Table 1. Frequentist Bayes factor and the Bayes factors under priors (a)-(c)for dyestuff data.

Frequentist	(a)	(b)	(c)
6.568	4.92424	8.67469	3.02869

Note that all the three Bayes factors constructed using noninformative priors (a)-(c) and the frequentist Bayes factor is a function of b. Figures 1 and 2 plot logarithm of Bayes factors against b for $m = 6$ and $m = 20$ (in each case $n_0 = 5$). It is clear that there is very good reconciliation of the Bayes factors under noninformative priors (a)-(c).

ADDITIONAL REFERENCES

Box, G. E. P. and Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis.* Wiley, New York.

Chernoff, H. (1954). On the distribution of the likelihood ratio, *J. Amer. Statist. Asso.* **25**, 573-578.

Datta, G.S. and Lahiri, P. (2000). A comparison of the Bayes factors and frequentist Bayes factor for a balanced one-way random effects model, *unpublished manuscript.*

REJOINDER

Bradley Efron and Alan Gous

This article was written under the following rule of thumb: no method that's been heavily used in serious statistical practice can be entirely wrong. The rule certainly applies to Fisherian hypothesis testing, but it also applies to Jeffreys and the BIC, leaving us to worry about Figure 1. The two scales of evidence seem to be giving radically different answers, even for sample sizes as small as $n = 100$.

Our paper localizes the disagreement to coherency, in this case sample size coherency, the key distinguishing feature of modern Bayesian philosophy. The BIC, along with any other methodology that acts coherently across different sample sizes, must share Figure 1's behavior, treating the smaller hypothesis M_o ever more favorably as n increases. Fisher's theory, which is usually presented with the sample size fixed, eschews sample size coherency in favor of a more aggressive demand for statistical power.

Professor Kass' good-natured commentary seems to share our ecumenical rule of thumb. His approach to the prosopagnosia data set, using BIC to screen possible working models before confronting the key inferential questions, is a reasonable beginning point. Predictably, it will be slanted toward smaller working models, with more of the nuisance variables deemed unimportant, than a similar screening based on standard hypothesis tests. In fact with $n = 20,000$, BIC could easily prefer *very* small models. Our own experience with the selenium analysis did not inspire confidence in the "biggest possible choice of n" rule, see the $n = 1312$ column of Table 8.

There is no question that hypothesis testing is greatly overused in statistical practice, and that many problems would be better phrased in terms of estimation, confidence intervals, or their Bayesian equivalents. There remains nevertheless a core of situations where hypothesis testing conveys the correct scientific attitude. The selenium experiment is a good example. The null hypothesis point $\theta = .5$ in (5.3) commands attention, and cannot be smoothly incorporated with its neighbors into a continuous prior density. Maybe we don't need a perfect delta function of prior belief at $\theta = .5$, a very narrow Gaussian curve serving just as well, but in any case the overall prior for θ must incorporate a bump near $\theta = .5$. Bumps make for mathematical difficulties, and neither Fisher's nor Jeffreys' model-selection methodology is as convincing as their estimation theories, but the hypothesis testing problem won't disappear just because it is awkward.

Professors Datta and Lahiri try various uninformative priors on an interesting components-of-variance example, partly to see how well our "frequentist Bayes factor" $B_{\text{freq}}(x)$, (2.30), works in a different setting. The big difference here is that the null hypothesis occurs at the extreme of the larger hypothesis instead of inside. We considered one such situation in Section 3.2, the one-sided hypothesis testing case. It is important to note that the approximation $B(\mathbf{x}) \doteq \widehat{B}(\mathbf{x})/\widehat{B}(\mathbf{y})$ that leads to $B_{\text{freq}}(x)$ must be modified in one-sided cases, the change going back to a one-sided application of Laplace's method in the Lemma of Section 2.2. The difference can be seen in the additional term $\log \Phi(x)/\Phi(y)$ in (3.9). Even without this correction, $B_{\text{freq}}(x)$ seems to be giving quite reasonable Bayes factors in Datta & Lahiri's two figures.

Kass also questions us about $B_{\text{freq}}(x)$ and its application to more complicated situations. One strong caveat: in higher dimensional problems, $d > 1$, Fisher's .90 choice for the breakeven quantile α_o is too low. This is the message of Figure 3, where for $d > 1$ we can't reconcile p-values with *any* prior unless α_o is increased. Table 9 suggests reasonable values of α_o for increasing dimension d. With this improvement to (2.30) kept in mind, $B_{\text{freq}}(x)$ seems quite useful. We haven't tried to extend it to nonnested testing situations, which are not easily handled by any frequentist methodology.

At its heart the Bayesian-frequentist controversy is an argument about principles,

not so much which principles as whether or not to use them at all. The Bayesian guiding principle is focussed on consistent decision-making across different frames of reference, sample-size coherency being a classic example. Examples of frequentist inconsistency, in which the Bayesian model-selection literature abounds, are apt to fall on deaf ears, frequentists being more focussed on just the problem at hand. Everyone uses Bayes rule, as Kass observes, but not everybody agrees on how it should be used in practice. Jeffreys' use in the model-selection problem leads to sample size coherency, while Fisher's use does not.

That brings us to Professor Kass' last and most difficult question: which model-selection paradigm do we teach our students? Fisher's scale seems perfectly suited to the common situation of fixed sample size and a straw-man null hypothesis that the investigators wish to disprove. However it is less satisfactory for more complicated problems involving multiple comparisons, data-mining, null hypotheses of genuine interest (as in the bioequivalence problem), or sequential decision making. Even slightly more complicated situations, like the selenium example, made us grateful for some Bayesian guidance in the form of $B_{\text{freq}}(x)$.

We are grateful to the discussants and the editors, both for producing this volume and allowing our participation. A preliminary version of this article was presented as the 1996 Morris DeGroot memorial lecture at Carnegie-Mellon University. Morris was a good-natured but always effective exponent of Bayesian thought, and this paper is dedicated to him.