Calibration of P-values for Testing Precise Null Hypotheses

by

Thomas Sellke
Purdue University

M.J. Bayarri
University of Valencia
and Duke University

James O. Berger
Duke University

Institute of Statistics and Decision Sciences
Durham, North Carolina 27708

Abstract

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 and phrases. Bayes factors; Bayesian robustness; Conditional frequentist error probabilities; Odds; Surprise.

1. Introduction

In statistical analysis of data $\mathbf{X}$, one is frequently working, at a given moment, with an entertained model or hypothesis $H_0 : \mathbf{X} \sim f(\mathbf{x})$, 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 $\mathbf{x}_{obs}$, with large values of T indicating less compatibility. The p-value is then defined as

$$p = \Pr(T(\mathbf{X}) \geq T(\mathbf{x}_{obs})). \tag{1.1}$$

In this paper, we assume that $f(\mathbf{x})$ is completely specified, so that the probability computation in (1.1) 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.

The density in H_0 can contain nuisance parameters, in which case the choice of an appropriate distribution for computation of (1.1) can be problematical; see Bayarri and Berger (1998, 1999) for discussion and recommendation of a preferred p-value. The calibration discussed herein can be directly applied to this preferred (and any other valid) p-value, so the restriction in this paper to a simple null hypothesis is primarily pedagogical. 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 papers, among them Edwards, Lindman, and Savage (1963), Berger and Sellke (1987), Berger and Delampady (1987), and Delampady and Berger (1990), the latter specifically considering the problem of testing fit when $T(\mathbf{X})$ is chosen to be the usual chi-squared statistic for fit. The focus of these works is that p-values are commonly thought to imply considerably greater evidence against H_0 than is actually warranted. In Section 2, we present two simple examples demonstrating this concern. The examples are presented as simulations that are easy to perform, even in introductory statistics classes. Indeed, we suggest that such simulations should be mandatory in any statistics course that presents p-values.

Because of the ubiquitous use of p-values, it seems desirable to provide a simple way to understand their evidentiary import. In Section 3 we discuss a simple calibration of p to achieve this. The calibration is easy to state: simply compute

$$B(p) = -e\,p\,\log(p)\,, \tag{1.2}$$

when $p < 1/e$, and interpret this as a lower bound on the odds (or Bayes factor) of H_0 to H_1. In terms of a frequentist error probability α (in rejecting H_0), the calibration is

$$\alpha(p) = (1 + [-e\,p\log(p)]^{-1})^{-1}. \tag{1.3}$$

Interestingly, this latter expression is exactly the same as the (default) posterior probability of H_0 that arises from use of the Bayes factor in (1.2) together with the assumption that H_0 and H_1 have equal prior probabilities of 1/2. Thus use of (1.3) has the additional pedagogical advantage that one need not fear misinterpretation of an error probability as the probability that the hypothesis is true; here, they coincide.

Table 1 presents various p-values and their associated calibrations. Thus $p = 0.05$ translates

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

Table 1: Calibration of p-values as odds (Bayes factors) and conditional error probabilities.

into odds $B(0.05) = 0.407$ (roughly 1 to 2.5) of H_0 to H_1, and frequentist error probability $\alpha(0.05) = 0.289$ in rejecting H_0. (The default posterior probability of H_0 would also be 0.289.) Clearly $p = 0.05$ does not indicate particularly strong evidence against H_0. Even $p = 0.01$ 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. The common misinterpretation of p-values

We present an extended example in this section, in order to emphasize that common interpretations of p-values are inappropriate. 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:

$$H_0 : D_i \text{ has negligible effect} \quad \text{versus} \quad H_1 : D_i \text{ has a non-negligible effect} . \tag{2.1}$$

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 ≈ 0.05 (or ≈ 0.01). The question we consider is: How strong is the evidence that the drug in question has a non-negligible effect?

DRUG	D1	D2	D3	D4	D5	D6
P-VALUE	0.41	**0.04**	0.32	0.94	**0.01**	0.28
DRUG	D7	D8	D9	D10	D11	D12
P-VALUE	0.11	**0.05**	0.65	**0.009**	0.09	0.66

Table 2: P-values corresponding to testing whether drug D_i has negligible effect.

To study this, we will simply collect all the p-values from a 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 $p \approx 0.05$ (D_2 and D_8 in Table 2) or $p \approx 0.01$ (D_5 and D_{10} in Table 2), and ask: What proportion of these tests have true H_0, i.e., ineffective drugs?

We shortly discuss the simulation to answer this question, but here is the basic and surprising conclusion, first established (theoretically) in Berger and Sellke (1987). Suppose it is known that, a priori, about 50% of the drugs tested have a negligible effect. (This is actually quite a neutral assumption; in some scenarios this percentage is likely to be much higher.) Then:

1. Of the D_i for which the p-value ≈ 0.05, *at least* 23% (and typically close to 50%) will have negligible effect.

2. Of the D_i for which the p-value ≈ 0.01, *at least* 7% (and typically close to 15%) will have negligible effect.

Similar results arise for other initial proportions of ineffective drugs. 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 the p-value is ≈ 0.05, is *at least* 12% (35%). The basic point is that a p-value of 0.05 can *never* reduce the initial proportion of true null hypotheses by more than a very modest factor.

The numbers above are based on the following simulation. Suppose that each test in (2.1) is based on normal data (known variance), with θ_j being the treatment mean for D_j, so that (2.1) 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

$$p = 2\left[1 - \Phi(T(\mathbf{x}_{obs}))\right] \quad ; \tag{2.2}$$

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 c.d.f. 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 0.05. For instance, one can look at those tests for which $0.04 \leq p \leq 0.05$. One then simply notes the proportion of such tests for which H_0 is true. An S+ code for carrying out this simulation is given in the Appendix, which also discusses some further details, such as choice of the alternatives θ_j.

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 $H_0 : |\theta_j| < \epsilon$, providing $\epsilon < \sigma_j/(4\sqrt{n_j})$. This is important because hypotheses such as $H_0 : \theta_j = 0$ are unlikely to ever be true exactly. (D_j will probably have *some* effect, even if only $\theta_j = 10^{-8}$.) Indeed, the hypothesis $H_0 : \theta_j = 0$ should really just be thought of as an approximation to a small interval null, and Berger and Delampady (1987) show that it is a good approximation if $\epsilon < \sigma_j/(4\sqrt{n_j})$. Thus, in practice, one must make the judgement that this condition will hold before formulating the test as that of $H_0 : \theta_j = 0$. 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. This fact is also the basis for resolution of the so-called Jeffreys Paradox (or Lindley's Paradox).

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 $0 < p < 0.05$. Indeed, the proportion of true nulls would then be in accordance with common intuition concerning p-values. The point, however, is that, if a study yields $p = 0.046$, this is the actual information, not the summary statement $0 < p < 0.05$. The two statements are very different from an evidentiary perspective, and replacing the former by the latter is simply an egregious mistake.

While the simulation visibly demonstrates that a p-value near 0.05 provides at best weak evidence against H_0, it does not indicate why this is so. The reason is basically that a p-value near .05 is essentially as likely to arise from H_1 as from H_0. 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 (2.2) fall that are generated from the null hypotheses, and also a histogram of the p-values generated under the alternative hypotheses.

Under the null hypotheses, p-values are well known to be $Uniform(0, 1)$; the histogram that would result from such p-values is represented in Figure 1 by the unshaded columns. Thus the probability that $0.1 < p < 0.2$ is 0.1, the probability that $0.04 < p < 0.05$ is 0.01, etc.

To make a histogram of the p-values in (2.2) under the alternative hypotheses, we must choose the n_j, σ_j, and θ_j. A variety of possible specifications are given in the Appendix; for

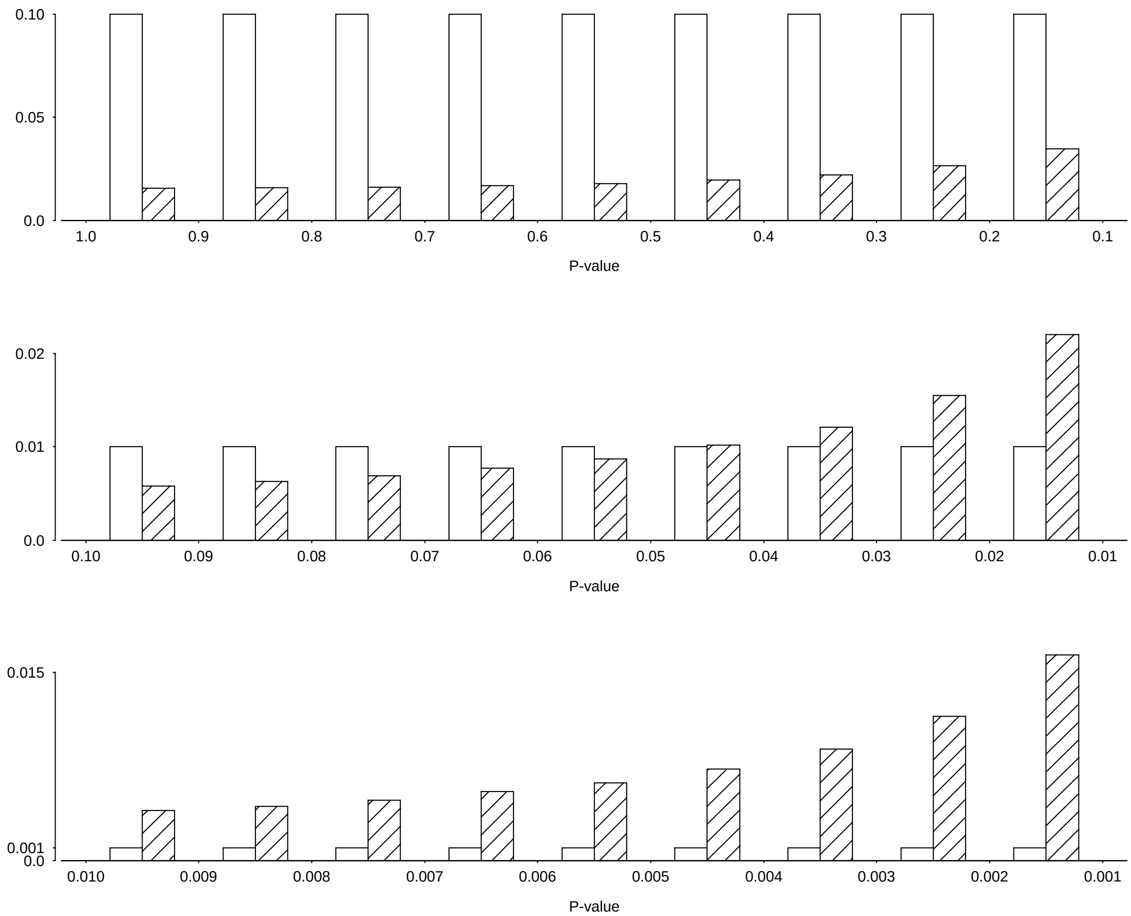

Figure 1: Distribution of p-values under the null hypotheses (unshaded columns) and under the alternative hypotheses (shaded columns).

illustrative purposes, we choose (as in the simulation in the Appendix) $n_j = 20$ and $\sigma_j = 1$ and let the θ_j be generated from a normal distribution with mean 0 and variance 2. (That this distribution is symmetric about 0 has no bearing on matters; the same histogram would result from generating the θ_j from the corresponding positive half-normal distribution.)

An easy computation shows that, for this choice of alternatives, the p-values will be distributed according to the c.d.f

$$2[1 - \Phi(\frac{1}{\sqrt{41}}\Phi^{-1}(1 - \frac{1}{2}p))] \ . \tag{2.3}$$

The corresponding histogram is given by the shaded columns in Figure 1. As expected, smaller values of p are more likely under the alternatives than 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 $(0.04, 0.05)$ is essentially equally likely to occur under the nulls as under the alternatives. Thus observing, say, $p = 0.046$ provides no evidence in favor of the null or the alternative. Even a p-value in the interval $(0.009, 0.010)$ is only about 4 times more likely to occur under the alternatives than under the nulls.

The natural question to ask is whether the qualitative nature of the phenomenon observed in Figure 1 is due to the particular choice we made for the alternatives. The answer is - no; this histogram is quite typical of what occurs. Indeed, it can be shown that, no matter *how* one chooses the n_j, σ_j and θ_j under the alternatives, *at most* 3.4% of the p-values will fall in the interval $(0.04, 0.05)$, so that a p-value near .05 provides *at most* 3.4 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 0.05 will be true.) And reasonable choices of the alternatives are much more likely to yield a histogram like Figure 1 than yield such extreme bounds. The clear message is that knowing that the data are 'rare' 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, p, that were given in (1.2) and (1.3), 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 $Uniform[0,1]$. (We write $p(\mathbf{X})$ to emphasize that p is now being treated as a random function of the data.) Alternatives would typically be 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:

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

Others have previously considered direct choice of alternatives for $p(\mathbf{X})$; see, for instance, Hodges (1992).

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. An example is that in Section 2; the density corresponding to (2.3) is

$$f(p) = \frac{1}{\sqrt{41}} \, \exp\{ \, \frac{1}{41} \left[\Phi^{-1}(1 - \frac{p}{2}) \right]^2 \} \quad ,$$

which is decreasing in p.

A class of alternatives for p that is very easy to work with is the class of $Be(\xi, 1)$ distributions, with $0 < \xi \leq 1$, so that the densities are nonincreasing:

$$f(p|\xi) = \xi \, p^{\xi - 1} \quad . \tag{3.1}$$

The uniform distribution (i.e., H_0) arises from the choice $\xi = 1$.

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

$$\underline{B} = \inf_{\text{all}\, \pi} B_\pi(p) = \frac{f(p|1)}{\sup_\xi \xi \, p^{\xi-1}} = -e \; p \; \log p \quad \text{for } p < e^{-1} \; , \tag{3.2}$$

and $\underline{B} = 1$ otherwise, which is the proposed calibration in (1.2). Of particular note is that this lower bound holds for *any* prior distribution on the alternative, and can hence be viewed as an objective lower bound on the odds of H_0 to H_1.

3.1.2. Bounds on the odds of H_0 to H_1 under decreasing failure rate

The Beta alternatives in Subsection 3.1.1 are a rather restricted class, and it is of interest to see if the bound in (3.2) holds 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. It can easily be checked that, if p has the $Be(\xi, 1)$ distribution given in (3.1), then

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

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

A natural requirement is that the distribution of Y have a decreasing (non-increasing) 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. In particular, this implies that, for any fixed p_0, the probability $Pr\{p < \frac{1}{2}|p < p_0\}$ increases as p_0 goes to 0; this is a natural condition implying that the mass under the alternative is appropriately concentrated near zero.

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

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

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

(3.2). This lower bound is thus valid over a very large class of nonparametric alternatives and priors.

In the remainder of this section, we present a 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. Usually, 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 $H_1 : f(t|\theta)$, under the prior $\pi(\theta)$. Let F and M denote the c.d.f.'s corresponding to f and m, respectively.

If p is defined as in (1.1), then it is straightforward to show that the survival function of $Y = -\log(p(X))$, under the alternative, is given by

$$Pr\{Y > y\} = Pr\{p < e^{-y}\} = 1 - M[F^{-1}(1 - e^{-y})], \tag{3.3}$$

so that its density is given by

$$f_1(y) = \frac{m[F^{-1}(1 - e^{-y})]}{e^y f[F^{-1}(1 - e^{-y})]} \quad . \tag{3.4}$$

The hazard rate function of Y is given by the ratio of (3.4) and (3.3), and can easily be seen to be nonincreasing if and only if

$$\frac{m(t)}{1 - M(t)} \Big/ \frac{f(t)}{1 - F(t)} \tag{3.5}$$

is nonincreasing. Thus the applicability of the bound in (3.2) can be assured by verification that (3.5) is nonincreasing.

Example 3.1 Consider the situation of Section 2, with i.i.d. $Normal(\theta, \sigma^2)$ data, $H_0 : \theta = 0$, $H_1 : \theta \neq 0$, 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 (3.5) is given by

$$R(t)/[c\ R(\frac{t}{c})], \tag{3.6}$$

where $c = (1 + nv^2/\sigma^2)^{1/2}$ and $R(t)$ is *Mill's ratio*, or the inverse of the hazard rate function of the standard normal. Figure 1 graphs the function in (3.6) for various values of c, and all appear to be decreasing to their limiting value $1/c^2$.

□

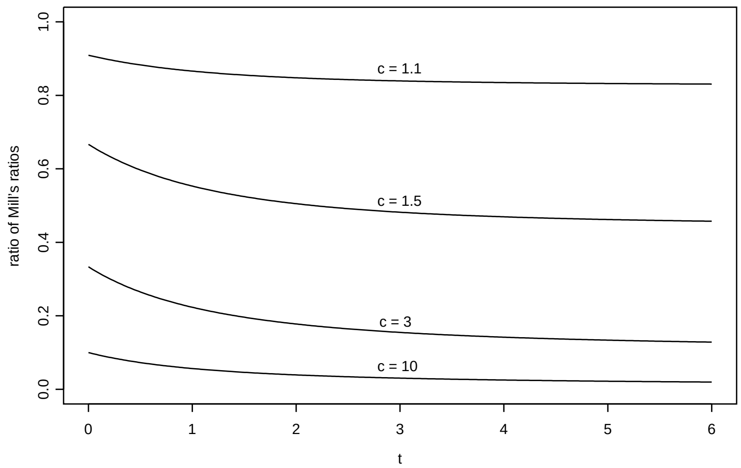

Figure 2: Plots of the ratio of Mill's ratios in (3.6).

3.1.3. Bounds on conditional frequentist error probabilities

The proposed calibration can also be seen to arise from a conditional frequentist perspective. The idea behind this approach, formalized in Kiefer (1977) and developed in Berger, Brown, and Wolpert (1994) and Berger, Boukai, and Wang (1997), is to find a conditioning statistic that measures the strength 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. In this section we show that a lower bound on the conditional error probability of Type I is given by (1.3) which thus becomes the suggested calibration for p-values in terms of frequentist error probabilities.

The analysis here follows the development of conditional frequentist testing in Berger, Brown, and Wolpert (1994). To test $H_0 : p \sim \text{Uniform}(0,1)$ versus $H_1 : p \sim \text{Beta}(\xi, 1)$, for a fixed ξ, $0 < \xi < 1$, the Bayes factor is easily seen to be

$$B(p) = \xi^{-1} p^{1-\xi}\,,$$

an increasing function of p. The distribution functions of B under the hypotheses are needed next. Clearly

$$Pr(B \leq b) = Pr\left[p \leq (b\,\xi)^{\frac{1}{1-\xi}}\right]$$

so that, under H_0 (where p has c.d.f. $F(s) = s$), the c.d.f. of B is

$$F_0(b) = (b\,\xi)^{\frac{1}{1-\xi}}$$

and, under H_1 (where p has c.d.f. $F(s) = s^{\xi}$), the c.d.f. of B is

$$F_1(b) = (b\,\xi)^{\frac{\xi}{1-\xi}}\,.$$

It can be numerically shown that $F_0(1) \le 1 - F_1(1)$, in which case the final needed quantity is given in Berger, Brown, and Wolpert (1994) as

$$a = F_0^{-1}[1 - F_1(1)] = \frac{1}{\xi}\left[1 - \xi^{\frac{\xi}{1-\xi}}\right]^{1-\xi}\,.$$

Letting CEP denote *conditional error probability*, the conditional frequentist test is then given as follows:

- If $B(p) \le 1$, reject H_0 and report CEP $\alpha_\xi(B) = \frac{B}{1+B}$.
- If $1 < B(p) < a$, take no decision.
- If $B(p) \ge a$, accept H_0 and report CEP $\beta_\xi(B) = \frac{1}{1+B}$.

Next, we compute $\inf_\xi \alpha_\xi(B)$. Since

$$\alpha_\xi(B) = \frac{B}{1+B} = \frac{1}{1+B^{-1}}$$

is an increasing function of B, it is clear that the minimum over ξ of $\alpha_\xi(B)$ is given by replacing B by its minimum over ξ, which is given in (3.2), resulting in the bound in (1.3).

Frequentists may well not agree with use of the *minimum* α, it being far more common to report the *maximum* in situations of nonconstant Type I error probability. Indeed, we would not disagree with this judgement, and would, instead, urge use of default conditional frequentist tests as proposed in Berger, Boukai, and Wang (1997), Dass and Berger (1998), and Dass (1998). However, recall that the purpose here was to calibrate a p-value, by at least putting it on an error probability 'scale,' and the given calibration achieves that goal. Another way of saying this is that reporting (1.3), while debatable from a frequentist perspective is, at least, far better than reporting the p-value itself.

For the conditional Type II error probability, $\beta_\xi(B) = 1/(1+B)$, it is clear that the lower bound on B from (3.2) becomes the upper bound

$$\beta_\xi(B) \geq \frac{1}{1 - e\,p\,\log(p)}\,.$$

Note, however, that one needs to also consider a when dealing with Type II error. That this is rarely a problem in practice is indicated by the fact that, for small values of p, it can be shown that $a \approx \log(\log(1/p))$, so that the *no-decision* region remains rather small.

Similar arguments can be made for the more general alternatives discussed in Subsection 3.1.2. Indeed, if the distribution of $Y = -\log(p)$ has nonincreasing failure rate, the arguments therein can be directly modified to obtain the same bounds as above on the conditional Type I and Type II error probabilities. The only real complication is that a is then no longer easily specified, but we suspect that the no-decision region would remain of negligible import and, in any case, it only affects Type II error under acceptance of the null.

3.2. Justification via parametric testing

It is natural to ask whether the bound $B \geq -ep\log p$ is also reasonable in parametric testing scenarios. Consider first the standard normal example.

Example 3.2 Consider the normal testing scenario in Example 3.1. Berger and Sellke (1987) provide lower bounds for the Bayes factor of H_0 to H_1 when $\pi(\theta)$ belongs to the following possible classes of priors:

$$\begin{aligned}
\Gamma_{Normal} &= \{\pi : \pi(\theta) = Normal(0, v^2), v > 0\} \\
\Gamma_{US} &= \{\pi : \pi(\theta) \text{ is unimodal and symmetrical about } 0\} \\
\Gamma_{Sym} &= \{\pi : \pi(\theta) \text{ is symmetrical about } 0\}\ \ .
\end{aligned}$$

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

p	0.1	0.05	0.01	0.001
$-ep\log p$	0.6259	0.4072	0.1252	0.01878
Γ_{Normal}	0.7007	0.4727	0.1534	0.02407
Γ_{US}	0.6393	0.4084	0.1223	0.01833
Γ_{Sym}	0.5151	0.2937	0.0730	0.00887

Table 3: Infimum of Bayes factors, p-values and their calibrations.

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 indicates 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.

$\square$

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 Subsection 3.1 would equally well apply to higher dimensional situations).

Example 3.3 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:

$$\begin{array}{rcl} \boldsymbol{\theta}|v^2 & \sim & N_k(\mathbf{0}, v^2\mathbf{I}) \\ \pi(v^2) & \text{is} & \text{a nondecreasing density on}(0, \infty). \end{array} \tag{3.7}$$

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 (3.7) can assign considerable mass elsewhere.

It is easy to see that finding the lower bound on the Bayes factor over the class in (3.7) is equivalent to finding the lower bound over the smaller class in which $\pi(v^2)$ is $Uniform(0, r)$, $r > 0$. The Bayes factor of H_0 to H_1, corresponding to this prior, is (for $k > 2$)

$$B_r = \frac{r\ b^a\ e^{-b}}{\Gamma(a)\ [\mathcal{G}(b|a,1) - \mathcal{G}(\frac{b}{1+r}|a,1)]}\,, \tag{3.8}$$

where $a = k/2 - 1$, $b = ||\mathbf{x}||^2/2$, and $\mathcal{G}(\cdot|a, b)$ 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 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.

$\square$

p	0.1	0.05	0.01	0.001
$-ep\log p$	0.6259	0.4072	0.1252	0.01878
$k=3$	0.6419	0.4281	0.1371	0.02101
$k=6$	0.6062	0.3989	0.1253	0.01894
$k=15$	0.5750	0.3748	0.1165	0.01748
$k=30$	0.5603	0.3643	0.1129	0.01695

Table 4: $\underline{B}$, p-values and their calibrations for various dimensions k.

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 (1.2) and (1.3) are an immediately useful tool, putting p-values on scales that can be more easily interpreted.

While 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 incurred through common interpretations of p-values.

While 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 (1998, 1999) for further discussion of this issue.

Acknowledgements

This work was supported, in part, by the National Science Foundation (U.S.A.) under Grants DMS-9303556 and DMS-9802261, and by the Ministry of Education and Culture (Spain) under Grant PB96-0776.

Appendix

In this Appendix, we provide the S+ code to simulate the proportion of times that the null hypothesis is true when $p \approx 0.05$ or $p \approx 0.01$. Specifically, `L` values of the usual normal T statistic, $T(X) = \sqrt{n}\,|\overline{X}|/\sigma$, are generated, the known standard deviation, `sigma`, and sample size, `n`, being inputs. (The values `sigma` = 1 and `n` = 20 are chosen below, but the specific choices are irrelevant and could vary from test to test; all that really matters is the choice of the $\eta_j = \sqrt{n_j}\,\theta_j/\sigma_j$.) Features that must be specified are `pi0`, the initial proportion of true nulls, and the `theta1`, the means under the alternatives. The simulation could be conducted with *any* desired sequence of alternative means, but the program below accommodates three interesting options: *(i)* all alternative `theta1` are fixed at the value `a`; *(ii)* the alternative `theta1` are randomly generated from a normal distribution with mean 0 and standard deviation `a`; *(iii)* the alternative `theta1` are randomly generated from a uniform distribution on the interval (`-a`, `a`). These three options are accessed by setting `dis` equal to 1, 2, and 3, respectively. `pro` returns the proportion of T-values in $(1.96, 2]$ (that is, with $p \approx 0.05$), and in (2.576, 2.616] (that is, with $p \approx 0.01$) for which the null hypothesis is true.

```
sigma <- 1                                  # standard deviation
n<-20                                       # sample size

pro <- function(pi0, L, a, dis)
 {
  L0 <-  round(L*pi0/100)                    # number of simulations from H0
  L1 <- L - L0                               # number of simulations from H1
  x0 <- rnorm(L0, 0, sigma/sqrt(n))                   # sample means from H0
  switch(dis,
         x1<-rnorm(L1, a, sigma/sqrt(n)),                         #one point
        {theta1 <- rnorm(L1, 0, a);
         x1 <- rnorm(theta1, sigma/sqrt(n))},                     #normal
        {theta1 <- runif(L1, -a, a);
         x1 <- rnorm(theta1, sigma/sqrt(n))}    )                 #uniform
  t0 <- abs(x0) * sqrt(n)/sigma                       #t's with H0 true
```

```
 t1 <- abs(x1) * sqrt(n)/sigma                      #t's with H1 true
 pr1<- 1/(1 + length(t1[1.96<t1 & t1<= 2])/length(t0[1.96<t0 & t0<= 2]))
 pr2<- 1/(1+length(t1[2.576<t1 & t1<=2.616])/length(t0[2.5766<t0 & t0<=2.616]))
 return (pr1*100, pr2*100)
}
```

When $p \approx 0.05$, it is interesting to note that the proportion of true nulls will usually exceed the initial proportion `pi0`, unless `a` is chosen carefully. Indeed, finding the value of `a` that minimizes the proportion of true nulls is an interesting exercise. For the three cases considered in the simulation and if the initial percentage of true nulls is 50%, the corresponding minimum percentages are *(i)* 23%; *(ii)* 32%; and *(iii)* 29%. These arise for values of `a` that are roughly 2 sample standard deviations from the null mean. Note that case *(i)* is the absolute minimum over all possible sequences of `theta1`.

References

[1] Bayarri, M. J., and Berger, J. O. (1998), "P-values for Composite Null Models," ISDS Discussion Paper 98-40, Duke University.

[2] Bayarri, M. J. and Berger, J. O. (1999), "Quantifying Surprise in the Data and Model Verification," to appear in *Bayesian Statistics 6*, eds. J. M. Bernardo, J. O. Berger, A.P. Dawid and A. F. M. Smith, Oxford: Oxford University Press.

[3] Berger, J. (1985), *Statistical Decision Theory and Bayesian Analysis, Second Edition*, New York: Springer-Verlag.

[4] Berger, J. (1994), "An Overview of Robust Bayesian Analysis" (with discussion), *Test*, **3**, 5–124.

[5] Berger, J., Boukai, B. and Wang, Y. (1997), "Unified Frequentist and Bayesian Testing of a Precise Hypothesis" (with discussion), *Statistical Science*, **12(3)**, 133–160.

[6] 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," *Annals of Statistics* **22**, 1787–1807.

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

[8] Berger, J. and Mortera, J. (1999), "Default Bayes Factors for Non-nested Hypothesis Testing," To appear in *Journal of the American Statistical Association*

[9] Berger, J. and Pericchi, L. (1996), "The intrinsic Bayes factor for model selection and prediction," *Journal of the American Statistical Association*, **91**, 109–122.

[10] Berger, J. and Pericchi, L. (1998), "Accurate and Stable Bayesian Model Selection: the Median Intrinsic Bayes Factor," *Sankhyā*, **B 60**, 1–18.

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

[12] 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.

[13] Dass, S. (1998), Unified Bayesian and Conditional Frequentist Testing Procedures. Ph.D. Thesis, Purdue University.

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

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

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

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

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

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

[20] 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, pp. 247-266, London: Oxford University Press.

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