function $\phi(i)$ proportional to $1/i(\log i)^2$. The true probability mass function is taken to be θ^* which differs from ϕ for small i and is equal to ϕ for all large i. The posterior has the unfortunate property of concentrating at ϕ rather than in neighborhoods of θ^*. From this inconsistency, we conclude that the Dirichlet prior does not locally match θ^*. Moreover, the Dirichlet prior assigns zero mass to the relative entropy neighborhood $\{\theta: \sum_i \theta^*(i) \log \theta^*(i)/\theta(i) < \varepsilon\}$ for ε sufficiently small.

Freedman and Diaconis have pointed out that ϕ and θ^* have infinite entropy $H(\theta^*) = \sum_i \theta^*(i) \log 1/\theta^*(i)$. One might think that the inconsistency is a result of the infinite entropy; however, even if certain finite entropy mass functions are used in the construction, inconsistency will still result. It is enough that θ^* and ϕ have tails proportional to $1/i^\alpha$ where $1 < \alpha < \frac{4}{3}$. (The verification of inconsistency closely parallels Sections 2 and 3 of Freedman and Diaconis, 1983). In Freedman (1963), finite entropy appears as part of a condition for consistency. We now know that the finite entropy assumption is extraneous. It is the *relative* entropy that matters for Bayes consistency.

In summary we have discussed some inadequacies of the Dirichlet prior as revealed by the analysis of Diaconis and Freedman and we have pointed toward stronger consistency and merging results obtainable for other priors.

REFERENCES

BARRON, A. R. (1985a). The strong ergodic theorem for densities: Generalized Shannon–McMillan–Breiman theorem. *Ann. Probab.* **13** 1292–1303.

BARRON, A. R. (1985b). Logically smooth density estimation. Ph.D. Thesis, Department of Electrical Engineering, Stanford University.

BARRON, A. R. (1986). On uniformly consistent tests and Bayes consistency. Preprint.

LO, A. Y. (1984). On a class of Bayesian nonparametric estimates: I. Density estimates. *Ann. Statist.* **12** 351–357.

The references to Diaconis, Freedman, and Schwartz are the same as in the paper under discussion.

DEPARTMENT OF STATISTICS
UNIVERSITY OF ILLINOIS
1409 W. GREEN STREET
URBANA, ILLINOIS 61801

JAMES BERGER

Purdue University

The very lucid paper of Diaconis and Freedman is full of stimulating ideas and discussion. The ideas fall roughly into three categories: (i) inconsistency of Bayes rule, (ii) frequentist-Bayesian interrelationships including the "what if" method, and (iii) new Bayesian devices and techniques. My comments will be grouped by these categories, and will be restricted (because of space considerations) solely to a Bayesian view of the situation.

1. Inconsistency of Bayes rules. The fact that *parametric* Bayesian analysis virtually always yields consistent estimators (Bayesian "stable estimation") may have, at one time, lulled Bayesians into believing that consistency was not a concern. Freedman (1963, 1965) disabused Bayesians of this notion (or at least should have), and this and the following paper provide convincing further evidence that Bayesians should be concerned with consistency in nonparametric and infinite parametric problems.

It is important to emphasize the nature of the inconsistency that can arise in these problems. Doob's theorem (see Corollary A2 in the appendix of the paper), shows that the posterior mean is consistent for θ in a set, Θ_0, of prior probability 1. The Freedman and Freedman–Diaconis results show, however, that there are θ very close to Θ_0 (indeed limit points of Θ_0) for which the posterior mean is not consistent. Thus the Dirichlet process prior concentrates on Θ_0 = {discrete distributions}, and for $\theta \in \Theta_0$ consistency problems are not to be expected (see also the commentary of H. Doss), but for $\theta \notin \Theta_0$ Diaconis and Freedman have constructed problems where inconsistency can result.

I feel this shows that a Bayesian has to be especially careful in constructing the prior for nonparametric or infinite parametric situations; in particular use of "convenient" priors may be more dangerous in nonparametric than in parametric Bayesian analysis. One could avoid the inconsistency problem by constructing the prior, μ, to concentrate on those θ deemed reasonable a priori (rather than settling for a convenient "dense" Θ_0, as the Dirichlet process prior does), or one could explicitly worry about consistency of the selected prior, but work is involved in either approach. In this respect it should be realized that the "tail-free" priors and "neutral-to-the-right" priors are positive steps in the latter direction; they are priors for which consistency has been verified in nonmixed problems. (Also, Freedman (1963) showed how such priors could be modified to incorporate subjective information.) Very little Bayesian research has been done on the alternate approach of developing priors which "live" on the right spaces, partly because of the calculational allure of priors such as the Dirichlet process priors; hopefully such development will now be forthcoming.

A general question I have for the authors is: How likely is it for one to encounter a consistency problem in practice? There are at least two reasons for asking this. The first is that many of the difficulties here seem to be caused by the concentration of the Dirichlet process prior on the set of discrete probability measures. From the beginning, many Bayesians (though certainly not all) have been very leery of the Dirichlet process (when used as a prior for all or continuous densities) because of this unnatural concentration. The weird occurrences in these two papers (and also the commentary of H. Doss), reinforce the notion that it is the Dirichlet process prior which is the main problem. There is, of course, the Freedman (1965) result, which shows that consistency will only tend to occur on a first category set, but the implication of this is unclear since (for instance) there exist first category sets on the real line whose complements have Lebesgue measure zero. Being consistent, except on a set of Lebesgue measure zero, would be quite satisfactory to many.

The second reason to wonder about the practical importance of these inconsistency results in nonparametric settings is that, often, the function or distribution being estimated is "nicer" than a typical element of the nonparametric class being considered; a too large nonparametric class is often assumed for mathematical convenience. Consistency may obtain at the realistic "nice" functions, as it does at the "nice" strongly unimodal densities, h, in the major location example of the paper.

I am not really trying to argue that the examples in these papers are artificial; indeed, one of the major strengths of the papers is that they exhibit inconsistency in relatively natural problems. Nevertheless, if the authors have developed any feel for the chance of encountering inconsistency in practice, it would be nice to hear.

Also, in this regard, the authors refer to analyses by Jeffreys, Fraser, Box and Tiao, and Johns, at least some of which are entirely parametric. It is unclear from the paper whether it is being claimed that these particular analyses can actually be inconsistent, or whether it is merely the case that related analyses, using (say) Dirichlet process priors, can be inconsistent. In the first case, there is obviously "evidence" that Bayesians are likely to encounter consistency difficulties.

Before leaving this subject, I feel compelled to also mention the other side of the coin. Although I do not think it was the intention of the paper to make any "Bayesian versus frequentist" value judgements, some may interpret the paper as an argument against Bayesian analysis. Such an interpretation must be tempered by the realization that consistency can also be a problem for frequentist procedures. Even more to the point from a practical perspective, the advantage in most finite sample situations of Bayesian analysis, as opposed to frequentist "large sample theory," is often not appreciated. There is a massive frequentist industry which derives large sample asymptotic results, and then "hopes" that the results work okay for finite samples. What is not commonly appreciated is that Bayes procedures will typically have the same large sample behavior, and yet are also probably reasonable for small samples. If one has a variety of "equivalent" large sample procedures, why not use one which is also constructed to be good for small samples, instead of simply choosing one "at random"?

Another apect of this "other side of the coin" is that it is precisely in high dimensional parametric and even nonparametric problems that it can be most crucial to utilize *subjective* prior information. It will be rare to have enough data to illuminate all dark corners of a high dimensional problem, and subjective input (including model development) is often unavoidable. As one frequentist-type example, consider Stein estimation in nonsymmetric multivariate settings. It is fairly well established (cf. Berger and Berliner, 1984) that one cannot avoid *subjectively* determining where and how one should "shrink" the least-squares estimator. And in nonparametrics there are often compelling reasons to attempt to subjectively specify the rough shape or at least the smoothness of the function or distribution to be estimated. Thus, while Bayesians may encounter unexpected difficulties (such as consistency) in these problems, the need and incentive for Bayesian input is greatly enlarged.

These last comments were not meant to prove anything. The point was merely to emphasize that frequentist analysis is by no means clearly superior to Bayesian analysis when considering the broad area of utilization of large sample theory.

2. Frequentist-Bayesian interrelationships. There are a large number of coincidental and technical relationships between frequentist and Bayesian analysis, many of which are mentioned in the paper and discussed extensively in the references therein. Though interesting, these relationships are not as important as the operational issue of when a Bayesian *should* make use of frequentist ideas. The related issue, of when a frequentist *must* make use of Bayesian methods, is a much lengthier topic, and will not be discussed here. (Some examples and references to this issue are given in the paper: others can be found in Berger and Wolpert (1984), Berger and Sellke (1984), and Berger (1985).)

The italicized words *should* and *must*, in the above paragraph, reflect my beliefs that a Bayesian can sometimes utilize frequentist ideas to make life easier, whereas a frequentist is often forced by reality to completely abandon ship. No attempt will be made to support the latter part of this statement, but I will digress to discuss the *robust Bayesian* motivation for the first part of the statement. This digression is somewhat out of place here, but my subsequent comments on the Bayesian uses of frequentist measures that are proposed in the paper would be otherwise unintelligible.

The robust Bayesian position can be roughly stated as follows: *An answer to a statistical problem is a good answer only if there is substantial reason to believe that the answer would approximately equal the posterior Bayes answer for any reasonable sampling model and prior distribution (and loss function in a decision problem) entertained.* Thus, suppose it is roughly felt that X is $N(\theta, 1)$, that θ is $N(0, 1)$, and that the loss in estimating θ is increasing in $|\theta - a|$. Then the Bayes estimate is $a^* = \frac{1}{2}x$. If $x = 1$ is observed, it can be seen that $a^* = \frac{1}{2}$ is a good answer, in that small reasonable variations in the model, prior, and loss do not change the Bayes estimate much. For $x = 5$, however, the situation is very different. Changing either of the distributions to, say, a similar Cauchy distribution will radically alter the Bayes estimate, so $a^* = 2.5$ is not necessarily a good estimate. No effort will be made to defend this robust Bayesian belief here; see Good (1983), Berger (1984), and Berger (1985) for such defence.

The most natural way to investigate Bayesian robustness is through what Leamer (1978) calls *global sensitivity analysis*: vary the model, prior, and loss over reasonable ranges and see what happens to the posterior Bayesian answer. (Recent works in this direction, which contain many other references, are Berger and Berliner (1983) and Berger (1985).)

The point of this aside is that the robust Bayesian definition of a good answer does not involve frequentist ideas in any way. The data, x, is always treated as known; the "variables" in the analysis are the model relating x and the unknown θ of interest, the prior for θ, and any loss to be considered. (Actually, Bayesians see little conceptual difference between models and priors. Also, we are consider-

ing here only the final "inference about θ" stage of the analysis. In topics such as experimental design, the data are not yet known and at least partly frequentist measures become necessary.) If robust Bayesians can be satisified that global sensitivity obtains, they will look no further.

From this viewpoint, frequentism may come into play only when global sensitivity is unattainable (due, say, to an inability to sufficiently refine the usually subjective inputs of model, prior, and loss), or is unverifiable (due to technical limitations in carrying out the global sensitivity study). In nonparametric or even high-dimensional parametric problems, both difficulties are present with a vengeance. It can be hard to perform *any* sensible Bayesian analysis, much less carry out an extensive sensitivity study. There are then various roles that frequentist ideas can play.

The role that is concentrated on in this paper is the "negative" one that bad frequency performance is often (but not always) an indicator of a definite lack of Bayesian robustness. A lack of consistency, for seriously entertained θ, would be perhaps the most drastic indication of such a lack of robustness. A number of other such frequentist indicators are discussed in Berger (1985).

Another frequentist-based tool that is discussed in the paper, as being of possible interest to Bayesians, is the "what if" method. Note, first of all, that the robust Bayesian viewpoint could be called a "what if" approach; *what if* the model, prior, and loss were changed in reasonable ways? The "what if" method discussed in the paper is quite different, however; it asks "What if we had observed different data?" The relevance of this to a robust Bayesian is not clear, since the robust Bayesian cares about sensitivity to assumptions only for the observed data. Thus, in our earlier simple example, the robust Bayesian can feel reasonably satisfied with his answer of $a^* = \frac{1}{2}$ when $x = 1$ is observed, and will not care that he might have been unhappy with his model or prior had he happened to observe $x = 5$. The changes that would be entertained in the model or prior, upon observing $x = 5$, will have little effect for $x = 1$. The general principle is that the Bayesian will not try to protect against features of the model or prior that are irrelevant for the data at hand. Note that this is part of the fundamental distinction between conditional and unconditional statistical analysis, a distinction which, to many, is much more crucial than use of a prior distribution.

We do not here defend the robust Bayesian version of "what if" as opposed to the frequentist version discussed in the paper; the goal has been simply to indicate that there is a crucial difference. Also, we would not state that the frequentist "what if" method is without value; it is just not clear when it can provide insight not available by a prior sensitivity study. In the example of inconsistency in estimating the location parameter, I would guess that, for a given large sample, the conclusion would be quite sensitive to the choice of the prior, so that a prior sensitivity study would reveal the problem. Even if this were not the case, one would probably not have to leave the given data to see a potential problem; calculating the posterior for subsets of the data would presumably reveal the oscillatory behavior of the posterior. (Of course, looking at such subsets is somewhat "frequentist what-iffish" in nature.)

The two situations in which it is clear that frequentist measures can be useful to a Bayesian are (i) when Bayesian calculations are very difficult compared to frequentist calculations, and (ii) when developing "automated" Bayesian procedures for use in (say) computer packages. Even then, interest (to a Bayesian) in a frequentist measure occurs primarily when it can be interpreted in a Bayesian fashion; the following is a standard example.

EXAMPLE. Let X denote the random observation in an experiment with unknown θ, and suppose that $C(x)$ (a subset of Θ for each x) is a $100(1-\alpha)\%$ confidence procedure; thus, for all θ,

$$P_\theta(C(X) \text{ contains } \theta) = 1 - \alpha. \tag{1}$$

If μ is a prior on Θ, it follows that

$$E^\mu P_\theta(C(X) \text{ contains } \theta) = 1 - \alpha. \tag{2}$$

Now, a Bayesian would be interested in the posterior probability that θ is in $C(x)$ (for the observed x, of course); denote this by $\delta_\mu(x)$. But it is easy to see that

$$E^m \delta_\mu(X) = E^\mu P_\theta(C(X) \text{ contains } \theta) = 1 - \alpha, \tag{3}$$

where m is the marginal distribution of X. But knowing that (3) holds, when α is small, is useful information to a Bayesian who has difficulty in working with $\delta_\mu(x)$ directly, in that it then seems very likely (with respect to m) that $\delta_\mu(x)$ is near 1. And this holds for *any* μ, so that Bayesian robustness seems likely to be present. There is, of course, no guarantee that $\delta_\mu(x)$ is near 1 for the actual x obtained, but there is certainly reason to be optimistic (when α is small).

For the two situations mentioned before the example, it is easy to see the value of (3). Examples exist (see Berger and Wolpert, 1984) where δ_μ is very difficult to calculate for *any* reasonable μ, and yet it is almost trivial to verify (1) (and hence (3)). Such examples are somewhat rare, but they do exist. And the attraction of (3) in "automated" statistics is that it can impart a feeling of Bayesian robustness without the need for a sophisticated sensitivity study (which users of automated procedures may not be able to perform). Note that (3) need not hold for all priors; it need hold only for the class of "reasonable" priors (cf. Morris, 1983).

The important distinction in the above use of frequentist measures, by a Bayesian, is that there is no desire to involve nonobserved x in the analysis. The frequentist measure merely provides a convenient route to a possibly useful Bayesian measure.

As a final comment on "automated procedures," the development of inherently robust Bayesian procedures is an important Bayesian research goal (see Berger, 1985). One major thrust of the Diaconis–Freedman program can be interpreted in this light; namely, the development of priors that are guaranteed to be consistent.

3. New Bayesian devices and techniques. Bayesians are always excited by new Bayesian tools, and at least two are discussed in the papers of Diaconis and Freedman. One is the technique by which difficult Bayesian calculations can be performed by a limiting argument. While related calculations have been carried out before, the very general discussion here (cf. Section 4 in the second paper) should prove very useful to Bayesians.

The second new tool is the derivative given in Theorem 4. For parametric classes of priors, the study of the derivative of posterior features of interest, with respect to the parameters of the prior, has come to be called *local sensitivity* (cf. Leamer (1978) and Polasek (1984)) and can indicate features of the prior that are particularly influential and which, hence, may require more careful consideration. For the most part, previous work has concentrated on local sensitivity of the posterior mean and covariance matrix in conjugate prior situations; the nonparametric generalizations in this paper (see Appendix B for the relevant formula for the posterior mean) are exciting developments. In line with the previous discussion on "robust Bayesianism," I am most excited about the use of these derivatives to indicate "directions" in which the answer is particularly sensitive to the prior input. In terms of the Gateaux derivative, this is somewhat more intuitive; letting the prior be $(1 - \varepsilon)\mu + \varepsilon\nu$, sending ε to zero, and finding ν for which the directional derivative is largest, may well indicate where additional prior elicitation efforts or sensitivity studies should be concentrated.

An additional attractive feature of using the Gateaux derivative is that it ties in well with the most promising formal approach to global sensitivity, which is to investigate the range of the Bayesian measure of interest as the prior ranges over the "ε-contamination" class $\{\mu = (1 - \varepsilon)\mu_0 + \varepsilon\nu\}$, where μ_0 is an elicited prior, ε reflects the possible inaccuracy in this specification, and ν is some class of plausible contaminations. The attractiveness of global sensitivity studies for this class is indicated by Huber (1973), Berger and Berliner (1983), and Berger (1985). The tie-in with local sensitivity via the Gateaux derivative might lead to a nice unification of Bayesian sensitivity theory.

REFERENCES

BERGER, J. (1984). The robust Bayesian viewpoint. In *Robustness of Bayesian Analyses* (J. Kadane, ed.). North-Holland, Amsterdam.

BERGER, J. (1985). *Statistical Decision Theory and Bayesian Analysis.* Springer, New York.

BERGER, J. and BERLINER, L. M. (1983). Robust Bayes and empirical Bayes analysis with ε-contaminated priors. Technical Report, Purdue University. To appear in *Ann. Statist.* **14.**

BERGER, J. and BERLINER, L. M. (1984). Bayesian input in Stein estimation and a new minimax empirical Bayes estimator. *J. Econometrics* **25** 87–108.

BERGER, J. and SELLKE, T. (1984). Testing a point null hypothesis: The irreconcilability of significance levels and evidence. Technical Report, Purdue University. To appear in *J. Amer. Statist. Assoc.* **81.**

BERGER, J. and WOLPERT, R. (1984). *The Likelihood Principle.* IMS Monograph Series, Hayward, Calif.

FREEDMAN, D. (1963). On the asymptotic behavior of Bayes estimates in the discrete case I. *Ann. Math. Statist.* **34** 1386–1403.

FREEDMAN, D. (1965). On the asymptotic behavior of Bayes estimates in the discrete case II. *Ann. Math. Statist.* **36** 454–456.

GOOD, I. J. (1983). *Good Thinking: The Foundations of Probability and Its Applications.* Univ. Minnesota Press, Minneapolis.

HUBER, P. (1973). The use of Choquet capacities in statistics. *Bull. Inst. Internat. Statist.* **45** 181–191.

LEAMER, E. E. (1978). *Specification Searches.* Wiley, New York.

MORRIS, C. (1983). Parametric empirical Bayes inference: Theory and applications (with Discussion). *J. Amer. Statist. Assoc.* **78** 47–65.

POLASEK, W. (1984). Multivariate regression systems: Estimation and sensitivity analysis of two-dimensional data. In *Robustness of Bayesian Analyses* (J. Kadane, ed.). North-Holland, Amsterdam.

DEPARTMENT OF STATISTICS
PURDUE UNIVERSITY
WEST LAFAYETTE, INDIANA 47907

MURRAY K. CLAYTON

University of Wisconsin

The two papers by Diaconis and Freedman which are under discussion contain a series of interesting and nicely presented results. The philosophical issues which they raise are thought-provoking and merit attention. Their papers also give a useful review touching on a number of topics of interest to frequentists and Bayesians.

For simplicity, in the ensuing comments I shall refer to Diaconis and Freedman (1986a) as DFa and Diaconis and Freedman (1986b) as DFb. My comments touch on three topics: the technical aspects of DFa, the philosophical implications of the results in DFb, and the extension of the "what if" method in DFb to Bayesian robustness.

The model (1.1) of DFa and the accompanying priors seem innocuous, and it is somewhat disconcerting that they can lead to inconsistency. Theorem 1 of DFa says that the posterior for θ will fail to converge even though h has a global maximum at 0. Theorem 3 states that using a symmetrized prior might not help; we can even get the posterior law of the data wrong. On the other hand, perhaps the consoling message from DFa is that if $\log \alpha'$ is convex, then in the setting of Theorem 1 the posterior for θ *will* converge. Less helpful is the fact that the posterior will converge if the (unknowable) density h is strongly unimodal.

The discretization results of Section 4 of DFa can be used to approximate the solutions to decision problems in the undominated case. In Clayton (1985), I used a form of discretization with a Dirichlet process prior to approximate the worth of optimal rules for a sequential problem. I conjectured in that paper that discretization could be used to construct nearly optimal rules. (The construction of *optimal* rules is practically impossible unless the Dirichlet parameter has a finite support.) It seems possible to use the results of Section 4 of DFa to prove that conjecture.

How important is this issue of inconsistency to a Bayesian? I think Diaconis and Freedman are right in DFb to consider separately the classical and subjective Bayesians, even though many Bayesians have the characteristics of both groups.