesis.*

Like classical statistical hypothesis tests, the tangible consequence of a Bayesian hypothesis test is often the rejection of one hypothesis, say H_0, in favor of the second, say H_1. In a Bayesian test, the null hypothesis is rejected if the posterior probability of H_1 exceeds a certain threshold. Given the prior odds between the hypotheses, this is equivalent to determining a threshold, say γ, over which the Bayes factor between H_1 and H_0 must fall in order to reject H_0 in favor of H_1. It is therefore of some practical interest to determine alternative hypotheses that maximize the probability that the Bayes factor from a test exceeds a specified threshold.

With this motivation and notation in place, a UMPBT(γ) may be formally defined as follows.

DEFINITION. A uniformly most powerful Bayesian test for evidence threshold $\gamma > 0$ in favor of the alternative hypothesis H_1 against a fixed null hypothesis H_0, denoted by UMPBT(γ), 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 $H_2 : \boldsymbol{\theta} \sim \pi_2(\boldsymbol{\theta})$:

$$\text{(5)} \qquad \mathbf{P}_{\boldsymbol{\theta}_t}\big[\text{BF}_{10}(\mathbf{x}) > \gamma\big] \geq \mathbf{P}_{\boldsymbol{\theta}_t}\big[\text{BF}_{20}(\mathbf{x}) > \gamma\big].$$

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

The remainder of this article is organized as follows. In the next section, UMPBTs are described for one-parameter exponential family models. As in the case of UMPTs, a general prescription for constructing UMPBTs is available only within this class of densities. Specific techniques for defining UMPBTs or approximate UMPBTs outside of this class are described later in Sections 4 and 5. In applying UMPBTs to one parameter exponential family models, an approximate equivalence between type I errors for UMPTs and the Bayes factors obtained from UMPBTs is exposed.

In Section 3, UMPBTs are applied in two canonical testing situations: the test of a binomial proportion, and the test of a normal mean. These two tests are perhaps the most common tests used by practitioners of statistics. The binomial test is illustrated in the context of a clinical trial, while the normal mean test is applied to evaluate evidence reported in support of the Higgs boson. Section 4 describes several settings outside of one parameter exponential family models for which UMPBTs exist. These include cases in which the nuisance parameters under the null and alternative hypothesis can be considered to be equal (though unknown), and situations in which it is possible to marginalize over nuisance parameters to obtain expressions for data densities that are similar to those obtained in one-parameter exponential family models. Section 5 describes approximations to UMPBTs obtained by specifying alternative hypotheses that depend on data through statistics that are ancillary to the parameter of interest. Concluding comments appear in Section 6.

2. One-parameter exponential family models. Assume that $\{x_1, \ldots, x_n\} \equiv \mathbf{x}$ are i.i.d. with a sampling density (or probability mass function in the case of discrete data) of the form

$$\text{(6)} \qquad f(x \mid \theta) = h(x) \exp\big[\eta(\theta)T(x) - A(\theta)\big],$$

where $T(x)$, $h(x)$, $\eta(\theta)$ and $A(\theta)$ are known functions, and $\eta(\theta)$ is monotonic. Consider a one-sided test of a point null hypothesis $H_0 : \theta = \theta_0$ against an arbitrary alternative hypothesis. Let γ denote the evidence threshold for a UMPBT(γ), and assume that the value of θ_0 is fixed.

LEMMA 1. *Assume the conditions of the previous paragraph pertain, and define $g_\gamma(\theta, \theta_0)$ according to*

$$(7) \qquad g_\gamma(\theta, \theta_0) = \frac{\log(\gamma) + n[A(\theta) - A(\theta_0)]}{\eta(\theta) - \eta(\theta_0)}.$$

In addition, define u to be 1 or −1 according to whether $\eta(\theta)$ is monotonically increasing or decreasing, respectively, and define v to be either 1 or −1 according to whether the alternative hypothesis requires θ to be greater than or less than θ_0, respectively. Then a UMPBT(γ) can be obtained by restricting the support of $\pi_1(\theta)$ to values of θ that belong to the set

$$(8) \qquad \arg\min_\theta uv g_\gamma(\theta, \theta_0).$$

PROOF. Consider the case in which the alternative hypothesis requires θ to be greater than θ_0 and $\eta(\theta)$ is increasing (so that $uv = 1$), and let θ_t denote the true (i.e., data-generating) parameter for **x** under (6). Consider first simple alternatives for which the prior on θ is a point mass at θ_1. Then

$$(9) \qquad \begin{aligned} \mathbf{P}_{\theta_t}(\mathrm{BF}_{10} > \gamma) &= \mathbf{P}_{\theta_t}\big[\log(\mathrm{BF}_{10}) > \log(\gamma)\big] \\ &= \mathbf{P}_{\theta_t}\left\{ \sum_{i=1}^n T(x_i) > \frac{\log(\gamma) + n[A(\theta_1) - A(\theta_0)]}{\eta(\theta_1) - \eta(\theta_0)} \right\}. \end{aligned}$$

It follows that the probability in (9) achieves its maximum value when the right-hand side of the inequality is minimized, regardless of the distribution of $\sum T(x_i)$.

Now consider composite alternative hypotheses, and define an indicator function s according to

$$(10) \qquad s(\mathbf{x}, \theta) = \mathrm{Ind}\left(\exp\left\{ [\eta(\theta) - \eta(\theta_0)] \sum_{i=1}^n T(x_i) - n[A(\theta) - A(\theta_0)] \right\} > \gamma \right).$$

Let θ^* be a value that minimizes $g_\gamma(\theta, \theta_0)$. Then it follows from (9) that

$$(11) \qquad s(\mathbf{x}, \theta) \le s(\mathbf{x}, \theta^*) \qquad \text{for all } \mathbf{x}.$$

This implies that

$$(12) \qquad \int_\Theta s(\mathbf{x}, \theta) \pi(\theta) \, d\theta \le s(\mathbf{x}, \theta^*)$$

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

$$(13) \qquad \mathbf{P}_{\theta_t}(\mathrm{BF}_{10} > \gamma) = \int_{\mathcal{X}} s(\mathbf{x}, \theta) f(\mathbf{x} \mid \theta_t) \, dx$$

is maximized by a prior that concentrates its mass on the set for which $g_\gamma(\theta, \theta_0)$ is minimized.

The proof for other values of (u, v) follows by noting that the direction of the inequality in (9) changes according to the sign of $\eta(\theta_1) - \eta(\theta_0)$. $\quad\square$

It should be noted that in some cases the values of θ that maximize $\mathbf{P}_{\theta_t}(\mathrm{BF}_{10} > \gamma)$ are not unique. This might happen if, for instance, no value of the sufficient statistic obtained from the experiment could produce a Bayes factor that exceeded the γ threshold. For example, it would not be possible to obtain a Bayes factor of 10 against a null hypothesis that a binomial success probability was 0.5 based on a sample of size $n = 1$. In that case, the probability of exceeding the threshold is 0 for all values of the success probability, and a unique UMPBT does not exist. More generally, if $T(x)$ is discrete, then many values of θ_1 might produce equivalent tests. An illustration of this phenomenon is provided in the first example.

2.1. *Large sample properties of UMPBTs.* Asymptotic properties of UMPBTs can most easily be examined for tests of point null hypotheses for a canonical parameter in one-parameter exponential families. Two properties of UMPBTs in this setting are described in the following lemma.

LEMMA 2. *Let $X_1, \ldots, X_n$ represent a random sample drawn from a density expressible in the form (6) with $\eta(\theta) = \theta$, and consider a test of the precise null hypothesis $H_0 : \theta = \theta_0$. Suppose that $A(\theta)$ has three bounded derivatives in a neighborhood of θ_0, and let θ^* denote a value of θ that defines a UMPBT(γ) test and satisfies*

$$(14) \qquad \frac{dg_\gamma(\theta^*, \theta_0)}{d\theta} = 0.$$

Then the following statements apply:

(1) *For some $t \in (\theta_0, \theta^*)$,*

$$(15) \qquad |\theta^* - \theta_0| = \sqrt{\frac{2\log(\gamma)}{nA''(t)}}.$$

(2) *Under the null hypothesis,*

$$(16) \qquad \log(\mathrm{BF}_{10}) \to N(-\log(\gamma), 2\log(\gamma)) \qquad as \ n \to \infty.$$

PROOF. The first statement follows immediately from (14) by expanding $A(\theta)$ in a Taylor series around θ^*. The second statement follows by noting that the weight of evidence can be expressed as

$$\log(\mathrm{BF}_{10}) = (\theta^* - \theta_0) \sum_{i=1}^{n} T(x_i) - n[A(\theta^*) - A(\theta_0)].$$

Expanding in a Taylor series around θ_0 leads to

$$(17) \quad \log(\mathrm{BF}_{10}) = \sqrt{\frac{2\log(\gamma)}{nA''(t)}} \left[\sum_{i=1}^{n} T(x_i) - nA'(\theta_0) - \frac{n}{2}A''(\theta_0)\sqrt{\frac{2\log(\gamma)}{nA''(t)}} \right] + \varepsilon,$$

where ε represents a term of order $O(n^{-1/2})$. From properties of exponential family models, it is known that

$$E_{\theta_0}[T(x_i)] = A'(\theta_0) \quad \text{and} \quad \mathrm{Var}_{\theta_0}(T(x_i)) = A''(\theta_0).$$

Because $A(\theta)$ has three bounded derivatives in a neighborhood of θ_0, $[A'(t) - A'(\theta_0)]$ and $[A''(t) - A''(\theta_0)]$ are order $O(n^{-1/2})$, and the statement follows by application of the central limit theorem. $\square$

Equation (15) shows that the difference $|\theta^* - \theta_0|$ is $O(n^{-1/2})$ when the evidence threshold γ is held constant as a function of n. In classical terms, this implies that alternative hypotheses defined by UMPBTs represent Pitman sequences of local alternatives [Pitman (1949)]. This fact, in conjunction with (16), exposes an interesting behavior of UMPBTs in large sample settings, particularly when viewed from the context of the Jeffreys–Lindley paradox [Jeffreys (1939), Lindley (1957); see also Robert, Chopin and Rousseau (2009)].

The Jeffreys–Lindley paradox (JLP) arises as an incongruity between Bayesian and classical hypothesis tests of a point null hypothesis. To understand the paradox, suppose that the prior distribution for a parameter of interest under the alternative hypothesis is uniform on an interval I containing the null value θ_0, and that the prior probability assigned to the null hypothesis is π_1. If π_1 is bounded away from 0, then it is possible for the null hypothesis to be rejected in an α-level significance test even when the posterior probability assigned to the null hypothesis exceeds $1 - \alpha$. Thus, the anomalous behavior exhibited in the example of Table 1, in which the null hypothesis was rejected in a significance test while being supported by the data, is characteristic of a more general phenomenon that may occur even in large sample settings. To see that the null hypothesis can be rejected even when the posterior odds are in its favor, note that for sufficiently large n the width of I will be large relative to the posterior standard deviation of θ under the alternative hypothesis. Data that are not "too far" from f_{θ_0} may therefore be much more likely to have arisen from the null hypothesis than from a density f_θ when θ is drawn uniformly from I. At the same time, the value of the test statistic based on the data may appear extreme given that f_{θ_0} pertains.

For moderate values of γ, the second statement in Lemma 2 shows that the weight of evidence obtained from a UMPBT is unlikely to provide strong evidence in favor of either hypothesis when the null hypothesis is true. When $\gamma = 4$, for instance, an approximate 95% confidence interval for the weight of evidence extends only between $(-4.65, 1.88)$, no matter how large n is. Thus, the posterior probability of the null hypothesis does not converge to 1 as the sample size grows. The null hypothesis is never fully accepted—nor the alternative rejected—when the evidence threshold is held constant as n increases.

This large sample behavior of UMPBTs with fixed evidence thresholds is, in a certain sense, similar to the JLP. When the null hypothesis is true and n is large, the probability of rejecting the null hypothesis at a fixed level of significance remains constant at the specified level of significance. For instance, the null hypothesis is rejected 5% of the time in a standard 5% significance test when the null hypothesis is true, regardless of how large the sample size is. Similarly, when $\gamma = 4$, the probability that the weight of evidence in favor of the alternative hypothesis will be greater than 0 converges to 0.20 as n becomes large. Like the significance test, there remains a nonzero probability that the alternative hypothesis will be favored by the UMPBT even when the null hypothesis is true, regardless of how large n is.

On a related note, Rousseau (2007) has demonstrated that a point null hypothesis may be used as a convenient mathematical approximation to interval hypotheses of the form $(\theta_0 - \varepsilon, \theta_0 + \varepsilon)$ if ε is sufficiently small. Her results suggest that such an approximation is valid only if $\varepsilon < o(n)$. The fact that UMPBT alternatives decrease at a rate of $O(n^{-1/2})$ suggests that UMPBTs may be used to test small interval hypotheses around θ_0, provided that the width of the interval satisfies the constraints provided by Rousseau.

Further comments regarding the asymptotic properties of UMPBTs appear in the discussion section.

3. Examples. Tests of simple hypotheses in one-parameter exponential family models continue to be the most common statistical hypothesis tests used by practitioners. These tests play a central role in many science, technology and business applications. In addition, the distributions of many test statistics are asymptotically distributed as standard normal deviates, which means that UMPBTs can be applied to obtain Bayes factors based on test statistics [Johnson (2005)]. This section illustrates the use of UMPBT tests in two archetypical examples; the first involves the test of a binomial success probability, and the second the test of the value of a parameter estimate that is assumed to be approximately normally distributed.

3.1. *Test of binomial success probability.* Suppose $x \sim \text{Bin}(n, p)$, and consider the test of a null hypothesis $H_0 : p = p_0$ versus an alternative hypothesis $H_1 : p > p_0$. Assume that an evidence threshold of γ is desired for the test; that is, the alternative hypothesis is accepted if $\text{BF}_{10} > \gamma$.

From Lemma 1, the UMPBT(γ) is defined by finding p_1 that satisfies $p_1 > p_0$ and

(18) $$p_1 = \arg\min_{p} \frac{\log(\gamma) - n[\log(1-p) - \log(1-p_0)]}{\log[p/(1-p)] - \log[p_0/(1-p_0)]}.$$

Although this equation cannot be solved in closed form, its solution can be found easily using optimization functions available in most statistical programs.

3.1.1. *Phase* II *clinical trials with binary outcomes.* To illustrate the resulting test in a real-world application that involves small sample sizes, consider a one-arm Phase II trial of a new drug intended to improve the response rate to a disease from the standard-of-care rate of $p_0 = 0.3$. Suppose also that budget and time constraints limit the number of patients that can be accrued in the trial to $n = 10$, and suppose that the new drug will be pursued only if the odds that it offers an improvement over the standard of care are at least 3:1. Taking $\gamma = 3$, it follows from (18) that the UMPBT alternative is defined by taking $H_1 : p_1 = 0.525$. At this value of p_1, the Bayes factor BF_{10} in favor of H_1 exceeds 3 whenever 6 or more of the 10 patients enrolled in the trial respond to treatment.

A plot of the probability that BF_{10} exceeds 3 as function of the true response rate p appears in Figure 1. For comparison, also plotted in this figure (dashed
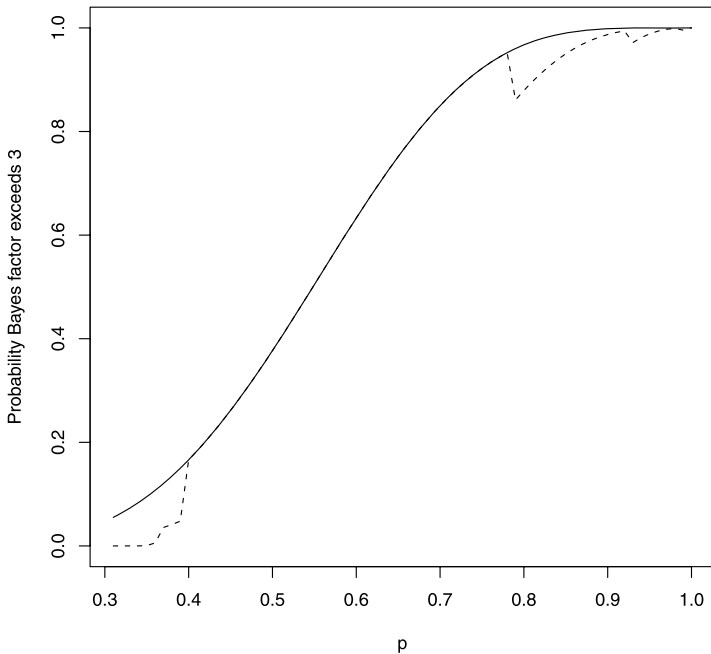

FIG. 1. *Probability that the Bayes factor exceeds* 3 *plotted against the data-generating parameter. The solid curve shows the probability of exceeding* 3 *for the UMPBT. The dashed curve displays this probability when the Bayes factor is calculated using the data-generating parameter.*

curve) is the probability that BF_{10} exceeds 3 when p_1 is set to the data-generating parameter, that is, when $p_1 = p_t$.

Figure 1 shows that the probability that BF_{10} exceeds 3 when calculated under the true alternative hypothesis is significantly smaller than it is under the UMPBT alternative for values of $p < 0.4$ and for values of $p > 0.78$. Indeed, for values of $p < 0.334$, there is no chance that BF_{10} will exceed 3. This is so because $(0.334/0.30)^x$ remains less than 3.0 for all $x \leq 10$. The decrease in the probability that the Bayes factor exceeds 3 for large values of p stems from the relatively small probability that these models assign to the observation of intermediate values of x. For example, when $p = 0.8$, the probability of observing 6 out 10 successes is only 0.088, while the corresponding probability under H_0 is 0.037. Thus $BF_{10} = 2.39$, and the evidence in favor of the true success probability does not exceed 3. That is, the discontinuity in the dashed curve at $p \approx 0.7$ occurs because the Bayes factor for this test is not greater than 3 when $x = 6$. Similarly, the other discontinuities in the dashed curve occur when the rejection region for the Bayesian test (i.e., values of x for which the Bayes factor is greater than 3) excludes another immediate value of x. The dashed and solid curves agree for all Bayesian tests that produce Bayes factors that exceed 3 for all values of $x \geq 6$.

It is also interesting to note that the solid curve depicted in Figure 1 represents the power curve for an approximate 5% one-sided significance test of the null hypothesis that $p = 0.3$ [note that $P_{0.3}(X \geq 6) = 0.047$]. This rejection region for the 5% significance test also corresponds to the region for which the Bayes factor corresponding to the UMPBT(γ) exceeds γ for all values of $\gamma \in (2.36, 6.82)$. If equal prior probabilities are assigned to H_0 and H_1, this suggests that a p-value of 0.05 for this test corresponds roughly to the assignment of a posterior probability between $(1.0/7.82, 1.0/3.36) = (0.13, 0.30)$ to the null hypothesis. This range of values for the posterior probability of the null hypothesis is in approximate agreement with values suggested by other authors, for example, Berger and Sellke (1987).

This example also indicates that a UMPBT can result in large type I errors if the threshold γ is chosen to be too small. For instance, taking $\gamma = 2$ in this example would lead to type I errors that were larger than 0.05.

It is important to note that the UMPBT does not provide a test that maximizes the expected weight of evidence, as equation (4) demonstrates. This point is illustrated in Figure 2, which depicts the expected weight of evidence obtained in favor of H_1 by a solid curve as the data-generating success probability is varied in (0.3, 1.0). For comparison, the dashed curve shows the expected weight of evidence obtained as a function of the true parameter value. As predicted by the inequality in (4), on average the UMPBT provides less evidence in favor of the true alternative hypothesis for all values of $p \in (0.3, 1.0)$ except $p = 0.525$, the UMPBT value.
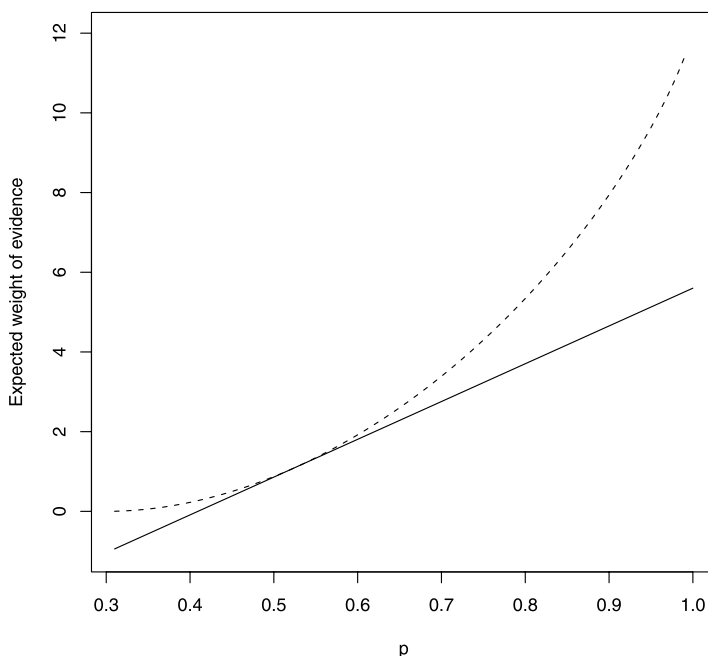
FIG. 2. *Expected weight of evidence produced by a* UMPBT(γ) *against a null hypothesis that* $p_0 = 0.3$ *when the sample size is* $n = 10$ *(solid curve), versus the expected weight of evidence observed using the data-generating success probability at the alternative hypothesis (dashed curve). The data-generating parameter value is displayed on the horizontal axis.*

3.2. *Test of normal mean,* σ^2 *known.* Suppose x_i, $i = 1, \ldots, n$ are i.i.d. $N(\mu, \sigma^2)$ with σ^2 known. The null hypothesis is $H_0 : \mu = \mu_0$, and the alternative hypothesis is accepted if $\text{BF}_{10} > \gamma$. Assuming that the alternative hypothesis takes the form $H_1 : \mu = \mu_1$ in a one-sided test, it follows that

$$\text{(19)} \qquad \log(\text{BF}_{10}) = \frac{n}{\sigma^2}\left[\bar{\mathbf{x}}(\mu_1 - \mu_0) + \frac{1}{2}(\mu_0^2 - \mu_1^2)\right].$$

If the data-generating parameter is μ_t, the probability that BF_{10} is greater than γ can be written as

$$\text{(20)} \qquad \mathbf{P}_{\mu_t}\left[(\mu_1 - \mu_0)\bar{\mathbf{x}} > \frac{\sigma^2 \log(\gamma)}{n} - \frac{1}{2}(\mu_0^2 - \mu_1^2)\right].$$

If $\mu_1 > \mu_0$, then the UMPBT(γ) value of μ_1 satisfies

$$\text{(21)} \qquad \underset{\mu_1}{\arg\min} \, \frac{\sigma^2 \log(\gamma)}{n(\mu_1 - \mu_0)} + \frac{1}{2}(\mu_0 + \mu_1).$$

Conversely, if $\mu_1 < \mu_0$, then optimal value of μ_1 satisfies

$$\text{(22)} \qquad \underset{\mu_1}{\arg\min} \, \frac{\sigma^2 \log(\gamma)}{n(\mu_1 - \mu_0)} + \frac{1}{2}(\mu_0 + \mu_1).$$
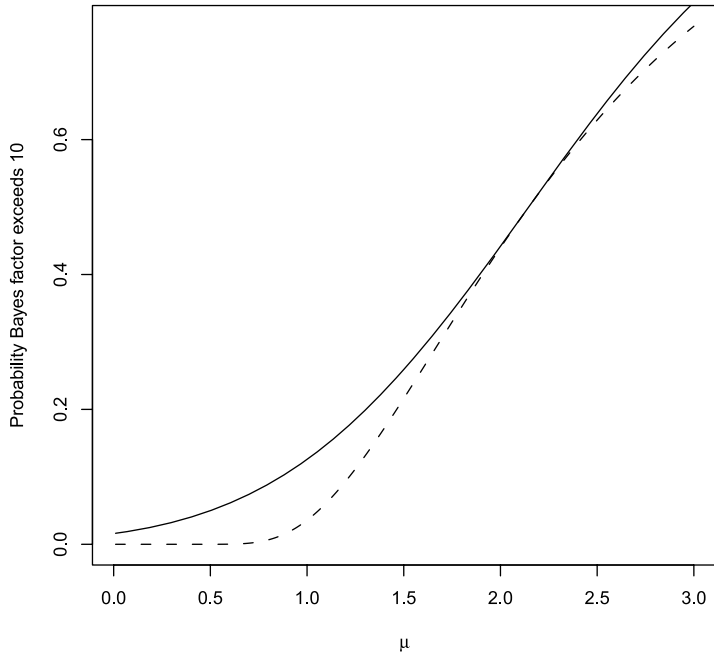
FIG. 3. *Probability that Bayes factor in favor of UMPBT alternative exceeds* 10 *when* $\mu_0 = 0$ *and* $n = 1$ (*solid curve*). *The dashed curve displays this probability when the Bayes factor is calculated under the alternative hypothesis that* μ_1 *equals the data-generating parameter* (*displayed on the horizontal axis*).

It follows that the UMPBT(γ) value for μ_1 is given by

$$(23) \qquad \mu_1 = \mu_0 \pm \sigma \sqrt{\frac{2 \log \gamma}{n}},$$

depending on whether $\mu_1 > \mu_0$ or $\mu_1 < \mu_0$.

Figure 3 depicts the probability that the Bayes factor exceeds $\gamma = 10$ when testing a null hypothesis that $\mu = 0$ based on a single, standard normal observation (i.e., $n = 1$, $\sigma^2 = 1$). In this case, the UMPBT(10) is obtained by taking $\mu_1 = 2.146$. For comparison, the probability that the Bayes factor exceeds 10 when the alternative is defined to be the data-generating parameter is depicted by the dashed curve in the plot.

UMPBTs can also be used to interpret the evidence obtained from classical UMPTs. In a classical one-sided test of a normal mean with known variance, the null hypothesis is rejected if

$$\bar{x} > \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}},$$

where α is the level of the test designed to detect $\mu_1 > \mu_0$. In the UMPBT, from (19)–(20) it follows that the null hypothesis is rejected if

$$\bar{x} > \frac{\sigma^2 \log(\gamma)}{n(\mu_1 - \mu_0)} + \frac{1}{2}(\mu_1 + \mu_0).$$

Setting $\mu_1 = \mu_0 + \sigma\sqrt{2\log(\gamma)/n}$ and equating the two rejection regions, it follows that the rejection regions for the two tests are identical if

$$(24) \qquad \gamma = \exp\left(\frac{z_\alpha^2}{2}\right).$$

For the case of normally distributed data, it follows that

$$(25) \qquad \mu_1 = \mu_0 + \frac{\sigma}{\sqrt{n}} z_\alpha,$$

which means that the alternative hypothesis places μ_1 at the boundary of the UMPT rejection region.

The close connection between the UMPBT and UMPT for a normal mean makes it relatively straightforward to examine the relationship between the p-values reported from a classical test and either the Bayes factors or posterior probabilities obtained from a Bayesian test. For example, significance tests for normal means are often conducted at the 5% level. Given this threshold of evidence for rejection of the null hypothesis, the one-sided γ threshold corresponding to the 5% significance test is 3.87, and the UMPBT alternative is $\mu_1 = \mu_0 + 1.645\sigma/\sqrt{n}$. If we assume that equal prior probabilities are assigned to the null and alternative hypotheses, then a correspondence between p-values and posterior probabilities assigned to the null hypothesis is easy to establish. This correspondence is depicted in Figure 4. For instance, this figure shows that a p-value of 0.01 corresponds to the assignment of posterior probability 0.08 to the null hypothesis.

3.2.1. *Evaluating evidence for the Higgs boson.* On July 4, 2012, scientists at CERN made the following announcement:

> We observe in our data clear signs of a new particle, at the level of 5 sigma, in the mass region around 126 gigaelectronvolts (GeV). (http://press.web.cern.ch/press/ PressReleases/Releases2012/PR17.12E.html).

In very simplified terms, the 5 sigma claim can be explained by fitting a model for a Poisson mean that had the following approximate form:

$$\mu(x) = \exp(a_0 + a_1 x + a_2 x^2) + s\phi(x; m, w).$$

Here, x denotes mass in GeV, $\{a_i\}$ denote nuisance parameters that model background events, s denotes signal above background, m denotes the mass of a new particle, w denotes a convolution parameter and $\phi(x; m, w)$ denotes a Gaussian density centered on m with standard deviation w [Prosper (2012)]. Poisson events
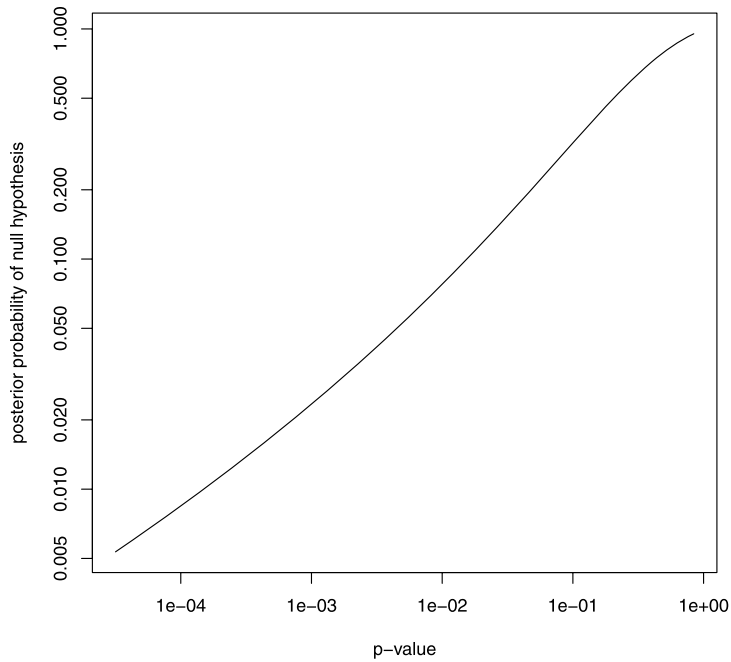
FIG. 4. *Correspondence between p-values and posterior model probabilities for a UMPBT test derived from a 5% test. This plot assumes equal prior probabilities were assigned to the null and alternative hypotheses. Note that both axes are displayed on the logarithmic scale.*

collected from a series of high energy experiments conducted in the Large Hadron Collider (LHC) at CERN provide the data to estimate the parameters in this stylized model. The background parameters $\{a_i\}$ are considered nuisance parameters. Interest, of course, focuses on testing whether $s > 0$ at a mass location m. The null hypothesis is that $s = 0$ for all m.

The accepted criterion for declaring the discovery of a new particle in the field of particle physics is the 5 sigma rule, which in this case requires that the estimate of s be 5 standard errors from 0 (http://public.web.cern.ch/public/).

Calculation of a Bayes factor based on the original mass spectrum data is complicated by the fact that prior distributions for the nuisance parameters $\{a_i\}$, m, and w are either not available or are not agreed upon. For this reason, it is more straightforward to compute a Bayes factor for these data based on the test statistic $z = \hat{s}/\text{se}(\hat{s})$ where $\hat{s}$ denotes the maximum likelihood estimate of s and $\text{se}(\hat{s})$ its standard error [Johnson (2005, 2008)]. To perform this test, assume that under the null hypothesis z has a standard normal distribution, and that under the alternative hypothesis z has a normal distribution with mean μ and variance 1.

In this context, the 5 sigma rule for declaring a new particle discovery means that a new discovery can only be declared if the test statistic $z > 5$. Using equation (24) to match the rejection region of the classical significance test to a UMPBT(γ)

implies that the corresponding evidence threshold is $\gamma = \exp(12.5) \approx 27{,}000$. In other words, a Bayes factor of approximately $\gamma = \exp(12.5) \approx 27{,}000$ corresponds to the 5 sigma rule required to accept the alternative hypothesis that a new particle has been found.

It follows from the discussion following equation (25) that the alternative hypothesis for the UMPBT alternative is $\mu_1 = 5$. This value is calculated under the assumption that the test statistic z has a standard normal distribution under the null hypothesis [i.e., $\sigma = 1$ and $n = 1$ in (23)]. If the observed value of z was exactly 5, then the Bayes factor in favor of a new particle would be approximately 27,000. If the observed value was, say 5.1, then the Bayes factor would be $\exp(-0.5[0.1^2 - 5.1^2]) = 44{,}000$. These values suggest very strong evidence in favor of a new particle, but perhaps not as much evidence as might be inferred by nonstatisticians by the report of a p-value of 3×10^{-7}.

There are, of course, a number of important caveats that should be considered when interpreting the outcome of this analysis. This analysis assumes that an experiment with a fixed endpoint was conducted, and that the UMPBT value of the Poisson rate at 126 GeV was of physical significance. Referring to (23) and noting that the asymptotic standard error of z decreases at rate $\sqrt{n}$, it follows that the UMPBT alternative hypothesis favored by this analysis is $O(n^{-1/2})$. For sufficiently large n, systematic errors in the estimation of the background rate could eventually lead to the rejection of the null hypothesis in favor of the hypothesis of a new particle. This is of particular concern if the high energy experiments were continued until a 5 sigma result was obtained. Further comments regarding this point appear in the discussion section.

3.3. *Other one-parameter exponential family models.* Table 2 provides the functions that must be minimized to obtain UMPBTs for a number of common exponential family models. The objective functions listed in this table correspond to the function $g_\gamma(\cdot, \cdot)$ specified in Lemma 1 with $v = 1$. The negative binomial is parameterized by the fixed number of failures r and random number of successes

TABLE 2
Common one parameter exponential family models for which UMPBT(γ) *exist*

Model	Test	Objective function
Binomial	$p_1 > p_0$	$\{\log(\gamma) - n\log[(1-p)/(1-p_0)]\}(\log\{[p(1-p_0)]/[(1-p)p_0]\})^{-1}$
Exponential	$\mu_1 > \mu_0$	$\{\log(\gamma) + n[\log(\mu_1) - \log(\mu_0)]\}[1/\mu_0 - 1/\mu_1]^{-1}$
Neg. Bin.	$p_1 > p_0$	$\{\log(\gamma) - r\log[(1-p_1)/(1-p_0)]\}[\log(p_1) - \log(p_0)]^{-1}$
Normal	$\sigma_1^2 > \sigma_0^2$	$\{2\sigma_1^2\sigma_0^2(\log(\gamma) + \frac{n}{2}[\log(\sigma_1^2) - \log(\sigma_0^2)])\}[\sigma_1^2 - \sigma_0^2]^{-1}$
Normal	$\mu_1 > \mu_0$	$[\sigma^2\log(\gamma)](\mu_1 - \mu_0)^{-1} + \frac{1}{2}(\mu_0 + \mu_1)$
Poisson	$\mu_1 > \mu_0$	$[\log(\gamma) + n(\mu_1 - \mu_0)][\log(\mu_1) - \log(\mu_0)]^{-1}$

$x = 0, 1, \ldots$ observed before termination of sampling. The other models are parameterized so that μ and p denote means and proportions, respectively, while σ^2 values refer to variances.

4. Extensions to other models. Like UMPTs, UMPBTs are most easily defined within one-parameter exponential family models. In unusual cases, UMPBTs can be defined for data modeled by vector-valued exponential family models, but in general such extensions appear to require stringent constraints on nuisance parameters.

One special case in which UMPBTs can be defined for a d-dimensional parameter $\boldsymbol{\theta}$ occurs when the density of an observation can be expressed as

$$
(26) \qquad f(x \mid \boldsymbol{\theta}) = h(x) \exp\left[\sum_{i=1}^{d} \eta_i(\boldsymbol{\theta}) T_i(x) - A(\boldsymbol{\theta})\right],
$$

and all but one of the $\eta_i(\boldsymbol{\theta})$ are constrained to have identical values under both hypotheses. To understand how a UMPBT can be defined in this case, without loss of generality suppose that $\eta_i(\boldsymbol{\theta})$, $i = 2, \ldots, d$ are constrained to have the same value under both the null and alternative hypotheses, and that the null and alternative hypotheses are defined by $H_0 : \theta_1 = \theta_{1,0}$ and $H_1 : \theta_1 > \theta_{1,0}$. For simplicity, suppose further that η_1 is a monotonically increasing function.

As in Lemma 1, consider first simple alternative hypotheses expressible as $H_1 : \theta_1 = \theta_{1,1}$. Let $\boldsymbol{\theta}_0 = (\theta_{1,0}, \ldots, \theta_{d,0})'$ and $\boldsymbol{\theta}_1 = (\theta_{1,1}, \ldots, \theta_{d,1})'$. It follows that the probability that the logarithm of the Bayes factor exceeds a threshold $\log(\gamma)$ can be expressed as

$$
\begin{aligned}
&\mathbf{P}\big[\log(\mathrm{BF}_{10}) > \log(\gamma)\big] \\
(27) \qquad &= \mathbf{P}\big\{\big[\eta_1(\theta_{1,1}) - \eta_1(\theta_{1,0})\big] T_1(\mathbf{x}) - \big[A(\boldsymbol{\theta}_1) - A(\boldsymbol{\theta}_0)\big] > \log(\gamma)\big\} \\
&= \mathbf{P}\bigg[T_1(\mathbf{x}) > \frac{\log(\gamma) + [A(\boldsymbol{\theta}_1) - A(\boldsymbol{\theta}_0)]}{[\eta_1(\theta_{1,1}) - \eta_1(\theta_{1,0})]}\bigg].
\end{aligned}
$$

The probability in (27) is maximized by minimizing the right-hand side of the inequality. The extension to composite alternative hypotheses follows the logic described in inequalities (11)–(13), which shows that UMPBT(γ) tests can be obtained in this setting by choosing the prior distribution of $\boldsymbol{\theta}_1$ under the alternative hypotheses so that it concentrates its mass on the set

$$
(28) \qquad \arg\min_{\theta} \frac{\log(\gamma) + [A(\boldsymbol{\theta}_1) - A(\boldsymbol{\theta}_0)]}{[\eta_1(\theta_{1,1}) - \eta_1(\theta_{1,0})]},
$$

while maintaining the constraint that the values of $\eta_i(\boldsymbol{\theta})$ are equal under both hypotheses. Similar constructions apply if η_1 is monotonically decreasing, or if the alternative hypothesis specifies that $\theta_{1,0} < \theta_{0,0}$.

More practically useful extensions of UMPBTs can be obtained when it is possible to integrate out nuisance parameters in order to obtain a marginal density

for the parameter of interest that falls within the class of exponential family of models. An important example of this type occurs in testing whether a regression coefficient in a linear model is zero.

4.1. *Test of linear regression coefficient, σ^2 known.* Suppose that

$$(29) \qquad \mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n),$$

where σ^2 is known, $\mathbf{y}$ is an $n \times 1$ observation vector, $\mathbf{X}$ an $n \times p$ design matrix of full column rank and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ denotes a $p \times 1$ regression parameter. The null hypothesis is defined as $H_0 : \beta_p = 0$. For concreteness, suppose that interest focuses on testing whether $\beta_p > 0$, and that under both the null and alternative hypotheses, the prior density on the first $p - 1$ components of $\boldsymbol{\beta}$ is a multivariate normal distribution with mean vector $\mathbf{0}$ and covariance matrix $\sigma^2 \boldsymbol{\Sigma}$. Then the marginal density of $\mathbf{y}$ under H_0 is

$$(30) \qquad m_0(\mathbf{y}) = (2\pi\sigma^2)^{-n/2} |\boldsymbol{\Sigma}|^{-1/2} |\mathbf{F}|^{-1/2} \exp\left(-\frac{R}{2\sigma^2}\right),$$

where

$$(31) \quad \mathbf{F} = \mathbf{X}'_{-p}\mathbf{X}_{-p} + \boldsymbol{\Sigma}^{-1}, \qquad \mathbf{H} = \mathbf{X}_{-p}\mathbf{F}^{-1}\mathbf{X}'_{-p}, \qquad R = \mathbf{y}'(\mathbf{I}_n - \mathbf{H})'\mathbf{y},$$

and $\mathbf{X}_{-p}$ is the matrix consisting of the first $p - 1$ columns of $\mathbf{X}$.

Let β_{p*} denote the value of β_p under the alternative hypothesis H_1 that defines the UMPBT(γ), and let $\mathbf{x}_p$ denote the pth column of $\mathbf{X}$. Then the marginal density of $\mathbf{y}$ under H_1 is

$$(32) \quad m_1(\mathbf{y}) = m_0(\mathbf{y}) \times \exp\left\{-\frac{1}{2\sigma^2}[\beta_{p*}^2\mathbf{x}'_p(\mathbf{I}_n - \mathbf{H})\mathbf{x}_p - 2\beta_{p*}\mathbf{x}'_p(\mathbf{I}_n - \mathbf{H})\mathbf{y}]\right\}.$$

It follows that the probability that the Bayes factor BF_{10} exceeds γ can be expressed as

$$(33) \qquad \mathbf{P}\left[\mathbf{x}'_p(\mathbf{I}_n - \mathbf{H})\mathbf{y} > \frac{\sigma^2 \log(\gamma)}{\beta_{p*}} + \frac{1}{2}\beta_{p*}\mathbf{x}'_p(\mathbf{I}_n - \mathbf{H})\mathbf{x}_p\right],$$

which is maximized by minimizing the right-hand side of the inequality. The UMPBT(γ) is thus obtained by taking

$$(34) \qquad \beta_{p*} = \sqrt{\frac{2\sigma^2 \log(\gamma)}{\mathbf{x}'_p(\mathbf{I}_n - \mathbf{H})\mathbf{x}_p}}.$$

The corresponding one-sided test of $\beta_p < 0$ is obtained by reversing the sign of β_{p*} in (34).

Because this expression for the UMPBT assumes that σ^2 is known, it is not of great practical significance by itself. However, this result may guide the specification of alternative models in, for example, model selection algorithms in which the

priors on regression coefficients are specified conditionally on the value of σ^2. For example, the mode of the nonlocal priors described in Johnson and Rossell (2012) might be set to the UMPBT values after determining an appropriate value of γ based on both the sample size n and number of potential covariates p.

5. Approximations to UMPBTs using data-dependent alternatives. In some situations—most notably in linear models with unknown variances—data dependent alternative hypotheses can be defined to obtain tests that are approximately uniformly most powerful in maximizing the probability that a Bayes factor exceeds a threshold. This strategy is only attractive when the statistics used to define the alternative hypothesis are ancillary to the parameter of interest.

5.1. *Test of normal mean, σ^2 unknown.* Suppose that x_i, $i = 1, \ldots, n$, are i.i.d. $N(\mu, \sigma^2)$, that σ^2 is unknown and that the null hypothesis is $H_0 : \mu = \mu_0$. For convenience, assume further that the prior distribution on σ^2 is an inverse gamma distribution with parameters α and λ under both the null and alternative hypotheses.

To obtain an approximate UMPBT(γ), first marginalize over σ^2 in both models. Noting that $(1 + a/t)^t \to e^a$, it follows that the Bayes factor in favor of the alternative hypothesis satisfies

$$(35) \qquad \mathrm{BF}_{10}(\mathbf{x}) = \left[\frac{\sum_{i=1}^{n}(x_i - \mu_0)^2 + 2\lambda}{\sum_{i=1}^{n}(x_i - \mu_1)^2 + 2\lambda} \right]^{n/2+\alpha}$$

$$(36) \qquad \approx \left[\frac{1 + (\bar{\mathbf{x}} - \mu_0)^2/s^2}{1 + (\bar{\mathbf{x}} - \mu_1)^2/s^2} \right]^{n/2+\alpha}$$

$$(37) \qquad \approx \exp\left\{ -\frac{n}{2s^2}[(\bar{\mathbf{x}} - \mu_1)^2 - (\bar{\mathbf{x}} - \mu_0)^2] \right\},$$

where

$$(38) \qquad s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{\mathbf{x}})^2 + 2\lambda}{n + 2\alpha}.$$

The expression for the Bayes factor in (37) reduces to (19) if σ^2 is replaced by s^2. This implies that an approximate, but data-dependent UMPBT alternative hypothesis can be specified by taking

$$(39) \qquad \mu_1 = \mu_0 \pm s\sqrt{\frac{2 \log \gamma}{n}},$$

depending on whether $\mu_1 > \mu_0$ or $\mu_1 < \mu_0$.

Figure 5 depicts the probability that the Bayes factor exceeds $\gamma = 10$ when testing a null hypothesis that $\mu = 0$ based on an independent sample of size $n = 30$ normal observations with unit variance ($\sigma^2 = 1$) and using (39) to set the value of μ_1 under the alternative hypothesis. For comparison, the probability that the
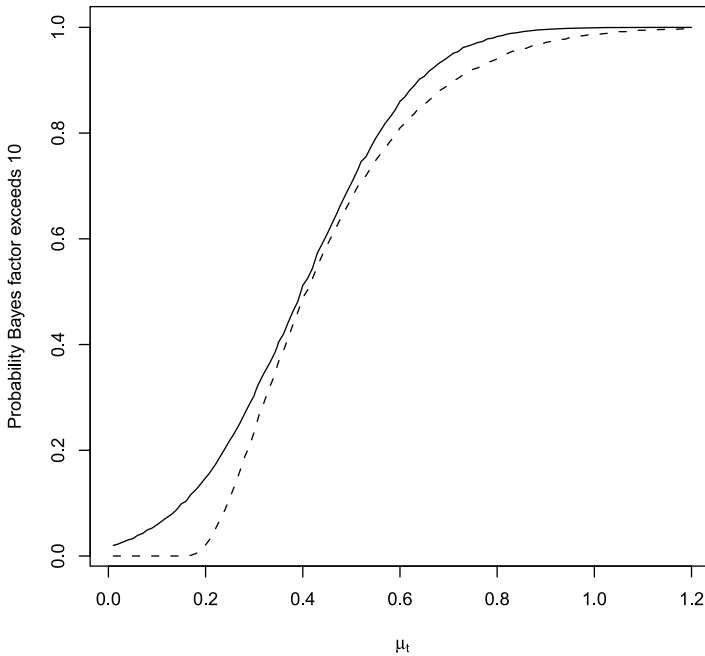
FIG. 5. *Probability that Bayes factor based on data-dependent, approximate UMPBT alternative exceeds 10 when $\mu_0 = 0$ and $n = 30$ (solid curve). The dashed curve displays this probability when the Bayes factor is calculated under the alternative hypothesis that μ_1 equals data-generating parameter (displayed on the horizontal axis) and $\sigma^2 = 1$ (the true value).*

Bayes factor exceeds 10 when the alternative is defined by taking $\sigma^2 = 1$ and μ_1 to be the data-generating parameter is depicted by the dashed curve in the plot. Interestingly, the data-dependent, approximate UMPBT(10) provides a higher probability of producing a Bayes factor that exceeds 10 than do alternatives fixed at the data generating parameters.

5.2. *Test of linear regression coefficient, σ^2 unknown.* As final example, suppose that the sampling model of Section 4.1 holds, but assume now that the observational variance σ^2 is unknown and assumed under both hypotheses to be drawn from an inverse gamma distribution with parameters α and λ. Also assume that the prior distribution for the first $p - 1$ components of $\boldsymbol{\beta}$, given σ^2, is a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\sigma^2 \Sigma$. As before, assume that $H_0 : \beta_p = 0$. Our goal is to determine a value β_{p*} so that $H_1 : \beta_p = \beta_{p*}$ is the UMPBT(γ) under the constraint that $\beta_p > 0$.

Define $\mathbf{y}_1 = \mathbf{y} - \mathbf{x}_p \beta_{p*}$ and let $\mathbf{y}_0 = \mathbf{y}$. By integrating with respect to the prior densities on σ^2 and the first $p - 1$ components of $\boldsymbol{\beta}$, the marginal density of the

data under hypothesis i, $i = 0, 1$ can be expressed as

$$(40) \qquad m_i(\mathbf{y}) = 2^\alpha \pi^{-n/2} |\boldsymbol{\Sigma}|^{-1/2} \frac{\lambda^\alpha}{\Gamma(\alpha)} \Gamma(n/2 + \alpha) |\mathbf{F}|^{-1/2} R_i^{-n/2-\alpha},$$

where $\mathbf{F}$ is defined in (31), and

$$(41) \qquad R_i = \mathbf{y}_i'(\mathbf{I}_n - \mathbf{H})\mathbf{y}_i + 2\lambda.$$

It follows that the Bayes factor in favor of H_1 can be written as

$$(42) \qquad \mathrm{BF}_{10} = \left[1 + \frac{\beta_{p*}^2 \mathbf{x}_p'(\mathbf{I}_n - \mathbf{H})\mathbf{x}_p - 2\beta_{p*}\mathbf{x}_p'(\mathbf{I}_n - \mathbf{H})\mathbf{y}}{R_0}\right]^{-n/2-\alpha}$$

$$(43) \qquad \approx \exp\left\{-\frac{1}{2s_p^2}\left[\beta_{p*}^2 \mathbf{x}_p'(\mathbf{I}_n - \mathbf{H})\mathbf{x}_p - 2\beta_{p*}\mathbf{x}_p'(\mathbf{I}_n - \mathbf{H})\mathbf{y}\right]\right\},$$

where

$$(44) \qquad s_p^2 = \frac{R_0}{n + 2\alpha}.$$

The UMPBT(γ) is defined from (43) according to

$$(45) \qquad \mathbf{P}(\mathrm{BF}_{10} > \gamma) = \mathbf{P}\left[\mathbf{x}_p'(\mathbf{I}_n - \mathbf{H})\mathbf{y} > \frac{s_p^2 \log(\gamma)}{\beta_{p*}} + \frac{1}{2}\beta_{p*}\mathbf{x}_p'(\mathbf{I}_n - \mathbf{H})\mathbf{x}_p\right].$$

Minimizing the right-hand side of the last inequality with respect to β_{p*} results in

$$(46) \qquad \beta_{p*} = \sqrt{\frac{2s_p^2 \log(\gamma)}{\mathbf{x}_p'(\mathbf{I}_n - \mathbf{H})\mathbf{x}_p}}.$$

This expression is consistent with the result obtained in the known variance case, but with s_p^2 substituted for σ^2.

6. Discussion. The major contributions of this paper are the definition of UMPBTs and the explicit description of UMPBTs for regular one-parameter exponential family models. The existence of UMPBTs for exponential family models is important because these tests represent the most common hypothesis tests conducted by practitioners. The availability of UMPBTs for these models means that these tests can be used to interpret test results in terms of Bayes factors and posterior model probabilities in a wide range of scientific settings. The utility of these tests is further enhanced by the connection between UMPBTs and UMPTs that have the same rejection region. This connection makes it trivial to simultaneously report both the p-value from a test and the corresponding Bayes factor.

The simultaneous report of default Bayes factors and p-values may play a pivotal role in dispelling the perception held by many scientists that a p-value of 0.05 corresponds to "significant" evidence against the null hypothesis. The preceding

sections contain examples in which this level of significance favors the alternative hypothesis by odds of only 3 or 4 to 1. Because few researchers would regard such odds as strong evidence in favor of a new theory, the use of UMPBTs and the report of Bayes factors based upon them may lead to more realistic interpretations of evidence obtained from scientific studies.

The large sample properties of UMPBTs described in Section 2.1 deserve further comment. From Lemma 2, it follows that the expected weight of evidence in favor of a true null hypothesis in an exponential family model converges to $\log(\gamma)$ as the sample size n tends to infinity. In other words, the evidence threshold γ represents an approximate bound on the evidence that can be collected in favor of the null hypothesis. This implies that γ must be increased with n in order to obtain a consistent sequence of tests.

Several criteria might be used for selecting a value for γ in large sample settings. One criterion can be inferred from the first statement of Lemma 2, where it is shown that the difference between the tested parameter's value under the null and alternative hypotheses is proportional to $[\log(\gamma)/n]^{1/2}$. For this difference to be a constant—as it would be in a subjective Bayesian test—$\log(\gamma)$ must be proportional to n, or $\gamma = \exp(cn)$ for some $c > 0$. This suggests that an appropriate value for c might be determined by calibrating the weight of evidence against an accepted threshold/sample size combination. For example, if an evidence threshold of 4 were accepted as the standard threshold for tests conducted with a sample size of 100, then c might be set to $\log(4)/100 = 0.0139$. This value of c leads to an evidence threshold of $\gamma = 16$ for sample sizes of 200, a threshold of 64 for sample sizes of 300, etc. From (24), the significance levels for corresponding z-tests would be 5%, 1% and 0.2%, respectively.

The requirement to increase γ to achieve consistent tests in large samples also provides insight into the performance of standard frequentist and subjective Bayesian tests in large sample settings. The exponential growth rate of γ required to maintain a fixed alternative hypothesis suggests that the weight of evidence should be considered against the backdrop of sample size, even in Bayesian tests. This is particularly important in goodness-of-fit testing where small deviations from a model may be tolerable. In such settings, even moderately large Bayes factors against the null hypotheses may not be scientifically important when they are based on very large sample sizes.

From a frequentist perspective, the use of UMPBTs in large sample settings can provide insight into the deviations from null hypotheses when they are (inevitably) detected. For instance, suppose that a one-sided 1% test has been conducted to determine if the mean of normal data is 0, and that the test is rejected with a p-value of 0.001 based on a sample size of 10,000. From (24), the implied evidence threshold for the test is $\gamma = 15$, and the alternative hypothesis that has been implicitly tested with the UMPBT is that $\mu = 0.023\sigma$. Based on the observation of $\bar{x} = 0.031\sigma$, the Bayes factor in favor of this alternative is 88.5. Although there are strong odds against the null, the scientific importance of this outcome may be

tempered by the fact that the alternative hypothesis that was supported against the null represents a standardized effect size of only 2.3%.

This article has focused on the specification of UMPBTs for one-sided alternatives. A simple extension of these tests to two-sided alternatives can be obtained by assuming that the alternative hypothesis is represented by two equally-weighted point masses located at the UMPBT values determined for one-sided tests. The Bayes factors for such tests can be written as

$$(47) \qquad \mathbf{P}\left[\frac{0.5m_l(\mathbf{x}) + 0.5m_h(\mathbf{x})}{m_0(\mathbf{x})} > \gamma\right],$$

where m_l and m_h denote marginal densities corresponding to one-sided UMPBTs. Letting $m_*(\mathbf{x}) = \max(m_l(\mathbf{x}), m_h(\mathbf{x}))$ for the data actually observed, and assuming that the favored marginal density dominates the other, it follows that

$$(48) \qquad \mathbf{P}\left[\frac{0.5m_l(\mathbf{x}) + 0.5m_h(\mathbf{x})}{m_0(\mathbf{x})} > \gamma\right] \approx \mathbf{P}\left[\frac{m_*(\mathbf{x})}{m_0(\mathbf{x})} > 2\gamma\right].$$

Thus, an approximate two-sided UMPBT(γ) can be defined by specifying an alternative hypothesis that equally concentrates its mass on the two one-sided UMPBT(2γ) tests.

Additional research is needed to identify classes of models and testing contexts for which UMPBTs can be defined. The UMPBTs described in this article primarily involve tests of point null hypotheses, or tests that can be reduced to a test of a point null hypothesis after marginalizing over nuisance parameters. Whether UMPBTs can be defined in more general settings remains an open question.

REFERENCES

BERGER, J. (2006). The case for objective Bayesian analysis. *Bayesian Anal*. **1** 385–402. MR2221271

BERGER, J. O. and PERICCHI, L. R. (1996). The intrinsic Bayes factor for model selection and prediction. *J. Amer. Statist. Assoc.* **91** 109–122. MR1394065

BERGER, J. O. and SELLKE, T. (1987). Testing a point null hypothesis: Irreconcilability of *P* values and evidence. *J. Amer. Statist. Assoc.* **82** 112–122.

BERGER, J. O. and WOLPERT, R. L. (1984). *The Likelihood Principle. Institute of Mathematical Statistics Lecture Notes—Monograph Series* **6**. IMS, Hayward, CA. MR0773665

EDWARDS, W., LINDMAN, H. and SAVAGE, L. (1963). Bayesian statistical inference for psychological research. *Psychological Review* **70** 193–242.

HOWSON, C. and URBACH, P. (2005). *Scientific Reasoning*: *The Bayesian Approach*, 3rd ed. Open Court, Chicago, IL.

JEFFREYS, H. (1939). *Theory of Probability*. Cambridge Univ. Press, Cambridge.

JOHNSON, V. E. (2005). Bayes factors based on test statistics. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **67** 689–701. MR2210687

JOHNSON, V. E. (2008). Properties of Bayes factors based on test statistics. *Scand. J. Stat.* **35** 354–368. MR2418746

JOHNSON, V. E. and ROSSELL, D. (2010). On the use of non-local prior densities in Bayesian hypothesis tests. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **72** 143–170. MR2830762

JOHNSON, V. E. and ROSSELL, D. (2012). Bayesian model selection in high-dimensional settings. *J. Amer. Statist. Assoc.* **107** 649–660. MR2980074

LEHMANN, E. L. and ROMANO, J. P. (2005). *Testing Statistical Hypotheses*, 3rd ed. Springer, New York. MR2135927

LINDLEY, D. (1957). A statistical paradox. *Biometrika* **44** 187–192.

MAYO, D. G. and SPANOS, A. (2006). Severe testing as a basic concept in a Neyman–Pearson philosophy of induction. *British J. Philos. Sci.* **57** 323–357. MR2249183

NEYMAN, J. and PEARSON, E. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference. *Biometrika* **20A** 175–240.

NEYMAN, J. and PEARSON, E. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **231** 289–337.

O'HAGAN, A. (1995). Fractional Bayes factors for model comparison. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **57** 99–118.

PITMAN, E. (1949). *Lecture Notes on Nonparametric Statistical Inference*. Columbia Univ., New York.

POPPER, K. R. (1959). *The Logic of Scientific Discovery*. Hutchinson, London. MR0107593

PROSPER (2012). Personal communication to news@bayesian.org.

ROBERT, C. P., CHOPIN, N. and ROUSSEAU, J. (2009). Harold Jeffreys's theory of probability revisited. *Statist. Sci.* **24** 141–172. MR2655841

ROUSSEAU, J. (2007). Approximating interval hypothesis: *p*-values and Bayes factors. In *Proceedings of the* 2006 *Valencia Conference* (J. Bernardo, M. Bayarri, J. Berger, A. Dawid, D. Heckerman, A. Smith and M. West, eds.) 1–27. Oxford Univ. Press, Oxford.

DEPARTMENT OF STATISTICS
TEXAS A&M UNIVERSITY
3143 TAMU
COLLEGE STATION, TEXAS 77843-3143
USA
E-MAIL: vjohnson@stat.tamu.edu