Calibration of p Values for Testing Precise Null Hypotheses

Thomas SELLKE, M. J. BAYARRI, and James O. BERGER

P values are the most commonly used tool to measure evidence against a hypothesis or hypothesized model. Unfortunately, they are often incorrectly viewed as an error probability for rejection of the hypothesis or, even worse, as the posterior probability that the hypothesis is true. The fact that these interpretations can be completely misleading when testing precise hypotheses is first reviewed, through consideration of two revealing simulations. Then two calibrations of a p value are developed, the first being interpretable as odds and the second as either a (conditional) frequentist error probability or as the posterior probability of the hypothesis.

KEY WORDS: Bayes factors; Bayesian robustness; Conditional frequentist error probabilities; Odds.

1. INTRODUCTION

In statistical analysis of data $\mathbf{X}$, one is frequently working, at a given moment, with an entertained model or hypothesis $\mathrm{H}_0 : \mathbf{X} \sim f(\mathbf{x})$; here we will consider the case where $f(\mathbf{x})$ is a continuous density. A statistic $T(\mathbf{X})$ is chosen to investigate compatibility of the model with the observed data $\mathbf{x}_{\text{obs}}$, with large values of T indicating less compatibility. The p value is then defined as

$$p = \Pr(T(\mathbf{X}) \geq T(\mathbf{x}_{\text{obs}})). \tag{1}$$

In this article, we assume that $f(\mathbf{x})$ is completely specified, so that the probability computation in (1) is under H_0. The null hypothesis is thus a "precise" hypothesis, as opposed to, say, the hypothesis that a treatment mean is less than zero. The results herein apply primarily to such precise hypotheses; see Casella and Berger (1987) and Berger and Mortera (1999) for discussion of the one-sided testing situation.

Often, of course, the density in H_0 will contain nuisance parameters, in which case computation of a p value can be considerably more involved. For review and discussion of appropriate ways to define a p value in this situation, see Bayarri and Berger (1999, 2000). The focus therein is in developing p values that are valid, in the sense of having a uniform distribution under H_0. The calibration of p values that is discussed in this article depends *only* on having a p value that is valid, so the restriction to a point null hypothesis is only done here for pedagogical reasons. Note, also, that alternative hypotheses, H_1, will be introduced as we proceed but alternatives play only a secondary role in the analysis since, in a sense, we will "optimize" over all reasonable alternatives.

The difficulty in interpretation of p values has been highlighted in many articles, among them Edwards, Lindman, and Savage (1963), Gibbons and Pratt (1975), Berger and Sellke (1987), Berger and Delampady (1987), Delampady and Berger (1990) (which specifically considers the problem of testing fit when $T(\mathbf{X})$ is chosen to be the usual chi-squared statistic for fit), and Schervish (1996); and has even reached the popular press (Matthews 1998).

A focus of these works (and the focus of this article) is on what could be termed the "p value fallacy," by which we mean the misinterpretation of a p value as either a direct frequentist error rate, the probability that the hypothesis is true in light of the data, or a measure of odds of H_0 to H_1. [The term "p value fallacy" was used, in the first of these senses, in the excellent articles Goodman (1999a,b).] Although standard textbooks typically warn against such interpretations, the warnings often go unheeded. Part of the purpose of this article is to provide simple examples (in Section 2) illustrating the p value fallacy, examples that are easy to use in even elementary courses so as to reinforce the verbal warnings against misinterpretation of p values.

Unfortunately, even direct illustrations of the p value fallacy are likely to have only a limited effect, unless students are also presented with suitable alternatives. In Section 3 we discuss two such alternatives that can be viewed as methods of calibrating p values so that they can be interpreted in either a Bayesian or a frequentist way. The calibrations are quite easy to state: for the Bayesian calibration, simply compute

$$B(p) = -e\,p\,\log(p)\,, \tag{2}$$

when $p < 1/e$, and interpret this as a lower bound on the odds provided by the data (or Bayes factor) for H_0 to H_1. (The final odds of H_0 to H_1 are found by multiplying the Bayes factor by the prior odds of H_0 to H_1.) In terms of frequentist Type I error probability α (in rejecting H_0), the calibration is

$$\alpha(p) = (1 + [-e\,p\log(p)]^{-1})^{-1}. \tag{3}$$

Although motivated by a pure frequentist argument in Section 3.1.2, this latter expression also has a Bayesian interpretation; it is the posterior probability of H_0 that arises

Thomas Sellke is Professor, Statistics Department, Purdue University, West Lafayette, IN 47907-1339. M. J. Bayarri is Professor, Department of Statistics and Operations Research, University of Valencia, Burjassot, Valencia 46100, Spain. James O. Berger is Arts and Sciences Professor, Institute of Statistics and Decision Sciences, Duke University, Durham, NC 27708-0251 (E-mail: berger@stat.duke.edu). This work was supported, in part, by the National Science Foundation (USA) under Grants DMS-9303556, DMS-9802261, and DMS-9971767, and by the Ministry of Education and Culture (Spain) under Grant PB96-0776. The authors are grateful to Lawrence Brown and Rui Paulo for helpful discussions, and to the Associate Editor and a referee for suggestions that considerably improved the article.

Table 1. Calibration of p Values as Odds (Bayes factors) and Conditional Error Probabilities

p	.2	.1	.05	.01	.005	.001
$B(p)$	.870	.625	.407	.125	.072	.0188
$\alpha(p)$	.465	.385	.289	.111	.067	.0184

from use of the Bayes factor in (2) together with the assumption that H_0 and H_1 have equal prior probabilities of $1/2$. Thus, use of (3) has the additional pedagogical advantage that one need not fear misinterpretation of a frequentist error probability as the probability that the hypothesis is true; here, they coincide.

Table 1 presents various p values and their associated calibrations. Thus, $p = .05$ translates into odds $B(.05) = .407$ (roughly 1 to 2.5) of H_0 to H_1, and frequentist error probability $\alpha(.05) = .289$ in rejecting H_0. (The default posterior probability of H_0 would also be .289.) Clearly $p = .05$ does not indicate particularly strong evidence against H_0. Even $p = .01$ corresponds to only about 8 to 1 odds against H_0. These calibrations will be formally motivated in Section 3, from a variety of perspectives.

2. ILLUSTRATIONS OF THE P VALUE FALLACY

In this section, we present an extended example that illustrates the p value fallacy. The example is presented in terms of a simulation, for two reasons. First, it is then accessible to even beginning statistics students, and can be used in introductory classes to convey the meaning of p values. Second, the use of simulation emphasizes the frequentist nature of these issues; we are not discussing a conflict between frequentist and Bayesian reasoning, but are exhibiting a fundamental property of p values that is apparent from any perspective.

Consider the situation in which experimental drugs $D_1, D_2, D_3, \ldots$ are to be tested. The drugs can be for the same illness (say, AIDS, common cold, etc.) or different illnesses. Each test will be thought of as completely independent; we simply have a series of tests so that we can explore the frequentist properties of p values. In each test, the following hypotheses are to be tested:

$$H_0 : D_i \text{ has negligible effect} \quad \text{versus} \quad H_1 : D_i \text{ has a non-negligible effect} . \quad (4)$$

Note that the null hypotheses, H_0, have special plausibility in these tests; many experimental drugs that are tested have "negligible effect," so that these null hypotheses could reasonably be true. [This is related to the earlier comment that we are only concerned with the testing of "precise" hypotheses. See Berger, Boukai, and Wang (1997) for further discussion.]

Suppose that one of these tests results in a p value $\approx .05$ (or $\approx .01$). The question we consider is: How strong is the evidence that the drug in question has a non-negligible effect? To study this, we will simply collect all the p values from a large number of such tests, and record how often the null hypothesis is true for p values at various levels. For instance, Table 2 shows hypothetical output from the first 12 tests. Suppose we focus on those tests, in a long series of tests, for which $p \approx .05$ (D_2 and D_8 in Table 2) or $p \approx .01$ (D_5 and D_{10} in Table 2), and ask: What proportion of these tests have true H_0; that is, ineffective drugs?

We shortly discuss the simulation to answer this question, but here is the basic and surprising conclusion for normal testing, first established (theoretically) by Berger and Sellke (1987). Suppose it is known, a priori, that about 50% of the drugs tested have a negligible effect. (We shortly consider the more general case.) Then:

1. Of the D_i for which the p value $\approx .05$, *at least* 23% (and typically close to 50%) will have negligible effect.
2. Of the D_i for which the p value $\approx .01$, *at least* 7% (and typically close to 15%) will have negligible effect.

Similar results arise for other initial proportions of ineffective drugs. Indeed, suppose that the initial proportion of ineffective drugs in the simulation is π_0. Then, among all those tests for which $p \approx .05$, a *lower bound* (derived by Berger and Sellke 1987) on the proportion of true nulls is given in Figure 1. For instance, if the initial proportion of true nulls is about $1/3$ (2/3), then the proportion of true nulls among those tests for which $p \approx .05$, is *at least* 12% (35%), and is typically (i.e., for most simulations) much larger.

The simulation we consider to represent this situation supposes that each test in (4) is based on normal data (known variance), with θ_j being the treatment mean for D_j, so that (4) is the test of $H_0 : \theta_j = 0$ versus $H_1 : \theta_j \neq 0$. One must choose π_0, the initial proportion of null hypotheses that are true, and also the values of θ_j under the alternative hypotheses. For each hypothesis, one then generates normal data with mean θ_j, and computes the corresponding p value, defined for the usual test statistic, $T(\mathbf{X}) = \sqrt{n_j}\,|\overline{X}_j|/\sigma_j$, as

$$p = 2\left[1 - \Phi(T(\mathbf{x}_{\text{obs}}))\right]; \quad (5)$$

here n_j, σ_j, and $\overline{X}_j$ are the sample size, standard deviation, and sample mean corresponding to the test of D_j, and Φ is the standard normal cdf. After doing this for a large series of tests, one looks at the subset of p values which are near a specified value, such as .05. For instance, one can look at those tests for which $.049 \leq p \leq .050$. (Any small interval near $p = .05$ would yield essentially the same answer.) One then simply notes the proportion of such tests for which H_0 is true. An applet that performs this simulation can be found at http:// www.stat.duke.edu/ ~berger/ p-values.html. The Web site also discusses numerous further details, such as choice of the alternatives θ_j. (Note that the lower bounds discussed above, and given in Figure 1, are true for *any*

Table 2. P Values Corresponding to Testing Whether Drug D_i has Negligible Effect

Drug	D1	D2	D3	D4	D5	D6
p value	.41	**.049**	.32	.94	**.01**	.28
Drug	D7	D8	D9	D10	D11	D12
p value	.11	**.05**	.65	**.009**	.09	.66

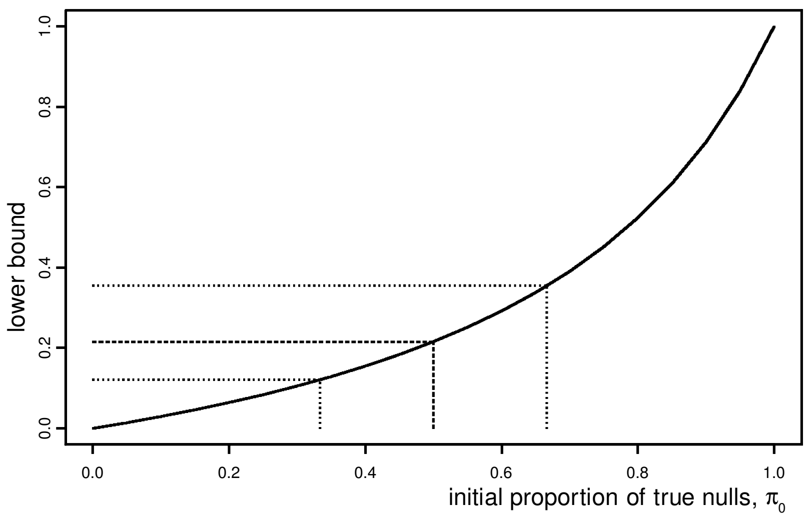

Figure 1. Lower Bound on the Proportion of True Nulls Among Those Tests for Which the p Value is Close to .05.

choice of alternatives, and most choices of alternatives will give answers substantially *higher* than these lower bounds.)

A large number of variants of this simulation could be performed. Having normal data is not crucial; the results would be qualitatively similar under most standard distributional assumptions. [See Berger and Sellke (1987) for some exceptions.] Likewise, the results would not qualitatively change if the null hypotheses were replaced by small interval nulls of the form $\mathrm{H}_0 : |\theta_j| < \epsilon$, providing $\epsilon < \sigma_j/(4\sqrt{n_j})$. This is important because hypotheses such as $\mathrm{H}_0 : \theta_j = 0$ are unlikely to ever be true exactly. (D_j will probably have *some* effect, even if only $\theta_j = 10^{-8}$.) Indeed, the hypothesis $\mathrm{H}_0 : \theta_j = 0$ should really just be thought of as an approximation to a small interval null, and Berger and Delampady (1987) showed that it is a good approximation if $\epsilon < \sigma_j/(4\sqrt{n_j})$. Thus, in practice, one must make the judgment that this condition will hold before formulating the test as that of $\mathrm{H}_0 : \theta_j = 0$. Note, also, that this condition will be violated for large enough n_j, so that a different analysis will be called for if the sample size is huge.

Another point of interest is that the answers obtained from the simulation would be quite different if one considered, say, the subset of all tests for which $0 < p < .05$. Indeed, if the initial proportion of true nulls in the above simulation were $1/2$, then, among those tests for which $0 < p < .05$, the proportion of true nulls would have the lower bound .048 (although, for nonextreme values of the alternative θ_j, the proportion of true nulls would be much higher). The point, however, is that, if a study yields $p = .049$, this is the actual information, not the summary statement $0 < p < .05$. The two statements are very different in terms of the information they convey, and replacing the former by the latter is simply an egregious mistake.

Although the simulation visibly demonstrates that a p value near .05 provides at best weak evidence against H_0, it does not indicate why this is so. The reason is basically that the probability of getting a p value near .05, when H_1 is true, cannot be much bigger than the probability of getting a p value near .05, when H_0 is true. To explicitly see this, consider a slightly different aspect of the above simulation. We will create a histogram that indicates where the p values in (5) fall that are generated from the null hypotheses, and also a histogram of the p values generated under the alternative hypotheses. For ease of assimilation, we give only the portion of the histogram corresponding to the range $.01 < p < .10$.

Under the null hypotheses, p values are well known to be $\mathrm{Uniform}(0, 1)$; the histogram that would result from such p values is represented in Figure 2 by the unshaded columns. Thus, the probability that $.01 < p < .02$ is .01.

To make a histogram of the p values in (5) under the alternative hypotheses, we must choose the n_j, σ_j, and θ_j. The distribution of p under the alternatives actually depends only on the $\xi_j = \sqrt{n_j}\theta_j/\sigma_j$. We consider the four cases (a) $\xi_j \equiv 1/2$, (b) $\xi_j \equiv 1$, (c) $\xi_j \equiv 2$, and (d) $\xi_j \equiv 4$. Figure 2 gives the corresponding histograms of p values (over the range $.01 < p < .10$); these are the shaded columns.

As expected, smaller values of p are more likely under the alternatives than under the nulls, but the degree to which this is so is rather modest for p values in common regions. For instance, a p value in the interval $(.04, .05)$ is essentially equally likely to occur under the nulls as under the alternatives when $\xi_j = .5$; is *less* likely to occur under the alternatives when $\xi_j = 4$; and is considerably more likely under the alternatives only in the case $\xi_j = 2$ (where the p value is 3.7 times more likely to have arisen from the alternative than the null). This last case is essentially the choice of alternatives that maximizes the probability of p being in the interval $(.04, .05)$ (as shown by Berger and Sellke 1987). Thus, no matter *how* one chooses the n_j, σ_j, and θ_j under the alternatives, *at most* 3.7% of the p values will fall in the interval $(.04, .05)$, so that a p value near .05 provides *at most* 3.7 to 1 odds in favor of H_1. (This is actually just a restatement of the earlier observation that, if 50% of the nulls are initially true, then *at least* 23% of those with a p value near .05 will be true.) And other choices of the alternatives are much more likely to yield a histogram like the other cases in Figure 2, rather than this extreme bound. The clear message is that knowing that the data are "rare"

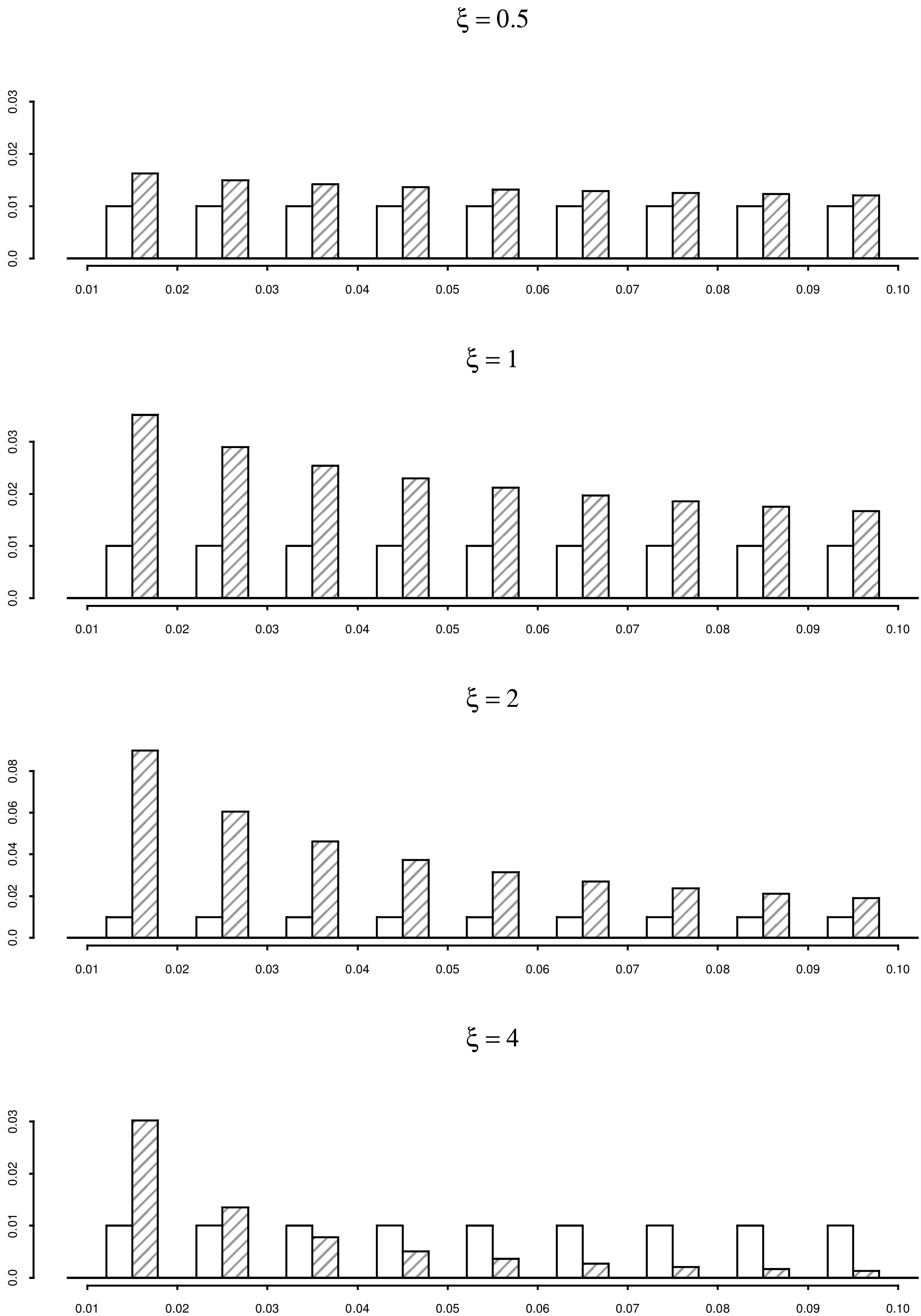

Figure 2. Distribution of p Values Under the Null Hypotheses (unshaded columns) and Under the Alternative Hypotheses (shaded columns) Over the Range .01 < p < .10. $\xi = \sqrt{n}\theta / \sigma$ is the standardized mean under the alternative.

under H_0 is of little use unless one determines whether or not they are also "rare" under H_1.

3. CALIBRATION OF P VALUES

In this section, the calibrations of a p value, that were given in (2) and (3), are developed. Motivations will be given in terms of nonparametric testing and parametric testing, from both Bayesian and frequentist perspectives. Since our goal is to interpret the calibrated p values as lower bounds on Bayes factors or conditional frequentist error probabilities, we have to explicitly consider alternatives to the null model.

3.1 Justification Via p Value Testing

3.1.1 Bounds on the Odds of H_0 *to* H_1 *Under Beta Alternatives*

In Section 2, we referred to the fact that, under the null hypothesis, the distribution of the p value, $p(\mathbf{X})$, is $\mathrm{Uniform}[0,1]$. [We write $p(\mathbf{X})$ to emphasize that p is now being treated as a random function of the data.] Alternatives are typically developed by considering alternative models for $\mathbf{X}$, as in Section 2, but the results then end up being quite problem specific. An attractive approach is to, instead, directly consider alternative distributions for p itself. Indeed, we shall suppose that, under H_1, the density of p is $f(p|\xi)$, where ξ is an unknown parameter. Thus, we will test:

$$H_0 : p \sim \mathrm{Uniform}(0,1) \quad \text{versus} \quad H_1 : p \sim f(p|\xi).$$

Others have previously considered direct choice of alternatives for $p(\mathbf{X})$; see, for instance, Hodges (1992), Donahue (1999), and Sackrowitz and Samuel-Cahn (1999). If the test statistic has been appropriately chosen so that large values of $T(\mathbf{X})$ would be evidence in favor of H_1, then the density of p under H_1 should be decreasing in p. A class of decreasing densities for p that is very easy to work with is the class of $\mathrm{Beta}(\xi,1)$ densities, for $0 < \xi \le 1$, given by

$$f(p|\xi) = \xi\, p^{\xi-1} \quad . \tag{6}$$

The uniform distribution (i.e., H_0) arises from the choice $\xi = 1$. We begin with this class because it is easy to follow the derivations of the calibrations in this case. A considerably more general class of alternatives is considered in Section 3.1.3.

The Bayes factor (or odds) of H_0 to H_1, for a given prior density $\pi(\xi)$ on this alternative, is

$$B_\pi(p) = \frac{f(p|1)}{\int_0^1 f(p|\xi)\pi(\xi)\, d\xi}.$$

Calculus shows that

$$\underline{B} = \inf_{\text{all } \pi} B_\pi(p) = \frac{f(p|1)}{\sup_\xi \xi\, p^{\xi-1}} = -e\; p\; \log p \quad \text{for} \quad p < e^{-1}\ , \tag{7}$$

and $\underline{B} = 1$ otherwise, which is the proposed calibration in (2). Of particular note is that this lower bound holds for *any* prior distribution on ξ, and can hence be viewed as an objective lower bound on the odds of H_0 to H_1 for the $\mathrm{Beta}(\xi,1)$ alternatives.

3.1.2 Bounds on Conditional Frequentist Error Probabilities

In this section, we develop the calibration in (3), using the conditional frequentist approach. The idea behind this approach, formalized in Kiefer (1977) and further developed in Berger, Brown, and Wolpert (1994), Wolpert (1995), and Berger, Boukai, and Wang (1997), is to find a conditioning statistic that measures the amount of evidence in the data (for or against the null hypothesis), and then to report error probabilities conditional on this statistic. The result is true frequentist error probabilities that are as data-dependent as p values. For the situation considered in Section 3.1.1, we will show that a lower bound on the conditional error probability of Type I is given by (3).

We begin with an example of conditional frequentist testing, to illustrate basic ideas and issues. The presentation that is adopted here is different than in the above articles, in part to clearly illustrate the options that are available and, in part, to emphasize the pure frequentist nature of the resulting procedure.

Assume that H_0 and H_1 are simple hypotheses (with absolutely continuous densities) and let S denote the statistic with respect to which conditioning is to be performed. It is most traditional, in conditional frequentist inference, to choose S to be an ancillary statistic, which here would mean that it has the same distribution under H_0 as under H_1. It will be seen, however, that other choices can be even more attractive.

A useful way to construct suitable S is to consider what we will call *evidential equivalence* statistics, E_0 and E_1, that have two purposes. First, H_0 will be accepted when $E_0 > E_1$ (i.e., when the "evidence" for H_0 is greater than that for H_1), and rejected otherwise. Next, define the conditioning statistic by $S = \max\{E_0, E_1\}$. Intuitively, data in the acceptance region, and for which $E_0 = s$, will be viewed as providing equivalent strength of evidence as data in the rejection region for which $E_1 = s$. Two interesting possible choices of the E_i are (a) likelihood ratios and (b) p values.

With S determined, one computes conditional Type I and Type II error probabilities as

$$\begin{aligned} \alpha(s) &= P_0(\text{Type I error}|S = s) \equiv P_0(E_0 \le E_1|S(X) = s) \\ \beta(s) &= P_1(\text{Type II error}|S = s) \\ &\qquad \equiv P_1(E_0 > E_1|S(X) = s), \end{aligned} \tag{8}$$

where P_0 and P_1 refer to probability under H_0 and H_1, respectively.

Example 1. Consider the special case of the situation in Section 3.1.1, in which it is desired to test $H_0 : p \sim \mathrm{Uniform}(0,1)$ versus $H_1 : p \sim \mathrm{Beta}(1/2,1)$. Noting that the density under the alternative is $(2\sqrt{p})^{-1}$, it follows that the likelihood ratio of H_0 to H_1 is $L(p) = 1/(2\sqrt{p})^{-1} = 2\sqrt{p}$. As p varies from 0 to 1, note that $L(p)$ varies from 0 to 2. We now consider four choices of the evidential equivalence statistics.

1. Ancillary conditioning: Choose $E_0 = L(p)$ and $E_1 = 2 - L(p)$. The intuition is that L ranges from 0 to 2, and

$L = 1$ is often viewed as conveying equal support for the two hypotheses; thus one might feel that, say, data for which $L = 3/2$ is equivalent, in terms of strength of evidence, to data for which $L = 1/2$. The main motivation for this choice, however, is that a basic calculation shows that the statistic $S = \max\{L(p), 2 - L(p)\}$ is an ancillary statistic, having the same distribution under H_0 as under H_1. Part of the folklore in statistics is that one should condition on ancillary statistics when they are available. Computing the resulting conditional error probabilities yields the following test:

$$\mathrm{T}^{\mathrm{A}} = \begin{cases} \text{if } p_{\mathrm{obs}} \le \frac{1}{4}, & \text{reject } \mathrm{H}_0 \text{ and report Type I} \\ & \text{conditional error probability} \\ & \alpha(p_{\mathrm{obs}}) = \sqrt{p_{\mathrm{obs}}}\,; \\ \text{if } p_{\mathrm{obs}} > \frac{1}{4}, & \text{accept } \mathrm{H}_0 \text{ and report Type II} \\ & \text{conditional error probability} \\ & \beta(p_{\mathrm{obs}}) = \frac{1}{2}. \end{cases} \quad (9)$$

The Type II conditional error probability in (9) is not satisfactory for two reasons. First, although $L(p_{\mathrm{obs}})$ varies as p_{obs} varies from $1/4$ to 1, $\beta(p_{\mathrm{obs}})$ remains constant. Furthermore, this constant is $1/2$, which suggests that one is doing no better than random choice of an hypothesis (at least from the perspective of Type II error).

2. *Intrinsic significance:* For the "symmetric" class of problems in which $L(p)$ has the same distribution under H_0 as does $1/L(p)$ under H_1, Birnbaum (1961) can be viewed as suggesting use of $E_0 = L(p)$, $E_1 = 1/L(p)$ and using $S = \max\{L(p), 1/L(p)\}$ to define a conditional frequentist test, calling the resulting conditional Type I error the "intrinsic significance level." This choice of S was seconded by Barnard in the discussion of Kiefer (1977). The resulting test for this example is as follows:

$$\mathrm{T}^{\mathrm{I}} = \begin{cases} \text{if } \; p_{\mathrm{obs}} \le \frac{1}{4}, & \text{reject } \mathrm{H}_0 \text{ and report Type I conditional error probability} \\ & \alpha(p_{\mathrm{obs}}) = \begin{cases} 1 & \text{if } \; 0 < p_{\mathrm{obs}} < \frac{1}{16} \\ (1 + (16 p_{\mathrm{obs}}^2)^{-1})^{-1} & \text{if } \; \frac{1}{16} < p_{\mathrm{obs}} < \frac{1}{4} \end{cases}; \\ \text{if } p_{\mathrm{obs}} > \frac{1}{4}, & \text{accept } \mathrm{H}_0 \text{ and report Type II conditional error probability} \\ & \beta(p_{\mathrm{obs}}) = (1 + 4 p_{\mathrm{obs}})^{-1}. \end{cases} \quad (10)$$

It is obviously unsuitable to report $\alpha(p_{\mathrm{obs}}) = 1$ when $p_{\mathrm{obs}} < 1/16$; that is, when the evidence against H_0 is strongest! This strange conditional error probability arose because the values of $E_1 = 1/L(p)$ range from 2 to ∞ over this range of p, and there are no data in the acceptance region for which $E_0 = L(p)$ can match these values. Hence S simply equals p for $p < 1/16$, and the conditioning in (7) is degenerate. Birnbaum (1961) did not actually recommend this test for situations such as this example that are not appropriately symmetric; our purpose in considering it is simply to show that this "natural" definition of evidential equivalence does not lead to fruitful conditional frequentist tests in general.

3. *p value conditioning:* In classical statistics, the most commonly used measure of evidence is the p value, so it is natural to consider choosing $E_0 = p_0$ and $E_1 = p_1$, where p_0 is the p value when testing H_0 versus H_1, and p_1 is the p value when testing H_1 versus H_0. Note that the use of p values in determining evidentiary equivalence is much weaker than their use as an absolute measure of significance. In particular, use of $E_i = \psi(p_i)$, where ψ is any strictly increasing function, would determine the same evidentiary equivalence; thus the criticisms of p values that we raised in earlier sections would not apply to their use here.

Defining $S = \max\{p_0, p_1\}$ avoids the problem incurred by the intrinsic conditioning statistic, since both p_i range continuously from 0 to 1 and one never has "unmatched" data. The resulting conditional frequentist test is:

$$\mathrm{T}^{\mathrm{P}} = \begin{cases} \text{if } p_{\mathrm{obs}} \le .382, & \text{reject } \mathrm{H}_0 \text{ and report Type I} \\ & \text{conditional error probability} \\ & \alpha(p_{\mathrm{obs}}) = (1 + \frac{1}{2} p_{\mathrm{obs}}^{-1/2})^{-1}; \\ \text{if } p_{\mathrm{obs}} > .382, & \text{accept } \mathrm{H}_0 \text{ and report Type II} \\ & \text{conditional error probability} \\ & \beta(p_{\mathrm{obs}}) = (1 + 2 p_{\mathrm{obs}}^{1/2})^{-1}. \end{cases} \quad (11)$$

These conditional error probabilities do not exhibit unnatural behavior for either small or large values of p_{obs}, so that T^{P} is quite attractive. There is a possible oddity for middle values of p_{obs}: one might make a decision with an error probability larger than .5. For instance, when $p_{\mathrm{obs}} = .36$, then the conclusion of T^{P} is to reject H_0 and report conditional error probability $\alpha(.36) = .55$. This possibility led Berger, Brown, and Wolpert (1994) to introduce a "no decision" region for this test, which eliminated the problem. The complication is arguably unnecessary, however, in that the situation occurs only with uninteresting data that provides no real evidence for, or against, H_0.

There is an additional startling fact about T^{P}: a direct application of Bayes's theorem shows that $\alpha(p_{\mathrm{obs}})$ and $\beta(p_{\mathrm{obs}})$ are precisely the Bayesian posterior probablities of H_0 and H_1, respectively, assuming the hypotheses have equal prior probabilities of $1/2$. Berger, Brown, and Wolpert (1994) showed that this equivalence holds generally when testing simple hypotheses.

4. Equal probability continuum conditioning: Kiefer (1977) suggested choosing S so that $\alpha(s) = \beta(s)$; this is natural from a minimax perspective. One can find the derivation of S and computation of the associated conditional error probabilities in Kiefer (1977) (although it is done for the variable $Y = -\log(p)$, which has an exponential distribution). The resulting test is

$$\mathrm{T}^{\mathrm{C}} = \begin{cases} \text{if } p_{\mathrm{obs}} \le .397, & \text{reject } \mathrm{H}_0 \text{ and report Type I} \\ & \text{conditional error probability} \\ & \alpha(p_{\mathrm{obs}}) = (1 + (p_{\mathrm{obs}}^{-3/4} - 1)^{1/3})^{-1}; \\ \text{if } p_{\mathrm{obs}} > .397, & \text{accept } \mathrm{H}_0 \text{ and report Type II} \\ & \text{conditional error probability} \\ & \beta(p_{\mathrm{obs}}) = (1 + (p_{\mathrm{obs}}^{-3/4} - 1)^{-1/3})^{-1}. \end{cases} \tag{12}$$

This conditioning has the nice property that it also avoids the difficulty of the intrinsic significance test: it guarantees "matching" data in both the rejection and acceptance regions. But the Type II conditional error probability has the undesirable property that $\beta(p_{\mathrm{obs}}) \to 0$ as $p_{\mathrm{obs}} \to 1$; this is highly unnatural because $L(1) = 2$, which hardly suggests that the decision to accept H_0 would be "error-free."

Note: In each of the above scenarios one could consider conditioning on $S = \min\{E_0, E_1\}$ rather than $S = \max\{E_0, E_1\}$. The motivation would be that, instead of equating evidence *in favor* of the two hypotheses, one equates evidence *against* them. For this example, computation shows that ancillary conditioning and intrinsic significance conditioning are unaffected by this change. However, p value conditioning with this choice of S yields quite different answers, but answers that are clearly unsatisfactory. Indeed, the resulting conditional error probabilities are such that $\alpha(p_{\mathrm{obs}}) \to 1/3$ as $L(p_{\mathrm{obs}}) \to 0$, while $\beta(p_{\mathrm{obs}}) \to 0$ as $L(p_{\mathrm{obs}}) \to 2$, neither of which is sensible. Hence, this choice of S should not be considered.

We now turn to the more general problem of interest, testing $\mathrm{H}_0 : p \sim \mathrm{Uniform}(0,1)$ versus $\mathrm{H}_1 : p \sim \mathrm{Beta}(\xi, 1)$. For $\xi = .5$, we saw in Example 1 that p value conditioning was clearly the preferred method of conditioning. For arbitrary choices of ξ, ancillary choices of S are available, but are quite complicated. Furthermore, for cases in which the computations could be performed (e.g., $\xi = 1/3$ or $\xi = 2/3$), the resulting conditional error probabilities exhibited very unsatisfactory behavior. The tests T^{I} and T^{C} can also be defined for general ξ, but exhibit exactly the same difficulties as observed for the case $\xi = 1/2$. In contrast, the p value conditioning yields reasonable conditional error probabilities for all values of ξ. In part, this is indicated by the fact that these conditional error probabilities have the simultaneous justification of being objective Bayesian posterior error probabilities; methodology that arises separately from pure frequentist and pure Bayesian arguments inherits the attractive properties of both schools. Thus, we henceforth consider only p value conditioning.

Determination of T^{P} for arbitrary fixed ξ, $0 < \xi < 1$, is identical to the analysis in Example 1. The likelihood ratio of H_0 to H_1 is now $L(p) = \xi^{-1} p^{1-\xi}$, and we again use the conditioning statistic $S = \max\{p_0, p_1\}$, where p_0 is the p value when testing H_0 versus H_1, and p_1 is the p-value when testing H_1 versus H_0. The resulting conditional frequentist test is:

$$\mathrm{T}^{\mathrm{P}} = \begin{cases} \text{if } p_{\mathrm{obs}} \le C, & \text{reject } \mathrm{H}_0 \text{ and report Type I} \\ & \text{conditional error probability} \\ & \alpha_\xi(p_{\mathrm{obs}}) = (1 + L(p_{\mathrm{obs}})^{-1})^{-1}; \\ \text{if } p_{\mathrm{obs}} > C, & \text{accept } \mathrm{H}_0 \text{ and report Type II} \\ & \text{conditional error probability} \\ & \beta_\xi(p_{\mathrm{obs}}) = (1 + L(p_{\mathrm{obs}}))^{-1}, \end{cases} \tag{13}$$

where C is the solution of the equation $C = 1 - C^{\xi}$.

The details of this test are not actually relevant for our purposes here. We need only the fact, following from (7), that, for $p_{\mathrm{obs}} < e^{-1}$,

$$\inf_{\xi} \alpha_\xi(p_{\mathrm{obs}}) = \left(1 + \frac{1}{\inf_\xi L(p_{\mathrm{obs}})}\right)^{-1} = \left(1 + \frac{1}{-e\, p_{\mathrm{obs}} \log(p_{\mathrm{obs}})}\right)^{-1}. \tag{14}$$

Recall that our goal was to provide a frequentist calibration for the common approach of reporting a p value when rejecting H_0. The frequentist test T^{P} will, upon rejecting H_0, report an error probability that is guaranteed to be bigger than the right hand side of (14). Hence this bound, which is that in (3), provides the desired calibration.

Many frequentists might feel that this calibration is too small, in that the actual frequentist error rate is larger than the bound. (Frequentists typically want to report upper bounds on the error probability, not lower bounds.) Indeed, when $p_{\mathrm{obs}} = .05$, all we are really saying is that the actual frequentist error probability is some number larger than the calibration $-e p_{\mathrm{obs}} \log(p_{\mathrm{obs}}) = .289$. For those not satisfied with this statement, and who want to produce a real frequentist error probability as opposed to a lower bound, we recommend use of the general conditional frequentist testing paradigm discussed by Berger, Boukai, and Wang (1997) and Dass and Berger (1998). (We do not recommend the alternative "solution" of saying that .289 cannot be used because it is *too small* as an error rate, so that the original $p_{\mathrm{obs}} = .05$ should be reported instead!)

3.1.3 Calibration for Nonparametric Alternatives With Decreasing Failure Rate

The Beta alternatives in Section 3.1.1 are a rather restrictive class, and it is of interest to see if the bounds in (7) and (14) hold more generally. Instead of working with p and its distribution $f(p|\xi)$, it is more convenient to consider $Y = -\log p$ and its distributions under the null and alternative hypotheses. If p has the $\mathrm{Beta}(\xi, 1)$ distribution given in (6), then

$$\Pr\{Y > y\} = \Pr\{p < e^{-y}\} = e^{-\xi y} \quad ,$$

so that Y has an Exponential(ξ) distribution (and, of course, the null hypothesis again obtains for $\xi = 1$).

A reasonable requirement is that the distribution of Y, under the alternative hypothesis, have a decreasing (nonincreasing) failure rate. This is equivalent to requiring that the distribution of $Y{-}y \mid Y{>}y$ be stochastically increasing with y. In terms of $p = e^{-y}$, the requirement of decreasing failure rate for Y means that the distribution of $\frac{p}{p_0} \mid p < p_0$ is stochastically decreasing with p_0. In particular, this implies that, for any fixed p_0, the probability $\Pr\{p < \frac{1}{2}p_0 | p < p_0\}$ increases as p_0 goes to 0; this is a natural condition implying that the mass under the alternative is appropriately concentrated near zero.

Lower bound on the Bayes factor: Assume, accordingly, that the failure rate function

$$h_1(y) = \frac{f_1(y)}{\int_y^\infty f_1(z)dz} \ ,$$

for the density, f_1, of Y under H_1, has a decreasing failure rate. Then

$$f_1(y) = h_1(y)\exp\left\{-\int_0^y h_1(z)dz\right\} \\ \leq h_1(y)\exp\{-yh_1(y)\} \quad ,$$

from which it follows that the Bayes factor of H_0 to H_1 satisfies

$$B = \frac{e^{-y}}{f_1(y)} \geq \frac{e^{-y}}{h_1(y)\exp\{-yh_1(y)\}} \geq e\,y\,e^{-y} \quad \text{for} \quad y \geq 1 \, ,$$

and $\underline{B} = 1$ otherwise, the inequalities being sharp. Since this lower bound holds for *any* density in the (now nonparametric) class of alternatives, it will also hold for any Bayes factor with respect to a prior over that class. Transforming back to p yields exactly the same bound as in (7). This lower bound is thus valid over a very large class of nonparametric alternatives and priors.

Lower bound on the conditional frequentist Type I error probability: The conditional frequentist argument for the nonparametric alternatives proceeds exactly as in Section 3.1.2. Indeed, if the density, $f_1(y)$ of $Y = -\log(p)$ has nonincreasing failure rate, the analogue of (14) is

$$\inf_{f_1} \alpha(y_{\mathrm{obs}}) = \left(1 + \frac{1}{e^{-y_{\mathrm{obs}}}/\sup_{f_1} f_1(y_{\mathrm{obs}})}\right)^{-1} \\ = \left(1 + \frac{1}{e y_{\mathrm{obs}} e^{-y_{\mathrm{obs}}}}\right)^{-1} . \quad (15)$$

Transforming back to p_{obs} yields (14) as the lower bound on the conditional Type I error probability.

Verifying the decreasing failure rate property: There is a relatively simple method for checking that Y has decreasing failure rate, given only the original densities of the test statistic $T(\mathbf{X})$ under H_0 and H_1, which will be denoted by $f_0(t)$ and $m(t)$, respectively. Let F_0 and M denote the cdf's corresponding to f_0 and m, respectively.

If p is defined as in (1), the survival function of $Y = -\log(p(X))$, under the alternative, is given by

$$\Pr\{Y > y\} \ = \ \Pr\{p < e^{-y}\} = 1 - M(F_0^{-1}(1 - e^{-y})), \quad (16)$$

so that its density is given by

$$f_1(y) = \frac{m(F_0^{-1}(1 - e^{-y}))}{e^y f_0(F_0^{-1}(1 - e^{-y}))} \quad . \quad (17)$$

The hazard rate function of Y is given by the ratio of (17) and (15). Differentiation shows that this hazard rate function is nonincreasing if and only if

$$\frac{m(t)}{1 - M(t)} \Big/ \frac{f_0(t)}{1 - F_0(t)} \quad (18)$$

is nonincreasing. Thus, the applicability of the bound in (7) can be assured by verification that (18) is nonincreasing.

In the Bayesian case, the density $m(t)$ will arise as the Bayesian marginal or predictive density $m(t) = \int f(t|\theta)\pi(\theta)\,d\theta$, corresponding to the alternative $\mathrm{H}_1 : f(t|\theta)$ and under the prior $\pi(\theta)$.

Example 2. Consider the situation of Section 2, with iid Normal(θ, σ^2) data, $\mathrm{H}_0 : \theta = 0$, $\mathrm{H}_1 : \theta \neq 0$, and $T(X) = \sqrt{n}|\bar{X}|/\sigma$. Suppose that the prior for θ under H_1 is Normal$(0, v^2)$. Then an easy computation shows that the ratio in (18) is given by

$$R(t)/\left[c\ R\left(\frac{t}{c}\right)\right] , \quad (19)$$

where $c = (1 + nv^2/\sigma^2)^{1/2}$ and $R(t) = (1 - \Phi(t)/\phi(t))$ (with Φ and ϕ denoting the standard normal cdf and density, respectively) is *Mill's ratio*, or the reciprocal of the hazard rate function of the standard normal. Figure 3 graphs the function in (19) for various values of c, and all appear to be decreasing to their limiting value $1/c^2$.

3.2 Bayesian Justification Via Parametric Testing

It is natural to ask whether the bound $B \geq -ep\log p$ is also reasonable in parametric testing scenarios involving composite alternatives. This is relatively easy to study from the Bayesian perspective, and so we restrict the analysis here to that situation. [For parametric conditional frequentist testing with composite alternatives, one would have to employ the more involved techniques of Berger, Boukai, and Wang (1997) and Dass and Berger (1998).] Consider first the standard normal example.

Example 3. Consider the normal testing scenario in Example 2. Berger and Sellke (1987) provided lower bounds for the Bayes factor of H_0 to H_1 when $\pi(\theta)$ belongs to the following possible classes of priors:

$$\Gamma_{\mathrm{Normal}} \ = \ \{\pi : \pi(\theta) = \mathrm{Normal}(0, v^2), v > 0\}$$

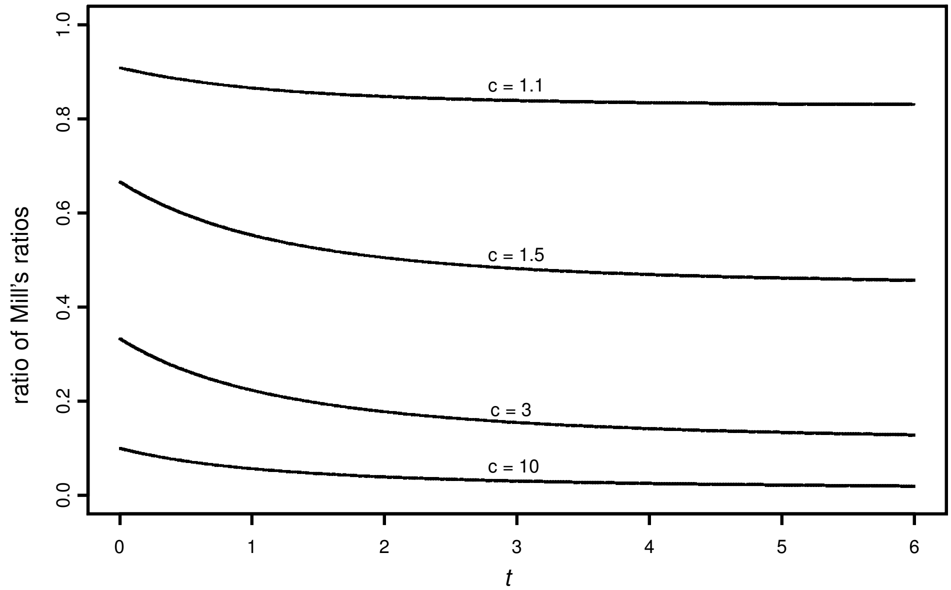

Figure 3. Plots of the ratio of Mill's ratios in (19).

$$\Gamma_{\mathrm{US}} = \{\pi : \pi(\theta) \text{ is unimodal and symmetrical about } 0\}$$

$$\Gamma_{\mathrm{Sym}} = \{\pi : \pi(\theta) \text{ is symmetrical about } 0\}.$$

Table 3 displays these lower bounds for various p values, along with the calibration $-ep\log p$.

A striking feature of Table 3 is the close agreement between the lower bounds on the Bayes factors for the class Γ_{US} and the proposed calibration, $-ep\log p$. This class of priors is often argued to contain all objective and sensible priors, so that the close agreement lends strong support to the appropriateness of the calibration. Incidentally, the close agreement also suggests that the hazard rate function for the alternatives at which the infimum is attained must be nearly constant, and this can indeed be shown numerically. The class Γ_{Sym} clearly falls outside the conditions under which the calibration bound is valid, but this is arguably a much too large class of priors.

The next example considers the multivariate normal situation. Comparisons between p values and Bayes factors can be difficult in higher dimensions, so this example is of considerable interest in indicating whether or not the proposed calibration is also reasonable in higher dimensions (although note that the nonparametric arguments of Section 3.1 would equally well apply to higher dimensional situations).

Example 4. Assume that the null model for the data $\mathbf{X} = (X_1, \ldots, X_k)$ is $N_k(\mathbf{0}, \mathbf{I})$ and that the alternative is $N_k(\boldsymbol{\theta}, \mathbf{I})$, where $\mathbf{I}$ is the $k \times k$ identity matrix. (Without loss of generality, we assume that there is only the single vector observation.) The prior distribution under the alternative is assumed to belong to the following class of scale mixtures of normals:

$$\boldsymbol{\theta}|v^2 \sim N_k(\mathbf{0}, v^2\mathbf{I})$$
$$\pi(v^2) \text{ is a nonincreasing density on } (0,\infty). \tag{20}$$

The reason we do not consider the conjugate class of $N_k(\mathbf{0}, v^2\mathbf{I})$ priors here is that such priors concentrate most of their mass very near the surface of the ball of radius $v\sqrt{k}$ in higher dimensions, which does not seem appropriate. In contrast, the priors in (19) can assign considerable mass elsewhere.

Finding the lower bound on the Bayes factor over the class in (19) is equivalent to finding the lower bound over the smaller class in which $\pi(v^2)$ is $\mathrm{Uniform}(0, r)$, $r > 0$. (This is so because any nondecreasing density can be written as a mixture of uniform distributions, and the linear functional $m(\mathbf{x}) = \int f(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|v^2)\pi(v^2)dv^2$ of $\pi(v^2)$ is thus maximized over these extreme points.) The Bayes factor of H_0 to H_1, corresponding to the uniform prior, is (for $k > 2$)

$$B_r = \frac{r\, b^a\, e^{-b}}{\Gamma(a)\,\left[\mathcal{G}(b|a,1) - \mathcal{G}(\frac{b}{1+r}|a,1)\right]}, \tag{21}$$

where $a = k/2 - 1$, $b = ||\mathbf{x}||^2/2$, and $\mathcal{G}(\cdot|a,1)$ is the Gamma distribution function with parameters a and 1. The infimum, $\underline{B}$, of B_r over r is then easy to compute numerically. Table

Table 3. Infimum of Bayes factors, p Values and Their Calibrations

p	.1	.05	.01	.001
$-ep\log p$	.6259	.4072	.1252	.01878
Γ_{Normal}	.7007	.4727	.1534	.02407
Γ_{US}	.6393	.4084	.1223	.01833
Γ_{Sym}	.5151	.2937	.0730	.00887

Table 4. $\underline{B}$, p Values, and Their Calibrations for Various Dimensions k

p	.1	.05	.01	.001
$-ep\log p$	.6259	.4072	.1252	.01878
$k = 1$	.7367	.5110	.1729	.02787
$k = 3$	.6419	.4281	.1371	.02101
$k = 6$	.6062	.3989	.1253	.01894
$k = 15$	.5750	.3748	.1165	.01748
$k = 30$	.5603	.3643	.1129	.01695

4 gives the values of $\underline{B}$ for various p values, p, and various dimensions, k. The calibration seems to maintain a very close similarity to the lower bounds on the Bayes factors for any dimension, lending considerable additional credibility to its use.

4. CONCLUSIONS

The most important conclusion is that, for testing "precise" hypotheses, p values should not be used directly, because they are too easily misinterpreted. The standard approach in teaching—of stressing the formal definition of a p value while warning against its misinterpretation—has simply been an abysmal failure. In this regard, the calibrations proposed in (2) and (3) are an immediately useful tool, putting p values on scales that can be more easily interpreted.

Although the proposed calibrations ameliorate the worst features of p values, they can themselves be criticized for being biased against the null hypothesis; recall that the calibrations arose from bounds on Bayes factors or conditional Type I error probabilities that were *least favorable* to the null hypothesis. That such bounds are still much larger than p values indicates the severe nature of the bias against a precise null that can arise due to the p value fallacy.

Although the calibrations are a considerable improvement over p values, this issue of bias against the null leads us to instead recommend objective Bayesian or conditional frequentist procedures, for situations when the alternative hypothesis is specified. References to the development of such procedures include, on the Bayesian side, Jeffreys (1961), Kass and Raftery (1995), O'Hagan (1995), and Berger and Pericchi (1996, 1998); and, on the conditional frequentist side, Berger, Brown, and Wolpert (1994), Berger, Boukai, and Wang (1997), Dass and Berger(1998), and Dass (1998).

One scenario in which we would definitely recommend use of the calibrations is when investigating fit to the null model, with no explicit alternative in mind. The lack of an alternative precludes use of the objective Bayesian or conditional frequentist procedures mentioned above. See Bayarri and Berger (1999, 2000) for further discussion of this issue.

[Received April 2000. Revised August 2000.]

REFERENCES

Bayarri, M. J., and Berger, J. O. (1999), "Quantifying Surprise in the Data and Model Verification," in *Bayesian Statistics 6*, eds. J. M. Bernardo, J. O. Berger, A.P. Dawid, and A. F. M. Smith, Oxford: Oxford University Press, pp. 53–82.

——— (2000), "*P*-values for Composite Null Models," *Journal of the American Statistical Association*, 95, 1127–1142.

Berger, J., Boukai, B., and Wang, Y. (1997), "Unified Frequentist and Bayesian Testing of a Precise Hypothesis" (with discussion), *Statistical Science*, 12, 133–160.

Berger, J. O., Brown, L. D., and Wolpert, R. L. (1994), "A Unified Conditional Frequentist and Bayesian Test for Fixed and Sequential Simple Hypothesis Testing," *The Annals of Statistics*, 22, 1787–1807.

Berger, J. O., and Delampady, M. (1987), "Testing Precise Hypothesis" (with discussion), *Statistical Science*, 2, 317–352.

Berger, J., and Mortera, J. (1999), "Default Bayes Factors for Non-nested Hypothesis Testing," *Journal of the American Statistical Association*, 94, 542–554.

Berger, J., and Pericchi, L. (1996), "The Intrinsic Bayes Factor for Model Selection and Prediction," *Journal of the American Statistical Association*, 91, 109–122.

——— (1998), "Accurate and Stable Bayesian Model Selection: the Median Intrinsic Bayes Factor," *Sankhyā*, Ser. B, 60, 1–18.

Berger, J. O., and Sellke, T. (1987), "Testing a Point Null Hypothesis: the Irreconcilability of *p*-Values and Evidence" (with discussion), *Journal of the American Statistical Association*, 82, 112–122.

Birnbaum, A. (1961), "On the Foundation of Statistical Inference: Binary Experiments," *Annals of Mathematical Statistics*, 32, 414–435.

Casella, G., and Berger, R. (1987), "Reconciling Bayesian and Frequentist Evidence in the One-sided Testing Problem" (with discussion), *Journal of the American Statistical Association*, 82, 106–111.

Dass, S. (1998), "Unified Bayesian and Conditional Frequentist Testing Procedures," unpublished Ph.D. Thesis, Purdue University.

Dass, S., and Berger, J. (1998), "Unified Bayesian and Conditional Frequentist Testing of Composite Hypotheses," ISDS Discussion paper 98-43, Duke University.

Delampady, M., and Berger, J. O., (1990), "Lower Bounds on Bayes Factors for Multinomial Distributions, With Application to Chi-squared Tests of Fit," *The Annals of Statistics*, 18, 1295–1316.

Donahue, R. (1999), "A Note on Information Seldom Reported Via the *P* Value," *The American Statistician*, 53, 303–306.

Edwards, W., Lindman, H., and Savage, L. J. (1963), "Bayesian Statistical Inference for Psychological Research," *Psychological Review*, 70, 193–242.

Jeffreys, H. (1961), *Theory of Probability*, London: Oxford University Press.

Gibbons, J., and Pratt, J. (1975), "*P* Values: Interpretation and Methodology," *The American Statistician*, 29, 20–25.

Goodman, S. (1999a), "Toward Evidence-Based Medical Statistics. 1: The *P*-Value Fallacy," *Annals of Internal Medicine*, 130, 995–1004.

——— (1999b), "Toward Evidence-Based Medical Statistics. 2: The Bayes Factor," *Annals of Internal Medicine*, 130, 1005–1013.

Hodges, J. (1992), "Who Knows What Alternative Lurks in the Heart of Significance Tests?," in *Bayesian Statistics 4*, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, London: Oxford University Press, pp. 247–266.

Kass, R. E., and Raftery, A. (1995), "Bayes Factors," *Journal of the American Statistical Association*, 90, 773–795.

Kiefer, J. (1977), "Conditional Confidence Statements and Confidence Estimators" (with discussion), *Journal of the American Statistical Association*, 72, 789–827.

Matthews, R. (1998), "The Great Health Hoax," in *The Sunday Telegraph*, September 13, 1998.

O'Hagan, A. (1995), "Fractional Bayes Factors for Model Comparisons," *Journal of the Royal Statistical Society*, Ser. B, 57, 99–138.

Sackrowitz, H., and Samuel-Cahn, E. (1999), "*P* Values as Random Variables—Expected *P* Values," *The American Statistician*, 53, 326–331.

Schervish, M. (1996), "*P* Values: What They Are and What They Are Not," *The American Statistician*, 50, 203–206.

Wolpert, R. L. (1995), "Testing Simple Hypotheses," in *Studies in Classification, Data Analysis, and Knowledge Organization* (vol. 7), eds. H. H. Bock and W. Polasek, Heidelberg: Springer-Verlag, pp. 289–297.