RECONCILING BAYESIAN AND FREQUENTIST EVIDENCE IN THE ONE-SIDED TESTING PROBLEM

George Casella
Cornell University

Roger L. Berger
North Carolina State University

Institute of Statistics Mimeo Series No. 1657
April, 1985

Reconciling Bayesian and Frequentist Evidence in the One-Sided Testing Problem

George Casella[1]
Cornell University

Roger L. Berger[2]
North Carolina State University

[1]Research supported by National Science Foundation Grant No. MCS81-02541

[2]Research supported in part by U. S. Army Research Office Grant DAAG 29-82-K-0168 at Florida State University. The United States Government is authorized to reproduce and distribute reprints for governmental purposes

George Casella is Associate Professor, Biometrics Unit, Cornell University, Ithaca, NY 14853. Roger L. Berger is Associate Professor, Department of Statistics, North Carolina State University, Raleigh, NC 27695. This research was done while Professor Casella was Visiting Associate Professor, Department of Statistics, North Carolina State University. The authors thank Raymond J. Carroll for expressing curiosity about the one-sided problem, and to Mark Berliner for helpful discussions.

Abstract

For the one-sided hypothesis testing problem it is shown that it is possible to reconcile Bayesian evidence against H_0, expressed in terms of the posterior probability that H_0 is true, with frequentist evidence against H_0, expressed in terms of the p-value. In fact, for many classes of prior distributions it is shown that the infimum of the Bayesian posterior probability of H_0 is either equal to or bounded above by the p-value. The results are in direct contrast to recent work of Berger and Sellke (1985) in the two-sided (point null) case, where it was found that the p-value is much less than the Bayesian infimum. Some comments on the point null problem are given.

I. Introduction

In the problem of hypothesis testing, 'evidence' can be thought of as a post-experimental (data-based) evaluation of the tenability of the null hypothesis, H_0. To a Bayesian, evidence takes the form of the posterior probability that H_0 is true, while to a frequentist, evidence takes the form of the p-value, or significance level, of the test. If the null hypothesis consists of a single point, it has long been known that these two measures of evidence can greatly differ. The famous paper of Lindley (1957) illustrates the possible discrepancy in the normal case.

The question of reconciling these two measures of evidence has been treated in the literature. For the most part, the two-sided (point null) problem has been treated, and the major conclusion has been that the p-value tends to overstate the evidence against H_0 (that is, the p-value tends to be smaller than a Bayesian posterior probability). Many references can be found in Shafer (1982). However Pratt (1965) does state that in the one-sided testing problem, the p-value is approximately equal to the posterior probability of H_0.

A slightly different approach to the problem of reconciling evidence was taken by DeGroot (1973). Working in a fairly general setting, DeGroot constructs alternative distributions and finds improper priors for which the p-value and posterior probability match. DeGroot assumes that the alternative distributions are stochastically ordered which, although he does not explicitly state it, essentially puts him in the one-sided testing problem.

Dickey (1977), in the two-sided problem, considers classes of priors, and examines the infimum of Bayesian evidence against H_0. As a measure of Bayesian evidence Dickey uses the "Bayes factor," which is closely related to the posterior probability of H_0. He also concludes that the p-value overstates the evidence against H_0, even when compared to the infimum of Bayesian evidence.

A recent paper by J. Berger and T. Sellke (1985) has approached the problem of reconciling evidence in a manner similar to Dickey's approach. For the Bayesian measure of evidence they consider the infimum, over a class of priors, of the posterior probability that H_0 is true. For many classes of priors it turns out that this infimum is much greater than the frequentist p-value, leading Berger and Sellke to conclude that, "... significance levels can be highly misleading measures of the evidence provided by the data against the null hypothesis."

Although their arguments are compelling, and may lead one to question the worth of p-values, their analyses are restricted to the problem of testing a point null hypothesis. If, in fact, the p-value is a misleading measure of evidence, discrepancies with Bayesian measures should emerge in other hypothesis testing situations.

The point null hypothesis is perhaps the most used and misused statistical technique. In particular, in the location parameter problem, the point null hypothesis is more the mathematical convenience rather than the statistical method of choice. Few experimenters, of whom we are aware, want to conclude "there is a difference." Rather, they are looking to conclude "the new treatment is better." Thus, for the most part, there is a direction of interest in almost any experiment, and saddling an experimenter with a two-sided test will not lead to the desired conclusions.

In this paper we consider the problem of reconciling evidence in the one-sided testing problem. We find, in direct contrast to the results of Berger and Sellke, that evidence can be reconciled. That is, for many classes of priors, the infimum of the Bayes posterior probability that H_0 is true is either equal to or bounded above by the p-value.

In Section 2 we present some necessary preliminaries, including the classes of priors we are considering and how they relate to those considered in the two-sided problem. Section 3 contains the main results concerning the relationship between Bayesian and frequentist evidence. Section 4 considers classes of priors that are biased toward H_0, and Section 5 contains comments about testing a point null hypothesis.

2. Preliminaries

We consider testing the hypotheses

$$H_0: \quad \theta \leq 0 \quad \text{vs.} \quad H_1: \quad \theta > 0 \tag{2.1}$$

based on observing $X = x$, where X has location density $f(x-\theta)$. Throughout this paper, unless explicitly stated, we assume that

i) $f(\cdot)$ is symmetric about zero

ii) $f(x - \theta)$ has monotone likelihood ratio (mlr).

Recall that i) and ii) imply that $f(\cdot)$ is unimodal.

If $X = x$ is observed, a frequentist measure of evidence against H_0 is given by the p-value

$$p(x) = P(X \geq x|\theta = 0) = \int_x^\infty f(t)dt \ . \tag{2.2}$$

A Bayesian measure of evidence, given a prior distribution $\pi(\theta)$, is the probability that H_0 is true given $X = x$,

$$P(H_0|x) = P(\theta \leq 0|x) = \frac{\int_{-\infty}^{0} f(x-\theta)\pi(\theta)d\theta}{\int_{-\infty}^{\infty} f(x-\theta)\pi(\theta)d\theta} \ . \tag{2.3}$$

Our major point of concern is whether these two measures of evidence can be reconciled, that is, can the p-value, in some sense, be regarded as a Bayesian measure of evidence. Since the p-value is based on the objective frequentist model, it seems apparent that, if reconciliation is possible, it must be based on impartial prior distributions. By impartial we mean that the prior distribution gives equal weight to both the null and alternative hypotheses.

Four reasonable classes of distributions are given by

$$
\begin{aligned}
G_A &= \{\text{all distributions giving mass } \tfrac{1}{2} \text{ to } (-\infty,0] \text{ and } (0,\infty)\} \\
G_S &= \{\text{all distributions symmetric about zero}\} \\
G_{US} &= \{\text{all unimodal distributions symmetric about zero}\} \\
G_{NOR} &= \{\text{all normal } (0,\tau^2) \text{ distributions, } 0 \leq \tau^2 < \infty\}.
\end{aligned}
\tag{2.4}
$$

For any class of priors, we can obtain a reasonably objective Bayesian measure of evidence by considering $\inf P(H_0|x)$, where the infimum is taken over a chosen class of priors. We can then examine the relationship between this infimum and $p(x)$. If there is agreement, we can conclude that Bayesian and frequentist evidence can be reconciled.

This development is, of course, similar to that of Berger and Sellke (1985), who consider the two-sided hypothesis test H_0: $\theta = 0$ vs. H_1: $\theta \neq 0$, using priors of the form

$$
\pi(\theta) = \begin{cases} \pi_0 & \text{if } \theta = 0 \\ (1-\pi_0)g(\theta) & \text{if } \theta \neq 0 \,, \end{cases}
$$

and allow $g(\cdot)$ to vary within a class of distributions, similar to the classes in (2.4). For any numerical calculations they choose $\pi_0 = \frac{1}{2}$, asserting that this provides an impartial prior distribution. We will return to this question in Section 5.

For testing H_0: $\theta \leq 0$ vs. H_1: $\theta > 0$, we will mainly be concerned with evidence based on observing $x > 0$. For $x < 0$, $p(x) > \frac{1}{2}$ and $\inf P(H_0|x) = \frac{1}{2}$, where the infimum is over any of the classes in (2.4). Thus, if $x < 0$, neither a frequentist nor a Bayesian would consider the data as having evidence against H_0, so ther is, in essence, nothing to be reconciled.

3. Symmetric Prior Distributions

In this section we consider prior distributions contained in the classes given in (2.4). Our goal is to calculate $\inf P(H_0|x)$ for each of these classes, and relate the answer to $p(x)$. In some cases we do not calculate $\inf P(H_0|x)$ exactly, but rather obtain an upper bound on the infimum. This is accomplished by calculating the infimum exactly for smaller classes of distributions.

For the one-sided testing problem, the class G_A is too large to be of use, as the following theorem shows.

Theorem 3.1: For the hypotheses in (2.1), if $x > 0$, then

$$\inf_{\pi \varepsilon G_A} P(H_0|x) = 0 .$$

Proof: Consider a sequence of priors

$$\pi_K(\theta) = \begin{cases} \frac{1}{2} & \text{if} \quad \theta = -k \\ g(\theta) & \text{if} \quad \theta > 0 \end{cases}$$

where $\int_0^\infty g(\theta)d\theta = \frac{1}{2}$. Then

$$P(H_0|x) = \frac{f(x + k)}{f(x+k) + \int_0^\infty f(x-\theta)g(\theta)d\theta} ,$$

and it is easy to see that $\lim_{k\to\infty} P(H_0|x) = 0$, establishing the result. □

Although we cannot obtain explicit answers for the class G_S, we can get some interesting results for the smaller class contained in G_S,

$$G_{2PS} = \{\text{all two-point distributions symmetric about } 0\}.$$

For $\pi\varepsilon G_{2PS}$, we have $\pi(\theta) = \frac{1}{2}$ if $\theta = \pm k$, and hence

$$P(H_0|x) = \frac{f(x+k)}{f(x-k)+f(x+k)} \quad . \tag{3.1}$$

Since f has mlr, it follows that, for $x > 0$, $P(H_0|x)$ is decreasing in k. Therefore, for $x > 0$,

$$\inf_{\pi\varepsilon G_{2PS}} P(H_0|x) = \lim_{k\to\infty} \frac{f(x+k)}{f(x-k)+f(x+k)} \quad . \tag{3.2}$$

The following theorem establishes that p(x) is an upper bound on this quantity.

<u>Theorem 3.2</u>: For the hypotheses in (2.1), if $x > 0$,

$$\inf_{\pi\varepsilon G_S} P(H_0|x) \le \inf_{\pi\varepsilon G_{2PS}} P(H_0|x) \le p(x) \tag{3.3}$$

<u>Proof</u>: The inequality

$$\frac{f(x+k)}{f(x-k)+f(x+k)} \le p(x)$$

is equivalent to

$$f(k-x)p(x) - f(k+x)(1-p(x)) \ge 0 \ .$$

To establish (3.3), it suffices to establish the weaker inequality,

$$\lim_{k\to\infty} [f(k-x)p(x) - f(k+x)(1-p(x))] \ge 0 \ . \tag{3.4}$$

Now

$$f(k-x)p(x) - f(k+x)(1-p(x))$$
$$= \int_0^\infty [f(k-x)f(x+z) - f(k+x)f(-x+z)]dz \ . \tag{3.5}$$

Let $x_1 = k-x$, $x_2 = k+x$, $\theta_1 = 0$, and $\theta_2 = k-z$. The integrand is of the form $f(x_1-\theta_1)f(x_2-\theta_2) - f(x_2-\theta_1)f(x_1-\theta_2)$. Since $x > 0$, $x_2 > x_1$, and $\theta_2 \geq \theta_1$ if and only if $k \geq z$. Thus the fact that f has mlr implies that the integrand is nonnegative if $z \leq k$ and nonpositive if $z > k$. It also follows from the assumptions on f that $\lim_{k\to\infty} f(k-x) = \lim_{k\to\infty} f(k+x)$ exists and equals zero. For $z > k$, $f(k-x)f(x+z) - f(k+x)f(-x+z) \leq 0$ so

$$|f(k-x)f(x+z) - f(k+x)f(-x+z)| \leq f(k+x)f(-x+z) \leq f(x)f(-x+z), \quad (3.6)$$

the last inequality following since f is unimodal and $x > 0$. Thus, by the dominated convergence theorem,

$$\lim_{k\to\infty} \int_k^\infty [f(k-x)f(x+z) - f(k+x)f(-x+z)]dz = 0 . \quad (3.7)$$

hence,

$$\begin{aligned} &\lim_{k\to\infty}[f(k-x)p(x) - f(k+x)(1 - p(x))] \\ &\quad = \lim_{k\to\infty}\int_0^k [f(k-x)f(x+z) - f(k+x)f(-x+z)]dz \geq 0 , \end{aligned} \quad (3.8)$$

establishing (3.4) and proving the theorem. □

The inequality between $\inf_{\pi\varepsilon G_{2PS}} P(H_0|x)$ and $p(x)$ is, in fact, strict in in many cases. Table 1 gives explicit expressions for some common distributions.

The Cauchy distribution, which does not have mℓr, does not attain its infimum at $k = \infty$ but rather at $k = (x^2+1)^{1/2}$. Even so, it is still the case that the p-value is greater than $\inf P(H_0|x)$ for the Cauchy distribution.

Table 1. P-values and inf $P(H_0|x)$ for the class of symmetric two-point distributions ($x > 0$)

Distribution	p(x)	inf $P(H_0\|x)$
normal	$1 - \Phi(x)$	0
double exponential	$\frac{1}{2} e^{-x}$	$(1 + e^{2x})^{-1}$
logistic	$(1 + e^{x})^{-1}$	$(1 + e^{2x})^{-1}$
Cauchy	$\frac{1}{2} - \frac{\tan^{-1} x}{\pi}$	$\frac{1 + (x-(x^2+1)^{\frac{1}{2}})^2}{2 + (x-(x^2+1)^{\frac{1}{2}})^2+(x+(x^2+1)^{\frac{1}{2}})^2}$

We now turn to the class of distributions G_{US}, where we again obtain the p-value as an upper bound on the infimum of the Bayesian evidence. We can, in fact, demonstrate equality between p(x) and inf $P(H_0|x)$ for two classes of distributions contained in G_{US}. We first consider

$$U_s = \{\text{all symmetric uniform distributions}\}.$$

Theorem 3.3: For the hypotheses in (2.1), if $x > 0$,

$$\inf_{\pi \varepsilon U_s} P(H_0|x) = p(x) .$$

Proof: Let $\pi(\theta)$ be uniform $(-k,k)$. Then

$$P(H_0|x) = \frac{\int_{-k}^{0} f(x-\theta)d\theta}{\int_{-k}^{k} f(x-\theta)d\theta} \tag{3.9}$$

and

$$\frac{d}{dk} P(H_0|x) = \left(\frac{f(x-k)+f(x+k)}{\int_{-k}^{k} f(x-\theta)d\theta}\right) \left[\frac{f(x+k)}{f(x-k)+f(x+k)} - P(H_0|x)\right] . \tag{3.10}$$

We will now establish that $P(H_0|x)$, as a function of k, has no minimum on the interior of $(0,\infty)$. Suppose $k = k_0$ satisfies

$$\frac{d}{dk} P(H_0|x)|_{k=k_0} = 0 \ . \tag{3.11}$$

It is straight forward to establish that the sign of the second derivative, evaluated at $k = k_0$, is given by

$$\operatorname{sgn} \frac{d^2}{dk^2} P(H_0|x) \ |_{k=k_0} = \operatorname{sgn} \ \frac{d}{dk} \frac{f(x+k)}{f(x-k)+f(x+k)} |_{k=k_0} \ . \tag{3.12}$$

Since f has mℓr, the ratio $f(x+k)/f(x-k)$ is decreasing in k for fixed $x > 0$. Therefore, the sign of (3.12) is always negative, so any interior extremum can only be a maximum. The minimum is obtained on the boundary, and it is straightforward to check that

$$\inf_{\pi \varepsilon U_s} P(H_0|x) = \lim_{k\to\infty} \frac{\int_{-k}^{0} f(x-\theta)d\theta}{\int_{-k}^{k} f(x-\theta)d\theta} = \int_{-\infty}^{0} f(x-\theta)d\theta = p(x) \ . \ \square$$

A similar result can be obtained for another class of distributions, G_{MU}, which consists of mixtures of symmetric uniform distributions. Let $\mathcal{G}$ be the set of all densities g on $[0,\infty)$ such that the scale parameter family $\{\sigma^{-1}g(k/\sigma), \sigma > 0\}$ has mlr in k. Define

$$G_{MU} = \{\pi: \ \pi(\theta) = \int_0^\infty (2k)^{-1} I_{(-k,k)}(\theta)\sigma^{-1} g(k/\sigma)dk, \ g\varepsilon\mathcal{G}, \ \sigma>0\} \ .$$

The class G_{MU} contains many familiar distributions symmetric about zero, including all normal and t distributions.

<u>Theorem 3.4</u>: For the hypotheses in (2.1), if $x > 0$,

$$\inf_{\pi\varepsilon G_{MU}} P(H_0|x) = p\ (x) \ . \tag{3.13}$$

<u>Proof</u>: Let $\pi(\theta)\varepsilon G_{MU}$. By interchanging the order of integration and using the symmetry of f we obtain

$$P(H_0|x) = \frac{\int_0^\infty (2k\sigma)^{-1} g(k/\sigma) \int_{-x-k}^{-x} f(z)dzdk}{\int_0^\infty (2k\sigma)^{-1} g(k/\sigma) \int_{-x-k}^{-x+k} f(z)dzdk} . \tag{3.14}$$

We first show that, for fixed g,

$$\inf_{0<\sigma<\infty} P(H_0|x) = \lim_{\sigma\to\infty} P(H_0|x) . \tag{3.15}$$

For notational convenience define

$$h(x,\sigma) = \int_0^\infty \sigma^{-1} g(y/\sigma) f(y-x)dy .$$

Since the denominator of (3.14) has derivative equal to $h(-x,\sigma)+h(x,\sigma) > 0$, it follows that

$$\text{sgn}\left[\frac{d}{d\sigma} P(H_0|x)\right] = \text{sgn}\left[\frac{h(-x,\sigma)}{h(-x,\sigma) + h(x,\sigma)} - P(H_0|x)\right] .$$

We now establish that if $P(H_0|x)$ has an extremum for $0 < \sigma < \infty$, that extremum must be a maximum. Suppose that $\sigma = \sigma_0$ satisfies

$$\frac{d}{d\sigma} P(H_0|x)\Big|_{\sigma=\sigma_0} = 0 .$$

Then

$$\text{sgn}\left[\frac{d^2}{d\sigma^2} P(H_0|x)\Big|_{\sigma=\sigma_0}\right] = \text{sgn}\left[\frac{d}{d\sigma} \frac{h(-x,\sigma)}{h(-x,\sigma)+h(x,\sigma)} \Big|_{\sigma=\sigma_0}\right]$$

$$= \text{sgn}\left[\frac{d}{d\sigma} \frac{h(-x,\sigma)}{h(x,\sigma)} \Big|_{\sigma=\sigma_0}\right] . \tag{3.16}$$

Since both f(k-x) and $\sigma^{-1}g(k/\sigma)$ have mlr, it follows from the Basic Composition Formula of Karlin (1968, p. 17) that $h(x,\sigma)$ also has mlr. Therefore, since $x > 0$, the sign of the last expression in (3.16) is negative, showing that any interior extremum must be a maximum. We therefore have

$$\inf_{0<\sigma<\infty} P(H_0|x) = \min\{\lim_{\sigma\to 0} P(H_0|x), \lim_{\sigma\to\infty} P(H_0|x)\} .$$

But from (3.14) it is easily verified, using l'Hopital's rule, that

$$\lim_{\sigma\to 0} P(H_0|x) = \frac{1}{2}, \quad \lim_{\sigma\to\infty} P(H_0|x) = p(x) < \frac{1}{2} .$$

Moreover, since we obtain the same infimum, p(x), regardless of the choice of $g \varepsilon G$, we have that

$$\inf_{\pi\varepsilon G_{MU}} P(H_0|x) = \inf_{g\varepsilon G} \inf_{0<\sigma<\infty} P(H_0|x) = \inf_{g\varepsilon G} p(x) = p(x) . \square$$

We can summarize the results of the above two theorems, and the relationship to G_{US} in the following corollary.

Corollary 3.1: For the hypotheses in (2.1), if $x > 0$,

$$\inf_{\pi\varepsilon G_{US}} P(H_0|x) \leq \inf_{\pi\varepsilon U_s} P(H_0|x) = \inf_{\pi\varepsilon G_{MU}} P(H_0|x) = p(x) .$$

This corollary is in striking contrast to the results of Berger and Sellke (1985). In the two sided problem with a point null hypothesis, they argued that using impartial prior distributions does not lead to any reconciliation between inf $P(H_0|x)$ and p(x). In fact, for the cases they considered, the Bayesian infimum was much greater that p(x). In contrast, we find that for classes of reasonable, impartial priors, such as G_{MU}, we obtain equality between inf $P(H_0|x)$ and p(x), showing that, in fact, p(x) is a conservative measure of evidence against the null hypothesis.

We close this section by examining two important special cases. In the first case we again obtain equality between $p(x)$ and $P(H_0|x)$.

Theorem 3.5: If $f(x-\theta) = (2\pi\sigma^2)^{-} \exp\{-\frac{1}{2}(x-\theta)^2/\sigma^2\}$, then for the hypothesis in (2.1), if $x > 0$,

$$\inf_{\pi\varepsilon G_{NOR}} P(H_0|x) = p(x) \quad .$$

Proof: The result is easily established by noting

$$P(H_0|x) = P\left(Z < \left(\frac{\tau^2}{\tau^2+\sigma^2}\right)^{\frac{1}{2}}\left(\frac{-x}{\sigma}\right)\right) \quad , \quad Z \sim n(0,1)$$

which attains its infimum at $\tau^2 = \infty$. □

We next consider the Cauchy distribution, to again examine the situation when the assumption of mlr does not obtain. For the class U_s, the symmetric uniform distributions, we calculate $\inf P(H_0|x)$ where $f(x-\theta) = [\pi(1+(x-\theta)^2)]^{-1}$. For $\pi(\theta)$ = Uniform $(-k,k)$ and it is straightforward to calculate

$$P(H_0|x) = \frac{\tan^{-1}(x+k) - \tan^{-1}(x)}{\tan^{-1}(x+k) - \tan^{-1}(x-k)} \quad .$$

For fixed $x > 0$, $P(H_0|x)$ is not monotone in k, but rather attains a unique minimum at a finite value of k. Table 2 lists the minimizing values of k, $\inf P(H_0|x)$, and the p-value for selected values of x.

Examination of Table 2 shows once again that $\inf P(H_0|x)$ is smaller than $p(x)$, this observation held true for more extensive calculations that are not reported here. Therefore, even in the case of the Cauchy distribution, the infimum of the Bayesian measure of evidence is smaller than the frequentist p-value.

Table 2. P-values and inf $P(H_0|x)$ for $X \sim$ Cauchy, infimum over U_s

x	k_{min}	p(x)	inf $P(H_o|x)$
.2	2.363	.437	.429
.4	2.444	.379	.363
.6	2.570	.328	.306
.8	2.727	.285	.260
1.0	2.913	.250	.222
1.2	3.112	.221	.192
1.4	3.323	.197	.168
1.6	3.541	.178	.148
1.8	3.768	.161	.132
2.0	3.994	.148	.119
2.5	4.572	.121	.094
3.0	5.158	.102	.077
3.5	5.746	.089	.065
4.0	6.326	.078	.056
5.0	7.492	.063	.044
10.0	13.175	.032	.020
25.0	29.610	.013	.007
50.0	56.260	.006	.004
75.0	82.429	.004	.002
100.0	108.599	.003	.002

4. Biased Prior Distributions

In this section we examine two cases where the prior distributions are biased toward H_0, and begin to see some of the reasons for the large discrepancies between Bayesian and frequentist evidence in the two-sided case.

Again consider H_0: $\theta \le 0$ vs. H_1: $\theta > 0$ where $X \sim n(\theta,\sigma^2), \sigma^2$ known. Consider the class of priors

$$G_{\theta_0} = \{n(\theta_0,\tau^2) \text{ distributions}, \ \theta_0 < 0 \text{ (fixed)}, \ 0 \le \tau^2 < \infty\}.$$

The class G_{θ_0} is clearly biased toward H_0, however, if we calculate $\inf P(H_0|x)$ over this class the result is again $p(x)$.

For any $\pi \varepsilon G_{\theta_0}$, it is easy to calculate

$$P(H_0|x) = P(Z \le -(\frac{\tau/\sigma}{(\sigma^2+\tau^2)^{1/2}}\, x + \frac{\sigma/\tau}{(\sigma^2+\tau^2)^{1/2}}\, \theta_0)) \ , \qquad (4.1)$$

where $Z \sim n(0,1)$. For $x > 0$, $P(H_0|x)$ is a decreasing function of τ^2, so the infimum is attained at $\tau^2 = \infty$:

$$\inf_{\pi \varepsilon G_{\theta_0}} P(H_0|x) = P(z < -x/\sigma) = p(x) \ . \qquad (4.2)$$

The effect of the bias for H_0 is diminished at τ^2 increases, resulting in a limit which is independent of θ_0. This is a different situation from the point-null case, where the prior probability on the point null is unaffected by any limiting operation.

We next consider a family of priors in which every member is biased toward H_0 by the same amount. Suppose that an experimenter is willing to assert, for every $k > 0$, it is q times more likely that $\theta \varepsilon (-k,0)$ than $\theta \varepsilon (0,k)$. This belief may be reflected in the prior

$$\pi(\theta) = \begin{cases} \frac{q}{k(1+q)} & -k < \theta \leq 0 \\ \frac{1}{k(1+q)} & 0 < \theta < k \end{cases} . \quad (4.3)$$

Let G_q denote the class of all of these priors. Then, by an argument similar to that used in Theorem 2.2, if $f(x-\theta)$ has mlr and $x > 0$, for testing H_0: $\theta \leq 0$ vs. H_1: $\theta > 0$ we have

$$\inf_{\pi \varepsilon G_q} P(H_0|x) = \frac{qp(x)}{qp(x)+(1-p(x))} . \quad (4.4)$$

The quantity in (4.4) is greater than $p(x)$ if $q > 1$ (prior biased toward H_0) and less than $p(x)$ if $q < 1$ (prior biased toward H_1). Therefore, (4.4) is a very reasonable measure of evidence, taking into account both prior beliefs and sample information. However, even in this biased case, we do not observe the same discrepancies as Berger and Sellke did in the point-null problem. For example, we might ask, "How large must q be in order that $\inf P(H_0|x)$ is twice as large as $p(x)$," in order to get some idea of how the bias for H_0 affects the measure of evidence. For $p = .01, .05, .1$, and various values of m, we can solve for q such that $\inf P(H_0|x) = mp$. Some values are given in Table 3.

For small m, q is approximately equal to m. However for larger values of m, q increases rapidly, showing that the prior must be very biased toward H_0 in order to achieve a large increase in $\inf P(H_0|x)$.

Table 3: For selected p-values, bias toward H_0 (q) necessary in order for inf $P(H_0|x) = mp(x)$

p=.1

m	2	4	6	8
q	2.25	6.00	13.50	36.00

p=.05

m	2	4	6	8	10	12	16
q	2.11	4.75	8.14	12.67	19.00	28.50	76.00

p=.01

m	2	4	6	8	10	20	50	75
q	2.02	4.13	6.32	8.61	11.00	24.75	99.00	297.00

5. Comments

For the problem of testing a one-sided hypothesis in a location parameter family, it is possible to reconcile evidence between the Bayesian and frequentist approaches. The frequency p-value is, in many cases, an upper bound on $P(H_0|x)$, showing that it is possible to regard the p-value as assessing "the probability that H_0 is true." Even though this phrase has no meaning within frequency theory, it has been argued that practitioners sometimes attach such a meaning to the p-value. The results in this paper show that, for testing a one-sided hypothesis, such a meaning can be attached to the p-value.

The discrepancies observed by Berger and Sellke in the two-sided (point null) case do not carry over to the problems considered here. This leads to the question of determining what factors are crucial in differentiating the two problems. It seems that if prior mass is concentrated at a point (or in a small interval), then discrepancies between Bayesian and frequentist measures will obtain. In fact, Berger and Sellke note that for testing H_0: $\theta = 0$ vs. H_1: $\theta > 0$, the p-value and the Bayesian infimum are quite different. (For example, for $X \sim n(\theta,1)$, an observed $x = 1.645$ will give a p-value of .05, while, if mass $\frac{1}{2}$ is concentrated at zero, $\inf P(H_0|x = 1.645) = .21$).

Seen in another light, however, it can be argued that placing a point mass of $\frac{1}{2}$ at H_0 is not representative of an impartial prior distribution. For the problem of testing H_0: $\theta \leq 0$ vs. H_1: $\theta > 0$, consider priors of the form

$$\pi(\theta) = \pi_0 h(\theta) + (1 - \pi_0) g(\theta) \qquad (5.1)$$

where π_0 is a fixed number, and $h(\theta)$ and $g(\theta)$ are proper priors on $(-\infty, 0)$ and $(0, \infty)$ respectively. It then follows that, for $x > 0$,

$$\sup_h P(H_0|x) = \sup_h \frac{\pi_0\int_{-\infty}^{0} f(x-\theta)h(\theta)d\theta}{\pi_0\int_{-\infty}^{0} f(x-\theta)h(\theta)d\theta + (1-\pi_0)\int_{0}^{\infty} f(x-\theta)g(\theta)d\theta} \quad (5.1)$$

$$= \frac{\pi_0 f(x)}{\pi_0 f(x)+(1-\pi_0)\int_{0}^{\infty} f(x-\theta)g(\theta)d\theta} \quad (5.2)$$

and the last expression is equal to $P(H_0|x)$ for the hypotheses H_0: $\theta = 0$ vs. H_1: $\theta > 0$ with prior $\pi(\theta) = \pi_0$ if $\theta = 0$ and $\pi(\theta) = (1 - \pi_0)g(\theta)$ if $\theta > 0$. Thus, concentrating mass on the point null hypothesis is biasing the prior in favor of H_0 as much as possible in this one-sided testing problem.

The calculation in (5.2) casts doubt on the reasonableness of regarding $\pi_0 = \frac{1}{2}$ as impartial. In fact, it is not clear to us if any prior that concentrates mass at a point can be viewed as an impartial prior.

Therefore, it is not surprising that the p-value and Bayesian evidence differ in the normal example given above. Setting $\pi_0 = \frac{1}{2}$ actually reflects a bias toward H_0, which is reflected in the Bayesian evidence.

To a Bayesian, the fact that evidence can be reconciled with the p-values allows for a Bayesian interpretation of a p-value and the possibility of regarding a p-value as an objective assessment of the probability that H_0 is true. It also, to a Bayesian, gives the p-value a certain amount of respectability. To a frequentist, the p-value (or significance level) has long been regarded as an objective assessment of the tenability of H_0, an interpretation that survives even within the Bayesian paradigm.

References

Berger, J., and Sellke, T. (1985), "Testing a Point Null Hypothesis: The Irreconcilability of Significance Levels and Evidence," Technical Report 84-25, Department of Statistics, Purdue University, West Lafayette, Ind.

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

Dickey, J. M. (1977), "Is the Tail Area Useful as an Approximate Bayes Factor," Journal of the American Statistical Association, 72, 138-142.

Karlin, S. (1968), Total Positivity, Volume I, Stanford, CA: Stanford University Press.

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

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

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