Testing a Point Null Hypothesis: The Irreconcilability of *P* Values and Evidence

JAMES O. BERGER and THOMAS SELLKE*

The problem of testing a point null hypothesis (or a "small interval" null hypothesis) is considered. Of interest is the relationship between the P value (or observed significance level) and conditional and Bayesian measures of evidence against the null hypothesis. Although one might presume that a small P value indicates the presence of strong evidence against the null, such is not necessarily the case. Expanding on earlier work [especially Edwards, Lindman, and Savage (1963) and Dickey (1977)], it is shown that actual evidence against a null (as measured, say, by posterior probability or comparative likelihood) can differ by *an order of magnitude* from the P value. For instance, data that yield a P value of .05, when testing a normal mean, result in a posterior probability of the null of *at least* .30 for *any* objective prior distribution. ("Objective" here means that equal prior weight is given the two hypotheses and that the prior is symmetric and nonincreasing away from the null; other definitions of "objective" will be seen to yield qualitatively similar results.) The overall conclusion is that P values can be highly misleading measures of the evidence provided by the data against the null hypothesis.

KEY WORDS: P values; Point null hypothesis; Bayes factor; Posterior probability; Weighted likelihood ratio.

1. INTRODUCTION

We consider the simple situation of observing a random quantity X having density (for convenience) $f(x \mid \theta)$, θ being an unknown parameter assuming values in a parameter space $\Theta \subset \mathbf{R}^1$. It is desired to test the null hypothesis $H_0 : \theta = \theta_0$ versus the alternative hypothesis $H_1 : \theta \neq \theta_0$, where θ_0 is a specified value of θ corresponding to a fairly sharply defined hypothesis being tested. (Although exact point null hypotheses rarely occur, many "small interval" hypotheses can be realistically approximated by point nulls; this issue is discussed in Sec. 4.) Suppose that a classical test would be based on consideration of some test statistic $T(X)$, where large values of $T(X)$ cast doubt on H_0. The P value (or observed significance level) of observed data, x, is then

$$p = \Pr_{\theta=\theta_0}(T(X) \geq T(x)).$$

Example 1. Suppose that $X = (X_1, \ldots, X_n)$, where the X_i are iid $\mathcal{N}(\theta, \sigma^2)$, σ^2 known. Then the usual test statistic is

$$T(X) = \sqrt{n}|\bar{X} - \theta_0|/\sigma,$$

where $\bar{X}$ is the sample mean, and

$$p = 2(1 - \Phi(t)),$$

where Φ is the standard normal cdf and

$$t = T(x) = \sqrt{n}|\bar{x} - \theta_0|/\sigma.$$

We will presume that the classical approach is the report of p, rather than the report of a (pre-experimental) Neyman–Pearson error probability. This is because (a) most statisticians prefer use of P values, feeling it to be important to indicate how strong the evidence against H_0 is (see Kiefer 1977), and (b) the alternative measures of evidence we consider are based on knowledge of x [or $t = T(x)$]. [For a comparison of Neyman–Pearson error probabilities and Bayesian answers, see Dickey (1977).]

There are several well-known criticisms of testing a point null hypothesis. One is the issue of "statistical" versus "practical" significance, that one can get a very small p even when $|\theta - \theta_0|$ is so small as to make θ equivalent to θ_0 for practical purposes. [This issue dates back at least to Berkson (1938, 1942); see also Good (1983), Hodges and Lehmann (1954), and Solo (1984) for discussion and history.] Also well known is "Jeffreys's paradox" or "Lindley's paradox," whereby for a Bayesian analysis with a fixed prior and for values of t chosen to yield a given fixed p, the posterior probability of H_0 goes to 1 as the sample size increases. [A few references are Good (1983), Jeffreys (1961), Lindley (1957), and Shafer (1982).] Both of these criticisms are dependent on large sample sizes and (to some extent) on the assumption that it is plausible for θ to equal θ_0 *exactly* (more on this later).

The issue we wish to discuss has nothing to do (necessarily) with large sample sizes for even exact point nulls (although large sample sizes do tend to exacerbate the conflict, the Jeffreys–Lindley paradox being the extreme illustration thereof). The issue is simply that p gives a very misleading impression as to the validity of H_0, from almost any evidentiary viewpoint.

Example 1 (Jeffreys's Bayesian Analysis). Consider a Bayesian who chooses the prior distribution on θ, which gives probability $\frac{1}{2}$ each to H_0 and H_1 and spreads the mass out on H_1 according to an $\mathcal{N}(\theta_0, \sigma^2)$ density. [This prior is close to that recommended by Jeffreys (1961) for testing a point null, though he actually recommended a Cauchy form for the prior on H_1. We do not attempt to defend this choice of prior here. Particularly troubling is the choice of the scale factor σ^2 for the prior on H_1, though it can be argued to at least provide the right "scale." See Berger (1985) for discussion and references.] It will be seen in Section 2 that the posterior probability, $\Pr(H_0 \mid x)$, of H_0 is given by

$$\Pr(H_0 \mid x) = (1 + (1 + n)^{-1/2} \exp\{t^2/[2(1 + 1/n)]\})^{-1}, \tag{1.1}$$

some values of which are given in Table 1 for various n and t (the t being chosen to correspond to the indicated

* James O. Berger is the Richard M. Brumfield Distinguished Professor and Thomas Sellke is Assistant Professor, Department of Statistics, Purdue University, West Lafayette, IN 47907. Research was supported by National Science Foundation Grant DMS-8401996. The authors are grateful to L. Mark Berliner, Iain Johnstone, Robert Keener, Prem Puri, and Herman Rubin for suggestions or interesting arguments.

Table 1. $Pr(H_0 \mid x)$ for Jeffreys-Type Prior

		n						
p	t	1	5	10	20	50	100	1,000
.10	1.645	.42	.44	.47	.56	.65	.72	.89
.05	1.960	.35	.33	.37	.42	.52	.60	.82
.01	2.576	.21	.13	.14	.16	.22	.27	.53
.001	3.291	.086	.026	.024	.026	.034	.045	.124

values of p). The conflict between p and $\Pr(H_0 \mid x)$ is apparent. If $n = 50$ and $t = 1.960$, one can classically "reject H_0 at significance level $p = .05$," although $\Pr(H_0 \mid x) = .52$ (which would actually indicate that the evidence *favors* H_0). For practical examples of this conflict see Jeffreys (1961) or Diamond and Forrester (1983) (although one can demonstrate the conflict with virtually any classical example).

Example 1 (An Extreme Bayesian Analysis). Again consider a Bayesian who gives each hypothesis prior probability $\frac{1}{2}$, but now suppose that he decides to spread out the mass on H_1 in the symmetric fashion that is *as favorable to* H_1 as possible. The corresponding values of $\Pr(H_0 \mid x)$ are determined in Section 3 and are given in Table 2 for certain values of t. Again the numbers are astonishing. Although $p = .05$ when $t = 1.96$ is observed, even a Bayesian analysis strongly biased toward H_1 states that the null has a .227 probability of being true, evidence against the null that would not strike many people as being very strong. It is of interest to ask just how biased against H_0 must a Bayesian analysis in this situation (i.e., when $t = 1.96$) be, to produce a posterior probability of $\Pr(H_0 \mid x) = .05$? The astonishing answer is that one must give H_0 an initial prior probability of .15 and then spread out the mass of .85 (given to H_1) in the symmetric fashion that *most supports* H_1. Such blatant bias toward H_1 would hardly be tolerated in a Bayesian analysis; but the experimenter who wants to reject need not *appear* so biased—he can just observe that $p = .05$ and reject by "standard practice."

If the symmetry assumption on the aforementioned prior is dropped, that is, if one now chooses the *unrestricted* prior most favorable to H_1, the posterior probability is still not as low as p. For instance, Edwards, Lindman, and Savage (1963) showed that, if each hypothesis is given initial probability $\frac{1}{2}$, the unrestricted "most favorable to H_1" prior yields

$$\Pr(H_0 \mid x) = [1 + \exp\{t^2/2\}]^{-1}, \tag{1.2}$$

the values of which are still substantially higher than p [e.g., when $t = 1.96$, $p = .05$ and $\Pr(H_0 \mid x) = .128$].

Table 2. $Pr(H_0 \mid x)$ for a Prior Biased Toward H_1

P Value (p)	t	$Pr(H_0 \mid x)$
.10	1.645	.340
.05	1.960	.227
.01	2.576	.068
.001	3.291	.0088

Example 1 (A Likelihood Analysis). It is common to perceive the comparative evidence provided by x for two possible parameter values, θ_1 and θ_2, as being measured by the likelihood ratio

$$l_x(\theta_1 : \theta_2) = f(x \mid \theta_1)/f(x \mid \theta_2)$$

(see Edwards 1972). Thus the evidence provided by x for θ_0 against some $\theta \neq \theta_0$ could be measured by $l_x(\theta_0 : \theta)$. Of course, we do not know which $\theta \neq \theta_0$ to consider, but a *lower bound* on the comparative evidence would be (see Sec. 3)

$$\underline{l}_x = \inf_{\theta} l_x(\theta_0 : \theta) = \frac{f(x \mid \theta_0)}{\sup_{\theta} f(x \mid \theta)} = \exp\{-t^2/2\}.$$

Values of $\underline{l}_x$ for various t are given in Table 3. Again, the lower bound on the comparative likelihood when $t = 1.96$ would hardly seem to indicate strong evidence against the null, especially when it is realized that maximizing the denominator over all $\theta \neq \theta_0$ is almost certain to bias strongly the "evidence" in favor of H_1.

The evidentiary clashes so far discussed involve either Bayesian or likelihood analyses, analyses of which a frequentist might be skeptical. Let us thus phrase, say, a Bayesian analysis in frequentist terms.

Example 1 (continued). Jeffreys (1980) stated, concerning the answers obtained by using his type of prior for testing a point null,

> These are not far from the rough rule long known to astronomers, i.e. that differences up to twice the standard error usually disappear when more or better observations become available, and that those of three or more times usually persist. (p. 452)

Suppose that such an astronomer learned, to his surprise, that many statistical users rejected null hypotheses at the 5% level when $t = 1.96$ was observed. Being of an open mind, the astronomer decides to conduct an "experiment" to verify the validity of rejecting H_0 when $t = 1.96$. He looks back through his records and finds a large number of normal tests of approximate point nulls, in situations for which the truth eventually became known. Suppose that he first noticed that, overall, about half of the point nulls were false and half were true. He then concentrates attention on the subset in which he is interested, namely those tests that resulted in t being between, say, 1.96 and 2. In this subset of tests, the astronomer finds that H_0 had turned out to be true 30% of the time, so he feels vindicated in his "rule of thumb" that $t \cong 2$ does *not* imply that H_0 should be confidently rejected.

In probability language, the "experiment" of the as-

Table 3. Bounds on the Comparative Likelihood

P Value (p)	t	Likelihood ratio lower bound ($\underline{l}_x$)
.10	1.645	.258
.05	1.960	.146
.01	2.576	.036
.001	3.291	.0044

tronomer can be described as taking a random series of true and false null hypotheses (half true and half false), looking at those for which t ends up between 1.96 and 2, and finding the limiting proportion of these cases in which the null hypothesis was true. It will be shown in Section 4 that this limiting proportion will be *at least* .22.

Note the important distinction between the "experiment" here and the typical frequentist "experiment" used to evaluate the performance of, say, the classical .05 level test. The typical frequentist argument is that, if one confines attention to the sequence of *true* H_0 in the "experiment," then only 5% will have $t \geq 1.96$. This is, of course, true, but is not the answer in which the astronomer was interested. He wanted to know what he should think about the truth of H_0 upon observing $t \cong 2$, and the frequentist interpretation of .05 says nothing about this.

At this point, there might be cries of outrage to the effect that $p = .05$ was never meant to provide an absolute measure of evidence against H_0 and any such interpretation is erroneous. The trouble with this view is that, like it or not, people do hypothesis testing to obtain evidence as to whether or not the hypotheses are true, and it is hard to fault the vast majority of nonspecialists for assuming that, if $p = .05$, then H_0 is very likely wrong. This is especially so since we know of no elementary textbooks that teach that $p = .05$ (for a point null) really means that there is at best very weak evidence against H_0. Indeed, most nonspecialists interpret p precisely as $\Pr(H_0 \mid x)$ (see Diamond and Forrester 1983), which only compounds the problem.

Before getting into technical details, it is worthwhile to discuss the main reason for the substantial difference between the magnitude of p and the magnitude of the evidence against H_0. The problem is essentially one of conditioning. The actual vector of observations is x, and $\Pr(H_0 \mid x)$ and l_x depend only on the evidence from the actual data observed. To calculate a P value, however, one effectively replaces x by the "knowledge" that X is in $A = \{y\colon T(y) \geq T(x)\}$ and then calculates $p = \Pr_{\theta=\theta_0}(A)$. Although the use of frequentist measures can cause problems, the main culprit here is the replacing of x itself by A. To see this, suppose that a Bayesian in Example 1 were told only that the observed x is in a set A. If he were initially "50-50" concerning the truth of H_0, if he were very uncertain about θ should H_0 be false, and if p were moderately small, then his posterior probability of H_0 would essentially *equal* p (see Sec. 4). Thus a Bayesian sees a drastic difference between knowing x (or t) and knowing only that x is in A.

Common sense supports the distinction between x and A, as a simple illustration shows. Suppose that X is measured by a weighing scale that occasionally "sticks" (to the accompaniment of a flashing light). When the scale sticks at 100 (recognizable from the flashing light) one knows only that the true x was, say, larger than 100. If large X casts doubt on H_0, occurrence of a "stick" at 100 should certainly be greater evidence that H_0 is false than should a true reading of $x = 100$. Thus there should be no surprise that using A in the frequentist calculation might cause a substantial overevaluation of the evidence against H_0. Thus Jeffreys (1980) wrote

> I have always considered the arguments for the use of P absurd. They amount to saying that a hypothesis that may or may not be true is rejected because a greater departure from the trial value was improbable; that is, that it has not predicted something that has not happened. (p. 453)

What is, perhaps, surprising is the magnitude of the overevaluation that is encountered.

An objection often raised concerning the conflict is that point null hypotheses are not realistic, so the conflict can be ignored. It is true that exact point null hypotheses are rarely realistic (the occasional test for something like extrasensory perception perhaps being an exception), but for a large number of problems testing a point null hypothesis is a good *approximation* to the actual problem. Typically, the *actual* problem may involve a test of something like $H_0 : |\theta - \theta_0| \leq b$, but b will be small enough that H_0 can be accurately approximated by $H_0 : \theta = \theta_0$. Jeffreys (1961) and Zellner (1984) argued forcefully for the usefulness of point null testing, along these lines. And, even if testing of a point null hypothesis were disreputable, the reality is that people do it all the time [see the economic literature survey in Zellner (1984)], and we should do our best to see that it is done well. Further discussion is delayed until Section 4 where, to remove any lingering doubts, small interval null hypotheses will be dealt with.

For the most part, we will consider the Bayesian formulation of evidence in this article, concentrating on determination of lower bounds for $\Pr(H_0 \mid x)$ under various types of prior assumptions. The single prior Jeffreys analysis is one extreme; the Edwards et al. (1963) lower bounds [in (1.2)] over essentially all priors with fixed probability of H_0 is another extreme. We will be particularly interested in analysis for classes of symmetric priors, feeling that any "objective" analysis will involve some such symmetry assumption; a nonsymmetric prior implies that there are specifically favored alternative values of θ.

Section 2 reviews basic features of the calculation of $\Pr(H_0 \mid x)$ and discusses the Bayesian literature on testing a point null hypothesis. Section 3 presents the various lower bounds on $\Pr(H_0 \mid x)$. Section 4 discusses more general null hypotheses and conditional calculations, and Section 5 considers generalizations and conclusions.

2. POSTERIOR PROBABILITIES AND ODDS

It is convenient to specify a prior distribution for the testing problem as follows: let $0 < \pi_0 < 1$ denote the prior probability of H_0 (i.e., that $\theta = \theta_0$), and let $\pi_1 = 1 - \pi_0$ denote the prior probability of H_1; furthermore, suppose that the mass on H_1 (i.e., on $\theta \neq \theta_0$) is spread out according to the density $g(\theta)$. One might question the assignment of a positive probability to H_0, because it will rarely be the case that it is thought possible for $\theta = \theta_0$ to hold exactly. As mentioned in Section 1, however, H_0 is to be understood as simply an approximation to the realistic hypothesis $H_0 : |\theta - \theta_0| \leq b$, and so π_0 is to be interpreted as the prior probability that would be assigned to $\{\theta : |\theta - \theta_0| \leq b\}$. A useful way to picture the actual prior in this case is as a smooth density with a sharp spike near θ_0. (To

a Bayesian, a point null test is typically reasonable only when the prior distribution is of this form.)

Noting that the marginal density of X is

$$m(x) = f(x \mid \theta_0)\pi_0 + (1 - \pi_0)m_g(x), \qquad (2.1)$$

where

$$m_g(x) = \int f(x \mid \theta)g(\theta)\, d\theta,$$

it is clear that the posterior probability of H_0 is given by (assuming that $f(x \mid \theta_0) > 0$)

$$\Pr(H_0 \mid x) = f(x \mid \theta_0) \times \pi_0/m(x)$$

$$= \left[1 + \frac{(1 - \pi_0)}{\pi_0} \times \frac{m_g(x)}{f(x \mid \theta_0)}\right]^{-1}. \qquad (2.2)$$

Also of interest is the *posterior odds ratio* of H_0 to H_1, which is

$$\frac{\Pr(H_0 \mid x)}{1 - \Pr(H_0 \mid x)} = \frac{\pi_0}{(1 - \pi_0)} \times \frac{f(x \mid \theta_0)}{m_g(x)}. \qquad (2.3)$$

The factor $\pi_0/(1 - \pi_0)$ is the *prior odds ratio*, and

$$B_g(x) = f(x \mid \theta_0)/m_g(x) \qquad (2.4)$$

is the *Bayes factor* for H_0 versus H_1. Interest in the Bayes factor centers around the fact that it does not involve the prior probabilities of the hypotheses and hence is sometimes interpreted as the actual odds of the hypotheses implied by the data alone. This feeling is reinforced by noting that B_g can be interpreted as the likelihood ratio of H_0 to H_1, where the likelihood of H_1 is calculated with respect to the "weighting" $g(\theta)$. Of course, the presence of g (which is a part of the prior) prevents any such interpretation from having a non-Bayesian reality, but the lower bounds we consider for $\Pr(H_0 \mid x)$ translate into lower bounds for B_g, and these lower bounds *can* be considered to be "objective" bounds on the likelihood ratio of H_0 to H_1. Even if such an interpretation is not sought, it is helpful to separate the effects of π_0 and g.

Example 1 (continued). Suppose that π_0 is arbitrary and g is again $\mathcal{N}(\theta_0, \sigma^2)$. Since a sufficient statistic for θ is $\bar{X} \sim \mathcal{N}(\theta, \sigma^2/n)$, we have that $m_g(\bar{x})$ is an $\mathcal{N}(\theta_0, \sigma^2(1 + n^{-1}))$ distribution. Thus

$$B_g(x)$$

$$= f(x \mid \theta_0)/m_g(\bar{x})$$

$$= \frac{[2\pi\sigma^2/n]^{-1/2} \exp\left\{-\frac{n}{2}(\bar{x} - \theta_0)^2/\sigma^2\right\}}{[2\pi\sigma^2(1+n^{-1})]^{-1/2} \exp\{-\frac{1}{2}(\bar{x} - \theta_0)^2/[\sigma^2(1+n^{-1})]\}}$$

$$= (1 + n)^{1/2} \exp\{-\tfrac{1}{2}t^2/(1 + n^{-1})\},$$

and

$$\Pr(H_0 \mid x) = [1 + (1 - \pi_0)/(\pi_0 B_g)]^{-1}$$

$$= \left[1 + \frac{(1 - \pi_0)}{\pi_0}(1 + n)^{-1/2} \times \exp\{\tfrac{1}{2}t^2/(1 + n^{-1})\}\right]^{-1},$$

which yields (1.1) for $\pi_0 = \frac{1}{2}$. [The Jeffreys–Lindley paradox is also apparent from this expression: if t is fixed, corresponding to a fixed P value, but $n \to \infty$, then $\Pr(H_0 \mid x) \to 1$ no matter how small the P value.]

When giving numerical results, we will tend to present $\Pr(H_0 \mid x)$ for $\pi_0 = \frac{1}{2}$. The choice of $\pi_0 = \frac{1}{2}$ has obvious intuitive appeal in scientific investigations as being "objective." (Some might argue that π_0 should even be chosen larger than $\frac{1}{2}$, since H_0 is often the "established theory.") Except for personal decisions (or enlightened true subjective Bayesian hypothesis testing) it will rarely be justifiable to choose $\pi_0 < \frac{1}{2}$; who, after all, would be convinced by the statement "I conducted a Bayesian test of H_0, assigning prior probability .1 to H_0, and my conclusion is that H_0 has posterior probability .05 and should be rejected"? We emphasize this obvious point because some react to the Bayesian-classical conflict by attempting to argue that π_0 should be made small in the Bayesian analysis so as to force agreement.

There is a substantial amount of literature on the subject of Bayesian testing of a point null. Among the many references to analyses with particular priors, as in Example 1, are Jeffreys (1957, 1961), Good (1950, 1958, 1965, 1967, 1983), Lindley (1957, 1961, 1965, 1977), Raiffa and Schlaifer (1961), Edwards et al. (1963), Smith (1965), Dickey and Lientz (1970), Zellner (1971, 1984), Dickey (1971, 1973, 1974, 1980), Lempers (1971), Leamer (1978), Smith and Spiegelhalter (1980), Zellner and Siow (1980), and Diamond and Forrester (1983). Many of these works specifically discuss the relationship of $\Pr(H_0 \mid x)$ to significance levels; other papers in which such comparisons are made include Pratt (1965), DeGroot (1973), Dempster (1973), Dickey (1977), Hill (1982), Shafer (1982), and Good (1984). Finally, the articles that find lower bounds on B_g and $\Pr(H_0 \mid x)$ that are similar to those we consider include Edwards et al. (1963), Hildreth (1963), Good (1967, 1983, 1984), and Dickey (1973, 1977).

3. LOWER BOUNDS ON POSTERIOR PROBABILITIES

3.1 Introduction

This section will examine some lower bounds on $\Pr(H_0 \mid x)$ when $g(\theta)$, the distribution of θ given that H_1 is true, is allowed to vary within some class of distributions G. If the class G is sufficiently large so as to contain all "reasonable" priors, or at least a good approximation to any "reasonable" prior distribution on the H_1 parameter set, then a lower bound on $\Pr(H_0 \mid x)$ that is not small would seem to imply that the data x do not constitute strong evidence against the null hypothesis $H_0 : \theta = \theta_0$. We will assume in this section that the parameter space is the entire real line (although most of the results hold with only minor modification to parameter spaces that are subsets of the real line) and will concentrate on the following four classes of g: $G_A = \{\text{all distributions}\}$, $G_S = \{\text{all distributions symmetric about } \theta_0\}$, $G_{US} = \{\text{all unimodal distributions symmetric about } \theta_0\}$, $G_{NOR} = \{\text{all } \mathcal{N}(\theta_0, \tau^2) \text{ distributions}, 0 \le \tau^2 < \infty\}$. Even though these G's are supposed to consist only of distributions on $\{\theta \mid \theta \neq \theta_0\}$, it will be convenient

to allow them to include distributions with mass at θ_0, so the lower bounds we compute are always attained; the answers are unchanged by this simplification, and cumbersome limiting notation is avoided. Letting

$$\underline{\Pr}(H_0 \mid x, G) = \inf_{g \in G} \Pr(H_0 \mid x)$$

and

$$\underline{B}(x, G) = \inf_{g \in G} B_g(x),$$

we see immediately from formulas (2.2) and (2.4) that

$$\underline{B}(x, G) = f(x \mid \theta_0)/\sup_{g \in G} m_g(x)$$

and

$$\underline{\Pr}(H_0 \mid x, G) = \left[1 + \frac{(1 - \pi_0)}{\pi_0} \times \frac{1}{\underline{B}(x, G)}\right]^{-1}.$$

Note that $\sup_{g \in G} m_g(x)$ can be considered to be an upper bound on the "likelihood" of H_1 over all "weights" $g \in G$, so $\underline{B}(x, G)$ has an interpretation as a lower bound on the comparative likelihood of H_0 and H_1.

3.2 Lower Bounds for G_A = {All Distributions}

The simplest results obtainable are for G_A and were given in Edwards et al. (1963). The proof is elementary and will be omitted.

Theorem 1. Suppose that a maximum likelihood estimate of θ [call it $\hat{\theta}(x)$], exists for the observed x. Then

$$\underline{B}(x, G_A) = f(x \mid \theta_0)/f(x \mid \hat{\theta}(x)),$$

and

$$\underline{\Pr}(H_0 \mid x, G_A) = \left[1 + \frac{(1 - \pi_0)}{\pi_0} \times \frac{f(x \mid \hat{\theta}(x))}{f(x \mid \theta_0)}\right]^{-1}.$$

[Note that $\underline{B}(x, G_A)$ is equal to the comparative likelihood bound, $\underline{l}_x$, that was discussed in Section 1 and hence has a motivation outside of Bayesian analysis.]

Example 1 (continued). An easy calculation shows that, in this situation,

$$\underline{B}(x, G_A) = e^{-t^2/2}$$

and

$$\underline{\Pr}(H_0 \mid x, G_A) = \left[1 + \frac{(1 - \pi_0)}{\pi_0} e^{t^2/2}\right]^{-1}.$$

For several choices of t, Table 4 gives the corresponding two-sided P values, p, and the values of $\underline{\Pr}(H_0 \mid x, G_A)$, with $\pi_0 = \frac{1}{2}$. Note that the lower bounds on $\Pr(H_0 \mid x)$ are considerably larger than the corresponding P values, in spite of the fact that minimization of $\Pr(H_0 \mid x)$ over $g \in G_A$ is "maximally unfair" to the null hypothesis. The last column shows that the ratio of $\underline{\Pr}(H_0 \mid x, G_A)$ to pt is rather stable. The behavior of this ratio is described in more detail by Theorem 2.

Table 4. Comparison of P Values and $\underline{Pr}(H_0 \mid x, G_A)$ When $\pi_0 = \frac{1}{2}$

P Value (p)	t	$\underline{Pr}(H_0 \mid x, G_A)$	$\underline{Pr}(H_0 \mid x, G_A)/(pt)$
.10	1.645	.205	1.25
.05	1.960	.128	1.30
.01	2.576	.035	1.36
.001	3.291	.0044	1.35

Theorem 2. For $t > 1.68$ and $\pi_0 = \frac{1}{2}$ in Example 1,

$$\underline{\Pr}(H_0 \mid x, G_A)/pt > \sqrt{\pi/2} \cong 1.253.$$

Furthermore,

$$\lim_{t \to \infty} \underline{\Pr}(H_0 \mid x, G_A)/pt = \sqrt{\pi/2}.$$

Proof. The limit result and the inequality for $t \geq 1.84$ follow from the Mills ratio-type inequality

$$1 - \frac{1}{y^2} < \frac{y\{1 - \Phi(y)\}}{\varphi(y)} < 1 - \frac{1}{3 + y^2}, \qquad y > 0.$$

The left inequality here is from Feller (1968, p. 175), and the right inequality can be proved by using a variant of Feller's argument. For $1.68 < t < 1.84$, the inequality of the theorem was verified numerically.

The interest in this theorem is that, for $\pi_0 = \frac{1}{2}$, we can conclude that $\Pr(H_0 \mid x)$ is *at least* (1.25) pt, for *any* prior; for large t the use of p as evidence against H_0 is thus particularly bad, in a proportional sense. [The actual difference between $\Pr(H_0 \mid x)$ and the P value, however, appears to be decreasing in t.]

3.3 Lower Bounds for G_S = {Symmetric Distributions}

There is a large gap between $\underline{\Pr}(H_0 \mid x, G_A)$ (for $\pi_0 = \frac{1}{2}$) and $\Pr(H_0 \mid x)$ for the Jeffreys-type single prior analysis (compare Tables 1 and 4). This reinforces the suspicion that using G_A unduly biases the conclusion against H_0 and suggests use of more reasonable classes of priors. Symmetry of g (for the normal problem anyway) is one natural objective assumption to make. Theorem 3 begins the study of the class of symmetric g by showing that minimizing $\Pr(H_0 \mid x)$ over all $g \in G_S$ is equivalent to minimizing over the class G_{2PS} = {all symmetric two-point distributions}.

Theorem 3.

$$\sup_{g \in G_{2PS}} m_g(x) = \sup_{g \in G_S} m_g(x),$$

so

$$\underline{B}(x, G_{2PS}) = \underline{B}(x, G_S)$$

and

$$\underline{\Pr}(H_0 \mid x, G_{2PS}) = \underline{\Pr}(H_0 \mid x, G_S).$$

Proof. All elements of G_S are mixtures of elements of G_{2PS}, and $m_g(x)$ is linear when viewed as a function of g.

Example 1 (continued). If $t \leq 1$, a calculus argument shows that the symmetric two-point distribution that strictly maximizes $m_g(x)$ is the degenerate "two-point" distribution putting all mass at θ_0. Thus $\underline{B}(x, G_S) = 1$ and $\underline{\Pr}(H_0$

Table 5. Comparison of P Values and $\underline{Pr}(H_0 \mid x, G_S)$ When $\pi_0 = \frac{1}{2}$

P Value (p)	t	$\underline{Pr}(H_0 \mid x, G_S)$	$\underline{Pr}(H_0 \mid x, G_S)/(pt)$
.10	1.645	.340	2.07
.05	1.960	.227	2.31
.01	2.576	.068	2.62
.001	3.291	.0088	2.68

$\mid x, G_S) = \pi_0$ for $t \leq 1$. (Since the point mass at θ_0 is not really a legitimate prior on $\{\theta \mid \theta \neq \theta_0\}$, this means that observing $t \leq 1$ actually constitutes evidence in *favor* of H_0 for any *real* symmetric prior on $\{\theta \mid \theta \neq \theta_0\}$.)

If $t > 1$, then $m_g(x)$ is maximized by a nondegenerate element of G_{2PS}. For moderately large t, the maximum value of $m_g(x)$ for $g \in G_{2PS}$ is very well approximated by taking g to be the two-point distribution putting equal mass at $\hat{\theta}(x)$ and at $2\theta_0 - \hat{\theta}(x)$, so

$$\underline{B}(x, G_S) \cong \frac{\varphi(t)}{\frac{1}{2}\varphi(0) + \frac{1}{2}\varphi(2t)} \cong 2 \exp\{-\tfrac{1}{2}t^2\}.$$

For $t \geq 1.645$, the first approximation is accurate to within 1 in the fourth significant digit, and the second approximation to within 2 in the third significant digit. Table 5 gives the value of $\underline{Pr}(H_0 \mid x, G_S)$ for several choices of t, again with $\pi_0 = \frac{1}{2}$.

The ratio $\underline{Pr}(H_0 \mid x, G_S)/\underline{Pr}(H_0 \mid x, G_A)$ converges to 2 as t grows. Thus the discrepancy between P values and posterior probabilities becomes even worse when one restricts attention to symmetric priors. Theorem 4 describes the asymptotic behavior of $\underline{Pr}(H_0 \mid x, G_S)/(pt)$. The method of proof is the same as for Theorem 2.

Theorem 4. For $t > 2.28$ and $\pi_0 = \frac{1}{2}$ in Example 1,

$$\underline{Pr}(H_0 \mid x, G_S)/pt > \sqrt{2\pi} \cong 2.507.$$

Furthermore,

$$\lim_{t \to \infty} \underline{Pr}(H_0 \mid x, G_S)/pt = \sqrt{2\pi}.$$

3.4 Lower Bounds for G_{US} = {Unimodal, Symmetric Distributions}

Minimizing $\Pr(H_0 \mid x)$ over all symmetric priors still involves considerable bias against H_0. A further "objective" restriction, which would seem reasonable to many, is to require the prior to be unimodal, or (equivalently in the presence of the symmetry assumption) nonincreasing in $|\theta - \theta_0|$. If this did not hold, there would again appear to be "favored" alternative values of θ. The class of such priors on $\theta \neq \theta_0$ has been denoted by G_{US}. Use of this class would prevent excessive bias toward specific $\theta \neq \theta_0$.

Theorem 5 shows that minimizing $\Pr(H_0 \mid x)$ over $g \in G_{US}$ is equivalent to minimizing over the more restrictive class $\mathfrak{U}_S$ = {all symmetric uniform distributions}. The point mass at θ_0 is included in $\mathfrak{U}_S$ as a degenerate case. (Obviously, each element of G_{US} is a mixture of elements of $\mathfrak{U}_S$. The proof of Theorem 5 is thus similar to that of Theorem 3 and will be omitted.)

Theorem 5.

$$\sup_{g \in G_{US}} m_g(x) = \sup_{g \in \mathfrak{U}_S} m_g(x),$$

so $\underline{B}(x, G_{US}) = \underline{B}(x, \mathfrak{U}_S)$ and $\underline{Pr}(H_0 \mid x, G_{US}) = \underline{Pr}(H_0 \mid x, \mathfrak{U}_S)$.

Example 1 (continued). Since $G_{US} \subset G_S$, it follows from our previous remarks that $\underline{B}(x, G_{US}) = 1$ and $\underline{Pr}(H_0 \mid x, G_{US}) = \pi_0$ when $t \leq 1$. If $t > 1$, then a calculus argument shows that the $g \in G_{US}$ that maximizes $m_g(x)$ will be nondegenerate. By Theorem 5, this maximizing distribution will be uniform on the interval $(\theta_0 - K\sigma/\sqrt{n}, \sigma_0 + K\sigma/\sqrt{n})$ for some $K > 0$. Let $m_K(\bar{x})$ denote $m_g(\bar{x})$ when g is uniform on $(\theta_0 - K\sigma/\sqrt{n}, \theta_0 + K\sigma/\sqrt{n})$. Since $\bar{X} \sim \mathfrak{N}(\theta, \sigma^2/n)$,

$$m_K(\bar{x}) = (\sqrt{n}/2\sigma K) \int_{\theta_0 - K\sigma/\sqrt{n}}^{\theta_0 + K\sigma/\sqrt{n}} f(\bar{x} \mid \theta)\, d\theta$$

$$= (\sqrt{n}/\sigma)(1/2K)[\Phi(K - t) - \Phi(-(K + t))].$$

If $t > 1$, then the maximizing value of K satisfies $\partial/\partial K) m_K(\bar{x}) = 0$, so

$$K[\varphi(K + t) + \varphi(K - t)] = \Phi(K - t) - \Phi(-(K + t)). \quad (3.1)$$

Note that

$$f(\bar{x} \mid \theta_0) = (\sqrt{n}/\sigma)\varphi\left(\frac{\bar{x} - \theta_0}{\sigma/\sqrt{n}}\right) = (\sqrt{n}/\sigma)\varphi(t).$$

Thus if $t > 1$ and K maximizes $m_K(\bar{x})$, we have

$$\underline{B}(x, G_{US}) = \frac{f(\bar{x} \mid \theta_0)}{m_K(\bar{x})} = \frac{2\varphi(t)}{\varphi(K + t) + \varphi(K - t)}.$$

We summarize our results in Theorem 6.

Theorem 6. If $t \leq 1$ in Example 1, then $\underline{B}(x, G_{US}) = 1$ and $\underline{Pr}(H_0 \mid x, G_{US}) = \pi_0$. If $t > 1$, then

$$\underline{B}(x, G_{US}) = \frac{2\varphi(t)}{\varphi(K + t) + \varphi(K - t)}$$

and

$$\underline{Pr}(H_0 \mid x, G_{US}) = \left[1 + \frac{(1 - \pi_0)}{\pi_0} \times \frac{(\varphi(K + t) + \varphi(K - t))}{2\varphi(t)}\right]^{-1},$$

where $K > 0$ satisfies (3.1).

For $t \geq 1.645$, a very accurate approximation to K can be obtained from the following iterative formula (starting with $K_0 = t$):

$$K_{i+1} = t + [2 \log(K_i/\Phi(K_i - t)) - 1.838]^{1/2}.$$

Convergence is usually achieved after only 2 or 3 iterations. In addition, Figures 1 and 2 give values of K and $\underline{B}$ for various values of t in this problem. For easier comparisons, Table 6 gives $\underline{Pr}(H_0 \mid x, G_{US})$ for some specific important values of t, and $\pi_0 = \frac{1}{2}$.

Comparison of Table 6 with Table 5 shows that

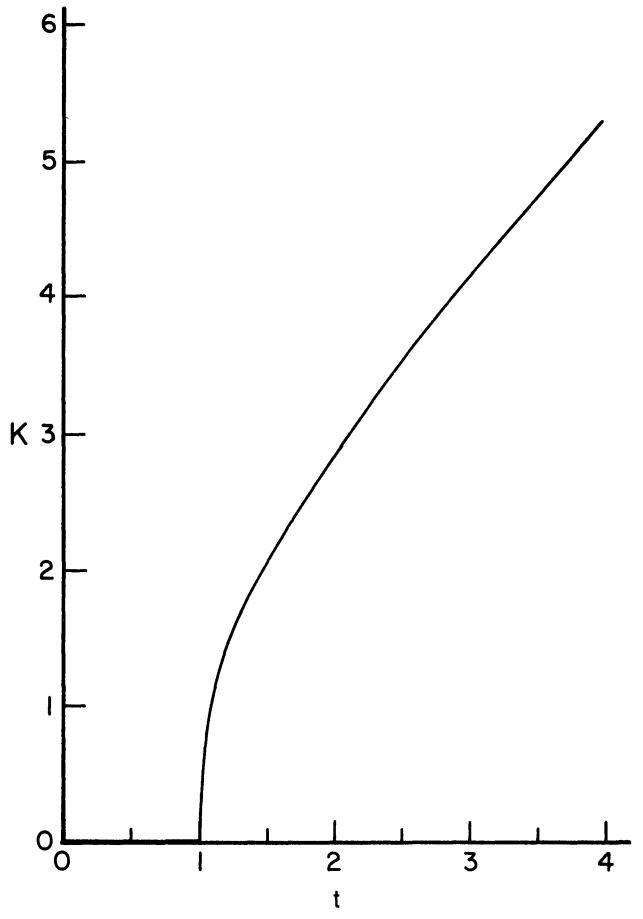

Figure 1. Minimizing Value of K When $G = G_{US}$.

$\underline{\Pr}(H_0 \mid x, G_{US})$ is only moderately larger than $\underline{\Pr}(H_0 \mid x, G_S)$ for P values of .10 or .05. The asymptotic behavior (as $t \to \infty$) of the two lower bounds, however, is very different, as the following theorem shows.

Theorem 7. For $t > 0$ and $\pi_0 = \frac{1}{2}$ in Example 1,

$$\underline{\Pr}(H_0 \mid x, G_{US})/(pt^2) > 1.$$

Furthermore,

$$\lim_{t\to\infty} \underline{\Pr}(H_0 \mid x, G_{US})/(pt^2) = 1.$$

Proof. For $t > 2.26$, the previously mentioned Mills ratio inequalities were used together with the easily verified (for $t > 2.26$) inequality $\underline{B}(x, G_{US}) > 2t\varphi(t)$. The inequality was verified numerically for $0 < t \leq 2.26$.

3.5 Lower Bounds for G_{NOR} = {Normal Distributions}

We have seen that minimizing $\Pr(H_0 \mid x)$ over $g \in G_{US}$ is the same as minimizing over $g \in \mathcal{U}_S$. Although using $\mathcal{U}_S$ is much more reasonable than using G_A, there is still some residual bias against H_0 involved in using $\mathcal{U}_S$. Prior opinion densities typically look more like a normal density or a Cauchy density than a uniform density. What happens when $\Pr(H_0 \mid x)$ is minimized over $g \in G_{NOR}$, that is, over scale transformations of a symmetric normal distribution, rather than over scale transformations of a symmetric uniform distribution? This question was investigated by Edwards et al. (1963, pp. 229–231).

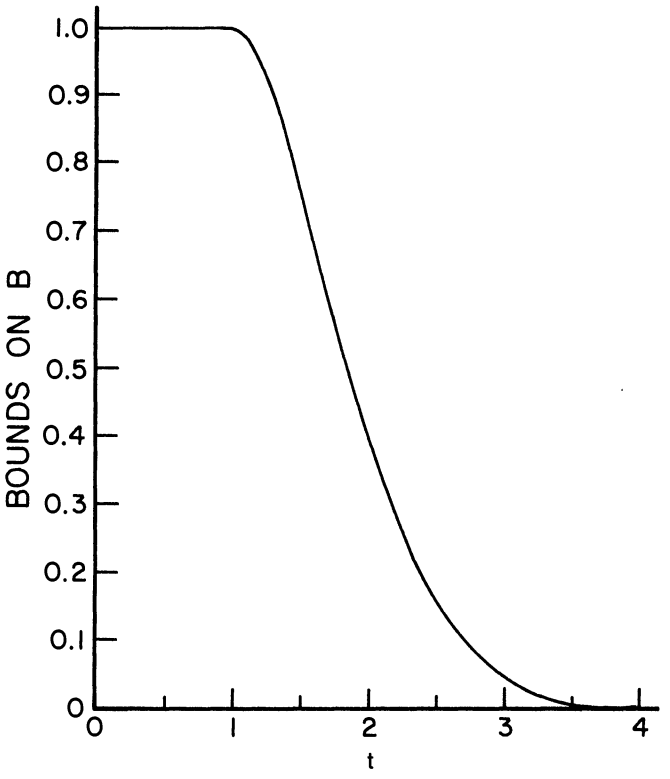

Figure 2. Values of $\underline{B}(x, G_{US})$ in the Normal Example.

Theorem 8. (See Edwards et al. 1963). If $t \leq 1$ in Example 1, then $\underline{B}(x, G_{NOR}) = 1$ and $\underline{\Pr}(H_0 \mid x, G_{NOR}) = \pi_0$. If $t > 1$, then

$$\underline{B}(x, G_{NOR}) = \sqrt{e}\, t\, e^{-t^2/2}$$

and

$$\underline{\Pr}(H_0 \mid x, G_{NOR}) = \left[1 + \frac{(1 - \pi_0)}{\pi_0} \times \frac{\exp\{t^2/2\}}{\sqrt{e}\, t}\right]^{-1}.$$

Table 7 gives $\underline{\Pr}(H_0 \mid x, G_{NOR})$ for several values of t. Except for larger t, the results for G_{NOR} are similar to those for G_{US}, and the comparative simplicity of the formulas in Theorem 8 might make them the most attractive lower bounds.

A graphical comparison of the lower bounds $\underline{B}(x, G)$, for the four G's considered, is given in Figure 3. Although the vertical differences are larger than the visual discrepancies, the closeness of the bounds for G_{US} and G_{NOR} is apparent.

Table 6. Comparison of P Values and $\underline{\Pr}(H_0 \mid x, G_{US})$ When $\pi_0 = \frac{1}{2}$

P Value (p)	t	$\underline{\Pr}(H_0 \mid x, G_{US})$	$\underline{\Pr}(H_0 \mid x, G_{US})/(pt^2)$
.10	1.645	.390	1.44
.05	1.960	.290	1.51
.01	2.576	.109	1.64
.001	3.291	.018	1.66

Table 7. Comparison of P Values and $\underline{Pr}(H_0 \mid x, G_{NOR})$ When $\pi_0 = \frac{1}{2}$

P Value (p)	t	$\underline{Pr}(H_0 \mid x, G_{NOR})$	$\underline{Pr}(H_0 \mid x, G_{NOR})/(pt^2)$
.10	1.645	.412	1.52
.05	1.960	.321	1.67
.01	2.576	.133	2.01
.001	3.291	.0235	2.18

4. MORE GENERAL HYPOTHESES AND CONDITIONAL CALCULATIONS

4.1 General Formulation

To verify some of the statements made in the Introduction, consider the Bayesian calculation of $\Pr(H_0 \mid A)$, where H_0 is of the form $H_0 : \theta \in \Theta_0$ [say, $\Theta_0 = (\theta_0 - b, \theta_0 + b)$] and A is the set in which x is known to reside (A may be $\{x\}$, or a set such as $\{x: \sqrt{n}|\bar{x} - \theta_0|/\sigma \geq 1.96\}$). Then, letting π_0 and π_1 again denote the prior probabilities of H_0 and H_1 and introducing g_0 and g_1 as the densities on Θ_0 and $\Theta_1 = \Theta_0^c$ (the complement of Θ_0), respectively, which describe the spread of the prior mass on these sets, it is straightforward to check that

$$\Pr(H_0 \mid A) = \left[1 + \frac{(1 - \pi_0)}{\pi_0} \times \frac{m_{g_1}(A)}{m_{g_0}(A)}\right]^{-1}, \quad (4.1)$$

where

$$m_{g_i}(A) = \int_{\Theta_i} \Pr_\theta(A) g_i(\theta)\, d\theta. \quad (4.2)$$

One claim made in the Introduction was that, if $\Theta_0 = (\theta_0 - b, \theta_0 + b)$ with b suitably small, then approximating

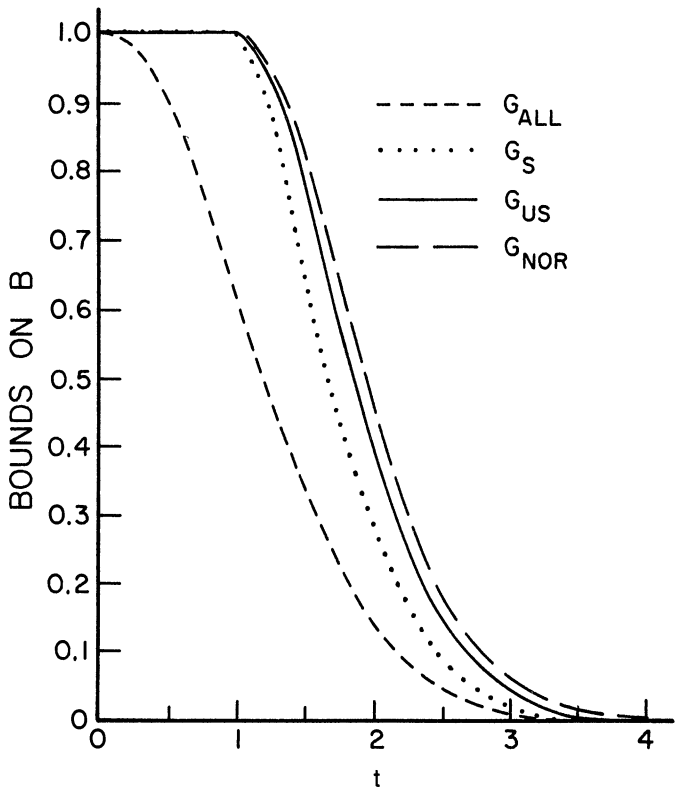

Figure 3. Values of $\underline{B}(x, G)$ in the Normal Example for Different Choices of G.

H_0 by $H_0 : \theta = \theta_0$ is a satisfactory approximation. From (4.1) and (4.2), it is clear that this will hold from the Bayesian perspective when $f(x \mid \theta)$ is approximately constant on Θ_0 [so $m_{g_0}(x) = \int_{\Theta_0} f(x \mid \theta) g_0(\theta)\, d\theta \cong f(x \mid \theta_0)$; here we are assuming that $A = \{x\}$]. Note, however, that g_1 is defined to give zero mass to Θ_0, which might be important in the ensuing calculations.

For the general formulation, one can determine lower bounds on $\Pr(H_0 \mid A)$ by choosing sets G_0 and G_1 of g_0 and g_1, respectively, calculating

$$\underline{B}(A, G_0, G_1) = \inf_{g_0 \in G_0} m_{g_0}(A) / \sup_{g_1 \in G_1} m_{g_1}(A), \quad (4.3)$$

and defining

$$\underline{\Pr}(H_0 \mid A, G_0, G_1) = \left[1 + \frac{(1 - \pi_0)}{\pi_0} \times \frac{1}{\underline{B}(A, G_0, G_1)}\right]^{-1}. \quad (4.4)$$

4.2 More General Hypotheses

Assume in this section that $A = \{x\}$ (i.e., we are in the usual inference model of observing the data). The lower bounds in (4.3) and (4.4) can be applied to a variety of generalizations of point null hypotheses and still exhibit the same type of conflict between posterior probabilities and P values that we observed in Section 3. Indeed, if Θ_0 is a small set about θ_0, the general lower bounds turn out to be essentially equivalent to the point null lower bounds. The following is an example.

Theorem 9. In Example 1, suppose that the hypotheses were $H_0 : \theta \in (\theta_0 - b, \theta_0 + b)$ and $H_1 : \theta \notin (\theta_0 - b, \theta_0 + b)$. If $|t - \sqrt{n}\, b/\sigma| \geq 1$ (which must happen for a classical test to reject H_0) and $G_0 = G_1 = G_S$ (the class of all symmetric distributions about θ_0), then $\underline{B}(x, G_0, G_1)$ and $\underline{\Pr}(H_0 \mid x, G_0, G_1)$ are *exactly the same as* $\underline{B}$ and $\underline{P}$ for testing the point null.

Proof. Under the assumption on b, it can be checked that the minimizing g_0 is the unit point mass at θ_0 [the interval $(\theta_0 - b, \theta_0 + b)$ being in the convex part of the tail of the likelihood function], whereas the maximization over G_1 is the same as before.

Another type of testing situation that yields qualitatively similar lower bounds is that of testing, say, $H_0 : \theta = \theta_0$ versus $H_1 : \theta > \theta_0$. It is assumed, here, that $\theta = \theta_0$ still corresponds to a well-defined theory to which one would ascribe probability π_0 of being true, but it is now presumed that negative values of θ are known to be impossible. Analogs of the results in Section 3 can be obtained for this situation; note, for instance, that $G = G_A = \{\text{all distributions}\}$ will yield the same lower bounds as in Theorem 1 in Section 3.2.

4.3 Posterior Probabilities Conditional on Sets

We revert here to considering $H_0 : \theta = \theta_0$ and use the general lower bounds in (4.3) and (4.4) to establish the two results mentioned in Section 1 concerning conditioning on sets of data. First, in the example of the "astronomer"

in Section 1, a lower bound on the long-run proportion of true null hypotheses is

$$\underline{\Pr}(H_0 \mid A) = \left[1 + \tfrac{1}{2} \times \frac{\sup_{g_1} m_{g_1}(A)}{P_{\theta_0}(A)}\right]^{-1},$$

where $A = \{x: 1.96 < t \leq 2.0\}$. Note that $\Pr_{\theta_0}(A) = 2[\Phi(2.0) - \Phi(1.96)] = .0044$, whereas

$$\sup_{g_1} m_{g_1}(A) = \sup_{\theta} \Pr_{\theta}(A) \cong \Phi(.02) - \Phi(-.02) = .016.$$

Hence $\underline{\Pr}(H_0 \mid A) \cong [1 + (.016)/(.0044)]^{-1} = .22$, as stated.

Finally, we must establish the correspondence between the P value and the posterior probability of H_0 when the data, x, are replaced by the cruder knowledge that $x \in A = \{y : T(y) \geq T(x)\}$. [Note that $\Pr_{\theta_0}(A) = p$, the P value.] A similar analysis was given in Dickey (1977). Clearly,

$$\underline{B}(A, G) = \Pr_{\theta_0}(A)/\sup_{g \in G} m_g(A)$$

$$= p/\sup_{g \in G} m_g(A),$$

so, when $\pi_0 = \frac{1}{2}$,

$$\underline{\Pr}(H_0 \mid A, G) = [1 + \sup_{g \in G} m_g(A)/p]^{-1}.$$

Now, for *any* of the classes G considered in Section 3, it can be checked in Example 1 that

$$\sup_{g \in G} m_g(A) = 1;$$

it follows that $\underline{\Pr}(H_0 \mid A, G) = (1 + p^{-1})^{-1}$, which for small p is approximately equal to p.

5. CONCLUSIONS AND GENERALIZATIONS

Comment 1. A rather fascinating "empirical" observation follows from graphing (in Example 1) $\underline{B}(x, G_{US})$ and the P value calculated at $(t - 1)^+$ [the positive part of $(t - 1)$] instead of t; this last will be called the "P value of $(t - 1)^+$" for brevity. Again, $\underline{B}(x, G_{US})$ can be considered to be a reasonable lower bound on the comparative likelihood measure of the evidence against H_0 (under symmetry and unimodality restrictions on the "weighted likelihood" under H_1). Figure 4 shows that this comparative likelihood (or Bayes factor) is close to the P value that would be obtained if we replaced t by $(t - 1)^+$. The implication is that the "commonly perceived" rule of thumb, that $t = 1$ means only mild evidence against H_0, $t = 2$ means significant evidence against H_0, $t = 3$ means highly significant evidence against H_0, and $t = 4$ means overwhelming evidence against H_0, should, at the very least, be replaced by the rule of thumb $t = 1$ means no evidence against H_0, $t = 2$ means only mild evidence against H_0, $t = 3$ means significant evidence against H_0, and $t = 4$ means highly significant evidence against H_0, and even this may be overstating the evidence against H_0 (see Comments 3 and 4).

Comment 2. We restricted analysis to the case of univariate θ, so as not to lose sight of the main ideas. We are currently looking at a number of generalizations to higher-dimensional problems. It is rather easy to see that the G_A bound is not very useful in higher dimensions, becoming very small as the dimension increases. (This is not unexpected, since concentrating all mass on the MLE under the alternative becomes less and less reasonable as the dimension increases.) The bounds for spherically symmetric (about θ_0) classes of priors (or, more generally, invariant priors) seem to be quite reasonable, however, comparable with or larger than the one-dimensional bounds.

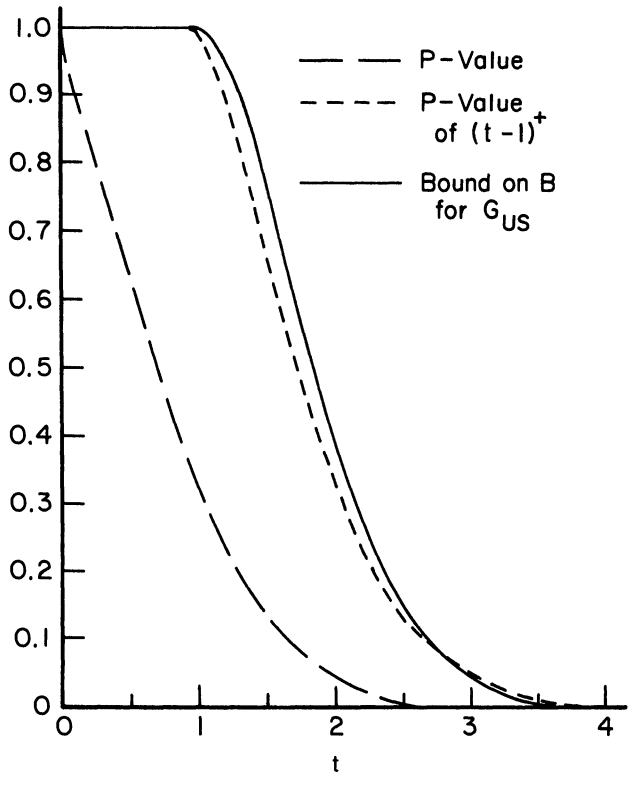

Figure 4. Comparison of $\underline{B}(x, G_{US})$ and P Values.

An alternative (but closely related) idea being considered for dealing with high dimensions is to consider the classical test statistic, $T(X)$, that would be used and replace $f(x \mid \theta)$ by $f_T(t \mid \theta)$, the corresponding density of T. In goodness-of-fit problems, for instance, $T(X)$ is often the chi-squared statistic, having a central chi-squared distribution under H_0 and a noncentral chi-squared distribution under contiguous alternatives (see Cressie and Read 1984). Writing the noncentrality parameter as η, we could reformulate the test as one of $H_0 : \eta = 0$ versus $H_1 : \eta > 0$ [assuming, of course, that contiguous alternatives are felt to be satisfactory; it seems likely, in any case, that the lower bound on $\Pr(H_0 \mid x)$ will be achieved by g concentrating on such alternatives]. Thus the problem has been reduced to a one-dimensional problem and our techniques can apply. Note the usefulness of much of classical testing theory to this enterprise; determining a suitable T and its distribution forms the bulk of a classical analysis and would also form the basis for calculating the bounds on $\Pr(H_0 \mid x)$.

Comment 3. What should a statistician desiring to test a point null hypothesis do? Although it seems clearly

unacceptable to use a P value of .05 as evidence to reject, the lower bounds on $\Pr(H_0 \mid x)$ that we have considered can be argued to be of limited usefulness; if the lower bound is large we know not to reject H_0, but if the lower bound is small we still do not know if H_0 can be rejected [a small lower bound not necessarily meaning that $\Pr(H_0 \mid x)$ is itself small]. One possible solution is to seek upper bounds for $\Pr(H_0 \mid x)$, an approach taken with some success in Edwards et al. (1963) and Dickey (1973). The trouble is that these upper bounds do require "nonobjective" subjective input about g. It seems reasonable, therefore, to conclude that we must embrace subjective Bayesian analysis, in some form, to reach sensible conclusions about testing a point null. Perhaps the most attractive possibility, following Dickey (1973), is to communicate $B_g(x)$ or $\Pr(H_0 \mid x)$ for a wide range of prior inputs, allowing the user to choose, easily, his own prior and also to see the effect of the choice of prior. In Example 1, for instance, it would be a simple matter in a given problem to consider all $\mathcal{N}(\mu, \tau^2)$ priors for g and present a contour graph of $B_g(x)$ with respect to the variables μ and τ^2. The reader of the study can then choose μ (often to equal θ_0) and τ^2 and immediately determine B or $\Pr(H_0 \mid x)$ (the latter necessitating a choice of π_0 also, of course). And by varying μ and τ^2 over reasonable ranges, the reader could also determine robustness or sensitivity to prior inputs. Note that the functional form of g will not usually have a great effect on $\Pr(H_0 \mid x)$ [replacing the $\mathcal{N}(\mu, \tau^2)$ priors by Cauchy priors would cause a substantial change only for very extreme x], so one can usually get away with choosing a convenient form with parameters that are easily accessible to subjective intuition. [If there was concern about the choice of a functional form for g, the more sophisticated robustness analysis of Berger and Berliner (1986) could be performed, an analysis that yields an interval of values for $\Pr(H_0 \mid x)$ as the prior ranges over all distributions "close" to an elicited prior.] General discussions of presentation of $\Pr(H_0 \mid x)$, as a function of subjective inputs, can be found in Dickey (1973) and Berger (1985).

Comment 4. If one insisted on creating a "standardized" significance test for common use (as opposed to the flexible Bayesian reporting discussed previously) it would seem that the tests proposed by Jeffreys (1961) are quite suitable. For small and moderate n in Table 1, $\Pr(H_0 \mid x)$ is not too far from the objective lower bounds in Table 6, say, indicating that the choice of a Jeffreys-type prior does not excessively bias the results in favor of H_0. As n increases, the exact $\Pr(H_0 \mid x)$ and the lower bound diverge, but this is due to the inadequacy of the lower bound (which does not depend on n).

Comment 5. Although for most statistical problems it is the case that, say, $\underline{\Pr}(H_0 \mid x, G_{US})$ is substantially larger than the P value for x, this need not always be so, as the following example demonstrates.

Example 2. Suppose that a single Cauchy $(\theta, 1)$ observation, X, is obtained and it is desired to test $H_0 : \theta = 0$ versus $H_1 : \theta \neq 0$. It can then be shown that (for $\pi_0 = \frac{1}{2}$)

$$\lim_{|x|\to\infty} \frac{\underline{B}(x, G_{US})}{P \text{ value}} = \lim_{|x|\to\infty} \frac{\underline{\Pr}(H_0 \mid x, G_{US})}{P \text{ value}} = 1,$$

so the P value does correspond to the evidentiary lower bounds for large $|x|$ (see Table 8 for comparative values when $|x|$ is small). Also of interest in this case is analysis with the priors $G_C = \{\text{all Cauchy distributions}\}$, since one can prove that, for $|x| \geq 1$ and $\pi_0 = \frac{1}{2}$,

$$\underline{B}(x, G_C) = \frac{2|x|}{(1 + x^2)} \quad \text{and} \quad \underline{\Pr}(H_0 \mid x, G_C) = \frac{2|x|}{(1 + |x|)^2}$$

[whereas $\underline{B}(x, G_C) = 1$ and $\underline{\Pr}(H_0 \mid x, G_C) = \frac{1}{2}$ for $|x| \leq 1$]. Table 8 presents values of all of these quantities for $\pi_0 = \frac{1}{2}$ and varying $|x|$.

Although it is tempting to take comfort in the closer correspondence between the P value and $\underline{\Pr}(H_0 \mid x, G_{US})$ here, a different kind of Bayesian conflict occurs. This conflict arises from the easily verifiable fact that, for *any fixed* g,

$$\lim_{|x|\to\infty} B_g(x) = 1 \quad \text{and} \quad \lim_{|x|\to\infty} \Pr(H_0 \mid x) = \pi_0, \tag{5.1}$$

so large x provides *no information* to a Bayesian. Thus, rather than this being a case in which the P value might have a reasonable evidentiary interpretation because it agrees with $\underline{\Pr}(H_0 \mid x, G_{US})$, this is a case in which $\underline{\Pr}(H_0 \mid x, G_{US})$ is itself highly suspect as an evidentiary conclusion.

Note also that the situation of a single Cauchy observation is not even irrelevant to normal theory analysis; the standard Bayesian method of analyzing the normal problem with unknown variance, σ^2, is to integrate out the nuisance parameter σ^2, using a noninformative prior. The resulting "marginal likelihood" for θ is essentially a t distribution with $(n - 1)$ degrees of freedom (centered at $\bar{x}$); thus if $n = 2$, we are in the case of a Cauchy distribution. As noted in Dickey (1977), it is actually the case that, for *any* n in this problem, the marginal likelihood is

Table 8. $\underline{B}$ and $\underline{\Pr}$ for a Cauchy Distribution When $\pi_0 = \frac{1}{2}$

P Value (p)	$\lvert x\rvert$	$\underline{B}(x, G_{US})$	$\underline{\Pr}(H_0 \mid x, G_{US})$	$\underline{B}(x, G_C)$	$\underline{\Pr}(H_0 \mid x, G_C)$
.50	1.000	.894	.472	1.000	.500
.20	3.080	.351	.260	.588	.370
.10	6.314	.154	.133	.309	.236
.05	12.706	.069	.064	.156	.135
.01	63.657	.0115	.0114	.031	.030
.0032	200	.0034	.0034	.010	.010

such that (5.1) holds. (Of course, the initial use of a noninformative prior for σ^2 is not immune to criticism.)

Comment 6. Since any unimodal symmetric distribution is a mixture of symmetric uniforms and a Cauchy distribution is a mixture of normals, it is easy to establish the interesting fact that (for *any* situation and *any* x)

$$\underline{B}(x, G_{US}) = \underline{B}(x, \mathcal{U}_S) \leq \underline{B}(x, G_{NOR}) \leq \underline{B}(x, G_C).$$

The same argument and inequalities also hold with G_C replaced by the class of all t distributions of a given degree of freedom.

[*Received January 1985. Revised October 1985.*]

REFERENCES

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

Berger, J., and Berliner, L. M. (1986), "Robust Bayes and Empirical Bayes Analysis with ε-Contaminated Priors," *The Annals of Statistics*, 14, 461–486.

Berkson, J. (1938), "Some Difficulties of Interpretation Encountered in the Application of the Chi-Square Test," *Journal of the American Statistical Association*, 33, 526–542.

——— (1942), "Tests of Significance Considered as Evidence," *Journal of the American Statistical Association*, 37, 325–335.

Cressie, N., and Read, T. R. C. (1984), "Multinomial Goodness-Of-Fit Tests," *Journal of the Royal Statistical Society*, Ser. B, 46, 440–464.

DeGroot, M. H. (1973), "Doing What Comes Naturally: Interpreting a Tail Area as a Posterior Probability or as a Likelihood Ratio," *Journal of the American Statistical Association*, 68, 966–969.

Dempster, A. P. (1973), "The Direct Use of Likelihood for Significance Testing," in *Proceedings of the Conference on Foundational Questions in Statistical Inference*, ed. O. Barndorff-Nielsen, University of Aarhus, Dept. of Theoretical Statistics, 335–352.

Diamond, G. A., and Forrester, J. S. (1983), "Clinical Trials and Statistical Verdicts: Probable Grounds for Appeal," *Annals of Internal Medicine*, 98, 385–394.

Dickey, J. M. (1971), "The Weighted Likelihood Ratio, Linear Hypotheses on Normal Location Parameters," *Annals of Mathematical Statistics*, 42, 204–223.

——— (1973), "Scientific Reporting," *Journal of the Royal Statistical Society*, Ser. B, 35, 285–305.

——— (1974), "Bayesian Alternatives to the F-Test and Least Squares Estimate in the Normal Linear Model," in *Studies in Bayesian Econometrics and Statistics*, eds. S. E. Fienberg and A. Zellner, Amsterdam: North-Holland, pp. 515–554.

——— (1977), "Is the Tail Area Useful as an Approximate Bayes Factor?," *Journal of the American Statistical Association*, 72, 138–142.

——— (1980), "Approximate Coherence for Regression Models With a New Analysis of Fisher's Broadback Wheatfield Example," in *Bayesian Analysis in Econometrics and Statistics: Essays in Honor of Harold Jeffreys*, ed. A. Zellner, Amsterdam: North-Holland, pp. 333–354.

Dickey, J. M., and Lientz, B. P. (1970), "The Weighted Likelihood Ratio, Sharp Hypotheses About Chances, the Order of a Markov Chain," *Annals of Mathematical Statistics*, 41, 214–226.

Edwards, A. W. F. (1972), *Likelihood*, Cambridge, U.K.: Cambridge University Press.

Edwards, W., Lindman, H., and Savage, L. J. (1963), "Bayesian Statistical Inference for Psychological Research," *Psychological Review*, 70, 193–242. [Reprinted in *Robustness of Bayesian Analyses*, 1984, ed. J. Kadane, Amsterdam: North-Holland.]

Feller, W. (1968), *An Introduction to Probability Theory and Its Applications* (Vol. 1, 3rd ed.), New York: John Wiley.

Good, I. J. (1950), *Probability and the Weighing of Evidence*, London: Charles W. Griffin.

——— (1958), "Significance Tests in Parallel and in Series," *Journal of the American Statistical Association*, 53, 799–813.

——— (1965), *The Estimation of Probabilities: An Essay on Modern Bayesian Methods*, Cambridge, MA: MIT Press.

——— (1967), "A Bayesian Significance Test for Multinomial Distributions," *Journal of the Royal Statistical Society*, Ser. B, 29, 399–431.

——— (1983), "*Good Thinking: The Foundations of Probability and Its Applications*," Minneapolis: University of Minnesota Press.

——— (1984), Notes C140, C144, C199, C200, and C201, *Journal of Statistical Computation and Simulation*, 19.

Hill, B. (1982), Comment on "Lindley's Paradox," by Glenn Shafer, *Journal of the American Statistical Association*, 77, 344–347.

Hildreth, C. (1963), "Bayesian Statisticians and Remote Clients," *Econometrika*, 31, 422–438.

Hodges, J. L., Jr., and Lehmann, E. L. (1954), "Testing the Approximate Validity of Statistical Hypotheses," *Journal of the Royal Statistical Society*, Ser. B, 16, 261–268.

Jeffreys, H. (1957), *Scientific Inference*, Cambridge, U.K.: Cambridge University Press.

——— (1961), *Theory of Probability* (3rd ed.), Oxford, U.K.: Oxford University Press.

——— (1980), "Some General Points in Probability Theory," in *Bayesian Analysis in Econometrics and Statistics*, ed. A. Zellner, Amsterdam: North-Holland, pp. 451–454.

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

Leamer, E. E. (1978), *Specification Searches: Ad Hoc Inference With Nonexperimental Data*, New York: John Wiley.

Lempers, F. B. (1971), *Posterior Probabilities of Alternative Linear Models*, Rotterdam: University of Rotterdam Press.

Lindley, D. V. (1957), "A Statistical Paradox," *Biometrika*, 44, 187–192.

——— (1961), "The Use of Prior Probability Distributions in Statistical Inference and Decision," in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley: University of California Press, pp. 453–468.

——— (1965), *Introduction to Probability and Statistics From A Bayesian Viewpoint* (Parts 1 and 2), Cambridge, U.K.: Cambridge University Press.

——— (1977), "A Problem in Forensic Science," *Biometrika*, 64, 207–213.

Pratt, J. W. (1965), "Bayesian Interpretation of Standard Inference Statements" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 27, 169–203.

Raiffa, H., and Schlaifer, R. (1961), *Applied Statistical Decision Theory*, Harvard University, Division of Research, Graduate School of Business Administration.

Shafer, G. (1982), "Lindley's Paradox," *Journal of the American Statistical Association*, 77, 325–351.

Smith, A. F. M., and Spiegelhalter, D. J. (1980), "Bayes Factors and Choice Criteria for Linear Models," *Journal of the Royal Statistical Society*, Ser. B, 42, 213–220.

Smith, C. A. B. (1965), "Personal Probability and Statistical Analysis," *Journal of the Royal Statistical Society*, Ser. A, 128, 469–499.

Solo, V. (1984), "An Alternative to Significance Tests," Technical Report 84-14, Purdue University, Dept. of Statistics.

Zellner, A. (1971), *An Introduction to Bayesian Inference in Econometrics*, New York: John Wiley.

——— (1984), "Posterior Odds Ratios for Regression Hypotheses: General Considerations and Some Specific Results," in *Basic Issues in Econometrics*, Chicago: University of Chicago Press, pp. 275–305.

Zellner, A., and Siow, A. (1980), "Posterior Odds Ratios for Selected Regression Hypotheses," in *Bayesian Statistics*, eds. J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith, Valencia: University Press, pp. 586–603.