Bayesian and Frequentist Model Selection

By

Bradley Efron and Alan Gous

Technical Report No. 193

May 1997

Prepared Under the Auspices

Of

Public Health Service Grant

2 R01 CA59039-21

Also supported by National Science Foundation Grant DMS95-04379 and issued as Technical Report No. 513, Department of Statistics, Stanford University

Division of Biostatistics

Stanford University

Stanford, California
Bayesian and Frequentist Model Selection

Bradley Efron and Alan Gous

Abstract

This paper concerns the relationship between Bayesian and frequentist model selection techniques. Some general theoretical similarities are discussed, particularly in the context of simple Gaussian hypothesis testing. We show that frequentist methods have a Bayesian justification, but one that assumes stronger priors than standard objective Bayes criteria such as the BIC. This difference has enormous practical consequences in large samples. Part of difference is historical, deriving from Fisher's and Jeffreys' competing scales of evidence, but a more important factor concerns the Bayesian concept of coherent behavior for increasing sample sizes. A recent investigation of selenium as a cancer preventative is used to show how p-values can be translated into Bayes factors.

The authors are grateful to Robert Tibshirani for helpful conversations, examples, and references.
Bayesian and Frequentist Model Selection

Bradley Efron and Alan Gous

1. Introduction. Here is a particularly simple model selection problem: we observe

\[ x \sim N(\theta, 1), \quad (1.1) \]

a normally distributed quantity with unknown expectation \( \theta \) and variance 1, and wish to choose between two models, a smaller one and a bigger one,

\[ M' : \theta = 0 \quad \text{versus} \quad M : \theta \neq 0. \quad (1.2) \]

Typically the choice between \( M' \) and \( M \) involves some question of scientific interest. In the example of Section 4, the question is whether or not ingesting selenium reduces cancer mortality.

The standard frequentist solution is based on an hypothesis test, with the test statistic \( S(x) = |x| \) being compared to its distribution under \( M' \). Values of \( S \) exceeding the .99 level 2.58 are considered strong evidence in favor of \( M \), while values near the .95 level 1.96 provide moderate evidence. The break-even point, of roughly equal evidence for \( M \) versus \( M' \), is near the two-sided .90 level 1.645.

Recent Bayesian literature provides a quite different answer to the model selection problem. Suppose that \( x \) in (1.1) is obtained as the sufficient statistic from a sample of \( n \) independent observations. Then the Bayesian Information Criterion (BIC), described in Section 2, suggests \( |x| = \sqrt{\log(n)} \) as the break-even point between evidence for \( M' \) versus \( M \). This differs from the frequentist value 1.645 by being much bigger in large samples,

\[ \begin{array}{cccc}
    n: & 10 & 100 & 1000 & 10000 \\
    \sqrt{\log(n)}: & 1.52 & 2.15 & 2.63 & 3.04 \\
\end{array} \quad (1.3) \]

and by the fact that it depends on \( n \) at all. Recent references include Kass and Wasserman (1995), Kass and Raftery (1995), O'Hagan (1995), and Raftery (1996), plus others discussed in Sections 2 and 4.

This paper concerns the relationship between Bayesian and frequentist model selection. Section 2 restates the Bayesian approach, based on "Bayes factors", in a way that is convenient for comparison with frequentist ideas. This comparison is made in Section 3. In the context of Gaussian problems like (1.1)--(1.2), we show that a frequentist model selector behaves much like a Bayesian who has rather strong prior opinions about the distribution of \( \theta \) in model \( M \). Section 4 discusses the practical merits and demerits of the frequentist and Bayesian approaches. The selenium experiment, Clark et al. (1996), is used to illustrate how the two approaches can be combined in
a complicated inferential situation. A summary of the paper’s main ideas appears at the end of Section 4.

Except in the selenium discussion of Section 4, we will concentrate on situations like (1.2) where there are only two models $M'$ and $M$, with $M'$ of smaller dimension and nested within $M$. However most of the ideas extend to more realistic model selection situations, such as the choice of predictor variables in a regression problem, where there can be many competing models having a lattice of inclusion relationships.

2. Bayesian Model Selection

2.1. Bayes Factors

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

$$x \sim f_\theta(x),$$

(2.1)

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

$$\theta \in M' \quad \text{versus} \quad \theta \in M, \quad \text{with} \quad M' \subset M. \quad (2.2)$$

A complete Bayesian analysis begins with prior probabilities for $M'$ and $M$,

$$\pi' = \text{prob}\{\theta \in M'\} \quad \text{and} \quad \pi = \text{prob}\{\theta \in M\},$$

(2.3)

and also with prior conditional densities for $\theta$ given each model,

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

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

Letting $\pi(x)$ and $\pi'(x)$ be the aposteriori probabilities for $M$ and $M'$ having observed $x$, Bayes’ rule gives

$$\frac{\pi(x)}{\pi'(x)} = \frac{\pi f(x)}{\pi' f'(x)},$$

(2.5)

where $f(x)$ and $f'(x)$ are the two marginal densities

$$f(x) = \int_{M} f_\theta(x) g(\theta) d\theta \quad \text{and} \quad f'(x) = \int_{M'} f_\theta(x) g'(\theta) d\theta. \quad (2.6)$$

The ratio

$$B(x) = \frac{f(x)}{f'(x)}$$

(2.7)
is called the Bayes factor in favor of $M$ compared to $M'$. 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,

$$\frac{\pi(x)}{\pi'(x)} = \frac{\pi}{\pi'} \cdot B(x).$$

The Bayesian model selection literature tends to focus on the case $\pi/\pi' = 1$ of equal prior odds, with the presumption that $\pi/\pi'$ 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 4.

An easy example of Bayes factor calculations applies to the one-dimensional Gaussian situation (1.1)-(1.2). The $M'$ prior distribution $g'(\theta)$ must put all of its probability on $\theta = 0$. For model $M$ we take $g(\theta)$ corresponding to a normal distribution $N(0, A)$. Then it is easy to show that

$$\log B(x) = \frac{A}{A + 1} \cdot \frac{x^2}{2} - \frac{\log(A + 1)}{2}.$$  

The break-even point $x_0$ where $B(x_0) = 1$ occurs at $x_0 = [((A + 1)/A) \cdot \log(A + 1)]^{1/2}$, e.g. $x_0 = 2.16$ for $A = 100$.

2.2. Objective Bayes Methods and the BIC

In most situations there will be no obvious choice for the prior densities $g'(\theta)$ and $g(\theta)$ in (2.4). Jeffreys (1935) 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.

A central result is the Bayesian Information Criterion. Let $\hat{B}(x)$ be the maximum likelihood ratio statistic,

$$\hat{B}(x) = \frac{f_M(x)}{f_{M'}(x)},$$

with $\hat{\theta}$ and $\hat{\theta}'$ being the maximum likelihood estimates (MLEs) for $\theta$ under models $M$ and $M'$. In repeated sampling situation, where the data $x$ in (2.1) represent a random sample of size $n$, the BIC approximates the Bayes factor $B(x)$ by

$$\log B(x) \approx \log \hat{B}(x) - \frac{\log(n)}{2} (m - m'),$$

with $m$ and $m'$ representing the number of free parameters in $M$ and $M'$. (2.11) is also known as the Schwarz criterion on the basis of its original derivation by Schwarz (1978).

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 $\theta$ as one observation out of a random sample of size $n$.

As an example of this approach suppose that the data is an independent and identically distributed (i.i.d.) sample of size $n$ from a normal distribution with unknown expectation $\theta_1$,

$$x_1, x_2, \ldots, x_n \sim N(\theta_1, 1),$$  

so that the sufficient statistic

$$x = \frac{\sum_{i=1}^{n} x_i}{\sqrt{n}} \sim N(\theta, 1) \quad [\theta = \sqrt{n}\theta_1]$$  

(2.13a)

as in (1.1). We want to choose between

$$\mathcal{M}': \theta = 0 \quad \text{versus} \quad \mathcal{M} : -\infty < \theta < \infty$$  

(2.13b)

as before. If we use a normal prior $g(\theta) \sim N(0, A)$ as in (2.9), then the BIC approach suggests choosing $A = n$, giving

$$\log B(x) = \frac{n}{n+1} \cdot \frac{x^2}{2} - \frac{\log(n+1)}{2} = \log \hat{B}(x) - \frac{\log(n)}{2},$$  

(2.14)

since $\hat{B}(x) = \exp(x^2/2)$. In this case $m = 1, m' = 0$, so (2.14) agrees with (2.11).

There seems to be a convenient objective Bayesian prior density for this situation: $g(\theta) = c$, a constant. The marginal densities in (2.6) are now $f(x) = \int_{-\infty}^{\infty} c \exp\{-\frac{1}{2}(x - \theta)^2\}/\sqrt{2\pi} \ d\theta = c$ and $f'(x) = \exp\{-\frac{1}{2}x^2\}/\sqrt{2\pi}$, giving

$$\log B(x) = \frac{x^2}{2} + \log(\sqrt{2\pi}c).$$  

(2.15)

However, we still have to select a value for the constant $c$. The BIC amounts to a rule for choosing $c, c = 1/\sqrt{2\pi n}$ in this case, which puts the break-even point at $x_0 = \sqrt{\log n}$.

Suppose we believed entirely in model $\mathcal{M}$ in (2.13), and wished only to estimate $\theta$. 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 $\mathcal{M}'$. A prior distribution for situation (2.2) is inherently bumpy around the smaller model $\mathcal{M}'$, rather than smooth as in estimation problems, making the Bayesian analysis more delicate.

2.3. Penalty Functions and the Least Favorable Point

Define the penalty function $C(x)$ by

$$\log B(x) = \log \hat{B}(x) - \log C(x).$$  

(2.16)
The idea here is that $C(x)$ penalizes the observed likelihood ratio $\hat{B}(x)$ to compensate for the greater number of free parameters in the bigger model. The BIC (2.11) uses penalty function

$$\log[C_{BIC}(x)] = \frac{\log(n)}{2}(m - m').$$  \hspace{1cm} (2.17)

The Akaike Information Criterion (AIC) uses

$$\log[C_{AIC}(x)] = m - m',$$  \hspace{1cm} (2.18)

a less severe penalty if $n$ exceeds 7. See for instance Section 1.3 of O’Hagan (1995). Section 3 shows a connection between standard frequentist hypothesis testing and the penalty function

$$\log[C_{freq}(x)] = \chi^{2(\alpha)}_{m-m'}/2,$$  \hspace{1cm} (2.19)

where $\chi^{2(\alpha)}_{k}$ indicates the $100\alpha$th percentile of the chi-squared distribution with $k$ degrees of freedom.

We will now discuss a device that helps elucidate the penalty function $C(x)$. Remember that $x$ gives MLE's $\hat{\theta}$ and $\hat{\theta}'$ under $\mathcal{M}$ and $\mathcal{M}'$ respectively. A data set $y$ is said to be the least favorable point corresponding to $x$ if $y$ gives MLE $\hat{\theta}'$ under both $\mathcal{M}$ and $\mathcal{M}'$. See Table 1. In situation (2.13a,b), $y = 0$. Notice that “point” here refers to an entire data set of the same type as the original data $x$. “Least favorable” is an appropriate name for $y$ since availability of the bigger model $\mathcal{M}$ does not improve the estimation of $\theta$.

<table>
<thead>
<tr>
<th>Model:</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mathcal{M}$</td>
</tr>
<tr>
<td>$x$</td>
</tr>
</tbody>
</table>

Table 1. MLE's for the original data $x$ and the least favorable point $y$.

From $B(x) = f(x)/f'(x)$ and $B(y) = f(y)/f'(y)$, (2.7), we obtain

$$\log B(x) = \log \left( \frac{r}{r'} \right) + \log B(y),$$  \hspace{1cm} (2.20)

where

$$r = f(x)/f(y) \quad \text{and} \quad r' = f'(x)/f'(y).$$  \hspace{1cm} (2.21)

The Appendix shows that under generalized linear model (GLM) circumstances, using Jeffreys’ invariant priors for $g$ and $g'$ makes

$$\log \left( \frac{r}{r'} \right) = \hat{B}(x)$$  \hspace{1cm} (2.22)
with second-order accuracy, that is with error a factor of size \( 1 + O_p(1/n) \). Then (2.20) gives
\[
\log B(x) \doteq \log \hat{B}(x) - \log B^{-1}(y) \tag{2.23}
\]
\( B^{-1}(y) = f'(y)/f(y) \) so that the penalty function is seen to be the Bayes factor in favor of the smaller model \( \mathcal{M}' \), at the least favorable point \( y \). In terms of (2.16),
\[
\log C(x) \doteq \log B^{-1}(y). \tag{2.24}
\]
The BIC penalty function (2.17) is equivalent to assigning \( B^{-1}(y) \) the value
\[
B^{-1}(y) = n^{(m - m')/2}, \tag{2.25}
\]
which says that for large sample sizes \( n \) we can get a substantial Bayes function in favor of the smaller hypothesis \( \mathcal{M}' \). On the other hand, the frequentist penalty factor (2.19) in favor of \( \mathcal{M}' \) stays bounded at \( \exp\{\chi_m^2/(a_0)/2\} \), no matter how large \( n \) becomes. This represents an important distinction between Bayesian and frequentist model selection, discussed further in Sections 3 and 4.

### 2.4. The Break–Even Point

Result (2.23) shows that to a good approximation the penalty function \( C(x) \) equals \( \log B^{-1}(y) \), and so has nearly the same value for all data sets \( x \) leading to the same least favorable point \( y \). This means that the difference \( \log \hat{B}(x) - \log B(x) \) is nearly constant for all such data sets \( x \), the difference being \( \log B^{-1}(y) \). A schematic illustration appears in Figure 1. All of this is made specific in the Gaussian and GLM situations that follow, but the specifics are not crucial to the discussion here. We will call \( x_0 \) a break-even point if it has least favorable point \( y \), and satisfies
\[
\log B(x_0) = 0 \tag{2.26}
\]
Then (2.23) applied to \( x_0 \) gives
\[
\log B^{-1}(y) \doteq \log \hat{B}(x_0), \tag{2.27}
\]
or as in (2.24)
\[
\log C(x) \doteq \log \hat{B}(x_0). \tag{2.28}
\]
We can now write (2.23) in a form that is convenient for the frequentist/Bayesian comparisons of sections 3 and 4,
\[
B(x) \doteq \hat{B}(x)/\hat{B}(x_0). \tag{2.29}
\]
The frequentist break-even point \( x_0 \) is determined from a standard percentile point of the test statistic. We can use the right-hand side of (2.29) to approximate \( B(x) \) in terms of likelihood ratio statistics, which are usually easy to calculate or at least to approximate.
Figure 1. The difference between $\log \hat{B}(x)$ and $\log B(x)$ is nearly constant for all $x$ having the same least favorable point $y$; the break-even point $x_0$ is defined to have $\log B(x_0) = 0$.

3. Frequentists as Bayesians

3.1. Fisher's Scale of Evidence

A frequentist who follows standard hypothesis testing rules is using a very familiar model selection procedure. What does this procedure look like from a Bayesian point of view? Berger (1985, Section 4.3) and Berger and Sellke (1984) say that it is not at all Bayesian, no matter which prior distributions might be assumed. Here we will argue that the frequentist procedure does in fact have a Bayesian justification, but one that implicitly assumes a more informative prior distribution than that for the BIC. In a sense the frequentist acts more Bayesian than the Bayesians, or at least the objective Bayesians, for model selection.
Frequentists use a scale of evidence set down by Fisher in the 1920's. The scale depends on the critical level $\alpha$ of a statistic $S(x)$ that measures deviations from the smaller model $\mathcal{M}'$, the "null hypothesis", the critical level being the probability under $\mathcal{M}'$ that $S(X)$ is less than its observed value $s = S(x)$. One minus the critical level is the usual $p$-value or significance level.

Table 2 gives Fisher's scale as it is commonly interpreted: $\alpha = .99$ is strong evidence in favor of $\mathcal{M}$ versus $\mathcal{M}'$, .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 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 $\chi^2$ indicate a real discrepancy". Section 20, Fisher (1954).

<table>
<thead>
<tr>
<th>critical level $\alpha$</th>
<th>.90</th>
<th>.95</th>
<th>.975</th>
<th>.99</th>
<th>.995</th>
<th>.999</th>
</tr>
</thead>
<tbody>
<tr>
<td>strength of evidence</td>
<td>border</td>
<td>moderate</td>
<td>substantial</td>
<td>strong</td>
<td>very strong</td>
<td>overwhelming</td>
</tr>
</tbody>
</table>

Table 2. Fisher's scale of evidence against the null hypothesis $\mathcal{M}'$. The critical level (one minus the $p$-value) is the $\mathcal{M}'$ probability of the test statistic being smaller than the value actually observed.

We can use formula (2.29) to transform critical levels, or $p$-values, into Bayes factors. The critical level $\alpha_0 = .90$ corresponds at least approximately to a Bayes factor of 1. We can define the frequentist break-even point $x_0$ to be a point such that

$$S(x_0) = S^{(\alpha_0)} = S^{(.90)}$$

the 90th percentile of $S(X)$ under $\mathcal{M}'$. Then

$$B(x) \doteq \hat{B}(x)/\hat{B}(x_0)$$

according to (2.29).

Things are particularly simple if the test statistic $S(x)$ is the likelihood ratio statistic $\hat{B}(x)$ itself. Wilks' theorem says that then

$$S(x_0) \doteq \exp\{\chi^2_{m-m'}/2\}$$

as in (2.19), so (3.2) gives

$$\log B(x) \doteq \log \hat{B}(x) - \chi^2_{m-m'}/2, \quad [\alpha_0 = .90]$$

Except in Gaussian cases where (3.3) is exact, (3.4) is less accurate than (3.2), but this usually tends to be unimportant compared to the rough interpretive nature of Bayes factors.
3.2. The One - Dimensional Gaussian Case

We now examine the connection between Fisher's scale and Bayes factors more explicitly in the context of the one-dimensional Gaussian case (1.1)-(1.2), or (2.13a,b). The optimum test statistic for two-sided testing is \( S(x) = |x| \), which is equivalent to Wilks' likelihood ratio statistic \( \tilde{B}(x) = \exp\{x^2/2\} \). In this case (3.4) becomes

\[
\log \tilde{B}(x) \doteq (x^2 - x_0^2)/2
\]

where \( x_0 = 1.645 \), the .90 critical value of \( |x| \).

Table 3 shows the Bayes factors (3.5) corresponding to increasing levels of evidence on Fisher's scale. Also shown is \( \pi(x) \), the a posteriori probability on the “alternative” hypothesis \( \mathcal{M} \) if the prior odds ratio \( \pi/\pi' \) equals 1, (2.8). Notice that the .95 critical level gives \( \pi(x) = .64 \), a little less than 2 to 1 odds in favor of \( \mathcal{M} \), so “moderate evidence” is just that.

<table>
<thead>
<tr>
<th>Evidence</th>
<th>borderline</th>
<th>moderate</th>
<th>substantial</th>
<th>strong</th>
<th>very strong</th>
<th>overwhelming</th>
</tr>
</thead>
<tbody>
<tr>
<td>Critical level ( \alpha ):</td>
<td>.90</td>
<td>.95</td>
<td>.975</td>
<td>.99</td>
<td>.995</td>
<td>.999</td>
</tr>
<tr>
<td>(</td>
<td>x</td>
<td>):</td>
<td>1.645</td>
<td>1.96</td>
<td>2.24</td>
<td>2.58</td>
</tr>
<tr>
<td>Bayes Factor ( B(x) ):</td>
<td>1</td>
<td>1.76</td>
<td>3.19</td>
<td>7.13</td>
<td>13.29</td>
<td>58.03</td>
</tr>
<tr>
<td>( \pi(x) ):</td>
<td>.50</td>
<td>.64</td>
<td>.76</td>
<td>.88</td>
<td>.93</td>
<td>.98</td>
</tr>
</tbody>
</table>

**Table 3.** Critical levels and corresponding Bayes factors (3.5); one-dimensional Gaussian case, two-sided testing; break-even point \( x_0 \) at \( \alpha_0 = .90 \) quantile of the test statistic \( |x| \); \( \pi(x) \) is the a posteriori probability of \( \mathcal{M} \) assuming prior probability 1/2.

Table 3 is based on taking Figure 1 literally. In the one-dimensional Gaussian case \( \log \tilde{B}(x) = x^2/2 \). Then \( \log B(x) \) is set equal to \( \log \tilde{B}(x) - C \), where \( C = x_0^2/2 \) with \( x_0 = 1.645 \). The least favorable point \( y = 0 \) has \( \log B(y) = -x_0^2/2 \), or \( B^{-1}(y) = 3.87 \). In terms of (3.2), \( B(x) = \tilde{B}(x)/3.87 \). An objective Bayesian analysis, perhaps beginning with the improper prior \( g(\theta) = c \) as in (2.15), would differ only in its choice of the break-even point \( x_0 \). The BIC penalty function \( \log C = (\log n)/2 \) gives \( x_0 = \sqrt{\log n} \) as in (1.3), with corresponding Bayes factor \( B(x) = \tilde{B}(x)/\sqrt{n} \).

We can strengthen the Bayesian/frequentist connection by finding a genuine prior density \( g(\theta) \), as opposed to an improper one, that gives the function \( B(x) \) ascribed to the frequentist in Table 3. Figure 2 shows that making \( g(\theta) \) uniform on \([-4.85, 4.85]\), denoted \( U[-4.85, 4.85] \), gives an excellent match. The average absolute error

\[
Q = \int_{-3.29}^{3.29} |B_{4.85}(x)/B(x) - 1| dx / 2 \cdot 3.29
\]

equals .011. Section 3.4 motivates the choice of 4.85 and shows that still better matching priors are possible. The optimum choice of \( g(\theta) \) is symmetrically supported on six \( \theta \) values,
Figure 2. Solid curve is $\log B(x) = (x^2 - x_0^2)/2, \ x_0 = 1.645$; dots are $\log B_{4.85}(x)$, the Bayes factor for prior density $g(\theta)$ uniform on $[-4.85, 4.85]$; calculated from (2.6), (2.7) with $g'(\theta)$ the delta function on $\theta = 0$.

\[
\begin{align*}
\theta: & \quad \pm .71 \quad \pm 2.18 \quad \pm 3.86 \\
g: & \quad .147 \quad .159 \quad .194 ,
\end{align*}
\]

and gives $Q = .00061$, see section 3.4.

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 $\mathcal{M}'$. 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 3.4,

\[
\begin{align*}
n: & \quad 10 \quad 100 \quad 1000 \quad 10000 \\
(\pi \cdot n/2)^{1/2}: & \quad 3.96 \quad 12.53 \quad 39.63 \quad 125.33 .
\end{align*}
\]

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.
3.3. One-Sided Testing

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

\[ \mathcal{M} : 0 < \theta. \]  

(3.8)

The improper prior \( g(\theta) = c \) on \( \mathcal{M} \) leads to Bayes factor

\[ \log B(x) = \frac{x^2 - x_0^2}{2} + \log \frac{\Phi(x)}{\Phi(x_0)}, \]  

(3.9)

\( \Phi(x) \) being the standard normal cumulative density function (cdf). The break-even point \( x_0 \) is related to the constant \( c \) by \( c = \exp(-x_0^2/2)/\Phi(x_0) \).

Table 4 is the equivalent of Table 3 for the one-sided case. Now the break-even point is at \( x_0 = 1.282 \), the .90 critical level of the one-sided test statistic \( S(x) = x \). Notice that the critical levels of Fisher's scale, .90, .95, .975, ..., produce nearly the same Bayes factors in both tables. The aposterori probabilities \( \pi(x) \) (assuming \( \pi/\pi' = 1 \) are the same to two digits.

<table>
<thead>
<tr>
<th>Critical level ( \alpha ):</th>
<th>.90</th>
<th>.95</th>
<th>.975</th>
<th>.99</th>
<th>.995</th>
<th>.999</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x ):</td>
<td>1.282</td>
<td>1.645</td>
<td>1.96</td>
<td>2.33</td>
<td>2.58</td>
<td>3.09</td>
</tr>
<tr>
<td>Bayes Factor ( B(x) ):</td>
<td>1</td>
<td>1.80</td>
<td>3.25</td>
<td>7.24</td>
<td>13.42</td>
<td>57.86</td>
</tr>
<tr>
<td>( \pi(x) ):</td>
<td>.50</td>
<td>.64</td>
<td>.76</td>
<td>.88</td>
<td>.93</td>
<td>.98</td>
</tr>
</tbody>
</table>

Table 4. Critical levels and Bayes factors (3.9) for one-sided testing, one-dimensional Gaussian case; break-even point at \( x_0 = 1.282 \), the .90 quantile of \( x \) under the null model; \( \pi(x) \) aposteriori probability of \( \mathcal{M} \) assuming prior probability \( 1/2 \).

The Bayes factors in Table 4 are closely approximated by taking the prior density \( g(\theta) \) to be uniform on \([0, 5.13]\). A Bayesian who began with the uniform prior \( U[-4.85, 4.85] \) 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 4. (Using 4.85 instead of 5.13 as the one-sided upper limit for \( g(\theta) \) gives a close match to (3.9) with \( x_0 = 1.243 = \Phi^{-1}(.893) \) instead of \( x_0 = \Phi^{-1}(.90) \).

Suppose we observe \( x = 1.96 \) and wish to estimate \( \gamma = \text{prob}\{\theta > 0|x\} \). The prior distribution appropriate to Table 4, having \( \pi/\pi' = 1 \) and \( g(\theta) \) uniform on \([0, 5.13]\), gives the aposteriori Bayes estimate \( \gamma_1 = .76 \). Table 3's prior, with \( \pi/\pi' = 1 \) for \( \mathcal{M} = (-\infty, \infty) \), and \( g(\theta) \) uniform on \([-4.85, 4.85]\), gives \( \gamma_2 = .64 \). This decrease is consistent with Bayesian theory since Table 3's prior puts only half as much probability on the positive axis.

Both \( \gamma_1 \) and \( \gamma_2 \) are much smaller than the estimate \( \gamma_3 = .975 \) 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 \( \mathcal{M} : \theta = 0 \), and estimation which does not. The estimation paradigm is often more appropriate

11
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 argument in the discussion following O'Hagan (1995).

Tables 3 and 4 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 axis half of the \( U[-4.85, 4.85] \) prior. We can also use the \( U[-4.85, 4.85] \) prior to investigate frequentist behavior in multiple testing situations, see Remark II in Section 5.

3.4. 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 \( \theta \) and covariance matrix the identity

\[
x \sim N_m(\theta, I),
\]

and that we wish to test

\[
\mathcal{M}' : \theta = 0 \quad \text{versus} \quad \mathcal{M} : \theta \in \mathbb{R}^m.
\]

We will denote \( z = ||x||, \theta = ||\theta|| \), and write \( B(z) \) instead of \( B(x) \), etc. The likelihood ratio statistic \( \tilde{B}(z) \) equals \( \exp(z^2/2) \) so that in terms of the likelihood ratio test (2.29) gives

\[
\log B(z) = (z^2 - x_0^2)/2 \quad \text{where} \quad x_0^2 = \chi_m^2(\alpha_0).
\]

We used (3.12) in the one-dimensional case (3.5), with \( \alpha_0 = .90 \), and showed that it agreed closely 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 \( B(z) \) in (3.12) is not close to being a Bayes factor for any genuine prior \( g(\theta) \).

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

\[
g(\theta) = \frac{\Gamma(m/2 + 1)}{\pi^{m/2} u^m} \quad \text{for} \quad 0 \leq \theta \leq u.
\]

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

\[
B_u(z) = \frac{c_m}{u^m} e^{z^2/2} F_m(u, x) \quad \text{where} \quad c_m = 2^{m/2} \Gamma(m/2 + 1),
\]

\[
F_m(u, x) = \text{prob}\{\chi_m^2(x^2) < u^2\},
\]

12
the probability that a non-central $\chi^2$ variate with $m$ degrees of freedom and non-centrality $x^2$ is less than $u^2$.

We would like (3.14) to match $B(x) = \exp(x^2 - x_0^2)/2$. Notice that

$$\frac{\partial}{\partial x} \log B_u(x) = \frac{\partial}{\partial x} \log B(x) + \frac{\partial}{\partial x} \log F_m(u, x).$$  \hspace{1cm} (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 $x_0 = [\chi_m^{2(\alpha_0)}]^{1/2}$ the break-even point satisfies $B_{u_0}(x_0) = 1$, or according to (3.14),

$$u_0 = [c_m e^{x_0^2/2} F_m(u_0, x_0)]^{1/m}. \hspace{1cm} (3.17)$$

For $m = 1$ and $x_0 = 1.645$ we get $u_0 = 4.85$, the value used in Section 3.2. The derivative $\frac{\partial}{\partial x} \log F_1(u_0, x)|_{x_0}$ in (3.16) is only $-0.0025$ in this case, accounting for the good match between $B(x)$ and $B_{4.85}(x)$ in Figure 2.

Suppose though that $m = 6$. Then $x_0 = [\chi_6^{2(\alpha_0)}]^{1/2} = 3.263$ and formula (3.17) gives $u_0 = 4.30$. In this case $\frac{\partial}{\partial x} \log F_6(u_0, x)|_{x_0} = -0.526$, compared to $\frac{\partial}{\partial x} \log B(x)|_{x_0} = 3.263$, so $B_u(x)$ is a poor match to $B(x)$. We might hope that by choosing $g(\theta)$ differently than the uniform prior (3.13) we could better match $\log B(x) = (x^2 - 3.263^2)/2$, but the next analysis shows not.

If $B_g(x)$ is the Bayes factor (2.6)-(2.7) corresponding to a spherically symmetric prior $g(\theta)$, and $B(x, \alpha_0) = \exp[(x^2 - x_0^2)/2]$ for $x_0^2 = \chi_m^{2(\alpha_0)}$ define

$$Q(g, \alpha_0) = \int_0^{x_{.999}} |B_g(x)/B(x; \alpha_0) - 1| \, dx/x_{.999},$$  \hspace{1cm} (3.18)

so $Q$ measures the average absolute deviation between $B_g(x)$ and $B(x; \alpha_0)$ over the range of Fisher's scale in Figure 2.

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 .15. In order to be able to match $B(x, \alpha_0)$ with accuracy $Q = .011$, the accuracy shown in Figure 2, we need to increase $\alpha_0$ to .96. This raises the break-even point to $x_0 = (\chi_6^{2(\alpha_0)})^{1/2} = 3.63$, rather than $(\chi_6^{2(\alpha_0)})^{1/2} = 3.26$, and decreases the Bayes factor $\exp((x^2 - x_0^2)/2)$ by a factor of 3.58.

We used Fisher's scale of evidence to set the frequentist break-even point at the .90 quantile of the test statistic. This definition turns out to be a good one for one-dimensional Gaussian testing problems: figure 3 shows that reducing $\alpha_0$ to say .80 makes it impossible to get an accurate match with any genuine Bayes factor, the minimum attainable $Q$ equalling .18. In higher dimensions $\alpha_0 = .90$ itself is too small. A frequentist who desires a proper Bayesian interpretation of p-values, as in (3.1)-(3.2), needs to set the breakdown quantile $\alpha_0$ higher.
Figure 3. Minimum possible value of average absolute deviation (3.18) as a function of $\alpha_0$ and dimension $m$; $m = .5$ is the one-sided one-dimensional problem.

There is another argument for saying that higher dimensions require greater values of $\alpha_0$. Suppose we believe that the penalty function $C(x)$ in (2.16) should only be a function of $d = m - m'$, the dimensional difference between $\mathcal{M}$ and $\mathcal{M}'$. Then by considering a nested sequence of models it is easy to show that

$$\log C(x) = \lambda d.$$  \hspace{1cm} (3.19)

The BIC (2.17) uses $\lambda = (\log n)/2$ while the AIC (2.15) uses $\lambda = 1$. However the frequentist penalty function $\log C_{\text{freq}}(x) = \chi^2_d(\alpha_0)/2$ does not satisfy (3.19).

We can enforce (3.19) by choosing $\lambda = \chi^2_1(\alpha_0)/2$ and then defining $\alpha_0(d)$, the break-even level for dimensional difference $d$, according to

$$\chi^2_d(\alpha_0(d))/2 = (\chi^2_1(\alpha_0)/2) \cdot d.$$  \hspace{1cm} (3.20)

Doing this makes $\alpha_0(d)$ increase with $d$, as shown in Table 5. The choice $\alpha_0 = .86$ is the minimum value that keeps $Q$ from Figure 3 reasonably small (about .05) for all dimensions.
\[
\begin{array}{ccccccc}
\text{d:} & 1 & 2 & 3 & 4 & 5 & 6 \\
1. \alpha_0 = .86: & .86 & .887 & .912 & .931 & .946 & .958 \\
2. \alpha_0 = .90: & .90 & .933 & .956 & .971 & .981 & .987 \\
\end{array}
\]

Table 5. Break-even level \( \alpha_0(d) \) for \( \alpha_0 = .86 \) or .90 in (3.20). Bottom line is \( B(x) = (x^2 - x_0^2)/2 \) for \( x = [\chi^2_d(0.99)]^{1/2} \) and \( x_0 = [\chi^2_d(\alpha_0(d))]^{1/2} \), \( \alpha_0(d) \) as in line 1.

All of this suggests making Fisher’s scale of evidence more stringent in higher dimensions. Doing so affects the Bayes factor calculation in interesting ways. The bottom line of Table 5 shows \( B(x) \) if we observe \( x \) at the 99th percentile, \( x = [\chi^2_d(0.99)]^{1/2} \), and use break-even level \( \alpha_0(d) \) from (3.26) with \( \alpha_0 = .86 \). \( B(x) \) first increases and then decreases as a function of dimension \( d = m - m' \), so the increases in \( \alpha_0(d) \) do not necessarily imply decreasing Bayes factors.

4. Discussion and Summary

We have discussed the theoretical similarities between frequentist and objective Bayes model selection procedures, but in practice the Bayesian approach tends to produce greater probabilities of selecting the smaller model \( \mathcal{M}' \). This is especially true when the sample size \( n \) is large. Most of the discrepancy has to do with the break-even point, which grows larger with \( n \) in the Bayesian framework, but it also arises from a more stringent scale of evidence suggested by Jeffreys (1961).

Table 6 shows Jeffreys’ scale of evidence, as amended by Kass and Raftery (1995), who also give Jeffreys’ original version. A comparison with table 3 shows how much more conservative Jeffreys is than Fisher in interpreting evidence against the smaller hypothesis \( \mathcal{M}' \). Fisher’s “moderate” .95 level is “not with more than a bare mention”, and even a .995 critical level does not reach Jeffreys’ threshold of strong evidence.

This comparison assumes Fisher’s choice of the .90 critical level for the break-even point \( x_0 \). If Jeffreys’ scale is combined with the BIC penalty function (2.17) (which can result in very large values of \( x_0 \) as in (1.3)), the model selection procedure becomes much more favorable to \( \mathcal{M}' \). Figure 4 shows this for the one-dimensional Gaussian case of Table 3. Observing \( x = 2.58 \), at the two-sided .99 level, is barely worth mentioning if \( n = 100 \), a shocking assertion to the medical scientist for whom a p-value of .01 settles the issue.

Part of the difference here is cultural. 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 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. There is nothing intrinsic to either Table 2 or Table 7 that makes one scale frequentist and the other Bayesian.
Bayes Factor:  

<table>
<thead>
<tr>
<th></th>
<th>&lt; 1</th>
<th>1 ——— 3</th>
<th>3 ——— 20</th>
<th>20 ——— 150</th>
<th>&gt; 150</th>
</tr>
</thead>
<tbody>
<tr>
<td>Evidence Against $\mathcal{M}'$:</td>
<td>negative than a bare mention</td>
<td>positive</td>
<td>strong</td>
<td>very strong</td>
<td></td>
</tr>
<tr>
<td></td>
<td>↑</td>
<td>↑ ↑ ↑</td>
<td>↑</td>
<td>↑</td>
<td></td>
</tr>
<tr>
<td>Table 3:</td>
<td>(.95)</td>
<td>(.975)(.99)(.995)</td>
<td>(.999)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 6. Jeffreys' scale of evidence for the interpretation of Bayes factors, as amended by Kass and Raftery (1995). Comparison with Table 3 (parentheses and arrows) shows that Jeffreys' scale is much more stringent than Fisher's.

Another difference concerns the status of $\mathcal{M}'$, the smaller hypothesis. $\mathcal{M}'$ is usually a straw man in Fisherian hypothesis testing, not a particularly interesting scientific theory in its own right, see Section 4 of Efron (1971). Formula (3.4) and Figure 1 show that the maximum possible Bayes factor in favor of $\mathcal{M}'$ using a standard likelihood ratio test is

$$B^{-1}(y) \asymp \exp\left\{ \chi_{m-m'}^2(\alpha_0)/2 \right\},$$  \hspace{1cm} (4.1)

equalling 3.87 for $m - m' = 1$ and $\alpha_0 = .90$. The Bayesian literature beginning with Jeffreys has treated $\mathcal{M}'$ and $\mathcal{M}$ more even-handedly. (For an interesting example see Good (1969) and also Efron (1971).) The BIC penalty function allows the data to provide strong evidence for $\mathcal{M}'$, as in (2.25). The difference here between Fisher and Jeffreys is again one of scientific context rather than of a fundamental frequentist-Bayesian disagreement.
Figure 4. Jeffreys' scale of evidence combined with the BIC choice $x_0 = [\log n]^{1/2}$ of break-even point, for the one-dimensional Gaussian problems of Table 3. Fisher's scale, shown at right, does not depend on $n$. Jeffreys' scale is much more conservative, especially for large $n$.

There is, however, a fundamental difference in philosophy that makes Bayesian model selection techniques more stringent than frequentist methods in large samples. This concerns the choice of the penalty function $C(x)$, or equivalently $B^{-1}(y)$, (2.24), which increases with $n$ in the Bayesian framework, as in (2.17). Why should it do so?

Consider the one-dimensional Gaussian situation (2.12–2.13a,b) where we observe $x_1, x_2, \cdots, x_n \sim i.i.d. N(\theta, 1)$ and wish to choose between $M' : \theta = 0$ and $M : \theta \in (-\infty, \infty)$, the same statistical problem as (2.13b). Suppose that $g_1(\theta)$ is the prior density for $\theta$ given model $M$. At the least favorable point $y = 0$ the Bayes factor in favor of $M'$ is

$$B^{-1}(0) = \sqrt{n} \varphi(0) / \int \varphi(\theta) g_1(\theta / \sqrt{n}) d\theta$$  \hspace{1cm} (4.2)

where $\varphi(\theta) = \exp(-\theta^2/2)/\sqrt{2\pi}$. This is obtained by reducing the data to the sufficient statistic $x = \sum_{i=1}^n x_i / \sqrt{n}$, (2.13a). For large $n$, (4.2) goes to

$$B^{-1}(0) \approx cn^{1/2} \quad \text{with} \quad c = \varphi(0) / g_1(0)$$  \hspace{1cm} (4.3)
assuming that $g_1(\theta)$ is positive and continuous near 0. The BIC value $B^{-1}(0) = n^{1/2}$ amounts to choosing $g_1(0) = 1/\sqrt{2\pi} = \varphi(0)$. Jeffreys' preferred a different choice, see Remark F.

Making $B^{-1}(0)$ grow at rate $cn^{1/2}$, or rate $n^{(m-m')/2}$ in higher dimensions, (2.25), might be called “sample size coherency”. The argument for it given above applies quite generally, the key assumption being that the model selection procedure at sample size $n$ should behave as if we began with a proper prior distribution $g_1(\theta_1)$ at sample size 1. There is nothing special about the BIC choice $c = 1$, see Remark F, but it seems to have gained ascendency in the Bayesian literature. Sequential coherency is considered by O'Hagan (1995) and some of his discussants.

Sample size coherency is an appealing property in situations where the statistician actually sees the data set growing in size. It is less compelling in the more usual situation where $n$ is fixed, and taking it literally can lead to Figure 4's situation, where even evidence at the .995 level is barely worth mentioning. Next we consider a medical example that shows how $p$-values can be transformed into Bayes factors using the methods of this paper. However the example also illustrates how differences can arise between the two points of view.

Table 7 shows total cancer mortality for a double-blind randomized trial of the trace element selenium taken as a cancer preventative. 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 mg. per day of selenium or an identical-looking placebo. The results from 1983 to 1989, labelled “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 $s \sim \text{binomial}(N, 1/2)$, splitting the probability atom at 7,

$$0.032 = \frac{1}{2N} \left\{ \sum_{i=0}^{s-1} \binom{N}{i} + \frac{1}{2} \binom{N}{s} \right\},$$

(4.4)

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 investigator’s attention was now focused on total cancer mortality. Lung, prostate, and colorectal cancer incidence were also listed as secondary endpoints, see Remark A.

We now consider selecting between the models

$\mathcal{M}':$ selenium has no effect on total cancer mortality

$\mathcal{M}:$ selenium has an effect on total cancer mortality.  

(4.5)
<table>
<thead>
<tr>
<th></th>
<th>1st Period</th>
<th></th>
<th>2nd Period</th>
<th></th>
<th>Combined</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Selenium</td>
<td>Placebo</td>
<td>Total</td>
<td>Selenium</td>
<td>Placebo</td>
<td>Total</td>
</tr>
<tr>
<td>Deaths:</td>
<td>7</td>
<td>16</td>
<td>23</td>
<td>22</td>
<td>41</td>
<td>63</td>
</tr>
<tr>
<td>p-value:</td>
<td>.032</td>
<td></td>
<td></td>
<td>.00843</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bayes factor:</td>
<td>1.44</td>
<td></td>
<td></td>
<td>4.49*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Adjusted</td>
<td></td>
<td></td>
<td></td>
<td>(7.64)*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bayes factor)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 7. Total cancer mortality in the selenium experiment, from Table 5 of Clark et al. (1996); p-values are one-sided, based on binomial distribution (4.4); Bayes factors $\exp(x^2 - z_0^2)/2$ with $x = \Phi^{-1}(1 - p$-value) and $z_0 = 1.645$; * One-sided Bayes factor, $z_0 = 1.282$; ** coherency adjustment (4.12).

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, as

$$s \sim \text{binomial}(N, \theta) \quad \text{with} \quad \begin{cases} M' : \theta = .5 \\ M : \theta \in [0, 1] \end{cases}$$

as the competing models. The dimensional difference is $m - m' = 1$.

How strong is the evidence against $M'$ and in favor of $M$? 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 (4.4) into an approximate normal deviate $x = \Phi^{-1}(1 - .00843) = 2.39$, and then calculates the Bayes factor for $M$ from the one-dimensional normal approximation (3.5)

$$B(s) \doteq \exp\{(2.39^2 - 1.645^2)/2\} = 4.49.$$  \hspace{1cm} (4.7)

Starting from the conventional prior odds ratio $\pi/\pi' = 1$ in (2.5) results in aposteriori probability $\pi(s) = .82$ for $M$. See Remark D.
We might feel entitled to use the one-sided break-even point 1.282 for the 2nd period analysis instead of 1.645, on the grounds that the 1st period results removed most of our apriori probability that \( \theta > .5 \). Then we get Bayes factor

\[
B(s) = \exp\{(2.39^2 - 1.282^2)/2\} = 7.65
\]

in favor of \( M : \theta < .5 \) versus \( M' : \theta = .5 \). Prior odds ratio \( \pi/\pi' = 1 \) now gives \( \pi(s) = .88 \) for \( M \).

We do not need to rely on the normal approximation (3.5). Let \( s_0 \) be the .05 critical value for a binomial\((N,.5)\) distribution, calculated by interpolation of the “split-atom” cdf (4.4), so \( s_0 \) is a break-even point at the two-sided .90 level for a binomial\((N,.5)\) test statistic. Then according to (3.2)

\[
B(s) = \frac{\hat{B}(s)/\hat{B}(s_0)}{s_0^{s/N}(N-s_0)^{N-s}}
\]

This gives 4.63 instead of 4.49 in (4.7), and 7.98 instead of 7.65 in (4.8). Normal approximations like (4.7) and (4.8) worked equally well for all of the Bayes factors in Table 7.

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

\[
\pi/\pi' = 1/4 .
\]

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 (4.7) is now

\[
B(s) = 25.20 .
\]

Bayes’ rule with \( \pi/\pi' = 1/4 \) gives posteriori probability \( \pi(s) = .86 \) for \( M \). Formula (4.9) gives \( B(s) = 26.40 \) and \( \pi(s) = .87 \).

However a reduction of \( B(s) = 25.20 \) to about

\[
B_{coher}(s) = 25.20/\sqrt{2} = 17.82
\]

is necessary if we wish to maintain sample size coherency. The argument for (4.12) is along the lines of (4.2)–(4.3), see Remark E. It relates to the question of what is meant by “sample size” in the BIC. Prior odds ratio \( \pi/\pi' = 1/4 \) now makes \( \pi(s) = .82 \).
All of our frequentist-cum-Bayesian analyses yield posterior probabilities for selenium being effective in the range .82 to .87. BIC analysis is predictively more conservative. The likelihood ratio statistic \( \tilde{B}(s) \) equals 104.0 for the combined data in Table 7, so the BIC Bayes factor (2.11) is

\[
B_{\text{BIC}}(s) = \frac{104.0}{\sqrt{n}} = 11.21
\]  \hspace{1cm} (4.13)

if \( n = 86 \), the combined total deaths, making \( \pi(s) \) only .74. If instead we take \( n = 1312 \), the number of subjects in the selenium trial, then \( B_{\text{BIC}}(s) = 2.87 \) and (4.10) gives \( \pi(s) = .42 \), actually favoring \( M' \).

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 (3.2) 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.

**Consistency** is one of the main virtues claimed for the BIC. In the situation where we choose \( x_1, x_2, \ldots, x_n \overset{\text{i.i.d.}}{\sim} N(\theta_1, 1) \) and wish to choose between \( M': \theta_1 = 0 \) versus \( M: \theta_1 \neq 0 \), the BIC estimate of the Bayes factor (2.14) is consistent: as \( n \to \infty \) then \( \log B(x) \to -\infty \) if \( \theta_1 = 0 \), while \( \log B(x) \to -\infty \) if \( \theta_1 \) has any non-zero value. The BIC penalty function \( C_n = \sqrt{n} \) makes this happen. However the same thing is true for any choice of \( C_n \) that goes to infinity with \( n \) at a rate slower than \( n \).

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 must beware of tacitly settling for very small power, as suggested by Figure 4. This is fundamentally a Bayesian point that seems to have been of greatest concern to frequentists, see Remark C.

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 techniques share a crucial property with the BIC: in a situation like (2.12)–(2.13) the amount of information in the prior density for \( \theta \) is small, on the order of \( 1/n \) of the information in the data, see Remark F.

Frequentist hypothesis testing can be described in Bayesian terms as in Section 3, but using priors for \( \theta \) that are considerably more informative when \( n \) is large. Comparing the \( U[-4.85, 4.85] \) prior of Figure 2 with (3.7) suggests that the frequentist prior has about \( 1/n_{\text{freq}} \) of the sample information for \( \theta \), where

\[
n_{\text{freq}} = \frac{2}{\sqrt[\pi]{4.85^2}} \approx 15.
\]  \hspace{1cm} (4.14)
The difference between $1/n$ and $1/n_{freq}$ has enormous practical consequences in large samples, dwarfing any other Bayesian/Frequentist philosophical differences.

**Summary**

The main point of this paper is the one just made: that frequentist model selection can be understood from a Bayesian point of view, but one where the prior distribution favors the bigger model, more so than with BIC or other objective Bayesian methods. The reasons for this are partly historical having to do with the difference between Fisher’s and Jeffreys’ scales of evidence, Tables 2 and 6.

More of the difference however arises from the Bayesian notion of sample size coherency, as discussed following (4.2), which implies that evidence against the smaller model must be judged more cautiously in larger samples. Sample size coherency is of dubious value in a fixed sample size experiment, but is more compelling when the statistician actually sees different sample sizes. The practical differences between frequentist and Bayesian methods can be enormous, as shown in Figure 4.

The technical arguments supporting this point of view involve the penalty function $C(x)$, (2.16), that computes the Bayes factor $B(x)$ from the likelihood ratio statistic $\hat{B}$ via the relationship

$$\log B(x) = \log \hat{B}(x) - \log C(x):$$

- Under GLM assumptions, the penalty function $C(x)$ is closely approximated by $B^{-1}(y)$, the Bayes factor in favor of the smaller hypothesis $\mathcal{M}'$ at the least favorable point for the bigger hypothesis $\mathcal{M}$, see (2.24) and Table 1.

- The BIC approximates $\log C(x)$ by $[\frac{1}{2} \log(n)] \cdot [m - m']$, where $n$ is the sample size and $m - m'$ is the dimensional difference between $\mathcal{M}$ and $\mathcal{M}'$, (2.17). Frequentist hypothesis testing using the likelihood ratio test approximates $\log C(x)$ by $\chi^2_{m-m'}/2$, (2.19), with $\alpha_0 = .90$ on Fisher's scale.

- This leads to a convenient formula for translating frequentist p-values into Bayes factors, $B(x) \cong \hat{B}(x)/\exp\{\chi^2_{m-m'}/2\}, (3.4)$. This was illustrated for the selenium data in Table 7.

- It is also helpful to state results in terms of the break-even point $x_0$, (satisfying $B(x_0) = 1$), Section (2.4): $B(x) \cong \hat{B}(x)/\hat{B}(x_0), (2.29)$, giving BIC approximation $B(x) \cong \hat{B}(x)/n^{(m-m')/2}$ from (2.24)–(2.25), and frequentist approximation (3.4). Figure 1 illustrates the relationship between $B(x), \hat{B}(x), C(x), x_0$, and $B^{-1}(y)$. In the one-dimensional Gaussian case (2.12)–(2.13), the BIC break-even point is $x_0 = \sqrt{\log(n)}$, (1.3), compared to $x_0 = 1.645$ for the standard frequentist approach.

- In the one-dimensional Gaussian case, a uniform prior over the interval $[-4.85, 4.85]$ gives a Bayes factor closely approximating the objective Bayes formula $\log B(x) = (x^2 - x_0^2)/2$, but having
the frequentist break-even point \( x_0 = 1.645 \). Figure 2. The corresponding "matching prior" for the BIC is uniform over \( \pm (\pi \cdot n/2)^{1/2} \), (3.7). Section 3.3 shows similar results for the one-sided case where the bigger model is \( \mathcal{M} : \theta > 0 \).

- The nice correspondence between frequentist and Bayesian model selection weakens in higher dimensional problems., section (3.4). To restore it we need to increase the break-even point on Fisher's scale above the \( \alpha_0 = .90 \) quantile of the test statistic. Figure 3 suggests an \( \alpha_0 \) of about .96 for a six-dimensional Gaussian testing problem. It also shows that even in the one-dimensional case we could not choose \( \alpha_0 \) much smaller than .90.

- Prior distributions for the model selection problem are inherently bumpy, having a bump of prior probability at \( \mathcal{M}' \), in contrast to the smooth priors used in the objective Bayes analysis of an estimation problem. This leads to much different estimates of quantities like \( \text{prob}\{\theta > 0|z\} \), as discussed in Section 3.3.

5. Remarks and Proofs.

**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 (4.41). 2nd period incidences were 21 versus 56, significance level \( 2.79 \cdot 10^{-5} \). Now the one-sided 2nd period Bayes factor corresponding to (4.8) is \( B(s) = 1482 \), while the equivalent of (4.12) gives \( B_{\text{coher}}(s) \doteq 1664 \). 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.** Is Fisher's scale of evidence too liberal? 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 \( \pi/\pi' \) in (2.8). 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.

Raftery (1995) argues for the superiority of Jeffreys' scale and BIC model selection in the social sciences. His main example concerns a multi-national 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 about \( 10^{-120} \). In this case the BIC penalty function (2.17) is severe enough to give a Bayes factor in favor of the sociometric model, compared to a saturated model.

The problem here may be one of putting too much strain on the model selection paradigm. Frequentist hypothesis testing followed by a confidence interval would show that the sociometric
model does not fit the data perfectly but that the deviations from the model are quite small (Remark D concerns a related example). Gelman and Rubin argue along these lines in the discussion following Raftery (1995). See also Diaconis and Efron (1985).

Remark C. Pre-experimental power calculations are an important part of frequentist model selection. A common prescription for the one-dimensional Gaussian situation (2.12)–(2.13) would be to require a sample size of \( n = (t_1/3.24)^2 \), where \( t_1 \) is a preliminary guess for the treatment effect \( \theta_t \). The value 3.24 = 1.96 + 1.28 results in .90 power for a two-sided .95 test, at \( \theta_t = t_1 \). This kind of calculation which has a strong Bayesian flavor, fits in well with the \( U[-4.85, 4.85] \) prior ascribed to the frequentist in Section 3.2. Notice that selecting \( n \) on the basis of the prior undercuts the BIC selection of the prior on the basis of \( n \).

Remark D. In (4.6), the two-sided formulation of the selenium problem, we calculated a Bayes factor of \( B(s) = 4.49 \) in favor of \( \mathcal{M} : \theta \leq 1 \) versus \( \mathcal{M}' : \theta = .5 \). However we were really interested in the one-sided alternative \( \theta < .5 \). To this end we can state the results following (4.6) as follows: the aposteriori probability of \( \mathcal{M} \) given the data is about \( \pi(s) = .82 \), and given that \( \mathcal{M} \) is true, the aposteriori probability that \( \theta < .5 \) is roughly \( \Phi(2.39) = .992 \). Here we are combining hypothesis testing with confidence intervals, or in Bayesian terms combining lumpy model selection priors with the smooth priors appropriate to estimation.

Remark E. Sequential observation of the data, even a small amount of it as in the selenium trial, spoils the agreement between frequentist and Bayesian model selection. Consider situation (2.12)–(2.13) with \( n = 2 \), where the statistician observes \( x_1, x_2 \overset{i.i.d.}{\sim} N(\theta_t, 1) \). Given the sufficient statistic \( x = (x_1 + x_2)/\sqrt{2} \), the frequentist might claim to be using the \( U[-4.85, 4.85] \) prior of Section 3.2 for \( \theta = \sqrt{2} \theta_t \), yielding Bayes factor \( B(x) = \exp\{(x^2 - x_0^2)/2\} \) in favor of \( \mathcal{M} \), with \( x_0 = 1.645 \).

A Bayesian, even one who accepts Fisher's scale of evidence, could object on the following grounds: if the frequentist only saw \( x_1 \) he or she would have used a \( U[-4.85, 4.85] \) prior for \( \theta_t \). Therefore the frequentist's prior for \( \theta = \sqrt{2} \theta_t \) should be uniform over the range \( \pm \sqrt{2} \cdot 4.85 \). But this prior gives a smaller Bayes factor, in fact \( B(x) \overset{\approx}{=} (1/\sqrt{2}) \exp\{(x^2 - x_0^2)/2\} \) for moderate values of \( x \). This is the coherency correction (4.12), and is easy to verify numerically or using the Laplace approximation (5.10).

The frequentist might also criticize the Bayesian's reasoning. When stated in terms of \( x \) and \( \theta \) as in (2.13a,b), the model selection problem does not depend on "\( n \)" in any way. Why should the fact that \( n = 2 \) pieces of information went into \( x \) affect the model selection procedure? The Bayesian argument is unassailable if we begin with a genuine prior for \( \theta_t \), but less so if the prior only expresses our ignorance of the situation.

Remark F. In situation (2.12)–(2.13), and in more general problems too, the BIC chooses a prior density having information \( 1/n \) for \( \theta \). Jeffreys' original proposal used a slightly more diffuse prior,
with information about $1/(\frac{1}{2}n)$, see Section 3 of Kass and Wasserman (1993). 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 G.** Two lines of argument led to the objective Bayes factor $\log B(x) = (x^2 - x_0^2)/2$, (3.5), for the one-dimensional normal problem: argument (2.20)-(2.24), based on penalty functions and the least-favorable point, and also calculation (2.15) beginning with the improper prior $g(\theta) = c$. However in the one-sided testing situation of Section (3.3) the two arguments no longer agree. In that case we used the improper prior calculation to get (3.9) for $\log B(x)$, and in fact a Bayes factor $\log B(x) = (x^2 - x_0^2)/2$ does not agree well with proper Bayes factors for values of $x$ near 0. The difficulty is caused by $\mathcal{M}' : \theta = 0$ lying at the boundary of $\mathcal{M} : \theta > 0$, instead of inside $\mathcal{M}$ as in the two-sided case.

**Remark H.** The $\mathcal{U}(-4.85,4.85)$ prior density figuring in the discussion of Section 3.2 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,

$$x_i \stackrel{i.i.d.}{\sim} N(\theta_i, 1), \quad i = 1, 2, \cdots, J,$$

and that we wish to test

$$\mathcal{M}' : \text{all } \theta_i = 0 \quad \text{versus} \quad \mathcal{M} : \text{one of the } \theta_i \text{ not zero}.$$  

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

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

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

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

$$\mathbf{x} = (x_1, 0, 0, \cdots, 0),$$

in which case

$$B(x) \doteq \sqrt{2\pi} e^{x^2/2} + J - 1 \quad \text{for} \quad 2J \cdot 4.85.$$

Using this approximation, the break-even point $x_0 = (x_0, 0, 0, \cdots, 0)$ occurs at

$$x_0^- = \pm \left(2 \log \left[ \frac{2J \cdot 4.85}{\sqrt{2\pi}} - (J - 1) \right] \right)^{1/2}.$$

Table 8 compares the Bayesian break-even point $x_0$ from (5.5) with the .90 critical point $x_0 = \Phi^{-1}(1 - .05/J)$ 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 (5.2)–(5.4).

<table>
<thead>
<tr>
<th>J:</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>6</th>
<th>8</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bonferroni:</td>
<td>1.96</td>
<td>2.13</td>
<td>2.24</td>
<td>2.39</td>
<td>2.50</td>
<td>2.58</td>
</tr>
<tr>
<td>(5.5):</td>
<td>1.95</td>
<td>2.13</td>
<td>2.25</td>
<td>2.41</td>
<td>2.52</td>
<td>2.60</td>
</tr>
</tbody>
</table>

Table 8. Simultaneous testing; comparison of critical point for the simultaneous .90 Bonferroni test with the break-even point (5.5) for the Bayesian analysis suggested by the \( \mathcal{U}[-4.85, 4.85] \) distribution; \( J \) is the number of simultaneous tests.

We conclude with the proof of (2.22), which led to the useful approximation \( \log C(x) \doteq \log B^{-1}(y) \) for the penalty function (2.24). Suppose that \( f_\eta(x) \) is an \( N \)-parameter exponential family of densities

\[
f_\eta(x) = e^{\eta^T x - \psi(\eta)}, \quad (5.6)
\]

with \( \eta \) being the \( N \)-dimensional natural parameter vector, \( x \) the \( N \)-dimensional sufficient statistic vector, and \( \psi(\eta) \) the normalizing function ("cumulative generating function") that makes \( f_\eta(x) \) integrate to one with respect to some carrier measure. Also let \( \mu \) be the expectation vector \( \mu = E_\eta(x) \).

We consider the General Linear Model situation where \( \mathcal{M} \) and \( \mathcal{M}' \) are nested exponential subfamilies of (5.6): letting \( M = (M_1, M_2) \) be an \( N \times m \) matrix partitioned into \( N \times m_1 \), and \( N \times m_2 \) parts, define

\[
\eta_\theta = M \theta = (M_1, M_2) \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} = M_1 \theta_1 + M_2 \theta_2, \quad (5.7)
\]

where the \( m \)-vector \( \theta \) has been correspondingly partitioned. The model \( \mathcal{M} \) is taken to be the \( m \)-dimensional exponential subfamily of (5.6).

\[
\mathcal{M} : \eta = \eta_\theta, \quad (5.8)
\]

and \( \mathcal{M}' \) is the \( m_1 \)-dimensional subfamily having \( \theta_2 = 0 \),

\[
\mathcal{M}' : \eta = M_1 \theta_1. \quad (5.9)
\]

Laplace’s method gives a second-order accurate approximation to the marginal density for model \( \mathcal{M} \),

\[
f(x) \doteq (2\pi)^{m/2} f_{\hat{\theta}}(x) g(\hat{\theta}) |I(\hat{\theta})|^{-1/2}, \quad (5.10)
\]

where \( \hat{\theta} \) is the MLE, \( g(\cdot) \) the density for \( \theta \) under model \( \mathcal{M} \), and \( I(\theta) \) the expected Fisher information for \( \theta \),

\[
I(\theta) = E_{\theta}\{-\frac{\partial^2}{\partial \theta^2} \log f_\theta(x)\}. \quad (5.11)
\]

26
As usually derived, for example in equation (5) of Kass and Raftery (1981), \( I(\hat{\theta}) \) in (5.10) is replaced by the observed Fisher information \( i_x(\hat{\theta}) = -\frac{\partial^2}{\partial \theta^2} \log f_\theta(x) |_{\hat{\theta}} \); but \( i_x(\hat{\theta}) = I(\hat{\theta}) \) in exponential families.

Following the definitions in Table 1 and (2.21), (5.10) gives
\[
\begin{align*}
  r & = \frac{f(x)}{f(y)} = \frac{f_\hat{\theta}(x) g'(\hat{\theta}) |I(\hat{\theta})|^{-1/2}}{f_\hat{\theta}(y) g'(\hat{\theta}) |I(\hat{\theta})|^{-1/2}}.
\end{align*}
\]  
(5.12)

Applying Laplace’s method to \( \mathcal{M}' \), and using obvious notation, gives
\[
\begin{align*}
  r' & = \frac{f'(x)}{f'(y)} = \frac{f_\hat{\theta}'(x) g'(\hat{\theta}) |I(\hat{\theta})|^{-1/2}}{f_\hat{\theta}'(y) g'(\hat{\theta}) |I(\hat{\theta})|^{-1/2}} = \frac{f_\hat{\theta}'(x)}{f_\hat{\theta}'(y)}.
\end{align*}
\]  
(5.13)

Then
\[
\begin{align*}
  \frac{r}{r'} & = \frac{f_\hat{\theta}(x) g'(\hat{\theta}) |I(\hat{\theta})|^{-1/2}}{f_\hat{\theta}'(x) g'(\hat{\theta}) |I(\hat{\theta})|^{-1/2}}.
\end{align*}
\]  
(5.14)

Finally, taking \( g(\theta) \) to be Jeffreys’ invariant prior \( |I(\theta)|^{1/2} \), we get
\[
\begin{align*}
  \frac{r}{r'} & = \frac{f_\theta(x)}{f_\theta'(x)} = \hat{B}(x)
\end{align*}
\]  
(5.15)
as in (2.22).

This argument is a variant of Good’s “device of imaginary results”, (1987), also used by Spiegelhalter and Smith (1982). The exponential family assumptions make (5.15) second-order accurate, erring by factor \( 1 + O(n^{-1}) \) in repeated sampling situations. Less restrictive assumptions lead to less accurate versions of (5.15). If for example we allow \( \mathcal{M}' \) to be a curved exponential subfamily of \( \mathcal{M} \) then (5.15) may only be first-order accurate, erring by factor \( 1 + O(n^{-1/2}) \), and similarly if we replace Jeffreys’ prior with some other slowly varying function \( g(\theta) \).

Notice that the components of the double ratio \( r/r' \), \( r = f(x)/f(y) \) and \( r' = f'(x)/f'(y) \), each involve only one of the two models. This makes approximations like (5.15) easier to justify since we never have to compare models of different dimensionalities.

References


