The Annals of Statistics
1984, Vol. 12, No. 4, 1151-1172

SPECIAL INVITED PAPER

BAYESIANLY JUSTIFIABLE AND RELEVANT FREQUENCY CALCULATIONS FOR THE APPLIED STATISTICIAN[1]

BY DONALD B. RUBIN

University of Chicago

A common reaction among applied statisticians is that the Bayesian statistician's energies in an applied problem must be directed at the a priori elicitation of *one* model specification from which an optimal design and all inferences follow automatically by applying Bayes's theorem to calculate conditional distributions of unknowns given knowns. I feel, however, that the applied Bayesian statistician's tool-kit should be more extensive and include tools that may be usefully labeled frequency calculations. Three types of Bayesianly justifiable and relevant frequency calculations are presented using examples to convey their use for the applied statistician.

1. Introduction. My purpose here is to discuss three important uses of frequency calculations for the applied Bayesian statistician: (1) for understanding, communicating and scientifically validating Bayesian statements, (2) for examining operating characteristics of Bayesian inferences derived from general models in order to understand the propriety of those models in a range of possible contexts, and (3) for monitoring the adequacy of specific models with fixed data sets. Before discussing these, I describe what I mean by the terms in the title of this paper, because their uses here may be somewhat idiosyncratic.

1.1 *Bayesian inference for the applied statistician.* First, consider the expression "Bayesian inference." By this I simply mean the method of statistical inference that draws conclusions by calculating conditional distributions of unknown quantities given (a) known quantities and (b) model specifications. Thus, in Bayesian inference, known quantities are treated as observed values of random variables and unknown quantities are treated as unobserved random variables; the conditional distribution of unknowns given knowns follows from applying Bayes's theorem to the model specifying the joint distribution of known and unknown quantities.

One important point in this last statement is that the plural form of "specifications" is intentional. If more than one model is being entertained, then more than one Bayesian inference is being entertained. For the applied Bayesian statistician, there is no need to arrive at one Bayesian inference, although such a goal may often be desirable.

Received May 1984; revised June 1984.

[1] This paper is the revised text of a Special Invited IMS lecture presented at the joint meeting of the IMS and ASA, August 21, 1982, Cincinnati, Ohio.

AMS 1980 *subject classifications.* 62-07, 62A15, 62F15, 62L10, 62P99.

Key words and phrases. Calibration, empirical Bayes, inference, model monitoring, operating characteristics, posterior predictive checks, stopping rules.

Another important point for the applied Bayesian statistician concerns what is meant by "known". In many practical problems, the number of characteristics that might be known for analysis is enormous (e.g., addresses, names and family histories of medical patients). Although some purist Bayesian positions might assert that every characteristic that is observable at essentially no cost must be treated as known, the more realistic applied position must be that there are costs associated with builiding complex models. Consequently "known" refers to values that are both available and considered worthwhile to include in model specifications.

Just as several specifications can be entertained by the applied Bayesian statistician, several definitions of what is known can be considered. Thus, for example, in a completely randomized experiment, initial analyses might assume no covariates are known, a second group of analyses might assume a few obviously relevant covariates are known, and subsequent analyses might assume several other less important covariates are also known. As more covariates are considered known, the model specifications become more complicated and more difficult to formulate, but have the potential payoffs that the inferences will be more precise and specific to subpopulations defined by the covariates regarded as known.

1.2 *Bayesianly justifiable and Bayesianly relevant.* A calculation is Bayesianly justifiable for the applied statistician if it follows the path just described, that is, if it treats known values as observed values of random variables, treats unknown values as unobserved random variables, and calculates the conditional distribution of unknowns given knowns and model specifications using Bayes's theorem. The calculation of a confidence level for a particular interval estimate is an example of a calculation that is not Bayesianly justifiable since it treats the known values of statistics as unobserved.

A Bayesianly relevant calculation for the applied statistician is one that helps the process of communicating and validating Bayesian answers, as well as the process of selecting model specifications upon which to condition. For instance, standard goodness-of-fit tests for models are Bayesianly relevant under this definition since they help select models, but they are not, at least as usually interpreted, Bayesianly justifiable since they treat observed values of statistics as unknown random variables. We will see later in Section 5, however, that often such tests, or modifications of them, can be interpreted so as to be Bayesianly justifiable. Of course, these descriptions of the terms *Bayesianly justifiable* and *Bayesianly relevant* are not mathematically precise definitions. Nonetheless, I believe that they are useful, and their intended meanings will become clearer in the context of examples described later.

1.3 *Frequency calculations.* By frequency calculations I mean probability calculations that are given relative frequency, as opposed to utility interpretations, where the relative frequencies of specified outcomes can be over a set of actually observable events, over a set of hypothetically observable events, or a set of unobservable events. For example, the distribution of the number of heads in twenty new tosses of a coin, having already observed the outcomes of ten tosses of the coin, is a frequency calculation whether the twenty tosses are actually carried out or merely contemplated. Furthermore, such a calculation is Bayesianly justifiable when the results of the twenty new tosses are unobserved and the distribution being calculated is conditional on the outcomes of the ten observed tosses as well as model specifications, e.g., independent, identically distributed (i.i.d.) Bernoulli trials. Even though the outcomes of the first ten tosses are observed, frequency calculations could be made concerning the outcomes of these first ten tosses or about the outcomes of the next twenty tosses ignoring the fact that the first ten tosses were observed; such frequency calculations would not be Bayesianly justifiable because they do not fix the first ten tosses at the observed outcomes.

As with previous descriptions of terms, this description of *frequency calculations* is not a precise definition. But again I believe that the term is a useful label for the collection of techniques I discuss.

1.4 *Outline.* Section 2 presents several reasons why the applied statistician should care about Bayesian methods of inference; four examples from my work are used to illustrate some of the ideas in the section. These examples also should help the reader understand the more abstract discussion in later sections by revealing the types of concrete applications of Bayesian statistics I have in mind. Section 3 describes simple frequency calculations that are useful in practice for understanding and communicating Bayesian statements, as well as frequency calculations that calibrate Bayesian statements by tying them to frequencies of real-world events. Section 4 discusses the use of frequency calculations to examine the operating characteristics of Bayesian inferences in order to guide the choice of models on which to base inferences. Section 5 describes the use of posterior predictive frequency distributions of test statistics to monitor the adequacy of specific model specifications with fixed data sets. Finally, Section 6 concludes with a few summary comments.

This paper is rather idiosyncratic in the sense that I rely almost entirely on examples from my own work to illustrate ideas and make no systematic attempt to connect my discussion to the vast statistical literature concerning the relationships among Bayesian, frequency and fiducial ideas. Also, in contrast to much of this work which describes how inferences should be conducted by idealized statisticians in an idealized world, I am concerned with activities of applied statisticians in the real world who are subject to constraints of finite resources, many problems to examine, and mixed expertise of consumers.

2. Why should the applied statistician care about Bayesian inference?

One reaction not uncommon among applied statisticians is that Bayesian methods are not really of much practical use; reacting to the emphasis on subjectivity, these statisticians feel that Bayesian thinking may be relevant to personal decision making, but that it is largely irrelevant to the public study of scientific questions. Several reasons why the applied statistician should be concerned with Bayesian methods are described here.

2.1 *Levels of randomness lead to better answers.* Since Bayesian models treat all unknowns as random variables, Bayesian models formulate distributions for parameters, and thereby naturally create models with multiple levels of randomness. Commonly, the answers that are derived from such models with levels of randomness are termed empirical Bayesian, but the essential feature here is not the label attached to the estimators but the Bayesian structure, which treats the unknown parameters of interest as random variables. The resultant extra flexibility generally leads to better answers by allowing borrowing of strength.

As a specific example, in Rubin (1980) I used a simple Bayesian model to create improved prediction equations of performance in law school for applicants at 82 law schools. The parameters of interest were the weights of undergraduate grade point average (UPGA) across the 82 law schools in equations predicting FYA (First Year Average) using weighted UGPA plus LSAT (Law School Aptitude Test). The data consisted of FYA, UGPA and LSAT for three years of attending students. Three basic methods for estimating the weight μ_i, $i = 1, \cdots, 82$ were considered:

(1) separate estimates for each of the 82 law schools based on least squares regression of FYA on LSAT and UGPA;

(2) a common estimate obtained by taking a weighted average of the separate least squares estimates in (1);

(3) Bayes/empirical Bayes estimates obtained by letting μ_i be i.i.d. $N(\mu_*, \sigma_*^2)$, then estimating μ_* and σ_*^2 by maximum likelihood, and finally estimating the μ_i by their posterior expectations given the maximum likelihood estimates of μ_* and σ_*^2; for each i, these estimates were between those in (1) and (2).

The results of the study were that, first, the Bayesian estimates predicted future FYA slightly better than the other methods, and second, the Bayesian estimates looked reasonable in contrast to the separate estimates, which fluctuated wildly from year to year and across law schools, and the common estimate, which allowed no variation in multipliers across law schools. Table 1 gives the multipliers of UGPA for three schools, three years of data, and methods (1) and (3).

Many other examples of the success of Bayesian methods of estimation exist, an early extensive one being Mosteller and Wallace (1964). More recent relevant references include: Lindley and Smith (1972), Novick, Jackson, Thayer, and Cole (1972), Efron and Morris (1975), Fay and Harriot (1979), Dempster, Rubin, and Tsutakawa (1981), DuMouchel and Harris (1983), Morris (1983), and Dempster, Selwyn, and Weeks (1983).

2.2 *Bayesian methods can provide straightforward answers in apparently difficult problems.* Modeling parameters not only generally provides better answers in complicated problems, it also can create simple answers in apparently complicated problems by allowing the pooling of many small pieces of information. A specific example conveys the essential idea.

Braun, Jones, Rubin, and Thayer (1983) investigated the evidence that pre-

TABLE 1

Multipliers of UGPA in prediction equations: FYA = LSAT + (?) × UGPA

Law school	Least Squares Multiplier of UGPA					Bayesian Multiplier of UGPA				
	1973	1974	1975	1973–74 pooled	1974–75 pooled	1973	1974	1975	1973–74 pooled	1974–75 pooled
7	2507	301	105	526	164	151	138	116	152	130
64	−24	49	153	3	116	100	111	119	92	120
11	179	118	98	149	107	150	121	115	140	115

diction equations for FYA in business school differ for white and black applicants; that is, perhaps different weightings of UGPA, GMAT-V (Graduate Management Admission Test-Verbal) and GMAT-Q (Quantitative) would result in more accurate predictions for the different applicants. This problem is similar in structure to the law school example in Section 2.2 with the addition of more predictor variables, and essentially the same algorithm can be used to obtain Bayes/empirical Bayes estimates. The new feature that makes this study apparently difficult is the paucity of data from which to estimate a prediction equation for black applicants: of the 59 schools in the study, 14 have no black students and 20 have between one and three black students. Although unique least squares equations cannot be estimated for these 20 schools, it is clear that they do provide some evidence about the relationship between FYA and the predictors, and Bayesian models can utilize this evidence to create estimated prediction equations for both whites and blacks in each of the 59 schools. As expected, schools with few data points have estimates similar to the pooled estimate and have large standard errors, whereas schools with many data points have estimates closer to their least squares estimates and have small standard errors. A more detailed summary of the results in Braun, Jones, Rubin and Thayer (1983) is given in Rubin (1983a).

2.3 *Users interpret answers Bayesianly.* Another reason for the applied statistician to care about Bayesian inference is that consumers of statistical answers, at least interval estimates, almost uniformly interpret them Bayesianly, that is as probability statements about the likely values of parameters. Consequently, the answers statisticians provide to consumers should be capable of being interpreted as approximate Bayesian statements—that is, statisticians' summary statements should be Bayesianly justifiable in the sense defined in Section 1.

A simple example conveys the central idea. Suppose $x_i \overset{\text{i.i.d.}}{\sim} N(\mu, 1)$, $i = 1, \cdots, N$. The standard interval estimate for μ is based on the statement that

$$(\bar{x} - \mu) \sim N(0, 1/\sqrt{N})$$

where $\bar{x}$ is the observed mean X in the sample. Thus the standard 95% interval for μ is

$$\bar{x} \pm 2/\sqrt{N}. \tag{2.1}$$

This interval is usually constructed and motivated as a 95% confidence interval; that is, treating $\bar{x}$ as a random variable with μ fixed, interval (2.1) includes μ in 95% of possible samples. However, in any applied problem with the observed value of $\bar{x}$ inserted in (2.1), the interval is—at least in my experience—nearly always interpreted Bayesianly, that is, as providing a fixed observed interval in which the unknown μ lies with 95% probability. This is, of course, the only interpretation of the observed 95% interval that *directly* addresses the essential question about μ, so it is *the* interpretation of primary interest to the consumer of statistical answers. If we as statisticians accept the fact that consumers will interpret such statements Bayesianly as well as accept the position that it is appropriate that consumers of statistics seek such answers, then we have an obligation to make sure that the answers we provide are Bayesianly justifiable.

2.4 *Exposure of scientific uncertainty.* In many cases, it will be difficult to arrive at a unique Bayesianly justifiable answer. That is, we will often discover that as we consider various plausible model specifications, all consistent with the observed data, the resulting Bayesian answers change in important ways. Although this sensitivity of inference to model specifications might be seen as a handicap of Bayesian models, I believe it is a virtue. If we view statistics as a discipline in the service of science, and science as being an attempt to understand (i.e., model) the world around us, then the ability to reveal sensitivity of conclusions from fixed data to various model specifications, all of which are scientifically acceptable, is equivalent to the ability to reveal boundaries of scientific uncertainty. When sharp conclusions are not possible without obtaining more information, whether it be more data, new theory, or deeper understanding of existing data and theory, then it must be scientifically valuable and appropriate to expose this sensitivity and thereby direct efforts to seek the particular information needed to sharpen conclusions.

As a specific illustrative example, in Rubin (1983b) I used a Bayesian model with Box-Cox (1964) transformations to normality to estimate the mean and median X in a finite population of 804 units from a random sample of 100 units. Each unit was a municipality in New York state in 1960 and X was its population (New York City was represented by its five boroughs). Using the Box-Cox reference prior for the transformation parameter, λ, led to relative posterior probabilities of .0000, .1372, .8628, .0000 for $\lambda = 0, -1/8, -1/4, -1/2$. Table 2 displays 95% posterior intervals for the population mean and median given λ. The intervals for the median X are insensitive to plausible values of λ because the implied middles of the distributions of X are nearly the same for the two values of λ. In contrast, the intervals for the mean X are quite sensitive to plausible values of λ because the implied right tails of the distributions of X are dramatically different for the values of λ. Such behavior is typical because estimation of the mean requires assumptions about the extreme tails of the distributions, which observed data can never directly address. Thus this initial Bayesian analysis revealed the extreme sensitivity of the mean to the value of λ. This led to a reformulated model that incorporated reasonable prior information and exhibited less sensitivity of the mean to the value of λ. In particular, suppose that no municipality can be larger than 5 million; then the 95% interval estimate for the mean X based on $\lambda = -1/4$ is $(1.2 \times 10^4, 4.2 \times 10^4)$ rather than the extremely wide interval displayed in Table 2.

TABLE 2

Conditional (on λ) 95% central posterior intervals based on Box-Cox transformation to normality

Estimand	$\lambda = -1/4$	$\lambda = -1/8$	**True Value in Finite Population**
Median X	$(1.3 \times 10^3, 2.4 \times 10^3)$	$(1.4 \times 10^3, 2.8 \times 10^3)$	1.7×10^3
Mean X	$(2.3 \times 10^4, 1.8 \times 10^{13})$	$(1.0 \times 10^4, 2.5 \times 10^4)$	1.7×10^4

2.5 *Computational directness.* A final reason why the applied statistician should care about Bayesian inference is less philosophical than the previous two and deals with the ability to analyze data in complicated problems. Continuing advances in computing mean that many analyses previously considered computationally hopeless even for large computing centers can now be handled quite easily, even by individuals without access to large mainframe computers. In particular, many Bayesian analyses for complicated models can be carried out on a fixed data set relatively simply and straightforwardly using Monte Carlo methods to simulate posterior distributions. Corresponding frequency-based inferences using equivalently rich models might still be prohibitively expensive because, in general, under each model and for all values of its parameters being contemplated, the frequency distribution of statistics would have to be generated just as if the procedure were being evaluated for all possible values of the parameters. This task can be substantially more arduous than the Bayesian's since the frequentist statistician is being required to simulate the distribution of data sets not seen for many values of nuisance parameters that may not be supported by the observed data. In contrast, the Bayesian data analyst fixes the observed data and calculates the posterior distribution of the unknown quantities of interest, averaging over nuisance parameters. Of course, the applied statistician may often wish to consider several models, but for each model the Bayesian leaves the observed values of data fixed and integrates over nuisance parameters.

As a specific example, in Rubin (1981) I analyzed the results of eight parallel randomized experiments of the effects of special "coaching" programs, for the SAT (Scholastic Aptitude Test); Table 3 gives summary statistics. Letting X_i be the estimated effect in school i and V_i be its associated squared standard error (treated as known), a Bayesian model was formulated in which the X_i given the true effects, μ_i, were $N(\mu_i, V_i)$, the μ_i were i.i.d. $N(\mu_*, \sigma_*^2)$, and the prior distribution on (μ_*, σ_*) was approximately proportional to a constant with $0 < \sigma_* < 100$. The posterior distribution of the μ_i was analytically complicated but easily stimulated by approximating the posterior distribution of σ_* by a step-function. Table 4 summarizes the results of 200 independent draws from the posterior distribution of the μ_i. The results appear to be reasonable and informative, and moreover were easily obtained.

Some might view the reliance on Monte Carlo methods as defective relative

TABLE 3
Effects of coaching programs on SAT-V scores in eight randomized experiments

School	Estimated treatment effect	Standard error of effect estimate
A	28	15
B	8	10
C	−3	16
D	7	11
E	−1	9
F	1	11
G	18	10
H	12	18

TABLE 4
Summary of 200 *simulated coaching effects of eight schools*

School	95 percent interval				
		50 percent interval			
			Median		
A	−2	6	11	17	36
B	−6	4	8	12	19
C	−10	3	7	11	22
D	−7	4	7	13	21
E	−9	3	7	11	16
F	−8	2	7	11	20
G	−1	6	9	14	24
H	−3	4	8	13	24

to mathematical analysis, even for the type of applied statistics just described. On the contrary, I believe that Monte Carlo methods, in a limited sense, can be superior to mathematical analysis for certain kinds of applied inference. Using Monte Carlo methods frees the applied statistician to explore a great variety of models with relative ease, and thus statisticians can pursue the scientific goals of matching models to data more effectively and with less algebraic digression than if mathematical analysis were the only tool; if a resultant model is considered of general interest, then additional mathematical analysis of it may of course be highly desirable and generate greater understanding of general quantitative relationships. But having performed a difficult mathematical analysis for a specific model, there exists an undeniable investment in that work, and consequently I suspect, an undeniable prejudice to use that model repeatedly, even when it is not truly relevant. Since less investment is needed to fit models by Monte Carlo than by analysis, there may well be less prejudice towards convenient yet inappropriate models. Bayesian statistics and Monte Carlo methods are ideally suited to the task of passing many models over one data set.

3. Frequency calculations useful for understanding and validating Bayesian statements. The first group of Bayesianly justifiable and relevant frequency calculations to be discussed involves the interpretation of Bayesian statements. First, frequency calculations are useful for understanding the manipulations involved in Bayesian analyses, and therefore useful for communicating Bayesian answers to more naive consumers; the technique presented here describes these manipulations by use of simulations from hypothetical superpopulations. Second, frequency calculations are useful for making Bayesian statements scientific, scientific in the sense of capable of being shown wrong by empirical test; here the technique is the calibration of Bayesian probabilities to the frequencies of actual events.

3.1 *Superpopulation frequency simulations.* Suppose the model for data set X is given by $f(X \mid \theta)p(\theta)$ where θ is the parameter whose posterior distribution is to be calculated. Of course, the posterior distribution of θ given observed X is calculated via Bayes's theorem, but how is the conceptual content of this theorem easily conveyed? Consider the following simple description.

Suppose we first draw equally likely values of θ from $p(\theta)$, and label these $\theta_1, \cdots, \theta_s$. The θ_j, $j = 1, \cdots, s$ can be thought of as representing the possible populations that might have generated the observed X. For each θ_j, we now draw an X from $f(X \mid \theta = \theta_j)$; label these $X_1, \cdots, X_s$. The X_j represent possible values of X_j that might have been observed under the full model $f(X \mid \theta)p(\theta)$. Now some of the X will look just like the observed X and many will not; of course, subject to the degree of rounding and the number of possible values of X, s might have to be very large in order to find generated X_j that agree with observed X, but this creates no problem for our conceptual experiment. Suppose we collect together all X_j that match the observed X, and then all θ_j that correspond to these X_j. This collection of θ_j represents the values of θ that could have generated the observed X; formally, this collection of θ values represents the posterior distribution of θ. An interval that includes 95% of these values of θ is a 95% probability interval for θ and has the frequency interpretation that under the model, 95% of populations that could have generated the data are included within the 95% interval.

3.2 *Frequency calibration.* In some formal sense, a Bayesian statement, such as a 95% posterior interval for an unknown, needs no justification since given the explicitly stated models, the statement follows from the laws of probability theory. This justification, however, is not very satisfying to the applied statistician for two related reasons. First, why should the models that are being conditioned upon be accepted? This issue is addressed in Sections 4 and 5. Second, what does the stated Bayesian probability, e.g. 95%, mean objectively or empirically? The question of tying the 95% to real world events is addressed here via the concept of frequency calibration.

A Bayesian is calibrated if his probability statements have their asserted coverage in repeated experience. For example, if $\{I_1, I_2, \cdots\}$ represents a series of 95% Bayes interval estimates for unknowns $\{\theta_1, \theta_2, \cdots\}$ from known data sets $\{X_1, X_2, \cdots\}$, then these statements are calibrated if 95% of them cover their unknowns and 5% do not. A subsequence of $\{I_1, I_2, \cdots\}$ is calibrated if 95% of those I_j in the subsequence cover their unknowns. For an interesting discussion of this idea, see Dawid (1982). Clearly, it is desirable for a Bayesian to be calibrated overall and for all subsequences defined by characteristics of the data sets.

If the Bayesian's models are correct, he will be calibrated overall and in all such subsequences. That is, if his models are correct $\Pr(\theta_j \in I_j \mid X_j) = .95$ for all j, and thus averaging over all data sets

$$\Pr(\theta_j \in I_j) = .95,$$

or averaging over all data sets with observed characteristric $Q = Q(X_j)$,

$$\Pr(\theta_j \in I_j \mid X_j \text{ satisfies } Q) = .95.$$

Thus, the Bayesian who uses realistic models can be expected to be calibrated overall and in any collection of cases with common observed characteristics. Consequently, the probabilities attached to Bayesian statements do have frequency interpretations that tie the statements to verifiable real world events.

We discuss the concept of calibration in more detail in the context of examining operating characteristics of Bayesian procedures.

4. Examining operating characteristics to select Bayesian models resulting in calibrated procedures. Many statisticians spend much of their professional time analyzing actual data; others spend much of their professional time investigating properties of inferences resulting from models in order to arrive at selected models that can be recommended for general use. If the world were such that all data analyses were to be performed only by expert Bayesian statisticians with (a) a full arsenal of statistical models that can be tuned to any situation, (b) full knowledge of the substantive area under study, and (c) essentially unlimited resources for each problem, then the need for Bayesian models for general consumption would not exist. Realistically, however, many data analyses are done by statistically relatively rather naive consumers of statistical techniques, or by statisticians with limited knowledge of the substantive area, or in a context of limited resources. Because of these constraints, a primary goal of much Bayesian statistical work must be the development and selection of models that perform well in rather general contexts. Frequency calculations that investigate the operating characteristics of Bayesian procedures are relevant and justifiable when investigating or recommending procedures for general consumption.

4.1 *Structure.* Suppose that the Bayesian statistician's model is $f(X \mid \theta)p(\theta)$ and that under this model, $I(X)$ is a 95% interval for θ; that is,

$$\frac{\int_{\theta \in I(X)} f(X \mid \theta)p(\theta)\, d\theta}{\int_{\theta} f(X \mid \theta)p(\theta)\, d\theta} = .95.$$

Further suppose that, as is always the case, the statistician's model $f(X \mid \theta)p(\theta)$ is chosen, to some extent, for computational simplicity and convenience, and is not precisely the model generating the kind of data in the field to which $I(X)$ will be applied. Suppose instead that in this field of application, data are generated according to the unknown model $g(X \mid \theta)q(\theta)$. The model $g(X \mid \theta)q(\theta)$ will be called the correct model; it is the model that the statistician would want to use to draw inferences because, in a frequency sense, it represents the distribution of (X, θ) that occurs in the ranges of examples to which the current procedures based on $f(X \mid \theta)p(\theta)$ will be recommended to be applied. If in this range of examples, inferences were based on $g(X \mid \theta)q(\theta)$, then all such inferences would be calibrated overall and in all subsequences defined by X.

The correct posterior distribution of θ is

$$\frac{g(X \mid \theta)q(\theta)}{\int_{\theta} g(X \mid \theta)q(\theta)\, d\theta}.$$

For given X, the probability coverage of $I(X)$ over this posterior distribution is

$$PC(X) = \frac{\int_{\theta \in I(X)} g(X \mid \theta) q(\theta)\, d\theta}{\int_\theta g(X \mid \theta) q(\theta)\, d\theta}. \tag{4.1}$$

If $g(X \mid \theta) = f(X \mid \theta)$ and $q(\theta) = p(\theta)$, then $PC(X) = .95$ for all X. In general, however, the coverage probability of $I(X)$, $PC(X)$, has a distribution over the examples to which $I(X)$ will be applied, a distribution defined by $g(X \mid \theta) q(\theta)$. For general consumption, it seems wise for the statistician to choose the model $f(X \mid \theta) p(\theta)$ used to derive $I(X)$ in such a way that the distribution of $PC(X)$ is tightly concentrated about .95 for a reasonably broad range of *plausible* $g(X \mid \theta) q(\theta)$. That is, it is wise to choose $f(X \mid \theta) p(\theta)$ so that consumers are nearly calibrated for any subsequence of studies to which $I(X)$ will be applied.

A much weaker requirement is that consumers be calibrated for the overall sequence, but not necessarily for every subsequence: $E[PC(X)] = .95$, where

$$E[PC(X)] = \int_X PC(X) \left\{ \int_\theta g(X \mid \theta) q(\theta)\, d\theta \right\} dX$$

or by (4.1)

$$E[PC(X)] = \int_X \int_{\theta \in I(X)} g(X \mid \theta) q(\theta)\, d\theta\, dX.$$

A procedure is conservatively calibrated overall when $PC(X)$ is typically at least .95, e.g., $E[PC(X)] \geq .95$. If $I(X)$ is calibrated overall with $f(X \mid \theta) = g(X \mid \theta)$ for all $q(\theta)$, i.e., for all θ, it is a confidence interval for θ under $f(X \mid \theta)$. This is a fairly weak statement in the absence of statements about calibration conditional on characteristics of the data, but it is not an unattractive property to a Bayesian.

Three examples are used to illustrate these ideas.

4.2 *Example 1—Simple normal example with noninformative priors.* Consider the standard normal set-up with $X = (x_1, \cdots, x_N)$ where $x_i \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$ $i = 1, \cdots, N$ defines $f(X \mid \theta)$, $\theta = (\mu, \sigma^2)$, and a uniform prior distribution on $(\mu, \ln \sigma^2)$ defines $p(\theta)$. Then the interval $I(X)$ for μ is essentially the standard 95% confidence interval for μ, $x \pm 2s/N^{1/2}$.

Suppose $g(X \mid \theta) = f(X \mid \theta)$ so that the normal specification is not in doubt. Then $E[PC(X)] = .95$ for all $q(\cdot)$. That is, the choice of the noninformative Jeffreys prior for (μ, σ^2), $p(\mu, \ln \sigma^2) \propto$ constant, implies that users will be calibrated overall as long as the data are normal. Notice that this is a frequency property of a Bayesian inference. Notice also that this overall calibration does not imply the user is calibrated for subsequences defined by the data, and thus is a rather weak statement by itself. For example, if we choose a $q(\theta)$ and a specific value for s^2, say s_0^2, and compare the distribution of $PC(X)$ when $s^2 \geq s_0^2$ with its distribution when $s^2 < s_0^0$, we will find that $PC(X)$ tends to be larger than .95 when $s^2 \geq s_0^2$ and smaller than .95 when $s^2 < s_0^2$. For $q(\theta)$ that is close to $p(\theta)$, this effect will, of course, be small. And without specific knowledge of the substantive field, it's difficult to imagine a choice for $p(\theta)$ other than

$p(\mu, \ln \sigma^2) \propto$ constant that gives $PC(X)$ more tightly concentrated about .95 for a range of plausible $q(\theta)$.

I suspect that this sort of calculation showing Bayesian answers to be calibrated for a range of plausible deviations from a specific model is the major reason for the acceptance of particular Bayesian models as useful for general consumption. For instance, I believe that the acceptance of Jeffreys's priors in many applied problems is more due to their yielding standard confidence intervals, that is their yielding approximate Bayesian procedures calibrated overall for any true prior, than to any invariance or informationless arguments such as those presented in Box and Tiao (1973). An interesting question that then immediately arises is how closely related are overall calibration and Jeffreys's priors? Recent work by Stein (1981) extending work by Welch and Peers (1963) and Welsh (1965) suggests that for scalar, θ, $I(X)$ created using Jeffrey's rule (1961) to create $p(\theta)$ is calibrated to order $1/N$. Of course, any prior with support for all θ in its parameter space is calibrated to order $1/\sqrt{N}$ under the usual asymptotic arguments based on the large sample normality of the likelihood function.

4.3 *Example 2—Empirical Bayes intervals.* Consider the following empirical Bayes model for the one-way analysis of variance structure, which is a simplification of that given in Morris (1983):

$f(X \mid \theta)$ is given by

$$x_i \sim N(\mu_i, 1) \quad i = 1, \cdots, k;$$

letting $\mu = (\mu_1, \cdots, \mu_k)$, the parameter θ is (μ, σ^2) where $p(\mu \mid \sigma^2)$ is given by

$$\mu_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2),$$

and $p(\sigma^2)$ is given by

$$\sigma^2 \sim \text{uniform } (0, \infty).$$

Under this model, Morris obtains 95% intervals $I_i(X)$ for each μ_i, which he terms empirical Bayes intervals for the μ_i.

Suppose that

$$g(X \mid \theta) = f(X \mid \theta) \quad \text{and} \quad q(\mu \mid \sigma^2) = p(\mu \mid \sigma^2),$$

but that $q(\sigma^2)$ does not necessarily equal $p(\sigma^2)$. Then Morris suggests that

$$E[PC_i(X)] \geq .95 \quad \text{for all} \quad q(\sigma^2),$$

where $PC_i(X)$ is the probability coverage of $I_i(X)$ for μ_i. That is, accepting the normality of the data and the i.i.d. normality of the μ_i, the resulting Bayes inferences for μ_i under $f(X \mid \theta)p(\theta)$ are conservatively calibrated overall. Thus, in cases where the normality of the empirical Bayes model seems reasonable, Morris's work suggests that the consumer using the statistician's model $f(X \mid \theta)p(\theta)$ can expect to be conservatively calibrated overall no matter what distribution $q(\sigma^2)$ is correct. Again, this is a fairly weak statement in the absence of statements about conditional calibration. Nevertheless, I do regard it as providing some support for the use of that prior distribution on σ^2 for some range of problems.

4.4 *Example* 3—*The effect of stopping rules.* Let $f(X \mid \theta)$ with $X = (x_1, x_2, \cdots)$ be defined by $x_i \overset{\text{i.i.d.}}{\sim} N(\theta, 1)$. Suppose values of x_i are collected sequentially following the rule that data will continue to be observed until either 100 values of x_i have been collected or $\bar{x}_n/\sqrt{n}$ (where $\bar{x}_n$ is the current sample mean, and n is the current number of observations) is greater than C where C is a fixed constant. That is, collect data until the test-statistic is big or 100 values are collected. Such rules might be used in medical research where the test-statistic addresses the possible effects of a new treatment and the study is to be terminated if there exists strong enough evidence that the new treatment is harmful.

For any fixed $p(\theta)$, n and $\bar{x}_n$, Bayesian inferences for θ are unaffected by the stopping rule, that is, they are the same as if the sample size n were fixed a priori or chosen by a random mechanism independent of X; for instance, letting $p(\theta)$ be Jeffreys's prior, we obtain the standard 95% interval for θ, $I(X) = \bar{x}_n \pm 2/\sqrt{n}$. Although this result might be viewed as suggesting that Bayesian inference is unaffected by such data-dependent stopping rules, an examination of the sampling distribution of $PC(X)$, which ideally should be identically .95, shows otherwise.

Rosenbaum and Rubin (1984) studied the distribution of $PC(X)$ for the standard interval $I(X)$, where $g(X \mid \theta) = f(X \mid \theta)$ and $q(\theta) = N(0, V)$ for $V = .01$, .1, 1, 100; $V = \infty$ corresponds to $p(\theta)$. Table 5 from Rosenbaum and Rubin (1984) presents results of simulations of the distribution of $PC(X)$ for stopping rules corresponding to $C = 2$, 1.5, .5. Values in the table are the 10% points of the distributions of $PC(X)$, i.e. 90% of the values of $PC(X)$ are larger than the values given. Thus, for $C = 2$ and $V = .01$, 90% of the values of $PC(X)$ were larger than .54, whereas for $C = .5$ and $V = 100$, 90% of the values of $PC(X)$ were larger than .95. Values in parentheses are the 10% points of the distributions for $PC(X)$ for random samples with the same sample-size distributions as obtained when using the stopping rules. That is, for fixed V and C, there is a distribution of sample sizes, n, using the stopping rule, and n is correlated with $PC(X)$; for example, when n is less than 100, we know $\bar{x}_n \geq C\sqrt{n}$. But random samples with the same size distribution can be drawn, and they would not have any correlation

TABLE 5

Estimated 10% *point of distribution of* $PC(X)$ *over* $g(X \mid \theta)q(\theta)$.
Results of stopping rule (*results for random samples of same size*)

C	*V*			
	.01	**.1**	**1**	**100**
2.0	.54(.88)	.78(.92)	.91(.94)	.95(.95)
1.5	.74(.89)	.84(.91)	.91(.95)	.95(.95)
0.5	.93(.89)	.92(.87)	.89(.93)	.95(.95)

between n and $PC(X)$. We see that for such random samples, 90% of the values of $PC(X)$ are larger than .88 for all cases considered.

This table demonstrates a danger when using procedures involving data dependent stopping rules: the consumer is calibrated overall for a less wide range of models. As V becomes smaller, the proportion of intervals $I(X)$ that lead to low $PC(X)$ increases far more rapidly with the data-dependent stopping rule than with a random sampling rule, even one with the same distribution of sample sizes. The reason for the different behavior of $PC(X)$ for the data-dependent and random stopping rules is simple: inferences under $p(\theta) = N(0, \infty)$ and under $g(\theta) = N(0, V)$, $V < \infty$, are more disparate when $\bar{x}_n$ is further from zero, and such data sets are more likely with the $\bar{x}_n > \sqrt{n}C$ stopping rule than the random stopping rule. Thus the data-dependent stopping rule leads to data sets exhibiting greater sensitivity of inference to changes in the specification of prior variance, and consequently to users being uncalibrated for a wider range of models.

Furthermore the stopping rule can create identifiable subsequences (i.e. $\bar{x}_n > \sqrt{n}\ C$) where $I(X)$ is relatively poorly calibrated, especially for small V. For instance, when $V = .01$ and $C = 2$, fifteen of 25 simulated samples stopped at sample size 100, and of these, 14 had $PC(X) \doteq .99$. All ten samples that stopped with $n < 100$ had $PC(X) < .9$, with the lowest values of $PC(X)$ in those samples with fewest observations.

These results do not necessarily mean that data dependent stopping rules should not be used, but rather that if they are used, more care is needed in assessing models, and therefore they should probably be used cautiously.

5. Model monitoring by posterior predictive checks. In Section 4 we calculated the operating characteristics of Bayes procedures derived under the model $f(X \mid \theta)p(\theta)$ over other distributions for (X, θ). This process is Bayesianly justifiable when considering which models to recommend for general use, that is, before X is observed. After observing data X, however, such calculations are no longer Bayesianly justifiable because the observed values in X should be conditioned upon. Some (perhaps, Berger, 1983) would argue that, nevertheless even with X observed, such frequency calculations are Bayesianly relevant because they help to select better models upon which to base inferences.

Whether or not the unconditional operating characteristics of Bayes procedures are Bayesianly relevant when drawing inferences for θ from observed data X, there are many other frequency calculations that are Bayesianly relevant with observed data in that they help select appropriate models upon which to base inferences. Standard goodness-of-fit tests, probability plots, examinations of residuals, all with accompanying frequentist p-values, can help the Bayesian arrive at specifications $f(X \mid \theta)p(\theta)$ that are consistent with observed data; that is, they help to select models that plausibly could generate the observed data, and thus are Bayesianly relevant. But in what sense are they Bayesianly justifiable?

5.1 *Bayesian framework for model monitoring.* Many such frequency calculations can be Bayesianly justifiable if conceptualized properly. The only requirement is that we condition on observed values and calculate the distribution of unobserved quantities. Consider the following scheme. Given observed data, X_{obs}, what would we expect to see in hypothetical replications of the study that generated X_{obs}? Intuitively, if the model specifications are appropriate, we would expect to see something similar to what we saw this time, at least similar in "relevant ways". This statement, which is essentially a fundamental premise of frequency inference, seems to me so basic that it needs no defense. Unlike the frequentist, the Bayesian, though, will condition on all observed values.

In order to apply the idea, we first need to define a statistic $T(X)$ that formalizes the notion "relevant ways". Next we need to define precisely what we mean by a replication of the current study. Having defined $T(X)$ and the replications, we then calculate the frequency distribution of $T(X)$ in the hypothetical future replications, where this distribution is conditional on both (a) the observed data X_{obs} and (b) the current model specification $f(X \mid \theta)p(\theta)$. This distribution—the model monitoring distribution or posterior predictive check distribution (Rubin, 1981, 1983a)—is the posterior predictive distribution of $T(X)$, "posterior" meaning conditional on observed values and "predictive" meaning the distribution of a future observable quantity. If the frequency distribution of $T(X)$ does not make the observed value of $T(X)$, $T(X_{\text{obs}}) = T_{\text{obs}}$, appear typical, where typical is usually defined by tail areas of the distribution of $T(X)$ beyond $T(X_{\text{obs}})$, then we may want to revise the model $f(X \mid \theta)p(\theta)$. The reason is that the model, in replications of the current study, does not generate data that are similar to the observed data, where similar is judged by comparing T_{obs} to the distribution of $T(X)$.

5.2 *Defining statistics.* The question of how to define statistics $T(X)$ seems not easy to address in a formal way. If an alternative model is at hand that can be used to suggest specific sufficient statistics, it may often be wiser from a Bayesian perspective to fit the alternative model within the Bayesian framework, and assess the relative fits of the original and alternative model from the posterior distribution of all parameters under an extended model which encompasses both the original and alternative models. In this case Dempster (1975) suggests computing the posterior distribution of the likelihood ratio for the original model vs. alternative model, as induced by the posterior distribution of all parameters under the extended model.

Monitoring models by use of frequency calculations seems most useful when the current model is possibly acceptable and no particular alternative model is compelling. Then statistics $T(X)$ should be chosen to be potentially revealing of lack of fit (for example, residuals, order statistics) rather than to be sufficient under some expanded model; examples are given shortly.

The use of convenient statistics, such as residuals, to monitor models is now an established part of sound statistical practice. An alternative to examining such statistics is to embed the current model in a richer and richer web of Bayesian models. Although possible in principle, this seems to be a hopelessly complex task to always implement in practice. It is usually substantially more difficult to create a new relevant model and perform a full Bayesian analysis under it than to check the distribution of a few cleverly chosen statistics. Certainly, when there is no indication that the current model is inadequate, the extra modeling effort may often be a poor expenditure of time.

5.3 *Defining replications.* The question of how to define replications is similar to the question of how to define statistics in that formality may not help much; the practical context of the problem usually defines the replications. In some problems, such replications lead to unconditional frequency checks such as described by Box (1980), where the monitoring distribution of $T(X)$ is the distribution of $T(X)$ implied by the prior distribution of (X, θ), $f(X \mid \theta)p(\theta)$. In other cases it may make more sense to regard θ as a fixed feature of the replications and consider the monitoring distribution of $T(X)$ to be that obtained taking the distribution of $T(X)$ given θ implied by $f(X \mid \theta)p(\theta)$, $\Pr(T(X) \mid \theta)$, and averaging it over the posterior distribution of θ, i.e., over the conditional distribution of θ given X_{obs}, $f(X_{\text{obs}} \mid \theta)p(\theta)/\int f(X_{\text{obs}} \mid \theta)p(\theta)\, d\theta$. For example, in a study of D drugs, interest may focus on the fit of the model for these drugs rather than for a sample of D new drugs drawn from the same population of drugs.

In still other cases, we will want to fix features of data as well as parameters. For example, in sample surveys we often may wish to fix features such as the sample size and pattern of missing data in addition to functions of θ. Then the monitoring distribution of $T(X)$ is conditional distribution of $T(X)$ given the fixed features in θ and X_{obs} (implied by $f(X \mid \theta)p(\theta)$) averaged over the posterior distribution of fixed features. Denoting the fixed features by the vector $K(X, \theta)$, the monitoring distribution of $T(X)$ is defined to be the conditional distribution of $T(X)$ given $K(X, \theta)$ averaged over the posterior distribution of $K(X, \theta)$, i.e., over the conditional distribution of $K(X, \theta)$ given X_{obs}. This gives our best (posterior given X_{obs} and current model specifications) guess of the distribution of $T(X)$ in future replications of the current study that are identical to the current study with respect to features defined by $K(X, \theta)$.

Model monitoring by posterior predictive distributions is ideally suited to complicated Bayesian models whose posterior distributions are calculated by Monte Carlo rather than by analysis because we need not be constrained in our choice of monitoring statistics to ones for which we can calculate distributions analytically.

I use three examples to illustrate these ideas.

5.4 *Example 1—Gauss vs. Cauchy.* Suppose ten values of x_i are observed where the statistician's model is the usual i.i.d. normal setup with

$$x_i \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2) \quad i = 1, \cdots, 10$$

$$\theta = (\mu, \ln \sigma^2)$$

$$p(\theta) \propto \text{constant}.$$

The observed values of $\bar{x}$ and s^2 are $\bar{x}_{\text{obs}}$ and s^2_{obs}. Consider the posterior predictive distribution of ten new draws of x_i with the same θ; that is, consider replications of the same study (same parameters, same sample size, new data).

The relevant distributions will be described as if we were simulating the replications. First, draw a value of σ^2, say σ_*^2 from its posterior distribution, which is $9s_{\text{obs}}^2$ over a χ^2 on 9 degrees of freedom. Second, draw a value of μ, say μ_*, from its posterior distribution given $\sigma^2 = \sigma_*^2$, which is normal with mean $\bar{x}_{\text{obs}}$ and variance $\sigma_*^2/10$. Third, draw new $x_1, \cdots, x_{10}$ as i.i.d. $N(\mu_*, \sigma_*^2)$. Repeat these three steps M times to generate M draws of $x_1, \cdots, x_{10}$ from their posterior predictive distribution with the same $\theta = (\mu, \ln \sigma^2)$ in the replications as in the actual study. (If θ were not fixed to be the same as in the actual study, the first two steps would be replaced with draws from the prior distribution of θ rather than its posterior distribution. This is impossible with an improper prior such as in this example; notice that Box (1980) uses proper priors for his predictive checks, which always draw θ from its *prior* distribution.)

In these M replications, the sufficient statistics $\bar{x}$ and s^2 will look approximately like the observed values x_{obs} and s_{obs}^2. However, this is not true for all functions of the generated data. Consider monitoring statistics such as the gaps between the order statistics; for specificity, consider the ratio of the first to second gap, $T(X) = [x_{(10)} - x_{(9)}]/[x_{(9)} - x_{(8)}]$. When the model that generated the observed data is normal, the gaps between the observed order statistics will be typical of the simulated gaps, since both will be generated under the normal distribution, and thus $T(X_{\text{obs}})$ will be typical of the M simulated values of $T(X)$. But when the model that generated the data is Cauchy, the observed gaps will be Cauchy gaps, and so $T(X_{\text{obs}})$ will not be typical of the M values of $T(X)$, which simulate the ratio of gaps assuming normality. Consequently, the simulated monitoring distribution of $T(X)$ addresses the adequacy of the statistician's model and does so in a Bayesianly justifiable manner since it conditions on all observed values.

5.5 *Example 2—Data from eight experiments.* Rubin (1981) presents a study consisting of randomized experiments of coaching for the SAT exams in eight schools, introduced here in Section 2.5. Letting x_i be the estimated effect of coaching in the ith school, the model for the data is

$$(x_i \mid \mu_i) \sim N(\mu_i, V_i) \quad i = 1, \cdots, 8$$

where V_i is known and

$$(\mu_i \mid \mu_*, \sigma_*^2) \sim N(\mu_*, \sigma_*^2)$$

$$(\mu_*, \sigma_*) \sim \text{uniform on } (0 < \sigma_* < 100).$$

Let $x_{i,\text{obs}}$ be the observed value of x_i, $i = 1, \cdots, 8$ in the study.

Suppose that we returned to these eight schools to reconduct similar randomized experiments; what sort of observed coaching effects would we expect to see? Since the schools are the same, and we would like to conduct experiments of the same size, we fix $(\mu_1, \cdots, \mu_8)$ and $(V_1, \cdots, V_8)$ in the model monitoring, and thus examine the distribution of $(x_1, \cdots, x_8)$ given $(\mu_1, \cdots, \mu_8)$ and $(V_1, \cdots, V_8)$, averaging over the posterior distribution of $(\mu_1, \cdots, \mu_8)$, i.e., averaging over the distribution of $(\mu_1, \cdots, \mu_8)$ given $(x_{1,\text{obs}}, \cdots, x_{8,\text{obs}})$ and $(V_1, \cdots, V_8)$. The normality of x_i given (μ_i, V_i) and the a priori exchangeability in the prior distribution of the μ_i were considered reasonable for reasons given in Rubin (1981). No strong a priori defense, however, could be made for the normal prior for the μ_i given (μ_*, σ_*^2) nor for the uniform prior for (μ_*, σ_*). For any exchangeable prior on the μ_i, the order statistics $x_{(1)}, \cdots, x_{(8)}$ with their associated $V_{(i)}$ are sufficient, and hence functions of these order statistics are reasonable choices for the monitoring statistics.

Figure 1 from Rubin (1981) plots 200 simulated values of $x_{(1)}$, $V_{(1)}$ where $x_{(1)}$ is the largest observed x_i and $V_{(1)}$ is the variance associated with the largest x_i. The arrows point out the observed value of $(x_{(1)}, V_{(1)})$, which is typical of the simulated $(x_{(1)}, V_{(1)})$. Figures for $x_{(j)}$, $V_{(j)}, j = 2, \cdots, 8$ portray a similar story in that they provide evidence that the assumed model generates order statistics for

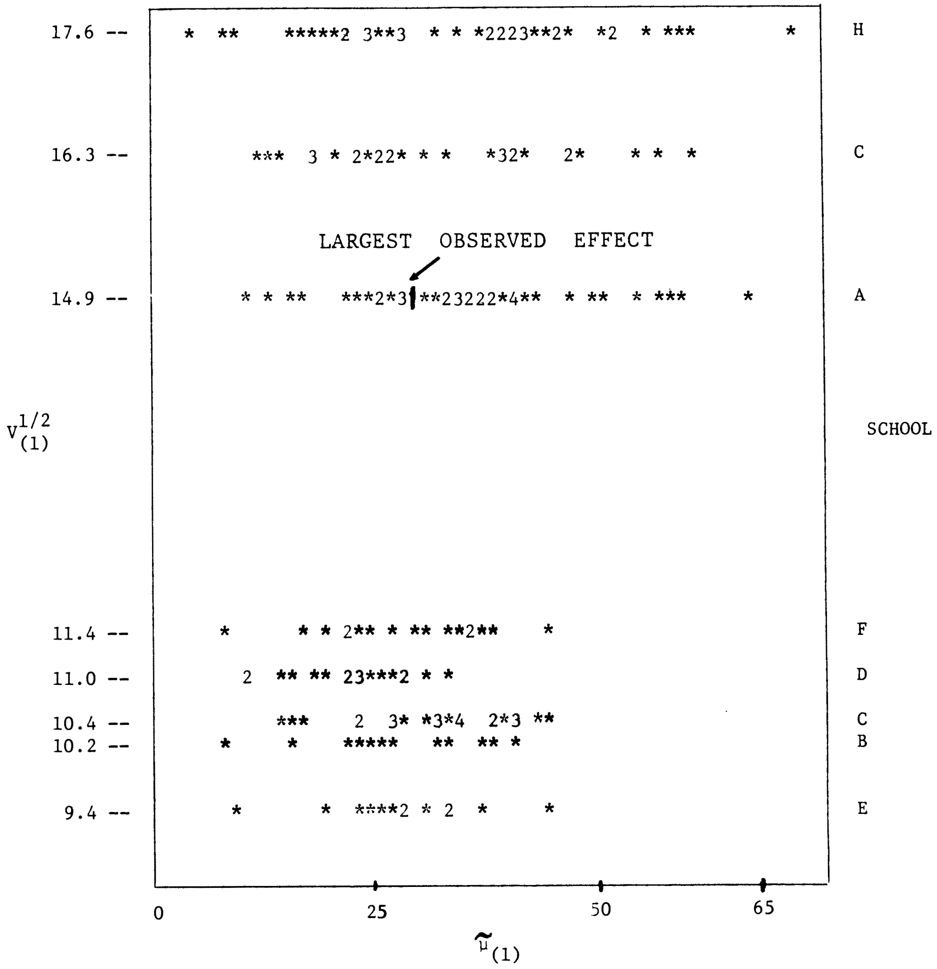

FIG. 1. *Joint distribution of largest estimated effect* $\tilde{\mu}_{(1)}$ *and its associated standard error,* $V_{(1)}^{1/2}$: 200 *simulated values.*

TABLE 6

Summary of 200 *simulations of the estimated effects for the eight schools*

Number of times that	$i = 1$	2	3	4	5	6	7	8
ith largest estimated effect occurred in the school with ith largest observed effect	41	25	18	19	24	23	30	27
and								
was larger than ith largest observed effect	28	11	10	10	8	16	13	5

the estimated effects quite similar to the observed effects; these figures are summarized in Table 6.

Of course, other statistics could be examined, but the essential idea of using model monitoring to detect inabilities of the assumed model to generate the observed data in this example is conveyed.

5.6 *Example* 3—*Fisher's randomization test for experiments.* The previous two examples treated the unknown parameters of the study as fixed features of the replications when calculating the monitoring distribution of statistics. In this example, we fix part of the data and show that Bayesian model monitoring can lead to Fisher's randomization test.

The setup is as follows: Suppose there are $2N$ units, with $x_i = (y_i, z_i)$ where y_i is an outcome variable and z_i indicates exposure to treatment 1 or treatment 0. The model $f(X \mid \theta)p(\theta)$ being entertained assumes no treatment effect, and thus under this model the observed value of y for the ith unit is the same whether it received treatment 1 ($z_i = 1$) or treatment 0 ($z_i = 0$); the model $f(X \mid \theta)p(\theta)$ further asserts that the experiment was completely randomized in the sense that $Y = (y_1, \cdots, y_N)$ and $Z = (z_1, \cdots, z_N)$ are independent and

$$\Pr(Z) = \begin{cases} 1/\binom{2N}{N} & \text{if } \sum Z_i = N \\ 0 & \text{otherwise.} \end{cases}$$

Consider the statistic that is the mean y for treatment 1 units minus the mean y for treatment 0 units:

$$T(X) = (1/N) \textstyle\sum_1^{2N} y_i z_i - (1/N) \sum_1^{2N} y_i(1 - z_i).$$

The observed value of $T(X)$, $T_{\text{obs}} = T(X_{\text{obs}})$ is the observed difference of means.

Suppose that interest is on the fit of the model for these units at this time; thus for monitoring purposes we fix the outcomes y_i, which, since all y_i are observed, means fixing all y_i in future replications at their current observed values. Then the monitoring distribution of $T(X)$ under $f(X \mid \theta)p(\theta)$ is simply the randomization distribution of $T(X)$ induced by $\Pr(Z)$ just as in Fisher's randomization test, and T_{obs} is compared to the distribution of $T(X)$ given Y just as in Fisher's randomization test.

Thus, the Bayesian justification of Fisher's randomization test is that it gives the posterior predictive distribution of the mean treatment difference under a model of no treatment effect and fixed units with fixed responses. If the observed mean difference is not typical of the monitoring distribution of differences (i.e., if a small p-value is obtained), then the null model is unacceptable and alternative models should be investigated. Although the frequentist can stop with a rejection of the null hypothesis, I believe that the Bayesian is obliged to seek and build a model that is acceptable to condition upon.

6. Conclusions. The applied statistician should be Bayesian in principle and calibrated to the real world in practice—appropriate frequency calculations help to define such a tie.

The applied statistician, when recommending Bayesian procedures for general consumption, should attempt to use specifications that lead to approximately calibrated procedures under reasonable deviations from those specifications—frequency calculations examining the operating characteristics of procedures are the basis for such judgments, where the more conditional the calibration the better.

The applied statistician should avoid models that are contradicted by observed data in relevant ways—frequency calculations for hypothetical replications can monitor a model's adequacy and help to suggest more appropriate models.

All three of these types of frequency calculations can be both Bayesianly justifiable and Bayesianly relevant. Such frequency calculations thus supplement the standard inferential Bayesian tool kit of prior assessment and prior to posterior calculation in a complementary way.

REFERENCES

BERGER, J. (1983). The robust Bayesian viewpoint. *Robustness in Bayesian Statistics* (J. Kadane, ed.). North-Holland, Amsterdam.

BOX, G. E. P. (1980). Sampling and Bayes' inference in scientific modelling and robustness. *J. Roy. Statist. Soc. A* **143** 383–430.

BOX, G. E. P. and COX, D. R. (1964). An analysis of transformations. *J. Roy. Statist. Soc. B* **36** 211.

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

BRAUN, H. I., JONES, D. J., RUBIN, D. B., and THAYER, D. T. (1983). Empirical Bayes estimation of coefficients in the general linear model with data of deficient rank. *Psychometrika* **48** 171–182.

DEMPSTER, A. P. (1975). A subjectivist look at robustness. *Bull. I. S. I. Proc. 40th Session* **46** 349–374.

DEMPSTER, A. P., RUBIN, D. B., and TSUTAKAWA, R. K. (1981). Estimation in covariance components models. *J. Amer. Statist. Assoc.* **76** 341–353.

DEMPSTER, A. P., SELWYN, M. R., and WEEKS, B. J. (1983). Combining historical and randomized controls for assessing trends in proportions. *J. Amer. Statist. Assoc.* **78** 221–227.

DUMOUCHEL, W. M. and HARRIS, J. E. (1983). Bayes methods for combining the results of cancer studies in humans and other species. *J. Amer. Statist. Assoc.* **78** 293–307.

EFRON, B. and MORRIS, C. (1975). Data analysis using Stein's estimator and its generalizations. *J. Amer. Statist. Assoc.* **70** 311–319.

FAY, R. E. and HARRIOT, R. A. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *J. Amer. Statist. Assoc.* **74** 269–277.

LINDLEY, D. V. and SMITH, A. F. M. (1972). Bayes estimates for the linear model. *J. Roy. Statist. Soc. B* **34** 1–41.

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

MOSTELLER, F. and WALLACE, D. L. (1964). *Inference and Disputed Authorship: The Federalist.* Addison-Wesley, Reading, Massachusetts.

NOVICK, M. R., JACKSON, P. H., THAYER, D. T., and COLE, N. S. (1972). Estimating multiple regressions in *m* groups: A cross-validation study. *British J. Math. Statist. Psychology* **25** 33–50.

ROSENBAUM, P. R. and RUBIN, D. B. (1984). Sensitivity of Bayes inference with data-dependent stopping rules. *The Amer. Statist.* **38** 106–109.

RUBIN, D. B. (1980). Using empirical Bayes techniques in the law school validity studies. *J. Statist. Assoc.* **75** 801–816.

RUBIN, D. B. (1981). Estimation in parallel randomized experiments. *J. Educ. Statist.* **6** 377–400.

RUBIN, D. B. (1983a). Some applications of Bayesian statistics to educational data. *The Statist.* **32** 55–68.

RUBIN, D. B. (1983b). A case study of the robustness of Bayesian methods of inference: Estimating the total in a finite population using transformations to normality. *Scientific Inference, Data Analysis and Robustness* 213–244. Academic, New York.

STEIN, C. M. (1981). On the coverage probability of confidence sets based on a prior distribution. Stanford Statistics Department Technical Report No. 180.

WELCH, B. L. (1965). On comparisons between confidence point procedures in the case of a single parameter. *J. Roy. Statist. Soc. B* **27** 1–8.

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** 328–329.

DEPARTMENT OF STATISTICS
THE UNIVERSITY OF CHICAGO
CHICAGO, ILLINOIS 60637