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

by

James Berger and Thomas Sellke
Purdue University

Technical Report #84-27

Department of Statistics
Purdue University

Revised October 1985

*Research Supported by the National Science Foundation Under Grant #DMS-8401996.

Title: Testing a Point Null Hypothesis: The Irreconcilability of P-values and Evidence*

Short Title: Testing a Point Null Hypothesis

*James O. Berger is Professor and Thomas Sellke is Assistant Professor, Department of Statistics, Purdue University, West Lafayette, IN 47907.
Research was supported by NSF Grant No. DMS-8401996. The authors are grateful to L. Mark Berliner, Iain Johnstone, Robert Keener, Prem Puri, and Herman Rubin for suggestions or interesting arguments.

Abstract

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 which yields a P-value of .05, when testing a normal mean, results 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.

1. INTRODUCTION

We consider the simple situation of observing a random quantity X having density (for convenience) $f(x|\theta)$, θ being an unknown parameter assuming values in a parameter space $\Theta \subset \mathbb{R}^1$. 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 Section 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

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

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

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

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

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

where Φ is the standard normal c.d.f. and

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

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 (i) most statisticians prefer use of P-values, feeling it to be important to indicate how strong the evidence against H_0 is (cf. Kiefer (1977)); and (ii) 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 Hodges and Lehmann (1954), Good (1983), 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 Jeffreys (1961), Lindley (1957), Shafer (1982), and Good (1983).) Both 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 <u>exactly</u> (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 a $\eta(\theta_0, \sigma^2)$ density. (This prior is close to that recommended by Jeffreys (1961) for testing a point null, though he actually recommends 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, $P(H_0|x)$, of H_0 is given by

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

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

Table 1. $P(H_0|x)$ For Jeffreys-Type Prior

		n						
p	t	1	5	10	20	50	100	1000
.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

The conflict between p and $P(H_0|x)$ is apparent. If n = 50 and t = 1.960, one can classically "reject H_0 at significance level p = .05," while $P(H_0|x)$ = .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 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 $P(H_0|x)$ are determined in Section 3, and are given in Table 2 for certain values of t.

Table 2. $P(H_0|x)$ For a Prior Biased Towards H_1

P-value (p)	t	$P(H_0|x)$
.10	1.645	.340
.05	1.960	.227
.01	2.576	.068
.001	3.291	.0088

Again the numbers are astonishing. Although p = .05 when t = 1.96 is observed, even a Bayesian analysis strongly biased towards H_1 states that the null has a .227 probability of being true, evidence against the null which 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, in order to produce a posterior probability of $P(H_0|x)$ = .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 towards 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 prior above is dropped, i.e. 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) shows that, if each hypothesis is given initial probability $\frac{1}{2}$, the unrestricted "most favorable to H_1" prior yields

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

the values of which are still substantially higher than p (e.g., when $t = 1.96$, $p = .05$ while $P(H_0|x) = .128$).

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

$$\ell_x(\theta_1:\theta_2) = f(x|\theta_1)/f(x|\theta_2)$$

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

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

Values of $\underline{\ell}_x$ for various t are given in Table 3.

Table 3. Bounds on the Comparative Likelihood

P-value (p)	t	Likelihood Ratio Lower Bound ($\underline{\ell}_x$)
.10	1.645	.258
.05	1.960	.146
.01	2.576	.036
.001	3.291	.0044

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 strongly bias 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) states, concerning the answers obtained 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."

Suppose 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 he first noticed that, overall, about half the point nulls were false and half were true. He then concentrates attention on the subset he is interested in, 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 H_0 should be confidently rejected.

In probability language, the "experiment" of the astronomer 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", than only 5% will have $t \geq 1.96$. This is, of course, true, but is not the answer the astronomer was interested in. 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 texts 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 $P(H_0|x)$ (cf. 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 $P(H_0|x)$ and $\underline{\ell}_x$ 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 $A = \{y: T(y) \geq T(x)\}$, and then calculates $p = P_{\theta=\theta_0}(A)$. 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 <u>equal</u> p (see Section 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 X is measured by a weighing scale which occassionally "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) says

> "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."

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 ESP 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) argue 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 (cf. 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 paper, concentrating on determination of lower bounds for $P(H_0|x)$ under various types of prior assumptions. The single prior Jeffreys analysis is one extreme; the Edwards, Lindman, and Savage (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 non-symmetric prior implies that there are specifically favored alternative values of θ.

Section 2 reviews basic features of the calculation of $P(H_0|x)$, and discusses the Bayesian literature on testing a point null hypothesis. Section 3 presents the various lower bounds on $P(H_0|x)$. 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 $\pi_1 = 1-\pi_0$ denote the prior probability of H_1; furthermore, suppose 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 $\theta = \theta_0$ is thought to possibly 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

$$m(x) = f(x|\theta_0)\ \pi_0 + (1-\pi_0)m_g(x), \tag{2.1}$$

where

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

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

$$P(H_0|x) = f(x|\theta_0) \cdot \pi_0/m(x)$$

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

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

$$\frac{P(H_0|x)}{1-P(H_0|x)} = \frac{\pi_0}{(1-\pi_0)} \cdot \frac{f(x|\theta_0)}{m_g(x)} . \qquad (2.3)$$

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

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

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 $P(H_0|x)$ 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 that 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 a $\mathcal{N}(\theta_0,\sigma^2(1+n^{-1}))$ distribution. Thus

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

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

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

and

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

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

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 $P(H_0|x) \to 1$ no matter how small the P-value.)

When giving numerical results, we will tend to present $P(H_0|x)$ 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 (personal communication) 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 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, Lindman, and Savage (1963), Smith (1965), Dickey and Leintz (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 $P(H_0|x)$ 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 papers which find lower bounds on B_g and $P(H_0|x)$ that are similar to those we consider include Edwards, Lindman and Savage (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 $P(H_0|x)$ 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 $P(H_0|x)$ which 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 which are subsets of the real line) and will concentrate on the following four classes of g:

$$
\begin{aligned}
G_A &= \{\text{all distributions}\},\\
G_S &= \{\text{all distributions symmetric about } \theta_0\},\\
G_{US} &= \{\text{all unimodal distributions symmetric about } \theta_0\},\\
G_{NOR} &= \{\text{all } \mathcal{N}(\theta_0,\tau^2) \text{ distributions, } 0 \leq \tau^2 < \infty\}.
\end{aligned}
$$

Even though these G's are supposed to consist only of distributions on $\{\theta|\theta\neq\theta_0\}$, it will be convenient to allow them to include distributions with mass at θ_0, so that the lower bounds we compute are always attained; the answers are unchanged by this simplification, and cumbersome limiting notation is avoided. Letting

$$
\underline{P}(H_0|x,G) = \inf_{g \in G} P(H_0|x)
$$

and

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

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

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

and

$$\underline{P}(H_0|x,G) = [1 + \frac{(1-\pi_0)}{\pi_0} \cdot \frac{1}{\underline{B}(x,G)}]^{-1}.$$

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 that $\underline{B}(x,G)$ 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, Lindman, and Savage (1963). The proof is elementary and will be omitted.

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

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

and

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

(Note that $\underline{B}(x,G_A)$ is equal to the comparative likelihood bound, $\underline{\ell}_x$, 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

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

For several choices of t, Table 2 gives the corresponding two-sided P-values, p, and the values of $\underline{P}(H_0|x, G_A)$, with $\pi_0 = \frac{1}{2}$.

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

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

Note that the lower bounds on $P(H_0|x)$ are considerably larger than the corresponding P-values, in spite of the fact that minimization of $P(H_0|x)$ over $g \in G_A$ is "maximally unfair" to the null hypothesis. The last column shows that the ratio of $P(H_0|x, G_A)$ to pt is rather stable. The behavior of this ratio is described in more detail by Theorem 2.

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

$$\frac{P(H_0|x,G_A)}{pt} > \sqrt{\frac{\pi}{2}} \cong 1.253.$$

Furthermore,

$$\lim_{t\to\infty} \frac{P(H_0|x,G_A)}{pt} = \sqrt{\frac{\pi}{2}}\,.$$

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

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

The left inequality here is from Feller (1968), page 175, and the right inequality can be proved 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 $P(H_0|x)$ 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 $P(H_0|x)$ and the P-value appears to be decreasing in t, however.)

3.3 Lower Bounds for G_S = {Symmetric Distributions}

There is a large gap between $\underline{P}(H_0|x, G_A)$ (for $\pi_0 = \frac{1}{2}$) and $P(H_0|x)$ 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 $P(H_0|x)$ over all $g \in G_S$ is equivalent to minimizing over the class

$$G_{2PS} = \{\text{all symmetric two-point distributions}\}.$$

Theorem 3.

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

so that

$$\underline{B}(x,G_{2PS}) = \underline{B}(x,G_S) \text{ and } \underline{P}(H_0|x,G_{2PS}) = \underline{P}(H_0|x,G_S).$$

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 which 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 $\underline{P}(H_0|x,G_S) = \pi_0$ for $t \leq 1$. (Since the point mass of θ_0 is not really a legitimate prior on

$\{\theta|\theta\neq\theta_0\}$, this means that observing $t \leq 1$ actually constitutes evidence in favor of H_0 for any real symmetric prior on $\{\theta|\theta\neq\theta_0\}$.)

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 that

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

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 $\underline{P}(H_0|x,G_S)$ for several choices of t, again with $\pi_0 = \frac{1}{2}$.

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

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

The ratio $\underline{P}(H_0|x,G_S)/\underline{P}(H_0|x,G_A)$ 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 $P(H_0|x,G_S)/(p\,t)$. 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,

$$\frac{\underline{P}(H_0|x,G_S)}{pt} > \sqrt{2\pi} \cong 2.507.$$

Furthermore,

$$\lim_{t\to\infty} \frac{\underline{P}(H_0|x,G_S)}{pt} = \sqrt{2\pi}.$$

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

Minimizing $P(H_0|x)$ 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 G_{US}. Use of this class would prevent excessive bias towards specific $\theta \neq \theta_0$.

Theorem 5 shows that minimizing $P(H_0|x)$ over $g \in G_{US}$ is equivalent to minimizing over the more restrictive class

$$\mathcal{U}_S = \{\text{all symmetric uniform distributions}\}.$$

The point mass at θ_0 is included in $\mathcal{U}_S$ as a degenerate case. (Obviously, each element of G_{US} is a mixture of elements of $\mathcal{U}_S$. The proof of Theorem 5 is thus similar to that of Theorem 3, and will be omitted.)

Theorem 5.

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

so that $\underline{B}(x, G_{US}) = \underline{B}(x, \mathcal{U}_S)$ and $\underline{P}(H_0|x, G_{US}) = \underline{P}(H_0|x, \mathcal{U}_S)$.

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

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

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

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

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

Note that $f(\bar{x}|\theta_0) = \frac{\sqrt{n}}{\sigma} \varphi\left(\frac{\bar{x}-\theta_0}{\sigma/\sqrt{n}}\right) = \frac{\sqrt{n}}{\sigma} \varphi(t)$. Thus if $t > 1$ and K maximizes $m_K(\bar{x})$, we have

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

We summarize our results in Theorem 6.

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

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

and

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

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 $K_0 = t$):

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

Convergence is usually achieved after only 2 or 3 iterations. Also, 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 $\underline{P}(H_0|x,G_{US})$ for some specific important values of t, and $\pi_0 = \frac{1}{2}$.

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

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

Comparison of Table 6 with Table 5 shows that $\underline{P}(H_0|x,G_{US})$ is only moderately larger than $\underline{P}(H_0|x,G_S)$ for P-values of .10 or .05. However, the asymptotic behavior (as $t\to\infty$) of the two lower bounds is very different, as the following theorem shows.

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

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

Furthermore,

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

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 $P(H_0|x)$ 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 like a uniform density. What happens when $P(H_0|x)$ 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 has been investigated by Edwards, Lindman and Savage (1963), pp. 229-231.

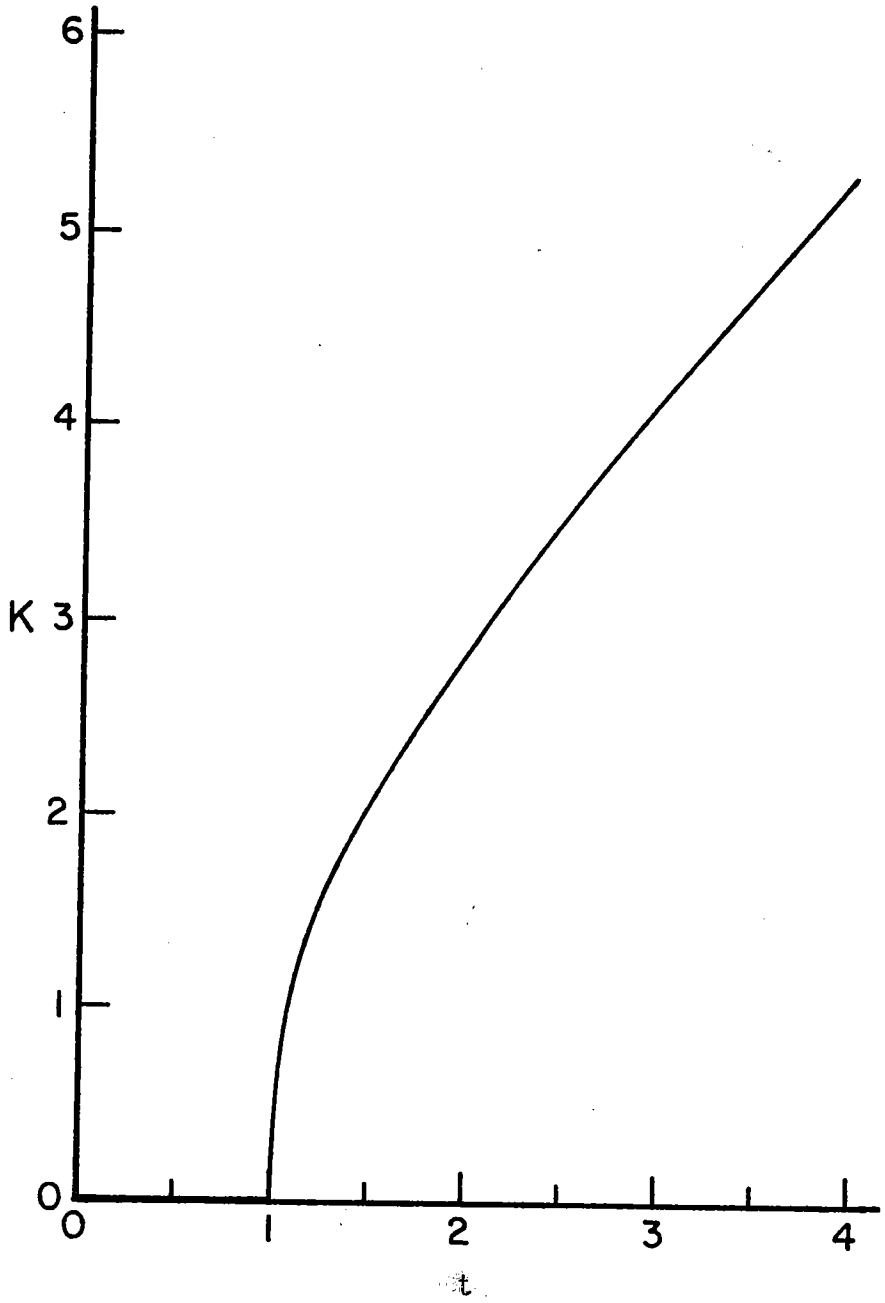

Figure 1. Minimizing Value of K When $G = G_{US}$.

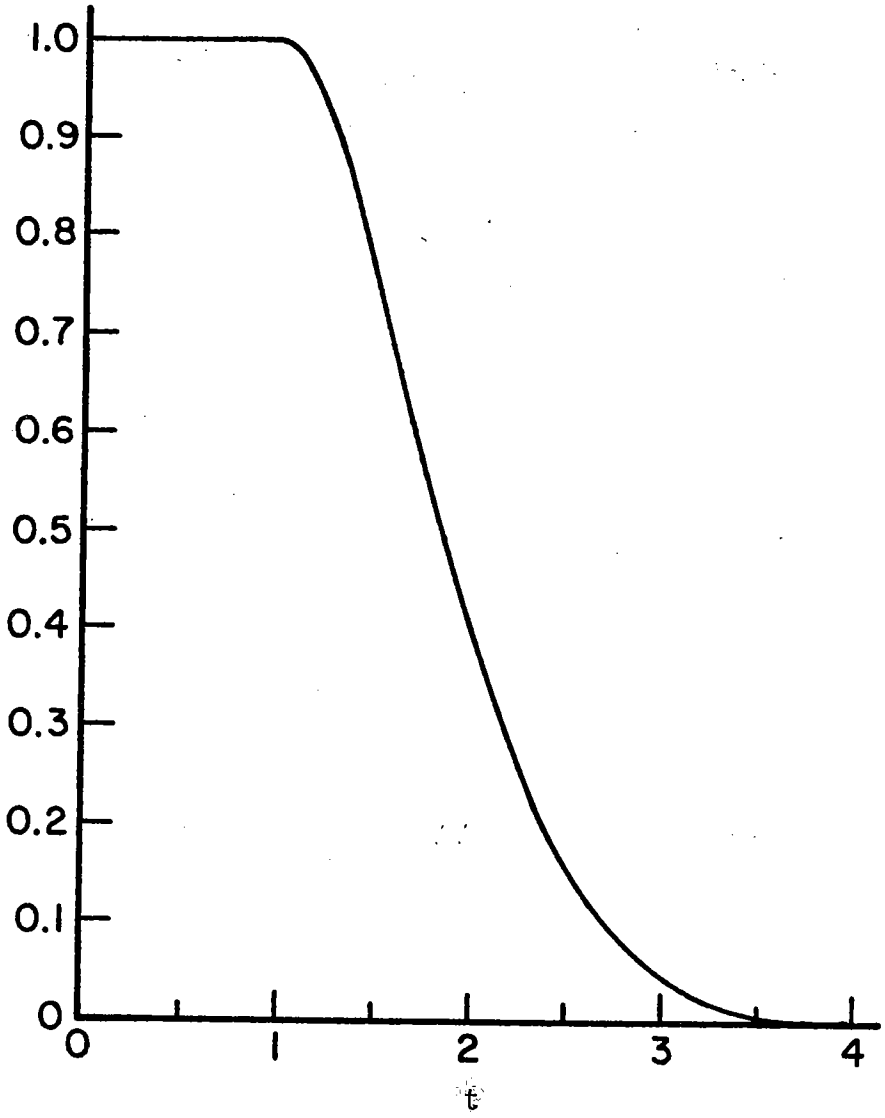

Figure 2. Values of $\underline{B}(x,G_{US})$ in the normal example.

Theorem 8. (Edwards, Lindman, and Savage (1963)). If $t \leq 1$ in Example 1, then $\underline{B}(x,G_{NOR}) = 1$ and $\underline{P}(H_0|x,G_{NOR}) = \pi_0$. If $t > 1$, then

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

and

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

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

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

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 considered, is given in Figure 3. Although the vertical differences are larger than the visual discrepencies, the closeness of the bounds for G_{US} and G_{NOR} is apparent.

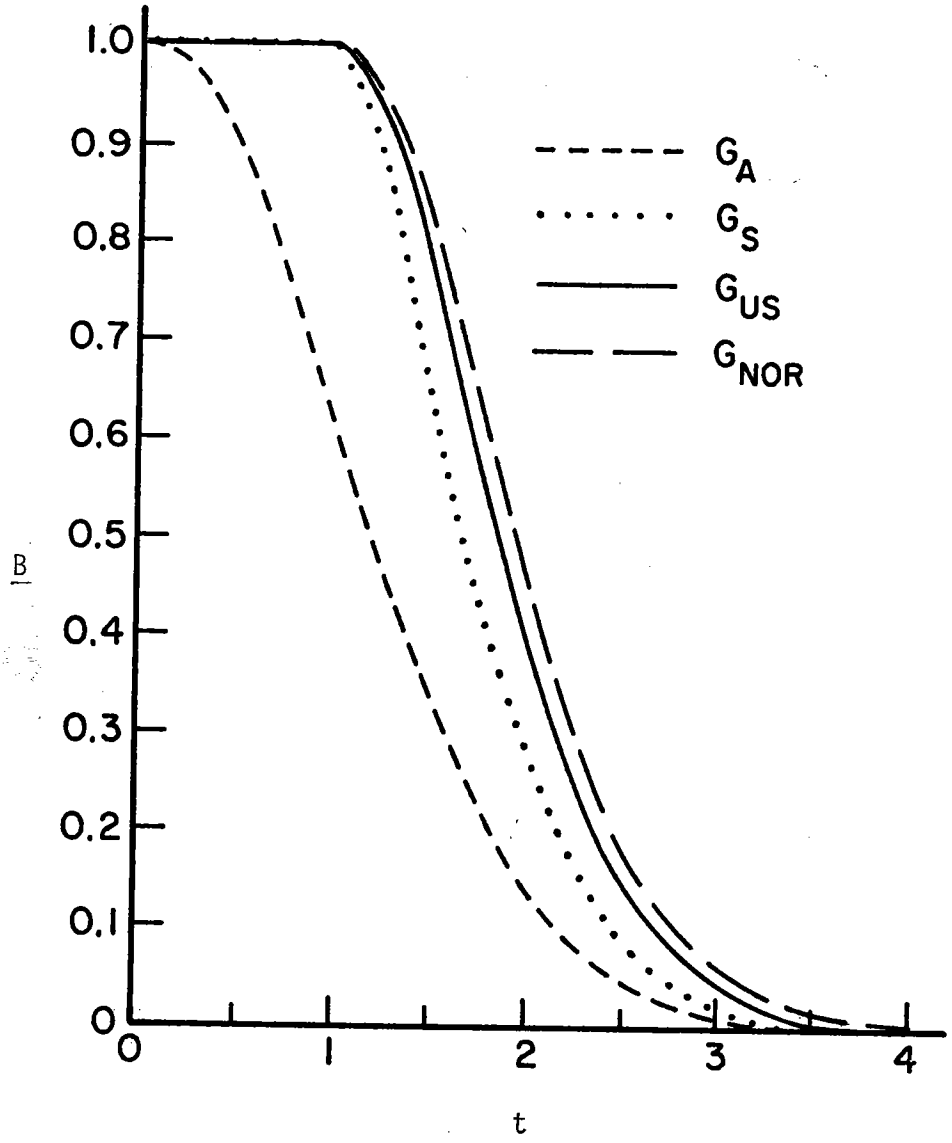

Figure 3. Values of $\underline{B}(x,G)$ in the normal example for different choices of G.

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 $P(H_0|A)$, 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

$$P(H_0|A) = \left[1 + \frac{(1-\pi_0)}{\pi_0} \cdot \frac{m_{g_1}(A)}{m_{g_0}(A)}\right]^{-1}, \tag{4.1}$$

where

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

One claim made in the introduction was that, if $\Theta_0 = (\theta_0-b, \theta_0+b)$ with b suitably small, then approximating 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 $f(x|\theta)$ is approximately constant on Θ_0 (so that $m_{g_0}(x) = \int_{\Theta_0} f(x|\theta)g_0(\theta)d\theta \cong f(x|\theta_0)$; 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 $P(H_0|A)$ by choosing sets G_0 and G_1 of g_0 and g_1, respectively, calculating

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

and defining

$$\underline{P}(H_0|A, G_0, G_1) = \left[1 + \frac{(1-\pi_0)}{\pi_0} \cdot \frac{1}{\underline{B}(A, G_0, G_1)}\right]^{-1}. \tag{4.4}$$

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 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 $\underline{P}(H_0|x, G_0, G_1)$ 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 (θ_0-b, θ_0+b) being in the

convex part of the tail of the likelihood function), while the maximization over G_1 is the same as before. □

Another type of testing situation which 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. Analogues 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 Subsection 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

$$\underline{P}(H_0|A) = \left[1 + \frac{1/2}{1/2} \cdot \frac{\sup_{g_1} m_{g_1}(A)}{P_{\theta_0}(A)} \right]^{-1},$$

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

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

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

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

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

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

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

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

Now, for <u>any</u> 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 $\underline{P}(H_0|A, G) = (1 + p^{-1})^{-1}$, 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

$$t = \begin{cases} 1 & \text{means only mild evidence against } H_0 \\ 2 & \text{means significant evidence against } H_0 \\ 3 & \text{means highly significant evidence against } H_0 \\ 4 & \text{means overwhelming evidence against } H_0, \end{cases}$$

should, at the very least, be replaced by the rule-of-thumb

$$t = \begin{cases} 1 & \text{means no evidence against } H_0 \\ 2 & \text{means only mild evidence against } H_0 \\ 3 & \text{means significant evidence against } H_0 \\ 4 & \text{means highly significant evidence against } H_0, \end{cases}$$

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

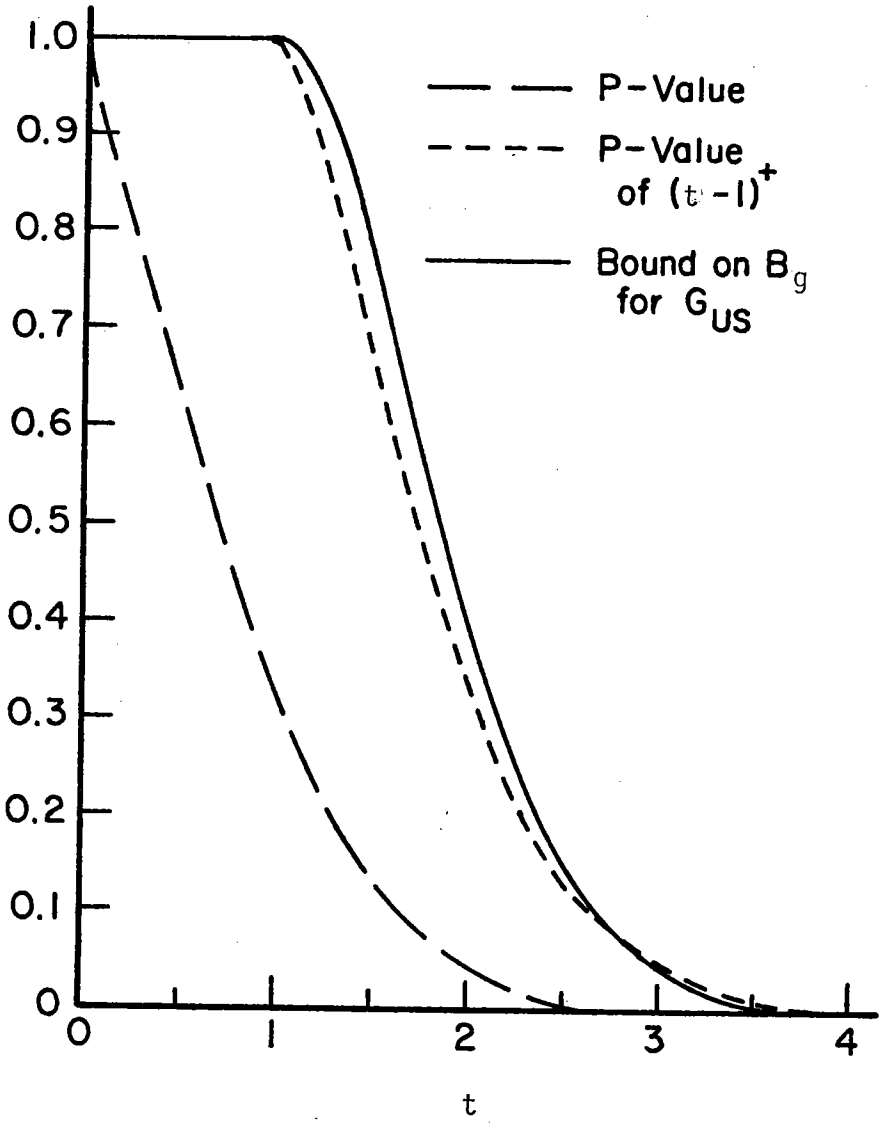

Figure 4. Comparison of $\underline{B}(x,G_{US})$ and P-values.

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 m.l.e. 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 to or larger than the one dimensional bounds.

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 $f(x|\theta)$ by $f_T(t|\theta)$, 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 (cf. Cressie and Read (1984)). Writing the noncentrality prarmeter 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 $P(H_0|x)$ 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 $P(H_0|x)$.

Comment 3. What should a statistician desiring to test a point null hypothesis do? While it seems clearly unacceptable to use a P-value of .05 as evidence to reject, the lower bounds on $P(H_0|x)$ 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 $P(H_0|x)$ is itself small). One possible solution is to seek upper bounds for $P(H_0|x)$, an approach taken with some success in Edwards, Lindman, and Savage (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, in order to reach sensible conclusions about testing a point null. Perhaps the most attractive possibility, following Dickey (1973), is to communicate $B_g(x)$ or $P(H_0|x)$ for a wide range of prior inputs, allowing the user to easily choose 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 $P(H_0|x)$ (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 $P(H_0|x)$ (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 which yields an interval of values for $P(H_0|x)$ as the prior ranges over all distributions "close" to an elicited prior.) General discussions of presentation of $P(H_0|x)$, as a function of subjective inputs, can be found in Dickey (1973), Berger (1985), and Berger and DasGupta (1985).

Comment 4. If one insisted on creating a "standardized" significance test for common use (as opposed to the flexible Bayesian reporting discussed above) it would seem that the tests proposed by Jeffreys (1961) are quite suitable. For small and moderate n in Table 1, $P(H_0|x)$ 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 $P(H_0|x)$ 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, $\underline{P}(H_0|x,\mathcal{G}_{US})$ 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 that it is desired to test H_0: $\theta = 0$ versus H_1: $\theta \neq 0$. It can then be shown that (for $\pi_0 = \frac{1}{2}$)

$$\lim_{|x|\to\infty} \frac{\underline{B}(x,\mathcal{G}_{US})}{\text{P-value}} = \lim_{|x|\to\infty} \frac{\underline{P}(H_0|x,\mathcal{G}_{US})}{\text{P-value}} = 1,$$

so that 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

$$\mathcal{G}_C = \{\text{all Cauchy distributions}\},$$

since one can prove that, for $|x| \geq 1$ and $\pi_0 = \frac{1}{2}$,

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

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

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

P-value (p)	$\lvert x\rvert$	$\underline{B}(x,G_{US})$	$\underline{P}(H_0\lvert x,G_{US})$	$\underline{B}(x,G_C)$	$\underline{P}(H_0\lvert x,G_C)$
.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

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

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

so that large x provides <u>no information</u> to a Bayesian. Thus, rather than this being a case where the P-value might have a reasonable evidentiary interpretation because it agrees with $\underline{P}(H_0|x,G_{US})$, this is a case where $\underline{P}(H_0|x,G_{US})$ 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 <u>any</u> n in this problem, the marginal likelihood is such that (5.1) holds. (Of course, the initial use of a noninformative prior for σ^2 is not immune to criticism.)

<u>Comment 6</u>. 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 <u>any</u> situation and <u>any</u> x)

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

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

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," To appear in the Annals of Statistics.

BERGER, J., and DAS GUPTA, A. (1985), "Bayesian Inference for a Normal Mean Using Robust Priors, Calculation of t Convolutions, and Scientific Communication," Technical Report 84-26, Department of Statistics, Purdue University, West Lafayette.

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.

BERKSON, J. (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 B, 46, 440-464.

DE GROOT, 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," Proceedings of the Conference on Foundational Questions in Statistical Inference, O. Barndorff-Nielsen et. al. (Eds.), University of Aarhus.

DIAMOND, G. A., and FORRESTER, J. S. (1983), "Clinical Trials and Statistical Verdicts: Probably 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 B, 35, 285-305.

_____(1974), "Bayesian Alternatives to the F-test and Least Squares Estimate in the Normal Linear Model," Studies in Bayesian Econometrics and Statistics, S. E. Fienberg and A. Zellner (Eds.), Amsterdam: North-Holland.

_____(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 Broadbalk Wheatfield Example," Bayesian Analysis in Econometrics and Statistics: Essays in Honor of Harold Jeffreys, A. Zellner (Ed.), Amsterdam: North-Holland.

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: 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, J. Kadane (Ed.), Amsterdam: North-Holland.)

FELLER, W. (1968), An Introduction to Probability Theory and Its Applications, Vol. 1, (3rd Edition), New York:Wiley.

GOOD, I. J. (1950), Probability and the Weighing of Evidence, London: Charles 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: M.I.T. Press.

_____(1967), "A Bayesian Significance Test for the Multinomial Distribution," Journal of the Royal Statistical Society 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'," 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 B, 16, 261-268.

JEFFREYS, H. (1957), Scientific Inference, Cambridge: Cambridge University Press.

_____(1961), Theory of Probability (3rd Edition), Oxford: Oxford University Press.

JEFFREYS, H. (1980), "Some General Points in Probability Theory," Bayesian Analysis in Econometrics and Statistics, A. Zellner (Ed.), Amsterdam: North-Holland, 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: 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," Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Berkeley: University of California Press.

_____(1965), Introduction to Probability and Statistics From a Bayesian Viewpoint (Parts 1 and 2), Cambridge: 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 B, 27, 169-203.

RAIFFA, H., and SCHLAIFER, R. (1961), Applied Statistical Decision Theory, Division of Research, Graduate School of Business Administration, Harvard University, Boston.

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

SMITH, C. A. B. (1965), "Personal Probability and Statistical Analysis," Journal of the Royal Statistical Society A, 128, 469-499.

SMITH, A. F. M. and SPIEGELHALTER, D. J. (1980), "Bayes Factors and Choice Criteria for Linear Models," Journal of the Royal Statistical Society B, 42, 213-220.

SOLO, V. (1984), "An Alternative to Significance Tests," Technical Report 84-14, Department of Statistics, Purdue University, West Lafayette.

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

_____(1984), "Posterior Odds Ratios for Regression Hypotheses: General Considerations and Some Specific Results," Basic Issues in Econometrics, Chicago: University of Chicago Press.

ZELLNER, A., and SIOW, A. (1980), "Posterior Odds Ratios for Selected Regression Hypotheses," Bayesian Statistics, J. M. Bernardo et. al. (Eds.), Valencia: University Press.

#84-27

Reprinted from: © 1987 American Statistical Association
Journal of the American Statistical Association
March 1987, Vol. 82, No. 397, Theory and Methods

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

JAMES O. BERGER and THOMAS SELLKE

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.

Journal of the American Statistical Association
March 1987, Vol. 82, No. 397, Theory and Methods

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

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

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 $A = \{y: T(y) \geq T(x)\}$ 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 $\mathfrak{N}(\theta_0, \sigma^2)$. Since a sufficient statistic for θ is $\bar{X} \sim \mathfrak{N}(\theta, \sigma^2/n)$, we have that $m_g(\bar{x})$ is an $\mathfrak{N}(\theta_0, \sigma^2(1 + n^{-1}))$ 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 = {all distributions}, G_S = {all distributions symmetric about θ_0}, G_{US} = {all unimodal distributions symmetric about θ_0}, G_{NOR} = {all $\mathfrak{N}(\theta_0, \tau^2)$ 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

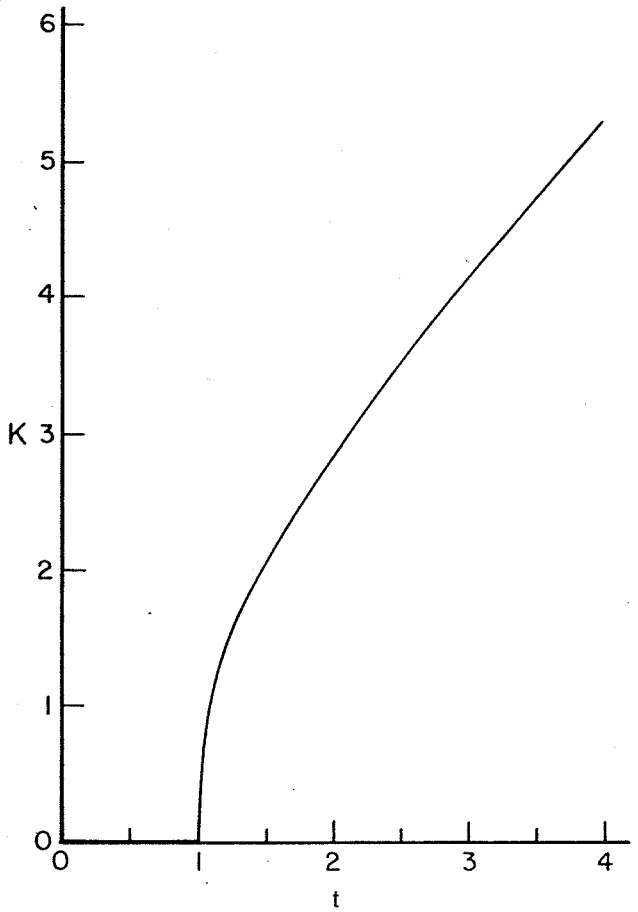

Figure 1. Minimizing Value of K When $G = G_{US}$.

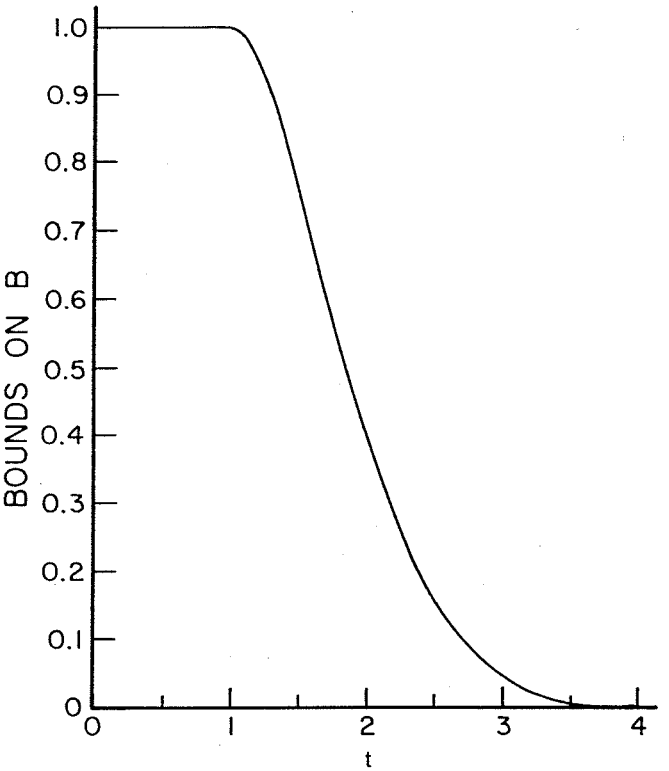

Figure 2. Values of $\underline{B}(x, G_{US})$ in the Normal Example.

$\underline{\Pr}(H_0 \mid x, G_{US})$ 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 $g \in \mathfrak{U}_S$. Although using $\mathfrak{U}_S$ is much more reasonable than using G_A, there is still some residual bias against H_0 involved in using $\mathfrak{U}_S$. 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).

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 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.

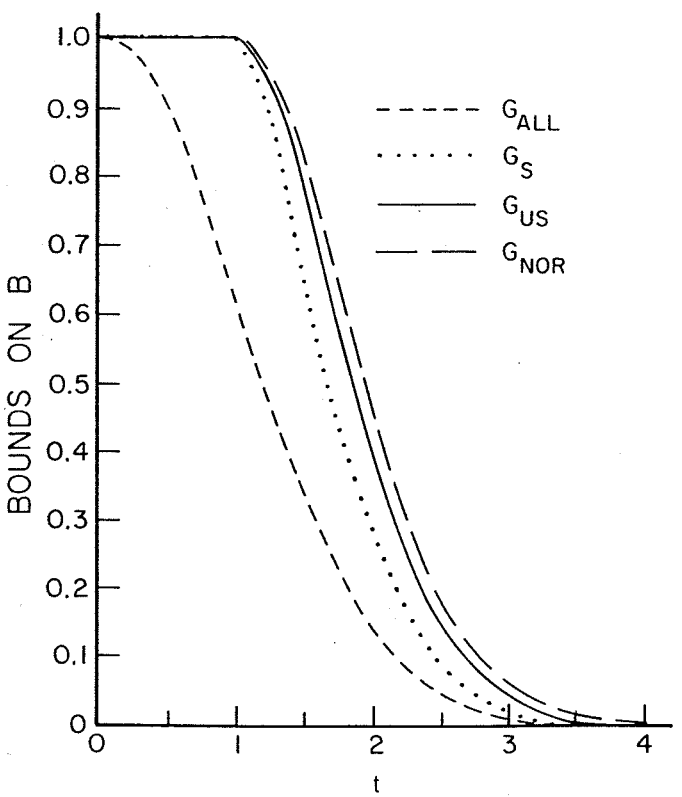

Figure 3. Values of $\underline{B}(x, G)$ in the Normal Example for Different Choices of G.

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 $G = G_A$ = {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

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

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.

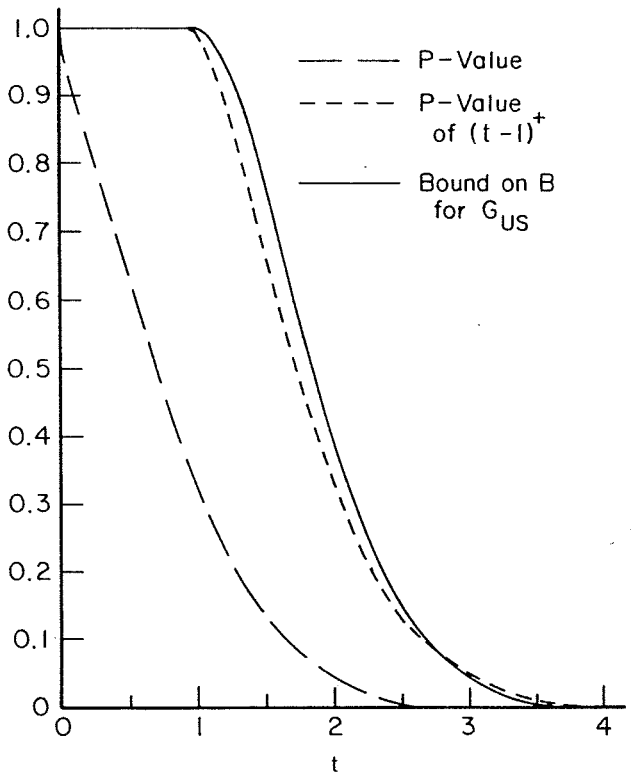

Figure 4. Comparison of $\underline{B}(x, G_{US})$ and P Values.

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 $\mathfrak{N}(\mu, \tau^2)$ 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 $\mathfrak{N}(\mu, \tau^2)$ 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.

Comment

JOHN W. PRATT*

1. BERGER AND SELLKE (AND EDWARDS, LINDMAN, AND SAVAGE)

When I was younger so much younger than today, I never needed anybody's help in any way, least of all the Beatles', and I usually found old fogeys' historical homilies distasteful. As my own fogeyhood impends, I find them just as distasteful, but more salutary. In this vein I must say that, despite the generous references in Berger and Sellke (B&S) and my previous looks at Edwards, Lindman, and Savage (1963) (EL&S), I realized only on recent rereading how much credit is due EL&S for formulating and resolving questions that illuminate the interpretation of P values in testing sharp null hypotheses (and much else). The extent and charm of their penetrating discussion and the progression ordering most of B&S's results are evident in this brief quotation from EL&S (p. 228) on testing the null hypothesis that a normal distribution with known variance has mean $\lambda = 0$.

> *Lower bounds on L.* An alternative when $u(\lambda \mid H_1)$ [the density on H_1] is not diffuse enough to justify stable estimation is to seek bounds on L [the likelihood ratio or Bayes factor in favor of H_0]. Imagine all the density under the alternative hypothesis concentrated at x, the place most favored by the data. The likelihood ratio is then
>
> $$L_{\min} = \frac{\phi(t)}{\phi(0)} = e^{-t^2/2}.$$
>
> This is of course the very smallest likelihood ratio that can be associated with t. Since the alternative hypothesis now has all its density on one side of the null hypothesis, it is perhaps appropriate to compare the outcome of this procedure with the outcome of a one-tailed rather than a two-tailed classical test. At the one-tailed classical .05, .01, and .001 points, $L_{\min}$ is .26, .066, and .0085, respectively. [This essentially covers Th. 1 and Tables 3 and 4 of B&S, in one-tailed form.] Even the utmost generosity to the alternative hypothesis cannot make the evidence in favor of it as strong as classical significance levels might suggest. Incidentally, the situation is little different for a two-tailed classical test and a prior distribution for the alternative hypothesis concentrated symmetrically at a pair of points straddling the null value [see B&S, Th. 3 and Tables 2 and 5]. If the prior distribution under the alternative hypothesis is required to be not only symmetric around the null value but also unimodal, which seems very safe for many problems, then the results [B&S, Ths. 5 and 6 and Table 6] are too similar to those obtained later for the smallest possible likelihood ratio obtainable with a symmetrical normal prior density to merit separate presentation here.

After giving results for normal priors (B&S, Th. 8 and Table 7), EL&S "conclude that a t of 2 or 3 may not be evidence against the null hypothesis at all, and seldom if ever justifies much new confidence in the alternative hypothesis" (p. 231) (see B&S, Comment 1).

It is not that B&S claim or sneak off with credit due others. Few are more aboveboard, and I have admired other writing by Berger, in particular his books, for both substance and referencing. But credit slides all too easily onto later authors even when they have no need or desire to steal it. EL&S is still must reading. Do not assume that later publications supersede or subsume it or let its introductory posture or exotic auspices deter you. It is reprinted in at least two books. Only 1% of it is quoted above. The other 99%, though not all so condensed, is also highly rewarding. Some of its subheadings on testing (the topic of half of it) are *Bernoullian example, Upper bounds on L, Haunts of χ^2 and F, Multidimensional normal measurements and a null hypothesis,* and *Some morals about testing sharp null hypotheses.*

B&S's spiraling exposition is helpful the first time around, but afterward I felt a need for more winding up than the graphs of Bayes factors in their Figure 3, even after the trivial but revealing addition of a graph of the comparable frequentist factor $p/(1 - p)$. In the top part of Table 1 here, I have collected and juxtaposed probabilities from B&S's tables (but not the Bayes factors or ratios to pt or pt^2), following A. S. C. Ehrenberg's precepts as best I could. The remaining three lines give $\Pr(H_0 \mid t)$ for a normal prior with variance equal to the sampling variance of the mean (B&S, Table 1 with $n = 1$), and for tight and diffuse priors, which may be viewed as extreme normals (with $n = 0$ and ∞, respectively). Thus the first column shows that the minimum posterior probability for a P value $p = .10$ is .205 when all priors are allowed and increases to .340, .390, and .412 as symmetry, unimodality, and normality restrictions are added. The excess over p and increase with more restrictions on the prior are proportionately even greater at smaller P values. Normality adds little to symmetry, as EL&S observed.

Not to leave well enough alone, I included a "large" t column with B&S's asymptotic formulas and two they happen to omit [where $2.07 = (\pi e/2)^{1/2}$ and $1.77 = \pi^{1/2}$]. They show that the first three are lower bounds for $t > 0$, $t > 2.28$, and $t > 0$, respectively (Theorems 2, 4, 7). The range where the fourth is a lower bound is $t > 2.72$ by my sketchy calculations. (For a normal prior with arbitrary n, the asymptotic formula is $\Pr(H_0 \mid t) = [(n + 1)\pi/2]^{1/2} e^{t^2/2(n+1)} tp$. The range of t where this is a lower bound depends on n. It cannot be a lower bound for all n and t, since it is not a lower bound for $t < 2.72$ in the EL&S worst case $n + 1 = t^2$.)

All the normal results hold for all sample sizes and all prior and sampling variances if n is defined as the ratio of the prior variance to the sampling variance of the mean rather than as the sample size. What I see as "troubling" about the scaling here (see B&S, p. 112) is only the importance of the height of the prior density under H_1 (near $\bar{X}$, say). Such trouble is inevitable in testing sharp null hypotheses, not a deficiency of the prior family. Since n is unrestricted, there is no troubling link between σ and

* John W. Pratt is Professor, Graduate School of Business Administration, Harvard University, Boston, MA 02163. The author is very grateful to Persi Diaconis and Arthur Schleifer, Jr. for helpful comments and to the Associates of the Harvard Business School for research support.

Journal of the American Statistical Association
March 1987, Vol. 82, No. 397, Theory and Methods

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

t	1.645	1.960	2.576	3.291	Large $\frac{2}{t}\phi(t)$	B&S Tables	B&S Theorems
P Value (p)	*.10*	*.05*	*.01*	*.001*			
Priors allowed							
All	.205	.128	.035	.0044	$1.25tp$	3, 4	1, 2
All symmetric	.340	.227	.068	.0088	$2.51tp$	2, 5	3, 4
Symmetric unimodal	.390	.290	.109	.018	t^2p	6	5–7
Symmetric normal	.412	.321	.133	.025	$2.07t^2p$	7	8
Normal var σ^2/n	.42	.35	.21	.086	$1.77e^{t^2/4}tp$	1	
Tight at θ_0	.5	.5	.5	.5			
Diffuse	1	1	1	1			

the prior variance of θ as there is for "conjugate" priors when σ is unknown.

The notion of choosing one or more classical or other insufficient statistics and basing a Bayesian analysis or comparison on them rather than on the whole data set (see B&S, Comment 2) is supported and explored at some length in Pratt (1965, sec. 2).

2. CASELLA AND BERGER (AND PRATT)

In certain one-sided cases, Casella and Berger (C&B, but a different Berger) show that the infimum of $\Pr(H_0 \mid x)$, the posterior probability of H_0, is as small as the P value, p, or smaller. Now a point that permeates EL&S is that, if small, a lower bound is almost useless since it doesn't say you will be anywhere near it. (Hence they seek upper bounds too.) In fact, however, not only is inf $\Pr(H_0 \mid x) \leqq$ or $= p$ but, more to the point, $\Pr(H_0 \mid x)$ itself is close to p in most ordinary one-sided testing problems if n is not small and the prior on θ is not jagged. This is obvious in particular for normal models and hence for procedures concordant with asymptotic likelihood theory. It is also obvious for flat priors in C&B's situation, that of a single observation (or test statistic) x with density known except for location. What C&B add is essentially that, in this situation, $\Pr(H_0 \mid x) < p$ is impossible if the prior is unimodal and the density symmetric with monotone likelihood ratios, but possible in many other cases. Their situation is unfortunately very special. Test statistics, even t and rank statistics, rarely have densities known except for location. Furthermore, for $n > 1$, a regular location family admits a single sufficient statistic only if it is normal with known variance (Kagan, Linnik, and Rao 1973), and otherwise attending to information besides the test statistic can either raise or lower $\Pr(H_0 \mid x)$. So where C&B take us is unclear but not far.

Having done the decent thing and quoted someone else, I will now do the fun thing and quote myself. In Pratt (1965, secs. 7 and 8) I did not merely "state that in the one-sided testing problem the p value can be approximately equal to the posterior probability of H_0" (C&B, p. 106). I *emphasized* the much more important point that it usually *will* be (without claiming novelty even then). I argued both via confidence limits as approximate posterior fractiles and, in location problems, via diffuse priors and independence of θ and $T - \theta$. Among my arguments for confidence limits as approximate posterior fractiles were one's natural reluctance to use them when they are not and asymptotic likelihood theory. I also mentioned Good's elegant argument (1950, 1958). If one-sided reconcilability is as little recognized as C&B suggest, at least I for one tried (both in 1965 and later). But the two-sided discrepancy may get more ink mainly because it is more subtle, surprising, and significant.

As to two-tailed P values, I would have been even more gloomy about the one-dimensional case if I had registered EL&S properly, but what I said in part, partly paraphrased, was "The only widely valid relation between a two-tailed P-value and a posterior probability of natural interest seems to be" that $\frac{1}{2}p$ sometimes has the foregoing one-sided interpretation. Although $1 - p$ "is often approximately the posterior probability that" $0 \leqq \theta \leqq 2\hat{\theta}$, this interval is not of natural interest. Its multidimensional counterpart is "even less so," and indeed depends on irrelevant particulars of the design and test statistic.

> In short, when the null hypothesis $\theta \leqq 0$ is tested against the alternative $\theta > 0$, where θ is one-dimensional and $\theta < 0$ is possible, the P-value is usually approximately the posterior probability that $\theta \leqq 0$. Most other situations where the P-value has a helpful interpretation can be recast in this form. Of course, $\theta \leqq 0$ can be replaced by $\theta \leqq \theta_0$ or $\theta \geqq \theta_0$. And while it is convenient to use P-values in the discussion, those who are interested only in whether or not the results are significant at some preselected level will find similar remarks apply. All the statements about the relation of P-values to posterior probabilities, or lack of it, can be seen easily to hold for a univariate or multivariate normal distribution with known variance or variance matrix. (Pratt 1965, p. 184)

Two technical points. C&B's Lemma 3.1 is an immediate consequence of the fact (subsumed in their proof) that the posterior obtained from a mixture of priors is a mixture of the posteriors obtained from each. The point is more familiar when mixing different models also: the posterior weights are the posterior probabilities of the components, which are of course proportional to their prior probabilities times their predictive densities. B&S (see Th. 3 and its proof) work directly with the Bayes factor and the predictive density, which is equivalent and simpler for the purpose.

C&B's Theorem 3.1 states less than they prove. As it is stated, all but the first sentence of the proof could be replaced by the observation that the inequality follows from Theorem 3.2 (whose proof is independent of Th. 3.1), or directly and easily by considering the uniform prior on $(-k, k)$ as $k \to \infty$ [Eq. (3.5) and the limit calculation at the end of the proof of Th. 3.2].

3. WHAT ABOUT THE PRIORS?

Are the minimizing priors "palatable"? If not, what then? The one-point prior most favorable to H_1 is clearly an exaggeration, but more palatable for one-sided than two-sided alternatives, as EL&S noted. The symmetric two-point prior is still worse for one-sided but somewhat better for two-sided alternatives. EL&S chose accordingly; their remark that one-point priors for one-sided alternatives are "little different" is borne out by halving the P values in B&S's Table 4 and comparing the result with Table 5 (or 2), most easily via the last column unless $p = .05$. All of the minimizing priors depend on the data, an unpalatable feature to most who care at all, and real opinions in one-sided problems would rarely be symmetric or improper. So real prior opinions will often be far from the minimizing opinions, which suggests that real posterior opinions may greatly exceed the lower bounds. This strengthens B&S's main point [because restricting the prior further can only increase the amount by which $\Pr(H_0 \mid x)$ exceeds p in the two-sided case], but points up the weakness of C&B's results in the one-sided case (where matters were already left indeterminate by their argument).

Unfortunately, to discredit a seriously entertained point null hypothesis, one needs something like a lower bound on the prior density in the region of maximum likelihood under the alternative. This appears directly in EL&S but only indirectly in B&S (Comment 3). To my mind it justifies EL&S in being even more cautious in their conclusion (quoted previously) than B&S in Comment 1. Any dimension-reducing hypothesis poses a similar troubling problem. Making such hypotheses approximate makes them more realistic but harder yet to analyze.

4. WHAT'S IT ALL ABOUT?

The broad question under discussion is an important one: what do frequentist inference procedures really accomplish, and what can statisticians of all stripes learn about them by viewing them through Bayesian glasses? The articles here give precise answers to well but narrowly posed subquestions about P values. If you are a Defender of Virtuous Testing or simply a Practical Person, you may feel that the subquestions do not represent the real issues well. But whatever your attitudes or Attitudes, the B&S–EL&S results can hardly comfort you, and I think should disturb you. And even if you can blink them completely—even if you are prepared to disavow any remotely posterior interpretation of P values or visibility through Bayesian glasses—you are not out of the woods. A vast literature discourses on all kinds of problems with hypothesis testing and P values for all kinds of purposes from all kinds of viewpoints: frequentist, Bayesian, logical, practical; for description, inference, decisions, conclusions; preliminary, simultaneous, final; choice of model, estimator, further sampling; and so on. It would be impolite to cite my several nibbles at the subject and invidious to select others, so I will trust the other discussants to suggest its scope. Domains where tests are acceptable may exist, but rejecting Bayesian arguments will not establish or enlarge them.

In summary, I see little major news here beyond what was known by 1963 (EL&S) or obvious by 1965 (Pratt). But every generation must rediscover old truths, and reviving, polishing, and amplifying them and even charting their backwaters are useful. If these articles help the world hear their messages, which I certainly agree with, well and good. If the world is ready for less stylized and precise but all the more disturbing messages about testing, better yet. Regardless, fogeyhood is fun!

ADDITIONAL REFERENCE

Kagan, A. M., Linnik, Yu. V., and Rao, C. R. (1973), *Characterization Problems in Mathematical Statistics* (translated from the Russian by B. Ramachandran), New York: John Wiley.

Comment

I. J. GOOD*

I was interested in both of these articles (which I shall call B&S and C&B) because Bayesian aspects of P values have fascinated me for more than 40 years. The topic will be taken more seriously now that it has hit JASA with two long articles, plus discussion, and the occasion will be all the easier to remember because two Bergers are involved. One result, I hope, will be that the conventional P value of approximately .05, when testing a simple statistical hypothesis H_0, will be correctly interpreted: *not as a good reason for rejecting H_0 but as a reason for obtaining more evidence provided that the original experiment was worth doing in the first place.*

In my opinion P values and Bayes factors are both here to stay, so the relationships between them need to be taken seriously. These relationships form a large part of the main

* I. J. Good is University Distinguished Professor, Department of Statistics, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061. This work was supported in part by National Institutes of Health Grant GM18770.

problem of pure rationality, namely to what extent Bayesian and non-Bayesian methods can be synthesized. (The main problem of *applied* rationality is how to preserve the human species.) My view is that the methods can be synthesized, because, contrary to the opinion of some radical Bayesians, I believe that P values are not entirely without merit. The articles by B&S and C&B contribute to this synthesis, although the *title* of B&S might suggest otherwise.

The relationships between P values and Bayes factors depend on the specific problem, on the background information (some of which is usually vague), on the sample size, on the model assumed, Bayesian or otherwise, and on the questions being asked. B&S and C&B consider distinct questions and, therefore, arrive at distinct solutions. Their problems can be described as significance testing and discrimination, respectively. I think that the article by C&B would have been improved if it had been slightly more friendly to B&S. Television commercials compare burgers, but they do not knock the simple statistical hypothesis. Both articles make useful contributions by careful considerations of inequalities satisfied by Bayes factors. My comments will be partly historical.

Sometimes it is adequate, as in B&S, to define a null hypothesis as $\theta = 0$ or as $|\theta| < \delta$, where δ is small [compare, e.g., Good 1950, p. 91]; sometimes (and this can be regarded as a generalization of the first case) the null hypothesis asserts that $\theta \leqslant 0$ with one or more priors conditional on this inequality; sometimes the initial or prior probability $\Pr(H_0)$ is (approximately) equal to $\frac{1}{2}$ as is usually assumed in both of the articles under discussion and by Jeffreys (1939); sometimes $\Pr(H_0)$ is far from $\frac{1}{2}$ (and of course the posterior probability of H_0 can, therefore, be arbitrarily smaller than a P value); sometimes we prefer to leave the estimation of $\Pr(H_0)$ to posterity and, therefore, try to summarize the evidence from the experimental outcome alone by a P value or by a Bayes factor (or by its logarithm the weight of evidence), both of which have the merit of not depending on $\Pr(H_0)$; and sometimes the priors conditioned on H_0 and on its negation H_1 are reasonably taken as "mirror reflections" in the origin, as is largely assumed by C&B. When testing a treatment that a scientist had previously claimed to be better than a standard one, we are apt to choose H_0 as $\theta = 0$ and H_1 as $\theta > 0$. This model shows more respect to the scientist than if we defined H_0 as $\theta \leqslant 0$ or H_1 as $\theta \neq 0$. Whether he deserves that much respect will again depend on circumstances.

Although the two articles deal with distinct problems, it is possible to produce models that include both problems and intermediate ones. I have worked out one such concrete example that more or less does this and that I shall describe briefly. For more details see Good (in press a). It is a special case of C&B (4.1), but I believe that it is general enough for most purposes.

Let X denote the mean of n random variables, iid, and each $N(\theta, \sigma^2)$, where σ^2 is known or well estimated from the sample. Our aim is to discriminate between $H_0 : \theta \leqslant 0$ and $H_1 : \theta > 0$.

Assume that the prior density of θ given H_i ($i = 0$ or 1) is the folded normal density

$$[(2/\pi)^{1/2}/\tau_i] \exp[-\theta^2/2\tau_i^2)], \tag{1}$$

where $\theta < 0$ if $i = 0$, and $\theta > 0$ if $i = 1$, but with τ_i having the log-Cauchy hyperprior density

$$\psi_i = \frac{\lambda_i}{\pi\tau_i\{\lambda_i^2 + [\log(\tau_i/a_i)]^2\}}. \tag{2}$$

This hyperprior provides a convenient way to give propriety to the familiar improper prior of Jeffreys and Haldane proportional to $1/\tau_i$. The upper and lower quartiles of (2) are $a_i e^{\lambda_i}$ and $a_i e^{-\lambda_i}$, so we can give τ_i a determinate value a_i by letting $\lambda_i \to 0$. In addition, we can determine a_i and λ_i by judging the quartiles.

For this two-level hierarchical Bayesian model we find, after a page of elementary calculus, that the Bayes factor *against H_0 provided by the observation x,* which by *definition* is $O[H_1 \mid (X = x)\ \&\ G]/O(H_1 \mid G)$, is equal to

$$B(H_1 : X = x \mid G) = \Psi_1/\Psi_0, \tag{3}$$

where O denotes odds (also sometimes called an odds ratio), G denotes what was given before X was observed, the colon is read "provided by the information that," the vertical stroke denotes "given" as usual, and

$$\Psi_i = \Psi_i(x, \sigma_n, a_i, \lambda_i) = \int_0^\infty (\sigma_n^2 + \tau^2)^{-1/2} \times \exp\left[\frac{-x^2/2}{\sigma_n^2 + \tau^2}\right]\phi\left[\frac{\varepsilon_i x\tau/\sigma_n}{(\sigma_n^2 + \tau^2)^{1/2}}\right]\psi_i(\tau; a_i, \lambda_i)\, d\tau, \tag{4}$$

where $\varepsilon_0 = 1$, $\varepsilon_1 = -1$, $\sigma_n^2 = \sigma^2/n$ is the variance of X, and ϕ is the error function

$$\phi(y) = (2\pi)^{-1/2} \int_y^\infty e^{-u^2/2}\, du. \tag{5}$$

The integrand in (4) is smooth and not difficult to calculate, so the Bayes factor can be presented as a program with six input parameters, x, σ_n, a_0, a_1, λ_0, and λ_1, and the user can try several priors.

The result contains several interesting special cases, including some results given by B&S and C&B, except that the Bayes factor of B&S will be one half of mine in the appropriate special case. (See my miscellaneous comment 2 below.)

For example, if we take $\lambda_0 = \lambda_1 = 0$, $a_0 = a_1 = \tau$, τ/σ_n large, and $\Pr(H_0) = \frac{1}{2}$, and let H_2 denote the hypothesis that $\theta = 0$, then

$$\Pr(H_0 \mid X = x) \approx \phi(x/\sigma_n) = P,$$

the single-tailed P value corresponding to the "null hypothesis" H_2. Note that H_2 is *not* H_0. We may also describe P as the *maximum* P value over all simple statistical hypotheses of the form $\theta = \theta_0$, where $\theta_0 \leqslant 0$ as in C&B. Because H_2 is not H_0 this case provides only a *partial* reconciliation of Bayesian and Fisherian methods, especially as it is only one of many possible cases, and for this reason I think that C&B have exaggerated. The result certainly does not, and C&B do not claim that it does,

justify the extraordinarily common error, mentioned in both articles, perpetrated by several reputable scientists ("nonspecialists," to quote B&S), of interpreting a P value as $\Pr(H_0 \mid X = x)$ even when H_0 is a point hypothesis. When I mentioned the prevalence of this error to Jim Dickey he pointed out that even Neyman had perpetrated it! [See Good (1984a).] (Most of my citations from now on will be to papers of which I have read every word.)

When $a_0 = 0$, $\lambda_1 = 0$ (so $\tau_1 = \mu_1$), a_1/σ_n is "large," and $x > 2\sigma_n$, we have the situation of B&S (Th. 2, apart from a factor of 2), and the Bayes factor against H_0 is approximately

$$B \approx 2n^{-1/2}(\sigma/\tau_1)e^{s^2/2}, \qquad (s = x/\sigma_n, \text{ the "sigmage"}) \quad (6)$$

$$= \frac{\sigma}{\tau_1 P}\left(\frac{2}{\pi n}\right)^{1/2}\left[\frac{1}{s+}\,\frac{1}{s+}\,\frac{2}{s+}\,\frac{3}{s+}\cdots\right] \quad (7)$$

by Laplace's continued fraction. [Compare Good (1967, p. 410).] Since s is a function of P, it follows that, *for a given value of P,* the Bayes factor against H_0 is proportional to $n^{-1/2}$, and this is usually true when H_0 is a simple statistical hypothesis. This may be called the root n effect and was perhaps first noticed by Jeffreys (1939, pp. 194 and 361–364). For some history of this and allied topics, see Good (1982a).

As a special case of (7) one could append a further column to Table 4 of B&S, giving the values of $O(x)/t$ [or B/t if it is not assumed that $\Pr(H) = \frac{1}{2}$)]. These values would be 1.414, 1.421, 1.391, and 1.350. They are nearly constant because the continued fraction is approximated by $1/s$. This observation is a slight modification of Theorem 2 in B&S.

The root n effect is closely related to the familiar "paradox," mentioned by C&B, that a tail-area pundit can cheat by optional stopping. This possibility is also implicit in Good (1950, p. 96) and was made crystal-clear by reference to the law of the iterated logarithm in Good (1955/1956, p. 13). This form of optional stopping is known as "sampling to a foregone conclusion." To prevent this form of cheating, and to justify to some extent the use of P values as measures of evidence, I proposed "standardizing" a tail-area probability P to sample size 100, by replacing P by $\min(\frac{1}{2}, n^{1/2}P/10)$ (Good 1982b). This proposal is an example of a Bayes/non-Bayes (or Bayes–Fisher) compromise, or "synthesis" as it was called by Good (1957, p. 862) and in lectures at Princeton University in 1955. An example for a multinomial problem was previously given by Good (1950, pp. 95–96). For other examples of the Bayes/non-Bayes synthesis see, for example, Good (in press b).

In most situations that I have seen, where one tests a point null hypothesis, the sample size n lies between 20 and 500, so if we think in terms of $n = 100$, the square root effect will not mislead us by more than a factor of $\sqrt{5}$ in either direction. This explains why I have found that a Bayes factor B' against a point null hypothesis on a given occasion is roughly inversely proportional to P. This leads to the useful harmonic-mean rule of thumb for combining "tests in parallel," that is, tests on the same data (Good 1958, 1984b). This rule of thumb is not precise, but it is much better than the dishonest precise procedure of selecting the test that best supports what you want to believe!

Miscellaneous comments.

1. B&S rightly emphasize the distinction between knowing that $P = P_0$ (or only just less) and knowing only that $P \leqslant P_0$. The latter statement is of course "unfair" to the null hypothesis when P is close to P_0 (Good 1950, p. 94). If a scientist reports only that $P < .05$ we are sometimes left wondering whether $P \approx .049$, in which case the scientist may have been *deliberately* misleading. Such a scientist might have been brought up not to tell fibs, without being told that a flam is usually worse than a fib. Or perhaps he was just brainwashed by an "official" Neyman–Pearson philosophy in an elementary textbook written with the help of a pair of scissors and a pot of glue and more dogmatic than either Neyman or Egon Pearson were. If Neyman had been dogmatic he would not have made the "nonspecialist's error," or error of the third kind, mentioned previously.

2. In the past, and frequently in conversation, I have used a rough rule that a P value of .05 is worth a Bayes factor of only about 4 when testing a simple statistical hypothesis (e.g., Good 1950, p. 94; 1983, p. 51). B&S get about half this value because they use a prior symmetric about $\theta = 0$ given H_1, whereas my rule is intended more for the case in which H_1 asserts that $\theta > 0$.

3. The topic of max factors, mentioned by B&S, without the cosmetic name, was also discussed in Good (1950, p. 91) as applied to multinomials, which of course includes binomials, and where the maximum weight of evidence (log-factor) is related to the chi-squared test. In the binomial case, the approximation given for the maximum weight of evidence (in "natural bans") again H_0 naturally agrees with the result $\frac{1}{2}t^2$ cited in Example 1 of B&S. Although in multivariate problems the max factor is much too large, the relationship to χ^2 shows the relevance to an aspect of the philosophy of the Bayes/non-Bayes or Bayes–Fisher synthesis, namely that even a poor Bayesian model can lead to a sensible non-Bayesian criterion (a point that I have made on several other occasions).

Sometimes a multivariate test can be reduced to a univariate one. B&S mention an example, and another example is that of a max factor that is useful because the maximization is over a *single* hyperparameter as in the mixed Dirichlet hierarchical Bayes approach to multinomials and contingency tables (e.g., Good 1976, p. 1170; Good and Crook 1974, p. 714).

4. In their concluding comments B&S state that when considering a simple statistical hypothesis H_0, by and large 2σ is weak evidence against H_0, 3σ is "significant," and so on. These conclusions agree roughly with Good (1957, p. 863), where I judged that the Bayes factor in favor of H_0 usually lies within a factor of 3 of $10P$. (This can break down if $P < 1/10{,}000$ and for very large sample sizes.)

5. The references in B&S cover much of the literature, and this will presumably be more true when the comments

are included. To aid in making the bibliography more complete I exercise the rights of a senior citizen and list 28 additional relevant publications of which I have read every word (10 of them are in the conscientious reference list of B&S): (a) items C73, C140, C144, C199, C200, C201, C209, C213, C214, and C217 in *Journal of Statistical Computation and Simulation* (1984); (b) Items 13 (pp. 91–96), 82, 127 (pp. 127–128), 174, 398 (p. 35), 416, 547, 603B (p. 61), 862, 1234 (pp. 140–143), 1278 (regarding Bernardo), 1320–C73, 1396 (pp. 342–343), 1444, and 1475–C144 in the bibliography (pp. 251–266) in Good (1983); (c) Good (1955/1956, p. 13; 1981; 1983, indexes; 1986; in press a,b). To these may be added the thesis of my student Rogers (1974) and a further reference relevant to C&B, Thatcher (1964).

ADDITIONAL REFERENCES

Good, I. J. (1955/1956), Discussion of "Chance and Control: Some Implications of Randomization," by G. S. Brown, in *Information Theory, Third London Symposium 1955*, London: Butterworth's, pp. 13–14.

——— (1957), "Saddle-Point Methods for the Multinomial Distribution," *Annals of Mathematical Statistics*, 28, 861–881.

——— (1976), "On the Application of Symmetric Dirichlet Distributions and Their Mixtures to Contingency Tables," *The Annals of Statistics*, 4, 1159–1189.

——— (1981), Discussion of "Posterior Odds Ratio for Selected Regression Hypotheses," by A. Zellner and A. Siow, *Trabajos de Estadistica y de Investigacion Operativa*, 32, No. 3, 149–150.

——— (1982a), Comment on "Lindley's Paradox," by Glenn Shafer, *Journal of the American Statistical Association*, 77, 342–344.

——— (1982b), "Standardized Tail-Area Probabilities" (C140), *Journal of Statistical Computation and Simulation*, 16, 65–66.

——— (1984a), "An Error by Neyman Noticed by Dickey" (C209), in "Comments, Conjectures, and Conclusions," *Journal of Statistical Computation and Simulation*, 20, 159–160.

——— (1984b), "A Sharpening of the Harmonic-Mean Rule of Thumb for Combining Tests 'in Parallel' " (C213), *Journal of Statistical Computation and Simulation*, 20, 173–176.

——— (in press a), "A Flexible Bayesian Model for Comparing Two Treatments," C272, *Journal of Statistical Computation and Simulation*, 26.

——— (in press b), "Scientific Method and Statistics," in *Encyclopedia of Statistical Science* (Vol. 8), eds. S. Kotz and N. L. Johnson, New York: John Wiley.

Good, I. J., and Crook, J. F. (1974), "The Bayes/Non-Bayes Compromise and the Multinomial Distribution," *Journal of the American Statistical Association*, 69, 711–720.

Jeffreys, H. (1939), *Theory of Probability* (1st ed.), Oxford, U.K.: Clarendon Press.

Rogers, J. M. (1974), "Some Examples of Compromises Between Bayesian and Non-Bayesian Statistical Methods," unpublished doctoral thesis, Virginia Polytechnic Institute and State University, Dept. of Statistics.

Thatcher, A. R. (1964), "Relationships Between Bayesian and Confidence Limits for Predictions" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 26, 176–192.

Comment

DAVID V. HINKLEY*

The authors have added an impressive array of technical results to the main body of work on this subject by Jeffreys, Lindley, and others. The sense of surprise in the first article suggests that statistical education is not as eclectic as one might wish. In my brief comments I should like to mention some of the general issues that should be considered in any broad discussion of significance tests.

First, the interpretation of P value as an error rate is unambiguously objective and does not in any way reflect the prior credibility of the null hypothesis. Rules of thumb aimed at calibrating P values to make them work like posterior probabilities cannot reflect the broad range of practical possibilities: in many situations the null hypothesis will be thought not to be true.

One area where null hypotheses have quite high prior probabilities is model checking, including both goodness-of-fit testing and diagnostic testing. Here specific alternative hypotheses may not be well formulated, and significance test P values provide one convenient way to put useful measures on a standard scale.

Rather different is the problem of choosing between two, or a few, separate families of models. Here the symmetric roles of the hypotheses seem to me to make significance testing very artificial. It would be better to adopt fair empirical comparisons, using cross-validation or bootstrap methods, or a full-fledged Bayesian calculation. The latter requires careful choice of prior distributions within each model to avoid inconsistencies.

Significance tests will sometimes be used for a nuisance factor, preliminary to the main test, as with the initial test for a cross-over effect in a comparative trial with cross-over design. Racine, Grieve, Fluhler, and Smith (1986) recently demonstrated the clear merits of a Bayesian approach in this context. If significance tests are to be useful, then they should have validity independent of the values of identifiable nuisance factors.

In general, for problems where the usual null hypothesis defines a special value for a parameter, surely it would be more informative to give a confidence range for that parameter. Note that some significance tests are not compatible with efficient confidence statements, simply

*David V. Hinkley is Professor, Department of Mathematics, University of Texas, Austin, TX 78712.

Journal of the American Statistical Association
March 1987, Vol. 82, No. 397, Theory and Methods

because a test contrast has been standardized by a null hypothesis standard error. Such a practice may be computationally convenient, as with score tests, but its negative features should not be overlooked.

One must agree that the operational interpretation of P values must be made relative to the amount of information available in the data, as expressed through ancillary statistics. Barnard (1982) argued cogently for this in the context of repeated significance tests, where a fixed cutoff for P values can lead to drastic loss of overall power. Of course confidence statements automatically account for available information, if proper conditioning is employed.

ADDITIONAL REFERENCES

Barnard, G. A. (1982), "Conditionality Versus Similarity in the Analysis of 2 × 2 Tables," in *Statistics and Probability: Essays in Honor of C. R. Rao*, eds. G. Kallianpur, P. R. Krishnaiah, and J. K. Ghosh, Amsterdam: North-Holland, pp. 59–65.

Racine, A., Grieve, A. P., Fluhler, H., and Smith, A. F. M. (1986), "Bayesian Methods in Practice: Experiences in the Pharmaceutical Industry" (with discussion), *Applied Statistics*, 35.

Comment

JAMES M. DICKEY*

What should our reaction be to the results announced in these two articles? What do they actually say to us, and what difference should it make in statistical practice? Before attempting to answer these questions, I would like to bring up a few relevant points.

Example 1, which runs through the Berger–Sellke article, is introduced by using the normal distribution, $\theta \sim \mathcal{N}(\theta_0, \sigma^2)$, as the conditional prior uncertainty given the alternative H_1. This distribution has the same variance as the sampling process. Consider, however, the generalization to an arbitrary prior variance, $\theta \sim \mathcal{N}(\theta_0, \tau^2)$, say $\tau^2 = \sigma^2/n^*$. In this notation, n/n^* represents the ratio $\tau^2/(\sigma^2/n)$ of the prior variance to the sampling variance of the sample mean. Unless I am mistaken, the expressions and tables in Sections 1 and 2 for the posterior probability $\Pr(H_0 \mid x)$ hold again for the more general case by merely replacing the variable n by n/n^* throughout. (The variable t retains its original definition in terms of the sample size n.) In many, if not most, areas of application, the conditional prior variance τ^2 is typically larger than the sampling variance σ^2. So the ratio n/n^* is larger than n, and one would find oneself looking further over in the right-hand (large-n) direction in Table 1 than if one pretended one's τ^2 equaled σ^2. In such applications, the effect touted here by Berger and Sellke is strengthened. The posterior probability of the null hypothesis tends not to be as small as the P value of the traditional test.

Theorems 2, 4, and 7 give lower bounds for the posterior probability of the null hypothesis in the case in which the corresponding prior probability π_0 is equal to $\frac{1}{2}$. Of course, the Bayes factor B, the ratio of posterior odds for H_0 to the corresponding prior odds $\pi_0/(1 - \pi_0)$, does not depend on π_0. Hence one is tempted to ask for versions of these theorems stated in terms of the Bayes factor. It is curious to see that the limits claimed for large t in these theorems do not appear in the accompanying tables as visible tendencies for increasing t. Rather, an opposite tendency, to move away from the limit, is exhibited. So it would seem that the limits are meaningless except for exorbitantly large values of t. (That is, meaningless in practice: H_0 would be strongly rejected by all methods before the limit would have any effect?) Have the authors done any investigating to see where the limits begin to take effect?

To my mind, the Casella–Berger article further supports the thesis of Berger and Sellke. Theorems 3.2 and 3.3 of Casella and Berger concern an infimum over a class of prior distributions. So the smallest corresponding posterior probability of one-sided H_0 equals the traditional P value, and this equality is attained for the extreme constant prior pseudodensity. That is, reasonable prior distributions give posterior probabilities for H_0 that are larger than the traditional P value, though perhaps not as much larger as in the case of a point null hypothesis.

By the way, the constant prior pseudodensity appears here in the second of its two legitimate roles in inference, as follows. Bayesian scientific reporting requires a report of the effect of the observed data on a whole range of prior distributions, keyed to context-meaningful prior uncertainties (Dickey 1973). "Noninformative" prior pseudodensities are sometimes useful for such reporting in two ways:

1. Such a prior can serve as a device to give a simple posterior distribution that approximates the posterior distributions from prior probability distributions expressing relevant context uncertainties. This approximation is quantified by L. J. Savage's "stable estimation" or "precise measurement" (Edwards, Lindman, and Savage 1963; Dickey 1976).

2. Such a prior can serve as a device to give bounds on posterior probabilities over classes of context-relevant prior distributions.

* James M. Dickey is Professor, School of Statistics, University of Minnesota, Minneapolis, MN 55455. This work was supported by National Science Foundation Research Grant DMS-8614793.

What should our attitude now be concerning P values? Berger and Sellke note that nonstatisticians tend to confuse the P value and the posterior probability of the null hypothesis. As pointed out in Good (1984), even the most respected statisticians can make the same mistake. The present works reinforce the distinction between sampling probability and posterior probability.

It has long seemed to me that the P value reports an interesting fact about the data. I once speculated to Dennis Lindley that the P value might offer a quicker and cruder form of inference than the Bayes factor. He replied by asking whether what I meant was analogous to comparing an orchestra with a tom-tom.

ADDITIONAL REFERENCES

Dickey, James M. (1976), "Approximate Posterior Distributions," *Journal of the American Statistical Association*, 71, 680–689.

Good, I. J. (1984), "An Error by Neyman Noticed by Dickey" (C209), in "Comments, Conjectures, and Conclusions," *Journal of Statistical Computation and Simulation*, 20, 159–160.

Comment

STEPHEN B. VARDEMAN*

Berger, Sellke, Casella, and Berger deserve our thanks for a most readable and thorough accounting of the problem of comparing p values and posterior probabilities of H_0. They have laid out in very clear fashion the history of the problem, a full array of technical points, and their arguments from the technical points to general conclusions. Their articles should help all of us, card-carrying Bayesians, militant frequentists, and fence-sitters like myself, to sort this issue out to our own satisfaction.

My view from the fence is that in spite of the fact that the articles are well done, there is nothing here very surprising or that carries deep philosophical implications. We all know that Bayesian and frequentist conclusions sometimes agree and sometimes do not, depending on the specifics of a problem. These articles seem to me to reinforce this truism. For example, I read the Casella/Berger Theorem 3.4, the argument behind it, and their subsequent discussion as confirmation that essentially anything can be possible for a posterior probability for H_0, *depending on how one is allowed to move prior mass around on* H_0 *and* H_1. (Of course, the simplest demonstration that nearly anything can be possible can be made by using arbitrary two-point priors in a composite versus composite case.)

Whether or not a Bayesian analysis can produce a small posterior probability for H_0 is largely a function of whether or not (staying within whatever rules are imposed by the problem structure and restrictions adopted for the prior) one can move the prior mass on H_0 "away from the data," at least as compared with the location of the prior mass on H_1. If this can be done, the posterior probability of H_0 can be made small, otherwise it cannot.

Take, for example, the Jeffreys–Lindley "paradox" discussed by Berger and Sellke. To maintain a p value that is constant with n (i.e., a constant value of t), one must send $\bar{X}_n$ (the data) to θ_0. The nonzero mass on H_0 is trapped at θ_0, while the mass on H_1 is all passed by as $\bar{X}_n \to \theta_0$. Why should anyone then be surprised that the posterior probability assigned to H_0 tends to 1?

Moving to a different point, I must say that I find the "spike at θ_0" feature of the priors used by Berger and Sellke and many before them to be completely unappealing. In fact, contrary to the exposition of Berger and Sellke, I think that the appeal of such priors *decreases* with increasing π_0. Unlike that of Casella and Berger, my objection has nothing to do with "impartiality" (indeed I question whether such a concept can have any real meaning), but is of a more elementary nature. The issue is simply that I do not believe that any scientist, when asked to sketch a distribution describing his belief about a physical constant like the speed of light, would produce anything like the priors used by Berger and Sellke. A unimodal distribution symmetric about the current best value? Probably. But with a spike or "extra" mass concentrated at θ_0? No.

Competent scientists do not believe their own models or theories, but rather treat them as convenient fictions. A small (or even 0) prior probability that the current theory is true is not just a device to make posterior probabilities as small as p values, it is the way good scientists think! The issue to a scientist is not whether a model is true, but rather whether there is another whose predictive power is enough better to justify movement from today's fiction to a new one. Scientific reluctance to change theories is appropriately quantified in terms of a cost structure, not by concentrating prior mass on H_0. In this regard, note that although the "spike at θ_0" priors are necessary to produce nontrivial Bayes rules (i.e., ones that sometimes "accept") for a zero–one type loss structure in the two-sided problem, other competing cost structures do not require them for a Bayesian formulation of the testing

* Stephen B. Vardeman is Professor, Statistics Department and Industrial Engineering Department, Iowa State University, Ames, IA 50011.

Journal of the American Statistical Association
March 1987, Vol. 82, No. 397, Theory and Methods

problem to be nontrivial. Consider, for example, a cost structure like

$$\text{cost}(\text{"reject,"}\ \theta) = k_1 - k_2(\theta - \theta_0)^2,$$

$$\text{cost}(\text{"accept,"}\ \theta) = k_3(\theta - \theta_0)^2$$

for positive constants k_1, k_2, and k_3. Here it is clearly possible to have $\Pr[H_0 \text{ is true} \mid \text{data}] = 0$ and at the same time have "accept" be the preferred decision.

A largely nontechnical observation that I feel obliged to make regarding both articles concerns word choice. I would prefer to see loaded words like "biased," "objective," and "impartial" left out of discussions of the present kind, albeit they are given local technical definitions. Too much of what *all* statisticians do, or at least talk about doing, is blatantly subjective for any of us to kid ourselves or the users of our technology into believing that we have operated "impartially" in any true sense. How does one "objectively" decide on a subject of investigation, what variable to measure, what instrument to use to measure it, what scale on which to express the result, what family of distributions to use to describe the response, etcetera, etcetera, etcetera? We can do what seems to us most appropriate, but we can *not* be objective and would do well to avoid language that hints to the contrary.

Having complimented the authors' thoroughness and clarity and expressed some skepticism regarding the depth of the implications that ought to be drawn from their results, I will close these remarks by pointing out what I found to be the most interesting issue they have raised. That is the role of conditioning in the stating of the strength of one's evidence against H_0. I have never been particularly comfortable while trying to convince elementary statistics students that having observed $t = 1.4$ they should immediately switch attention to the event $[|t| \geq 1.4]$. Although I am unmoved to abandon the practice, I do find it interesting that Berger and Sellke see this as the main point at which standard practice goes astray.

Comment

C. N. MORRIS*

These two articles address an extremely important point, one that needs to be understood by all statistical practitioners. I doubt that it is. Let us dwell on a simple realistic example here to see that the Berger–Sellke result is correct in spirit, although case-specific adjustments can be used in place of their lower bounds, and that the Casella–Berger infimum, although computed correctly, is too optimistic for most practical situations.

Example. Mr. Allen, the candidate for political Party A will run against Mr. Baker of Party B for office. Past races between these parties for this office were always close, and it seems that this one will be no exception—Party A candidates always have gotten between 40% and 60% of the vote and have won about half of the elections.

Allen needs to know, for $\theta \equiv$ the proportion of voters favoring him today, whether $H_0 : \theta < .5$ or $H_1 : \theta > .5$ is true. A random sample of n voters is taken, with Y voters favoring Allen. The population is large and it is justifiable to assume that $Y \sim \text{Bin}(n, \theta)$, the binomial distribution. The estimate $\hat{\theta} = Y/n$ will be used.

Question. Which of three outcomes, all having the same p value, would be most encouraging to candidate Allen?

(a) $Y = 15$, $n = 20$, $\hat{\theta} = .75$;

(b) $Y = 115$, $n = 200$, $\hat{\theta} = .575$;

or

(c) $Y = 1{,}046$, $n = 2{,}000$, $\hat{\theta} = .523$.

Facts. The p values are all about .021, with values of $t \equiv (\hat{\theta} - .5)\sqrt{n}/\sigma$, $\sigma \doteq .5$, being 2.03, 2.05, and 2.03. Standard 95% confidence intervals are (.560, .940), (.506, .644), and (.501, .545), respectively. (For the application with $n = 20$, exact binomial calculations are made, and continuity corrections are used for t throughout.)

This problem is modeled as $\hat{\theta} \sim N(\theta, \sigma^2/n)$, given θ, with $\sigma^2 = .25$ known, from binomial considerations. The two hypotheses are taken to be, with $\theta_0 \equiv .5$, $H_0 : \theta < \theta_0$ versus $H_1 : \theta > \theta_0$ (θ_0 is given essentially zero probability). We use the conjugate normal prior distribution, and because of information about past elections, we take $\theta \sim N(\theta_0, \tau^2)$ with $\tau = .05$ so that $\Pr(H_0) = \Pr(H_1) = \frac{1}{2}$ a priori (as both articles assume), and so very probably, .4

* C. N. Morris is Professor, Department of Mathematics and Center for Statistical Sciences, University of Texas, Austin, TX 78712. Support for this work was provided by National Science Foundation Grant DMS-8407876.

Table 1. Data, p Values, Posterior Probabilities, and Power at $\theta_1 = .55$ for the Three Surveys

Survey	(a)	(b)	(c)
n	20	200	2,000
$\hat{\theta}$	.750	.575	.523
t	2.03	2.05	2.03
p value	.021	.020	.021
C_n	.408	.816	.976
$\Pr(H_0 \mid t)$	.204	.047	.024
Power(@ 1.645)	.115	.409	.998
Power(@ t)	.057	.262	.993

$\leq \theta \leq .6$. Then t is the usual test statistic, and the p value is $\Phi(-t)$.

A standard calculation yields

$$\Pr(H_0 \mid \hat{\theta}) = \Phi(-C_n t) \tag{1}$$

with

$$C_n^2 \equiv \tau^2/(\tau^2 + \sigma^2/n) = n/(n + \sigma^2/\tau^2). \tag{2}$$

Note that the probability given in (1) decreases as n increases, in contrast to Jeffreys's formula reported in Table 1 of Berger and Sellke.

The results for the three surveys are reported in Table 1 here.

Survey (a) is far less comforting to Allen than is (b), which is less so than (c). Only for (c), with $C_n = .976$, does $P(H_0 \mid t)$ closely approximate the p value of .021. It is understood in making this assertion that winning and losing are the only items of interest, victory margin being irrelevant (in a real setting, this would be untrue if there were time to influence votes further).

Of course, other results might follow from the same data, but different information. If the election were not expected to be close, for example, if $\tau = .25$ were reasonable, then $C_{20} = .91$ and the p value .021 would be near $\Pr(H_0 \mid t)$ even for $n = 20$. Indeed, this is the Casella–Berger result for the normal distribution setting, that $\Pr(H_0 \mid t)$ diminishes as $\tau \to \infty$ to its minimum $\Phi(-t)$, the p value; check (1) and (2) to see this. Their result is correct, but irrelevant when one knows that τ is bounded above in such a way that C_n is substantially less than unity for all reasonable τ.

The key to understanding these results from any perspective, Bayesian or non-Bayesian, is that the result $\hat{\theta} = .75$ for Survey (a) is not much more likely for the values of θ that one expects to obtain under H_1 than it is if H_0 is true. That is, taking $\theta_1 = .55$ as a typical value for H_1, $\Pr(\hat{\theta} \geq .75 \mid \theta = \theta_1)$ is 5.7% for Survey (a), and it only rises to 12.6% when $\theta_1 = .60$, the largest tenable value for θ. To generalize, and perhaps to explain intuitively when p values fail to reflect probabilities, we note that rare event concepts underlie p value reasoning, but that

if a rare event for H_0 occurs that also is rare for typical H_1 values, it provides little evidence for rejecting H_0 in favor of H_1.

The final two rows of Table 1 provide the powers for the one-tailed tests in each survey at $\theta_1 = .55$, first for test size .05 (rejecting H_0 if $t \geq 1.645$) and in the latter row for test size $\Phi(-t)$, the p value. These power formulas then are $\Phi(\sqrt{n}\delta - 1.645)$ and $\Phi(\sqrt{n}\delta - t)$, respectively, defining $\delta \equiv (\theta_1 - \theta_0)/\sigma$ as the signal-to-noise ratio. Here $\theta_0 = .50$ and $\delta = .1$. We see from Table 1 that

the p value corresponds to $Pr(H_0 \mid t)$ only when good power obtains at typical H_1 parameter values.

I qualify this statement, however, here and in later remarks, by requiring that the parameter space H_1 include the interval between θ_0 and θ_1. Otherwise, in the simple H_0 versus simple H_1 case, for example, there would be excellent power at $\theta_1 = \theta_0 + \delta\sigma$ when δ is large, but at $t = \delta\sqrt{n}/2$, $\hat{\theta} = \theta_0 + \delta\sigma/2$, one has $\Pr(H_0 \mid t) = \frac{1}{2}$, even with a statistically significant test statistic.

Practical statisticians, be they Bayesian or frequentist, have to assess the possible "typical" values θ_1 in H_1 when they design experiments, if only for the purpose of making power calculations to justify the sample size. If we label θ_1 as a typical value when it falls one (prior) standard deviation above the null value θ_0, $\theta_1 = \theta_0 + \tau$, then $C_n^2 = n\delta^2/(1 + n\delta^2)$.

Thus

$$t^* \equiv C_n t \tag{3}$$

is the "corrected" standardized statistic, since then $\Pr(H_0 \mid t) = \Phi(-t^*) = p$ value if t^* had been observed in place of t. Tables of the normal distribution can be applied directly to t^*. In the survey example, taking $t^* = 1.645$ for 5% significance, values of $t = t^*/C_n$ equaling 4.03, 2.01, and 1.69 would be required for $n = 20$, 200, 2,000. Such corrections t^* are in the spirit of the Berger–Sellke rule of thumb for modifying standardized test statistics, but go further because they also incorporate the particular features of each problem.

The essential distinction between the results for two-sided tests and one-sided tests, considered by the authors of these two articles and various others before them, seems not to depend on the number of sides of the test, but on whether all prior probability mass is allowed to slip off to infinity. When that cannot happen, and it automatically cannot in two-sided situations, the p value will tend to be too low. Otherwise, Casella–Berger type results will obtain and p values will be more appropriate. The heuristics of the one-sided survey example are relevant to the Berger–Sellke situation, but the example could easily have been extended to their two-sided situation at the cost of increased complexity.

When significant power is available at reasonable alternatives in H_1, p values will work well. But otherwise they generally overstate evidence. Thus they usually would be reliable for the primary hypotheses in well-designed (for good power) experiments, surveys, and observational studies. But for hypotheses of secondary interest, and

when on "fishing expeditions" with data from unplanned studies, adjustments to t values like those suggested by Berger and Sellke or in formula (3) are mandatory. These facts need to be better understood by the wide population of individuals doing data analyses or interpreting the reports of such analyses. They need to be taught in introductory courses, perhaps when the power of tests is introduced, and should be recognized by the editors of journals that report empirical work in terms of significance tests and p values.

Rejoinder

JAMES O. BERGER and THOMAS SELLKE

We thank all discussants for their interesting comments. Our rejoinder will rather naturally emphasize any disagreements or controversy, and thus will be mainly addressed to the non-Bayesians. We are appreciative of the expressed disagreements, including those of Casella and Berger, since one of our hopes was to provoke discussion of these issues in the profession. These are not dead issues, in the sense of being well known and thoroughly aired long ago; although the issues are not new, we have found the vast majority of statisticians to be largely unaware of them. We should also mention that the commentaries contain many important additional insights with which we agree but will not have the space to discuss adequately. Before replying to the official discussants, we have several comments on the Casella–Berger article.

1. COMMENTS ON THE CASELLA–BERGER ARTICLE

First, we would like to congratulate Casella and Berger on an interesting piece of work; particularly noteworthy was the establishment of the P value as the attained *lower bound* on the posterior probability of the null for many standard one-sided testing situations. It was previously well known that the P value was the limit of the posterior probabilities for increasingly vague priors, but that it is typically the lower bound was not appreciated. And the less common examples, where the lower bound is even smaller than the P value, are certainly of theoretical interest.

Our basic view of the Casella–Berger article, however, is that it pounds another nail into the coffin of P values. To clarify why, consider what it is that makes a statistical concept valuable; of primary importance is that the concept must convey a well-understood and sensible message for the vast majority of problems to which it is applied. Statistical models are valuable, because they can be widely used and yield similar interpretations each time they apply. The notion of 95% "confidence" sets (we here use "confidence" in a nondenominational sense) is valuable, because, for most problems, people know how to interpret them (conditional counterexamples aside). But what can be said about P values? Well, they can certainly be defined for the vast majority of testing problems, but do they give a "sensible message"? In our article we argued that they do not give a sensible message for testing a precise null hypothesis, but one could make the counterargument that this is merely a calibration problem. The P value is after all (usually) a one-to-one monotonic function of the posterior probability of the null, and one could perhaps calibrate or "learn how to interpret P values." This is

Journal of the American Statistical Association
March 1987, Vol. 82, No. 397, Theory and Methods

possible, however, only if the calibration is fairly simple and *constant*. In our article we mentioned one well-known source of nonconstancy in interpretation of the P value: as the sample size increases in testing precise hypotheses, a given P value provides less and less real evidence against the null. One could perhaps argue that a different calibration can be found for each sample size. But now Casella and Berger have also demonstrated that one must calibrate by the nature of the problem. For one-sided testing, a P value is often roughly equivalent to evidence against H_0, whereas for testing a precise hypothesis a P value must typically be multiplied by a factor of 10 or more to yield the same evidential interpretation. And these are *not* the only two possibilities. Indeed, suppose that the null hypothesis is an interval of the form $H_0 : |\theta - \theta_0| \leq C$. If C is near 0, one is effectively in the point null situation, and as C gets large the situation becomes similar to one-sided testing. For C in between, there is a continuum of different possible "calibrations."

Although somewhat less important than the sample size and C, the dimension of the problem and the distribution being considered can also necessitate different calibrations between P values and "evidence against H_0." The bottom line is simple: the concept of a P value is faulty, in that it does not have a reasonable direct interpretation as to evidence against H_0 over the spectrum of testing problems. It may be useful to identify when P values are (and are not) sensible measures of evidence, so as to allow reappraisal of those scientific results that have been based on P values, but the future of the concept in statistics is highly questionable.

Another issue raised in the article of Casella and Berger has to do with the validity of precise hypothesis testing. It is implied in Section 1 that one-sided tests are more useful in practice, and in Section 4 that placing mass near a point can be considered as "biasing the result in favor of H_0"; the practical import of our results is thus questioned. This issue is complicated by the fact that, in practice, many testing problems are erroneously formulated as tests of point null hypotheses. There is undeniably a huge number of such tests performed, but how many should be so formulated?

One answer to this objection is simply to note that we have little professional control over misformulations in statistics; we do, however, have some control over the statistical analysis performed for a given formulation. It is awkward to argue that a bad analysis of a given formulation is okay because the formulation is often wrong.

At a deeper level, it is possible even to argue the other way on the question of proper formulations of testing; one can argue that it is actually precise nulls that encompass the majority of "true" testing problems. This argument notes that most one-sided testing problems have to do with things like deciding whether a treatment has a positive or negative effect, or which of two treatments is best. The point is that, in such problems, what is typically really desired is an evaluation of how *large* the effect is or how *much* better one treatment is than another. Such problems are more naturally formulated as estimation or decision problems, and the appropriateness of testing is then debatable.

Precise hypotheses, on the other hand, ideally relate to, say, some precise theory being tested. Of primary interest is whether the theory is right or wrong; the amount by which it is wrong may be of interest in developing alternative theories, but the initial question of interest is that modeled by the precise hypothesis test.

In such problems the key fact is that there *is* real belief that the null hypothesis could be approximately true. If I am an experimenter conducting a test that will show, hopefully, that vitamin C has a beneficial effect on the common cold, I had better officially entertain the hypothesis that its effect is essentially negligible. In other words, I should not take the prior mass assigned to "no positive effect" and spread it out equally over all $\theta \leq 0$; this does not correspond to the reality that most people may be quite ready to believe that vitamin C is not harmful, yet give substantial weight to a belief in no or little effect. Such situations require substantial prior mass near 0.

We present the previous argument about what is "practical hypothesis testing" only halfheartedly. The huge variety of applications in which P values are used (see Cox 1977) makes questionable any claim that only "one type" of situation need be considered from a practical perspective. Whether most situations are one-sided, have a precise null hypothesis, or are really decision problems is irrelevant; our basic statistical theory should handle all.

2. REPLY TO HINKLEY

Hinkley defends the P value as an "unambiguously objective error rate." The use of the term "error rate" suggests that the frequentist justifications, such as they are, for confidence intervals and fixed α-level hypothesis tests carry over to P values. This is not true. Hinkley's interpretation of the P value as an error rate is presumably as follows: the P value is the Type I error rate that would result if this observed P value were used as the critical significance level in a long sequence of hypothesis tests [see Cox and Hinkley (1974, p. 66): "Hence [the P value] is the probability that we would mistakenly declare there to be evidence against H_0, were we to regard the data under analysis as being just decisive against H_0."] This hypothetical error rate does not conform to the usual classical notion of "repeated-use" error rate, since the P value is determined only once in this sequence of tests. The frequentist justifications of significance tests and confidence intervals are in terms of how these procedures perform when used repeatedly.

Can P values be justified on the basis of how they perform in repeated use? We doubt it. For one thing, how would one measure the performance of P values? With significance tests and confidence intervals, they are either right or wrong, so it is possible to talk about error rates. If one introduces a decision rule into the situation by saying that H_0 is rejected when the P value $\leq .05$, then of course the classical error rate is .05, but the expected P value given rejection is .025, an average understatement of the error rate by a factor of two.

In the absence of an unambiguous interpretation of P values as a repeated-use error rate, we have most frequently heard P values defended as a measure of the evidence against H_0, via an "either H_0 is true or a rare event has occurred" argument. It is for this reason that we concentrated on evaluating P values in terms of whether they really are effective in conveying information about the strength of the evidence against H_0. We acknowledge the difficulty in defining "evidence" in an absolute (non-Bayesian) sense, and for this reason we considered a variety of notions of evidence in the article, including lower bounds on the Bayes factor (or weighted likelihood ratio). Indeed, the lower bound on the Bayes factor strikes us as having a true claim to being "unambiguously objective," since it depends on no prior inputs at all (Th. 1) or only on a symmetry assumption (Th. 3) and yet relates to a valid (conditional) measure of evidence.

We indicated in Comment 2 that the results can be extended to goodness-of-fit testing and yield much the same conclusions, even when the alternative hypotheses are not well formulated. One can find lower bonds over essentially arbitrary alternatives within the chi-squared testing framework. Thus, whether or not the P value can really be considered as a standard scale, its interpretation in terms of evidence against H_0 should be sharply qualified.

We would disagree with the idea that usual confidence ranges for a parameter are more informative than posterior probabilities of hypotheses, when the null hypothesis defines a special value for a parameter. As an example, the density (on $\mathbf{R}^1$)

$$f(x|\theta) = (1 + \varepsilon) - 4\varepsilon|x - \theta|, \qquad \text{for } |x - \theta| \le \tfrac{1}{2},$$

will yield, as a usual 95% confidence set for small ε,

$$C(x) = (x - .475, x + .475);$$

but if $\theta = 0$ is a special value and $x = .48$ is observed, we would be loathe to reject $H_0 : \theta = 0$, since

$$f(.48|0)/\sup_{\theta \in C(.48)} f(.48|\theta) \ge (1 - \varepsilon)/(1 + \varepsilon).$$

The point is that a special parameter value outside a confidence set can have virtually the same likelihood as any parameter value inside a confidence set, and we would then argue that the data do not indicate rejection of the special parameter value. This phenomenon also occurs in the normal testing problem we discuss, though to a lesser degree.

We are wholeheartedly in agreement that proper conditioning must be employed. To us, however, this is even more important in testing than with confidence sets. We feel that refusing to "condition" on the actual data x, and instead using the set A of "as or more extreme" values, causes more harm in statistical practice than other failures to condition.

3. REPLY TO VARDEMAN

Our major disagreement seems to center again on the issue of concentrating prior mass near θ_0. We argued previously that (a) in examples such as the "vitamin C" example, one often does have mass near θ_0, and (b) even if H_0 is a fair-sized interval, the contradiction occurs (the agreement of posterior probabilities with P values only occurring in the limiting case in which H_0 is a very large interval with prior mass "uniformly" distributed over it).

Perhaps less controversy would have ensued if we had used Bayes factors or weighted likelihood ratios as our central measure. The argument then avoids the loaded issue of "prior beliefs" and simply says "how does the support of the data for H_0, given by the likelihood $f(x \mid \theta_0)$, compare with the support of the data for H_1, given by some average of $f(x \mid \theta)$ over θ in H_1." This is the Bayes factor, with g being the averaging measure on H_1, and the various theorems find bounds on the Bayes factor over g. If θ_0 has no distinction, as in the scenario of Casella and Berger, one probably does not care if $f(x \mid \theta_0)$ is a substantial fraction of the weighted likelihood of H_1; on the other hand, if θ_0 has the distinction of being a particular value for which it is desired to assess the evidence for or against, it is hard to ignore a comparatively large value of $f(x \mid \theta_0)$. We chose not to emphasize this "likelihood" argument, because we have found that the interpretation of observed likelihood ratios as direct evidence (and not just as inputs into a classical test) is less familiar to many classical statisticians than is the use of posterior probabilities as evidence.

This also relates to the issue of our agreed-upon discomfort at replacing $t = 1.4$ by the event $[|t| \ge 1.4]$. In the normal case (and most others), $f(1.4 \mid \theta_0)$ is a substantial fraction of any reasonable average of $f(1.4 \mid \theta)$ over H_1. On the other hand, $\Pr([|t| \ge 1.4] \mid \theta_0)$ is much smaller than reasonable averages of $\Pr([|t| \ge 1.4] \mid \theta)$ over H_1. Thus, by likelihood reasoning, there is also a great difference between knowing precisely that $t = 1.4$ and knowing only that $|t| \ge 1.4$; the latter would yield much greater evidence against H_0.

Another illustration of the conditioning aspect of the problem is described in our story about the "astronomer" in Section 1. We would really like to see an explanation, written for this astronomer, as to why he should believe that $t = 1.96$ is substantial evidence against H_0. The general point is that any method of conditionally measuring evidence that we have considered indicates that the replacement of $t = 1.4$ by $[|t| \ge 1.4]$ is the source of the huge discrepancies; and the replacement has no real justification except that of "convenience." One of the purposes of this article was to indicate a common statistical situation in which it is essential to condition properly, feeling that the issue of conditioning *is* one of the deepest and most important issues in statistics.

We applaud Vardeman's leanings toward decision-theoretic formulations, though we have argued that one should not completely abandon the possibility of stating how much the data support a special value θ_0. We also are not particularly at ease with the use of words like "objective," but we use them out of a certain defensive posture. Many statisticians feel that it is possible and essential to be objective; whether or not this really is possible, we

would argue that the closest one can come to objectivity is through the types of conditional analyses we have discussed. (See Comment 3 for our views concerning the actual possibility of objectivity.)

4. REPLY TO DICKEY

The observation that n in Table 1 can, in general, be replaced by the ratio of the prior and sampling variances is a useful fact (pointed out also by Pratt). It is interesting that the accuracy of the point null formulation (i.e., the appropriateness of the approximation of a realistic small interval null by a point) depends on $\sigma/\sqrt{n}$ but not on τ^2; thus if τ^2 is indeed larger than σ^2, one can move to the right in the table without increased worry concerning the soundness of the formulation.

The asymptotic t arguments are given for completeness, but it is true that the asymptotics take effect for t too large to be of much interest. We agree with all other comments, except that the equating of a P value with a tom-tom strikes us as somewhat overly positive.

5. REPLY TO PRATT

We are in complete agreement that Edwards, Lindman, and Savage (1963) (EL&S) contained the essence of our article. Indeed, had EL&S not been so mysteriously ignored for so long, our contribution would have been mainly a presentation of Theorem 5, its ramifications, and the results in Section 4. Because very few people we talked to were aware of the results in EL&S, however, a general review seemed to be in order. We feel that the result of Theorem 5 is a substantive advance for two reasons. First, although the results for G_{US} are not greatly different from those for G_N, this is not apparent a priori; non-Bayesians tend to be very wary of a result established for only normal priors, so verifying that the same answer holds qualitatively for all unimodal symmetric priors can substantially enhance the impact of the basic phenomenon. Second, the techniques for working with large classes, such as G_{US}, are important in general Bayesian sensitivity studies, and we hoped that the application here would indicate the possibilities and kindle interest. Finally, the result on interval hypotheses in Section 4 is valuable for both sociological and scientific purposes.

Pratt's Table 1 and the subsequent comments and insights are all of value. We agree with his later comment that our Comment 1 is probably not cautious enough; it was given with the simple hope that a not-too-terrible rule of thumb might be able to drive out a terrible rule of thumb.

6. REPLY TO GOOD

There is virtually nothing in this interesting set of comments with which we disagree. We would probably have to align ourselves with the radical Bayesians, however, in that we remain unconvinced that P values have any merit. The number of "rules of thumb" that have to be learned to "calibrate" properly P values in the various possible testing situations is so large that it strikes us as simply unwieldy to continue to use them. Why not just shift over to Bayes factors (or bounds on the Bayes factors)? We would agree that often (though not always) a P value of .05 is an indication that more evidence should be obtained.

We thank Good for the additional references; we tried, but knew we must have missed some.

7. REPLY TO MORRIS

Morris raises a number of interesting issues that bear on the comparison of the one-sided and precise null testing situations. For ease in discussion, it is helpful to consider a precise null version of the example of Morris.

Example. Consider a paired comparison experiment in which two new treatments will be screened. The outcome for each subject pair is a 0 or 1, depending on which treatment is judged to be superior. Let θ denote the probability of obtaining a 1, and let n denote the number of (independent) pairs in the experiment. These are two new treatments, and it is judged that there is a substantial probability ($\frac{1}{2}$, say) that they are both ineffective, which would correspond to a θ very near $\frac{1}{2}$. All past experiments with similar treatments have indicated that, when there are treatment effects, θ ranges between .4 and .6. Indeed (as in the Morris example), suppose that we view it reasonable to model this θ, a priori (conditional on there being treatment effects), as having an $\mathcal{N}(\frac{1}{2}, (.05)^2)$ distribution. Assuming that the normal approximation for $\hat{\theta}$ is valid, the entire model above falls within the framework of our article, with $X \doteq \hat{\theta} \sim \mathcal{N}(\theta, .25/n)$, the desire to test $H_0 : \theta = \frac{1}{2}$ versus $H_1 : \theta \neq \frac{1}{2}$, $\pi_0 = \frac{1}{2}$, and $g(\theta)$ being the $\mathcal{N}(\frac{1}{2}, (.05)^2)$ density.

The difference between this problem and that of Morris is, of course, that there *is* substantial reason to suspect $\theta = \frac{1}{2}$; in a voting situation there is no reason to single our $\theta = \frac{1}{2}$ as deserving positive prior mass. (We implicitly assume that n is not enormous; the real hypothesis of "no treatment effects" would be accurately modeled as $H_0 : |\theta - \frac{1}{2}| \leq \varepsilon$, and if n is enormous it can be inaccurate to approximate this by $H_0 : \theta = \frac{1}{2}$.)

By using an easy modification of formula (1.1), we can calculate the posterior probability of H_0 for each of the situations in Table 1 of Morris. The results for $n = 20$, $n = 200$, and $n = 2{,}000$, respectively, are .436, .302, and .387; compare these with the posterior probabilities found by Morris of .204, .047, and .024, respectively. Note, in particular, the huge difference for $n = 2{,}000$.

The example here makes clear that the insightful comments of Morris, although valid for the situation in which no special mass is to be assigned to a point θ_0, need not be valid for the precise null situation. For instance, the comment "the P value corresponds to $\Pr(H_0 \mid t)$ only when good power obtains at typical H_1 parameter values" may be valid for nonprecise nulls but is false for precise nulls; the powers at $\theta = .55$ for our example are very near 1

when $n = 2{,}000$, yet the P value differs drastically from the posterior probability of H_0.

The necessary distinction between precise nulls and imprecise nulls only reinforces the exhortation (with which we completely agree), in the last paragraph of Morris's comment, to the effect that it is crucial for all statisticians and scientists using P values to learn exactly what P values do and do not convey about the evidence against H_0 in the wide variety of testing problems to which they are applied.

ADDITIONAL REFERENCES

Cox, D. R. (1977), "The Role of Significance Tests," *Scandinavian Journal of Statistics*, 4, 49–70.

Cox, D. R., and Hinkley, D. V. (1974), *Theoretical Statistics*, London: Chapman & Hall.