Bayesian Testing of Precise Hypotheses *

by

James Berger
Purdue University

and

Mohan Delampady
Purdue University and
University of British Columbia

Technical Report #86-38

Department of Statistics
Purdue University

August 1986

* Research supported by the National Science Foundation, Grant DMS-8401996.
Testing Precise Hypotheses

James 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 possibility 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, chi-squared tests, scientific communication.

1 Introduction and Basics

1.1 Basics and Measures of Evidence

Suppose $X$ having density $f(x|\theta)$ is observed, with $\theta$ being an unknown element of the parameter space $\Theta$, 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 3 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

$$\alpha = P_{\theta_0} (|T(X)| \geq |t|).$$

(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

$$B = \frac{f(x|\theta_0)}{m_g(x)},$$

(2)

where

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

(3)

For a Bayesian, $g$ would be the prior density for $\theta$, 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 $\pi_0$, then the posterior probability of $H_0$ is

$$P(H_0|x) = \left[ 1 + \frac{(1-\pi_0)}{\pi_0} \frac{m_g(x)}{f(x|\theta_0)} \right]^{-1}$$

$$= \left[ 1 + \frac{(1-\pi_0)}{\pi_0} \cdot \frac{1}{B} \right]^{-1}.$$  

(4)

**Example 1.** Suppose we observe $X \sim N(\theta, \sigma^2/\pi)$, where $\sigma^2$ is known. Then, letting

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

one obtains the usual $P$–value as

$$\alpha = 2 \left[ 1 - \Phi(\lambda) \right],$$

where $\Phi$ is the standard normal c.d.f.

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

$$B = \sqrt{1 + \rho^{-2}} \exp \left\{ -\frac{1}{2} \left( \frac{\lambda - \rho \eta}{(1 + \rho^2) - \eta^2} \right) \right\},$$

(6)
where \( \rho = \sigma / (\sqrt{n} \tau) \) and \( \eta = (\theta_0 - \mu) / \tau \). Note that \( \mu \) will often be chosen to be \( \theta_0 \) (so as to have a symmetric "weight function"), in which case

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

The posterior probability of \( H_0 \) can be found from these formulas and (4), provided \( \pi_0 \) is specified.

As a specific example, suppose \( \mu = \theta_0, \ \tau = \sigma, \) and \( \pi_0 = 1/2. \) For various \( t \) and \( n, \) the various measures of evidence are given in Table 1. There \( P \) stands for \( P(H_0|x). \)

<table>
<thead>
<tr>
<th>( t )</th>
<th>( \alpha )</th>
<th>( n )</th>
<th>( 1 )</th>
<th>( 5 )</th>
<th>( 10 )</th>
<th>( 20 )</th>
<th>( 50 )</th>
<th>( 100 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.645</td>
<td>.10</td>
<td></td>
<td>.72</td>
<td>.42</td>
<td>.79</td>
<td>.44</td>
<td>.89</td>
<td>.47</td>
</tr>
<tr>
<td>1.960</td>
<td>.05</td>
<td></td>
<td>.54</td>
<td>.35</td>
<td>.49</td>
<td>.33</td>
<td>.59</td>
<td>.37</td>
</tr>
<tr>
<td>2.576</td>
<td>.01</td>
<td></td>
<td>.27</td>
<td>.21</td>
<td>.15</td>
<td>.13</td>
<td>.16</td>
<td>.14</td>
</tr>
<tr>
<td>3.291</td>
<td>.001</td>
<td></td>
<td>.10</td>
<td>.09</td>
<td>.03</td>
<td>.03</td>
<td>.02</td>
<td>.02</td>
</tr>
</tbody>
</table>

**1.2 Motivation**

Here are some commonly held opinions which 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, \( P(H_0|x) \) is from 5 to 50 times larger than the P-value, \( \alpha. \) 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, while the classical statistician would "reject \( H_0 \) at significance level \( \alpha = .05 \)." Admittedly, classical statisticians will warn against interpreting \( \alpha \) 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.)

**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 \( \pi_0 \) and \( g \). One can argue that \( \pi_0=1/2 \) is the objective choice for the prior probability of \( H_0 \), or can objectively avoid the choice of \( \pi_0 \) by concentrating on the Bayes factor, \( B \). There is no choice of \( g \), however, which can claim to be objective. One might argue that \( g \) should be symmetric about \( \theta_0 \) (at least when the parameter space is the entire real line), and perhaps \( g \) should be nonincreasing in \(|\theta - \theta_0|\) (to avoid treating values of \( \theta \) other than \( \theta_0 \) as special). And 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 \((\theta_0, \tau^2) \) (the form preferred by Jeffreys). Unfortunately, the choice of the scale factor, \( \tau \), 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) \( \tau \) the value of \( \rho \) will typically be small, so that

\[
B \approx \frac{\sqrt{n\tau}}{\sigma} \exp\left(-\frac{1}{2}t^2\right),
\]
and (with $\tau_0 = 1/2$),

$$P(H_0|x) \cong \frac{1}{1 + \frac{\sigma}{\sqrt{nr}} \exp\left(\frac{1}{2} t^2\right)}.$$

Thus $\tau$ has a dramatic effect on the Bayesian or likelihood answer. Furthermore, letting $\tau^2 \to \infty$ so that $g$ becomes “noninformative” is ridiculous, since then $P(H_0|x) \to 1$. Thus a Bayesian must, at a minimum, subjectively specify $\tau^2$, and there is no default value which “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 ($\theta_0, \sigma^2$) default $g$, while 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 can not 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 $\pi_0$ 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, while perhaps objective, is nearly meaningless, and that the actual meaning usually ascribed to a P-value by practitioners contains hidden and extreme bias.


The motivation for this opinion is that significance testing ignores a crucial question, namely “how far is $\theta$ from $\theta_0$?” Having statistically significant evidence
that \( \theta \neq \theta_0 \) might be irrelevant if \( \theta \) and \( \theta_0 \) are within, say, \( 10^{-10} \) of each other. So, the argument goes, one should simply find (say) a 95\% confidence set for \( \theta \). If \( \theta_0 \) 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 \( \theta \) and \( \theta_0 \).

This opinion is wrong, because it ignores the supposed special nature of \( \theta_0 \). A point can be outside a 95\% confidence set, yet not be 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 \( \theta_0 \). 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 \( \alpha \) and the conditional measures \( B \) or \( P(H_0|x) \) 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 \( P(H_0|x) \) over wide classes of \( g \), and observing that even these lower bounds are much larger than \( \alpha \). Thus, if \( G \) is a class of densities under consideration, we will investigate

\[
B = \inf_{g \in G} B = \frac{f(x|\theta_0)}{\sup_{g \in G} m_g(x)} \tag{8}
\]

and

\[
P(H_0|x) = \inf_{g \in G} P(H_0|x) = \left[ 1 + \frac{(1 - \pi_0)}{\pi_0} \cdot \frac{1}{B} \right]^{-1}. \tag{9}
\]

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, though 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 apriori 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 $B$ or $P(H_0|x)$ are calculated by choosing that $g \in G$ which is most favorable to $H_1$; to minimize this bias towards $H_1$, one should try to choose $G$ as small as possible, consistent with apriori beliefs.

In this light, choosing $G = \{ \text{all densities} \}$ seems certainly too extreme, allowing for severe bias towards $H_1$. It is astonishing that, even for this choice, $B$ and $P(H_0|x)$ (for $\pi_0 = 1/2$) are still often substantially larger than the $P$-value (see Edwards, Lindman and Savage (1963)), indicating that even extreme bias towards $H_1$ 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

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 $P(H_0|\alpha)$ over $g \in G$. They were the first to make the extent of the conflict between P-values and posterior probabilities unambiguously clear. Berger and Sellke (1987) was 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 $P$ and $B$. 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 $B$ and $P$ 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" $B$; 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 which 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 which can be exactly modeled as \( \theta = \theta_0 \). One might feel that hypotheses such as

\[
H_0 : \text{A subject has no ESP,}
\]

or

\[
H_0 : \text{Talking to plants has no effect on their growth,}
\]

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

More common precise hypotheses, such as

\[
H_0 : \text{Vitamin C has no effect on the common cold,}
\]

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

\[
H_0 : |\theta - \theta_0| \leq \epsilon \quad \text{versus} \quad H_1 : |\theta - \theta_0| > \epsilon,
\]

(11)

where \( \epsilon \) is “small”. It is then of interest to determine when one can approximate (11) by the point null test

\[
H_0^* : \theta = \theta_0 \quad \text{versus} \quad H_1^* : \theta \neq \theta_0.
\]

2.1 Classical Approximation by Point Nulls

From a classical perspective, there is no difficulty in determining when \( \epsilon \) 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

\[
\alpha_\epsilon = \sup_{\theta : |\theta - \theta_0| \leq \epsilon} P_\theta(|T(X)| \geq |t|).
\]

Thus one can seek conditions under which \( \alpha_\epsilon \approx \alpha_0 \) (the P-value corresponding to \( H_0^* \)).
Example 2. Suppose one observes \( \bar{X} \sim N(\theta, \sigma^2/n) \), where \( \sigma^2 \) is known, and that it is desired to test (11). Again defining \( T(\bar{X}) = \sqrt{n}(\bar{X} - \theta_0)/\sigma \), calculation gives

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

Suppose that one is interested in (say) determining when

\[
\alpha_0 \geq (0.9)\alpha_\epsilon
\]

(so that approximating the interval null by a point leads to no more than a 10% error in the P-value). The values \( \epsilon^* = \epsilon\sqrt{n}/\sigma \) which achieve this are those values less than or equal to the entries in Table 2 (for various \( t \)).

<table>
<thead>
<tr>
<th>( t )</th>
<th>1.645</th>
<th>1.96</th>
<th>2.576</th>
<th>2.807</th>
<th>3.29</th>
<th>3.89</th>
</tr>
</thead>
<tbody>
<tr>
<td>P-value</td>
<td>0.10</td>
<td>0.05</td>
<td>0.01</td>
<td>0.005</td>
<td>0.001</td>
<td>0.0001</td>
</tr>
<tr>
<td>Bound on ( \epsilon^* )</td>
<td>0.257</td>
<td>0.221</td>
<td>0.173</td>
<td>0.160</td>
<td>0.138</td>
<td>0.117</td>
</tr>
</tbody>
</table>

Thus, as long as \( \epsilon \) is no more than \( \frac{1}{4} \) to \( \frac{1}{6} \) of a sample standard deviation, the use of a point null will cause at most 10% error in the calculated P-value (for moderate \( t \)).

2.2 Bayesian Approximation by Point Nulls

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

\[
\pi_0 = \int_{\Omega} \pi(\theta) \, d\theta, \quad g_0(\theta) = \frac{1}{\pi_0} \cdot \pi(\theta) I_{\Omega}(\theta), \text{ and}
\]

\[
g_1(\theta) = \frac{1}{(1 - \pi_0)} \cdot \pi(\theta) I_{\overline{\Omega}}(\theta). \tag{12}
\]

Thus \( \pi_0 \) is the prior probability assigned to \( H_0 \), \( g_0 \) is the conditional density of \( \theta \) given that \( H_0 \) is true, and \( g_1 \) is the conditional density of \( \theta \) given that \( H_1 \) is true. Typically, \( g_0 \) will be a sharply spiked density, while \( g_1 \) will be rather diffuse.
It is usually fairly easy to specify $\pi_0$, 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 $\pi_0$ 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 $\epsilon$ and to choose $g_0$; it is simply hard to make fine distinctions about small sets with large probability.

Because of the difficulties in choosing $\epsilon$ 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 $\pi_0$ to $\{\theta = \theta_0\}$, and gives conditional prior density $g(\theta)$ to $\{\theta \neq \theta_0\}$; we assume that

\begin{equation}
(i) \quad g(\theta) \propto g_1(\theta) \quad \text{on} \quad |\theta - \theta_0| > \epsilon;
(ii) \quad \lambda = \int_\Omega g(\theta) \, d\theta \quad \text{is suitably small.}
\end{equation}

The idea here is that one assigns the same probability to $H_0^*$ as to $H_0$, and specifies $g(\theta)$ 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 $\epsilon$ and $g_0$.

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

\begin{equation}
B = \frac{\int_\Omega f(\bar{x} | \theta) g_0(\theta) \, d\theta}{\int_\Omega f(\bar{x} | \theta) g_1(\theta) \, d\theta},
\end{equation}

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

\begin{equation}
\hat{B} = \frac{f(\bar{x} | \theta_0)}{m_g(\bar{x})},
\end{equation}

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 $\epsilon^* = \epsilon \sqrt{n} / \sigma$ and $t = \sqrt{n}(\bar{x} - \theta_0) / \sigma$ as in Example 2, and define

\begin{equation}
\gamma = \frac{1}{2 \epsilon^* \phi(t)} [\Phi(t + \epsilon^*) - \Phi(t - \epsilon^*)] - 1,
\end{equation}

where $\phi$ and $\Phi$ are the standard normal density and c.d.f., respectively.
Theorem 1. Suppose that \( \pi(\theta) \) and \( g(\theta) \) are unimodal and symmetric about \( \theta_0 \), and that \( |t| \geq 1 \), \( \epsilon^* < |t| - 1 \), and \( \hat{B} \leq (1 + \gamma)^{-1} \). Then

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

where

\[
-\lambda \leq \frac{\lambda (\hat{B} - 1)}{1 - \lambda \hat{B}} \leq \rho \leq \frac{\gamma + \lambda(1 + \gamma)(\hat{B} - 1)}{1 - \lambda \hat{B}(1 + \gamma)} \leq \gamma.
\] (17)

Proof: See the Appendix. \( \square \)

If both \( \lambda \) and \( \gamma \) are small, then the error in approximating \( B \) by \( \hat{B} \) will be small. The following table gives, for various common \( t \) and associated P–values, the maximum \( \epsilon^* \) for which \( \gamma \leq .1 \).

| Table 3: \( \epsilon^* \) 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 \( \epsilon^* \) | 0.62           | 0.47           | 0.33           | 0.30           | 0.25           | 0.20           |

Note also that a crude bound on \( \lambda \) is

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

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 \( \tau^2 \). Then

\[
\lambda \leq \sqrt{\frac{2}{\pi n}} \cdot \frac{\epsilon^*\sigma}{\tau}.
\]

This is less than .1 if \( n \geq 64(\epsilon^*\sigma/\tau)^2 \). Thus, if \( t = 1.96 \) is observed, \( \sigma = \tau, \epsilon^* = 1/4 \), 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 \( \epsilon = \epsilon^*\sigma/\sqrt{n} \) is the half-width of \( \Omega \), these numbers suggest that the point null approximation to \( H_0 \) will be reasonable so long as \( \Omega \) 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 $\pi_0$ and $g$, data $\bar{x}_n$ which yields (for each sample size $n$) a fixed $P$-value $\alpha$, will result in

$$P(H_0^* | \bar{x}_n) \rightarrow 1 \text{ as } n \rightarrow \infty,$$

no matter how small $\alpha$ 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 $n \rightarrow \infty$, $\epsilon = \epsilon^* \sigma / \sqrt{n} \rightarrow 0$. Thus, a precise hypothesis $H_0 : |\theta - \theta_0| \leq \epsilon_0$ will fail to be approximable by $H_0^* : \theta = \theta_0$ when $n$ gets very large. In fact, for $H_0 : |\theta - \theta_0| \leq \epsilon_0$, it can be shown for fixed $\pi(\theta)$ that $\bar{x}_n$ which yields (for each sample size $n$) a fixed $P$-value $\alpha$, will often result in

$$P(H_0 | \bar{x}_n) \rightarrow \alpha \text{ as } n \rightarrow \infty,$$

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 $\alpha$), 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)).

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 $\pi(\theta | x, \Omega)$, 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 $\pi(\theta | x, \Omega)$ would communicate the location of $\theta$ 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 \( \pi_0 \), 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, while 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 \( \theta_0 \) (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 \( \theta_0 \) 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 \( g \) which will excessively bias the lower bounds \( B \) and \( P \) towards \( H_1 \). 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

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

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

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

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

It follows from Theorem 2 that $B$ and $P(H_0|x)$ (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 $X \sim N_p(\theta, I)$, where $X = (X_1, X_2, \ldots, X_p)$ and $\theta = (\theta_1, \theta_2, \ldots, \theta_p)$. It is desired to test

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

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

$$T(X) = |X - \theta^0|^2,$$

which has a $\chi^2_p$ distribution under $H_0$. Therefore the $P$-value of the data $x$ is

$$\alpha = P\left(\chi^2_p \geq T(x)\right).$$

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

$$B = \frac{\exp(-\frac{1}{2} |x - \theta^0|^2)}{\sup_{k}\frac{1}{V(k)} \int_{|\theta-\theta^0| \leq k} \exp(-\frac{1}{2} |\theta - x|^2)\,d\theta}.$$  \hspace{1cm} (18)

Numerical values in Table 4 were computed using the above result for different dimensions and different $P$-values. Here $\alpha$ is the $P$-value, $P$ is the lower bound.
Table 4: Lower Bounds for Spherically Symmetric Unimodal Densities

<table>
<thead>
<tr>
<th>Dimension</th>
<th>$\alpha = .001$</th>
<th>$\alpha = .01$</th>
<th>$\alpha = .05$</th>
<th>$\alpha = .10$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$P$</td>
<td>$B$</td>
<td>$P$</td>
<td>$B$</td>
</tr>
<tr>
<td>1</td>
<td>.0179</td>
<td>.0182</td>
<td>.1093</td>
<td>.1227</td>
</tr>
<tr>
<td>2</td>
<td>.0141</td>
<td>.0143</td>
<td>.0891</td>
<td>.0978</td>
</tr>
<tr>
<td>3</td>
<td>.0118</td>
<td>.0119</td>
<td>.0827</td>
<td>.0902</td>
</tr>
<tr>
<td>4</td>
<td>.0113</td>
<td>.0114</td>
<td>.0783</td>
<td>.0850</td>
</tr>
<tr>
<td>5</td>
<td>.0098</td>
<td>.0099</td>
<td>.0761</td>
<td>.0824</td>
</tr>
<tr>
<td>6</td>
<td>.0096</td>
<td>.0097</td>
<td>.0747</td>
<td>.0807</td>
</tr>
<tr>
<td>7</td>
<td>.0095</td>
<td>.0096</td>
<td>.0738</td>
<td>.0797</td>
</tr>
<tr>
<td>8</td>
<td>.0094</td>
<td>.0095</td>
<td>.0731</td>
<td>.0789</td>
</tr>
<tr>
<td>9</td>
<td>.0093</td>
<td>.0094</td>
<td>.0725</td>
<td>.0782</td>
</tr>
<tr>
<td>10</td>
<td>.0093</td>
<td>.0094</td>
<td>.0721</td>
<td>.0777</td>
</tr>
<tr>
<td>15</td>
<td>.0092</td>
<td>.0093</td>
<td>.0699</td>
<td>.0752</td>
</tr>
<tr>
<td>20</td>
<td>.0092</td>
<td>.0093</td>
<td>.0692</td>
<td>.0743</td>
</tr>
<tr>
<td>30</td>
<td>.0091</td>
<td>.0092</td>
<td>.0685</td>
<td>.0735</td>
</tr>
<tr>
<td>40</td>
<td>.0091</td>
<td>.0092</td>
<td>.0680</td>
<td>.0730</td>
</tr>
<tr>
<td>$\infty$</td>
<td>.0090</td>
<td>.0091</td>
<td>.0678</td>
<td>.0727</td>
</tr>
</tbody>
</table>

on the posterior probability of $H_0$ for $\pi_0 = 1/2$, and $B$ is the lower bound on the Bayes factor.

Note the dramatic discrepancy between $\alpha$ and the lower bounds. When $p = 1$ and $\alpha = .05$, for instance, $B = .4092$; thus the weighted likelihood of $H_1$ is at most $2^{1/2}$ times that of $H_0$. A likelihood ratio of $2^{1/2}$ 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 $H_0$. And a $P$-value of $\alpha = .01$, often considered very strong evidence against $H_0$, corresponds to $B = .1227$, 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 \( \theta \), however, it is not possible to use such a class. Nevertheless, it is possible to find reasonable lower bounds on \( B \) and \( P(H_0|x) \). 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 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 \( \theta_0 \); this is illustrated in subsection 3.2.3. A third possibility is to consider the class of all conjugate densities having mean \( \theta_0 \); this is illustrated in subsection 3.2.4. (Choosing \( \Gamma \) 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 \( \theta \). 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 that there is no natural symmetry in this problem unless \( \theta_0 = 1/2 \). This even makes difficult the definition of a P-value. We will use the "intrinsic significance level" obtained by choosing \( T(X) = 1/f(X|\theta_0) \) in (1). This leads to defining the P-value of an observation \( x \) as

\[
\alpha = P_{\theta=\theta_0}(\{y : f(y|\theta_0) \leq f(x|\theta_0)\}).
\]

3.2.2 A Natural Transformed Symmetric Class

When \( \theta_0 = 1/2 \) it is easy to define a notion of prior symmetry; simply choose all densities which are symmetric about the point \( 1/2 \). 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 \( \theta \). 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

\[
u(\theta) = \frac{\theta - \theta_0}{\sqrt{\theta(1-\theta)}}.
\]

(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 \(\theta\), yields the density

\[
g(\theta) = h(u(\theta)) \frac{du(\theta)}{d\theta}.
\]  

( Since \(u(\theta)\) is an increasing function of \(\theta\), the Jacobian can be written \(\frac{du(\theta)}{d\theta}\) instead of \(|\frac{du(\theta)}{d\theta}|\).) Note that \(\theta_0\) is the median of the density \(g\), since

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

Let \(G_u\) 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.**

\[
P_{US} = \inf_{g \in G_u} P(H_0|x) = (1 + \frac{1}{\pi_0} \sup_k \frac{1}{2k} \int_{-k}^k l(u) du) \left(\frac{1}{(n)(\theta_0^x)(1 - \theta_0)^{n-x}}\right)^{-1},
\]

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

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

**Proof:** See the Appendix. \(\square\)

**Remark:** The supremum over \(k\) is actually attained for some \(k\), since \(\frac{1}{2k} \int_{-k}^k l(u) du\) converges to \(l(0) > 0\) as \(k \to 0\) and vanishes at \(\infty\).

It is a simple numerical computation to obtain \(P_{US}\) using Theorem 3. For \(\pi_0 = \frac{1}{2}\) and selected values of \(\theta_0, n\) and \(x\), these lower bounds are tabulated, along with the corresponding \(P\)-values, in Table 5.

The differences between \(\alpha\) and \(P_{US}\) are of the same magnitude here as in the normal situation of Section 3. For \(\alpha = .01, .05, .10\), the values of \(P_{US}\) from Table 4 (in one dimension) were .109, .290, and .392, respectively, all similar to the
corresponding \( P_{US} \) 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 \( \alpha \) and \( P(H_0|x) \). 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 \( \alpha \) and \( P \).

<table>
<thead>
<tr>
<th>( \alpha )</th>
<th>( n )</th>
<th>( x )</th>
<th>( p^0 )</th>
<th>( \maximizing k )</th>
<th>( P_{US} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0090</td>
<td>50</td>
<td>11</td>
<td>.40</td>
<td>4.4851</td>
<td>0.0794</td>
</tr>
<tr>
<td>0.0100</td>
<td>20</td>
<td>9</td>
<td>.20</td>
<td>3.2502</td>
<td>0.1201</td>
</tr>
<tr>
<td>0.0101</td>
<td>20</td>
<td>14</td>
<td>.40</td>
<td>4.3694</td>
<td>0.0770</td>
</tr>
<tr>
<td>0.0106</td>
<td>10</td>
<td>7</td>
<td>.30</td>
<td>4.3190</td>
<td>0.0951</td>
</tr>
<tr>
<td>0.0118</td>
<td>20</td>
<td>4</td>
<td>.50</td>
<td>5.1514</td>
<td>0.0642</td>
</tr>
<tr>
<td>0.0128</td>
<td>10</td>
<td>4</td>
<td>.10</td>
<td>2.8777</td>
<td>0.1244</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>16</td>
<td>.20</td>
<td>2.5784</td>
<td>0.2786</td>
</tr>
<tr>
<td>0.0507</td>
<td>25</td>
<td>3</td>
<td>.30</td>
<td>4.3252</td>
<td>0.2210</td>
</tr>
<tr>
<td>0.0534</td>
<td>35</td>
<td>12</td>
<td>.20</td>
<td>2.5329</td>
<td>0.2867</td>
</tr>
<tr>
<td>0.0541</td>
<td>40</td>
<td>10</td>
<td>.40</td>
<td>3.2164</td>
<td>0.2741</td>
</tr>
<tr>
<td>0.0920</td>
<td>50</td>
<td>9</td>
<td>.10</td>
<td>2.0479</td>
<td>0.3687</td>
</tr>
<tr>
<td>0.0960</td>
<td>15</td>
<td>6</td>
<td>.20</td>
<td>2.2632</td>
<td>0.3496</td>
</tr>
<tr>
<td>0.0987</td>
<td>30</td>
<td>20</td>
<td>.50</td>
<td>2.8149</td>
<td>0.3257</td>
</tr>
<tr>
<td>0.1011</td>
<td>10</td>
<td>7</td>
<td>.40</td>
<td>3.1644</td>
<td>0.2969</td>
</tr>
<tr>
<td>0.1021</td>
<td>45</td>
<td>19</td>
<td>.30</td>
<td>2.3423</td>
<td>0.3624</td>
</tr>
<tr>
<td>0.1094</td>
<td>10</td>
<td>2</td>
<td>.50</td>
<td>3.7602</td>
<td>0.2780</td>
</tr>
</tbody>
</table>

### 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 \( \theta_0 \) is quite reasonable. In fact, the class of symmetric unimodal densities, about the point \( \theta_0 \) 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 $\theta_0$. Then
\[ G_1 = \{ g : \int_0^{\theta_0} g(\theta) \, d\theta = 1/2 \}. \]

Let $G_0$ be the set of all unimodal densities on the interval $[0, 1]$ with median and mode at $\theta_0$. That is,
\[ G_0 = \{ g : \int_0^{\theta_0} g(\theta) \, d\theta = 1/2; \text{ and } g \text{ is nondecreasing on } [0, \theta_0], \text{ nonincreasing on } [\theta_0, 1] \}. \]

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

**Proof:** See the Appendix. \qed

**Theorem 5.**
\[
\sup_{g \in G_0} \int_0^1 \binom{n}{x} \theta^x(1 - \theta)^{n-x} g(\theta) \, d\theta = \frac{1}{2}\left[ \binom{n}{x} \theta_0^x(1 - \theta_0)^{n-x} + \frac{1}{k} \int_{a(k)}^{b(k)} \binom{n}{x} \theta^x(1 - \theta)^{n-x} \, d\theta \right],
\]

where
\[
\{a(k), b(k)\} = \begin{cases} 
\{\theta_0 - k, \theta_0\} & \text{if } \frac{x}{n} \leq \theta_0; \\
\{\theta_0, \theta_0 + k\} & \text{if } \frac{x}{n} > \theta_0.
\end{cases}
\]

**Proof:** See the appendix. \qed

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 $P_{SM*}$, $P_{SM*\theta_0}$ were computed for some selected values of $n, x, \theta_0$ 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.
Table 6: Lower Bounds for Densities with Specified Median, Mode

| $\alpha$  | $n$ | $x$ | $p^0$ | $P_{SMx}$ | $P_{SM_{x|M_0}}$ |
|-----------|-----|-----|-------|-----------|------------------|
| 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           |

3.2.4 Conjugate Class with Mean $\theta_0$

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

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

and the mean is $\frac{a}{a+b}$. Therefore, the class of conjugate priors to consider is that which consists of all $Beta(a, b)$ distributions such that $\frac{a}{a+b} = \theta_0$, 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.

\[ P_C = \inf_{\theta \in \Theta} P(H_0 | x) \]

\[ = \left[ 1 + \frac{(1 - \pi_0)}{\pi_0} \sup_{c > 0} \frac{\Gamma(c) \Gamma(x + \theta_0) \Gamma(n - x + c(1 - \theta_0))}{\Gamma(c\theta_0) \Gamma(c(1 - \theta_0)) \Gamma(n + c \theta_0) \Gamma(1 - \theta_0)^{n-x}} \right]^{-1}. \]

Proof: See the appendix. \qed

The lower bounds on the posterior probability of \( H_0 \) given \( x \) can easily be computed numerically, for specified values of \( \theta_0, n, x \) and \( \pi_0 \). For \( \pi_0 = 1/2 \) and some selected values of \( n, x \), these lower bounds, denoted \( P_C \), are tabulated in Table 7 along with the corresponding P-values. Again, we defer discussion until Section 3.2.5.

Table 7: Lower Bounds for Conjugate Densities

<table>
<thead>
<tr>
<th>( \alpha )</th>
<th>( n )</th>
<th>( x )</th>
<th>( p^\theta )</th>
<th>( P_C )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0090</td>
<td>50</td>
<td>11</td>
<td>.40</td>
<td>0.0981</td>
</tr>
<tr>
<td>0.0100</td>
<td>20</td>
<td>9</td>
<td>.20</td>
<td>0.1771</td>
</tr>
<tr>
<td>0.0101</td>
<td>20</td>
<td>14</td>
<td>.40</td>
<td>0.1064</td>
</tr>
<tr>
<td>0.0118</td>
<td>20</td>
<td>4</td>
<td>.50</td>
<td>0.0858</td>
</tr>
<tr>
<td>0.0120</td>
<td>45</td>
<td>10</td>
<td>.10</td>
<td>0.2211</td>
</tr>
<tr>
<td>0.0493</td>
<td>50</td>
<td>16</td>
<td>.20</td>
<td>0.3313</td>
</tr>
<tr>
<td>0.0505</td>
<td>15</td>
<td>1</td>
<td>.30</td>
<td>0.1956</td>
</tr>
<tr>
<td>0.0507</td>
<td>25</td>
<td>3</td>
<td>.30</td>
<td>0.2414</td>
</tr>
<tr>
<td>0.0541</td>
<td>40</td>
<td>10</td>
<td>.40</td>
<td>0.3016</td>
</tr>
<tr>
<td>0.0556</td>
<td>15</td>
<td>4</td>
<td>.10</td>
<td>0.4223</td>
</tr>
<tr>
<td>0.0960</td>
<td>15</td>
<td>6</td>
<td>.20</td>
<td>0.4123</td>
</tr>
<tr>
<td>0.0980</td>
<td>25</td>
<td>5</td>
<td>.10</td>
<td>0.4779</td>
</tr>
<tr>
<td>0.0987</td>
<td>30</td>
<td>20</td>
<td>.50</td>
<td>0.3565</td>
</tr>
<tr>
<td>0.1000</td>
<td>35</td>
<td>15</td>
<td>.30</td>
<td>0.4328</td>
</tr>
<tr>
<td>0.1011</td>
<td>10</td>
<td>7</td>
<td>.40</td>
<td>0.3458</td>
</tr>
<tr>
<td>0.1094</td>
<td>10</td>
<td>2</td>
<td>.50</td>
<td>0.3163</td>
</tr>
</tbody>
</table>
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 “$\alpha$” 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 “$\alpha$”; 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.

<table>
<thead>
<tr>
<th>$\alpha$</th>
<th>Normal bounds</th>
<th>range of binomial lower bounds</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>C</td>
</tr>
<tr>
<td>.01</td>
<td>.11</td>
<td>.09–.22</td>
</tr>
<tr>
<td>.05</td>
<td>.29</td>
<td>.20–.42</td>
</tr>
<tr>
<td>.10</td>
<td>.39</td>
<td>.32–.48</td>
</tr>
</tbody>
</table>

Clearly the conjugate priors bounds tend to be the largest, followed by the US and SMeMo bounds which 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 as 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 which attempt to spread mass on both sides of $\theta_0$, 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.

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 1/2. 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 which are representable as exact point nulls. The argument continues with voicing of the suspicion that the conflict between P-values and \( P(H_0|x) \) is due to the assignment of the positive mass \( \pi_0 \) 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, \( \pi_0 \), assigned to \( \theta = \theta_0 \) is simply the mass corresponding to the precise hypothesis \( H_0 : |\theta - \theta_0| \leq \epsilon \). For true precise hypotheses, such as \( H_0 : \text{"Vitamin C has negligible effect on the common cold"} \), 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 \( \pi_0 \), and can be thought of as the evidence (in a likelihood ratio sense) provided by the data against \( H_0 \). And we have seen that \( B \) and \( P(H_0|x) \) 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 \( H_0 : \theta \leq \theta_0 \) versus \( H_1 : \theta > \theta_0 \), where there is no particularly strong belief that \( \theta \) is near \( \theta_0 \). 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 which 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 $H_0 : |\theta - \theta_0| < \epsilon$ versus $H_1 : \theta > \theta_0 + \epsilon$) 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 $P(H_0|x)$, 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 $f(x|\theta)$—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 $X \sim N(\theta, \sigma^2/n)$, $\sigma^2$ known, and that it is desired to test

$$H_0 : |\theta - \theta_0| \leq \epsilon \text{ against } H_1 : |\theta - \theta_0| > \epsilon.$$ 

Let $T = \sqrt{n}(X - \theta_0)/\sigma$, and $t_\alpha$ be the critical value (which depends on $\epsilon$) such that

$$\alpha = P_{\theta_0+\epsilon}(|T(X)| \geq t_\alpha).$$
Figure 1: $P(H_0|x)$ For Interval Nulls
Were we to observe $T = t_\alpha$, we would report $\alpha$ as the P-value. To compare this with the posterior probability when $t_\alpha$ is observed, consider priors $\pi$ for $\theta$ which (i) are symmetric about $\theta_0$; (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

$$P(H_0|x) \equiv \inf_{\pi \in US} P_\pi(H_0|x)$$

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

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

When $\epsilon^*$ (and hence $\epsilon$) is zero, the test is that of a point null, and $P(H_0|x)$ 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 $\epsilon^* \to \infty$ (i.e., $H_0$ becomes more diffuse) $P(H_0|x) \to \alpha$. (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 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 $\theta$ from $\theta_0$.

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 $\theta_0$. Only measures such as $B$ or $P(H_0|x)$ 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 \( \theta_0 \), which is outside, say, a 95% confidence set \( C \), is often not too much smaller than the "average" likelihood of \( \theta \) in \( C \), and there is then no strong justification for rejecting \( \theta_0 \).

For hypothesis testing problems with a special point \( \theta_0 \) (or special small interval \( (\theta_0 - \epsilon, \theta_0 + \epsilon) \)), we would urge reporting both the Bayes factor, \( B \), against \( \theta_0 \) and a confidence or credible region, \( C \). The Bayes factor communicates the evidence in the data against \( \theta_0 \), 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 \( \chi^2 \)-test of fit.

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

\[
H_0 : F = F_0 \text{ versus } H_1 : F \neq F_0,
\]
where \( F_0 \) is a specified distribution. The standard test procedure for this problem is the \( \chi^2 \) test of fit.

\textbf{\( \chi^2 \)-Test Procedure:} 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 \( z = (z_1, \cdots, z_m)' \) denote these frequencies; thus \( z_i \) = the number of \( X_i \)'s in \( (a_{i-1}, a_i] \). Let

\[
\begin{align*}
p_i &= F'(a_i) - F'(a_{i-1}) = P_F(a_{i-1} < X \leq a_i), \\
p_i^0 &= F_0'(a_i) - F_0'(a_{i-1}) = P_{F_0}(a_{i-1} < X \leq a_i),
\end{align*}
\]

and

\[
\begin{align*}
p &= (p_1, \cdots, p_m)', \\
p^0 &= (p_1^0, \cdots, p_m^0)'.
\end{align*}
\]

Then the \( \chi^2 \) test procedure is to calculate the test statistic,

\[
t = \sum_{i=1}^{m} \frac{(z_i - N p_i^0)^2}{N p_i^0},
\]

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

\[
\alpha = P(T \geq t).
\]

But reducing the observations to the vector \( z \) of frequencies really means that we are testing

\( H_0 : p = p^0 \) versus \( H_1 : p \neq p^0, \)

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

It is shown in Delampady and Berger (1987) that lower bounds on \( B \) and \( P(H_0|z) \), 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 non-parametric 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, \( \alpha \), has the actual frequentist interpretation that a long series of \( \alpha \)-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 \( t \equiv 2 \) 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 \( t \equiv 2 \), 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 which 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

\[
E = \{ \text{possible data } x: |T(x)| \geq |T(x_0)| \}.
\]

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 \( P(H_0 | E) \) is frequently very close to the P–value, and hence much smaller than \( P(H_0 | x_0) \). 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 p. 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 \( X \sim \text{Binomial} (n, \theta) \) 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

\[
\frac{f(220|0.5)}{\sup_{\theta} f(220|\theta)} = \frac{f(220|0.5)}{f(220|0.55)} = 0.135 \approx \frac{1}{7.5},
\]

where \( f(x|\theta) \) 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 \( P(H_0|x) \) at (or very close to) \( \theta_0 \) 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 \( \theta \) being the mean difference and \( H_0 : \theta = 0 \) corresponding (typically) to both drugs being completely ineffective, then the size of \( P(H_0|x) \) 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 \( \theta_0 \) 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) and 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 yet been formulated; while 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 which is believed to be true apriori, in the sense that there is a much sharper concentration of prior beliefs about this model than elsewhere (in "model space"). And P-values are not necessarily terrible, as measures of evidence, in the absence of such apriori 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 $\theta$ 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 which are being addressed in this paper.

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

**Method 1. Use the Lower Bounds $B$ and $P$**

It would be a substantial improvement over $P$-values to report, as the evidence against $H_0$, the lower bound $B$ in (8) or $P(H_0|x)$ in (9) (the latter for, say, $\pi_0 = \frac{1}{2}$) 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, since 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 towards $H_1$, 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 $\theta_0$. (Conjugate densities may not exist or be well-defined; but any low-dimensional class of priors having mean, or median, equal to $\theta_0$ 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 $g$ which have mean $\theta_0$ 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.
Method 2. Use A Conventional Prior

Though vastly superior to P-values, B and 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, B and 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, B and P are suspect (though 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 B or P.

Note, especially, the use of $P(H_0|x)$ for $\pi_0 = \frac{1}{2}$ 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," while full subjective Bayesian analysis is decried 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 $P(H_0|x)$ 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: $\pi_0$, the probability of $H_0$, and (say) the quartiles of $g$, the conditional prior distribution assuming that $H_0$ is false. (And $\pi_0$ 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 $X_1, \ldots, X_n$, distributed as i.i.d. $N(\theta, \sigma^2)$ random variables, $\theta$ and $\sigma^2$ 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 $\mu$ and $\tau$, where $\mu$ is the median of $g$ and $\mu \pm \tau$ are the quartiles. (Thus $\mu$ would be the guess for $\theta$, 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 $\theta$ and $\sigma^2$, respectively), one can construct a contour graph of $B$ as a function of $\mu$ and $\tau$. See Figure 2. Thus an individual who believes that if $H_1$ is true then $\theta$ 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 $\tau$ 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 $\tau$ is about 0.55, even though the P−value was 0.05.

ACKNOWLEDGMENT

Research for the article was supported by the National Science Foundation, Grant DMS-8401996, 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.
Figure 2: Contours of $B$ as a Function of $|\mu - \bar{x}|$ and $\tau$

When $\bar{x} = 20.93$, $s = 37.71$, and $n = 15$. 

$|\mu - \bar{x}|$

$\tau$
References


Appendix

Proof of Theorem 1: Since $g_0$ is unimodal and symmetric about $\theta_0$, it can be represented as a mixture of uniform densities,

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

so that

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

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

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

This can be rewritten

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

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

Next, observe that, on $\overline{\Omega}$,

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

so that

$$ \int_{\overline{\Omega}} f(\bar{x}|\theta)g_1(\theta) d\theta = \frac{1}{(1 - \lambda)} \int_{\overline{\Omega}} f(\bar{x}|\theta)g(\theta) d\theta = \left[ m_g(\bar{x}) - \int_\Omega f(\bar{x}|\theta)g(\theta) d\theta \right] = m_g(\bar{x})(1 + \psi_2), $$

where

$$ \psi_2 = \frac{\lambda}{(1 - \lambda)} \left[ 1 - \frac{1}{m_g(\bar{x})} \int_{\overline{\Omega}} f(\bar{x}|\theta)(\frac{g(\theta)}{\lambda}) d\theta \right]. $$
Observing that \( g(\theta)/\lambda \) is a unimodal symmetric density on \( \Omega \), it follows as in the argument leading to (23) that

\[
\int_{\mathbb{R}} f(\theta|\theta)(\frac{g(\theta)}{\lambda}) d\theta = f(\theta|\theta_0) (1 + \psi_1^*),
\]

where \( 0 \leq \psi_1^* \leq \gamma \). Hence

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

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

\[
B = \frac{f(\theta|\theta_0)(1 + \psi_1)}{m_\theta(x)(1 + \psi_2)} = \hat{B}(1 + \rho),
\]

where

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

Algebra, together with the bounds \( 0 \leq \hat{B} \leq (1 + \gamma)^{-1} \), yield (17). \( \square \)

Proof of Theorem 3:

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

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 \( U \) be the class of all unimodal symmetric densities \( h \). Then

\[
\sup_{g \in G} \int_0^1 \binom{n}{x} \theta^x(1 - \theta)^{n-x} g(\theta) d\theta = \sup_{h \in U} \int_{-\infty}^{\infty} \binom{n}{x} \theta(u)^x(1 - \theta(u))^{n-x} h(u) du
\]

\[
= \sup_k \frac{1}{2k} \int_{-k}^k \binom{n}{x} \theta(u)^x(1 - \theta(u))^{n-x} du, \tag{26}
\]

44
from Theorem 2, which proves the result.

\[ \sup_{g \in \mathcal{G}} \int_0^1 \left( \frac{n}{x} \right) \theta^z (1 - \theta)^{n-x} g(\theta) d\theta = \sup_{h_1} \int_0^{\theta_0} \left( \frac{n}{x} \right) p^z (1 - p)^{n-x} h_1(\theta) d\theta \]

\[ + \sup_{h_2} \int_{\theta_0}^1 \left( \frac{n}{x} \right) \theta^z (1 - \theta)^{n-x} h_2(\theta) d\theta, \]

where \( h_1 \geq 0 \) on \([0, \theta_0]\), \( h_2 \geq 0 \) 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 \( \frac{\theta}{n} \leq \theta_0 \). Then

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

\[ \sup_{h_2} \int_{\theta_0}^1 \left( \frac{n}{x} \right) \theta^z (1 - \theta)^{n-x} h_2(\theta) d\theta = \frac{1}{2} \left( \frac{n}{x} \right) \theta_0^z (1 - \theta_0)^{n-x}, \]

since \( \left( \frac{n}{x} \right) \theta^z (1 - \theta)^{n-x} \) has its maximum at \( \hat{\theta} = \frac{\theta}{n} \) and decreases monotonically on both sides. In the other case, where \( \frac{\theta}{n} > \theta_0 \), the roles of \( h_1 \) and \( h_2 \) are reversed. \( \square \)

**Proof of Theorem 5.** Assume that \( \frac{\theta}{n} \leq \theta_0 \); the other case is similar. Let

\[ I = \sup_{g \in \mathcal{G}} \int_0^1 \left( \frac{n}{x} \right) \theta^z (1 - \theta)^{n-x} g(\theta) d\theta. \]

Then

\[ I = \sup_{h_1} \int_0^{\theta_0} \left( \frac{n}{x} \right) \theta^z (1 - \theta)^{n-x} h_1(\theta) d\theta + \sup_{h_2} \int_{\theta_0}^1 \left( \frac{n}{x} \right) \theta^z (1 - \theta)^{n-x} h_2(\theta) d\theta, \]

where \( h_1 \geq 0 \) and nondecreasing on \([0, \theta_0]\), \( h_2 \geq 0 \) 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

\[ I = \sup_{g \in \mathcal{G}} \int_0^1 \left( \frac{n}{x} \right) \theta^z (1 - \theta)^{n-x} g(\theta) d\theta \]

\[ = \sup_{h_1} \int_0^{\theta_0} \left( \frac{n}{x} \right) \theta^z (1 - \theta)^{n-x} h_1(\theta) d\theta + \frac{1}{2} \left( \frac{n}{x} \right) \theta_0^z (1 - \theta_0)^{n-x}, \]

45
since $\binom{n}{x} \theta^{x}(1 - \theta)^{n-x}$, in this case, is decreasing monotonically on both sides of $\frac{x}{n} \leq \theta_0$. 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

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

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

$$\sup_{h_3} \int_{0}^{2\theta_0} g(x, n, \theta) h_3(\theta) d\theta = \frac{1}{2r} \int_{\theta_0-r}^{\theta_0+r} g(x, n, \theta) d\theta$$

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

since unimodal symmetric densities are mixtures of uniforms. \hfill \Box

**Proof of Theorem 6.**

$$\inf_{\pi \in \mathcal{C}} P^\pi_\pi(H_0|x) = \left[1 + \frac{(1 - \pi_0)}{\pi_0} \frac{\sup_{\pi \neq \pi_0} \int \frac{f(x|\pi) \pi(\pi)}{f(x|\theta_0)} d\pi}{\sup_{\pi \neq \pi_0} \int \frac{f(x|\pi) \pi(\pi)}{f(x|\theta_0)} d\pi} \right]^{-1}.$$

Now for $\theta \neq \theta_0$

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

Therefore

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

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

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

$$= \sup_{c > 0} \frac{\Gamma(c)}{\Gamma(c\theta_0) \Gamma(c(1 - \theta_0))} \frac{\Gamma(x+c\theta_0) (n-x+c(1 - \theta_0))}{\Gamma(n+c)}$$

which proves the Theorem. \hfill \Box