Testing Precise Hypotheses

James O. Berger; Mohan Delampady

Statistical Science, Vol. 2, No. 3 (Aug., 1987), 317-335.

Stable URL:
http://links.jstor.org/sici?sici=0883-4237%28198708%292%3A3%3C317%3ATPH%3E2.0.CO%3B2-W

Statistical Science is currently published by Institute of Mathematical Statistics.

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at http://www.jstor.org/about/terms.html. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at http://www.jstor.org/journals/ims.html.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

JSTOR is an independent not-for-profit organization dedicated to creating and preserving a digital archive of scholarly journals. For more information regarding JSTOR, please contact support@jstor.org.

http://www.jstor.org/
Mon Apr 4 11:51:17 2005

Statistical Science
1987, Vol. 2, No. 3, 317–352

Testing Precise Hypotheses

James O. Berger and Mohan Delampady

Abstract. Testing of precise (point or small interval) hypotheses is reviewed, with special emphasis placed on exploring the dramatic conflict between conditional measures (Bayes factors and posterior probabilities) and the classical P-value (or observed significance level). This conflict is highlighted by finding lower bounds on the conditional measures over wide classes of priors, in normal and binomial situations, lower bounds, which are much larger than the P-value; this leads to the recommendation of several alternatives to P-values. Results are also given concerning the validity of approximating an interval null by a point null. The overall discussion features critical examination of issues such as the probability of objective testing and the possibility of testing from confidence sets.

Key words and phrases: Point null hypothesis, P-value, Bayes factor, posterior probability, objectivity, robust Bayesian analysis, Jeffreys's paradox, binomial tests, χ^2 tests, scientific communication.

1. INTRODUCTION AND BASICS

1.1 Basics and Measures of Evidence

Suppose X having density $f(x \mid \theta)$ is observed, with θ being an unknown element of the parameter space Θ, and that it is desired to test H_0: $\theta = \theta_0$ versus H_1: $\theta \neq \theta_0$. (In Section 2 this will be argued to be a good approximation to many realistic scenarios concerning testing of a precise hypothesis.) We consider and compare three measures of evidence against H_0, the classical P-value, the weighted likelihood ratio or Bayes factor and the Bayesian posterior probability of H_0.

P-value. Let $T(X)$ be a test statistic, extreme values of which are deemed to be evidence against H_0. If $X = x$ is observed, with corresponding $t = T(x)$, the P-value (or observed significance level) is

$$\alpha = P_{\theta_0}(\mid T(X) \mid \geq \mid t \mid). \tag{1}$$

Bayes factor. Let $g(\theta)$ be a continuous density on $\{\theta \neq \theta_0\}$. Then the Bayes factor, or weighted likelihood ratio of H_0 to H_1, is

$$B = \frac{f(x \mid \theta_0)}{m_g(x)}, \tag{2}$$

where

$$m_g(x) = \int f(x \mid \theta) g(\theta)\, d\theta. \tag{3}$$

For a Bayesian, g would be the prior density for θ, conditional on H_1 being true. For a likelihoodist, g might be thought of merely as some weight function to allow the computation of an average likelihood for H_1. B might then be called a "weighted likelihood ratio" for the two hypotheses. Its interpretation is similar to that of a usual likelihood ratio; e.g., a value of $B = 1/10$ means that H_1 is supported ten times as much by the data as is H_0.

Posterior probability. If a Bayesian specifies, in addition to g, the prior probability of H_0, to be denoted by π_0, then the posterior probability of H_0 is

$$\begin{aligned} P(H_0 \mid x) &= \left[1 + \frac{(1-\pi_0)}{\pi_0} \frac{m_g(x)}{f(x \mid \theta_0)}\right]^{-1} \\ &= \left[1 + \frac{(1-\pi_0)}{\pi_0} \frac{1}{B}\right]^{-1}. \end{aligned} \tag{4}$$

Example 1. Suppose we observe $\bar{X} \sim N(\theta, \sigma^2/n)$, where σ^2 is known. Then, letting

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

one obtains the usual P-value as

$$\alpha = 2[1 - \Phi(\mid t \mid)], \tag{5}$$

where Φ is the standard normal cumulative distribution function (cdf).

James O. Berger is the Richard M. Brumfield Distinguished Professor of Statistics at Purdue University, West Lafayette, Indiana 47907. Mohan Delampady is Assistant Professor, Department of Statistics, University of British Columbia, Vancouver, British Columbia V6T 1W5, Canada.

An easy to analyze density g is the $N(\mu, \tau^2)$ density. Calculation yields that

$$(6) \quad B = \sqrt{1+\rho^{-2}} \exp\left\{-\frac{1}{2}\left[\frac{(t-\rho\eta)^2}{(1+\rho^2)} - \eta^2\right]\right\},$$

where $\rho = \sigma/(\sqrt{n}\tau)$ and $\eta = (\theta_0 - \mu)/\tau$. Note that μ will often be chosen to be θ_0 (so as to have a symmetric "weight function"), in which case

$$(7) \quad B = \sqrt{1+\rho^{-2}} \exp\left\{-\frac{1}{2}\left[\frac{t^2}{(1+\rho^2)}\right]\right\}.$$

The posterior probability of H_0 can be found from these formulas and (4), provided π_0 is specified.

As a specific example, suppose $\mu = \theta_0$, $\tau = \sigma$ and $\tau_0 = ½$. For various t and n, the various measures of evidence are given in Table 1. There P stands for $P(H_0 \mid x)$.

1.2 Motivation

Here are some commonly held opinions that this testing scenario strongly contraindicates:

Opinion 1. *Classical and Likelihood or Bayesian Answers Typically Agree.* Bayesian answers can, of course, vary markedly depending on the choice of prior, so that a more precise phrasing of this opinion would be that classical answers will agree with some sensible ("objective") Bayesian analysis. Such is simply not the case in the testing of precise hypotheses. This is indicated in Table 1 where, for instance, $P(H_0 \mid x)$ is from 5 to 50 times larger than the P-value, α. The Bayesian analysis here is close to that recommended by Jeffreys (1961) as a "standard" Bayesian significance test. (Jeffreys chose a Cauchy form for the prior, but this makes a substantial difference only when $|t|$ is large.) Thus, if $n = 50$ and $t = 1.960$, Jeffreys would conclude that H_0 has probability .52 of being true, although the classical statistician would "reject H_0 at significance level $\alpha = .05$." Admittedly, classical statisticians will warn against interpreting α as the probability that H_0 is true, but surely the classicist feels that $\alpha = .05$ is reasonable cause to doubt H_0, in marked contrast to the Bayesian conclusion. This is perhaps the simplest problem where the Bayesian and classical statistician are in fundamental practical disagreement, and as such, the problem deserves intense study. (Our label "classical statistician" is admittedly ambiguous; there are statisticians who consider themselves to be "classical," and yet do not view P-values as meaningful measures of evidence, and there are Bayesians who view P-values as useful measures of evidence; see Section 5.)

TABLE 1
Measures of evidence, normal example

t	α	n = 1		5		10		20		50		100	
		B	P	B	P	B	P	B	P	B	P	B	P
1.645	.10	.72	.42	.79	.44	.89	.47	1.27	.56	1.86	.65	2.57	.72
1.960	.05	.54	.35	.49	.33	.59	.37	.72	.42	1.08	.52	1.50	.60
2.576	.01	.27	.21	.15	.13	.16	.14	.19	.16	.28	.22	.37	.27
3.291	.001	.10	.09	.03	.03	.02	.02	.03	.03	.03	.03	.05	.05

Opinion 2. *Objective Bayesian Analyses Are Always Possible.* Having automatic statistical procedures available is certainly advantageous for certain users. The frequent criticism of Bayesian analysis, to the effect that (nonautomatic) prior specification is required, is effectively answered by the well developed and very successful Bayesian approach using noninformative ("objective") priors (see Berger (1985) for references). Do such objective Bayesian methods always exist, however? The answer is no, and the precise null testing situation is a prime example in which objective procedures do not exist.

To see this, note that a Bayesian must specify π_0 and g. One can argue that $\pi_0 = ½$ is the objective choice for the prior probability of H_0, or can objectively avoid the choice of π_0 by concentrating on the Bayes factor, B. There is no choice of g, however, that can claim to be objective. One might argue that g should be symmetric about θ_0 (at least when the parameter space is the entire real line), and perhaps that g should be nonincreasing in $|\theta - \theta_0|$ (to avoid treating values of θ other than θ_0 as special). For many problems the exact functional form of g can be shown to be rather irrelevant, so that one might argue in Example 1, say, that g could be taken to be $N(\theta_0, \tau^2)$ or maybe *Cauchy* (θ_0, τ^2) (the form preferred by Jeffreys). Unfortunately, the choice of the scale factor, τ, for g has a large effect on the answer. This can be seen in Example 1 by looking at (7); for interesting (moderate to large) τ the value of ρ will typically be small, so that

$$B \cong \frac{\sqrt{n}\tau}{\sigma} \exp\left(-\frac{1}{2}t^2\right),$$

and (with $\pi_0 = ½$),

$$P(H_0 \mid x) \cong \frac{1}{1 + (\sigma/\sqrt{n}\tau)\exp(½t^2)}.$$

Thus, τ has a dramatic effect on the Bayesian or likelihood answer. Furthermore, letting $\tau^2 \to \infty$ so that g becomes "noninformative" is ridiculous, because then $P(H_0 \mid x) \to 1$. Thus, a Bayesian must, at a minimum, subjectively specify τ^2, and there is no default value that "lets the data speak for itself."

In light of this fact, the derivations of "automatic" Bayesian significance tests in Jeffreys (1961) (see also Zellner and Siow, 1980, and Smith and Spiegelhalter, 1980) are of considerable interest. Both proceed by

arguing that "if one must specify a default g for automatic use, then a good such g is $\cdots$". In Example 1, Jeffreys argued for a *Cauchy* (θ_0, σ^2) default g, although Smith and Spiegelhalter argued for a constant default g (but a particular constant); these default g actually often give very similar answers.

We agree that, if one must produce an automatic Bayesian significance test, then the Jeffreys, Zellner-Siow or Smith-Spiegelhalter tests are quite satisfactory. Furthermore, we feel that automatic use of such tests is vastly superior to automatic use of P-values, for reasons to be made clear later. Nevertheless, we would argue that either test imposes a particular and highly informative g on the user, and as such cannot claim to be objective.

This should *not* be interpreted as lending support to the P-value, because its formal definition in (1) appears to be objective; appearances can be deceiving. For instance, in Example 1 when $t = 1.96$, a posterior probability of 0.05 can only be achieved (among symmetric unimodal weight functions g, say) by choosing π_0 to be 0.11 or smaller. It would certainly not be "objective" to state that the posterior probability of H_0 is 0.05, hiding the fact that most of this "evidence" is due to the prior probability being only 0.11. Yet many users of P-values do interpret 0.05 as providing 19 to 1 evidence against H_0, i.e., interpret it as a posterior probability. We feel that the correct interpretation of a P-value, although perhaps objective, is nearly meaningless, and that the actual meaning usually ascribed to a P-value by practitioners contains hidden and extreme bias.

Opinion 3. *Testing Is Somewhat Irrelevant; One Should Concentrate on Confidence Sets, Testing from Them If Necessary.* The motivation for this opinion is that significance testing ignores a crucial question, namely "how far is θ from θ_0?" Having statistically significant evidence that $\theta \neq \theta_0$ might be irrelevant if θ and θ_0 are within, say, 10^{-10} of each other. So, the argument goes, one should simply find (say) a 95% confidence set for θ. If θ_0 is not in this set it can be rejected, and looking at the set will provide a good indication as to the actual magnitude of the difference between θ and θ_0.

This opinion is wrong, because it ignores the supposed special nature of θ_0. A point can be outside a 95% confidence set, yet not be so strongly contraindicated by the data. Only by calculating a Bayes factor (or related conditional measure) can one judge how well the data supports a distinguished point θ_0. Further discussion of this will be given in Section 4.3.

1.3 Lower Bounds on Conditional Measures

A non-Bayesian might be tempted to dismiss the conflict between α and the conditional measures B or $P(H_0 \mid x)$ in Table 1 by arguing that the difference is simply due to the specific g that was chosen. That this is not the case can be seen by looking at lower bounds on B and $P(H_0 \mid x)$ over wide classes of g, and observing that even these lower bounds are much larger than α. Thus, if G is a class of densities under consideration, we will investigate

$$\underline{B} = \inf_{g \in G} B = \frac{f(x \mid \theta_0)}{\sup_{g \in G} m_g(x)} \tag{8}$$

and

$$\begin{aligned} \underline{P}(H_0 \mid x) &= \inf_{g \in G} P(H_0 \mid x) \\ &= \left[1 + \frac{(1 - \pi_0)}{\pi_0} \cdot \frac{1}{\underline{B}}\right]^{-1}. \end{aligned} \tag{9}$$

Interesting classes of G to consider include $G = \{\text{all densities}\}$, $G = \{\text{all conjugate densities with mean } \theta_0\}$,

$$G = \{\text{all densities symmetric about } \theta_0 \text{ and nonincreasing in } |\theta - \theta_0|\} \tag{10}$$

and variations thereof. (By "conjugate densities" here we will mean "textbook" conjugate priors, although other versions could be considered.) The most interesting choices for G are those that are neither too big nor too small. One wants G to include all g that are *a priori* reasonable (in order for the lower bounds to be valid lower bounds), but a too large G might result in uselessly small bounds. Another way of saying this is that $\underline{B}$ or $\underline{P}(H_0 \mid x)$ are calculated by choosing that $g \in G$, which is *most favorable* to H_1; to minimize this bias toward H_1, one should try to choose G as small as possible, consistent with *a priori* beliefs.

In this light, choosing $G = \{\text{all densities}\}$ seems certainly too extreme, allowing for severe bias toward H_1. It is astonishing that, even for this choice, $\underline{B}$ and $\underline{P}(H_0 \mid x)$ (for $\pi_0 = ½$) are still often substantially larger than the P-value (see Edwards, Lindman and Savage (1963)), indicating that even extreme bias toward H_1 in a Bayesian analysis results in less evidence against H_0 than would appear to have been obtained by the P-value.

Choosing $G = \{\text{all conjugate densities with mean } \theta_0\}$ actually works quite well, but one might fear that too many sensible g are omitted to trust the resulting lower bounds. The class in (10) strikes a nice balance between these two extremes. In situations such as Example 1, it is natural to argue (on grounds of objectivity) that g should be symmetric and nonincreasing in $|\theta - \theta_0|$. (Either of these properties could be dropped, without qualitatively changing the results.) Virtually any "objective" weight function that would be proposed is in this class, and no density in the class is clearly ridiculous. Hence, we will tend to use (10), or variants of it, in our analyses.

1.4 History and Overview

There is 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, 1985, 1986), Lindley (1957, 1961, 1965, 1977), Raiffa and Schlaiffer (1961), Edwards, Lindman and Savage (1963), Smith (1965), Zellner (1971, 1984), Dickey (1971, 1973, 1974, 1980), Lempers (1971), Leamer (1978), Smith and Spiegelhalter (1980), Zellner and Siow (1980), Diamond and Forrester (1983) and Gómez and de la Horra Navarro (1984). Many of these works specifically discuss the relationship of $P(H_0 \mid x)$ to significance levels; other papers in which such comparisons are made include Pratt (1965), DeGroot (1973), Dempster (1973), Dickey (1977), Bernardo (1980), Hill (1982), Shafer (1982) and Good (1984). Finally, the papers that find lower bounds on B and $P(H_0 \mid x)$ that are similar to those we consider include Edwards, Lindman and Savage (1963), Hildreth (1963), Good (1967, 1983, 1984), Dickey (1973, 1977), Berger (1986), Berger and Sellke (1987), Casella and Berger (1987), Delampady (1986a, 1986b, 1986c) and Delampady and Berger (1987).

Edwards, Lindman and Savage (1963) deserves special mention, being the first to approach the problem from the viewpoint of finding lower bounds on B and $P(H_0 \mid x)$ over $g \in G$. They were the first to make the extent of the conflict between P-values and posterior probabilites unambiguously clear. Berger and Sellke (1987) were the first to utilize classes G, such as (10), which we tend to prefer and, along with the adjoining paper by Casella and Berger (1987) and associated discussion, explores the extent and meaning of the conflict between P-values and posterior probabilities in the univariate normal problem and certain extensions.

The purpose of this paper is to further explore the issues raised in Section 1.2. Of primary importance is further discussion of the conflict between P-values and conditional measures of evidence. Section 3.1 briefly reviews the normal theory situation for context. One can very reasonably inquire, however, as to the extent of the conflict for nonnormal (and nonsymmetric) problems.

To alleviate concerns that the conflict is special to the normal problem, we consider in Section 3.2 the binomial problem. In the binomial problem, there is no natural definition of symmetry, and it is not clear how to develop classes G for determination of $\underline{P}$ and $\underline{B}$. We will consider a variety of different classes G and show that the answers are qualitatively similar for the G considered but all differ markedly from the P-value.

It should be emphasized that $\underline{B}$ and $\underline{P}$ are of interest in their own right. Lower bounds on the evidence against H_0 can immediately demonstrate, without having to go through a detailed Bayesian analysis, that the evidence against H_0 is weak. Indeed, it is tempting to replace use of "automatic" P-values by the far less misleading "automatic" $\underline{B}$; problems with doing so are, however, discussed in Section 5, which describes our general recommendations.

Section 2 is an important preliminary to the subject. It attempts to quantify when it is reasonable to model a hypothesis testing problem as a test of a point null hypothesis. Thus, it essentially defines the class of problems that is being addressed. As an immediate application, a resolution of Jeffreys's paradox is offered. Finally, a description of the situation from a "robust Bayesian" viewpoint is described; this may be the part of the paper of most interest to a Bayesian.

In Section 4, a number of common objections to the analysis (mainly defenses of P-values) are considered.

2. PRECISE HYPOTHESES AND POINT NULLS

It is rare, and perhaps impossible, to have a null hypothesis that can be exactly modeled as $\theta = \theta_0$. One might feel that hypotheses such as

H_0: A subject has no ESP,

or

H_0: Talking to plants has no effect on their growth,

are representable as exact (and believable) point nulls, but, even here, minor biases in the experiments designed to test hypotheses will usually prevent exact representations as points.

More common precise hypotheses, such as

H_0: Vitamin C has no effect on the common cold,

are clearly not meant to be thought of as exact point nulls; surely vitamin C has *some* effect, although perhaps a very miniscule effect. Thus, in reality, precise hypotheses are better represented as tests of, say,

$$(11) \quad H_0\colon |\theta - \theta_0| \le \varepsilon \quad \text{versus} \quad H_1\colon |\theta - \theta_0| > \varepsilon,$$

where ε is "small." It is then of interest to determine when one can approximate (11) by the point null test

$$H_0^*\colon \theta = \theta_0 \quad \text{versus} \quad H_1^*\colon \theta \ne \theta_0.$$

2.1 Classical Approximation by Point Nulls

From a classical perspective, there is no difficulty in determining when ε is small enough so that the test in (11) can be approximated by H_0^* versus H_1^*. The definition of a P-value for (11) would be

$$\alpha_\varepsilon = \sup_{\theta:|\theta-\theta_0|\le\varepsilon} P_\theta(|T(X)| \ge |t|).$$

Thus, one can seek conditions under which $\alpha_\varepsilon \cong \alpha_0$ (the P-value corresponding to H_0^*).

Example 2. Suppose one observes $\bar{X} \sim N(\theta, \sigma^2/n)$, where σ^2 is known, and that it is desired to test (11). Again defining $T(\bar{X}) = \sqrt{n}(\bar{X} - \theta_0)/\sigma$, calculation gives

$$\alpha_\varepsilon = 2 - \left[\Phi\left(|t| - \frac{\varepsilon\sqrt{n}}{\sigma}\right) + \Phi\left(|t| + \frac{\varepsilon\sqrt{n}}{\sigma}\right)\right].$$

Suppose that one is interested in (say) determining when

$$\alpha_0 \geq (0.9)\alpha_\varepsilon$$

(so that approximating the interval null by a point leads to no more than a 10% error in the P-value). The values $\varepsilon^* = \varepsilon\sqrt{n}/\sigma$ that achieve this are those values less than or equal to the entries in Table 2 (for various t). Thus, as long as ε is no more than ¼ to ⅙ of a sample standard deviation, the use of a point null will cause at most 10% error in the calculated P-value (for moderate t). See also Hodges and Lehmann (1954).

2.2 Bayesian Approximation by Point Nulls

A Bayesian, when considering a test such as (11) with ε small, has in mind a prior density, $\pi(\theta)$, which is continuous but sharply spiked near θ_0. It is convenient to define $\Omega = \{\theta: |\theta - \theta_0| \leq \varepsilon\}$, $\bar{\Omega}$ = Complement of Ω,

$$\pi_0 = \int_\Omega \pi(\theta)\, d\theta, \quad g_0(\theta) = \frac{1}{\pi_0}\pi(\theta)I_\Omega(\theta), \tag{12}$$
$$g_1(\theta) = \frac{1}{(1-\pi_0)}\pi(\theta)I_{\bar{\Omega}}(\theta).$$

Thus, π_0 is the prior probability assigned to H_0, g_0 is the conditional density of θ given that H_0 is true, and g_1 is the conditional density of θ given that H_1 is true. Typically, g_0 will be a sharply spiked density, although g_1 will be rather diffuse.

It is usually fairly easy to specify π_0, by thinking of the prior probability of the original hypothesis (e.g., H_0: Vitamin C has no (or negligible) effect). Likewise, specification of $g_1(\theta)$, the conditional density given that H_0 is false, is reasonably tractable. (Note also that specification of π_0 can be avoided through use of the Bayes factor, and we will be addressing elimination of g_1 through use of lower bounds over g_1.) It can be very difficult, however, to choose ε and to choose g_0; it is simply hard to make fine distinctions about small sets with large probability.

TABLE 2
Bounds on ε^* yielding 10% error in the P-value

t	1.645	1.96	2.576	2.807	3.29	3.89
P-value	0.10	0.05	0.01	0.005	0.001	0.0001
Bound on ε^*	0.257	0.221	0.173	0.160	0.138	0.117

Because of the difficulties in choosing ε and g_0, it is of interest to determine when one can approximate (11) by the point null test of H_0^* versus H_1^*, where the "prior" assigns mass π_0 to $\{\theta = \theta_0\}$, and gives conditional prior density $g(\theta)$ to $\{\theta \neq \theta_0\}$; we assume that

$$\begin{aligned} &\text{(i)} \quad g(\theta) \propto g_1(\theta) \quad \text{on} \quad |\theta - \theta_0| > \varepsilon; \\ &\text{(ii)} \quad \lambda = \int_\Omega g(\theta)\, d\theta \quad \text{is suitably small.} \end{aligned} \tag{13}$$

The idea here is that one assigns the same probability to H_0^* as to H_0, and specifies $g(\theta)$ by using the same intuition that would have been used to specify $g_1(\theta)$. We are hoping the answer obtained from this simple test is close to that which would have been obtained for the original formulation, (11), thus avoiding the difficult assessment of ε and g_0.

We will consider the case where we are to observe $\bar{X} \sim N(\theta, \sigma^2/n)$, where σ^2 is known. Letting $f(\bar{x}\,|\,\theta)$ denote this density, the exact Bayes factor for testing H_0 versus H_1 in (11) is

$$B = \frac{\int_\Omega f(\bar{x}\,|\,\theta)g_0(\theta)\, d\theta}{\int_{\bar{\Omega}} f(\bar{x}\,|\,\theta)g_1(\theta)\, d\theta}, \tag{14}$$

while the Bayes factor for testing H_0^* versus H_1^* is

$$\hat{B} = \frac{f(\bar{x}\,|\,\theta_0)}{m_g(\bar{x})}, \tag{15}$$

where $m_g(\bar{x}) = \int f(\bar{x}\,|\,\theta)g(\theta)\, d\theta$. The following theorem gives conditions under which $\hat{B} \cong B$. For use in this theorem, define $\varepsilon^* = \varepsilon\sqrt{n}/\sigma$ and $t = \sqrt{n}(\bar{x} - \theta_0)/\sigma$ as in Example 2, and define

$$\gamma = \frac{1}{2\varepsilon^*\phi(t)}[\Phi(t + \varepsilon^*) - \Phi(t - \varepsilon^*)] - 1, \tag{16}$$

where ϕ and Φ are the standard normal density and cdf, respectively.

THEOREM 1. *Suppose that $\pi(\theta)$ and $g(\theta)$ are unimodal and symmetric about θ_0, and that $|t| \geq 1$, $\varepsilon^* < |t| - 1$, and $\hat{B} \leq (1 + \gamma)^{-1}$. Then*

$$B = \hat{B}(1 + \rho),$$

where

$$\begin{aligned} -\lambda &\leq \frac{\lambda(\hat{B} - 1)}{1 - \lambda\hat{B}} \leq \rho \\ &\leq \frac{\gamma + \lambda(1 + \gamma)(\hat{B} - 1)}{1 - \lambda\hat{B}(1 + \gamma)} \leq \gamma. \end{aligned} \tag{17}$$

PROOF. See the Appendix.

If both λ and γ are small, then the error in approximating B by $\hat{B}$ will be small. Table 3 gives, for various common t and associated P-values, the maximum ε^* for which $\gamma \leq .1$. Note also that a crude bound on λ is

$$\lambda = \int_\Omega g(\theta)\, d\theta \leq 2\varepsilon g(\theta_0) = \frac{2\varepsilon^*\sigma}{\sqrt{n}}\, g(\theta_0).$$

Thus, for moderately large n, this will also be quite small.

Example 3. Suppose g is taken to be $N(\theta_0, \tau^2)$, with the user asked to specify τ^2. Then

$$\lambda \leq \sqrt{\frac{2}{\pi n}} \cdot \frac{\varepsilon^*\sigma}{\tau}.$$

This is less than .1 if $n \geq 64(\varepsilon^*\sigma/\tau)^2$. Thus, if $t = 1.96$ is observed, $\sigma = \tau$, $\varepsilon^* = 1/4$ and $n \geq 4$, the error in approximating B by $\hat{B}$ (note that $\hat{B}$ is given by (15)) is no more than 10%.

Recalling that $\varepsilon = \varepsilon^*\sigma/\sqrt{n}$ is the half-width of Ω, these numbers suggest that the point null approximation to H_0 will be reasonable so long as Ω is one-half a sample standard deviation $(\sigma/\sqrt{n})$ in width or smaller. (This is substantially stronger than the related result in Dickey (1976), which can be used to verify the accuracy of the point null approximation, providing Ω is no more than 1/10 a sample standard deviation in width.) Note also the essential agreement of this result with the related result for a frequentist approximation that was discussed in Section 2.1.

2.3 Jeffreys's Paradox

"Jeffreys's paradox" or "Lindley's paradox" (cf. Jeffreys, 1961, and Lindley, 1957) concerns the fact that, in testing H_0^* versus H_1^* with a fixed π_0 and g, data $\bar{x}_n$, which yields (for each sample size n) a fixed P-value α, will result in

$$P(H_0^* \mid \bar{x}_n) \to 1 \quad \text{as} \quad n \to \infty,$$

no matter how small α is. Thus, when n is very large, a Bayesian test will frequently yield a posterior probability of H_0^* near one, even when the P-value is very small. This has been much discussed (cf. Bernardo, 1980, and Shafer, 1982), but is of questionable relevance because, as $n \to \infty$, $\varepsilon = \varepsilon^*\sigma/\sqrt{n} \to 0$. Thus, a precise hypothesis H_0: $|\theta - \theta_0| \leq \varepsilon_0$ will fail to be approximable by H_0^*: $\theta = \theta_0$ when n gets very large. In fact, for H_0: $|\theta - \theta_0| \leq \varepsilon_0$, it can be shown for fixed $\pi(\theta)$ that $\bar{x}_n$, which yields (for each sample size n) a fixed P-value α, will often result in

$$P(H_0 \mid \bar{x}_n) \to \alpha \quad \text{as} \quad n \to \infty,$$

in marked contrast to Jeffreys's paradox. The reason for this is that, as the likelihood function becomes concentrated at the edge of the interval null (where it must be located for the P-value to be α), the interval null becomes, effectively, a half line; and, for one-sided testing, P-values and posterior probabilities are often similar (cf. Pratt, 1965, and Casella and Berger, 1987).

TABLE 3
ε^* yielding 10% upper error

t	1.645	1.96	2.576	2.807	3.29	3.89
P-value	0.10	0.05	0.01	0.005	0.001	0.0001
Bound on ε^*	0.62	0.47	0.33	0.30	0.25	0.20

2.4 Robust Bayesian Interpretation

Imagine being faced with a test of a precise hypothesis. A satisfactory Bayesian *output* for many purposes (see also Sections 4 and 5) would be B, the Bayes factor against H_0, together with $\pi(\theta \mid x, \bar{\Omega})$, the posterior density conditional on H_0 being false. In the vitamin C example, B would measure how strongly the data support H_0, while $\pi(\theta \mid x, \bar{\Omega})$ would communicate the location of θ should H_0 be wrong. From these, most decisions or conclusions could be made.

One has a "robust Bayesian" conclusion if the answers are not highly dependent on uncertain inputs. For testing a precise hypothesis, the actual "width" of H_0 and form of the prior in this region are typically very uncertain inputs. Section 2.2 indicates that these specifications are, however, avoidable if one can make the crude judgment that the width is less than half a sample standard deviation. Also, by use of B one can even avoid (at this stage of the analysis) specification of π_0, the prior probability of H_0. The only necessary prior specification is thus g_1, the prior density assuming H_0 is false. In many situations this can be fairly easily specified. And if not, one can often present answers for a variety of g_1 or calculate bounds on the answers over plausible classes of g_1; examples will be given later. (Note that, for large n, the condition that the width of the null interval be less than half the sample standard deviation will be violated. Robustness can then be lacking (see Rubin, 1971).

3. LOWER BOUNDS ON BAYES FACTORS AND POSTERIOR PROBABILITIES

This section will explore the conflict between P-values and conditional measures that was discussed in "Opinion 1" of Section 1.2. Section 3.1 reviews the normal situation, whereas Section 3.2 considers the much more difficult binomial case. Section 3.2 is included to demonstrate the generality of the conflict being discussed, but can be skipped by readers interested only in the broad picture.

3.1 Lower Bounds for Symmetric Unimodal G

Often, as when dealing with the multivariate normal distribution with covariance matrix a multiple of the identity, the problem is symmetric about θ_0 (more precisely, is orthogonally invariant). A very natural class of weight functions g to consider is then the class given by (10), of unimodal, symmetric about θ_0 densities. Either symmetry or unimodality (or sometimes both) could be dropped (see Berger and Sellke, 1987), but recall that we have to perform a delicate balancing act: we want to allow all reasonable ("objective"?) g into the class G, but disallow unreasonable g that will excessively bias the lower bounds $\underline{B}$ and $\underline{P}$ toward H_1. The class G in (10) is a reasonable balance.

This class is also relatively easy to work with, because of the following standard result. We assume, in this section, that the parameter space is R^p.

THEOREM 2. *The supremum over G in* (10) *of*

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

is attained at a uniform distribution on a sphere of some radius k_0. In other words,

$$\sup_{g \in G} m_g(x) = \sup_k \frac{1}{V(k)} \int_{|\theta - \theta_0| \le k} f(x \mid \theta)\, d\theta,$$

where $V(k)$ is the volume of a sphere of radius k.

It follows from Theorem 2 that $\underline{B}$ and $\underline{P}(H_0 \mid x)$ (see (8) and (9)) can be calculated by simple one-dimensional maximizations. This is a delightful simplification; the original space G is very complex. This result was first utilized in Berger and Sellke (1987). The following application to the multivariate normal distribution is from Delampady (1986a).

Example 4. *Testing a p-Variate Normal Mean.* Suppose $\mathbf{X} \sim N_p(\boldsymbol{\theta}, I)$, where $\mathbf{X} = (X_1, X_2, \cdots, X_p)$ and $\boldsymbol{\theta} = (\theta_1, \theta_2, \cdots, \theta_p)$. It is desired to test

$$H_0\colon \boldsymbol{\theta} = \boldsymbol{\theta}^0 \quad \text{against} \quad H_1\colon \boldsymbol{\theta} \neq \boldsymbol{\theta}^0,$$

where $\boldsymbol{\theta}^0 = (\theta_1^0, \theta_2^0, \cdots, \theta_p^0)$ is a specified vector. The classical significance test statistic is

$$T(\mathbf{X}) = |\mathbf{X} - \boldsymbol{\theta}^0|^2,$$

which has a χ_p^2 distribution under H_0. Therefore, the P-value of the data $\mathbf{x}$ is

$$\alpha = P(\chi_p^2 \ge T(\mathbf{x})).$$

Using Theorem 2, the lower bound on the Bayes factor over the class G is

$$\underline{B} = \frac{\exp(-\tfrac{1}{2} |\mathbf{x} - \boldsymbol{\theta}^0|^2)}{\sup_k (1/V(k)) \int_{|\boldsymbol{\theta} - \boldsymbol{\theta}^0| \le k} \exp(-\tfrac{1}{2} |\boldsymbol{\theta} - \mathbf{x}|^2)\, d\boldsymbol{\theta}}. \tag{18}$$

Numerical values in Table 4 were computed by using the above result for different dimensions and different P-values. Here α is the P-value, $\underline{P}$ is the lower bound on the posterior probability of H_0 for $\pi_0 = \tfrac{1}{2}$ and $\underline{B}$ is the lower bound on the Bayes factor.

Note the dramatic discrepancy between α and the lower bounds. When $p = 1$ and $\alpha = .05$, for instance, $\underline{B} = .4092$; thus, the weighted likelihood of H_1 is *at most* 2½ times that of H_0. A likelihood ratio of 2½ is not particularly strong evidence, particularly when it is a bound. However, it is customary in practice to view $\alpha = .05$ as strong evidence against H_0. A P-value of $\alpha = .01$, often considered very strong evidence against H_0, corresponds to $\underline{B} = .1227$, indicating that H_1 is *at most* 8 times as likely as H_0. The message is simple: common interpretation of P-values, in terms of evidence against precise hypotheses, are faulty; Bayes factors or posterior probabilities are typically at least an order of magnitude larger. Note that a natural generalization of the symmetry assumption is to an assumption of invariance, when a testing problem is suitably invariant under a group of transformations. See Delampady (1986c) for such a generalization and examples.

3.2 A Nonsymmetric Situation: Binomial Testing

3.2.1 Introduction

We have seen that the class of symmetric unimodal densities is sensible and easy to work with. If the testing problem is not naturally symmetric in θ, however, it is not possible to use such a class. Nevertheless, it is possible to find reasonable lower bounds on B and $P(H_0 \mid x)$. A natural way to proceed is to consider a transformed version of the problem in which symmetry is plausible and proceed as before. This is

TABLE 4
Lower bounds for spherically symmetric unimodal densities

Dimension	$\alpha = .001$		$\alpha = .01$		$\alpha = .05$		$\alpha = .10$	
	$\underline{P}$	$\underline{B}$	$\underline{P}$	$\underline{B}$	$\underline{P}$	$\underline{B}$	$\underline{P}$	$\underline{B}$
1	.0179	.0182	.1093	.1227	.2904	.4092	.3916	.6437
2	.0141	.0143	.0891	.0978	.2582	.3481	.3630	.5699
3	.0118	.0119	.0827	.0902	.2458	.3259	.3505	.5396
4	.0113	.0114	.0783	.0850	.2390	.3141	.3435	.5232
5	.0098	.0099	.0761	.0824	.2350	.3072	.3391	.5131
6	.0096	.0097	.0747	.0807	.2321	.3023	.3359	.5058
7	.0095	.0096	.0738	.0797	.2302	.2990	.3335	.5004
8	.0094	.0095	.0731	.0789	.2286	.2963	.3318	.4966
9	.0093	.0094	.0725	.0782	.2273	.2942	.3303	.4932
10	.0093	.0094	.0721	.0777	.2264	.2927	.3292	.4908
15	.0092	.0093	.0699	.0752	.2233	.2875	.3255	.4826
20	.0092	.0093	.0692	.0743	.2214	.2844	.3235	.4782
30	.0091	.0092	.0685	.0735	.2193	.2809	.3209	.4725
40	.0091	.0092	.0680	.0730	.2183	.2793	.3200	.4706
∞	.0090	.0091	.0678	.0727	.2180	.2788	.3189	.4682

illustrated in Subsection 3.2.2. A second possibility is to consider the class of all prior densities g, which have median and mode at θ_0; this is illustrated in Subsection 3.2.3. A third possibility is to consider the class of all conjugate densities having mean θ_0; this is illustrated in Subsection 3.2.4. (Choosing G to be all densities was considered in Edwards, Lindman and Savage (1963), but as mentioned in Section 1, we feel that the resulting lower bounds are too seriously biased in favor of H_1 to be very useful.)

The problem that will be considered in this section is that of testing a binomial parameter. Thus, X will have a binomial distribution with parameters n and θ. The problem of interest is to test H_0: $\theta = \theta_0$ against H_1: $\theta \neq \theta_0$, where $0 < \theta_0 < 1$ is a specified quantity. It is clear there is no natural symmetry in this problem unless $\theta_0 = ½$. This even makes difficult the definition of a P-value. We will use the "intrinsic significance level" obtained by choosing $T(X) = 1/f(X \mid \theta_0)$ in (1). This leads to defining the P-value of an observation x as

$$\alpha = P_{\theta=\theta_0}(\{y : f(y \mid \theta_0) \leq f(x \mid \theta_0)\}).$$

3.2.2 A Natural Transformed Symmetric Class

When $\theta_0 = ½$ it is easy to define a notion of prior symmetry; simply choose all densities which are symmetric about the point ½. It is not clear how to define symmetry otherwise, however. A natural way to obtain a notion of symmetry is to consider symmetry in a suitable transformation of the parameter θ. One such transformation is suggested by the normal approximation to the binomial likelihood function. Thus, if H_0: $\theta = \theta_0$ is to be tested, it may be reasonable to specify symmetry in the variable

$$u(\theta) = \frac{\theta - \theta_0}{\sqrt{\theta(1-\theta)}}; \tag{19}$$

note that this has a range of $(-\infty, \infty)$, unlike $\theta \in (0, 1)$. Let h be a nonnegative, unimodal symmetric function about the origin (the symmetric unimodal density of u). Transforming back to θ, yields the density

$$g(\theta) = h(u(\theta)) \frac{du(\theta)}{d\theta}. \tag{20}$$

(Because $u(\theta)$ is an increasing function of θ, the Jacobian can be written $(du(\theta)/d\theta)$ instead of $|(du(\theta)/d\theta)|$.) Note that θ_0 is the median of the density g, because

$$\int_0^{\theta_0} g(\theta)\, d\theta = \int_{-\infty}^{0} h(u)\, du = \frac{1}{2}.$$

Let G_{US} be the set of all densities of the form given by (20). Using Theorem 2, a simple expression for the lower bound on the posterior probability of H_0: $\theta = \theta_0$ can be obtained as follows.

THEOREM 3.

$$\underline{P}_{US} = \inf_{g \in G_{US}} P(H_0 \mid x) = \left(1 + \frac{(1-\pi_0)}{\pi_0} \frac{\sup_k (1/2k) \int_{-k}^{k} l(u)\, du}{\binom{n}{x}(\theta_0)^x (1-\theta_0)^{n-x}}\right)^{-1}, \tag{21}$$

where $l(u) = \binom{n}{x}\theta(u)^x(1 - \theta(u))^{n-x}$ *and* $\theta(u)$ *is the inverse function of* $u(\theta)$, *given by*

$$\theta(u) = \frac{\theta_0 + (u^2/2) + u\sqrt{(u^2/4) + \theta_0(1-\theta_0)}}{1+u^2}.$$

PROOF. See the Appendix.

REMARK. The supremum over k is actually attained for some k, since $(1/2k) \int_{-k}^{k} l(u)\, du$ converges to $l(0) > 0$ as $k \to 0$ and vanishes at ∞.

It is a simple numerical computation to obtain $\underline{P}_{US}$ by using Theorem 3. For $\pi_0 = ½$ and selected values of θ_0, n and x, these lower bounds are tabulated along with the corresponding P-values in Table 5.

The differences between α and $\underline{P}_{US}$ are of the same magnitude here as in the normal situation of Section 3. For $\alpha = .01, .05, .10$, the values of $\underline{P}_{US}$ from Table 4 (in one dimension) were .109, .290 and .392, respectively, all similar to the corresponding $\underline{P}_{US}$ in Table 5. Thus, the discreteness of X and the lack of natural symmetry of the parameter space do not markedly affect the large discrepancy between α and $P(H_0 \mid x)$. Of course, the notion of symmetry that was used here could be questioned, but other notions of symmetry were tried in Delampady (1986a) and gave, if anything, larger discrepancies between α and $\underline{P}$.

TABLE 5
Lower bounds for transformed symmetric densities

α	n	x	p^0	Maximizing k	$\underline{P}_{US}$
0.0090	50	11	.40	4.4851	0.0794
0.0100	20	9	.20	3.2502	0.1201
0.0101	20	14	.40	4.3694	0.0770
0.0106	10	7	.30	4.3190	0.0951
0.0118	20	4	.50	5.1514	0.0642
0.0128	10	4	.10	2.8777	0.1244
0.0493	50	16	.20	2.5784	0.2786
0.0507	25	3	.30	4.3252	0.2210
0.0534	35	12	.20	2.5329	0.2867
0.0541	40	10	.40	3.2164	0.2741
0.0920	50	9	.10	2.0479	0.3687
0.0960	15	6	.20	2.2632	0.3496
0.0987	30	20	.50	2.8149	0.3257
0.1011	10	7	.40	3.1644	0.2969
0.1021	45	19	.30	2.3423	0.3624
0.1094	10	2	.50	3.7602	0.2780

3.2.3 Classes with Specified Mode and Median

In a number of situations, the class of densities g, which have their median and mode at θ_0, is quite reasonable. In fact, the class of symmetric unimodal densities about the point θ_0 is a subclass of this set. Study of this class of densities will help us judge if the definition of symmetry suggested in the previous section is reasonable.

Let G_1 be the set of all densities on [0, 1] with median at θ_0. Then,

$$G_1 = \left\{ g: \int_0^{\theta_0} g(\theta)\, d\theta = \frac{1}{2} \right\}.$$

Let G_0 be the set of all unimodal densities on the interval [0, 1] with median and mode at θ_0. That is,

$$G_0 = \left\{ g: \int_0^{\theta_0} g(\theta)\, d\theta = \frac{1}{2}; \text{ and } g \text{ is nondecreasing on } [0, \theta_0], \text{ and nonincreasing on } [\theta_0, 1] \right\}.$$

Then $G_0 \subset G_1$. Of these two classes, G_0 may be the most reasonable; G_1 allowing unrealistic concentration at particular points. Lower bounds on the posterior probability of the null hypothesis, H_0: $\theta = \theta_0$, in the binomial case are once again considered.

THEOREM 4.

$$\sup_{g \in G_1} \int_0^1 \binom{n}{x} \theta^x (1-\theta)^{n-x} g(\theta)\, d\theta = \frac{1}{2}\left[\binom{n}{x} \theta_0^x (1-\theta_0)^{n-x} + \binom{n}{x} \left(\frac{x}{n}\right)^x \left(1 - \frac{x}{n}\right)^{n-x} \right].$$

PROOF. See the Appendix.

THEOREM 5.

$$\sup_{g \in G_0} \int_0^1 \binom{n}{x} \theta^x (1-\theta)^{n-x} g(\theta)\, d\theta = \frac{1}{2}\left[\binom{n}{x} \theta_0^x (1-\theta_0)^{n-x} + \sup_k \frac{1}{k} \int_{a(k)}^{b(k)} \binom{n}{x} \theta^x (1-\theta)^{n-x}\, d\theta \right],$$

where

$$\{a(k), b(k)\} = \begin{cases} \{\theta_0 - k, \theta_0\} & \text{if } (x/n) \le \theta_0, \\ \{\theta_0, \theta_0 + k\} & \text{if } (x/n) > \theta_0. \end{cases}$$

PROOF. See the Appendix.

Using the expressions given in Theorem 4 and 5, the lower bounds on the posterior probability of H_0 may be calculated. These lower bounds denoted, respectively, as $\underline{P}_{\text{SMe}}$, $\underline{P}_{\text{SMeMo}}$, were computed for some selected values of n, x, θ_0 when $\pi_0 = 1/2$, and are tabulated along with the corresponding P-values in Table 6. We defer discussion of the table until Section 3.2.5.

3.2.4 Conjugate Class with Mean θ_0

The class of conjugate g with mean $E^g(\theta) = \theta_0$ are studied here, and the relevant lower bounds are obtained. For the binomial distribution, the beta distributions form a family of conjugate distributions. The density of *beta*(a, b) with parameters $a > 0$, $b > 0$, is given by

$$\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1}(1-\theta)^{b-1}, \qquad 0 < \theta < 1,$$

and the mean is $(a/(a+b))$. Therefore, the class of conjugate priors to consider is that which consists of all *beta*(a, b) distributions such that $(a/(a+b)) = \theta_0$, or equivalently such that $a = c\theta_0$, $b = c(1-\theta_0)$ for some $c > 0$. Let G_C denote the class of all such densities.

THEOREM 6.

$$\underline{P}_C = \inf_{g \in G_C} \underline{P}(H_0 \mid x) = \left[1 + \frac{(1-\pi_0)}{\pi_0} \times \sup_{c>0} \frac{\Gamma(c)\Gamma(x+\theta_0)\Gamma(n-x+c(1-\theta_0))}{\Gamma(c\theta_0)\Gamma(c(1-\theta_0))\Gamma(n+c)(\theta_0)^x(1-\theta_0)^{n-x}} \right]^{-1}.$$

PROOF. See the Appendix.

TABLE 6
Lower bounds for densities with specified median, mode

α	n	x	p^0	$\underline{P}_{\text{SMe}}$	$\underline{P}_{\text{SMeMo}}$
0.0093	25	11	.20	0.0469	0.1080
0.0101	20	14	.40	0.0471	0.0807
0.0106	10	7	.30	0.0613	0.1559
0.0107	30	12	.20	0.0767	0.1330
0.0118	20	4	.50	0.0398	0.0730
0.0479	40	3	.20	0.1394	0.2072
0.0493	50	16	.20	0.1933	0.2444
0.0507	25	3	.30	0.1559	0.2224
0.0534	35	12	.20	0.2003	0.2513
0.0541	40	10	.40	0.1933	0.2526
0.0980	25	5	.10	0.3314	0.3789
0.0987	30	20	.50	0.2362	0.2909
0.1000	35	15	.30	0.3013	0.3442
0.1011	10	7	.40	0.2154	0.2728
0.1094	10	2	.50	0.2026	0.2653

The lower bounds on the posterior probability of H_0 given x can easily be computed numerically, for specified values of θ_0, n, x and π_0. For $\pi_0 = ½$ and some selected values of n, x, these lower bounds, denoted $\underline{P}_C$ are tabulated in Table 7 along with the corresponding P-values. Again, we defer discussion until Section 3.2.5.

Good (1950, 1958, 1967) and Edwards, Lindman and Savage (1963) contain results related to the lower bounds on H_0 over the class of conjugate densities that are presented here.

3.2.5 Comparisons and Conclusions

The major results of this section are now summarized and compared. Table 8 very briefly summarizes all of the earlier tables. Here "α" is the P-value, "range" is the range of the lower bounds on the posterior probability of H_0 for the binomial distribution as the corresponding P-value varies around "α"; C stands for the class of conjugate priors, US for the class of priors that are unimodal symmetric in the transformed parameter, SMeMo for the class of priors with specified median and mode and finally, SMe stands for the class of priors with specified median.

Clearly the conjugate priors bounds tend to be the largest, followed by the US and SMeMo bounds that are similar, with the SMe bounds being the smallest. This ordering was to be expected, because the corresponding classes of densities are roughly inversely related to size. Our own preference is for the US bounds, because of our feeling that the US class of densities is an excellent compromise between being too big (and hence too biased against H_0) and too small.

One observation of interest is that the three classes that attempt to spread mass on both sides of θ_0, namely C, US and SMeMo, all yield moderately similar lower bounds. The bounds for SMe are substantially smaller only because this class allows the mass from one side to be concentrated in the middle, a not terribly reasonable eventuality.

Note also, that the bounds obtained for the normal distribution in Table 4 are similar to the first three binomial lower bounds. This gives considerable support to the notion that the discrepancy between P-values and posterior probabilities in testing precise hypotheses is a general phenomenon.

TABLE 7
Lower bounds for conjugate densities

α	n	x	p^0	$\underline{P}_C$
0.0090	50	11	.40	0.0981
0.0100	20	9	.20	0.1771
0.0101	20	14	.40	0.1064
0.0118	20	4	.50	0.0858
0.0120	45	10	.10	0.2211
0.0493	50	16	.20	0.3313
0.0505	15	1	.30	0.1956
0.0507	25	3	.30	0.2414
0.0541	40	10	.40	0.3016
0.0556	15	4	.10	0.4223
0.0960	15	6	.20	0.4123
0.0980	25	5	.10	0.4779
0.0987	30	20	.50	0.3565
0.1000	35	15	.30	0.4328
0.1011	10	7	.40	0.3458
0.1094	10	2	.50	0.3163

TABLE 8
Summary of lower bounds

α	Normal bounds	Range of binomial lower bounds			
		C	US	SMeMo	SMe
.01	.11	.09–.22	.06–.13	.07–.16	.04–.07
.05	.29	.20–.42	.22–.29	.21–.25	.14–.20
.10	.39	.32–.48	.28–.37	.27–.38	.20–.33

4. COMMON REJOINDERS

Many studies, such as those in Section 3, have been performed for a wide variety of testing situations involving a precise null hypothesis. The overwhelming conclusion is that P-values are typically at least an order of magnitude smaller than Bayes factors or posterior probabilities for H_0. This would indicate that say, claiming that a P-value of .05 is significant evidence against a precise hypothesis is sheer folly; the actual Bayes factor may well be near 1, and the posterior probability of H_0 near ½. Needless to say, supporters of P-values will marshall a number of arguments against such dismissal of P-values. In this section we consider, and reply to, a number of such rejoinders.

4.1 Rejoinder 1: Point Nulls Are Unreasonable

This argument is basically that it is rare or impossible to encounter hypotheses that are representable as exact point nulls. The argument continues with voicing of the suspicion that the conflict between P-values and $P(H_0 \mid x)$ is due to the assignment of the positive mass π_0 to a single point.

Section 2 effectively rebuts this argument. It shows that point nulls are often reasonable, as an approximation to fuzzy precise nulls, and that the mass, π_0, assigned to $\theta = \theta_0$ is simply the mass corresponding to the precise hypothesis H_0: $|\theta - \theta_0| \leq \varepsilon$. For true precise hypotheses, such as H_0: "vitamin C has negligible effect on the common cold," the assignment of a positive mass to the hypothesis is perfectly natural.

Also, the above argument ignores the role of B. The Bayes factor does not depend on π_0, and can be thought of as the evidence (in a likelihood ratio sense)

provided by the data against H_0. We have seen that B and $P(H_0 \mid x)$ convey essentially the same message concerning P-values.

Sometimes it is argued (cf. Casella and Berger, 1987) that the important testing problems are one-sided tests of the form H_0: $\theta \le \theta_0$ versus H_1: $\theta > \theta_0$, where there is no particularly strong belief that θ is near θ_0. For a reply to this point, see the rejoinder by Berger and Sellke (1987), where it is argued that the only problems which probably should be formulated as testing problems are those involving precise (i.e., small interval) hypotheses. (One-sided testing situations are typically better handled as decision problems.) Of course, every statistician must judge for himself or herself how often precise hypotheses actually occur in practice. At the very least, however, we would argue that all types of tests should be able to be properly analyzed by statistics.

Note also that it is not the two-sided feature of our testing formulation that is the cause of the discrepancy between the P-value and the posterior probability. Testing H_0: $\theta = \theta_0$ versus H_1: $\theta > \theta_0$ (or H_0: $|\theta - \theta_0| \le \varepsilon$ versus H_1: $\theta > \theta_0 + \varepsilon$) yields Bayes factors and posterior probabilities with qualitatively the same behavior as the two-sided case. The key feature is that of H_0 being precise, as opposed to diffuse in the Casella-Berger (1987) sense.

4.2 Rejoinder 2: The P-Value Is Just a Data Summary, Which We Can Learn To Properly Calibrate

Typically, the P-value is a monotonic function of the actual evidence against H_0 (either B or $P(H_0 \mid x)$, say), and one can argue that, through experience, one can learn how to interpret P-values. There are several obstacles to such "calibration" of P-values, however, including:

(i) It is dependent on sample size—see Table 1 for an illustration.

(ii) The interpretation of P-values can depend strongly on the model $f(x \mid \theta)$—see Berger and Sellke (1987) for an illustration.

(iii) The interpretation depends strongly on the stopping rule used—see Berger and Berry (1987) for illustration.

(iv) The interpretation depends strongly on the type of null hypothesis being tested. In particular, the degree to which it is precise, as opposed to diffuse, has a very large effect, as the following generalization of Example 2 from Section 2.1 demonstrates. (See Good (1986) for a related analysis).

Example 2 (continued). Suppose $\bar{X} \sim N(\theta, \sigma^2/n)$, σ^2 known, and that it is desired to test

$$H_0: |\theta - \theta_0| \le \varepsilon \quad \text{against} \quad H_1: |\theta - \theta_0| > \varepsilon.$$

Let $T = \sqrt{n}(\bar{X} - \theta_0)/\sigma$ and t_α be the critical value (which depends on ε) such that

$$\alpha = P_{\theta_0+\varepsilon}(|T(X)| \ge t_\alpha).$$

Were we to observe $T = t_\alpha$, we would report α as the P-value. To compare this with the posterior probability when t_α is observed, consider priors π for θ that (i) are symmetric about θ_0; (ii) are nonincreasing in $|\theta - \theta_0|$; and (iii) give prior probability .5 to H_0. Denote this class by US. In Delampady (1986b) it is shown that

$$\underline{P}(H_0 \mid x) \equiv \inf_{\pi \in \mathrm{US}} P^\pi(H_0 \mid x)$$

$$= \left(1 + \sup_r \frac{\Phi(r - t_\alpha) - \Phi(-r - t_\alpha) - 2b\phi(t_\alpha)}{2(r - 2\varepsilon^* + b)\phi(t_\alpha)}\right)^{-1},$$

where $b = [\Phi(\varepsilon^* - t_\alpha) - \Phi(-\varepsilon^* - t_\alpha)]/2\phi(t_\alpha)$, $\varepsilon^* = \varepsilon\sqrt{n}/\sigma$ and ϕ and Φ are the standard normal density and cdf, respectively. Figure 1 presents $\underline{P}(H_0 \mid x)$ as a function of ε^*, when the P-value is fixed at $\alpha = .05$ (i.e., for each ε^*, x is assumed to be such that $T(x) = t_{.05}$). Here "LENGTH" stands for the standardized length ε^* of the half interval, and "BOUND" denotes $\underline{P}(H_0 \mid x)$.

When ε^* (and hence ε) is zero, the test is that of a point null, and $\underline{P}(H_0 \mid x)$ is the lower bound discussed in Section 3; note that it is much larger than the P-value, $\alpha = .05$. On the other hand, as $\varepsilon^* \to \infty$ (i.e., H_0 becomes more diffuse) $\underline{P}(H_0 \mid x) \to \alpha$. (This is essentially the result of Casella and Berger, 1987.)

The above example demonstrates that the interpretation of a P-value, as evidence against H_0, depends crucially on the nature of H_0. But if the interpretation depends on H_0, the sample size, the density and the

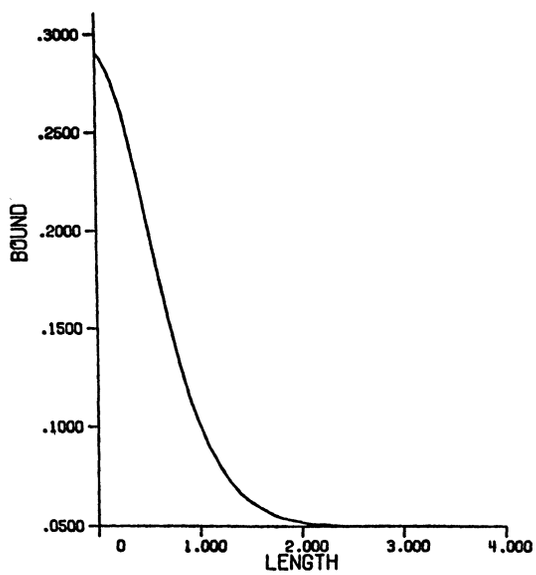

FIG. 1. $\underline{P}(H_0 \mid x)$ *for interval nulls.*

stopping rule, all in crucial ways, it becomes ridiculous to argue that we can intuitively learn to properly calibrate P-values.

4.3 Rejoinder 3: Just Use Confidence Intervals

This was Opinion 3 discussed in Section 1.2. "Just determine a confidence region (or Bayesian credible region, perhaps with an objective prior)," so the argument goes, "and draw conclusions directly from the region." The chief advantage of a confidence or credible region, of course, is that it indicates the magnitude of the discrepancy of θ from θ_0.

This argument is not unreasonable when H_0 is diffuse, but is wrong when H_0 is precise. Confidence regions and "objective" Bayesian credible regions generally correspond to very diffuse prior distributions, and are simply inappropriate if there is a special value θ_0. Only measures such as B or $P(H_0 \mid x)$ can indicate the strength of evidence against a particular value specified by H_0: $\theta = \theta_0$. To put this another way, the likelihood of a special point θ_0, which is outside, say, a 95% confidence set C, is often not too much smaller than the "average" likelihood of θ in C, and there is then no strong justification for rejecting θ_0.

For hypothesis testing problems with a special point θ_0 (or special small interval $(\theta_0 - \varepsilon, \theta_0 + \varepsilon)$), we would urge reporting *both* the Bayes factor, B, against θ_0 and a confidence or credible region, C. The Bayes factor communicates the evidence in the data against θ_0, and C indicates the magnitude of the possible discrepancy.

4.4 Rejoinder 4: There May Be No Alternatives

Suppose the hypothesis of interest is that $X \sim F_0$, but that no alternatives to F_0 are specified. It might still be possible to specify a statistic $T(X)$ to measure discrepancy of the data with F_0, and one could then calculate a P-value against F_0. Without explicit alternatives, however, no Bayes factor or posterior probability could be calculated. Thus, the argument goes, one has no recourse but to use the P-value.

A number of Bayesian responses to this argument have been raised (cf. Berger and Wolpert, 1984); here we concentrate on responding in terms of the discussion in this paper. If, indeed, it is the case that P-values for precise hypotheses essentially *always* drastically overstate the actual evidence against H_0 when the alternatives are known, how can one argue that no problem exists when the alternatives are not known? To the contrary, what we have learned about testing precise hypotheses when we have alternatives, should serve as overwhelming evidence that a small P-value against a precise hypothesis simply may not indicate strong cause to doubt the hypothesis.

Often there are, in fact, alternatives lurking in the background, which can be used to calculate Bayes factors or posterior probabilities for H_0. Consider, for instance, the following example (from Delampady and Berger, 1987), concerning the χ^2 test of fit.

Example 5. *χ^2 Test of Fit.* Consider a statistical experiment in which N independent and identically distributed random quantities $X_1, X_2, \cdots, X_N$ are observed from a distribution F. The problem is to test the hypothesis

$$H_0\colon F = F_0 \quad \text{versus} \quad H_1\colon F \neq F_0,$$

where F_0 is a specified distribution. The standard test procedure for this problem is the χ^2 test of fit.

χ^2 Test Procedures. First, a partition $\{a_i\}_{i=0}^m$ of the real line is considered. Then the frequencies of the N observations in this partition are found. Let $\mathbf{z} = (z_1, \cdots, z_m)'$ denote these frequencies; thus, z_i = the number of X_i's in $(a_{i-1}, a_i]$. Let

$$\begin{aligned} p_i &= F(a_i) - F(a_{i-1}) \\ &= P_F(a_{i-1} < X \le a_i), \\ p_i^0 &= F_0(a_i) - F_0(a_{i-1}) \\ &= P_{F_0}(a_{i-1} < X \le a_i), \end{aligned}$$

and

$$\mathbf{p} = (p_1, \cdots, p_m)',$$
$$\mathbf{p}^0 = (p_1^0, \cdots, p_m^0)'.$$

Then the χ^2 test procedure is to calculate the test statistic,

$$t = \sum_{i=1}^m \frac{(z_i - Np_i^0)^2}{Np_i^0},$$

and compute the P-value assuming a χ^2_{m-1} distribution for T, as

$$\alpha = P(T \ge t).$$

But reducing the observations to the vector $\mathbf{z}$ of frequencies really means that we are testing

$$H_0\colon \mathbf{p} = \mathbf{p}^0 \quad \text{versus} \quad H_1\colon \mathbf{p} \neq \mathbf{p}^0,$$

where $\mathbf{z}$ has a *multinomial* $(N, \mathbf{p})$ distribution. Thus, there really are implied alternatives, and one can calculate B or $P(H_0 \mid \mathbf{z})$ for this precise null testing problem.

It is shown in Delampady and Berger (1987) that lower bounds on B and $P(H_0 \mid x)$, analogous to those discussed in this paper, can be found for this multinomial testing problem. For all "objective" classes of priors considered, these lower bounds are once again an order of magnitude larger than the P-value.

Of course, not all nonparametric tests can so easily be reduced to parametric tests amenable to Bayesian analysis. The example thus mainly serves to reinforce skepticism of the argument that P-values are okay if no alternatives have been specified.

4.5 Rejoinder 5: P-Values Have a Valid Frequentist Interpretation

This rejoinder is simply not true. P-values are *not* a repetitive error rate, at least in any real sense. A Neyman-Pearson error probability, α, has the actual frequentist interpretation that a long series of α level tests will reject no more than $100\alpha\%$ of true H_0, but the data-dependent P-values have no such interpretation. P-values do not even fit easily into any of the conditional frequentist paradigms.

Furthermore, any type of a repetitive "error rate" can be accused of addressing the wrong question. Most hypothesis tests are set up to attempt to answer the question: "In light of the data, do I have reason to think that H_0 is false." To see, in a long run repetitive scenario, that the P-value can be a misleading answer to this question, consider the following example from Berger and Sellke (1987).

Example 6. Jeffreys (1980) states, concerning the answers obtained using his prior for testing a point null,

> "These are not far from the rough rule known to astronomers, i.e., that differences up to twice 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 can be shown that this limiting proportion will be *at least* .22, and is usually much larger.

Note the important distinction between the "experiment" here and the typical frequentist "experiment" used to evaluate the performance of, say, the $\alpha = .05$ 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 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 $\alpha = .05$ says nothing about this.

4.6 Rejoinder 6: The P-Value Is a Measure of Surprise

It is true that, were H_0 true, we would not expect to observe data that yield a small P-value. The issue, however, is whether or not the actual magnitude of the P-value can be given a quantitative interpretation in terms of the evidence against H_0. We feel that no reasonable such interpretation can be given; at best, the P-value can only serve as a crude indicator that something surprising is going on, perhaps acting as spur to carry out a meaningful Bayes factor or posterior probability calculation.

Some statisticians argue that the implied logic concerning a small P-value is compelling: "Either H_0 is true and a rare event has occurred, or H_0 is false." One could again argue against this reasoning as addressing the wrong question, but there is a more obvious major flaw: the "rare event" whose probability is being calculated under H_0 is *not* the event of observing the actual data x_0, but the event

$$E = \{\text{possible data } x\colon |T(x)| \geq |T(x_0)|\}.$$

The inclusion of all data "more extreme" than the actual x_0 is a curious step, and one which we have seen no remotely convincing justification for. Indeed, there are at least the following two arguments against such inclusion:

(i) There is a *vast* difference between being told that $X = x_0$ and being told only that $X \in E$. Intuitively, the difference is clear; the latter seems to be substantially stronger evidence against H_0. This is quantified in Berger and Sellke (1987), where it is demonstrated that $P(H_0 \mid E)$ is frequently very close to the P-value and, hence, much smaller than $P(H_0 \mid x_0)$. See also Good (1984), Note C200. Note that the "logic of surprise" cannot differentiate between x_0 and E (at least, in its usual forms of implementation).

(ii) The logic of surprise cannot separate evidence against the hypothesis from "unlucky" data. It is not hard to construct examples (cf. Example 19 on page 202 of Berger, 1985) in which data can be extremely surprising yet not contraindicate H_0.

This matter is of enough importance that one more example (this one similar to examples in Edwards, Lindman and Savage, 1963) is in order.

Example 7. An ESP experiment is conducted to see if the subject can forecast the outcome of the flip

of a fair coin. A fixed sample of 400 flips is taken, and the subject is correct 220 times. Letting $X = \#$ correct guesses out of $n = 400$ trials, with $\theta =$ probability of a correct guess, clearly $X \sim$ binomial (n, θ) and testing H_0: $\theta = 0.5$ versus H_1: $\theta \neq 0.5$ is of interest. The P-value of $x = 220$ (two-sided) is almost exactly .05 (note that the one-sided P-value would be 0.025), yet

$$\frac{f(220 \mid 0.5)}{\sup_\theta f(220 \mid \theta)} = \frac{f(220 \mid 0.5)}{f(220 \mid 0.55)} = 0.135 \cong \frac{1}{7.5},$$

where $f(x \mid \theta)$ is the binomial density. This likelihood ratio is an absolute lower bound on the Bayes factor, and so the evidence against H_0 can be no more than 7.5 to 1. Thus, even though we are surprised to see $x = 220$, and the amount of our surprise might be representable as 20 to 1 or 40 to 1 (two-sided or one-sided), we can explicitly calculate that the evidence against H_0 is at most 7.5 to 1. Use of the class (10) of "objective" weight functions would yield a lower bound on the Bayes factor of about 1/2.5, causing even greater doubt about the relevance of "surprise" to "evidence against H_0."

4.7 Rejoinder 7: Decision-Theoretic Analysis Is Necessary in Testing Problems

A frequent attempt to dismiss the conflict between P-values and Bayes factors is to argue that neither is relevant: one should instead quantify losses in incorrectly accepting or rejecting H_0 and perform a decision analysis. We certainly agree that this is a good idea, but would argue that the basic point we have made is still relevant. In particular, knowing that there is a probability mass $P(H_0 \mid x)$ at (or very close to) θ_0 can be very important in the decision analysis. For instance, if one is testing a current scientific theory H_0, the size of this point mass may be all important in the decision problem. Or, if one is conducting a screening test of two completely new drugs, with θ being the mean difference and H_0: $\theta = 0$ corresponding (typically) to both drugs being completely ineffective, then the size of $P(H_0 \mid x)$ will be very important in the decision problem concerning whether to proceed with more extensive testing. The point is that one cannot always ignore the special nature of θ_0 in decision analyses.

5. WHAT SHOULD BE DONE?

First and foremost, when testing precise hypotheses, formal use of P-values should be abandoned. Almost anything will give a better indication of the evidence provided by the data against H_0.

Before discussing alternatives to P-values, comments about informal uses of P-values are in order. It is often vigorously argued that P-values are valuable in informal stages of model development. (By "model" here we mean the entire stochastic structure, possibly including the prior or priors. This is even argued by many Bayesians (cf. Dempster, 1971, 1973; Box, 1980).) The argument is that, at a given point in the model development process, the currently entertained model is similar to a precise null, but alternatives have not been formulated; although Bayes factors cannot then be calculated, it is nevertheless desirable to have mechanisms to determine whether or not the present model is satisfactory. The P-value, it is argued, can be useful in making this decision.

This situation is very similar to that discussed in Section 4.4, in which we argued that the lack of alternatives does not in any way save the P-value. There is, however, an important difference here: the precise null being considered at an interim stage in the model formulation process is *not* usually a model that is believed to be true *a priori*, in the sense that there is a much sharper concentration of prior beliefs about this model than elsewhere (in "model space"). P-values are not necessarily terrible, as measures of evidence, in the absence of such *a priori* concentration. Of course, even here a small P-value should not, by itself, lead to rejection of the model; it may lead to a search for alternatives, but, once the alternatives are formulated, final decisions should be based on Bayes factors or posterior probabilities.

The above discussion points out the care that must be taken in formulating hypotheses and thinking about testing. With P-values, it matters very little whether, say, one formulates a problem as a one-sided test or as a two-sided test of a point null (assuming, of course, that θ is a real-valued parameter); the P-value can change by a factor of two, but this is relatively minor (in the grand scope of things). Bayes factors for the two formulations typically differ by at least an order of magnitude, however. Thus, it is crucial to distinguish between precise hypotheses that are just stated for convenience and have no special prior believability, and precise hypotheses which do correspond to a concentration of prior belief. It is the latter that are being addressed in this paper.

Returning to "what should be done," several possibilities deserve consideration:

Method 1. *Use the Lower Bounds* $\underline{B}$ *and* $\underline{P}$. It would be substantial improvement over P-values to report, as the evidence against H_0, the lower bound $\underline{B}$ in (8) or $\underline{P}(H_0 \mid x)$ in (9) (the latter for, say, $\pi_0 = ½$) with G chosen appropriately. This would, at least, result in the reporting of numbers that are (usually) of the correct order of magnitude.

The choice of G is somewhat arbitrary, but, as we have seen, there is often considerable robustness with respect to this choice. The bounds based on broad

classes such as (10) (e.g., (18) for the normal problem and (21) for the binomial problem) have the appeal of arguably being absolute "objective" lower bounds.

It is probably *not* optimal, however, to strive for absolute lower bounds, because we are imagining use of these numbers as actual *evidence against* H_0. A lower bound may well be unreasonably small, especially if G is chosen to be a large class. Recall that using a lower bound as the reported evidence is *bias toward* H_1, which is usually considered to be undesirable. (Indeed, the only reason we might actually encourage use of the lower bounds, with their attendant biases, is that they are much less misleading than are P-values.)

An attractive intermediate class, G, is the class of all standard conjugate densities with mean equal to θ_0. (Conjugate densities may not exist or be well-defined; but any low-dimensional class of priors having mean, or median, equal to θ_0 and having a wide range of variances would likely be reasonable.) Lower bounds for this class (see, for example, Section 3.2.4) are often not too extreme and tend to be almost trivial to calculate. This is because the marginal density, $m_g(x)$, typically has a closed form representation if g is conjugate, and the class of all conjugate g that have mean θ_0 is usually a small (often one-) dimensional class; the minimizations in (8) or (9) are then easy (often one-variable) minimization problems. Several other examples can be found in Edwards, Lindman and Savage (1963).

Method 2. *Use a Conventional Prior.* Although vastly superior to P-values, $\underline{B}$ and $\underline{P}$ can be misleadingly small. An indication of this is that all lower bounds discussed herein do *not* depend on the sample size. But Table 1 indicates that, for actual fixed prior opinions, $\underline{B}$ and $\underline{P}$ should eventually be increasing in n (indeed, roughly as a multiple of $\sqrt{n}$); cf. Jeffreys (1961), Lindley (1961) and Good (1984, Notes C144 and C200). Hence, especially for large n, $\underline{B}$ and $\underline{P}$ are suspect (although see Section 2).

This suggests simply choosing a prior density g in some conventional fashion. Note that, as argued in Section 1.2, there is no objective or noninformative choice for g that can be made. The conventional choices for g that are referenced in Section 1.2 are all reasonable, but it would take too long to discuss their differing motivations. We would feel comfortable with the use of any of the conventional choices, especially if augmented with presentation of $\underline{B}$ or $\underline{P}$.

Note, especially, the use of $P(H_0 \mid x)$ for $\pi_0 = ½$ and any of the conventional g. This requires *no* subjective inputs of the user, and is understandable as the final probability of H_0, in light of the data. It is as easy to use as a P-value, much easier to understand, and much less likely to be misinterpreted. The point of emphasizing this is that such automatic Bayesian methods are often criticized as being "arbitrary," although full subjective Bayesian analysis is described as being too hard. But the user domains are different. For those who have the capability, we would urge a fully subjective Bayesian analysis; for those requiring an automatic method, at least the automatic $P(H_0 \mid x)$ for a conventional prior will be much better than the automatic P-value.

Method 3. *Subjective Bayesian Analysis.* Fully subjective Bayesian analysis is often quite easy in this problem. The reason is that, as indicated in Section 2, only a few key features of the prior are typically required: π_0, the probability of H_0, and (say) the quartiles of g, the conditional prior distribution assuming that H_0 is false. (π_0 can be omitted if one uses the Bayes factor.) The exact functional form of g is frequently irrelevant, so that one can choose, say, a conjugate form to do the analysis.

Furthermore, graphical displays of, say, B as a function of the parameters of g can be used to communicate scientific results to a wide variety of different users with different prior opinions (cf. Dickey, 1973).

Example 8. Suppose we observe $X_1, \cdots, X_n$, distributed as iid. $N(\theta, \sigma^2)$ random variables, θ and σ^2 unknown. It is desired to test H_0: $\theta = 0$ versus H_1: $\theta \neq 0$. The actual experiment has $n = 15$, $\bar{x} = 20.93$, $s = 37.71$ and $t = \sqrt{n}\bar{x}/s = 2.15$. The P-value for this t test is 0.05.

To perform a subjective Bayesian analysis, it suffices to specify μ and τ, where μ is the median of g and $\mu \pm \tau$ are the quartiles. (Thus, μ would be the guess for θ, assuming H_0 to be false, etc.) Then, following the analysis in Section 4.10 of Berger (1985) (based on independent Cauchy and noninformative priors for θ and σ^2, respectively), one can construct a contour graph of B as a function of μ and τ. See Figure 2. Thus, an individual who believes that if H_1 is true then θ is likely to be between 10 and 50 (specifically, who chooses $\mu = 30$, so $|\mu - \bar{x}| = 9.07$ and $\tau = 20$) will conclude that $B \cong 0.38$ (i.e., the evidence against H_0 is only 1 to 2.5). An individual with $\mu = 30$, but τ only partly specified, say $10 \leq \tau \leq 30$, will conclude that $0.26 \leq B \leq 0.51$. Note that, for the "objective" choice $\mu = 0$, the lower bound on B over all τ is about 0.55, even though the P-value was 0.05.

APPENDIX

PROOF OF THEOREM 1. Since g_0 is unimodal and symmetric about θ_0, it can be represented as a mixture of uniform densities,

$$g_0(\theta) = \int_0^c \frac{1}{2a} I_{(\theta_0 - a, \theta_0 + a)}(\theta) F(da),$$

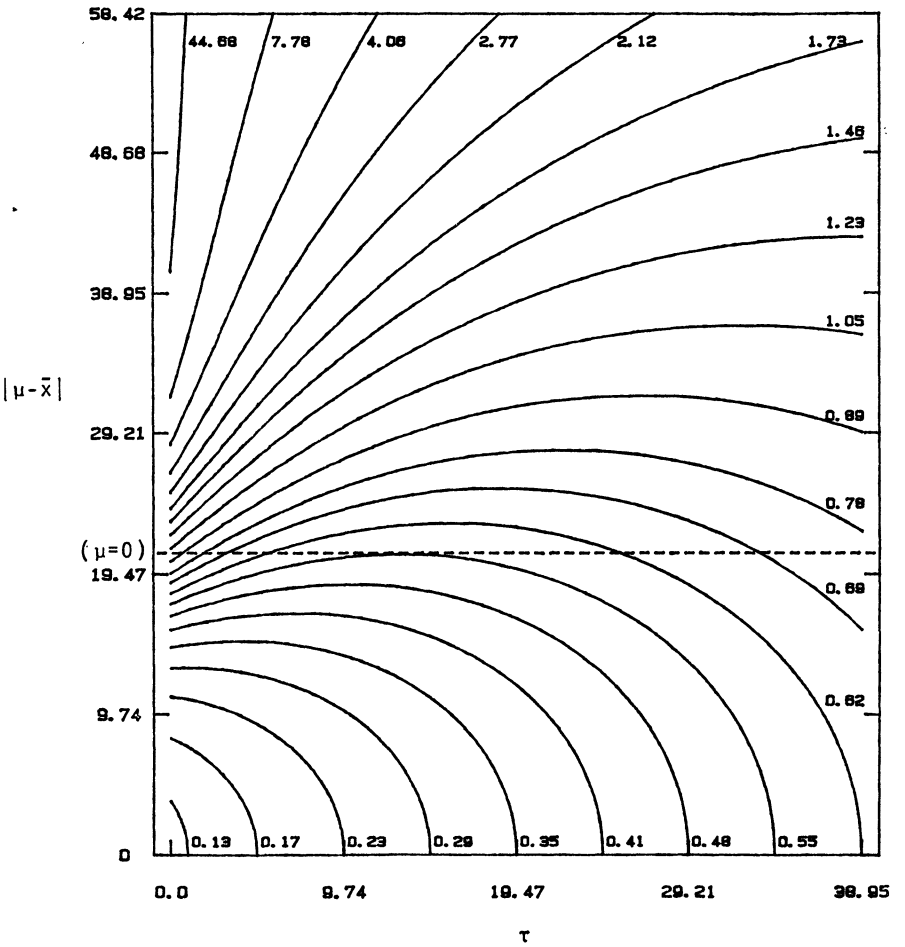

FIG. 2. *Contours of B as a function of* $|\mu - \bar{x}|$ *and* τ *when* $\bar{x} = 20.93$, $s = 37.71$ *and* $n = 15$.

so that

$$\int_\Omega f(\bar{x} \mid \theta) g_0(\theta)\, d\theta = \int_0^\varepsilon \left[\frac{1}{2a} \int_{\theta_0 - a}^{\theta_0 + a} f(\bar{x} \mid \theta)\, d\theta\right] F(da). \tag{22}$$

For $|t| \geq 1$ and $a < \varepsilon = \varepsilon^* \sigma/\sqrt{n}$ with $\varepsilon^* < |t| - 1$, it is easy to see that the bracketed integrand in (23) is an increasing function of a. Hence,

$$f(\bar{x} \mid \theta_0) \leq \int_\Omega f(\bar{x} \mid \theta) g_0(\theta)\, d\theta \leq \frac{1}{2\varepsilon} \int_\Omega f(\bar{x} \mid \theta)\, d\theta.$$

This can be rewritten

$$\int_\Omega f(\bar{x} \mid \theta) g_0(\theta)\, d\theta = f(\bar{x} \mid \theta_0)(1 + \psi_1), \tag{23}$$

where $0 \leq \psi_1 \leq \gamma$.

Next, we observe that, on $\bar{\Omega}$,

$$g_1(\theta) = \frac{g(\theta)}{(1 - \lambda)},$$

so that

$$\begin{aligned} \int_{\bar{\Omega}} f(\bar{x} \mid \theta) g_1(\theta)\, d\theta &= \frac{1}{(1 - \lambda)} \int_{\bar{\Omega}} f(\bar{x} \mid \theta) g(\theta)\, d\theta, \\ &= \left[m_g(\bar{x}) - \int_\Omega f(\bar{x} \mid \theta) g(\theta)\, d\theta\right] \\ &= m_g(\bar{x})(1 + \psi_2), \end{aligned} \tag{24}$$

where

$$\psi_2 = \frac{\lambda}{(1 - \lambda)}\left[1 - \frac{1}{m_g(\bar{x})} \int_{\bar{\Omega}} f(\bar{x} \mid \theta)\left(\frac{g(\theta)}{\lambda}\right) d\theta\right].$$

Observing that $g(\theta)/\lambda$ is a unimodal symmetric density on Ω, it follows as in the argument leading to (23) that

$$\int_{\bar{\Omega}} f(\bar{x} \mid \theta)\left(\frac{g(\theta)}{\lambda}\right) d\theta = f(\bar{x} \mid \theta_0)(1 + \psi_1^*),$$

where $0 \le \psi_1^* \le \gamma$. Hence,

$$\frac{\lambda}{(1-\lambda)}[1-\hat{B}(1+\gamma)] \le \psi_2 \le \frac{\lambda}{(1-\lambda)}[1-\hat{B}]. \tag{25}$$

Thus, combining (23), (24) and (25), we get

$$B = \frac{f(\bar{x} \mid \hat{\theta}_0)(1+\psi_1)}{m_g(\bar{x})(1+\psi_2)} = \hat{B}(1+\rho),$$

where

$$\frac{1}{1+\lambda(1-\lambda)^{-1}[1-\hat{B}]} \le (1+\rho) \le \frac{(1+\gamma)}{1+\lambda(1-\lambda)^{-1}[1-\hat{B}(1+\gamma)]}.$$

Algebra, together with the bounds $0 \le \hat{B} \le (1+\gamma)^{-1}$, yields (17). □

PROOF OF THEOREM 3.

$$\inf_{g \in G_{\mathrm{US}}} P(H_0 \mid n, x) = \left(1 + \frac{(1-\pi_0)}{\pi_0} \frac{\sup_{g \in G_{\mathrm{US}}} \int_0^1 \binom{n}{x} \theta^x (1-\theta)^{n-x} g(\theta)\, d\theta}{\binom{n}{x} \theta_0^x (1-\theta_0)^{n-x}}\right)^{-1},$$

where G_{US} consists of all g such that

$$g(\theta) = h(u(\theta)) \frac{du(\theta)}{d\theta}$$

and h is unimodal symmetric about 0. Let H be the class of all unimodal symmetric densities h. Then,

$$\begin{aligned} \sup_{g \in G} \int_0^1 \binom{n}{x} \theta^x (1-\theta)^{n-x} g(\theta)\, d\theta &= \sup_{h \in H} \int_{-\infty}^{\infty} \binom{n}{x} \theta(u)^x (1-\theta(u))^{n-x} h(u)\, du \\ &= \sup_k \frac{1}{2k} \int_{-k}^{k} \binom{n}{x} \theta(u)^x (1-\theta(u))^{n-x}\, du, \end{aligned} \tag{26}$$

from Theorem 2, which proves the result. □

PROOF OF THEOREM 4.

$$\begin{aligned} \sup_{g \in G_1} \int_0^1 \binom{n}{x} \theta^x (1-\theta)^{n-x} g(\theta)\, d\theta &= \sup_{h_1} \int_0^{\theta_0} \binom{n}{x} p^x (1-p)^{n-x} h_1(\theta)\, d\theta \\ &\quad + \sup_{h_2} \int_{\theta_0}^1 \binom{n}{x} \theta^x (1-\theta)^{n-x} h_2(\theta)\, d\theta, \end{aligned}$$

where $h_1 \ge 0$ on $[0, \theta_0]$, $h_2 \ge 0$ on $[\theta_0, 1]$ and $\int_0^{\theta_0} h_1(\theta)\, d\theta = 1/2 = \int_{\theta_0}^1 h_2(\theta)\, d\theta$. Let $x/n \le \theta_0$. Then,

$$\sup_{h_1} \int_0^{\theta_0} \binom{n}{x} \theta^x (1-\theta)^{n-x} h_1(\theta)\, d\theta = \frac{1}{2}\binom{n}{x}\left(\frac{x}{n}\right)^x \left(1 - \frac{x}{n}\right)^{n-x},$$

$$\sup_{h_2} \int_{\theta_0}^1 \binom{n}{x} \theta^x (1-\theta)^{n-x} h_2(\theta)\, d\theta = \frac{1}{2}\binom{n}{x} \theta_0^x (1-\theta_0)^{n-x},$$

because $\binom{n}{x}\theta^x(1-\theta)^{n-x}$ has its maximum at $\hat{\theta} = x/n$ and decreases monotonically on both sides. In the other case, where $x/n > \theta_0$, the roles of h_1 and h_2 are reversed. □

PROOF OF THEOREM 5. Assume that $x/n \le \theta_0$; the other case is similar. Let

$$I = \sup_{g \in G_0} \int_0^1 \binom{n}{x} \theta^x (1-\theta)^{n-x} g(\theta)\, d\theta.$$

Then

$$\begin{aligned} I &= \sup_{h_1} \int_0^{\theta_0} \binom{n}{x} \theta^x (1-\theta)^{n-x} h_1(\theta)\, d\theta \\ &\quad + \sup_{h_2} \int_{\theta_0}^1 \binom{n}{x} \theta^x (1-\theta)^{n-x} h_2(\theta)\, d\theta, \end{aligned}$$

where $h_1 \ge 0$ and nondecreasing on $[0, \theta_0]$, $h_2 \ge 0$ and nonincreasing on $[\theta_0, 1]$, and $\int_0^{\theta_0} h_1(\theta)\, d\theta = 1/2 = \int_{\theta_0}^1 h_2(\theta)\, d\theta$. Therefore,

$$\begin{aligned} I &= \sup_{g \in G_0} \int_0^1 \binom{n}{x} \theta^x (1-\theta)^{n-x} g(\theta)\, d\theta \\ &= \sup_{h_1} \int_0^{\theta_0} \binom{n}{x} \theta^x (1-\theta)^{n-x} h_1(\theta)\, d\theta \\ &\quad + \frac{1}{2}\binom{n}{x} \theta_0^x (1-\theta_0)^{n-x}, \end{aligned}$$

because $\binom{n}{x}\theta^x(1-\theta)^{n-x}$, in this case, is decreasing monotonically on both sides of $(x/n) \le \theta_0$. Hence,

$$I = \sup_{h_3} \int_0^{2\theta_0} g(x, n, \theta) h_3(\theta)\, d\theta + \frac{1}{2}\binom{n}{x} \theta_0^x (1-\theta_0)^{n-x},$$

where

$$g(x, n, \theta) = \begin{cases} \binom{n}{x} \theta^x (1-\theta)^{n-x} & \text{for } \theta \le \theta_0, \\ \binom{n}{x} (2\theta_0 - \theta)^x (1 - (2\theta_0 - \theta))^{n-x} & \text{for } \theta > \theta_0, \end{cases}$$

and $h_3 \geq 0$ is unimodal and symmetric about θ_0, $\int_0^{2\theta_0} h_3(\theta)\, d\theta = \frac{1}{2}$. However,

$$\sup_{h_3} \int_0^{2\theta_0} g(x, n, \theta) h_3(\theta)\, d\theta$$

$$\doteq \sup_r \frac{1}{4r} \int_{\theta_0 - r}^{\theta_0 + r} g(x, n, \theta)\, d\theta$$

$$= \sup_r \frac{1}{2r} \int_{\theta_0 - r}^{\theta_0} \binom{n}{x} \theta^x (1-\theta)^{n-x}\, d\theta,$$

because unimodal symmetric densities are mixtures of uniforms. □

PROOF OF THEOREM 6.

$$\inf_{\pi \in C} P^\pi(H_0 \mid x)$$

$$= \left[1 + \frac{(1-\pi_0)}{\pi_0} \frac{\sup_{\pi \in C} \int_{\theta \neq \theta_0} f(x \mid \theta) \pi(\theta)\, d\theta}{f(x \mid \theta_0)}\right]^{-1}.$$

Now for $\theta \neq \theta_0$

$$\pi(\theta) I(\theta \neq \theta_0)$$

$$= \frac{\Gamma(c)}{\Gamma(c\theta_0)\Gamma(c(1-\theta_0))} \theta^{c\theta_0 - 1}(1-\theta)^{c(1-\theta_0)-1}.$$

Therefore,

$$\sup_{\pi \in C} \int_{\theta \neq \theta_0} f(x \mid \theta) \pi(\theta)\, d\theta$$

$$= \sup_{c>0} \left\{ \int_{\theta \neq \theta_0} \binom{n}{x} \theta^x (1-\theta)^{n-x} \times \frac{\Gamma(c)}{\Gamma(c\theta_0)\Gamma(c(1-\theta_0))} \theta^{c\theta_0 - 1}(1-\theta)^{c(1-\theta_0)-1}\, d\theta \right\}$$

$$= \sup_{c>0} \left\{ \frac{\Gamma(c)}{\Gamma(c\theta_0)\Gamma(c(1-\theta_0))} \binom{n}{x} \times \int_0^1 \theta^{x + c\theta_0 - 1}(1-\theta)^{n-x+c(1-\theta_0)-1}\, d\theta \right\}$$

$$= \sup_{c>0} \left\{ \frac{\Gamma(c)}{\Gamma(c\theta_0)\Gamma(c(1-\theta_0))} \binom{n}{x} \times \frac{\Gamma(x + c\theta_0)\Gamma(n - x + c(1-\theta_0))}{\Gamma(n+c)} \right\},$$

which proves the theorem. □

ACKNOWLEDGMENTS

Research for this article was supported by the National Science Foundation Grant DMS-84-01996 and by a David Ross Grant at Purdue University. The authors are grateful to Morris DeGroot for many perceptive comments and suggestions, to Patricia Pepple for spotting two errors and to Tsai Hung Fan for the development of Figure 2.

REFERENCES

BERGER, J. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer, New York.

BERGER, J. (1986). Are P-values reasonable measures of accuracy. In *Pacific Statistical Congress* (I. S. Francis et al., eds.). North-Holland, Amsterdam.

BERGER, J. and BERRY, D. (1987). The relevance of stopping rules in statistical inference. In *Statistical Decision Theory and Related Topics IV* (S. Gupta and J. Berger, eds.). Springer, New York. To appear.

BERGER, J. and SELLKE, T. (1987). Testing a point null hypothesis: The irreconcilability of P-values and evidence (with discussion). *J. Amer. Statist. Assoc.* **82** 112–139.

BERGER, J. O. and WOLPERT, R. L. (1984). *The Likelihood Principle*. IMS, Hayward, Calif.

BERNARDO, J. M. (1980). A Bayesian analysis of classical hypothesis testing. In *Bayesian Statistics* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.) 605–618. University Press, Valencia.

BOX, G. E. P. (1980). Sampling and Bayes inference in scientific modeling and robustness (with discussion). *J. Roy. Statist. Soc. Ser. A* **143** 383–430.

CASELLA, G. and BERGER, R. L. (1987). Reconciling Bayesian and frequentist evidence in the one-sided testing problem. *J. Amer. Statist. Assoc.* **82** 106–111.

DEGROOT, M. H. (1973). Doing what comes naturally: Interpreting a tail area as a posterior probability or as a likelihood ratio. *J. Amer. Statist. Assoc.* **68** 966–969.

DELAMPADY, M. (1986a). Testing a precise hypothesis: Interpreting P values from a robust Bayesian viewpoint. Ph.D. thesis, Purdue Univ.

DELAMPADY, M. (1986b). Lower bounds on Bayes factors for interval null hypotheses. Technical Report 86-35, Dept. Statistics, Purdue Univ.

DELAMPADY, M. (1986c). Lower bounds on Bayes factors for invariant testing situations. Technical Report 86-36, Dept. Statistics, Purdue Univ.

DELAMPADY, M. and BERGER, J. (1987). Lower bounds on posterior probabilities for multinomial and χ^2 tests. Technical Report 86-37, Dept. Statistics, Purdue Univ.

DEMPSTER, A. P. (1971). Model searching and estimation in the logic of inference. In *Foundations of Statistical Inference* (V. P. Godambe and D. A. Sprott, eds.). Holt, Rinehart, and Winston, Toronto.

DEMPSTER, A. P. (1973). The direct use of likelihood for significance testing. In *Proc. of the Conference on Foundational Questions in Statistical Inference* (O. Barndorff-Nielsen et al., eds.) 335–352. Dept. Theoretical Statistics, Univ. Aarhus.

DIAMOND, G. A. and FORRESTER, J. S. (1983). Clinical trials and statistical verdicts: Probable grounds for appeal. *Ann. Internal Med.* **98** 385–394.

DICKEY, J. M. (1971). The weighted likelihood ratio, linear hypotheses on normal location parameters. *Ann. Math. Statist.* **42** 204–223.

DICKEY, J. M. (1973). Scientific reporting. *J. Roy. Statist. Soc. Ser. B* **35** 285–305.

DICKEY, J. M. (1974). Bayesian alternatives to the F test and least squares estimate in the linear model. In *Studies in Bayesian Econometrics and Statistics* (S. E. Fienberg and A. Zellner, eds.) 515–554. North-Holland, Amsterdam.

DICKEY, J. M. (1976). Approximate posterior distributions. *J. Amer. Statist. Assoc.* **71** 680–689.

DICKEY, J. M. (1977). Is the tail area useful as an approximate Bayes factor? *J. Amer. Statist. Assoc.* **72** 138–142.

DICKEY, J. M. (1980). Approximate coherence for regression model inference— with a new analysis of Fisher's Broadback Wheatfield example. In *Bayesian Analysis in Econometrics and Statistics: Essays in Honor of Harold Jeffreys* (A. Zellner, ed.) 333–354. North-Holland, Amsterdam.

EDWARDS, W., LINDMAN, H. and SAVAGE, L. J. (1963). Bayesian statistical inference for psychological research. *Psychol. Rev.* **70** 193–242. Reprinted in *Robustness of Bayesian Analysis* (J. B. Kadane, ed.). North-Holland, Amsterdam, 1984.

GÓMEZ VILLEGAS, M. A. and DE LA HORRA NAVARRO, J. (1984). Aproximacion de factores Bayes. *Cuad. Bioestadist.* **2** 355–361.

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

GOOD, I. J. (1958). Significance tests in parallel and in series. *J. Amer. Statist. Assoc.* **53** 799–813.

GOOD, I. J. (1965). *The Estimation of Probabilities: An Essay on Modern Bayesian Methods.* M.I.T. Press, Cambridge, Mass.

GOOD, I. J. (1967). A Bayesian significance test for the multinomial distribution. *J. Roy. Statist. Soc. Ser. B* **29** 399–431.

GOOD, I. J. (1983). *Good Thinking: The Foundations of Probability and Its Applications.* Univ. Minnesota Press, Minneapolis.

GOOD, I. J. (1984). Notes C140, C144, C199, C200 and C201. *J. Statist. Comput. Simulation* **19**.

GOOD, I. J. (1985). Weight of evidence: A brief survey. In *Bayesian Statistics 2* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.) 249–270. North-Holland, Amsterdam.

GOOD, I. J. (1986). A flexible Bayesian model for comparing two treatments. *J. Statist. Comput. Simulation* **26** 301–305.

HILDRETH, C. (1963). Bayesian statisticians and remote clients. *Econometrica* **31** 422–438.

HILL, B. (1982). Comment on "Lindley's paradox," by G. Shafer. *J. Amer. Statist. Assoc.* **77** 344–347.

HODGES, J. L., JR. and LEHMANN, E. L. (1954). Testing the approximate validity of statistical hypotheses. *J. Roy. Statist. Soc. Ser. B* **16** 261–268.

JEFFREYS, H. (1957). *Scientific Inference.* Cambridge Univ. Press, Cambridge.

JEFFREYS, H. (1961). *Theory of Probability*, 3rd ed. Oxford Univ. Press, London.

JEFFREYS, H. (1980). Some general points in probability theory. In *Bayesian Analysis in Econometrics and Statistics: Essays in Honor of Harold Jeffreys* (A. Zellner, ed.) 451–453. North-Holland, Amsterdam.

LEAMER, E. E. (1978). *Specification Searches: Ad Hoc Inference with Nonexperimental Data.* Wiley, New York.

LEMPERS, F. B. (1971). *Posterior Probabilities of Alternative Linear Models.* Univ. Rotterdam Press, Rotterdam.

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

LINDLEY, D. V. (1961). The use of prior probability distributions in statistical inference and decisions. *Proc. Fourth Berkeley Symp. Math. Statist. Probab.* **1** 453–468. Univ. California Press.

LINDLEY, D. V. (1965). *An Introduction to Probability and Statistics from a Bayesian Viewpoint* **1, 2**. Cambridge Univ. Press, Cambridge.

LINDLEY, D. V. (1977). A problem in forensic science. *Biometrika* **64** 207–213.

PRATT, J. W. (1965). Bayesian interpretation of standard inference statements (with discussion). *J. Roy. Statist. Soc. Ser. B* **27** 169–203.

RAIFFA, H. and SCHLAIFFER, R. (1961). *Applied Statistical Decision Theory.* Division of Research, Graduate School of Business Administration, Harvard Univ.

RUBIN, H. (1971). A decision-theoretic approach to the problem of testing a null hypothesis. In *Statistical Decision Theory and Related Topics* (S. S. Gupta and J. Yackel, eds.). Academic, New York.

SHAFER, G. (1982). Lindley's paradox (with discussion). *J. Amer. Statist. Assoc.* **77** 325–351.

SMITH, A. F. M. and SPIEGELHALTER, D. J. (1980). Bayes factors and choice criteria for linear models. *J. Roy. Statist. Soc. Ser. B* **42** 213–220.

SMITH, C. A. B. (1965). Personal probability and statistical analysis. *J. Roy. Statist. Soc. Ser. A* **128** 469–499.

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

ZELLNER, A. (1984). Posterior odds ratios for regression hypotheses: General considerations and some specific results. In *Basic Issues in Econometrics* (A. Zellner, ed.) 275–305. Univ. Chicago Press, Chicago.

ZELLNER, A. and SIOW, A. (1980). Posterior odds ratios for selected regression hypotheses. In *Bayesian Statistics* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.) 585–603. University Press, Valencia.

Comment

D. R. Cox

The use and misuse of significance tests has been a popular topic of comment for many years and from many points of view. Although the matter could hardly be said to be underdiscussed in the recent literature, Professors Berger and Delampady have made a valuable addition to that literature by their careful account of the relation between P-values and the posterior probabilities of "precise" null hypotheses, a matter first raised many years ago by H. Jeffreys. The extension of the discussion to include broad classes of priors is particularly striking.

For those taking an eclectic view of statistical theory the comparison of different approaches to the same or similar problems is important, sometimes soothing and occasionally constructively alarming. What is one to make of the present comparisons? The authors are in no doubt.

(i) "Rejoinder 5. P-values have a valid frequentist interpretation. This rejoinder is simply not true" (Section 4.5).

D. R. Cox is Professor of Statistics, SERC Senior Research Fellow, Department of Mathematics, Imperial College, London SW7 2BZ, United Kingdom.