The Annals of Statistics
1986, Vol. 14, No. 1, 1–26

SPECIAL INVITED PAPER

ON THE CONSISTENCY OF BAYES ESTIMATES

BY PERSI DIACONIS[1] AND DAVID FREEDMAN[2]

Stanford University and University of California, Berkeley

We discuss frequency properties of Bayes rules, paying special attention to consistency. Some new and fairly natural counterexamples are given, involving nonparametric estimates of location. Even the Dirichlet prior can lead to inconsistent estimates if used too aggressively. Finally, we discuss reasons for Bayesians to be interested in frequency properties of Bayes rules. As a part of the discussion we give a subjective equivalent to consistency and compute the derivative of the map taking priors to posteriors.

1. Consistency of Bayes rules. One of the basic problems in statistics can be put this way. Data is collected following a probability model with unknown parameters; the parameters are to be estimated from the data. Often, there is prior information about the parameters, for example, their probable sign or order of magnitude. Many statisticians express such information in the form of a prior probability over the unknown parameters. Estimates based on prior probabilities will be called Bayes estimates in what follows.

This paper studies points of contact between frequentist and Bayesian statistics. We derive frequency properties of Bayes estimates and suggest a Bayesian interpretation for some frequentist computations. Our main concern is consistency: as more and more data are collected, will the Bayes estimates converge to the true value for the parameters?

If the underlying probability mechanism has only a finite number of possible outcomes (tossing a coin or die) and the prior probability does not exclude the true parameter values as impossible, it has long been known that Bayes estimates are consistent. As will be discussed below, if the underlying mechanism allows an infinite number of possible outcomes (e.g., estimation of an unknown probability on the integers), Bayes estimates can be inconsistent: as more and more data comes in, some Bayesian statisticians will become more and more convinced of the wrong answer. The class of tail-free and Dirichlet priors was introduced to insure consistency in such settings. We present examples showing that mechanical extension of such priors to other very similar settings leads to inconsistent estimates.

In Section 2 we review other points of contact between the mathematics of frequentist and Bayesian statistics. In Section 3 we offer two Bayesian uses for

Received March 1984; revised August 1985.
[1]Research partially supported by National Science Foundation Grant MCS80-24649.
[2]Research partially supported by National Science Foundation Grant MCS83-01812.
AMS 1980 *subject classifications.* 62A15, 62E20
Key words and phrases. Consistency, location problem, Dirichlet prior, merging of opinions, foundations, uniformities.

frequentist computations. The first is a subjective equivalent to consistency involving intersubjective agreement. The second uses frequency computations as a way of thinking about priors. As part of the discussion, we compute the derivative of the map taking priors to posteriors. Mathematical details are given in the appendices.

To define things, consider a family of probabilities $\{Q_\theta: \theta \in \Theta\}$ on a space $\mathcal{X}$. Write Q_θ^∞ for the infinite product measure on $\mathcal{X}^\infty$ which makes the coordinate random variables $X_1, X_2, \ldots$, independent with common distribution Q_θ. We will assume throughout that $\mathcal{X}$ and Θ are Borel subsets of complete separable metric spaces. Let μ be a prior probability on Θ. Let P_μ be the joint distribution of the parameter and the data:

$$P_\mu(A \times B) = \int_A Q_\theta^\infty(B)\,\mu(d\theta)$$

for Borel sets A and B. The *posterior* is the P_μ-distribution of the parameter θ given the data $X_1, \ldots, X_n$; we denote this by $\mu_n(d\theta|X_1, \ldots, X_n)$. The usual Bayes estimate is just the mean of the posterior.

Here are a few typical examples:

Coin-tossing. $\mathcal{X}$ has two points, H and T for heads and tails, respectively. The parameter space Θ can be taken as the unit interval: $\theta \in \Theta$ is the probability of a head. Then $\mathcal{X}^\infty$ is the space of sequences of heads and tails; Q_θ^∞ makes the sequence independent; in any position there is chance θ of getting a head and $1 - \theta$ of getting a tail. We call this a "θ-coin." Informally, P_μ may be described as follows: choose θ at random from μ, then toss a θ-coin. Of course, the posterior $\mu_n(d\theta|X_1, \ldots, X_n)$ has a density with respect to the prior, proportional to the likelihood function:

$$\theta^S(1-\theta)^{n-S} \Big/ \left[\int \theta^S(1-\theta)^{n-S}\mu(d\theta)\right],$$

where S is the number of heads and $n - S$ the number of tails. Thus, the Bayes estimate is

$$\int \theta^{S+1}(1-\theta)^{n-S}\mu(d\theta) \Big/ \int \theta^S(1-\theta)^{n-S}\mu(d\theta).$$

Rolling a die. This is the same as coin tossing, except that $\mathcal{X} = \{1, 2, 3, 4, 5, 6\}$, and Θ is the "6-simplex," all sequences $\theta_1, \ldots, \theta_6$ of length six whose terms are nonnegative and add to one:

$$\theta_i \geq 0 \quad \text{and} \quad \sum_{i=1}^{6} \theta_i = 1.$$

Rolling an infinite die. The same, except $\mathcal{X} = \{1, 2, 3, \ldots\}$ and Θ is an infinite-dimensional simplex. Despite the superficial similarity, all the paradoxes of inconsistency already appear in this case. From our point of view, this is the simplest natural example of a nonparametric problem: estimating an unknown probability on the positive integers. On the other hand, we are still in the dominated case, so the posterior can still be computed in the usual way: μ_n denotes this familiar version of the posterior.

The real line. Now $\mathcal{X} = (-\infty, \infty)$, and Θ is the set of all probabilities on $\mathcal{X}$. This is another nonparametric problem: estimating an unknown probability on the line. There is an additional complication: Θ is undominated. The posterior still exists but in general there is no explicit formula for it. There will usually be many versions of the posterior and this causes some technical difficulties.

We will use the weak-star topology: if μ_n and μ are probabilities on Θ, then $\mu_n \to \mu$ iff $\int f\,d\mu_n \to \int f\,d\mu$ for all bounded continuous functions f on Θ. We denote point mass at θ by δ_θ. Thus, $\mu_n \to \delta_\theta$ iff $\mu_n(U) \to 1$ for every neighboarhood U of θ.

In the coin tossing case, Θ is the unit interval; $\mu_n \to \mu$ weak-star if and only if $\mu_n[0, x] \to \mu[0, x]$ for all intervals such that $\mu\{x\} = 0$. Turn next to the infinite die. Then $\mu_n \to \mu$ if and only if

$$\mu_n\{\theta|\theta_1 \le x_1 \text{ and } \theta_2 \le x_2 \ldots \text{ and } \theta_k \le x_k\}$$
$$\to \mu\{\theta|\theta_1 \le x_1 \text{ and } \theta_2 \le x_2 \ldots \text{ and } \theta_k \le x_k\}$$

for all k and all x's such that

$$\mu\{\theta|\theta_1 = x_1 \text{ or } \theta_2 = x_2 \ldots \text{ or } \theta_k = x_k\} = 0.$$

Finally, take the line. Changing the imagery a little, $\mu_n \to \mu$ if and only if

$$\int_\Theta \left[\prod_{i=1}^k \int f_i(x)\,\theta(dx)\right]\mu_n(d\theta) \to \int_\Theta \left[\prod_{i=1}^k \int f_i(x)\,\theta(dx)\right]\mu(d\theta)$$

for all k and all bounded continuous functions f_i on $\mathcal{X}$.

Consistency. The pair (θ, μ_n) is *consistent* if for Q_θ^∞-almost all sequences, the posterior μ_n converges to point mass at θ in the weak-star topology. The weak-star topology has the fewest open sets of any natural topology, so it is fairly easy for posteriors to be consistent. Minor technical difficulties apart, if (θ, μ_n) is consistent in our sense, the Bayes estimate for θ—in the sense of a posterior mean—will be consistent too. We say μ_n is *consistent* if (θ, μ_n) is consistent for all θ. Notice that consistency depends to some extent on the version selected for μ_n. Often, there is only one natural version, as in the dominated case. Sometimes, a natural version can be selected on the basis of continuity in the data: see Zabell (1979) and Tjur (1980). Sometimes, however, there is no sensible way to resolve the ambiguity. When there is a natural version of μ_n, as in the dominated case, we will say that (θ, μ) is consistent rather than (θ, μ_n).

In the coin-tossing example, what does it mean for (θ, μ_n) to be consistent? Just that a Bayesian with prior μ, who happens to be tossing a θ-coin, will eventually find this out: his posterior will concentrate in smaller and smaller intervals around θ as more and more data comes in. Likewise for the infinite die. More specifically, (θ, μ) is consistent if and only if for any positive integer k and any small positive ε,

$$\mu_n\{N_{k\varepsilon}|X_1, \ldots, X_n\} \to 1 \quad \text{as } n \to \infty \quad \text{a.e. } Q_\theta^\infty,$$

where

$$N_{k\varepsilon} = \{\phi\colon \phi \in \Theta \text{ and } |\phi_i - \theta_i| < \varepsilon \text{ for } i = 1, \ldots, k\}.$$

For coin-tossing, (θ, μ) is consistent iff μ assigns positive mass to every open interval around θ. For the infinite die the situation is much more complicated—and that is what the present discussion is all about.

Doob (1948) proved a very general theorem on consistency: one implication is that (θ, μ_n) is consistent for μ-almost all θ. Thus, a Bayesian with prior μ can be sure that the posterior will converge. (This does not depend on the version, because null sets do not matter here.) Doob's work has been extended by Breiman, Le Cam, and Schwartz (1964). However, a frequentist using a Bayes rule will want to know for which θ's the rule is consistent.

For smooth, finite-dimensional families, (θ, μ) is consistent if and only if θ is in the support of μ. See Freedman (1963) or Schwartz (1965). (The support is the smallest closed set of probability 1.) But the assumption of finite dimensionality is really needed. For example, take $\mathcal{X}$ to be the positive integers. Take Θ to be the set of all probabilities on $\mathcal{X}$. Take θ to be the geometric distribution with parameter $\frac{1}{4}$. Freedman (1963) constructed a prior μ with the following properties:

- Every open neighborhood of θ has positive μ-probability.
- For θ^∞-almost all sequences, the posterior converges to point mass at a geometric distribution with parameter $\frac{3}{4}$.

This example is generic in a topological sense: for most priors μ and most parameters θ, the pair (θ, μ) is inconsistent. To make this precise, we need the notion of "category." A set is of the first category if it is contained in a countable union of closed, nowhere dense sets. First-category sets are the topological equivalents of null sets. Put the weak-star topology on $\pi(\Theta)$, the set of priors on Θ. Then the set of consistent pairs (θ, μ) is of the first category in $\Theta \times \pi(\Theta)$. See Freedman (1965).

Tail-free and Dirichlet priors. The existence of such counterexamples suggests the need for some careful investigation. Are there priors consistent for all parameters? (We will call such priors "consistent.") For countable $\mathcal{X}$, Freedman suggested using tail-free and Dirichlet priors, showing that these priors are consistent at all parameters. This work was extended by Fabius (1964), who showed how to construct consistent tail-free priors for any complete, separable metric $\mathcal{X}$. These ideas were further developed by Ferguson (1973, 1974) with an excellent survey of the literature. Other extensions include "neutral-to-the-right priors." See Doksum (1974).

Here is a brief description of tail-free priors. For definiteness, consider the problem of estimating an unknown probability θ on the positive integers. Write θ_i for the ith coordinate of θ. We visualize the prior as randomly selecting a probability on the positive integers. Even more crudely, what the prior does is to randomly distribute a total mass of 1 among the integers. Thus, θ_1 is the randomly selected mass assigned to the integer 1, while θ_2 is the mass assigned to 2, and so on.

The simplest tail-free prior on the integers can be described by "stick-breaking." Let $S_1, S_2, \ldots,$ be independent and uniform over $[0, 1]$. Think of a stick of unit length. Break off a piece of length $\theta_1 = S_1$. This leaves a remaining piece of length $1 - \theta_1$. Now break off $\theta_2 = S_2(1 - \theta_1)$, $\theta_3 = S_3(1 - \theta_1 - \theta_2)$, and so forth.

Freedman (1963) suggested a useful extension: the distribution of any finite number of the θ_j, say $\theta_1, \ldots, \theta_N$, can be specified in an essentially arbitrary way. Then the prior is completed by stick-breaking; the "cuts" S_i are independent but not necessarily uniform or even identically distributed. The S_i take values in $(0, 1)$ and have arbitrary distributions with two restrictions: S_i falls into any open interval with positive probability and $\Sigma E(S_i) = \infty$. The cuts are used to distribute mass inductively as follows: suppose mass $m = \sum_{j=1}^{N} \theta_j < 1$ has been assigned and mass $1 - m$ remains; now mass $\theta_{N+1} = (1 - m)S_{N+1}$ is assigned to $N + 1$ and mass

$$1 - m - (1 - m)S_{N+1} = (1 - m)(1 - S_{N+1})$$

is left for the next move, which is carried out using S_{N+2}, and so on.

The motivation is as follows. A Bayesian may have a reasonably clear opinion about θ_i for some i's. However, it seems unlikely that such an opinion can be carefully quantified for all i. Freedman's extension of stick-breaking allows a Bayesian to approximate any prior by one consistent at all parameter values; indeed, any prior can be approximated (weak-star) by specifying the distribution of a finite number of θ_i.

Early users of "stick-breaking" were Banach (1964), Kahane and Salem (1958), and Eberlein (1962). Kahane and Salem studied sums $S = \sum_{i=1}^{\infty} r_i X_i$; the X_i are iid, taking the values 0 or 1 with probability $\frac{1}{2}$ each; the r_i are nonnegative and sum to 1. When is the distribution of S absolutely continuous? Kahane and Salem prove that for "almost all" sequences $(r_1, r_2, \ldots)$ the law of S is absolutely continuous with an L^2 density; "almost all" is relative to a stick-breaking measure. Banach and Eberlein were interested in a natural integral over the function spaces l_2 and l_1 and used stick-breaking as a key ingredient.

Dirichlet priors are tail-free; the S_i having certain beta distributions. Usually, Dirichlet priors are parametrized in terms of a finite measure α on the observation space $\mathcal{X}$, which is for the moment general. The Dirichlet prior $D(\alpha)$ with base measure α can be characterized as follows. Partition the observation space $\mathcal{X}$ into a finite number of sets $A_1, A_2, \ldots, A_k$. Consider a probability θ on the observation space, selected at random from $D(\alpha)$. Then $\theta(A_1), \ldots, \theta(A_k)$ are random variables with respect to $D(\alpha)$. These have a Dirichlet distribution on the k-simplex, with parameter vector $(\alpha(A_1), \ldots, \alpha(A_k))$. More concretely, these are distributed like

$$U_1/S, \ldots, U_k/S,$$

where $S = U_1 + \cdots + U_k$, the U_i being independent gamma variables with shape parameter $\alpha(A_i)$. When the observation space is the integers, α is specified by a countable collection of numbers $\alpha_1, \alpha_2, \ldots$. Let $\|\alpha\| = \Sigma\alpha_i$. By assumption, $\|\alpha\| < \infty$. For a Dirichlet prior with parameter measure α on the positive

integers, the cut S_i has a beta distribution with the two parameters $\alpha_1 + \cdots + \alpha_i$ and $\|\alpha\| - (\alpha_1 + \cdots + \alpha_i)$.

Dirichlet priors have been suggested for use in a wide variety of problems. Also, some writers have considered using mixtures of Dirichlet priors; see Antoniak (1974) or Dalal and Hall (1980, 1983). It is natural to ask whether mixtures of Dirichlet priors are consistent. The first step is easy; any mixture of a finite number of consistent priors is consistent. However, in Freedman and Diaconis (1983), we showed that a countable mixture of Dirichlet priors can be inconsistent. A starting point of our construction is an example showing that a mixture of a Dirichlet prior and a point mass at a certain long-tailed probability is inconsistent. Such priors are similar to the ones suggested by Jeffreys (1967) for Bayesian hypothesis testing. On the positive side, we showed that if the mass of the parameter measures are bounded, then the mixture is consistent.

The inconsistent priors in Freedman (1963) were constructed with malice aforethought. Bayesian reaction seems to be, "Oh, nobody would ever use a prior like that." See, for example, the remarks in Box, Leonard, and Wu (1983, page xi). That a Jeffreys-style prior is inconsistent should therefore be of interest. Moreover, in Diaconis and Freedman (1986), we give examples of priors suggested by practicing Bayesians, which turn out to be inconsistent. That paper is fairly technical and the following heuristic discussion may be helpful.

The location problem. Consider estimating an unknown location parameter θ with squared error as loss. The observations are modelled as

$$X_i = \theta + \varepsilon_i, \qquad i = 1, 2, \ldots .$$

The ε_i are independent disturbance terms with a common distribution G. If G has known density g and a prior μ is put on θ, then the Bayes estimate is the mean of the posterior distribution:

$$\hat{\theta} = \frac{\int \theta \prod g(x_i - \theta)\, \mu(d\theta)}{\int \prod g(x_i - \theta)\, \mu(d\theta)}. \tag{1.1}$$

If $\mu(d\theta)$ is taken as Lebesgue measure, then this becomes the Pitman estimator. If the density g in (1.1) is unknown, it can be estimated from the data. Often, g is assumed to belong to some parametric family. Fraser (1976), Box and Tiao (1973, Chapters 3 and 4), and Johns (1979) all propose estimators of that general type, with a prior distribution on the parameters of the family. Such estimators can be inconsistent; the argument is like that in Freedman and Diaconis (1982b) or Diaconis and Freedman (1986) as outlined below.

A nonparametric approach to estimating G is also natural. This involves putting a prior on θ and a prior on the law G of ε_i. Dalal (1979a, b) has suggested using a Dirichlet prior for G. We will now show that for a Dirichlet prior with a Cauchy base measure α, the Bayes estimates are inconsistent.

To avoid identifiability problems, we will assume that the law G of ε is symmetric. To put a prior on symmetric G's, we symmetrize the Dirichlet as follows: if G is a distribution function for a random variable X, let G^- be the distribution function of $-X$ and let $\overline{G} = \frac{1}{2}(G + G^-)$, so $\overline{G}$ is symmetric. Let $\overline{D}_\alpha$

be the law of $\overline{G}$, where G has a Dirichlet distribution with base measure α on $\mathbb{R}$. This is a symmetrized Dirichlet. The construction was first suggested by Dalal (1979a, b). For further discussion see Hannum and Hollander (1983). We assume that the base measure α has density α' and write $\|\alpha\| = \alpha(\mathbb{R})$. We make θ and G independent; our prior for θ has density f on $\mathbb{R}$ while G is chosen from $\overline{D}_\alpha$.

The posterior distribution of θ is computed in Lemma 3.1 of Diaconis and Freedman (1986). For simplicity, we only discuss the Bayes estimate here.

THEOREM 1. *Suppose that $X_1, \ldots, X_n$ are all distinct. Let $\theta_{ij} = \frac{1}{2}(X_i + X_j)$ and suppose θ_{ij} are distinct. The Bayes estimate θ is*

$$\hat{\theta}_n = \left[\int \theta\omega(\theta)\, d\theta + \sum_{i<j} \theta_{ij}\omega_{ij}\right] \Bigg/ \left[\int \omega(\theta)\, d\theta + \sum_{i<j} \omega_{ij}\right]$$

with

$$\omega(\theta) = \|\alpha\| f(\theta) \prod_{k=1}^{n} \alpha'(X_k - \theta),$$

$$\omega_{ij} = \tfrac{1}{2}\left[f(\theta_{ij})/\alpha'(\delta_{ij})\right] \prod_{k=1}^{n} \alpha'(X_k - \theta_{ij}),$$

$$\delta_{ij} = \tfrac{1}{2}(X_i - X_j).$$

For the next result, take f to be standard normal and α to be Cauchy. We suppose that in fact the ε_i have a density h which is symmetric about 0, with a strict maximum at 0; further, h is infinitely differentiable and vanishes off a compact interval.

THEOREM 2. *For some h, the Bayes estimate is inconsistent. Indeed, for large n, there is probability near $\frac{1}{2}$ that $\hat{\theta}_n$ is close to γ and probability near $\frac{1}{2}$ that $\hat{\theta}_n$ is close to $-\gamma$, where $\gamma \neq 0$ depends on h.*

REMARKS. (1) Theorem 2 is valid when α is any t-distribution.

(2) When sampling from a continuous density, X_i and θ_{ij} will be distinct with probability 1.

(3) The Bayes estimate $\hat{\theta}$ is a convex combination of two other estimates:

$$\hat{\theta}_1 = \int \theta\omega(\theta)\, d\theta \Big/ \int \omega(\theta)\, d\theta$$

and

$$\hat{\theta}_2 = \sum_{i<j} \theta_{ij}\omega_{ij} \Big/ \sum_{i<j} \omega_{ij}.$$

The first is a Bayes estimate, as at (1.1), when sampling from known density α'. The second is a weighted average of the θ_{ij}, which estimate θ from the pair X_i and X_j. If α' is one of the standard unimodal densities such as normal or Cauchy, the weight ω_{ij} is relatively large when θ_{ij} is in the center of the X_k's. It turns out

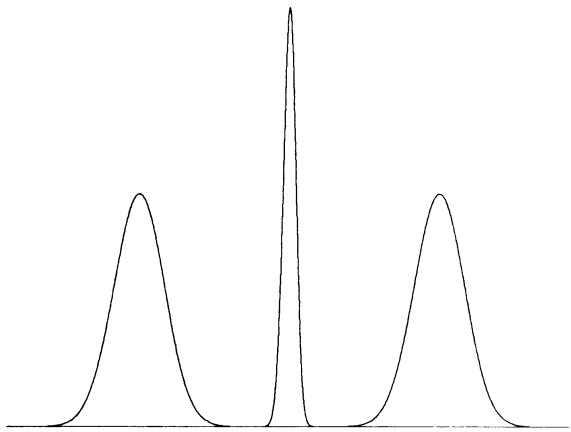

FIG. 1. *The counterexample density h.*

that $\hat{\theta}_1$ dominates $\hat{\theta}_2$ as $n \to \infty$ for some h; for others $\hat{\theta}_2$ will dominate. Here, we focus on the first case.

(4) The h constructed in Theorem 2 is trimodal, as in Figure 1. It has a unique maximum at 0 but the two other modes matter. If desired, h can be chosen strictly positive on the interior of its interval of support.

(5) Any of the classical estimators, such as the mean or the median, will be consistent in this situation, so the Bayes estimates do worse than available frequentist procedures.

(6) In Diaconis and Freedman (1986), we argue that the Bayes estimate is consistent for any strongly unimodal density h: strong unimodality means that h increases up to its unique maximum and then decreases. Further, we show that if $\log \alpha'$ is convex, then the Bayes estimate is consistent for any symmetric h.

(7) Doss (1983a, b, 1984) has carried out similar computations for neutral to the right and tail-free priors. He has also introduced some other methods of symmetrizing. Very roughly, his results parallel ours; it does not seem possible to find a consistent prior concentrated in a neighborhood of the normal location family.

(8) A subjective rationale for mixtures of location models is discussed in Freedman and Diaconis (1982a).

We now sketch the argument for inconsistency. As explained in Remark 3, the Bayes estimate is essentially

$$\hat{\theta}_1 = \int \theta \bar{\omega}(\theta)\, d\theta \Big/ \int \bar{\omega}(\theta)\, d\theta$$

with $\bar{\omega}(\theta) = \prod_{i=1}^{n} 1/[1 + (X_i - \theta)^2]$, the X_i being the data. We have

$$\bar{\omega}(\theta) = \exp\left\{ - \sum_{i=1}^{n} \log\left[1 + (X_i - \theta)^2\right] \right\} \doteq \exp\left\{ -n E\left[\log\left(1 + (X_1 - \theta)^2\right)\right] \right\}.$$

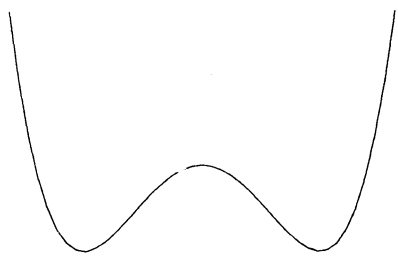

FIG. 2. *The function H.*

In the approximation, the sum has been replaced by its mean under the true sampling distribution. Now consider

$$H(\theta) = E\{\log[1 + (X - \theta)^2]\}.$$

If X takes only the two values $\pm a$ where $a > 1$, then H has a local maximum at 0, a global minimum at $\pm\gamma$, and tends to infinity at $\pm\infty$. Of course, H is symmetric. See Figure 2.

Now we estimate the integrals in the numerator and denominator of $\hat{\theta}_1$ by Laplace's method. As a function, $\bar{\omega}(\theta)$ is close to $\exp[-nH(\theta)]$, so only θ's near $\pm\gamma$ matter, and as n tends to infinity, $\hat{\theta}_1$ oscillates between $+\gamma$ and $-\gamma$. The two-point distribution of X can be smoothed out to the density h shown in Figure 1.

This argument is like the one used in Freedman and Diaconis (1982b) to show inconsistency of M-estimators. It has been objected that the counterexample density h is not in the "support" of the Dirichlet, since the Dirichlet chooses a discrete measure with probability 1. A technical response is that the Dirichlet assigns positive mass to every open set of probabilities, so h is in the support—the smallest closed set of full prior mass. A broader response is that the Dirichlet assigns zero mass to any particular probability. After the fact, therefore, any troublesome probability can be differentiated in some qualitative way from a Borel set of full prior mass.

The foregoing discussion has all been asymptotic. What are the implications with finite samples? Consider an estimation problem with a large amount of data —and a large number of parameters. A Bayes rule will do well for most parameter values, "most" being defined relative to the prior measure itself. On the other hand, a set of parameters which is large as judged by one measure may be quite small as judged by another, so Bayes rules may be quite unsatisfactory for a practical frequentist. Bayesians usually argue that the data will swamp the prior—but this may not happen in high-dimensional inference problems, or it may occur, but very slowly. Our view is that the oscillation at infinity will even show up with large, finite samples—in high-dimensional problems.

On the other hand, even the most dedicated subjectivist will usually not insist on quantifying all the precise details of a prior opinion in a high-dimensional situation. Our results indicate that small changes in the details of a prior can lead to Bayes rules with much better operating characteristics.

Quantifying these ideas is hard and that is why we give asymptotic results. We hope to study the finite-sample problem elsewhere, and in particular, we hope to quantify the extent to which small changes in the prior can make small changes in the posterior but big improvements in rates of convergence. The derivative of the posterior with respect to the prior is a relevant concept and will be discussed in Section 3.

2. Other connections between Bayesians and frequentists. Frequentists often discuss Bayes rules for the following reasons. First, the complete class theorem, as in Wald (1950), Le Cam (1955), Stein (1955), Sacks (1963), or Brown (1981), implies that all admissible procedures are approximately Bayes. Similarly, all minimax procedures are approximately Bayes. So not much is lost by confining attention to Bayes procedures. Thus, Bayes procedures are convenient, tractable, and close to optimal. For instance, in some problems confidence sets are difficult to obtain. Welch and Peers (1963), Hinkley (1980), and Stein (1981) suggest using regions of high posterior mass, the prior being chosen so that Bayesian and frequentist coverage probabilities agree to several terms in an asymptotic expansion. Similarly, Bayesian techniques are used as a way of eliminating nuisance parameters. Berk (1970) gives examples of this in sequential analysis.

The common thread is a kind of pragmatic use of Bayes methods by frequentists: theory and convenience suggest the use of Bayes rules. As long as a Bayes rule is to be used, one may as well work with a prior that concentrates on a plausible part of the parameter space.

Second, Bayesian techniques can be helpful in proving frequentist theorems. For example, consider estimating the mean of a univariate normal with known scale and squared error loss. Any admissible estimator must be an analytic function of the observations. The only available proof uses complete class theorems of Sacks (1963) and Stein (1955) to represent the estimator as a formal Bayes rule. It is easy to show that formal Bayes rules are suitably smooth. Another example of this sort is in Matthes and Truax (1967).

Naturally, frequentists have been interested in frequentist properties of Bayes procedures. Theorems dating back to Laplace (1774) show that the posterior distribution can be approximated by the distribution of the maximum likelihood estimator. Modern versions of these theorems can be found in Bernstein (1934), Von Mises (1964), Johnson (1967, 1970), Le Cam (1982), or Ghosh et al. (1982). Bayesians use these results to show that standard frequentist procedures are nearly Bayes. See Lindley (1965).

There are many other points of contact between the frequentist and Bayesian schools. However, there is often some incompatibility between Bayes rules and frequentist desiderata. For example, Blackwell and Girschick (1954) followed by Blackwell and Bickel (1967) and Noorbaloochi and Meedan (1983) have shown

that there are essentially no unbiased Bayes procedures. Other points of difference are emphasized in survey articles by Pratt (1965), Cornfield (1969), Lindley (1972), and Neyman (1977). Savage (1972) gives a review.

Sufficiency gives a point of technical contact. Kolmogorov (1942) introduced the notion of Bayesian sufficiency—a statistic is Bayes sufficient if for any prior the posterior only depends on the data through the sufficient statistic. For smooth, finite-dimensional problems, Bayesian sufficiency is the same as the frequentist concept. Recently, Blackwell and Ramamoorthi (1982) give an infinite-dimensional example where the two notions disagree.

Points of agreement arise in Bayesian discussions of robustness as in Berger (1984) or Kadane and Chuang (1978). Here frequentist properties of Bayes procedures are used as a means of protection from naive specification of the prior. A similar compromise was suggested by Hodges and Lehmann (1952).

3. Bayesian interpretations of consistency. It is useful to separate Bayesians into two groups: we will call them "classical" and "subjectivist." Classical Bayesians, like Laplace or Bayes himself, seemed to believe there is a true but unknown parameter which is to be estimated from data. This parameter is part of an objective probability model for the data. Prior opinion about the parameter is expressed as a probability distribution. Subjective Bayesians like de Finetti and Savage reject such ideas; for them, probabilities represent degrees of belief and there are no objective probability models. (Freedman used to be a classical Bayesian, while Diaconis is a subjectivist.)

Consistency properties of Bayes rules are clearly of interest to classical Bayesians; as data comes in, the posterior should converge to point mass at the true parameter. We will now argue that frequency properties of Bayes rules are also of interest to subjectivists. The first reason has to do with "intersubjective agreement." In some circumstances, Bayesians learn from experience, so opinions based on very different priors will merge as data accumulates; the data swamps the prior. We will now argue that consistency is equivalent to merging of intersubjective opinions under certain conditions.

A general result of this type was provided by Blackwell and Dubins (1962). To state their result, assume that P and Q are probabilities governing a process $(X_1, X_2, \ldots)$. Let P_n and Q_n be regular conditional probabilities for the future $(X_{n+1}, X_{n+2}, \ldots)$ given the past $(X_1, X_2, \ldots, X_n)$. We think of P and Q as the priors of two Bayesians, and assume P and Q are mutually absolutely continuous: $P \equiv Q$, the Bayesians agree on what is possible or impossible. Blackwell and Dubins show that almost surely, P_n and Q_n merge in variation distance.

Reverting to the context of Section 1, assume that $\theta \to Q_\theta$ is a homeomorphism and consider two Bayesians with priors μ and ν, respectively. Now $\mu \equiv \nu$ if and only if $P_\mu \equiv P_\nu$, and $\mu \perp \nu$ if and only if $P_\mu \perp P_\nu$. So the Blackwell–Dubins result does not apply when $\mu \perp \nu$, as is usually the case with Dirichlet priors. However, even if merging in variation distance does not happen, "weak-star merging" is still a possibility. We say two sequences $\{\alpha_n\}$, $\{\beta_n\}$ of probabilities *merge weak-star* if and only if α_n and β_n become indistinguishable from the point of view of integrating bounded continuous functions: more precisely,

$R(\alpha_n, \beta_n) \to 0$ for every weak-star continuous function R defined on pairs of probabilities satisfying $R(\alpha, \alpha) = 0$. As an example, R might be a metric for weak-star convergence; or $R(\alpha, \beta)$ might be $\int f\, d\alpha - \int f\, d\beta$.

We show that the version μ_n is consistent if and only if P_μ and P_ν merge weak-star for any ν, in the following sense. Recall that P_μ is the joint distribution of θ and $X_1, X_2, \ldots$. Let $P_{\mu n}$ be the law of $X_{n+1}, X_{n+2}, \ldots$, given $X_1, \ldots, X_n$, determined by the formula

$$P_{\mu n} = \int Q_\theta^\infty \mu_n(d\theta).$$

A Bayesian with prior μ holds the opinion $P_{\mu n}$ about the future $X_{n+1}, X_{n+2}, \ldots$, after seeing the data $X_1, \ldots, X_n$.

THEOREM 3. *Suppose $\theta \to P_\theta$ is continuous, one-to-one, and continuously invertible. Suppose the version μ_n is consistent. Let ν be any other prior. As $n \to \infty$, $P_{\mu n}$ and $P_{\nu n}$ merge weak-star along P_ν-almost all sequences. Conversely, if $P_{\mu n}$ and $P_{\nu n}$ merge for all ν, then μ_n is consistent.*

Informally, the prior μ is consistent if and only if any other subjectivist with prior ν is sure that P_μ and P_ν will merge on the future, as more and more data comes in. Theorem 3 is proved in Appendix A. Similar issues were considered by Lockett (1971).

This completes our discussion of "intersubjective agreement." We turn to our second reason for thinking that frequency properties should interest subjectivists as a way of specifying priors. The idea is simple: after specifying a prior distribution, generate imaginary data sequences, compute the posterior, and consider whether the posterior would be an adequate representation of the updated prior. This is quite close to the checks for coherence proposed by de Finetti and Savage; see de Finetti (1974, pages 229–246) for examples and Savage (1971) for a review. This use of fictitious samples was suggested by the "device of imaginary results;" Good (1950, page 35) uses this device as a method of roughly quantifying a prior in difficult situations. We call it the "what if" method: what if the data came out that way?

Take the prior suggested by Dalal (1979b) for use in symmetric location problems as discussed in Section 1. If data were generated from the density h of Theorem 2, the posterior would oscillate and never converge to the center of symmetry. The "what if" method strongly suggests modifying the prior. We give further examples and discussion in Diaconis and Freedman (1983).

The "what if" method suggests various interesting mathematical problems. For example, to what extent do the Bayes estimates determine the prior? Diaconis and Ylvisaker (1979) show that the posterior mean of the natural parameter in an exponential family determines the prior under suitable regularity conditions. Diaconis and Ylvisaker (1985) give counterexamples in location problems.

For a fixed observation x, Bayes theorem gives the relationship between the posterior and the prior. This defines a map from measures to measures. In the

next theorem we calculate the derivative of this map and the norm of the derivative. This helps to identify data sets x where small changes in the prior cause large changes in the posterior. These may be the most informative x's for the "what if" method. (Of course, they may also be unlikely x's.)

In the theorem, probabilities are considered as a subset of all signed measures, with distance defined by the variation norm: the distance between μ and ν is the total mass of the signed measure $\mu - \nu$. Equivalently,

$$\|\mu - \nu\| = \int \left| \frac{d\mu}{d\sigma} - \frac{d\nu}{d\sigma} \right| d\sigma,$$

where σ is any dominating σ-finite measure, e.g., $\sigma = \mu + \nu$. With this norm, the measures form a Banach space.

Suppose $\{Q_\theta\colon \theta \in \Theta\}$ is a "dominated family;" there exists a σ-finite measure λ on X with all the Q_θ's absolutely continuous with respect to λ. Let $f(x|\theta)$ be the density of Q_θ with respect to λ: we write $dQ_\theta = f(\cdot|\theta)\, d\lambda$. We will assume that f is jointly measurable and $\sup_\theta f(x|\theta) < \infty$ for every x. For a probability μ on Θ, define

$$T(\mu) = N(\mu)/D(\mu),$$

where

$$N(\mu)\,(d\theta) = f(x|\theta)\mu(d\theta) \quad \text{and} \quad D(\mu) = \int_\Theta f(x|\theta)\mu(d\theta).$$

Thus, $N(\mu)$ is a measure and $D(\mu)$ is a number; $T(\mu)$ is the posterior distribution. We confine ourselves to the x with $D(\mu) > 0$, a set of P_μ-measure 1. The dependence on x is suppressed for the moment. The map $\mu \to T(\mu)$ takes priors into posteriors. It has a derivative $\dot{T}_\mu$ at μ. This $\dot{T}_\mu$ is a linear map on signed measures such that

$$T(\mu + \delta) = T(\mu) + \dot{T}_\mu(\delta) + o(\|\delta\|) \quad \text{as } \|\delta\| \to 0,$$

where δ is a signed measure with signed mass 0. The norm of $\dot{T}_\mu$ will be used as a measure of the effect of a small change in μ. This is defined by

$$\|\dot{T}_\mu\| = \sup_{\|\delta\|=1} \|\dot{T}_\mu(\delta)\|.$$

Let $\sup_0 f(x|\theta) = \sup_\theta\{ f(x|\theta)\colon \mu\{\theta\} = 0\}$, the sup over an empty set being 0 by convention. Under the given conditions:

THEOREM 4. (a) $\dot{T}_\mu(\delta) = [N(\delta)/D(\mu)] - [N(\mu)D(\delta)/D(\mu)^2]$.
(b) $\|\dot{T}_\mu\| \le \sup_\theta f(x|\theta)/D(\mu)$.
(c) $\|\dot{T}_\mu\| \ge \sup_0 f(x|\theta)/D(\mu)$.

Theorem 4 is proved in Appendix B. In (a), the quantity $N(\mu)$ is a signed measure, normalized by $D(\mu)$. Likewise, $N(\mu)$ is a signed measure, normalized by the factor $D(\delta)/D(\mu)^2$. Thus, $\dot{T}_\mu(\delta)$ is a signed measure. As is easily seen, this

measure has signed mass 0. For many priors the upper and lower bounds of (b) and (c) coincide. Then conclusion (b) has a simple interpretation in terms of likelihood ratios: the x's where the posterior is most sensitive to small changes in the prior are the x's which have high ratio of objectivist likelihood to subjectivist likelihood. These x's are the ones where the "what if" method will be most informative: at such x, small changes in the prior can make big changes in the posterior.

As an example, consider a normal location problem with known variance σ^2. Without loss, suppose there is only one observation. Let μ be a normal prior for the location parameter θ. Suppose that μ has mean μ_0 and variance σ_0^2. Then $D(\mu)$ at x turns out to be the normal density at x with mean μ_0 and variance $\sigma_0^2 + \sigma^2$. The norm computed in (b) is

$$\left[\left(\sigma^2 + \sigma_0^2\right)/\sigma^2\right]^{1/2} \exp\left\{\tfrac{1}{2}(x - \mu_0)^2 \Big/ \left(\sigma^2 + \sigma_0^2\right)\right\}.$$

This is large when x is far from μ_0. We have carried out similar computations for the other standard one-dimensional exponential families with conjugate priors. The results are similar: values of x far from the mean of the prior lead to large norms.

On the other hand, for examples of the type considered in Freedman and Diaconis (1983), the posterior concentrates at some distance from the maximum likelihood estimate, so $\int f\, d\mu$ will be many orders of magnitude smaller than $\max f$ for most data sets, relative to the true sampling distribution. Such data sets will have high leverage, and a posterior, which may seem on reflection to be unsatisfactory, so the "what if" method might prompt revision of an inconsistent prior. This completes our discussion of the "what if" method and with it Bayesian defense of frequentist analysis.

4. Conclusion. We return now to the big picture. There is a probability model for data and some of the parameters are to be estimated. A statistician who really has a sharp prior probability distribution for these parameters should use it, according to Bayes theorem; inconsistency on a null set of *a priori* probability zero is an irrelevant nuisance. On this point, there seems to be general agreement in the statistical community.

Often, a statistician has prior information about a problem (say as to the rough order of magnitude of a key parameter), but does not really have a sharply defined prior probability distribution. Many different distributions would have the right qualitative features and a Bayesian typically chooses one on the basis of mathematical convenience. In smooth, low-dimensional problems, this ought to help, and anyway cannot lead to disaster, because the data will swamp the details of the prior.

Unfortunately, in high-dimensional problems, arbitrary details of the prior can really matter; indeed, the prior can swamp the data, no matter how much data you have. That is what our examples suggest, and that is why we advise against the mechanical use of Bayesian nonparametric techniques.

APPENDIX A

Merging of posteriors and the weak-star topology. We suppose Θ is a Borel subset of a complete separable metric space—a Borel set for short. Next, for each $\theta \in \Theta$, we have a probability Q_θ on the Borel subsets of another Borel set $\mathscr{X}$. Thus, Θ is the parameter space and $\mathscr{X}$ the observation space. The map $\theta \to Q_\theta$ is assumed to be 1–1 and Borel. For a prior probability μ on Θ, we write P_μ for the probability on $\Theta \times \mathscr{X}^\infty$ defined by

$$P_\mu(A \times B) = \int_A Q_\theta^\infty(B)\,\mu(d\theta),$$

where A is Borel in Θ and B is Borel in $\mathscr{X}^\infty$. As is easily verified, if H is a Borel subset of $\Theta \times \mathscr{X}^\infty$, then

$$P_\mu(H) = \int_\Theta (\delta_\theta \times Q_\theta^\infty)(H)\,\mu(d\theta). \tag{A.1}$$

As usual, δ_θ is point mass at θ.

Fix a version $\mu_n = \mu_n(d\phi|x_1, \ldots, x_n)$ of the P_μ-law of θ given $X_1 = x_1, \ldots, X_n = x_n$. Here, as elsewhere, θ denotes the coordinate map

$$(\theta, x_1, x_2, \ldots) \to \theta.$$

The posterior μ_n may be considered as a function on $\Theta \times \mathscr{X}^\infty$, or just $\mathscr{X}^\infty$, or even $\mathscr{X}^n$, according to convenience.

Consistency of (θ, μ_n) depends not only on μ, but also on the choice of the version μ_n, which is well defined only a.e. An artificial example may clarify the point.

EXAMPLE A.1. Let $\Theta = [0,1]$, $\mathscr{X} = [0,1]$, $Q_\theta = \delta_\theta$, and let μ be Lebesgue measure on Θ. Let

$$\mu_n\{d\phi|x_1, \ldots, x_n\} = \delta_{x_1}$$

unless $x_1 = 0$ in which case

$$\mu_n\{d\phi|x_1, \ldots, x_n\} = \delta_{1/2}.$$

This μ_n is a (stupid) version of the law of θ given $X_1, \ldots, X_n$ because

$$P_\mu\{X_1 = 0, \ldots, X_n = 0\} = 0.$$

For this version, μ_n is consistent except at $\theta = 0$ when $\mu_n \to \delta_{1/2}$.

Bayes estimates of an unknown probability on the line provide a less artificial example, where $\{Q_\theta\}$ is not dominated by a σ-finite measure. Then, the posterior can be changed on a set of measure zero, so consistency is determined for μ almost all θ but not for all θ.

The next topic is the weak-star topology for probability measures on Θ. Recall that Θ is a separable metric space by assumption, but Θ need not be complete, or compact, or locally compact: for example, Θ might be the irrational numbers in the unit interval. Let $\mathscr{C}$ be the set of bounded, continuous functions on Θ. If α_n

and α are probabilities on Θ, then $\alpha_n \to \alpha$ weak-star if and only if $\int f\,d\alpha_n \to \int f\,d\alpha$ for all $f \in \mathscr{C}$.

A technical problem is that $\mathscr{C}$ will not usually be separable. However, it is enough to consider only the uniformly continuous $f \in \mathscr{C}$: to be specific, let Θ_0 be a countable dense subset of Θ. For each $\theta_0 \in \Theta_0$ and nonnegative rationals $r < s$ and arbitrary rationals u, v, define a function $f = f_{r,s,u,v}$ on Θ as follows:

$$f(\theta) = u \quad \text{for } \theta \text{ with } \rho(\theta_0, \theta) \le r$$
$$= v \quad \text{for } \theta \text{ with } \rho(\theta_0, \theta) \ge s$$

while

$$f(\theta) = u\frac{s - \rho(\theta_0, \theta)}{s - r} + v\frac{\rho(\theta_0, \theta) - r}{s - r}$$

for θ with $r \le \rho(\theta_0, \theta) \le s$. Here, ρ is the given metric on Θ. Clearly, f is bounded and uniformly continuous. Let $\mathscr{C}_0 = \{f\}$, a countable collection. Let $\mathscr{C}_\vee = \{f_1 \vee \cdots \vee f_k$: for $f_i \in \mathscr{C}_0$ and $k = 1, 2, \ldots\}$. Let $\mathscr{C}_\wedge = \{f_1 \wedge \cdots \wedge f_k$: for $f_i \in \mathscr{C}_0$ and $k = 1, 2, \ldots\}$. Here $\vee$ and $\wedge$ are pointwise max and min, respectively. Now $\mathscr{C}_\vee$ and $\mathscr{C}_\wedge$ are countable subsets of $\mathscr{C}$, and pointwise dense, as follows.

LEMMA A.1. *If $f \in \mathscr{C}$, there is a sequence $f_k \in \mathscr{C}_\vee$ with $f_k \uparrow f$ pointwise; and another sequence $g_k \in \mathscr{C}_\wedge$ with $g_k \downarrow \mathscr{C}$ pointwise.*

PROOF. Use the method of exhaustion: for instance, if $f \in \mathscr{C}$, then $f = \inf\{g$: $g \in \mathscr{C}_0$ and $g \ge f\}$. □

COROLLARY A.1. *If $\int f\,d\alpha_n \to \int f\,d\alpha$ for all $f \in \mathscr{C}_\vee \cup \mathscr{C}_\wedge$, then $\alpha_n \to \alpha$ weak-star.*

We can now prove Doob's theorem in our context.

COROLLARY A.2. *$\mu_n \to \delta_\theta$ weak-star, P_μ-almost surely.*

PROOF. Fix an $f \in \mathscr{C}_\vee \cup \mathscr{C}_\wedge$. By Corollary A.1 it is enough to prove that $\int f\,d\mu_n \to f(\theta)$ a.e. P_μ. But $\int f\,d\mu_n$ is a version of $E\{f(\theta)|X_1 \ldots X_n\}$, so $\int f\,d\mu_n \to E\{f(\theta)|X_1, X_2, \ldots\}$ a.e. by martingales. The last step is to prove identifiability: that θ can be computed measurably from all the Xs, a.e. P_μ.

The identifiability argument will only be sketched: it is at this point that "1–1-ness" matters. Since $\theta \to Q_\theta$ is 1–1 and Borel, so is the inverse: see Kuratowski (1958, Section 35.5). So we only need to compute Q_θ from $X_1, X_2, \ldots$. Using Corollary A.1 again, we only need to compute $\int g\,dQ_\theta$ for $g \in \mathscr{C}_\wedge \cup \mathscr{C}_\vee$. But

$$\int g\,dQ_\theta = \lim_{n\to\infty} \frac{1}{n}[g(X_1) + \cdots + g(X_n)] \tag{A.2}$$

a.e. by the law of large numbers.

In a bit more detail, let H_g be the subset of $\Theta \times \mathcal{X}^\infty$ where (A.2) holds. Then H_g is Borel, and $\delta_\theta \times Q_\theta^\infty(H_g) = 1$ for all θ by the law of large numbers. So $P_\mu(H_g) = 1$ by (A.1). □

The following lemma and corollary are used later.

LEMMA A.2. *A Borel set is homeomorphic to a dense Borel subset of a compact metric set.*

PROOF. This standard fact can be proved by appealing to Urysohn's lemma (Kuratowski, 1958, page 119) and then (Kuratowski, 1958, page 397). □

COROLLARY A.3. $\theta \to Q_\theta^\infty$ *is weak-star continuous.*

PROOF. Let f be bounded continuous on $\mathcal{X}^\infty$ and

$$g(\theta) = \int_{\mathcal{X}^\infty} f \, dQ_\theta^\infty.$$

We have to show that g is continuous.

CASE 1. $\mathcal{X}$ is compact. Then f can be approximated by finite linear combinations of functions $\prod_{i=1}^k f_i(x_i)$ with f_i bounded continuous on $\mathcal{X}$.

THE GENERAL CASE. Embed $\mathcal{X}$ as a dense Borel subset of the compact metric space $(\bar{\mathcal{X}}, \rho)$ and metrize $\mathcal{X}$ by ρ: see Lemma A.2. The bounded, ρ-uniformly continuous functions on $\mathcal{X}$ are those which extend to continuous functions on $\bar{\mathcal{X}}$. Now use Corollary A.1 to reduce the general case to Case 1, the point being that a function g which is both upper and lower semicontinuous is continuous. □

The next topic is "weak-star merging." Let α_n and β_n be probabilities on Θ; the sequences $\{\alpha_n\}$ and $\{\beta_n\}$ *merge weak-star* if and only if $R(\alpha_n, \beta_n) \to 0$ for all continuous functions R which satisfy $R(\alpha, \alpha) = 0$. Some possible choices for R:

$$R(\alpha, \beta) = \int f \, d\alpha - \int f \, d\beta \quad \text{for a fixed bounded continuous } f,$$

(A.3) $$R(\alpha, \beta) = \lambda(\alpha, \beta) \quad \text{with } \lambda \text{ a metric for the weak-star topology,}$$

$$R(\alpha, \beta) = T(\alpha) - T(\beta) \quad \text{for a fixed bounded continuous function } T.$$

For the next theorem, we also need

$$P_{\mu n} = \int_\Theta Q_\theta^\infty \mu_n(d\theta).$$

This is the "predictive" distribution of $X_{n+1}, X_{n+2}, \ldots$, given $X_1, \ldots, X_n$. Of course, $P_{\mu n} = P_{\mu n}(\cdot | x_1 \ldots x_n)$ is a probability on $\mathcal{X}^\infty$, depending on the first n data points.

We will prove somewhat more than Theorem 3. To state the result, let $G \subset \Theta \times \mathcal{X}^\infty$ be the set of $(\theta, x_1, x_2, \ldots)$ such that $\mu_n(d\phi|x_1 \ldots x_n) \to \delta_\theta$ weak-star. Notice that G depends on μ_n and μ_n is consistent if and only if $(\delta_\theta \times Q_\theta^\infty)(G) = 1$ for all θ.

THEOREM A.1. *Let $\theta \to Q_\theta$ be 1–1 and Borel. Fix a prior μ on Θ and a version μ_n of the posterior. The following three conditions are equivalent.*

(i) *μ_n is consistent.*

(ii) *$P_\nu(G) = 1$ for all probabilities ν on Θ.*

(iii) *$\mu_n(d\phi|x_1 \ldots x_n)$ and $\nu_n\{d\phi|x_1 \ldots x_n\}$ merge weak-star as $n \to \infty$, with P_ν probability one, for all probabilities ν on Θ.*

Suppose that $\theta \to Q_\theta$ is continuous and has a continuous inverse. Then (i) *is also equivalent to*

(iv) *$P_{\mu n}$ and $P_{\nu n}$ merge weak-star as $n \to \infty$, with P_ν probability one, for all probabilities ν on Θ.*

PROOF. Fix ν. We will prove that (i) $\Rightarrow$ (ii) $\Rightarrow$ (iii). The first implication is trivial: from (A.1),

$$P_\nu(G) = \int (\delta_\theta \times Q_\theta^\infty)(G)\nu(d\theta) \tag{A.4}$$

and $(\delta_\theta \times Q_\theta^\infty)(G) = 1$ for all θ from the definition of consistency. For the next implication, suppose (ii) for ν. We will argue that a.e. P_ν,

$$\mu_n \to \delta_\theta, \tag{A.5}$$

$$\nu_n \to \delta_\theta. \tag{A.6}$$

The notation may be a bit confusing: θ is being used for the coordinate function $(\theta, x_1, x_2, \ldots) \to \theta$ and its value θ. It would follow from (A.5–A.6) that $R(\mu_n, \nu_n) \to 0$. Now (A.5) holds on G, which has P_ν-probability 1, by assumption. And (A.6) holds a.e. P_ν by Doob's theorem, Corollary A.2 above. Thus, (i) $\Rightarrow$ (ii) $\Rightarrow$ (iii). Clearly, (iii) $\Rightarrow$ (i): take ν to be point mass and R as in (A.3).

Now assume that $\theta \to Q_\theta$ is 1–1 and continuous. We will argue (i) $\Rightarrow$ (iv). Indeed, suppose (i). We will show that a.e. P_ν,

$$P_{\mu n} \to Q_\theta^\infty, \tag{A.7}$$

$$P_{\nu n} \to Q_\theta^\infty. \tag{A.8}$$

For (A.7), let f be bounded and continuous on $\mathcal{X}^\infty$. By definition,

$$\int_{\mathcal{X}^\infty} f\, dP_{\mu n} = \int_\Theta \int_{\mathcal{X}^\infty} f\, dQ_\theta^\infty \mu_n(d\theta).$$

Now g: $\theta \to \int f\, dQ_\theta^\infty$ is a bounded, continuous function on Θ, in view of Corollary A.3. In view of (i), $\int g\, d\mu_n \to g(\theta)$ a.e. Q_θ^∞ for all θ. Thus, $\int g\, d\mu_n \to g(\theta)$ a.e. P_ν. This proves (A.7) via Corollary A.1, letting f run through the analogs of $\mathcal{C}_\vee$ and $\mathcal{C}_\wedge$ on $\mathcal{X}^\infty$. For (A.8) by Doob's theorem, $\int f\, dP_\nu \to g(\theta)$ a.e. P_ν, and the rest of the argument is the same. Thus, (i) $\Rightarrow$ (iv) in the presence of continuity.

Finally, assume (iv), and let $\nu = \delta_\theta$, so $P_\nu = \delta_\theta \times Q_\theta^\infty$ and $P_{\nu n} = Q_\theta^\infty$ a.e. Q_θ^∞. Also, $P_{\mu n} = \int Q_\phi^\infty \mu_n(d\phi) \to Q_\theta^\infty$ a.e. Q_θ^∞ by (iv). Thus $U(\mu_n) \to U(\delta_\theta)$ a.e. Q_θ^∞ and $\mu_n \to \delta_\theta$ a.e. Q_θ^∞, by Proposition A.1 below, where U is defined. □

REMARKS. (1) Consider a Bayesian with prior μ who chooses μ_n as the posterior. Consider a second Bayesian with the generic prior ν. Now G is the set where the first Bayesian gets the parameters θ right. Condition (ii) in the theorem is that any second Bayesian will be sure that the first Bayesian gets θ right—whatever θ may be. Condition (iii) is that any second Bayesian is sure that his posterior will merge weak-star with that of the first Bayesian. (This is well defined, since P_ν-null sets do not matter here: any version ν_n can be used.) Condition (iv) is that any second Bayesian is sure that his conditional opinion of the future given the past will merge weak-star with that of the first Bayesian.

(2) The implications (iii) $\Rightarrow$ (i) or (iv) $\Rightarrow$ (i) are valid with much weaker notions of merging. Essentially, only one R is needed, provided R vanishes on the diagonal and is positive off the diagonal. Subfamilies of functions R lead to different notions of merging:

(i) $\int f\,d\alpha_n - \int f\,d\beta_n \to 0$ for every bounded continuous f.

(ii) $\lambda(\alpha_n, \beta_n) \to 0$ for λ metrizing the weak star topology.

(iii) $T(\alpha_n) - T(\beta_n) \to 0$ for every continuous function on the probabilities.

These notions are all different, even on $\mathbb{R}$. For example, let $\alpha_n = \delta_n$, $\beta_n = \delta_{n+1/n}$, with δ_x a point mass at x. Then $\lambda(\alpha_n, \beta_n) \to 0$ for Prohorov's metric, but there is a bounded continuous f with $\int f\,d\alpha_n = 1$, $\int f\,d\beta_n = 0$. Thus (ii) is different from (i). To see that (iii) is different from (ii), take $\alpha_n = \delta_n$, $\beta_n = (1 - (1/n))\delta_n + (1/n)\delta_{n+1}$. Take $T(\alpha) = \alpha(1) + \alpha(2)^2 + \cdots + \alpha(n)^n + \cdots$. Now $T(\alpha_n) \equiv 1$ but $T(\beta_n) \to 1/e$.

The notion of merging we use implies merging in the above senses and is equivalent to merging in the finest uniformity compatible with the weak star topology (see Kelley, 1955, Chapter 6). For further discussion of these issues, see Diaconis and Freedman (1984) and Dudley (1966, 1968).

(3) The continuity conditions are needed to conclude (iv) from (i). To see this, take $\Theta = [0,1]$ and $\mathcal{X} = \{0,1\}$ and let

$$f(\theta) = \begin{cases} \frac{1}{2}\theta & \text{for } \theta \in \left[0, \frac{1}{2}\right), \\ \theta & \text{for } \theta = \frac{1}{2}, \\ \frac{1}{2} + \frac{1}{2}\theta & \text{for } \theta \in \left(\frac{1}{2}, 1\right]. \end{cases}$$

Let Q_θ be $f(\theta)$-coin tossing. Let μ be the uniform distribution on Θ. Straightforward analysis shows that μ is consistent, but the predictive distribution $P_{\mu n}$ does not merge with $P_{\nu n}$ when $\nu = \delta_{1/2}$. Lockett (1971) gives a similar example involving geometric variables.

(4) The continuity conditions are needed to conclude (i) from (iv). To see this, take $\Theta = [0, 2\pi)$ and $\mathcal{X}$ as the unit circle. Let $F(\theta) = e^{i\theta}$ map Θ onto $\mathcal{X}$. This F is 1–1 and continuous, but does not have a continuous inverse at $\xi = (1,0) \in \mathcal{X}$.

Define $Q_\theta = \delta_{F(\theta)}$. Let μ be uniform on Θ. Define a posterior, maliciously, as

$$\mu_n(\cdot|x_1 \ldots x_n) = \delta_{F^{-1}(x_1)} \quad \text{if } x_1 \neq \xi$$
$$= \delta_{2\pi - 1/n} \quad \text{if } x_1 = \xi.$$

So $(0, \mu_n)$ is not consistent. But $P_{\mu n} \to Q_\theta^\infty$ a.e. Q_θ^∞, for all θ, even $\theta = 0$, and so merges with $P_{\nu n}$ a.e. P_ν, for any ν.

The following is required to complete the proof of Theorem A.1. If S is a Borel set, we write $\pi(S)$ for the set of probabilities on S, endowed with the weak-star topology. The set $\pi(S)$ is Borel too. Let S_0 be a Borel subset of S. Let $\mathscr{P}_0 = \{\mu | \mu \in \pi(S) \text{ and } \mu\{S_0\} = 1\}$. Then $\mathscr{P}_0$ is a Borel subset of $\pi(S)$. Let M map $\mathscr{P}_0$ onto $\pi(S_0)$ as follows: if $\alpha \in \mathscr{P}_0$, then $M\alpha$ is the restriction of α to the Borel subsets of S_0. Thus, M maps probabilities on S to probabilities on S_0. We write $M = M(S, S_0)$ to show the dependence on the spaces S and S_0.

LEMMA A.3. *M is a homeomorphism of $\mathscr{P}_0$ onto $\pi(S_0)$.*

PROOF. Use Corollary A.1. □

Using Lemma A.3, we can view $\mathscr{X}$ as a dense Borel subset of the compact metric space $(\bar{\mathscr{X}}, \rho)$; metrize $\mathscr{X}$ by ρ. Then $\pi(\mathscr{X})$ can be viewed as a Borel subset of the compact metric set $\pi(\bar{\mathscr{X}})$. For μ a probability on Θ, let

$$U(\mu) = \int_\Theta Q_\theta^\infty \, \mu(d\theta).$$

So $U(\mu)$ is a probability on $\mathscr{X}^\infty$.

PROPOSITION A.1. *U is a homeomorphism of $\pi(\Theta)$ into $\pi(\mathscr{X}^\infty)$.*

PROOF. Let $\bar{\Theta} = \pi(\bar{\mathscr{X}})$. Let $\bar{U}$ map $\pi(\bar{\Theta})$ into $\pi(\bar{\mathscr{X}}^\infty)$ as follows:

$$\bar{U}(\mu) = \int_{\bar{\Theta}} \theta^\infty \, \mu(d\theta).$$

Then $\bar{U}$ is continuous, and 1–1 by de Finetti's theorem (Diaconis and Freedman, 1980, Appendix). Since $\pi(\bar{\Theta})$ is compact, $\bar{U}^{-1}$ is continuous. Next, suppose $\mathscr{X}$ is a Borel subset of $\bar{\mathscr{X}}$, Θ is a Borel subset of $\pi(\mathscr{X})$, and $Q_\theta = \theta$. In this setting, $U(\mu) = \int_\Theta \theta^\infty \, \mu(d\theta)$. Now apply Lemma A.3 twice, with $M_1 = M(\bar{\Theta}, \Theta)$ and $M_2 = M(\bar{\mathscr{X}}^\infty, \mathscr{X}^\infty)$. Then $U = M_2 \circ \bar{U} \circ M_1^{-1}$ is a homeomorphism on $\pi(\Theta)$. To see that the composition makes sense, let $\mu \in \pi(\theta)$ and $\nu = M_1^{-1}\mu$; then $\nu \in \pi(\bar{\Theta})$ and $\nu(\Theta) = 1$, so $\bar{U}(\nu)(\mathscr{X}^\infty) = 1$ because $\theta(\mathscr{X}) = 1$ for $\theta \in \Theta$.

The general case is almost immediate: $\theta \to Q_\theta$ being continuous and 1–1 on Θ, the image $\tilde{\Theta}$ under this mapping of Θ is a Borel subset of $\pi(\mathscr{X})$: see Kuratowski (1958). Apply the previous argument with $\tilde{\Theta}$ in place of Θ. Let M_0 map $\pi(\Theta)$ onto $\pi(\tilde{\Theta})$ by the recipe $(M_0\mu)(A) = \mu\{\theta\colon Q_\theta \in A\}$ for Borel $A \subset \tilde{\Theta}$. This M_0 is a homeomorphism because $\theta \to Q_\theta$ is. Now $U = M_2 \circ \bar{U} \circ M_1^{-1} \circ M_0$ is a homeomorphism too. □

APPENDIX B

The derivative of the posterior with respect to the prior. Proceeding heuristically for a moment, the derivative of the ratio $T(\mu) = N(\mu)/D(\mu)$ is

$$\dot{T}_\mu = \left[D(\mu)\dot{N}_\mu - N(\mu)\dot{D}_\mu\right]\Big/D(\mu)^2.$$

Now $N(\mu)$ is linear: $N(\mu + \delta) = N(\mu) + N(\delta)$. So $\dot{N}_\mu = N(\cdot)$. Likewise, $D(\mu)$ is a linear functional, so $\dot{D}_\mu = D(\cdot)$. For example, in our sense, the derivative of the linear function ϕ: $x \to 3x$ is just ϕ, because $\phi(x + h) = \phi(x) + \phi(h)$.

The upshot is $\dot{T}_\mu = R_\mu$, where

$$R_\mu(\cdot) = \frac{N(\cdot)}{D(\mu)} - \frac{N(\mu)}{D(\mu)^2}D(\cdot).$$

This is part (a) of Theorem 4. For a rigorous proof, we must show that for any δ with signed mass 0,

$$T(\mu + \delta) = T(\mu) + R_\mu(\delta) + o(\|\delta\|).$$

The difference $T(\mu + \delta) - T(\mu) - R_\mu(\delta)$ is easily seen to equal

$$-\left(N(\delta) - N(\mu)\frac{D(\delta)}{D(\mu)}\right)\left(\frac{D(\delta)}{D(\mu)\,D(\mu + \delta)}\right).$$

The norm of this is smaller than

$$\left|\frac{D(\delta)}{D(\mu)\,D(\mu + \delta)}\right|\left\{\|N(\delta)\| + \|N(\mu)\|\frac{|D(\delta)|}{D(\mu)}\right\}. \tag{B.1}$$

Let $C = \sup_\theta f(x|\theta) < \infty$ by assumption. Then $|D(\delta)| \le C\|\delta\|$ and $\|N(\delta)\| \le C\|\delta\|$. Further, $D(\mu + \delta)$ tends to $D(\mu)$ as δ tends to 0. It follows that the bound (B.1) is smaller than $C(\mu)\|\delta\|^2$ for $\|\delta\|$ small. This completes the proof of part (a).

To prove part (b) of Theorem 4, fix x and write f for $\theta \to f(x|\theta)$. Let $\bar{f} = f/D(\mu)$ so $\int \bar{f}\,d\mu = 1$. Then

$$d\dot{T}_\mu(\delta) = \bar{f}\,d\delta - \left(\int \bar{f}\,d\delta\right)\bar{f}\,d\mu.$$

Choose σ to dominate both $|\delta|$ and μ, e.g., $\sigma = |\delta| + \mu$. Then $d\delta = \dot{\delta}\,d\sigma$ and $d\mu = \dot{\mu}\,d\sigma$ and

$$d\dot{T}_\mu(\delta) = \left[\dot{\delta} - \left(\int \bar{f}\,d\delta\right)\dot{\mu}\right]\bar{f}\,d\sigma.$$

We must show

$$\|\dot{T}_\mu\| \le \|\delta\| \cdot \sup_\theta \bar{f}. \tag{B.2}$$

Indeed,

$$\int \dot{\delta}\bar{f}\,d\sigma = \int \bar{f}\,d\delta \quad \text{and} \quad \int \dot{\mu}\bar{f}\,d\sigma = \int \bar{f}\,d\mu = 1.$$

Thus, $\dot{T}_\mu(\delta)$ has signed mass 0, namely, $\dot{T}_\mu(\delta)[\Theta] = \int \bar{f}\, d\sigma - \int \bar{f}\, d\sigma = 0$. It follows that

$$\text{(B.3)} \qquad \|\dot{T}_\mu(\delta)\| = 2\int \left[\dot{\delta} - \left(\int \bar{f}\, d\delta\right)\dot{\mu}\right]^+ \bar{f}\, d\sigma = \int \left|\dot{\delta} - \left(\int \bar{f}\, d\delta\right)\dot{\mu}\right| \bar{f}\, d\sigma.$$

Assume without loss of generality that $\int \bar{f}\, d\delta \geq 0$. For $a \geq 0$ and real d, clearly $(d - a)^+ \leq d^+$. So

$$\|\dot{T}_\mu(\delta)\| \leq 2\int \dot{\delta}^+ \bar{f}\, d\sigma \leq 2\left(\int \dot{\delta}^+\, d\sigma\right)(\sup \bar{f}).$$

But

$$\int \dot{\delta}^+\, d\sigma = \tfrac{1}{2}\int |\dot{\delta}|\, d\sigma = \tfrac{1}{2}\|\delta\|$$

because $\int \dot{\delta}\, d\sigma = \int d\delta = 0$. This completes the proof of (B.2), and hence part (b) of Theorem 4.

To prove part (c) of Theorem 4, we must show that for every $\varepsilon > 0$ there is signed measure δ with signed mass 0 and total mass 1, satisfying

$$\|\dot{T}_\mu(\delta)\| \geq (1 - \varepsilon)\sup_0 \bar{f}.$$

Choose θ_0 with $\mu\{\theta_0\} = 0$ and $f(\theta_0) > (1 - \varepsilon)\sup_0 \bar{f}$. Let $\delta = \frac{1}{2}(\delta_{\theta_0} - \mu)$, where δ_{θ_0} is point mass at θ_0. Let $\sigma = \mu + \delta$. Then the rightmost expression in (B.3) can be evaluated, and is $\bar{f}(\theta_0)$, so

$$\|\dot{T}_\mu(\delta)\| = \bar{f}(\theta_0) \geq (1 - \varepsilon)\sup_0 \bar{f}. \qquad \square$$

REMARKS. (1) Ordinarily, $\sup_0 f = \sup f$, so the theorem determines $\|\dot{T}_\mu\|$. If e.g., $\mu\{\theta_0\} > 0$ and $f(\theta_0) > \sup\{f(\theta)\colon \theta \neq \theta_0\}$, then $\sup f > \sup_0 f$. In this case, $\|\dot{T}_\mu\|$ is hard to compute. However, it can be shown that $\|\dot{T}_\mu\| = \sup\|\dot{T}_\mu(\delta)\|$ where $\delta = \frac{1}{2}\delta_{\theta_1} - \frac{1}{2}\delta_{\theta_2}$ and $\theta_1 \neq \theta_2$ vary over Θ.

(2) If δ is required to be absolutely continuous with respect to μ, a similar argument shows that $\|\dot{T}_\mu(\delta)\| \leq \mu$-ess. sup $\bar{f}$, the inequality being sharp if e.g., μ is continuous.

(3) We have chosen to differentiate in the set of signed measures. Our strong derivative is called the "Frechet derivative." Another standard way of perturbing μ is to consider the mixture $(1 - \varepsilon)\mu + \varepsilon\nu$ for some probability ν as ε tends to zero: the "Gateaux derivative." The mixture can be written as $\mu + \varepsilon(\nu - \mu)$ and the notions of derivative coincide—for bounded likelihood functions.

(4) Similar computations can be carried out for Bayes rules. If θ is a real parameter, the mapping M

$$\mu \to \frac{\int \theta f(x|\theta)\, \mu(d\theta)}{\int f(x|\theta)\, \mu(d\theta)}$$

has derivative

$$\dot{M}_\mu(\cdot) = \frac{N_1(\cdot)}{D(\mu)} - \frac{N_1(\mu)}{D(\mu)^2} D(\cdot),$$

where $N_1(\delta) = \int\theta f(x|\theta)\,\delta(d\theta)$ and $D(\delta) = \int f(x|\theta)\,\mu(d\theta)$ as before. The norm of $\dot{M}_\mu$ is computed as follows. Let $c = N_1(\mu)/D(\mu)$. This is the Bayes rule based on μ. Let $g(\theta) = (\theta - c)f(x|\theta)$. Define range $g = \sup g - \inf g$. Then

$$\|\dot{M}_\mu\| = \tfrac{1}{2}\,\text{range}\,g.$$

(5) Theorem 4 assumes a dominated family. In the undominated case, the derivative need not exist in our strong sense. The difficulties can already be seen in the following simple example: take $\mathcal{X}$ and Θ to be the real line mod 10. Let $Q_\theta\{x\} = \frac{1}{2}$ if $x = \theta \pm 1$; suppose the prior μ for θ has continuous density f on Θ. the posterior for θ given x is supported at $x \pm 1$ with mass

$$T(\mu)\{\theta\} = \begin{cases} \dfrac{f(x-1)}{f(x-1)+f(x+1)} & \text{if } \theta = x-1, \\ \dfrac{f(x+1)}{f(x-1)+f(x+1)} & \text{if } \theta = x+1. \end{cases} \tag{B.4}$$

The map T is norm continuous at no x. To see this, consider a sequence of continuous prior densities f_n converging to f in variation distance but pointwise at no point. More specifically, let $s_n = 1 + \frac{1}{2} + \cdots + 1/n$. Let g_n on the line vanish to the left of $s_n - (2/n)$; increase linearly to the value 1 at s_n; decrease linearly to zero at $s_n + (2/n)$; and vanish to the right of that value. Wrap g_n around the line mod 10; let f_n be the sum of f and the wrapped g_n, normalized to be a density. Clearly $f_n \to f$ in L_1, but $f_n(x) \to f(x)$ for no x.

The argument is only sketched. Fix a real number θ with $0 \le \theta < 10$; for any integer k, the real number $k + \theta$ wraps to the same point θ in Θ. For infinitely many n, for some $k = k_n$, we have $s_n \le k + \theta < s_{n+1}$. Then $g_n(k + \theta) \ge \frac{1}{2}$, because $s_{n+1} - s_n < 1/n$. For such n, we have

$$f_n(\theta) > \left[f(\theta) + \frac{1}{2}\right]\frac{n}{n+2}$$

because $\int g_n = 2/n$. Since g_n has only one bump, this can be either at $x + 1$ or $x - 1$, but not both. Thus, for any x, for infinitely many n, we have both the following relations

$$f_n(x+1) > \left[f(x+1) + \frac{1}{2}\right]\frac{n}{n+2},$$

$$f_n(x-1) = f(x-1)\frac{n}{n+2}.$$

So f_n does not converge pointwise, and from (B.4) the map T is not continuous.

In this example, the posterior is Gateaux differentiable: A derivative can be calculated by considering $(1 - \varepsilon)f + \varepsilon g$ as ε tends to zero. The Gateaux derivative exists quite generally, as we will show elsewhere.

(6) In the situation of Theorem 4, the same result holds if the weak-star topology is used instead of the norm topology, provided f is bounded continuous in θ.

(7) A related computation is contained in Huber's (1973) discussion of Bayesian robustness.

REFERENCES

ANTONIAK, C. (1974). Mixtures of Dirichlet processes with application to Bayesian non-parametric problems. *Ann. Statist.* **2** 1152–1174.

BANACH, S. (1964). The Lebesgue integral in abstract spaces. In S. Saks, *Theory of the Integral.* 2nd ed. Dover, New York.

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

BERK, R. (1970). Stopping times of SPRTS—based on exchangeable models. *Ann. Math. Statist.* **41** 979–990.

BERNSTEIN, S. (1934). *Theory of Probability.* Moscow. (Russian).

BLACKWELL, D. and BICKEL, P. (1967). A note on Bayes estimates. *Ann. Math. Statist.* **38** 1907–1911.

BLACKWELL, D. and DUBINS, L. (1962). Merging of opinions with increasing information. *Ann. Math. Statist.* **33** 882–886.

BLACKWELL, D. and GIRSCHICK, M. A. (1954). *Theory of Games and Statistical Decisions.* Wiley, New York.

BLACKWELL, D. and RAMAMOORTHI, R. V. (1982). A Bayes, but not classically sufficient statistic. *Ann. Statist.* **10** 1025–1026.

BOX, G. E. P., LEONARD, T. and WU, C. F. (1983). *Scientific Inference, Data Analysis, and Robustness.* Academic, New York.

BOX, G. E. P. and TIAO, G. C. (1973). *Bayesian Inference in Statistical Analysis.* Addison-Wesley, Reading, Massachusetts.

BREIMAN, L., LE CAM, L. and SCHWARTZ, L. (1964). Consistent estimates and zero-one sets. *Ann. Math. Statist.* **35** 157–161.

BROWN, L. (1981). *Lecture Notes on Decision Theory.* Cornell Univ.

CORNFIELD, J. (1969). The Bayesian outlook and its applications. *Biometrics* **25** 617–657.

DALAL, S. (1977). Some contributions to Bayes nonparametric decision theory. Ph.D. thesis, Dept. Statist., Univ. Rochester, New York.

DALAL, S. R. (1979a). Dirichlet invariant processes and applications to nonparametric estimation of symmetric distribution functions. *Stochastic Process. Appl.* **9** 99–107.

DALAL, S. R. (1979b). Nonparametric and robust estimation of location. In *Optimizing Methods in Statistics* (J. S. Rustagi, ed.) 141–166. Academic, New York.

DALAL, S. R. and HALL, G. (1980). On approximating parametric Bayes models by nonparametric Bayes models. *Ann. Statist.* **8** 664–672.

DALAL, S. R. and HALL, W. J. (1983). Approximating priors by mixtures of natural conjugate priors. *J. Roy. Statist. Soc. B* **45** 278–286.

DE FINETTI, B. (1974). *Probability, Induction, Statistics.* Wiley, New York.

DIACONIS, P. and FREEDMAN, D. (1980). de Finetti's theorem for Markov chains. *Ann. Probab.* **8** 115–130.

DIACONIS, P. and FREEDMAN, D. (1983). Frequency properties of Bayes' rules. In *Scientific Inference, Data Analysis, and Robustness* (G. E. P. Box, T. Leonard, and C. F. Wu, eds.). Academic, New York.

DIACONIS, P. and FREEDMAN, D. (1984). Weak-star uniformities. Technical Report, Stanford Univ.

DIACONIS, P. and FREEDMAN, D. (1986). On inconsistent Bayes estimates of location. *Ann. Statist.* **14** 68–87.

DIACONIS, P. and YLVISAKER, D. (1979). Conjugate priors for exponential families. *Ann. Statist.* **7** 269–281.

DIACONIS, P. and YLVISAKER, D. (1985). Quantifying prior opinion. In *Bayesian Statistics* **2** (J. Bernardo, M. DeGroot, D. Lindley, and A. Smith, eds.). 133–156. North-Holland, Amsterdam.

DOKSUM, K. (1974). Tail-free and neutral random probabilities and their posterior distributions. *Ann. Probab.* **2** 183–201.

DOOB, J. L. (1948). Application of the theory of martingales. *Coll. Int. du C.N.R.S. Paris* 22–28.

DOSS, H. (1984). Bayesian estimation in the symmetric location problem. *Z. Wahrsch. verw. Gebiete* **68** 127–147.

DOSS, H. (1985a). Bayesian nonparametric estimation of the median; Part I: Computation of the estimates. *Ann. Statist.* **13** 1432–1444.

DOSS, H. (1985b). Bayesian nonparametric estimation of the median; Part II: Asymptotic properties of the estimates. *Ann. Statist.* **13** 1445–1464.

DUDLEY, R. (1966). Convergence of Baire measures. *Studia Math.* **27** 251–268.

DUDLEY, R. (1968). Distances of probability measures and random variables. *Ann. Math. Statist.* **39** 1563–1572.

EBERLEIN, W. F. (1962). An integral over function space. *Can. J. Math.* **14** 379–384.

FABIUS, J. (1964). Asymptotic behavior of Bayes estimates. *Ann. Math. Statist.* **35** 846–856.

FERGUSON, T. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1** 209–230.

FERGUSON, T. (1974). Prior distributions on spaces of probability measures. *Ann. Statist.* **2** 615–629.

FRASER, D. A. S. (1976). Necessary analysis and adaptive inference. *J. Amer. Statist. Assoc.* **71** 99–113.

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.

FREEDMAN, D. and DIACONIS, P. (1982a). de Finetti's theorem for symmetric location families. *Ann. Statist.* **10** 184–189.

FREEDMAN, D. and DIACONIS, P. (1982b). On inconsistent M-estimators. *Ann. Statist.* **10**, 454–461.

FREEDMAN, D. and DIACONIS, P. (1983). On inconsistent Bayes estimates in the discrete case. *Ann. Statist.* **11** 1109–1118.

GHOSH, J. K., SINHA, B. K. and JOSHI, S. N. (1982). Expansions for posterior probability and integrated Bayes risk. In *Statistical Decision Theory and Related Topics III*, **1**, 403–456 (S. Gupta and J. Berger, eds.). Academic, New York.

GOOD, I. J. (1950). *Probability and the Weighing of Evidence*. Griffin, London.

HANNUM, R. and HOLLANDER, M. (1983). Robustness of Ferguson's Bayes estimator of a distribution function. *Ann. Statist.* **11** 632–639, 1267.

HINKLEY, D. V. (1980). Likelihood as approximate pivotal distribution. *Biometrika* **67** 287–292.

HODGES, J. L. and LEHMANN, E. (1952). The use of previous experience in reaching statistical decisions. *Ann. Math. Statist.* **23** 396–407.

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

JEFFREYS, H. (1967). *Theory of Probability*, 3rd ed. Clarendon, Oxford.

JOHNS, M. V. (1979). Robust Pitman-like estimators. In *Robustness in Statistics* (R. Launer and G. Wilkensen, eds.) 49–60. Academic, New York.

JOHNSON, R. A. (1967). An asymptotic expansion for posterior distributions. *Ann. Math. Statist.* **38** 1899–1906.

JOHNSON, R. A. (1970). Asymptotic expansions associated with posterior distributions. *Ann. Math. Statist.* **41** 851–864.

KADANE, J. B. and CHUANG, D. T. (1978). Stable decision problems. *Ann. Statist.* **6** 1095–1110.

KAHANE, J. P. and SALEM, R. (1958). Sur la convolution d'une infinité de distributions de Bernoulli. *Colloq. Math.* **6**, 193–202.

KELLEY, J. L. (1955). *General Topology*. Van Nostrand, Princeton.

KOLMOGOROV, A. N. (1942). Definition of center of dispersion and measure of accuracy to form a finite number of observations (Russian). *Izv. Akad. Nauk SSSR Ser. Mat.* **6** 3–32.

KURATOWSKI, C. (1958). *Topologie* **1**, 4th ed. Hafner, New York.

LAPLACE, P. S. (1774). Mémoire sur la probabilité des causes par les événements. *Mémoires de mathematique et de physique presentés à l'académie royale des sciences, par divers savants, & lûs dans ses assemblés* **6** [Reprinted in Laplace's *Oeuvres completes* **8** 27–65, English translation by S. Stigler, Technical Report 164, Dept. Statist., Univ. Chicago (September, 1984)].

LE CAM, L. (1955). An extension of Wald's theory of statistical decision functions. *Ann. Math. Statist.* **26** 69–81.

LE CAM, L. (1982). On the risk of Bayes estimates. In *Statistical Decision Theory and Related Topics III* **2** (S. Gupta and J. Berger, eds.) 121–138. Academic, New York.

LINDLEY, D. V. (1965). *Introduction to Probability and Statistics from a Bayesian Point of View*

I, II. University Press, Cambridge.
LINDLEY, D. V. (1972). *Bayesian Statistics—A Review.* SIAM, Philadelphia.
LOCKETT, J. L. (1971). *Convergence in Total Variation of Predictive Distributions: Finite Horizon.* Unpublished Ph.D. dissertation, Dept. Statist., Stanford Univ.
MATTHES, T. K. and TRUAX, D. R. (1967). Tests of composite hypotheses for the multivariate exponential family. *Ann. Math. Statist.* **38**, 681–697.
NEYMAN, J. (1977). Frequentist probability and frequentist statistics. *Synthese* **36** 97–129.
NOORBALOOCHI, S. and MEEDEN, G. (1983). Unbiasedness as the dual of being Bayes. *J. Amer. Statist. Assoc.* **78** 619–623.
PRATT, J. (1965). Bayesian interpretation of standard inference statements. *J. Roy. Statist. Soc. B* **27** 169–203.
SACKS, J. (1963). Generalized Bayes solutions in estimation problems. *Ann. Math. Statist.* **34** 751–768.
SAVAGE, L. J. (1971). Elicitations of personal probability and expections. *J. Amer. Statist. Assoc.* **66** 783–801.
SAVAGE, L. J. (1972). *The Foundations of Statistics.* Dover, New York.
SCHWARTZ, L. (1965). On Bayes' procedures. *Z. Wahrsch. verw. Gebiete* **4** 10–26.
STEIN, C. S. (1955). A necessary and sufficient condition for admissibility. *Ann. Math. Statist.* **26** 518–522.
STEIN, C. S. (1981). On the coverage probability of confidence sets based on a prior distribution. Technical Report 180, Dept. Statist., Stanford Univ.
TJUR, T. (1980). *Probability Based On Randon Measures.* Wiley, New York.
VON MISES, R. (1964). *Mathematical Theory of Probability and Statistics.* (H. Geiringer, ed.). Academic, New York.
WALD, A. (1950). *Statistical Decision Functions.* Wiley, New York.
WELCH, B. L. and PEERS, H. W. (1963). On formulas for confidence points based on integrals of weighted likelihoods. *J. Roy. Statist. Soc. B.* **25** 318–329.
ZABELL, S. (1979). Continuous versions of regular conditional distributions. *Ann. Probab.* **7** 159–165.

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA
BERKELEY, CALIFORNIA 94720

DISCUSSION

ANDREW R. BARRON[1]

University of Illinois, Urbana

1. General remarks. Diaconis and Freedman have demonstrated some advantages and pitfalls of Bayesian inference. In summary, their results include the inconsistency of location estimates based on a Dirichlet prior; the equivalence of weak consistency and weak merging of posteriors; and an analysis of the sensitivity of the posterior to changes in the prior. In this discussion, we provide additional insight and point toward new developments. It is argued that the Dirichlet is a poor choice of prior because the Dirichlet mixture has a likelihood which is exponentially smaller than every product likelihood. We give conditions

[1]Work supported in part by NSF Grant ECS 82-11568 at Stanford University.