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.

on the P integrals, ... it appears that there is not much difference in the practical recommendations." Qualitatively and roughly, Bayes factors may be brought to approximate agreement with frequentist tests when the prior on the alternative shrinks at rate $O(n^{-1/2})$ toward the null value. This was made precise by Andrews (1994). Efron and Gous show that, in fact, the prior may be chosen so as to obtain a very close agreement in one dimension, but the agreement is not so close in higher dimensions. They rightly put their finger on sample size coherency as the key restriction of the Bayesian approach to testing and model selection, and their discussion and remarks on this point are helpful. They present an example of hypothesis testing in which the conclusions to be drawn are important, and they apparently like the way their B_{freq} measures the evidence. What remains, however, is the deep and vexing problem of model selection.

I have recently come across a data set on prosopagnosia—a disorder in which patients have difficulty recognizing faces—where the goal is to better understand the cognitive deficiency. We developed a generalized ROC analysis in which probit regression is used with 7 explanatory variables defined by the experimental design, the goal being to make inferences about several contrasts. There are a total of just over 20,000 binary observations on roughly 40 subjects and to examine the contrasts of interest, something must be done to take account of the remaining explanatory variables. Thus, we have a familiar problem of "model selection," in quotes because one may well decide not to literally select a model.

My own preference would be to use BIC to sift through a wide variety of models (perhaps using MCMC), drawing conclusions based on the potentially many models that have substantial posterior probability according to BIC (i.e., prob = $\exp(BIC)/(1+\exp(BIC))$). However, we run into a problem, which the authors allude to toward the end of their Section 5: What value of the sample size n should be used in the penalty term? The difficulty is most easily appreciated in the one-way ANOVA context with k replications in I groups. At first glance, there are two possibilities: n could be the total number of observations $n = I \cdot k$ or, if we view each replication as an observation on an I-dimensional vector, we might instead take $n = k$. In many examples, the results are strikingly different. The same problem arises in our probit experimental design. I think the choices $n = I \cdot k$ and $n = k$ are both intuitively reasonable. It may seem preferable to take $n = k$, but keep in mind that in linear regression one would usually take n to be the total number of observations. Thus, if we take $n = k$ for replicated designs we create a discontinuity: adding very small epsilons to the 1's and 0's in the design matrix for the linear regression version of ANOVA creates a non-replicated linear regression that is essentially identical to the ANOVA design, yet we would be using a very differenct value of n than in the ANOVA design and would thus get very different results.

One potential way out of this bind is to begin by regarding the sample size as being as large as possible; in the one-way ANOVA setting we would take $n = I \cdot k$. In the terminology of Kass and Wasserman (1995), this would correspond to using a "unit-information prior" with n units of information in the sample. We might then ask how many units of information the prior should contain. That is, we might consider the possibility that the prior might contain c/n units of information. In the one-way ANOVA setting, we could choose $c = 1$ to obtain a BIC-based analysis with penalty term depending on $\log(I \cdot k)$, or we could instead choose $c = I$ to obtain a penalty term involving $\log(k)$. Furthermore, there is now an additional possibility of estimating c from the data—finding the value of c that provides the best fit. An interesting version of this approach to model selection (albeit more involved than I am describing here) has been proposed by George and Foster (2000).

This is a somewhat lengthy background to a simple question for the authors: Would they extend their B_{freq} to more complicated contexts? Can, for instance, their formula (2.30) be regarded as a general model-selection criterion, and would they apply it even in non-nested situations (which is crucial for applications of the kind I have mentioned)?

I'd like to come back to Jeffreys' remark (about the garden-variety small-sample problems most commonly treated) that his methods and Fisher's might not differ much in practice, and to a point that continues to bother me, at least a little. There is a pretty substantial body of work comparing Bayes factors to significance tests, following from Jeffreys and Savage (especially Edwards, Lindman, and Savage, 1963), including work by Berger and various colleagues. To me, the most important practical take-home message is that .05 really doesn't represent much evidence against the null. The authors provide a nice quote from Fisher that seems to give his perspective pretty clearly on the use of .05 as an appropriate conventional cutoff value, as well as setting the stage for their use of $\alpha_0 = .9$ as the breakeven point. Another relevant reference, however, is Fisher's famous description (in the introduction to his *Experimental Design*) of "A Lady Tasting Tea." There, he rejects the design involving 6 cups of tea, prefering one involving 8, because the p-value from exact identification with 6 cups would be $p = .05$ and this he considered insufficient evidence. In a similar vein, Gosset wrote (Pearson, 1938):

> [A test] doesn't in itself necessarily prove that the sample is not drawn randomly from the population even if the chance is very small, say .00001: what it does is to show that if there is any alternative hypothesis which will explain the occurrence of the sample with a more reasonable probability, say .05 ... you will be very much more inclined to consider that the original hypothsis is not true."

The point here is that Gosset picked the .05 level to correspond to a "reasonable" tail-area for an *acceptable* hypothesis. In my elementary statistics classes we can and sometimes do get into extended discussions about sample size, and the arbitrariness of any cutoff value, as well as the virtues of estimation rather than testing. However, statistics students—like everyone else—need to have clear and concise guidelines they can remember. I tell students to interpret $p > .05$ as "no evidence against H_0" and $p < .05$ as "some evidence against H_0" but I frequently emphasize that a p near .05 would be pretty shaky. In doing this I think I am offering a bit more conservative advice than is traditional, but it seems to me to be good advice that is more or less consistent with its Gosset/Fisher origins while being enlightened by subsequent analysis. So, my last question is this: How has any of this reconsideration of the foundations affected their thinking, and what scale of evidence do they tell their students to use?

Let me now, with great pleasure, thank the authors for a stimulating article.

ADDITIONAL REFERENCES

Edwards, W., Lindman, H., and Savage, L.J. (1963). Bayesian statistical inference for psychological research. *Psychol. Rev.* **70**, 193-242.

Efron, B. (1986). Why isn't everyone a Bayesian? *Amer. Statist.* **40**, 1-11.

George, E.I. and Foster, D.P. (2000). Calibration and empirical Bayes variable selection. *Biometrika* **87**, 731-748.

Pearson, E.S. (1938) "Student" as statistician, in William Sealy Gosset, 1876-1937. *Biometrika* **30**, 210-250.

G. S. Datta and P.Lahiri

University of Georgia and University of Nebraska-Lincoln

Starting from a simple testing hypothesis problem, the authors have very successfully demonstrated how one can possibly reconcile two apparently different approaches (Fisher's and Jeffreys') to model selection. In the process, they have proposed a new frequentist Bayes factor for model selection.

G. S. Datta is Professor, Department of Statisitcs, University of Georgia, Athens, GA 30602, U.S.A; email: gauri@stat.uga.edu. P. Lahiri is Milton Mohr Distinguished Professor of Statisitcs, Department of Mathematics and Statistics, University of Nebraska-Lincoln, Lincoln, NE 68588-0323, U.S.A; email: plahiri@math.unl.edu. Datta was partially supported by NSF grants SBR-9705145 and DMS-0071642 and by NSA. Lahiri was supported in part by NSF grant SBR-9978145. The authors thank Mr. Shijie Chen for computing help.