Model Selection
IMS Lecture Notes - Monograph Series (2001) Volume 38

Scales of Evidence for Model Selection: Fisher versus Jeffreys

Bradley Efron and Alan Gous

Stanford University and Cariden Technologies, Inc.

Abstract

Model selection refers to a data-based choice among competing statistical models, for example choosing between a linear or a quadratic regression function. The most popular model selection techniques are based on interpretations of p-values, using a scale originally suggested by Fisher: .05 is moderate evidence against the smaller model, .01 is strong evidence, etc. Recent Bayesian literature, going back to work by Jeffreys, suggests a quite different answer to the model selection problem. Jeffreys provided an interpretive scale for Bayes factors, a scale which can be implemented in practice by use of the BIC (Bayesian Information Criterion.) The Jeffreys scale often produces much more conservative results, especially in large samples, so for instance a .01 p-value may correspond to barely any evidence at all against the smaller model. This paper tries to reconcile the two theories by giving an interpretation of Fisher's scale in terms of Bayes factors. A general interpretation is given which works fine when checked for the one-dimensional Gaussian problem, where standard hypothesis testing is seen to coincide with a Bayesian analysis that assumes stronger (more informative) priors than those used by the BIC. This argument fails in higher dimensions, where Fisher's scale must be made more conservative in order to get a proper Bayes justification.

Bradley Efron is Max H. Stein Professor of Statistics, Department of Statistics, Sequia Hall, Stanford University, Paolo Alto, CA 94305-4065, U.S.A; email: brad@stat.Stanford.EDU. Alan Gous is Director of Research & Development, Cariden Technologies, Inc., 1733 Woodside Road, Suite 310, Redwood City, CA 94061, U.S.A.; email: alan@cariden.com.

Contents

1 Introduction

In model selection problems the statistician must use the data to choose between discrete alternative models, for instance which explanatory variables to include in a regression analysis. A particularly simple example involves a single Gaussian observation

$$x \sim N(\theta, 1). \tag{1.1}$$

and the choice between two models, a smaller one inside a bigger one,

$$\mathcal{M}_0 : \theta = 0 \quad \text{versus} \quad \mathcal{M} : \theta \neq 0. \tag{1.2}$$

This example shows up frequently in the model selection literature, and we will use it here for comparison of different methods.

By far the most widely-used model selection methods are based on hypothesis tests. A test statistic $S(\mathbf{x})$ depending on the observed data $\mathbf{x}$ is evaluated, $S(\mathbf{x}) = s$, and the *critical level* α is calculated,

$$\alpha = \text{prob}_{\mathcal{M}_0}\{S(\mathbf{x}) < s\} \tag{1.3}$$

(so $1 - \alpha$ equals the p-value or significance level). Here we suppose that larger values of S indicate stronger evidence against the smaller hypothesis $\mathcal{M}_0$, as with the optimal test statistic $S(\mathbf{x}) = |x|$ for situation (1.1)-(1.2). This leaves us with the problem of evaluating the strength of evidence for $\mathcal{M}$ and against $\mathcal{M}_0$. Frequentists use a scale of evidence set down by Fisher in the 1920's. Table 1 gives Fisher's scale as it is commonly interpreted: $\alpha = .99$ is strong evidence in favor of $\mathcal{M}$ versus $\mathcal{M}_0$, .95 is moderate evidence, etc. The borderline of neutral evidence is somewhere around $\alpha = .90$. Fisher, discussing chi-square tests, states it this way, in terms of $P = 1 - \alpha$: "If P is between .1 and .9 there is certainly no reason to suspect the hypothesis tested. If it is below .02 it

critical level α:	**.90**	**.95**	**.975**	**.99**	**.995**	**.999**
strength of evidence for $\mathcal{M}$:	borderline	moderate	substantial	strong	very strong	overwhelming

Table 1. Fisher's scale of evidence against hypothesis $\mathcal{M}_0$. The critical level (one minus the p-value) is the $\mathcal{M}_0$ probability of the test statistic being smaller than the value actually observed.

is strongly indicated that the hypothesis fails to account for the whole of the facts. We shall not often be astray if we draw a conventional line at .05, and consider that higher values of χ^2 indicate a real discrepancy." Section 20, Fisher (1954).

Recent Bayesian literature provides a different answer to the model selection problem. Attention is focussed on the *Bayes Factor* $B(\mathbf{x})$,

$$\frac{\text{prob}\{\mathcal{M}|\mathbf{x}\}}{\text{prob}\{\mathcal{M}_0|\mathbf{x}\}} = B(\mathbf{x})\frac{\text{prob}\{\mathcal{M}\}}{\text{prob}\{\mathcal{M}_0\}}, \tag{1.4}$$

the ratio between the aposteriori and apriori odds in favor of the bigger model $\mathcal{M}$, defined more carefully in Section 2. Jeffreys (1961) suggested a scale of evidence for interpreting Bayes factors, presented in slightly amended form in Table 2. For example $B(\mathbf{x}) = 10$ indicates positive but not strong evidence in favor of $\mathcal{M}$.

The calculation of $B(\mathbf{x})$ requires a full Bayesian prescription of the prior distribution on $\mathcal{M}$ and $\mathcal{M}_0$, as described in Section 2.1. This is not available in most situations, but Jeffreys suggested rules for objectively selecting priors in the absence of apriori knowledge, see Kass and Wasserman (1996). These rules have led to convenient data-based estimates of the Bayes factor $B(\mathbf{x})$, the most popular of which is the BIC (Bayesian Information Criterion), Schwarz (1978). Let $\widehat{B}(\mathbf{x})$ be Wilks' maximum likelihood ratio statistic, the ratio of the maximized likelihoods for $\mathcal{M}$ compared to $\mathcal{M}_0$. In a repeated sampling situation, where the data $\mathbf{x}$ comprise a random sample of size n, the BIC approximates the Bayes factor by $B_{\text{BIC}}(\mathbf{x})$,

$$\log\, B_{\text{BIC}}(\mathbf{x}) = \log\, \widehat{B}(\mathbf{x}) - \frac{d}{2}\log(n), \tag{1.5}$$

where d is the difference in the number of free parameters between $\mathcal{M}$ and $\mathcal{M}_0$. The rationale for this formula is discussed in Section 2.

As an example, suppose $\mathbf{x}$ is an i.i.d. (independent and identically distributed) Gaussian sample of size n,

$$x_1, x_2, \ldots x_n \overset{iid}{\sim} N(\theta_1, 1). \tag{1.6}$$

Bayes Factor $B(\mathbf{x})$:	< 1	$1 - 3$	$3 - 20$	$20 - 150$	> 150
Evidence for $\mathcal{M}$:	negative	not worth more than a bare mention	positive	strong	very strong

Table 2. Jeffreys' scale of evidence for the interpretation of Bayes factors, as amended by Kass and Raftery (1995).

and we wish to select between

$$\mathcal{M}_0 : \theta_1 = 0 \quad \text{versus} \quad \mathcal{M} : \theta_1 \neq 0. \tag{1.7}$$

This problem reduces to (1.1), (1.2) by defining

$$x = \sqrt{n}\,\bar{x} = \sum_1^n x_i/\sqrt{n} \quad \text{and} \quad \theta = \sqrt{n}\,\theta_1, \tag{1.8}$$

since x is a sufficient statistic. In this case $d = 1$, $\widehat{B}(\mathbf{x}) = x^2/2$, and

$$\log B_{\mathrm{BIC}}(\mathbf{x}) = \frac{x^2}{2} - \frac{\log(n)}{2}. \tag{1.9}$$

We can now make a disturbing comparison, of a type first emphasized in Lindley's 1957 paper: Jeffreys' scale of evidence, as implemented by the BIC, often leads to much more conservative decisions than Fisher's scale. A data set with p-value .01 (critical level .99) may be "barely worth mentioning" on Jeffreys' scale, a shocking assertion to the medical scientist for whom a .01 significance level settles the issue. Figure 1 illustrates the comparison for situation (1.6)-(1.7).

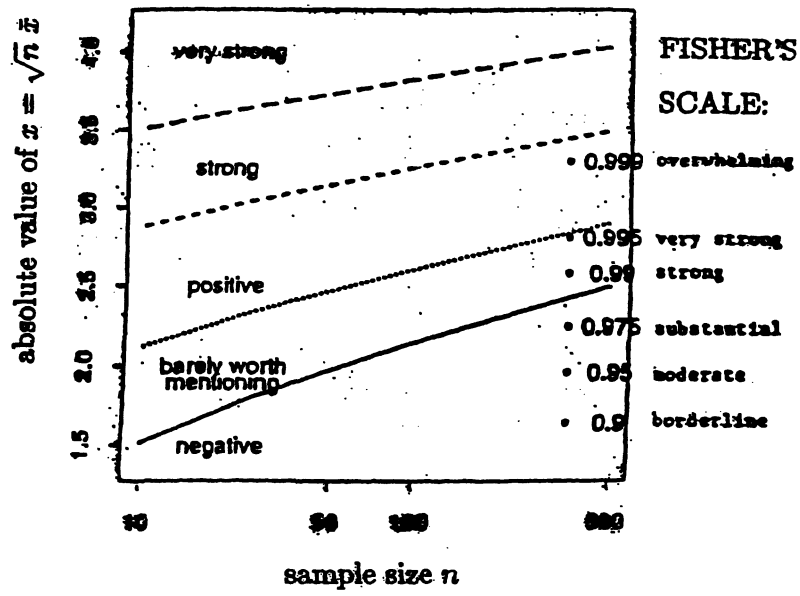

Figure 1 Jeffreys' scale of evidence for $\mathcal{M}$ as implemented by BIC; for Gaussian samples as in (1.6)-(1.9); Fisher's scale, shown at right, does not depend on n. Jeffreys' scale is much more conservative for large sample sizes. For example, $x = 2.58$ with $n = 100$ is strong Fisherian evidence, but barely worth mentioning for Jeffreys.

How can the two scales of evidence give such different results? We will try to answer the question by showing that frequentist model selection can be understood from a Bayesian point of view, but one where the prior distribution favors the bigger model more strongly than with BIC or other objectivist Bayes techniques. The reasons for this are partly historical, perhaps relating to Fisher's and Jeffreys' different scientific environments and different attitudes concerning the "null hypothesis" $\mathcal{M}_0$. More of the difference though arises from the Bayesian notion of sample size coherency, discussed in Section 4, which implies that evidence against the smaller model must be judged more cautiously in larger samples, as seen in Figure 1.

Section 2 reviews Bayes factors and develops a useful approximation formula for their comparison,

$$B(\mathbf{x}) \doteq \widehat{B}(\mathbf{x})/\widehat{B}(\mathbf{y}), \tag{1.10}$$

where $\mathbf{y}$ is a "breakeven point", that is a data set for which the Bayes factor $B(\mathbf{y})$ equals 1. Fisher's scale is interpreted as putting the breakeven point at the 90th percentile of the test statistic, e.g., at $y = 1.645$ in situation (1.1), (1.2), while the BIC puts it at $\sqrt{\log(n)}$, (1.9).

There is a healthy literature on the comparison of Bayesian and frequentist model selection. Some of this concerns the interpretation of a p-value as $\text{prob}\{\mathcal{M}_0|\mathbf{x}\}$. Berger and Selke (1987) show that this interpretation must be wrong for two-sided testing situations like (1.1)-(1.2), with the p-value underestimating $\text{prob}\{\mathcal{M}_0|\mathbf{x}\}$ as in Lindley's paradox, while Casella and Berger (1987) show that it is reasonable for one-sided problems. Berger, Boukai, and Wang (1997) reconcile p-values with Bayesian posterior probabilities using a conditional inference approach.

Andrews' (1994) results are closer to the point of view taken here. He shows that in a certain asymptotic sense there is a monotone relationship between p-values and Bayes factors. Section 3 gives some comparisons with Andrews' results. Section 5 concerns a small but illustrative example of model selection, combining the Bayesian and frequentist points of view. The paper ends with a series of remarks in section 6, and a brief summary.

Our main goal here is to reconcile Fisherian hypothesis testing with Jeffreys' Bayes-factor theory, at least as far as reconciliation is possible, and then to pinpoint the nature of the remaining differences. Technical issues will be kept to a minimum, with simple examples used to make many of the main points.

2 Approximations for Bayes Factors

The BIC formula (1.5) provides an objective approximation to the Bayes factor $B(\mathbf{x})$, objective in the sense that it is entirely data-based and does not involve subjective appraisals of prior probabilities. This section discusses a class of such approximation formulas based on the convenient representation (1.10). Our comparison of Fisher's and Jeffreys' scales of evidence will take place within this class.

2.1 Bayes Factors and the BIC

Suppose that we observe data $\mathbf{x}$ distributed according to a parametric family of densities $f_\theta(\mathbf{x})$,

$$\mathbf{x} \sim f_\theta(\mathbf{x}), \tag{2.1}$$

θ being an unknown parameter vector. We wish to choose between a smaller model and a bigger model for θ,

$$\theta \in \mathcal{M}_o \quad \text{versus} \quad \theta \in \mathcal{M}, \quad \text{with} \quad \mathcal{M}_o \subset \mathcal{M}. \tag{2.2}$$

A complete Bayesian analysis begins with prior probabilities for $\mathcal{M}_o$ and $\mathcal{M}$,

$$\pi_o = \text{prob}\{\theta \in \mathcal{M}_o\} \quad \text{and} \quad \pi = \text{prob}\{\theta \in \mathcal{M}\}, \tag{2.3}$$

$\pi_o + \pi = 1$, and also with prior conditional densities for θ given each model,

$$g_o(\theta) \quad \text{and} \quad g(\theta) \tag{2.4}$$

for the densities of θ given $\mathcal{M}_o$ and $\mathcal{M}$ respectively. We assume that $g(\theta)$ puts probability zero on $\mathcal{M}_o$ so that it is not necessary to subtract $\mathcal{M}_o$ from $\mathcal{M}$ as was done in (1.2). Sometimes it is slightly more convenient to take $\mathcal{M}_o \subset \mathcal{M}$ as in (2.2).

Letting $\pi(\mathbf{x})$ and $\pi_o(\mathbf{x})$ be the aposteriori probabilities for $\mathcal{M}$ and $\mathcal{M}_o$ having observed $\mathbf{x}$, Bayes' rule gives

$$\frac{\pi(\mathbf{x})}{\pi_o(\mathbf{x})} = \frac{\pi}{\pi_0} \cdot \frac{f(\mathbf{x})}{f_0(\mathbf{x})}, \tag{2.5}$$

where $f(\mathbf{x})$ and $f_o(\mathbf{x})$ are the two marginal densities

$$f(\mathbf{x}) = \int_{\mathcal{M}} f_\theta(\mathbf{x}) g(\theta) d\theta \quad \text{and} \quad f_o(\mathbf{x}) = \int_{\mathcal{M}_o} f_\theta(\mathbf{x}) g_o(\theta) d\theta. \tag{2.6}$$

The ratio

$$B(\mathbf{x}) = \frac{f(\mathbf{x})}{f_o(\mathbf{x})} \tag{2.7}$$

is called the *Bayes factor* in favor of $\mathcal{M}$ compared to $\mathcal{M}_o$. Kass and Raftery (1995) provide a nice overview of Bayes factors, including their origins in the work of Jeffreys, Good, and others.

Bayes' rule (2.5) relates the posterior and prior odds ratio through the Bayes factor,

$$\frac{\pi(\mathbf{x})}{\pi_o(\mathbf{x})} = \frac{\pi}{\pi_o} \cdot B(\mathbf{x}). \tag{2.8}$$

The Bayesian model selection literature tends to focus on the case $\pi/\pi_o = 1$ of equal prior odds, with the presumption that π/π_o will be appropriately adjusted in specific applications. This paper proceeds in the same spirit, except for some of the specific frequentist/Bayesian comparisons of Section 5.

In most situations there will be no obvious choice for the prior densities $g_o(\theta)$ and $g(\theta)$ in (2.4). Jeffreys (1935, 1961) suggested objective rules for choosing priors in model selection problems. This has led to a substantial current literature, Kass and Raftery's (1995) bibliography listing more than 150 papers. The BIC approximation (1.5) can be thought of as a Bayes factor based on Jeffreys' objective priors, but we will see that other approximations, closer in effect to Fisher's scale of evidence can be described in the same way.

A clear derivation of the BIC appears in Section 2 of Kass and Wasserman (1995). They follow Smith and Spiegelhalter's (1980) approach in which Bayesian objectivity is interpreted to mean a prior distribution having the same amount of information for θ as one observation out of a random sample of size n. As an example, consider the repeated-sampling Gaussian problem (1.6)-(1.7), with $x_1, x_2, \ldots x_n \overset{\text{i.i.d.}}{\sim} N(\theta_1, 1)$. If we use a Gaussian prior $g(\theta_1) = N(0, A)$ then Smith and Spiegelhalter's approach suggests taking $A = 1$. The exact Bayes factor is easily calculated,

$$\log B(\mathbf{x}) = \frac{n}{n+1} \frac{x^2}{2} - \frac{\log(n+1)}{2}, \quad [x = \sqrt{n}\,\bar{x}] \tag{2.9}$$

which approaches $\log B_{\text{BIC}}(\mathbf{x}) = [x^2 - \log(n)]/2$, (1.9), as n grows large.

There seems to be an obvious objective Bayesian prior density for this situation: $g(\theta_1) = c$, a constant. This gives a Bayes factor $\int cf_{\theta_1}(\mathbf{x})d\theta_1/f_o(\mathbf{x})$ of the form

$$\log B(\mathbf{x}) = \frac{x^2}{2} - \frac{1}{2}\log(\frac{n}{2\pi c^2}). \tag{2.10}$$

However because we are dealing with an improper density function we still have to select a value for the constant c. The BIC chooses $c = 1/\sqrt{2\pi}$. The breakeven point for (2.10), where $B(\mathbf{x}) = 1$, occurs at $x = [\log(n/2\pi c^2)]^{1/2}$, so the BIC choice amounts to setting the breakeven point at $[\log(n)]^{1/2}$.

Suppose we believe entirely in model $\mathcal{M}$ in (1.6)-(1.7), and wish only to estimate θ. Then the constant prior density $g(\theta) = c$ gives answers agreeing with the usual frequentist confidence intervals, and the Bayesian objectivist does not have to worry about the choice of c. Model selection is inherently more awkward and difficult than estimation, part of the trouble coming from the different dimensionalities of $\mathcal{M}$ and $\mathcal{M}_o$. A prior distribution for situation (2.2) is inherently bumpy around the smaller model $\mathcal{M}_o$, rather than smooth as in estimation problems, making the Bayesian analysis more delicate. See Remark D in section 6.

2.2 A Useful Approximation Lemma

Our comparison of frequentist and Bayesian model selection relies on formula (1.10), $B(\mathbf{x}) \doteq \widehat{B}(\mathbf{x})/\widehat{B}(\mathbf{y})$, $\mathbf{y}$ a breakeven point, which will now be derived in somewhat more general form. We will work in an exponential family setting where $\mathcal{M}$ is a multiparameter exponential family of densities $f_\theta(\mathbf{x})$, and $\mathcal{M}_o \subset \mathcal{M}$ is also an exponential family. For example, $\mathcal{M}$ might be a logistic regression model with m predictors while $\mathcal{M}_0$, is a submodel in which $m_o < m$ of the predictors appear.

Besides the observed data point $\mathbf{x}$ consider a second point $\mathbf{x}^\dagger$, which for now can be any point in the domain of the exponential family. Definition (2.7) gives

$$B(\mathbf{x}) = \frac{B(\mathbf{x})}{B(\mathbf{x}^\dagger)}B(\mathbf{x}^\dagger) = \frac{R}{R_o}B(\mathbf{x}^\dagger),$$

where R and R_o are ratios of marginal densities (2.6),

$$R = \frac{f(\mathbf{x})}{f(\mathbf{x}^\dagger)} \quad \text{and} \quad R_o = \frac{f_o(\mathbf{x})}{f_o(\mathbf{x}^\dagger)}. \tag{2.11}$$

Section (1.2) of O'Hagen (1995) reviews some of the history of equation (2.10).

In what follows we will show that $R/R_o = B(\mathbf{x})/B(\mathbf{x}^\dagger)$ is well-approximated by the ratio of likelihood ratio statistics,

$$B(\mathbf{x})/B(\mathbf{x}^\dagger) \doteq \widehat{B}(\mathbf{x})/\widehat{B}(\mathbf{x}^\dagger), \tag{2.12}$$

where

$$\widehat{B}(\mathbf{x}) = \frac{f_{\widehat{\theta}}(\mathbf{x})}{f_{\widehat{\theta}_o}(\mathbf{x})} \quad \text{and} \quad \widehat{B}(\mathbf{x}^\dagger) = \frac{f_{\widehat{\theta}^\dagger}(\mathbf{x}^\dagger)}{f_{\widehat{\theta}_o^\dagger}(\mathbf{x}^\dagger)}. \tag{2.13}$$

Here $(\widehat{\theta}, \widehat{\theta}_o)$ are the MLEs under $\mathcal{M}$ and $\mathcal{M}_o$ for data point $\mathbf{x}$, and similarly $(\widehat{\theta}^\dagger, \widehat{\theta}_o^\dagger)$ for $\mathbf{x}^\dagger$. Notice that R involves only densities from $\mathcal{M}$, and R_o involves only densities from $\mathcal{M}_o$, which makes the approximation theory easy: we need never compare densities from spaces of different dimensions, which is particularly helpful in working with improper priors such as $g(\theta) =$ constant. This tactic is an example of Good's (1947, 1951) "device of imaginary results" used, differently, by Smith and Spiegelhalter (1980, 1982), Kass and Wasserman (1995) and Pettit (1992).

Information matrices play a central role in the derivation of (2.12). The observed and expected Fisher information matrices for θ are defined in terms of the log likelihood second derivative matrix $\ddot{\ell}_{\mathbf{x}}(\theta) = \frac{\partial^2}{\partial\theta^2} \log f_\theta(\mathbf{x})$,

$$i_{\mathbf{x}}(\theta) = -\ddot{\ell}_{\mathbf{x}}(\theta) \quad \text{and} \quad I(\theta) = E_\theta\{i_{\mathbf{x}}(\theta)\}. \tag{2.14}$$

In exponential families $i_{\mathbf{x}}(\theta) = I(\theta)$, so $i_{\mathbf{x}}(\theta)$ does not depend upon $\mathbf{x}$. In repeated sampling situations such as (1.6), $I(\theta)$ is proportional to the sample size n. The order of magnitude error rates that follow refer to n, but they are valid beyond the repeated sampling framework. In a logistic regression situation, for instance, we could take n to be trace $\{I(\theta)\}$ and the bounds would remain valid under mild regularity conditions on the choice of the covariate vectors. In practice "sample size" is a difficult concept to define and causes trouble in applications of the BIC, see Section 5 here and Section 5 of Kass and Wasserman (1995).

For estimation problems (but not for model selection) Jeffreys suggested using the invariant prior density, usually improper,

$$g(\theta) = c \cdot |I(\theta)|^{\frac{1}{2}}, \tag{2.15}$$

c being an arbitrary positive constant, Kass and Wasserman (1996). This is particularly convenient for deriving (1.10) as the following lemma shows.

Lemma Suppose $\mathcal{M}_o \subset \mathcal{M}$ are exponential families and that the prior densities (2.4) are

$$g_o(\theta) = c_o \cdot |I_o(\theta_o)|^{\frac{1}{2}} \quad \text{and} \quad g(\theta) = c \cdot |I(\theta)|^{\frac{1}{2}}, \tag{2.16}$$

$I_o(\theta_o)$ being the $\mathcal{M}_o$ Fisher information matrix at point θ_o in $\mathcal{M}_o$. Then the ratio $R/R_o = B(\mathbf{x})/B(\mathbf{x}^\dagger)$ (2.11) is approximated to second order by $\widehat{B}(\mathbf{x})/\widehat{B}(\mathbf{x}^\dagger)$, (2.13),

$$\frac{B(\mathbf{x})}{B(\mathbf{x}^\dagger)} = \frac{\widehat{B}(\mathbf{x})}{\widehat{B}(\mathbf{x}^\dagger)}[1 + O(n^{-1})]. \tag{2.17}$$

Proof For any smooth prior density $g(\theta)$, Laplace's method, Tierney, Kass and Kadane (1989), gives

$$f(\mathbf{x}) = (2\pi)^{m/2} f_{\widehat{\theta}}(\mathbf{x}) g(\widehat{\theta})\ |I(\widehat{\theta})|^{-\frac{1}{2}}(1 + O(n^{-1})), \tag{2.18}$$

$m = \dim(\mathcal{M})$, and similarly for the other combinations of parameter values and data points appearing in (2.13). Therefore

$$R = \frac{f_{\widehat{\theta}}(\mathbf{x}) g(\widehat{\theta})\ |I(\widehat{\theta})|^{-\frac{1}{2}}}{f_{\widehat{\theta}^\dagger}(\mathbf{x}^\dagger) g(\widehat{\theta}^\dagger)\ |I(\widehat{\theta}^\dagger)|^{\frac{1}{2}}}\,[1+O(n^{-1})] \quad \text{and} \quad R_o = \frac{f_{\widehat{\theta}_o}(\mathbf{x}) g_o(\widehat{\theta}_o)\ |I_o(\widehat{\theta}_o)|^{-\frac{1}{2}}}{f_{\widehat{\theta}_o^\dagger}(\mathbf{x}^\dagger)\ g_o(\widehat{\theta}_o^\dagger)\ |I_o(\widehat{\theta}_o)\ |^{-\frac{1}{2}}}\,[1+O(n^{-1})]. \tag{2.19}$$

Jeffreys' priors (2.16) result in $R \doteq f_{\widehat{\theta}}(\mathbf{x})/f_{\widehat{\theta}^\dagger}(\mathbf{x}^\dagger)$ and $R_o \doteq f_{\widehat{\theta}_o}(\mathbf{x})/f_{\widehat{\theta}_o^\dagger}(\mathbf{x}^\dagger)$, regardless of the two constants c_o and c in (2.16), verifying (2.17).

Combining (2.10) with the lemma shows that

$$B(\mathbf{x}) \doteq \frac{\widehat{B}(\mathbf{x})}{\widehat{B}(\mathbf{x}^\dagger)} B(\mathbf{x}^\dagger), \tag{2.20}$$

"$\doteq$" indicating second-order accuracy as in (2.17). Various choices of $\mathbf{x}^\dagger$ are made in Section 2.3, leading to convenient approximations for $B(\mathbf{x})$. For all such choices we will insist that $\mathbf{x}^\dagger$ have the same $\mathcal{M}_o$ MLE as $\mathbf{x}$,

$$\widehat{\theta}_o^\dagger = \widehat{\theta}_o. \tag{2.21}$$

This has the following benefit: *if $\mathcal{M}_o$ is an exponential family and $\widehat{\theta}_o^\dagger = \widehat{\theta}_o$, then $R_o = f_o(\mathbf{x})/f_o(\mathbf{x}^\dagger)$ exactly equals $f_{\widehat{\theta}_o}(\mathbf{x})/f_{\widehat{\theta}_o}(\mathbf{x}^\dagger)$, no matter what the prior $g_o(\theta_o)$ may*

be, reducing assumptions (2.16) to $g(\theta) = c\,|I(\theta)|^{\frac{1}{2}}$, and removing part of the error in approximation (2.17). This follows easily from the exponential family fact that (2.21) makes

$$\frac{f_{\theta_o}(\mathbf{x})}{f_{\theta_o}(\mathbf{x}^\dagger)} = \frac{f_{\widehat{\theta}_o}(\mathbf{x})}{f_{\widehat{\theta}_o}(\mathbf{x}^\dagger)} \qquad \text{for all } \theta_o \in \mathcal{M}_o. \tag{2.22}$$

Approximation (2.12), $B(\mathbf{x})/B(\mathbf{x}^\dagger) \doteq \widehat{B}(\mathbf{x})/\widehat{B}(\mathbf{x}^\dagger)$, tends to be highly accurate even in small samples. Table 3 shows part of a regression example with $0-1$ response data taken from Finney (1947), and used by Kass and Wasserman (1995). A total of $n = 39$ (predictor, response) pairs (v_i, x_i) are available, and we consider testing model $\mathcal{M}$, that the probability p_i of $x_i = 1$ follows the linear logistic form

$$\text{logit}(p_i) = \theta_o + \theta_1 v_i, \tag{2.23}$$

versus $\mathcal{M}_o$: $\theta_1 = 0$. The MLE of $\mathbf{p} = (p_1, p_2, \ldots, p_{39})$ under $\mathcal{M}_o$ is

$$\widehat{\mathbf{p}}_o = (.513, .513, \ldots, .513), \tag{2.24}$$

and we will take

$$\mathbf{x}^\dagger = .365\ \widehat{\mathbf{p}}_o + .635\ \mathbf{x}. \tag{2.25}$$

This is a breakeven point for testing $\theta_1 = 0$, as defined in Section 2.3.

v:	157	154	110	88	90	85	78	104	95	95	90	74	78
x:	1	1	1	1	1	1	0	0	0	0	0	0	0
v:	115	88	136	151	93	123	126	60	98	113	118	120	78
x:	1	1	1	1	1	0	1	0	0	0	0	1	0
v:	126	98	128	120	143	137	104	104	108	90	98	88	111
x:	1	0	1	0	1	0	0	1	1	1	0	0	1

Table 3 Logistic regression example; $n = 39$ cases of predictor v and dichotomous response x. From Finney (1947) in an experiment concerning vasoconstriction; v is x_1 from his table 1.

We can evaluate $B(\mathbf{x})$ and $B(\mathbf{x}^\dagger)$ directly by numerical integration of (2.6) over the Jeffreys' prior (2.15), obtaining

$$\frac{B(\mathbf{x})}{B(\mathbf{x}^\dagger)} = 8.888 \quad \text{compared to} \quad \frac{\widehat{B}(\mathbf{x})}{\widehat{B}(\mathbf{x}^\dagger)} = 8.841, \tag{2.26}$$

an error of only half a percent. This impressive accuracy reflects the fact that in a certain practical sense discussed in Remark I of Section 6, approximation (2.12) is third-order accurate.

The combination of Jeffreys' prior densities with exponential families makes the approximation $B(\mathbf{x})/B(\mathbf{x}^\dagger) \doteq \widehat{B}(\mathbf{x})/\widehat{B}(\mathbf{x}^\dagger)$ at least second-order accurate. Less restrictive assumptions lead to less accurate versions of (2.17) and (2.20). If we allow $\mathcal{M}_o$ to be a *curved* exponential subfamily of $\mathcal{M}$ then (2.17) may only be first-order accurate, erring by factor $1 + O(n^{-1/2})$, and similarly if Jeffreys' prior is replaced by some other slowly varying function $g(\theta)$. All of this is of more theoretical than practical importance. Much bigger differences between the frequentist and Bayesian methods arise from their different choices of the breakeven point $\mathbf{y}$ in (1.10).

2.3 The Breakeven Point

Definition A *breakeven point* $\mathbf{y}$ is a data set having the same $\mathcal{M}_o$ MLE as the observed data set $\mathbf{x}$, (2.21), and satisfying $B(\mathbf{y}) = 1$.

It follows from (2.20) that

$$B(\mathbf{x}) \doteq \widehat{B}(\mathbf{x})/\widehat{B}(\mathbf{y}) \tag{2.27}$$

so if we can find $\mathbf{y}$ we can compute a good approximation to the Bayes factor $B(\mathbf{x})$. This is especially convenient in Fisher's framework. A critical level of $\alpha_o = .90$ corresponds to the Fisherian breakeven point, so we can take $\mathbf{y}$ to be a point such that the test statistic $S(\mathbf{y})$ equals its $\mathcal{M}_o$ 90th percentile,

$$S(\mathbf{y}) = S^{(.90)}. \tag{2.28}$$

Things are particularly simple if the test statistic $S(\mathbf{x})$ is the likelihood ratio statistic $\widehat{B}(\mathbf{x})$ itself. Wilks' theorem says that $2\log(S^{.90})$ is approximated by the 90th percentile of a χ^2_d random variable, $d = \dim(\mathcal{M}) - \dim(\mathcal{M}_o)$,

$$S^{(.90)} \doteq \exp\{\chi^{2(.90)}_d/2\} \tag{2.29}$$

Then (2.27) provides an estimate of "$B_{\text{freq}}(\mathbf{x})$", the effective frequentist Bayes factor,

$$\log\, B_{\text{freq}}(\mathbf{x}) \doteq \log\, \widehat{B}(\mathbf{x}) - \chi_d^{2(.90)}/2. \tag{2.30}$$

Notice that formula (2.30) does not require explicit calculation of the breakeven point $\mathbf{y}$. It is less accurate than (2.27), see Remark E, but the difference is small in practice.

For the one-dimensional Gaussian situation (1.6)-(1.7), relation (2.27) says that an objective Bayes factor in favor of $\mathcal{M}$ should be of the form

$$\log\, B(\mathbf{x}) \doteq (x^2 - y^2)/2 \qquad (x = \sqrt{n}\,\bar{x}) \tag{2.31}$$

for some choice of the breakeven value y. The frequentist choice is $y_{\text{freq}} = 1.645$, the 90th percentile of $|x|$, while the BIC formula (1.9) uses $y_{\text{BIC}} = \sqrt{\log\,(n)}$. Table 4 shows that y_{BIC} crosses the frequentist value at sample size $n = 15$, growing larger at a rate that causes the dramatic differences seen on the right side of Figure 1. Section 4 discusses the Bayesian rationale for increasing the breakeven point as n gets bigger.

n:	10	15	100	1000	10000
$\sqrt{\log\,(n)}$:	1.52	1.645	2.15	2.63	3.04

Table 4 Breakeven point for BIC in Gaussian problem (1.6)-(1.7), as a function of sample size n. It equals frequentist value 1.645 at $n = 15$.

For another comparison consider selecting between $\mathcal{M}_o$ and $\mathcal{M}$ in the logistic regression situation of Table 3 , (2.23), which has $\widehat{B}(\mathbf{x}) = 34.28$. Applying definitions (2.30) and (1.5) gives

$$B_{\text{freq}}(\mathbf{x}) = 8.84 \quad \text{and} \quad B_{\text{BIC}}(\mathbf{x}) = 5.49, \tag{2.32}$$

showing substantial disagreement even at sample size $n = 39$.

There is another interesting choice for $\mathbf{x}^\dagger$ in the approximation formula $B(\mathbf{x}) \doteq [\widehat{B}(\mathbf{x})/\widehat{B}(\mathbf{x}^\dagger)]\, B(\mathbf{x}^\dagger)$:

Definition: A *least favorable point* $\mathbf{z}$ is a data set having its MLE under both $\mathcal{M}_o$ and $\mathcal{M}$ equal $\widehat{\theta}_o$, the $\mathcal{M}_o$ MLE for the actual data set $\mathbf{x}$. The point $\mathbf{z}$ is least favorable to the bigger hypothesis in the sense that the availability of $\mathcal{M}$ does not change the MLE from its $\mathcal{M}_o$ value. In examples (1.1), (1.2), and (1.6), (1.7), $z = 0$.

We have $\widehat{B}(\mathbf{z}) = f_{\widehat{\theta}_o}(\mathbf{z})/f_{\widehat{\theta}_o}(\mathbf{z}) = 1$, so (2.20) gives $B(\mathbf{x}) \doteq \widehat{B}(\mathbf{x}) \cdot B(\mathbf{z})$, or equivalently

$$\log\ B(\mathbf{x}) \doteq \log\ \widehat{B}(\mathbf{x}) - \log\ B^{-1}(\mathbf{z}), \tag{2.33}$$

where $B^{-1}(\mathbf{z}) = 1/B(\mathbf{z})$ is the Bayes factor for $\mathcal{M}_o$ compared to $\mathcal{M}$. This has the following interpretation: to obtain the log Bayes factor in favor of $\mathcal{M}$, *penalize* the corresponding log likelihood ratio by the log of the Bayes factor in favor of $\mathcal{M}_o$ at the least favorable point. Comparing (2.33) with (2.27) shows that

$$B^{-1}(\mathbf{z}) \doteq \widehat{B}(\mathbf{y}), \tag{2.34}$$

so the penalty against $\log\ \widehat{B}(\mathbf{x})$ also equals the log of the likelihood ratio statistic at the breakeven point.

From (1.5) and (2.30) we see that the BIC and frequentist penalties are

$$\frac{d}{2}\log\,(n) \quad \text{and} \quad \log\ S^{(.90)} \doteq \chi_d^{2(.90)}/2 \tag{2.35}$$

respectively. The BIC penalty amounts to taking $B^{-1}(\mathbf{z}) = n^{d/2}$ in (2.33), so that at the least favorable point for $\mathcal{M}$ there can be a large Bayes factor in favor of $\mathcal{M}_o$. By contrast, in the frequentist framework their can never be a large Bayes factor for $\mathcal{M}_o$, the maximum possible factor being $\exp\{\chi_d^{2(.90)}/2\}$, equaling 3.87 for $d = 1$.

3 Frequentists As Bayesians

Section 2 argues that objective Bayes factors should be of the form $B(\mathbf{x}) \doteq \widehat{B}(\mathbf{x})/\widehat{B}(\mathbf{y})$, and that Fisher's scale of evidence amounts to putting the breakeven point $\mathbf{y}$ at a value $\mathbf{y}^{(.90)}$ such that $S(\mathbf{y}^{(.90)}) = S^{(.90)}$, the 90th percentile of the test statistic S. Taken together these arguments suggest defining the frequentist Bayes factor to be $B_{\text{freq}}(\mathbf{x}) \doteq \widehat{B}(\mathbf{x})/\widehat{B}(\mathbf{y}^{(.90)})$, or for convenience the cruder approximation (2.30).

For the one-dimensional Gaussian situation (1.6)-(1.7) with $x = \sqrt{n}\,\bar{x}$, we have

$$B_{\text{freq}}(\mathbf{x}) = \exp((x^2 - y^2)/2), \quad y = 1.645. \tag{3.1}$$

Table 5 shows $B_{\text{freq}}(\mathbf{x})$ for values of x corresponding to Fisher's scale in Table 1. For example, $|x| = 2.58$, which is strong evidence against $\mathcal{M}_o$ on Fisher's scale, corresponds to Bayes factor 7.13, giving aposteriori probability .88 for $\mathcal{M}$ assuming equal prior probabilities on $\mathcal{M}$ and $\mathcal{M}_o$. Expressing the frequentist results in terms of $B_{\text{freq}}(\mathbf{x})$, instead of comparing than with $B_{\text{BIC}}(\mathbf{x})$ as in Figure 1, reduces the discrepancy between Fisher's

Fisher:	borderline	moderate	substantial	strong	very strong	overwhelming
α:	**.90**	**.95**	**.975**	**.99**	**.995**	**.999**
$\|x\|$:	1.645	1.96	2.24	2.58	2.81	3.29
$B_{\text{freq}}(\mathbf{x})$:	1	1.76	3.19	7.13	13.29	58.03
$\pi(\mathbf{x})$:	.50	.64	.76	.88	.93	.98

Table 5. Frequentist Bayes factors corresponding to critical levels on Fisher's scale; one-dimensional Gaussian case, (3.1); $\pi(\mathbf{x})$ is prob$\{\mathcal{M}|\mathbf{x}\}$ assuming equal prior probabilities for $\mathcal{M}$ and $\mathcal{M}_o$.

and Jeffreys' scales of evidence, though Jeffreys' scale remains somewhat more favorable to $\mathcal{M}_o$.

$B_{\text{freq}}(\mathbf{x})$ is an answer to the question "what kind of Bayesians are frequentist model selectors?" We can fortify our belief in this answer by finding a genuine prior density $g(\theta)$ on $\mathcal{M}$, as opposed to Jeffreys' improper prior, that gives Bayes factors close to $B_{\text{freq}}(\mathbf{x})$. This section shows that such priors exist in the one-dimensional Gaussian situation (1.6)-(1.7), and in its one-sided version, but not in higher dimensional Gaussian problems. These results have a close connection to the work of Andrews (1994), discussed below.

3.1 The One-Dimensional Gaussian Case

We consider the simple problem (1.1), (1.2), which is equivalent to the repeated sampling version (1.6), (1.7). Figure 2 compares $B_{\text{freq}}(\mathbf{x})$, (3.1), with $B_{4.85}(\mathbf{x})$, the actual Bayes factor(2.7) when $g(\theta) = \mathcal{U}[\pm\, 4.85]$, the uniform density for θ on $[-4.85, 4.85]$. (Since $\mathcal{M}_o$ consists of the single point $\theta = 0$, the other conditional density $g_o(\theta_o)$ in (2.4) must put all of its probability on zero). We see that the match is excellent. The average absolute error over the range of Fisher's scale,

$$Q = \int_{-3.29}^{3.29} |B_{4.85}(\mathbf{x})/B_{\text{freq}}(\mathbf{x}) - 1|\; dx/(2 \cdot 3.29) \tag{3.2}$$

is only 0.011. Section 3.3 motivates the choice 4.85 and shows that still better matching priors are possible. The optimum choice among symmetric priors $g(\theta)$ is supported on six points,

$$\begin{array}{lccc} \theta: & \pm.71 & \pm2.18 & \pm3.86 \\ g: & .147 & .159 & .194 \end{array} \tag{3.3}$$

and gives $Q = .00061$, the minimum possible Q value according to the linear programming theory of Section 3.3.

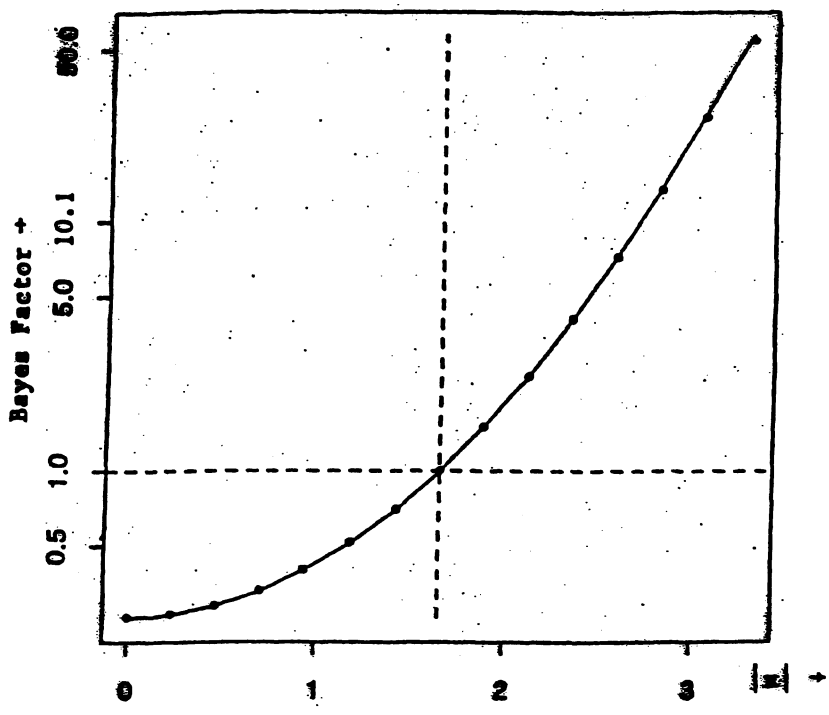

Figure 2 Comparison of log $B_{\text{freq}}(\mathbf{x})$, (solid curve, from (3.1)), with log $B_{4.85}(\mathbf{x})$ (dots) the genuine Bayes factor for the one-dimensional Gaussian situation if $g(\theta)$ is uniform on $[-4.85, 4.85]$.

All of this says that the frequentist is behaving like a somewhat unobjective Bayesian: the prior distribution $g(\theta)$ on the alternative hypothesis $\mathcal{M}$ is entirely supported within a few standard errors of the null hypothesis $\mathcal{M}_o$. By contrast, the BIC criterion is nearly equivalent to using a prior density $g(\theta)$ uniformly distributed over $\pm\ (\pi \cdot n/2)^{1/2}$, see formula (3.17) of Section (3.3),

$$\begin{array}{lccccc} n: & 10 & 15 & 100 & 1000 & 10000 \\ (\pi \cdot n/2)^{1/2}: & 3.96 & 4.85 & 12.53 & 39.63 & 125.33 \end{array} \tag{3.4}$$

The $n^{1/2}$ growth in the range of support for $g(\theta)$ is rooted in notions of Bayesian coherency as discussed in Section 4.

As in Table 4, the BIC and frequentist results agree at $n = 15$. To put it another way, the frequentist is effectively using a prior density $g(\theta)$ having about 1/15th of the data's information for estimating θ within model $\mathcal{M}$, while the BIC prior has only 1/nth of the information. See Remark H of Section 6.

The results in this section are closely related to those of Andrews (1994). Andrews

considers the asymptotic analysis of model-selection problem (2.1), (2.2) when the prior density $g(\theta)$ in (2.8) is shrinking toward $\mathcal{M}_o$ at rate $1/\sqrt{n}$. For a class of elliptically symmetric $g(\theta)$ densities chosen to match the hypothesis-testing situation, he shows (in our notation) that $B(\mathbf{x})$ is asymptotically a function of $\widehat{B}(\mathbf{x})$. This amounts to defining a frequentist Bayes factor, like our $B_{\text{freq}}(\mathbf{x})$ though the specific form is a little different.

In situation (1.6)-(1.8) Andrews' theory would suggest

$$\log\ B_{\text{freq}}(\mathbf{x}) = \frac{A}{A+1}\,\frac{x^2}{2} - \frac{1}{2}\,\log\,(A+1), \tag{3.5}$$

as in his equation (2.6), this being the actual Bayes factor $B(\mathbf{x})$ starting from the Gaussian prior $g(\theta) \sim N(0, A)$. He recommends choosing the prior variance A so that $B_{\text{freq}}(\mathbf{x})$ equals one at the usual acceptance point for a hypothesis test. If Andrews' acceptance point is put at critical level .90 then we need $A = 11.0$ in (3.5).

In our previous language this amounts to giving the prior 1/11th of the data's information content, reasonably close to our value of 1/15th. The difference comes from Andrews' use of proper Gaussian-like priors rather than the improper Jeffreys' priors used here. Jeffreys' priors give a better match to objective Bayes methods like the BIC (which was not Andrews' purpose of course), via the lemma of Section 2.2, while avoiding most of the asymptotics.

3.2 One-Sided Testing

We can consider the one-sided version of the one-dimensional Gaussian problem (1.1)-(1.2) by changing the bigger model to

$$\mathcal{M}:\ \theta > 0. \tag{3.6}$$

The improper prior $g(\theta) = c$ on $\mathcal{M}$ gives Bayes factor,

$$B(x) = c\ \Phi(x)/\varphi(x) \tag{3.7}$$

where $\Phi(x)$ and $\varphi(x)$ are the standard normal cdf and density. For y a breakeven point, $B(y) = 1$, we can write

$$\log\ B(x) = \log\ \{B(x)/B(y)\} = \frac{x^2 - y^2}{2} + \log\frac{\Phi(x)}{\Phi(y)}. \tag{3.8}$$

The frequentist breakeven point is now $y = 1.282$, the .90 critical level of the one-sided test statistic $S(x) = x$, giving

$$\log B_{\text{freq}}(x) = \frac{x^2 - y^2}{2} + \log \frac{\Phi(x)}{\Phi(y)} \qquad (y = 1.282) \tag{3.9}$$

as the one-sided version of (3.1). This can also be derived from a version of the lemma at (2.17) which takes account of the fact that in the one-sided case $\mathcal{M}_o$ is an extreme point of $\mathcal{M}$ rather than an interior point. (This changes (2.18), the Laplace approximation for $f(\mathbf{x})$.) Andrews (1994) gives similar formulas, for example at his equation (2.7).

Table 6 is the equivalent of Table 5 for the one-sided case. Notice that the critical levels of Fisher's scale, .90, .95, .975, ..., produce nearly the same Bayes factors in both tables. The aposteriori probabilities $\pi(x)$ (assuming $\pi/\pi_0 = 1$) are the same to two digits.

The Bayes factors in Table 6 are closely approximated by taking the prior density $g(\theta)$ to be uniform on $[0, 5.13]$. However a Bayesian who began with the uniform prior $\mathcal{U}[-4.85, 4.85]$ appropriate to the two-sided situation, and then decided that $\theta < 0$ was apriori impossible, would get Bayes factors $B(x)$ not much different than those in Table 6: Using 4.85 instead of 5.13 as the one-sided upper limit for $g(\theta)$ gives an excellent match to the version of (3.8) having breakeven point at $y = \Phi^{-1}(.893)$ instead of $\Phi^{-1}(.90)$.

Tables 5 and 6 show that a frequentist going from a two-sided to a one-sided Gaussian testing problem does so in reasonably coherent Bayesian fashion, essentially by cutting off the negative half of the $\mathcal{U}[-4.85, 4.85]$ prior. We can also use the $\mathcal{U}[-4.85, 4.85]$ prior to investigate frequentist behavior in multiple testing situations, see Remark J in Section 6.

Critical level α	.90	.95	.975	.99	.995	.999
x:	1.282	1.645	1.96	2.33	2.58	3.09
$B_{\text{freq}}(x)$:	1	1.80	3.25	7.24	13.42	57.86
$\pi(x)$:	.50	.64	.76	.88	.93	.98

Table 6. Critical levels and Bayes factors (3.9) for one-sided testing, one-dimensional Gaussian case; breakeven point at $y = 1.282$, the .90 quantile of x under $\mathcal{M}_o$; $\pi(x)$ aposteriori probability of $\mathcal{M}$ assuming prior probability 1/2.

3.3 Multidimensional Gaussian Testing

The Bayesian justification for Fisher's scale of evidence is less satisfactory in higher dimensional testing problems. Suppose that we observe an m-dimensional Gaussian vector with unknown expectation vector θ and covariance matrix the identity

$$\mathbf{x} \sim N_m(\boldsymbol{\theta}, \mathbf{I}), \tag{3.10}$$

and that we wish to test

$$\mathcal{M}_o : \boldsymbol{\theta} = \mathbf{0} \quad \text{versus} \quad \mathcal{M} : \boldsymbol{\theta} \in R^m . \tag{3.11}$$

We will denote $x = \|\mathbf{x}\|$, $\theta = \|\boldsymbol{\theta}\|$, and write $B(x)$ instead of $B(\mathbf{x})$ in the case where $B(\mathbf{x})$ depends only on x, etc. The likelihood ratio statistic $\widehat{B}(x)$ equals $\exp(x^2/2)$ so that (2.27) gives

$$\log B(x) \doteq (x^2 - y^2)/2 \quad \text{where} \quad y^2 = \chi_m^{2(\alpha_0)} . \tag{3.12}$$

Here α_0 is the frequentist breakeven critical level, $\alpha_0 = .90$ on Fisher's scale. The arguments of this section extend easily to the case where $\boldsymbol{\theta}$ is partitioned as $\boldsymbol{\theta} = (\boldsymbol{\theta_0}, \boldsymbol{\theta_1})$, and $\mathcal{M}_o$ is $\boldsymbol{\theta_1} = \mathbf{0}$. We can also take $\mathbf{x} \sim N_m(\boldsymbol{\theta}, \sigma^2\mathbf{I})$ with σ^2 estimated independently from $\widehat{\sigma}^2 \sim \sigma^2\chi^2$, see Remark E of Section 6.

We used (3.12) in the one-dimensional case, with $\alpha_0 = .90$, and showed that it agreed well with a proper Bayesian analysis, starting from $g(\theta)$ uniform on [-4.85, 4.85]. The trouble in higher dimensions is that if we choose $\alpha_0 = .90$, then $B(x)$ in (3.12) is not close to being a Bayes factor for any genuine prior $g(\theta)$.

To show this we begin with $g(\boldsymbol{\theta})$, the density of $\boldsymbol{\theta}$ given $\mathcal{M}$, uniform on a disk of radius u,

$$g(\boldsymbol{\theta}) = \frac{\Gamma(m/2+1)}{\pi^{m/2}u^m} \qquad \text{for} \qquad 0 \le \theta \le u . \tag{3.13}$$

The constant in (3.13) makes g integrate to 1. Using definitions (2.6)-(2.7), and remembering that $g_0(\boldsymbol{\theta})$ is a delta function at zero, it is easy to show that the resulting Bayes factor is

$$B_u(x) = \frac{c_m}{u^m} e^{x^2/2} F_m(u, x) \qquad [c_m = 2^{m/2}\Gamma(m/2+1)] , \tag{3.14}$$

where

$$F_m(u,x) = \text{prob}\{\chi^2_m(x^2) < u^2\} \ , \tag{3.15}$$

the probability that a non-central χ^2 variate with m degrees of freedom and non-centrality x^2 is less than u^2.

We would like (3.14) to match the objective Bayes formula (3.12) $B(x) = \exp(x^2 - y^2)/2$. Notice that

$$\frac{\partial}{\partial x}\log B_u(x) = \frac{\partial}{\partial x}\log B(x) + \frac{\partial}{\partial x}\log F_m(u,x) \ . \tag{3.16}$$

The last term is always negative so $\frac{\partial}{\partial x}\log B_u(x) < \frac{\partial}{\partial x}\log B(x)$.

The value $u = u_0$ that makes $y = [\chi^{2(\alpha_0)}_m]^{1/2}$ the breakeven point satisfies $B_{u_0}(y) = 1$, or according to (3.14),

$$u_0 = [c_m e^{y^2/2} F_m(u_0, y)]^{1/m} \ . \tag{3.17}$$

For $m = 1$ and $y = 1.645$ we get $u_0 = 4.85$, the value used in Section 3.1. The second term $\frac{\partial}{\partial x}\log F_1(u_0,x)|_y$ in (3.16) is only -.0025 in this case, accounting for the good match between $B(x)$ and $B_{4.85}(x)$ in figure 2. Numerical calculations show that this match breaks down as the dimension m increases: the last term in (3.16) grows large, spoiling the agreement between $B_u(x)$ and $B(x)$. We might hope to save things by choosing $g(\boldsymbol{\theta})$ differently, but the next analysis shows not.

If $B_g(x)$ is the Bayes factor (2.6)-(2.7) corresponding to a spherically symmetric prior $g(\boldsymbol{\theta})$, and $B_{\text{freq}}(x,\alpha_0) = \exp[(x^2 - y^2)/2]$ for $y^2 = \chi^{2(\alpha_0)}_m$ define

$$Q(g,\alpha_0) \doteq \int_0^{x_{.999}} |B_g(x)/B_{\text{freq}}(x;\alpha_0) - 1| dx \ / x_{.999} \ , \tag{3.18}$$

$x_{.999} = (\chi^{2(.999)}_m)^{1/2}$, so Q measures the average absolute deviation between $B_g(x)$ and $B_{\text{freq}}(x;\ \alpha_0)$ over the range of Fisher's scale in Table 1.

Linear programming techniques allow us to minimize $Q(g;\ \alpha_0)$ over all possible choices of a spherically symmetric prior distribution $g(\theta)$. The results are shown in Figure 3. We see for example that for $\alpha_0 = .90$ and dimension $m = 6$, the minimum possible value of Q is .17. In order to match $B_{\text{freq}}(x,\alpha_0)$ with accuracy $Q = .011$, the accuracy shown in Figure 2, we need to increase α_0 to .96. This raises the breakeven

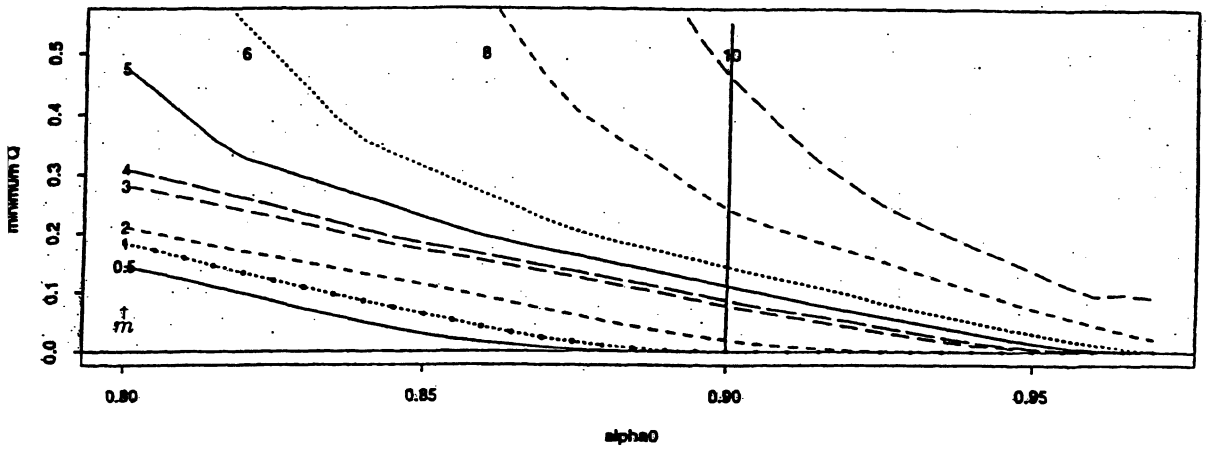

Figure 3. Minimum possible value of average absolute deviation (3.18) as a function of α_0 and dimension m; $m = .5$ is the one-sided one-dimensional problem.

point to $y = (\chi_6^{2(\alpha_0)})^{1/2} = 3.63$, rather than $(\chi_6^{2(.90)})^{1/2} = 3.26$, and decreases the Bayes factor $\exp((x^2 - y^2)/2)$ by a factor of 3.58. Some details of these calculations appear in the Appendix.

We used Fisher's scale of evidence to set the frequentist breakeven point at the .90 quantile of the test statistic. This definition turns out to be near the minimum possible for one-dimensional Gaussian testing problems: the $m = 1$ curve in Figure 3 shows that reducing α_0 to say .80 would make it impossible to get an accurate match with any genuine Bayes factor, the minimum attainable Q being .18. In higher dimensions $\alpha_0 = .90$ itself is too small. A frequentist who desires a proper Bayesian interpretation of p-values needs to set the breakeven quantile α_0 higher. A recipe for doing so appears in Remark G of Section 6.

4 Sample Size Coherency

The BIC assesses evidence in favor of the bigger model $\mathcal{M}$ more cautiously as the sample size n grows larger. The behavior is rooted in notions of coherent decision making, namely that the Bayesian specifications (2.3)-(2.4), whatever they may be, should stay the same for all sample sizes. This simple but important principle, which we have been calling "sample size coherency", causes most of the disagreement between Fisher's and Jeffreys' methods.

It is particularly easy to illustrate sample size coherency for the one-dimensional Gaussian example (1.6)-(1.7). Suppose that at sample size $n = 1$ we used density $g_1(\theta_1)$ for "g" in (2.4) (remembering that in this case g_0 puts all of its probability on zero.). Then, according to the coherency argument, we should still use $g_1(\theta_1)$ at sample size n. Making transformations (1.8),

$$x = \sqrt{n}\,\bar{x} \sim N(\theta, 1), \qquad \theta = \sqrt{n}\,\theta_1\ , \tag{4.1}$$

restores us to situation (1.1), (1.2), now with

$$g(\theta) = g_1(\theta_1/\sqrt{n})/\sqrt{n}\ . \tag{4.2}$$

Equation (4.2) says that $g(\theta)$ is dilating at rate $\sqrt{n}$, and that the crucial value $g(0)$ is going to zero at rate $1/\sqrt{n}$. If $g_1(0)$ is positive and $g_1(\theta)$ is continuously differentiable at zero then it is easy to show that the breakeven point y occurs at

$$y \doteq [\log \frac{n}{2\pi g_1(0)^2}]^{\frac{1}{2}}\ . \tag{4.3}$$

Choosing $g_1(\theta)$ to be a standard $N(0,1)$ density gives $y \doteq \sqrt{\log(n)}$, which is the BIC breakeven point, (1.9). Variations of this argument are familiar in the Bayesian model selection literature, as in Section 3 of Kass and Wasserman (1995).

Frequentist model selection does not obey sample size coherency. Using Fisher's scale in the standard way amounts to using a fixed prior density no matter what the sample size may be, for instance $g(\theta) \doteq \mathcal{U}[\pm\, 4.85]$ in situation (1.6)-(1.8).

We have now reached a clear and unresolvable distinction between the Fisher and Jeffreys approaches to model selection. Which one is right? Here are some of the arguments that have been put forward:

Consistency Under quite general conditions BIC will select the correct model as $n \to \infty$, with $\log B_{\mathrm{BIC}}(\mathbf{x})$ going to $-\infty$ if $\theta \in \mathcal{M}_o$ and $+\infty$ if $\theta \notin \mathcal{M}_o$. The BIC penalty $-\frac{d}{2}\log(n)$ in (2.35) makes this happen (as would any other penalty function going to infinity at a rate slower than n.) The frequentist uses a fixed penalty $\doteq \chi^{2(.90)}_d/2$ and does not achieve consistency; even if $\theta \in \mathcal{M}_o$ there is always a fixed probability, .05 in the most common formulation, of selecting $\mathcal{M}$.

Power Consistency is not much of a comfort in a fixed sample size experiment. To use Neyman-Pearson terminology, what we really want is good size and good power too. Fisher's approach tends to aggressively maximize power by being satisfied with critical

levels (i.e. sizes) in the .95 range. Jeffreys' methods are aimed at a more equitable balance. In using them the statistician risks settling for small power, as suggested by Figure 1.

Pre-experimental power calculations are an important part of frequentist model selection. A common prescription for the one-dimensional Gaussian situation (1.6)-(1.7) is to require sample size $n = (t_1/3.24)^2$, where t_1 is a preliminary guess for the treatment effect θ_1. The constant 3.24 results in .90 power for a two-sided .95 test, at $\theta_1 = t_1$. This kind of calculation, which has a strong Bayesian flavor, fits in well with the $\mathcal{U}[\pm\ 4.85]$ prior ascribed to the frequentist in Section 3.1. Selecting n on the basis of the prior reverses the BIC selection of the prior on the basis of n.

Sufficiency Transformations (1.8) restore problem (1.6), (1.7) to form (1.1), (1.2) no matter what n may be. Why should our assessment of evidence for $\mathcal{M}$ versus $\mathcal{M}_o$ depend in any way upon n? This could be the frequentist's riposte to the Bayesian argument for sample size coherency. It is strengthened by the difficulties of defining "n" in practical problems, see Section 5. The Bayesian argument is unassailable if we begin with a genuine prior but less so if "$g(\theta)$" expresses only a lack of prior knowledge.

Very Large Sample Sizes Raftery (1995) argues for the superiority of Jeffreys' scale and BIC model selection in the social sciences. His main example concerns a multinational social mobility study with data on $n = 113,556$ subjects. An appealingly simple sociometric model explains 99.7% of the deviance but still is rejected by the standard likelihood ratio test at a p-value of 10^{-120}. In this case the BIC penalty function is severe enough to give a Bayes factor in favor of the sociometric model, compared to a saturated model.

Many statisticians intuitively agree with Raftery that something goes wrong with Fisher's scale of evidence when n is very large, and that evidence against $\mathcal{M}_o$ should indeed be judged more cautiously in large samples. A counterargument says that Raftery is putting too much strain on the model selection paradigm: standard hypothesis tests and confidence intervals would show that the sociometric model does not fit the data perfectly but that the deviations from the model are quite small. Gelman and Rubin argue along these lines in the discussion following Raftery (1995), see also Diaconis and Efron (1985).

The Role of $\mathcal{M}_o$ The smaller model $\mathcal{M}_o$ is usually a straw man in Fisher's program, not an interesting scientific theory in its own right, see Section 4 of Efron (1971). Jeffreys' scale shows more respect for $\mathcal{M}_o$, perhaps on the Occam's razor principle that simpler hypotheses are preferred whenever tenable. This shows up in Table 5 (where

the distorting effect of sample size coherency has been avoided by use of the frequentist breakeven value $y = 1.645$). Comparing $B_{\text{freq}}(x)$ with Jeffreys' scale, Table 2, Fisher's "moderate" evidence is "not worth more than a brief mention" for Jeffreys', "strong" is only "positive", etc.

This is more a difference in scientific context than a fundamental Bayesian-frequentist disagreement. Fisher worked in an agricultural field station where sample sizes were small and the data were noisy. Jeffreys' hard-science background suggests more abundant data, better structured models, and a more stringent standard of evidence. It is conceivable that had Jeffreys worked at Rothamsted he would have adjusted his scale downward, conversely for Fisher in geophysics.

Perhaps no single scale of evidence can serve satisfactorily in all contexts. In practice Jeffreys' scale, as opposed to Fisher's, tends to favor $\mathcal{M}_o$, and if implemented by the BIC will do so with increasing force as n increases.

The BIC is not the last word in objective Bayes model selection, though it seems to be the most popular method. Kass and Wasserman (1996) review a variety of techniques developed mostly since 1990 under such names as uninformative priors, reference priors, intrinsic priors, and fractional Bayes factors. All of these methods come close to obeying sample size coherency, and demonstrate BIC-like behavior in large samples.

Sample size coherency is an appealing principle in situations where the statistician actually sees the data set growing in size. It is less compelling in the more common case of fixed n, and taking it literally can lead to the situation seen in Figure 1 where evidence at the .995 level is barely worth mentioning. The selenium experiment of Section 5 has aspects of both fixed and changing sample sizes.

5 The Selenium Experiment

This section uses our methods to produce a combined Bayesian-frequentist analysis of a small but inferentially challenging data set. Table 7 shows total cancer mortality for a double-blind randomized trial of the trace element selenium taken as a cancer preventative, Clark et al. (1996). The original purpose of the trial was to test selenium's ability to prevent the recurrence of carcinoma of the skin. 1312 subjects, all of whom had suffered previous skin cancers, were recruited beginning in 1983 and received either 200 mmg per day of selenium or an identical-looking placebo. The results from 1983 to 1989, labeled "1st Period" in Table 7, did not show any skin cancer reduction in the selenium group. However total cancer mortality, mainly from lung, prostate, and colorectal cancers, did suggest a moderately significant reduction. The p-value of .032 shown in Table 7 is the one-sided binomial probability of seeing 7 or less occurrences of

	Selenium	Placebo	Total	s/N	**p-value**
1st Period:	7	16	23	7/23	**.032**
2nd Period:	22	41	63	22/63	**.00843**
Combined:	29	67	86	29/86	**.00124**

Table 7 Total cancer mortality in the selenium experiment; p-values are one-sided, based on binomial distribution (5.1). Data from Table 5 of Clark et al. (1996).

$s \sim$ binomial $(N, 1/2)$, splitting the probability atom at 7,

$$.032 = \frac{1}{2^N}\Big[\sum_{i=0}^{s-1}\binom{N}{i} + \frac{1}{2}\binom{N}{s}\Big], \tag{5.1}$$

with $N = 23$ and $s = 7$.

New funding was obtained, allowing a second trial period from 1990-1993. At the beginning of this period total cancer mortality was officially listed as a "secondary endpoint". The primary endpoint remained skin cancer, but given the results of the 1st period it seems fair to assume that the investigators' attention was now focused on total cancer mortality. Lung, prostate, and colorectoral cancer incidence were also listed as secondary endpoints, see Remark A.

We now consider selecting between the models

$$\begin{aligned} \mathcal{M}_o &: \text{ selenium has no effect on total cancer mortality} \\ \mathcal{M} &: \text{ selenium has an effect on total cancer mortality.} \end{aligned} \tag{5.2}$$

By conditioning on N the total number of cancer deaths in both groups, $N = 23$ for the first period and 63 for the second period, we can model s, the number of deaths in the selenium group, by

$$s \sim \text{binomial}(N, \theta) \qquad \text{with} \begin{cases} \mathcal{M}_o : \ \theta = .5 \\ \text{versus} \\ \mathcal{M} : \ \theta \neq .5 \end{cases} \tag{5.3}$$

as the competing models. We will also consider the one-sided version $\mathcal{M} : \ \theta < .5$. In either case we are dealing with a one-dimensional problem, $\dim(\mathcal{M}) - \dim(\mathcal{M}_o) = 1$.

How strong is the evidence for $\mathcal{M}$ and against $\mathcal{M}_o$? The combined-data p-value of .00124, even doubled for two-sided testing, is between "very strong" and "overwhelming" on Fisher's scale. However, this ignores the data-mining aspects of the experiment, which

used the first period's outcome to change the focus of interest. Restating the frequentist results in terms of Bayes factors helps clarify the strength of evidence for selenium's cancer-preventing ability.

The simplest approach considers only the 2nd period data since it was then that attention was focused on total cancer mortality. A very quick way of doing the calculations transforms the one-sided binomial p-value .00843, computed as in (5.1) into an approximate normal deviate $x = \Phi^{-1}(1 - .00843) = 2.39$, and then calculates the Bayes factor for $\mathcal{M}$ from the one-dimensional Gaussian approximation formula (3.1),

$$B_{\text{freq}}(s) \doteq \exp\{(2.39^2 - 1.645^2)/2\} = 4.49 \; . \tag{5.4}$$

Starting from the conventional prior odds ratio $\pi/\pi_0 = 1$, the Bayes rule results in aposteriori probability $\pi(s) = .82$ for $\mathcal{M}$. Remark C of Section 6 shows that this is nearly the same as the aposteriori probability for the event of actual interest $\{\theta < .5\}$, even though we began the analysis with $\mathcal{M} : \theta \neq .5$. Remark B shows that the Gaussian approximation (5.4) works quite well in this case.

We might instead begin our analysis with the one-sided model $\mathcal{M} : \theta < .5$, on the grounds that the 1st period results removed most of our apriori probability on $\theta > .5$. This entitles us to use the one-sided formula (3.9) for the Bayes factor, giving $B_{\text{freq}}(s) = 8.42$, nearly double (5.4), and $\pi(s) = .89$. These results are shown in Table 8.

Instead of focusing on the 2nd period we might consider the combined data for both periods. If so we need to adjust our inferences to account for the fact that total cancer

	$\hat{B}$	B_{freq}		B_{BIC}	
		2-sided	1-sided	$n = N$	$n = 1312$
2nd Period:	18.36	4.49	8.42	2.31	0.51
$[\pi(s)]$:		**[.82]**	**[.89]**	**[.70]**	**[.34]**
Combined:	103.8	25.14		11.19	2.87
$[\pi(s)]$:		**[.86]**		**[.74]**	**[.42]**

Table 8 Approximate Bayes factors for the Selenium experiment, as explained in text; boldface numbers are aposteriori probabilities for selenium having an effect, assuming $\pi/\pi_o = 1$ for 2nd period, $\pi/\pi_o = 1/4$ for combined data.

mortality was not the original primary endpoint. We will do this by setting the prior odds ratio to be

$$\pi/\pi_0 = 1/4 \, . \tag{5.5}$$

This is rather arbitrary of course but it cannot be wildly optimistic: after the 1st period results, which yielded a Bayes factor of only 1.44, the investigators effectively raised total cancer mortality to the status of primary endpoint, presumably with odds ratio near the conventional value 1 we have been using.

The combined data has $s = 29$ and $N = 86$. The two-sided normal approximation used in (5.4) is now

$$B_{\text{freq}}(s) = 25.14 \, . \tag{5.6}$$

Bayes' rule with $\pi/\pi_0 = 1/4$ gives aposteriori probability $\pi(s) = .86$ for $\mathcal{M}$. One-sided testing gives $B(s) = 47.46$ and $\pi(s) = .92$, but now we lack scientific justification for a one-sided analysis.

All of our frequentist-cum-Bayesian analyses yielded aposteriori probabilities $\pi(s)$ in the range .82 to .89 for selenium being efficacious in reducing cancer deaths. Perhaps this seems disappointing given the striking p-values in Table 7, but as Table 5 shows this is what we get from "strong" evidence on Fisher's scale. As far as Jeffreys' scale is concerned, $B_{\text{freq}}(s)$ never gets stronger than "positive" (remembering to divide by 4 for the combined data).

BIC analysis is predictably more pessimistic about selenium's efficacy. The BIC Bayes factor (1.5) for the combined data is

$$B_{\text{BIC}}(s) = 103.8/\sqrt{n} \tag{5.7}$$

Taking $n = 86$, the number of deaths, gives $B_{\text{BIC}}(s) = 11.19$, and $\pi(s) = .74$ starting from $\pi/\pi_0 = 1/4$. If we take $n = 1312$, the number of subjects in the study then $\pi(s) = .42$. Raftery (1986) makes a good argument for preferring $n = 86$ to 1312, but in general there is not a firm prescription for "n". If the data was collected in pairs should "n" be $n/2$? Kass and Wasserman (1995) aptly characterize the sample size question as "subtle but important", see also Lauritzen's commentary on O'Hagan (1995). These difficulties are avoided in the frequentist formulation, at the expense of ignoring sample size coherency.

Sample size coherency is unimportant in the usual fixed size experiment where the frequentist approach operates to best advantage. The selenium experiment is somewhere intermediate between fixed sample size and sequential. If we think of it as occurring in two stages, then the sample size coherency argument suggests dividing the 2nd period Bayes factors by $\sqrt{2}$, giving $B_{\text{freq}}(s) = 5.95$ and $\pi(s) = .86$ for the one-sided analysis. This kind of Bayesian correction is not much different than the standard frequentist approach to multiple testing situations, see Remark J.

6 Remarks

Remark A. Lung, prostate, and colorectal cancer incidence rates were also flagged as important secondary endpoints for the 2nd period of the selenium trial. Incidence of all three together was 17 in the selenium group versus 29 in the placebo group during the 1st period, giving binomial significance level .040 according to (5.1). 2nd period incidences were 21 versus 56, significance level $2.79 \cdot 10^{-5}$. Now the one-sided 2nd period Bayes factor corresponding to $B_{\text{freq}}(s) = 8.42$ in Table 8 is 1642. The very strong 2nd period results are a reminder that the two periods differ in the amount of selenium experienced by the treatment group.

Remark B. We do not need to rely on the Gaussian approximation (3.1) for the selenium analysis. Let y be the .95 percentile point for a binomial$(N, .5)$ distribution, calculated by interpolation of the "split-atom" cdf as in (5.1), so y is a .90 breakeven point for two-sided testing. Then according to (2.27)

$$B_{\text{freq}}(s) \doteq \widehat{B}(s)/\widehat{B}(y) = \frac{s^s(N-s)^{N-s}}{y^y(N-y)^{N-y}}\,. \tag{6.1}$$

This gives 4.63 instead of 4.49 for the two-sided 2nd period value of B_{freq} and 26.80 instead of 25.14 for the combined data. We see that (3.1) works quite well here.

Remark C In the two-sided formulation (5.2), 2nd period data, we calculated $B_{\text{freq}}(s) = 4.49$ in favor of $\mathcal{M}: \theta \neq .5$ versus $\mathcal{M}_o: \theta = .5$. However, we are really interested in the one-sided alternative $\theta < .5$. To this end we can state the results as follows: the aposteriori probability of $\mathcal{M}: \theta \neq .5$ is $\pi(s) = .82$, and given that $\mathcal{M}$ is true, the aposteriori probability that $\theta < .5$, (using $g(\theta) = \mathcal{U}[\pm\, 4.85]$) is about $\Phi(2.39) = .992$. This gives .81 for the aposteriori probability of $\{\theta < .5\}$.

Remark D Suppose that in situation (1.1), (1.2) we observe $x = 1.96$ and wish to estimate $\gamma = \text{prob}\{\theta > 0 \mid x\}$. The prior distribution of Section 3.2 appropriate to

one-sided model selection, $\pi/\pi_0 = 1$ and $g(\theta)$ uniform on $[0, 5.13]$, gives aposteriori Bayes estimate $\gamma_1 = .76$. Section 3.1's prior for two-sided model selection, $\pi/\pi_0 = 1$ and $g(\theta)$ uniform on $[-4.85, 4.85]$, gives $\gamma_2 = .64$. This decrease is reasonable since the second prior puts only half as much probability on the positive axis.

Both γ_1 and γ_2 are much smaller than the value $\gamma_3 = .975$ we get using the standard objective prior for estimation, which has $g(\theta)$ constant over $(-\infty, \infty)$, with no special treatment for $\theta = 0$. This is the difference between model selection, which puts a bump of probability on, or at least near, $\mathcal{M}_o : \theta = 0$, and estimation which does not. The estimation paradigm is often more appropriate than model selection. If we are trying to choose between constant, linear, or quadratic regression functions for prediction purposes, then there may not be any reason to assign bumps of prior probability to zero values of the regression coefficients $\beta_0, \beta_1, \beta_2$. Efron and Tibshirani (1997) consider discrete selection problems from the "smooth prior" point of view. See also Lindley and O'Hagan's disagreement in the discussion following O'Hagan (1995).

Remark E The multidimensional Gaussian situation (3.10), (3.11) can be generalized to its more common form where we observe

$$\mathbf{x} \sim N_m(\boldsymbol{\theta}, \sigma^2 \mathbf{I}) \quad \text{independent of} \quad \widehat{\sigma}^2 \sim \sigma^2 \chi^2_q / q \,, \tag{6.2}$$

$\boldsymbol{\theta} = (\boldsymbol{\theta}_0, \boldsymbol{\theta}_1)$, $\dim(\boldsymbol{\theta}_1) = d$, and wish to select between

$$\mathcal{M}_o : \boldsymbol{\theta}_1 = \mathbf{0} \quad \text{versus} \quad \mathcal{M} : \boldsymbol{\theta} \in R^m \,. \tag{6.3}$$

Applying the lemma (2.17) in form (2.27), (2.28) gives estimated Bayes factor

$$B(\mathbf{x}, \widehat{\sigma}^2) \doteq \left[\frac{1 + \frac{d}{q} F_{d,q}}{1 + \frac{d}{q} F^{(\alpha_0)}_{d,q}} \right]^{\frac{d+q}{2}} , \tag{6.4}$$

where $F_{d,q}$ is the usual F statistic for testing $\mathcal{M}_o$ versus $\mathcal{M}$, and α_0 is the breakeven quantile, $\alpha_0 = .90$ on Fisher's scale. The shortcut formula (2.30) amounts to replacing $F^{(\alpha_0)}_{d,q}$ with its limit as $q \to \infty$.

Remark F Is Fisher's scale of evidence too liberal, as charged in the Bayesian literature? The answer depends at least partly on scientific context, though much of the contextual effect can be mitigated by an honest choice of the prior odds ratio π/π_0. Biomedical research has employed Fisher's scale literally millions of times, with generally

good results. In crucial situations the scale may be implicitly tightened, for instance by the F.D.A. requirement of *two* .95 significant studies to qualify a new drug.

We have argued that Fisherian hypothesis testing has a reasonable Bayesian interpretation, at least if one is willing to forego sample size coherency. Section (3.3) shows the Bayesian interpretation breaking down in multidimensional testing problems, with the suggestion that Fisher's choice of breakeven quantile $\alpha_0 = .90$ needs to be increased.

Remark G Here is another argument for using bigger values of α_0 in higher dimensions. The BIC penalty function $d \cdot (\log\ n)/2$, (2.35), is linear in the dimensional difference d. By considering a nested sequence of models it is easy to show that this must always be the case if the penalty function depends only on d and n. However, the frequentist penalty function $\chi_d^{2(\alpha_0)}/2$ is not linear in d.

We can enforce linearity by replacing .90 in approximation (2.30) with $\alpha_o(d)$, where $\alpha_0(d)$ satisfies

$$\chi_d^{2(\alpha_0(d))} = d \cdot \chi_1^{2(\alpha_0(1))}\ , \tag{6.5}$$

$\alpha_0(1)$ being the breakeven quantile in dimension 1. Doing so makes $\alpha_0(d)$ increase with d, as shown in Table 9. The choice $\alpha_0(1) = .86$ is the minimum that keeps the discrepancy measure Q, (3.18) reasonably small (about .05) for dimensions 1-6.

d:	1	2	3	4	5	6
$\alpha_0(1) = .86$:	.86	.887	.912	.931	.946	.958
$\alpha_0(1) = .90$:	.90	.933	.956	.971	.981	.987

Table 9 Quantile $\alpha_0(d)$ satisfying linearity relationship (6.5); for two choices of $\alpha_0(1)$.

Remark H In the Gaussian situation (1.6)-(1.7), and in more general problems too, the BIC depends on a prior density $g(\theta)$ having $1/n$ of the data's information for estimating θ. Jeffreys' original proposal used a slightly more diffuse prior with information about $1/(\frac{\pi}{2}n)$, see Section 3 of Kass and Wasserman (1995). Berger and Pericchi (1993) suggest the value $1/(n/n_0)$, where $n_0 \geq 1$ is a fixed small constant. O'Hagan's (1995) proposal leads to larger values of n_0, for robustness purposes, an idea criticized in Berger and Mortera's discussion.

Remark I The calculations leading to the lemma's result (2.17) can be carried out more accurately by using a higher-order version of Laplace's method as in the appendix of Tierney and Kadane (1986). We get

$$\frac{B(\mathbf{x})}{B(\mathbf{x}^\dagger)} = \frac{\widehat{B}(\mathbf{x})}{\widehat{B}(\mathbf{x}^\dagger)}\left[1 + \frac{a(\widehat{\theta}) - a(\widehat{\theta}^\dagger)}{n} + O(n^{-2})\right], \tag{6.6}$$

where $a(\theta)$ is $O(1)$. In applications like the logistic regression example (2.26), $\widehat{\theta}^\dagger$ is $O_p(n^{-\frac{1}{2}})$. The interesting uses of (6.6), the ones where $B(\mathbf{x})$ is of moderate size, also have $a(\widehat{\theta})$ of order $O_p(n^{-\frac{1}{2}})$, so $a(\widehat{\theta}) - a(\widehat{\theta}^\dagger)$ is $O_p(n^{-\frac{1}{2}})$ in (6.6). This suggests that approximation (2.17) is actually third-order accurate, giving the kind of good small-sample results seen in (2.26). However, this argument fails for the standard asymptotic calculation where the true θ is a fixed point in $\mathcal{M}$, in which case $a(\widehat{\theta}) - a(\widehat{\theta}^\dagger)$ is $O_p(1)$.

Remark J The $\mathcal{U}[\pm 4.85]$ prior density figuring in the discussion of Section 3.1 also gives a rough Bayesian justification for the standard frequentist approach to multiple testing. Suppose that we observe J independent Gaussian variates, each having a possibly different expectation,

$$x_i \overset{i.i.d.}{\sim} N(\theta_i, 1), \qquad i = 1, 2, \ldots, J, \tag{6.7}$$

and that we wish to test

$$\begin{aligned} &\mathcal{M}_o: \quad \text{all} \quad \theta_i = 0 \quad \text{versus} \\ &\mathcal{M}: \quad \text{one of the} \quad \theta_i \quad \text{not zero.} \end{aligned} \tag{6.8}$$

For the conditional prior distribution of $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_J)$ given model $\mathcal{M}$, we take an equal mixture of the J distributions

$$g_i(\boldsymbol{\theta}): \quad \theta_i \sim \mathcal{U}(-4.85, 4.85) \quad \text{and} \quad \theta_j = 0 \quad \text{for} \quad j \neq i, \tag{6.9}$$

$i = 1, 2, \ldots, J$. If $J = 2$, $g(\boldsymbol{\theta})$ is a cross-shaped distribution.

It is easy to calculate the Bayes factor for $\mathcal{M}$ versus $\mathcal{M}_o$ if the observed data vector is

$$\mathbf{x} = (x_1, 0, 0, \ldots, 0), \tag{6.10}$$

in which case

$$B(\mathbf{x}) \doteq \sqrt{2\pi}\,\frac{e^{x_1^2/2} + J - 1}{2J \cdot 4.85}. \tag{6.11}$$

J:	2	3	4	6	8	10
Bonferroni:	1.96	2.13	2.24	2.39	2.50	2.58
(6.11):	1.95	2.13	2.25	2.41	2.52	2.60

Table 10. Simultaneous testing; comparison of critical point for the simultaneous .90 Bonferroni test with the breakeven point (6.11) for the Bayesian analysis suggested by the $\mathcal{U}[\pm\, 4.85]$ distribution; J is the number of simultaneous tests.

Using this approximation, the breakeven point $\mathbf{y} = (y, 0, 0, \ldots, 0)$ occurs at

$$y = \pm\, \{2 \log [\frac{2J \cdot 4.85}{\sqrt{2\pi}} - (J-1)]\}^{1/2} . \tag{6.12}$$

Table 10 compares the Bayesian breakeven point y from (6.12) with the .90 critical point $y = \Phi^{-1}(1-.05/J)$ of the usual frequentist Bonferroni procedure for J simultaneous tests. Once again we see that the frequentist is behaving in a reasonably Bayesian fashion, although caution is warranted here because of the somewhat special choices made in (6.8)-(6.10).

Summary Frequentist hypothesis testing, as interpreted on Fisher's scale of evidence, is the most widely used model selection technique. Jeffreys' theory of Bayes factors, as implemented by objective formulas like the BIC, implies that Fisher's scale is badly biased against smaller models, especially in large samples. This paper compares the two theories by giving an interpretation of Fisher's scale in terms of Bayes factors, as far as that is possible, along these lines:

• An ideal form for objective Bayes factors is developed in Section 2, $B(\mathbf{x}) \doteq \widehat{B}(\mathbf{x})/\widehat{B}(\mathbf{y})$, with $\mathbf{y}$ a breakeven point: $B(\mathbf{y}) = 1$.

• Fisher's theory is interpreted as putting the breakeven point at the 90th percentile of the test statistic, leading to a prescription $B_{\text{freq}}(\mathbf{x})$ for the frequentists' implied Bayes factor, as in (3.1).

• For the one-dimensional Gaussian problem Figure 2 shows that $B_{\text{freq}}(\mathbf{x})$ is close to the actual Bayes factor if the prior density on the bigger model is uniform over $[-4.85, 4.85]$. This portrays the frequentist as a somewhat unobjective Bayesian. By contrast, the BIC amounts to choosing the prior density uniform over $\pm\, [\pi \cdot n/2]^{1/2}$, getting wider as the sample size n increases. Roughly speaking, the frequentist assigns the prior $1/15$ of the data's information, compared to $1/n$ for the BIC.

• The BIC, and other objectivist Bayesian methods, behave this way because of "sample-size coherency", the principle that a Bayesian specification for the model selection problem should be consistent over all possible sample sizes. Section 4 discusses arguments for and against this principle.

• Sample size coherency causes most of the disagreement between Fisher's and Jeffreys' methods, but even after correcting for it, Jeffreys' scale remains somewhat more inclined toward the smaller hypothesis $\mathcal{M}_0$. This is more a matter of scientific context than frequentist/Bayesian dispute.

• The argument that $B_{\text{freq}}(\mathbf{x})$ is close to being a proper Bayes factor weakens in higher dimensional problems. To restore it we need to increase the breakeven point on Fisher's scale above the $\alpha_0 = .90$ quantile of the test statistic. Figure 3 suggests an α_0 of about .96 for a six-dimensional Gaussian testing problem. It also shows that in the one-dimensional case we could not choose α_0 much smaller than .90.

• $B_{\text{freq}}(\mathbf{x})$ provides a Bayesian interpretation for standard p-values, as exemplified in Section 5's analysis of the selenium data.

Appendix

Numerical methods were used in Section 3.3 to find the spherically symmetric prior g on the alternative of the m-dimensional Gaussian test (3.10)-(3.11) yielding a Bayes factor B_g most closely matching the effective frequentist Bayes factor, B_{freq}, defined in (2.30). We will describe the calculations in detail here. The one-sided test (labelled $m = 0.5$ in Figure 3) may be treated similarly, as will be shown later.

As in Section 3.3 we put $x = \|\mathbf{x}\|$ and $\theta = \|\boldsymbol{\theta}\|$, and write $B_g(x)$ instead of $B_g(\mathbf{x}), g(\theta)$ instead of $g(\boldsymbol{\theta})$, etc. In this notation, we define the following objective function, a measure of distance between the Bayes factors B_g and B_{freq}:

$$Q(g, \alpha) = \int_0^{x_{.999}} |B_g(x)/B_{\text{freq}}(x; \alpha_0) - 1| dx/x_{.999} \ . \tag{A.1}$$

Here $x_{.999} = (\chi^{2(.999)}_m)^{\frac{1}{2}}$. Other objective functions are possible; see below. The function Q is convenient because the problem can be formulated as a linear program. Formally, we want to solve the following:

$$\underset{g}{\text{minimize}} \qquad Q(g, \alpha_0)$$

(O1) subject to 1. g is a density on $\mathcal{R}^m$, spherically symmetric around 0,

2. $B_g(y) = B_{\text{freq}}(y; \alpha_0) = 1$, where $y^2 = \chi_m^{2(\alpha_0)}$.

The second constraint restricts consideration to priors g which result in a Bayesian test with the same "breakeven point" y (see Section 2.3) as the frequentist test.

The g minimizing (O1) for dimension $m = 1$ and $\alpha_0 = .90$ is given in (3.3). The minimum $q(g, \alpha_0)$ over a range of m and α_0 are shown in Figure 3.

Writing the spherically symmetric density g as $g(\boldsymbol{\theta}) \equiv h(\theta)$ we have

$$\int_0^\infty h(\theta)\theta^{m-1}d\theta = \frac{\Gamma(m/2)}{2\pi^{m/2}} . \tag{A.2}$$

If $m = 1$, for example, $\int_0^\infty h(\theta)d\theta = 1/2$.

An expression for B_{freq} is given in (3.12). We can calculate

$$B_g(x) = \int_0^u s_x(\theta)h(\theta)d\theta , \tag{A.3}$$

where

$$s_x(\theta) = \theta^{m-1}e^{-\theta^2/2}H(x\theta),\ \theta \geq 0 , \tag{A.4}$$

$$H(z) = \frac{2\pi^{(m-1)/2}}{\Gamma((m-1)/2)} \int_0^\pi (\sin\theta)^{m-2}e^{z\cos\theta}d\theta\ z \geq 0 . \tag{A.5}$$

To solve (O1) numerically we will discretize over the ranges of both x and θ. Let $\theta_1, \ldots, \theta_{n_\theta}$ be n_θ equally-spaced values from 0 to u, and $x_1, \ldots, x_{n_x}$ be n_x equally-spaced values from 0 to $x_{.999}$. We will approximate h by an n_θ-vector γ, with $h(\theta_j) = \gamma_j, j = 1, \ldots, n_\theta$.

Define the n_θ-vectors $b_i, i = 1, \ldots, n_x$, to have entries

$$b_{ij} = s_{x_i}(\theta_j)/B_{\text{freq}}(x_i, \alpha_0),\ j = 1, \ldots, n_\theta , \tag{A.6}$$

evaluated using (3.12), (7.4), and (7.5). The function H in (A.5) can be evaluated using numerical integration. Then a discrete approximation to Q is

$$Q'(\gamma; \alpha_0) = \sum_{i=1}^{n_x} |b_i^T\gamma - 1| , \tag{A.7}$$

so a discrete version of (O1) is

$$\text{(O2)} \qquad \begin{array}{ll} \underset{\gamma}{\text{minimize}} & Q'(\gamma;\alpha_0) \\ \text{subject to} & 1a.\ \gamma_j \geq 0, j = 1,\ldots,n_\theta\ . \\ & 1b.\ \sum_{j=1}^{n_\theta} \gamma_j \theta_j^{m-1} = \Gamma(m/2)/(2\pi^{m/2}) \\ & 2.\ \sum_{j=1}^{n_\theta} \gamma_j s_y(\theta_j) = 1, \text{where } y^2 = \chi_m^{2(\alpha_0)}. \end{array}$$

Constraint 1 in (O1) has been split into two parts; part 1b. follows from (A.2).

To express (O2) in the standard form of a linear program we can introduce dummy variables $t_1,\ldots,t_{n_x}$, and write (O2) as

$$\text{(O3)} \qquad \begin{array}{ll} \underset{\gamma}{\text{minimize}} & \sum_{i=1}^{n_x} t_i \\ \text{subject to} & \text{1a., 1b., and 2. in (O2), and} \\ & 3.\ t_i > b_i^T\gamma - 1, t_i > 1 - b_i^T\gamma,\ i = 1,\ldots,n_x\ . \end{array}$$

One-sided testing

The optimization problem (O2) may be adapted to the one-sided testing case of Section 3.2 with the following modifications. We now have B_{freq} as in (3.9), so that the breakeven point becomes $y = \Phi^{-1}(\alpha_0)$, and the upper limit of integration in (A.1) is $x_{.999} = \Phi^{-1}(.999)$. The b_{ij} in (A.6) are calculated using (3.9) and

$$s_\theta(x) = e^{-\theta^2/2 + x\theta}\ . \tag{A.8}$$

Finally, since g (and γ) in this case has mass 1 on the positive real line, we replace constraint 1b. of (O2) with

$$\text{1b.}\ \sum_{j=1}^{n_\theta} \gamma_j = 1\ .$$

Choice of the objective function:

Note that since B_{freq} is $O(e^{x^2})$ for large x, while B_g is only $O(e^x)$ for any g, the two functions can, at most, remain close up to some finite x. We therefore restrict the integration in (A.1) below $x_{.999}$, Table 1's upper bound on plausible values for x under the null model.

Other objective functions could be considered. The squared-error criterion

$$Q_1(g,\alpha) = \int_0^{x_{.999}} (B_g(x)/B_{\text{freq}}(x;\alpha_0) - 1)^2 dx/x_{.999} \qquad (A.9)$$

for example results in a quadratic, rather than linear, program. Besides computational simplicity, we prefer Q to Q_1 because Q penalizes values of B_g far from B_{freq} less heavily, and so matches the two functions more closely near $x = y$ where their values are equal. This is the region of primary interest.

The objective function could compare the Bayes factors on the log scale, using, for example,

$$Q_2(g,\alpha) = \int_0^{x_{.999}} |\log B_g(x) - \log B_{\text{freq}}(x;\alpha_0)| dx/x_{.999} \; . \qquad (A.10)$$

Minimization over g is far more difficult in this case though, since Q_2, unlike Q or Q_1, is not convex in g.

REFERENCES

Andrews, D. (1994). The large sample correspondence between classical hypothesis tests and Bayesian posterior odds tests. *Econometrica* **62**, 1207-1232.

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

Berger, J.O. and Sellke, T. (1987). Testing a point null hypothesis: the irreconciliability of p-values and evidence. *J. Amer. Statist. Asso.* **82**, 112-135.

Berger, J.O. and Pericchi, L.R. (1993). The intrinsic Bayes factor for model selection. Technical Report 93-43C, Dept. of Statistics, Purdue Univ.

Berger, J., Boukai, B., and Wang, T. (1997). Unified frequentist and Bayesian testing of a precise hypothesis. *Statist. Sci.* **12**, 133-160.

Casella, G. and Berger, R. (1987). Reconciling Bayesian and frequentist evidence in the one-sided testing problem. *J. Amer. Statist. Asso.* **82**, 106-111.

Clark, L.C. Combs, J.F., Turnbull, B.W., et al. (1996). Effects of Selenium supplementation for cancer prevention in patients with carcinoma of the skin. *Jour. Amer. Medical Assoc.* **276**, 1957-1963, Editorial p. 1964-1965.

Efron, B. (1971). Does an observed sequence of numbers follow a simple rule? (Another look at Bode's Law). *J. Amer. Statist. Asso.* **66**, 552-568.

Efron, B. and Tibshirani, R. (1997). The problem of regions. Technical Report, Dept. of Statistics and Biostatistics, Stanford Univ.

Efron, B. and Diaconis, P. (1985). Testing for independence in a two-way table: new interpretations of the chi-square statistic (with discussion and rejoinder). *Ann. Stat.* **13**, 845-913.

Finney, D. (1947). The estimation from individual records of the relationship between dose and quantal response. *Biometrika* **34**, 320-335.

Fisher, R.A. (1954). *Statistical Methods for Research by Workers*, (12th ed). Haffner, New York.

Good, I.J. (1947 and 1950). *Probability and the Weighing of Evidence.* Griffen, London.

Good, I.J. (1969). A subjective evaluation of Bode's law and an objective test for approximate numerical rationality. *J. Amer. Statist. Asso.* **64**, 23-66.

Jeffreys, H. (1935). Some tests of significance treated by the theory of probability. *Proc. Cambridge Philosophical Soc.* **31**, 203-222.

Jeffreys, H. (1961). *Theory of Probability*, 3rd ed. Oxford University Press.

Kass, R.E. and Raftery, A.E. (1995). Bayes factors. *J. Amer. Statist. Asso.* **90**, 773-795.

Kass, R.E. and Wasserman, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *J. Amer. Statist. Asso.* **90**, 928-934.

Kass, R.E. and Wasserman, L. (1996). The selection of prior distributions by formal rules. *J. Amer. Statist. Asso.* **91**, 1343-1377.

O'Hagan, A. (1995). Fractional Bayes factors for model comparison. *J. Roy. Statist. Soc. Ser. B* **57**, 99-138.

Pettit, L. (1992). Bayes factors for outlier models using the device of imaginary observations. *J. Amer. Statist. Asso.* **87**, 541-545.

Raftery, A.E. (1986). A note on Bayes factors for log-linear contingency table models with vague prior information. *J. Roy. Statist. Soc. Ser. B* **48**, 249-250.

Raftery, A.E. (1995). Bayesian model selection in social research. *Sociology Methodology, 1995* (P.V. Mauden, ed.) 111-195, Blackwells, Cambridge, Mass.

Raftery, A.E. (1996). Approximate Bayes factors and accounting for model uncertainty in generalized linear models. *Biometrika* **83**, 251-266.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461-464.

Smith, A.F. and Spiegelhalter, D.J. (1980). Bayes factors and choice criteria for linear models. *J. Roy. Statist. Soc. Ser. B* **42**, 213-220.

Spiegelhalter, D.J. and Smith, A.F. (1982). Bayes factors for linear and log-linear models with vague prior information. *J. Roy. Statist. Soc. Ser. B* **44**, 377-387.

Tierney, L. and Kadane, J. (1986). Accurate approximation for posterior moments and marginal densities. *J. Amer. Statist. Asso.* **81**, 82-86.

Tierney, L., Kass, R., and Kadane, J. (1989). Fully exponential Laplace approximations to expectations and variances of nonpositive functions. *J. Amer. Statist. Asso.* **84**, 710-716.

DISCUSSION

R.E. Kass

Carnegie-Mellon University

At this moment in the history of statistics, there seems to be less interest in the great Bayesian/frequentist divide than there had been in the nineteen seventies and eighties, when Efron (1986) asked, "Why isn't everyone a Bayesian?" We are all eager to get on to solving the many challenges of contemporary data analysis. Yet, we have our foundational conscience speaking to us; it continues to prod, with occasional welcome reminders from papers such as this one by Efron and Gous. How can these two great paradigms co-exist in peace? Where are the resolutions? What conflicts are irresolvable? And where does this leave us?

To me, the issues raised in this paper continue to be interesting. I find the authors' discussion clear and their new results informative. On the other hand, there are those in the Bayesian camp who see little relevance of all this to things they care about. Nearly all statisticians I have come across, regardless of philosophical persuastion, freely admit to *thinking* Bayesianly. Among the converted, however, there is a kind of Cartesian credo: "I think Bayesianly, therefore I am Bayesian." The impatience of the true believers comes in part from their taking the next step: "I think Bayesianly, therefore I must place all of my statistical work within the Bayesian paradigm."

A second, equally fundamental difficulty many Bayesians (and some frequentists) have with the perspective articulated in this paper, is in the importance it places on hypothesis testing and model selection. As the authors note, a recent version of this dissenting point of view is in Gelman and Rubin's discussion of Raftery (1995).

One might say that a major practical goal of this paper is to dissect Jeffreys's remark that his methods and Fisher's would rarely lead to different conclusions (Jeffreys, 1961, p. 435): "In spite of the difference in principle between my tests and those based

R. E. Kass is Professor and the Head , Department of Statistics, Carnegie-Mellon University, Pittsburgh, PA 15213-3890, U.S.A; email: kass@stat.cmu.edu.