Statistical Science
2004, Vol. 19, No. 1, 58–80
DOI 10.1214/088342304000000116

The Interplay of Bayesian and Frequentist Analysis

M. J. Bayarri and J. O. Berger

Abstract. Statistics has struggled for nearly a century over the issue of whether the Bayesian or frequentist paradigm is superior. This debate is far from over and, indeed, should continue, since there are fundamental philosophical and pedagogical issues at stake. At the methodological level, however, the debate has become considerably muted, with the recognition that each approach has a great deal to contribute to statistical practice and each is actually essential for full development of the other approach. In this article, we embark upon a rather idiosyncratic walk through some of these issues.

Key words and phrases: Admissibility, Bayesian model checking, conditional frequentist, confidence intervals, consistency, coverage, design, hierarchical models, nonparametric Bayes, objective Bayesian methods, p-values, reference priors, testing.

CONTENTS

M. J. Bayarri is Professor, Department of Statistics, University of Valencia, Av. Dr. Moliner 50, 46100 Burjassot, Valencia, Spain (e-mail: susie.bayarri@uv.es). J. O. Berger is Director, Statistical and Applied Mathematical Sciences Institute, and Professor, Duke University, P.O. Box 90251, Durham, North Carolina 27708-0251, USA (e-mail: berger@stat.duke.edu).

1. INTRODUCTION

Statisticians should readily use both Bayesian and frequentist ideas. In Section 2 we discuss situations in which simultaneous frequentist and Bayesian thinking is essentially required. For the most part, however, the situations we discuss are situations in which it is simply extremely useful for Bayesians to use frequentist methodology or frequentists to use Bayesian methodology.

The most common scenarios of useful connections between frequentists and Bayesians are when no external information (other than the data and model itself) is to be introduced into the analysis—on the Bayesian side, when "objective prior distributions" are used. Frequentists are usually not interested in subjective, informative priors, and Bayesians are less likely to be interested in frequentist evaluations when using subjective, highly informative priors.

We will, for the most part, avoid the question of whether the Bayesian or frequentist approach to statistics is "philosophically correct." While this is a valid question, and research in this direction can be of

fundamental importance, the focus here is simply on methodology. In a related vein, we avoid the question of what is "pedagogically correct." If pressed, we would probably argue that Bayesian statistics (with emphasis on objective Bayesian methodology) should be the type of statistics that is taught to the masses, with frequentist statistics being taught primarily to advanced statisticians, but that is not an issue for this paper.

Several caveats are in order. First, we primarily focus on the Bayesian and frequentist approaches here; these are the most generally applicable and accepted statistical philosophies, and both have features that are compelling to most statisticians. Other statistical schools, such as the *likelihood* school (see, e.g., Reid, 2000), have many attractive features and vocal proponents, but have not been as extensively developed or utilized as the frequentist and Bayesian approaches.

A second caveat is that the selection of topics here is rather idiosyncratic, being primarily based on situations and examples in which we are currently interested. Other Bayesian–frequentist synthesis works (e.g., Pratt, 1965; Barnett, 1982; Rubin, 1984; and even Berger, 1985a) focus on a quite different set of situations. Furthermore, we almost completely ignore many of the most time-honored Bayesian–frequentist synthesis topics, such as empirical Bayes analysis. Hence, rather than being viewed as a comprehensive review, this paper should be thought of more as a personal view of current interesting issues in the Bayesian–frequentist synthesis.

2. INHERENTLY JOINT BAYESIAN–FREQUENTIST SITUATIONS

There are certain statistical scenarios in which a joint frequentist–Bayesian approach is arguably required. As illustrations of this, we first discuss the issue of design—in which the notion should not be controversial—and then discuss the basic meaning of frequentism, which arguably should be (but is not typically perceived as) a joint frequentist–Bayesian endeavor.

2.1 Design or Preposterior Analysis

Frequentist design focuses on planning of experiments—for instance, the issue of choosing an appropriate sample size. In Bayesian analysis this is often called *preposterior analysis*, because it is done before the data is collected (and, hence, before the posterior distribution is available).

EXAMPLE 2.1. Suppose $X_1, \ldots, X_n$ are i.i.d. Poisson random variables with mean θ, and that it is desired to estimate θ under the weighted squared error loss $(\hat{\theta} - \theta)^2/\sqrt{\theta}$ and using the classical estimator $\hat{\theta} = \bar{X}$. This estimator has frequentist expected loss $\mathrm{E}_\theta[(\bar{X} - \theta)^2/\sqrt{\theta}] = \sqrt{\theta}/n$.

A typical design problem would be to choose the sample size n so that the expected loss is less than some prespecified limit C. (An alternative formulation might be to minimize $C + nc$, where c is the cost of an observation, but this would not significantly alter the discussion here.) This is clearly not possible, for all θ; hence we must bring prior knowledge about θ into play.

A primitive recommendation that one often sees, in such situations, is to make a "best guess" θ_0 for θ, and then choose n so that $\sqrt{\theta_0}/n \le C$; that is, choose $n \approx \sqrt{\theta_0}/C$. This is needlessly dogmatic, in that one rarely believes particularly strongly in a particular value θ_0.

A common primitive recommendation in the opposite direction is to choose an upper bound θ_U for θ, and then choose n so that $\sqrt{\theta_U}/n \le C$; that is, choose $n \approx \sqrt{\theta_U}/C$. This is needlessly conservative, in that the resulting n will typically be much larger than needed.

The Bayesian approach to the design question is to elicit a subjective prior distribution $\pi(\theta)$ for θ, and then to choose n so that $\int \frac{\sqrt{\theta}}{n}\pi(\theta)\,d\theta \le C$; that is, choose $n \approx \int \sqrt{\theta}\pi(\theta)\,d\theta/C$. This is a reasonable compromise between the above two extremes and will typically result in the most reasonable values of n.

Classical design texts often focus on the very special situations in which the design criterion is constant in the unknown model parameter θ, and hence fail to clarify the philosophical centrality of Bayesian issues in design. The basic fact is that, before experimentation, one knows neither the data nor θ, and so expectations over both (i.e., both frequentist and Bayesian expectations) are needed for design. See Chaloner and Verdinelli (1995) and Dawid and Sebastiani (1999).

A very common situation in which design evaluation is not constant is classical testing, in which the sample size is often chosen to achieve a given power at a specified value θ' of the parameter under the alternative hypothesis. Again, specifying a specific θ' is very crude when viewed from a Bayesian perspective. Far more reasonable for a classical tester would be to specify a prior distribution for θ under the alternative, and consider the average power with respect to this

distribution. (More controversial would be to consider an average Type I error.)

2.2 The Meaning of Frequentism

There is a sense in which essentially everyone should ascribe to frequentism:

FREQUENTIST PRINCIPLE. In repeated practical use of a statistical procedure, the long-run average actual accuracy should not be less than (and ideally should equal) the long-run average reported accuracy.

This version of the frequentist principle is actually a joint frequentist–Bayesian principle. Suppose, for instance, that we decide it is relevant to statistical practice to *repeatedly* use a particular statistical model and procedure—for instance, a 95% classical confidence interval for a normal mean. This procedure will, in practice, be used on a series of different problems involving a series of different normal means with a corresponding series of data. Hence, in evaluating the procedure, we should simultaneously be averaging over the differing means and data.

This is in contrast to textbook statements of the frequentist principle which tend to focus on fixing the value of, say, the normal mean, and *imagining* repeatedly drawing data from the given model and utilizing the confidence procedure repeatedly on this data. The word *imagining* is emphasized, because this is solely a thought experiment. What is done in practice is to use the confidence procedure on a series of different problems—not use the confidence procedure for a series of repetitions of the *same* problem with different data (which would typically make no sense in practice).

Neyman himself repeatedly pointed out (see, e.g., Neyman, 1977) that the motivation for the frequentist principle is in its use on differing real problems, and not imaginary repetitions for one problem with a fixed "true parameter." Of course, the reason textbooks typically give the latter (philosophically misleading) version is because of the convenient mathematical fact that if, say, a confidence procedure has 95% frequentist coverage for each fixed parameter value, then it will necessarily also have 95% coverage when used repeatedly on a series of differing problems. Thus (as with design), whenever the frequentist evaluation is constant over the parameter space, one does not need to also do a Bayesian average over the parameter space; but, conceptually, it is the combined frequentist–Bayesian average that is practically relevant.

The impact of this "real" frequentist principle thus arises when the frequentist evaluation of a procedure is not constant over the parameter space. Here is an example.

EXAMPLE 2.2. *Binomial confidence interval.* Brown, Cai and DasGupta (2001, 2002) considered the problem of observing $X \sim \text{Binomial}(n, \theta)$ and determining a 95% confidence interval for the unknown success probability θ. We consider here the special case of $n = 50$, and two confidence procedures. The first is $C^J(x)$, defined as the "Jeffreys equal-tailed 95% confidence interval," given by

$$C^J(x) = \big(q_{0.025}(x), q_{0.975}(x)\big), \tag{2.1}$$

where $q_\alpha(x)$ is the αth-quantile of the Beta$(x + 0.5, 50.5 - x)$ distribution. The second confidence procedure we consider is the "modified Jeffreys equal-tailed 95% confidence interval," given by

$$C^{J*}(x) = \begin{cases} \big(q_{0.025}(x), q_{0.975}(x)\big), & \text{if } x \neq 0 \\ & \text{and } x \neq n, \\ \big(0, q_{0.975}(x)\big), & \text{if } x = 0, \\ \big(q_{0.025}(x), 1\big), & \text{if } x = n. \end{cases} \tag{2.2}$$

For the moment, simply consider these as formulae for confidence intervals; we later discuss their motivation.

Brown, Cai and Dasgupta (2001) provide the graph of the coverage probability of C^{J*} given in Figure 1. Note that, while roughly close to the target 95%, the coverage probability varies considerably as a function of θ, going from a high of 1 at $\theta = 0$ and $\theta = 1$ to a low of 0.884 at $\theta = 0.049$ and $\theta = 0.951$. A "textbook frequentist" might then assert that this is only an 88.4% confidence procedure, since the coverage cannot be guaranteed to be higher than this limit. But would the "practical frequentist" agree with this?

The practical frequentist evaluates how C^{J*} would work for a sequence $\{\theta_1, \theta_2, \ldots, \theta_m\}$ of parameters (and corresponding data) encountered in a series of real problems. If m is large, the law of large numbers guarantees that the coverage that is actually experienced will be the average of the coverages obtained over the sequence of problems. Thus we should be considering averages of the coverage in Figure 1 over sequences of θ_j.

One could, of course, choose the sequence of θ_j to all be 0.049 and/or 0.951, but this is not very realistic. One might consider global averages with respect to sequences generated from prior distributions $\pi(\theta)$, but a "practical frequentist" presumably does not want to

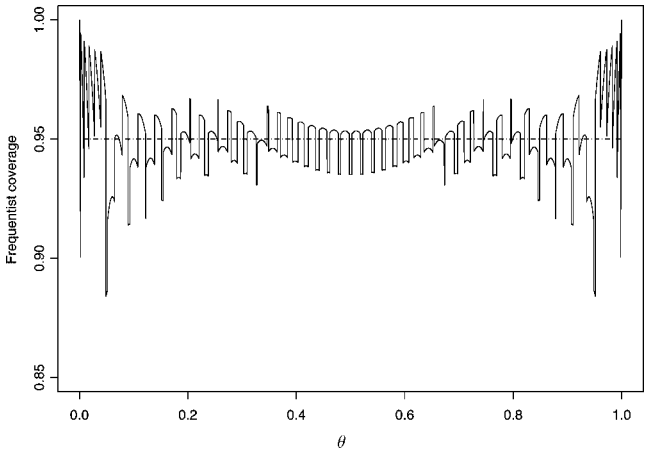

FIG. 1. *Frequentist coverage of the C^J* intervals, as a function of θ when $n = 50$.*

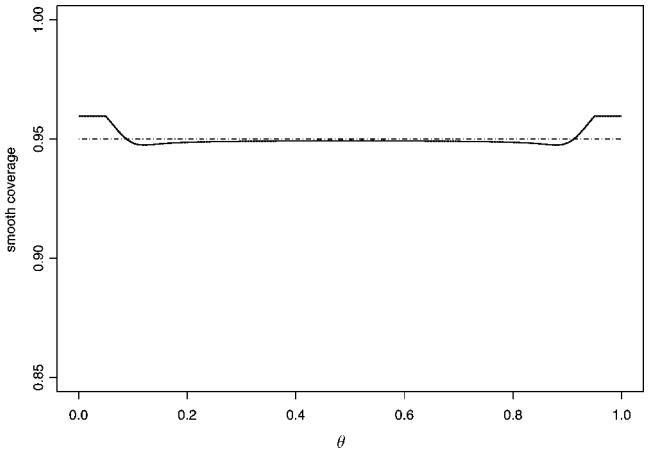

FIG. 2. *Local average coverage of the C^J* intervals, as a function of θ when $n = 50$ and $\varepsilon = 0.05$.*

spend much time thinking about prior distributions. One plausible solution is to look at *local average coverage*, defined via a local smoothing of the binomial coverage function. A convenient computational kernel for this problem, when smoothing at a point θ is desired and when a smoothing kernel having standard deviation ε is desired (so that $\theta \pm 2\varepsilon$ can roughly be thought of as the range over which the smoothing is performed), is the Beta$(a(\theta), a(1-\theta))$ distribution $k_{\varepsilon,\theta}(\cdot)$ where

$$a(\theta) = \begin{cases} 1 - 2\varepsilon, & \text{if } \theta \le \varepsilon, \\ [\theta(1-\theta)\varepsilon^{-2} - 1]\theta, & \\ & \text{if } \varepsilon < \theta < 1 - \varepsilon, \\ \dfrac{1}{\varepsilon} - 3 + 2\varepsilon, & \text{if } \theta \ge 1 - \varepsilon. \end{cases} \tag{2.3}$$

Writing the standard frequentist coverage as $1 - \alpha(\theta)$, this leads to the *ε-local average coverage*

$$\begin{aligned} 1 - \alpha_\varepsilon(\theta) &= \int_0^1 [1 - \alpha(\theta)] k_{\varepsilon,\theta}(\theta)\, d\theta \\ &= \sum_{x=0}^{n} \binom{n}{x} \big[\Gamma\big(a(\theta) + a(1-\theta)\big) \\ &\qquad \cdot \Gamma\big(a(\theta) + x\big) \\ &\qquad \cdot \Gamma\big(a(1-\theta) + n - x\big)\big] \\ &\qquad \cdot \big[\Gamma(a(\theta))\Gamma\big(a(1-\theta)\big) \\ &\qquad \cdot \Gamma\big(a(\theta) + a(1-\theta) + n\big)\big]^{-1} \\ &\quad \cdot \int_{C^J*} \mathrm{Beta}\big(\theta | a(\theta), a(1-\theta)\big)\, d\theta, \end{aligned}$$

the last equation following from the standard expression for the beta–binomial predictive distribution [and with Beta$(\theta | a(\theta), a(1-\theta))$ denoting the beta density with given parameters].

For the binomial example we are considering, this is graphed in Figure 2, for $\varepsilon = 0.05$. (Such a value of ε could be interpreted as implying that one is sure that the sequence of practical problems, for which the binomial confidence interval will be used, has θ_j varying by at least ± 0.05.) Note that this local average coverage is always close to 0.95, so that a practical frequentist would be quite pleased with the confidence interval.

One could imagine a textbook frequentist arguing that, sometimes, a particular value, such as $\theta = 0.049$, could be of special interest in repeated investigations, the value perhaps corresponding to some important physical theory concerning θ that science will repeatedly investigate. In such a situation, however, it is arguably not appropriate to utilize confidence intervals; that there is a special value of θ of interest should be acknowledged via some type of testing procedure. Even if there were a distinguished value of θ and it was erroneously handled by finding a confidence interval, the practical frequentist has one more arrow in his or her quiver: it is not likely that a series of experiments investigating this particular physical theory would all choose the same sample size, so one should consider "practical averaging" over sample size. For instance, suppose sample sizes would vary between 40 and 60 for the binomial problem we have been considering. Then one could reasonably consider average coverage over these sample sizes, the result of which is given in Figure 3. While not always as close to 0.95 as was the local average coverage, it would still strike most people as reasonable to call C^J* a 95% confidence interval when averaged over reasonable sample sizes.

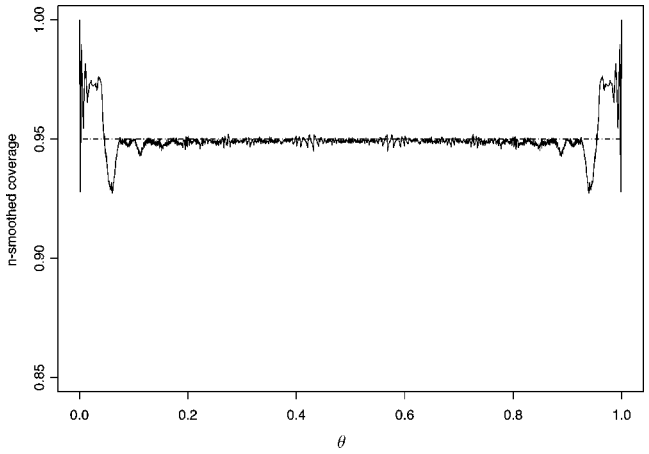

FIG. 3. *Average coverage over n between* 40 *and* 60 *of the C^{J*} intervals, as a function of θ.*

A similar idea concerning "local averages" of frequentist properties was employed by Woodroofe (1986), who called the concept "very weak expansions." Brown, Cai and DasGupta (2002), for the binomial problem, considered the average coverage defined as the smooth part of their asymptotic expansion of coverage, yielding a result similar to that in Figure 2. Rousseau (2000) took a different approach, considering slight adjustment of the Bayesian intervals through randomization to achieve the correct frequentist coverage.

So far the discussion has been in terms of the practical frequentist acknowledging the importance of considering averages over θ. We would also claim, however, that Bayesians should ascribe to the above version of the frequentist principle. If (say) a Bayesian were to repeatedly construct purported 90% credible intervals in his or her practical work, yet they only contained the unknowns about 70% of the time, something would be seriously wrong. A Bayesian might feel that the practical frequentist principle will automatically be satisfied if he or she does a good Bayesian job of separately analyzing each individual problem, and hence that it is not necessary to specifically worry about the principle, but that does not mean that the principle is invalid.

EXAMPLE 2.3. In this regard, let us return to the binomial example to discuss the origin of the confidence intervals $C^J(x)$ and $C^{J*}(x)$. The intervals $C^J(x)$ arise as the Bayesian equal-tailed credible sets obtained from use of the Jeffreys prior (see Jeffreys, 1961) $\pi(\theta) \propto \theta^{-1/2}(1-\theta)^{-1/2}$ for θ. (In particular, the intervals are formed by the upper and lower $\alpha/2$-quantiles of the resulting posterior distribution for θ.) This is the prior that is customary for an objective Bayesian to use for the binomial problem. (See Section 3.4.3 for further discussion of the Jeffreys prior.) Note that, because of the derivation of the credible set from the objective Bayesian perspective, there is strong reason to believe that "conditionally on the given situation and data, the accuracy assignment of 95% is reasonable." See Section 3.2.2 for discussion of conditional performance.

The frequentist coverage of the intervals $C^J(x)$ is given in Figure 4. A pure frequentist might well be concerned with the raw coverage of this credible interval because it goes to zero at $\theta = 0$ and $\theta = 1$. A moment's reflection reveals why this is the case: the equal-tailed Bayesian credible intervals purposely exclude values in the left and right tails of the posterior distribution and, hence, will always exclude $\theta = 0$ and $\theta = 1$. The modification of this interval employed in Brown, Cai and DasGupta (2001) is $C^{J*}(x)$ in (2.2): for the observations $x = 0$ or $x = n$, one simply extends the Jeffreys equal-tailed credible intervals to include 0 or 1. Of course, from a conditional Bayesian perspective, these intervals then have posterior probability 0.975, so a Bayesian would no longer call them 95% credible intervals.

While the raw frequentist coverage of the Jeffreys equal-tailed credible intervals might seem unappealing, their 0.05-local average coverage is excellent, virtually the same as that in Figure 2 for the modified interval; indeed, the difference is not visually apparent, so that we do not separately include a graph of this local average coverage. Hence the practical frequentist would be quite happy with use of $C^J(x)$, even if it has low coverage right at the endpoints.

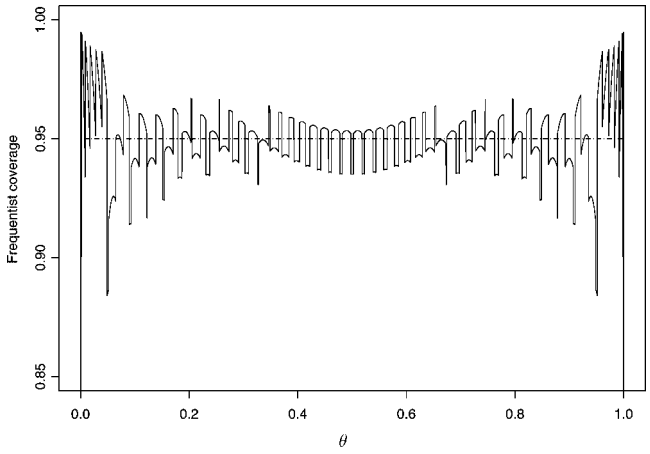

FIG. 4. *Coverage of the C^J intervals, as a function of θ when $n = 50$.*

The issue of Bayesians achieving good pure frequentist coverage near a finite boundary of a parameter space is an interesting issue; our guess is that this is often not possible. In the above example, for instance, whether a Bayesian includes, or excludes, $\theta = 0$ or $\theta = 1$ in a credible interval is rather arbitrary and will depend on, for example, a choice such as that between an equal-tailed or highest posterior density (HPD) interval. (The HPD intervals for $x = 0$ and $x = n$ would include $\theta = 0$ and $\theta = 1$, respectively.) Furthermore, this choice will typically lead to either 0 frequentist coverage or coverage of 1 at the endpoints, unless something unnatural to a Bayesian, such as randomization, were incorporated. Hence the recognition of the centrality to frequentist practice of some type of average coverage, rather than pointwise coverage, can be important in such problems to achieve simultaneously acceptable Bayesian and frequentist performance.

2.3 Empirical Bayes, Gamma Minimax, Restricted Risk Bayes

Several approaches to statistical analysis have been proposed which are inherently a mixture of Bayesian and frequentist analysis. These approaches have lengthy histories and extensive literature and we so we can do little more here than simply give pointers to the areas.

Robbins (1955) introduced the *empirical Bayes* approach, in which one specifies a class of prior distributions Γ, but assumes that the prior is otherwise unknown. The data is then used to help determine the prior and/or to directly find the optimal Bayesian answer. Frequentist reasoning was intimately involved in Robbins' original formulation of empirical Bayes, and in significant implementations of the paradigm, such as Morris (1983) for hierarchical models. More recently, the name empirical Bayes is often used in association with approximate Bayesian analyses which do not specifically involve frequentist measures. (Simply using a maximum likelihood estimate of a hyperparameter does not make a technique frequentist.) For modern reviews of empirical Bayes analysis and previous references, see Carlin and Louis (2000) and Robert (2001).

In the *gamma minimax* approach, one again has a class Γ of possible prior distributions and considers the frequentist Bayes risk (the expected loss over both the data and unknown parameters) of the Bayes procedure for priors in the class. One then chooses that prior which minimizes this frequentist Bayes risk. For examples and references, see Berger (1985a) and Vidakovic (2000).

In the *restricted risk Bayes* approach, one has a single prior distribution, but can only consider statistical procedures whose frequentist risk (expected loss) is constrained in some fashion. The idea is that one can utilize the prior information, but in a way that will be guaranteed to be acceptable to the frequentist who wants to limit frequentist risk. (See Berger, 1985a, for discussion and earlier references.) This approach is actually not "inherently Bayesian–frequentist," but is more what could be termed a "hybrid" approach, in the sense that it seeks some type of formal compromise between Bayesian and frequentist positions. There have been many other attempts at such compromises, but none has seemed to significantly affect statistical practice.

There are many other important areas in which joint frequentist–Bayesian evaluation is used. Some were even developed primarily from the Bayesian perspective, such as the *prequential approach* of Dawid (cf. Dawid and Vovk, 1999).

3. ESTIMATION AND CONFIDENCE INTERVALS

In statistical estimation (including development of confidence intervals), objective Bayesian and frequentist methods often give similar (or even identical) answers in standard parametric problems with continuous parameters. The standard normal linear model is the prototypical example: frequentist estimates and confidence intervals coincide exactly with the standard objective Bayesian estimates and credible intervals. Indeed, this occurs more generally in situations that exhibit an "invariance structure," provided objective Bayesians use the "right-Haar prior density"; see Berger (1985a), Eaton (1989) and Robert (2001) for discussion and earlier references.

This dual frequentist–Bayesian interpretation of many textbook estimation procedures has a number of important implications, not the least of which is that much of standard textbook statistical methodology (and standard software) can alternatively be presented and described from the objective Bayesian perspective. In particular, one can teach much of elementary statistics from this alternative perspective, without changing the procedures that are taught.

In more complicated situations, it is still usually possible to achieve near-agreement between frequentist and Bayesian estimation procedures, although this may require careful utilization of the tools of both. A number of situations requiring such cross-utilization of tools are discussed in this section.

3.1 Computation with Hierarchical, Multilevel or Mixed Model Analysis

With the advent of Gibbs sampling and other Markov chain Monte Carlo (MCMC) methods of analysis (cf. Robert and Casella, 1999), it has become relatively standard to deal with models that go under any of the names listed in the above title as Bayesian methods. This popularity of the Bayesian methods is not necessarily because of their intrinsic virtues, but rather because the Bayesian computation is now much easier than computation via more classical routes. See Hobert (2000) for an overview and other references.

On the other hand, any MCMC method relies fundamentally on frequentist reasoning to do the computation. An MCMC method generates a sequence of simulated values $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_m$ of an unknown quantity $\boldsymbol{\theta}$, and then relies upon a law of large numbers or ergodic theorem (both frequentist) to assert that $\bar{\boldsymbol{\theta}}_m = \frac{1}{m}\sum_{i=1}^{m} \boldsymbol{\theta}_i \to \boldsymbol{\theta}$. Furthermore, diagnostics for MCMC convergence are almost universally based on frequentist tools. There is a purely Bayesian way of looking at such computation problems, which goes under the heading "Bayesian numerical analysis" (cf. Diaconis, 1988a; O'Hagan, 1992), but in practice it is typically much simpler to utilize the frequentist reasoning.

In conclusion for much of modern statistical analysis in hierarchical models, we already see an inseparable joining of Bayesian and frequentist methodology.

3.2 Assessment of Accuracy of Estimation

Frequentist methodology for point estimation of unknown model parameters is relatively straightforward and successful. However, assessing the accuracy of the estimates is considerably more challenging and is a problem for which frequentists should draw heavily on Bayesian methodology.

3.2.1 *Finding good confidence intervals in the presence of nuisance parameters.* Confidence intervals for a model parameter are a common way of indicating the accuracy of an estimate of the parameter. Finding good confidence intervals when there are nuisance parameters is very challenging within the frequentist paradigm, unless one utilizes objective Bayesian methodology, in which case the frequentist problem becomes relatively straightforward. Indeed, here is a rather general prescription for finding confidence intervals using objective Bayesian methods:

- Begin with a "reasonable" objective prior distribution. (See Section 3.4 for discussion of objective priors, and note that a "reasonable" objective prior may well depend on which parameter is the parameter of interest.)
- By simulation, obtain a (large) sample from the posterior distribution of the parameter of interest:
  Option 1. If a predetermined confidence interval $C(X)$ is of interest, simply approximate the posterior probability of the interval by the fraction of the samples from the posterior distribution that fall in the interval.
  Option 2. If the confidence interval is not predetermined, find the $\alpha/2$ upper and lower fractiles of the posterior sample; the interval between these fractiles approximates the $100(1-\alpha)\%$ equal-tailed posterior credible interval for the parameter of interest. (Alternative forms for the confidence set can be considered, but the equal-tailed interval is fine for most applications.)
- Assert that the obtained interval is the frequentist confidence interval, having frequentist coverage given by the posterior probability of the interval.

There is a large body of theory, discussed in Section 3.4, as well as considerable practical experience, supporting the validity of constructing frequentist confidence intervals in this way. Here is one example from the "practical experience" side.

EXAMPLE 3.1. *Medical diagnosis* (*Mossman and Berger,* 2001). Within a population for which $p_0 = \Pr(\text{Disease } D)$, a diagnostic test results in either a Positive $(+)$ or Negative $(-)$ reading. Let $p_1 = \Pr(+|\text{patient has } D)$ and $p_2 = \Pr(+|\text{patient does not have } D)$. By Bayes' theorem,

$$\theta = \Pr(D|+) = \frac{p_0 p_1}{p_0 p_1 + (1-p_0)p_2}.$$

In practice, the p_i are typically unknown, but for $i = 0, 1, 2$ there are available (independent) data x_i having Binomial(n_i, p_i) densities. It is desired to find a $100(1-\alpha)\%$ confidence set for θ that has good conditional and frequentist properties.

A simple objective Bayesian approach to this problem is to utilize the Jeffreys priors $\pi(p_i) \propto p_i^{-1/2} \cdot (1-p_i)^{-1/2}$ for each of the p_i, and compute the $100(1-\alpha)\%$ equal-tailed posterior credible interval for θ. A suitable implementation of the algorithm presented above is as follows:

- Draw random p_i from the Beta$(x_i + \frac{1}{2}, n_i - x_i + \frac{1}{2})$ posterior distributions, $i = 0, 1, 2$.
- Compute the associated

$$\theta = \frac{p_0\, p_1}{p_0\, p_1 + (1-p_0)\, p_2}$$

for each random triplet.
- Repeat this process 10,000 times.
- The $\alpha/2$ and $1-\alpha/2$ fractiles of these 10,000 generated θ form the desired confidence interval. [In other words, simply order the 10,000 values of θ, and let the confidence interval be the interval between the $(10{,}000 \times \frac{\alpha}{2})$th and $(10{,}000 \times \frac{1-\alpha}{2})$th values.]

The proposed objective Bayesian procedure is clearly simple to use, but is the resulting confidence interval a satisfactory frequentist interval? To provide perspective on this question, note that the above problem has also been studied in the frequentist literature, using standard log-odds and delta-method procedures to develop confidence intervals, as well as more sophisticated approaches such as the Gart–Nam (Gart and Nam, 1988) procedure. For a description of these classical methods, as applied to this problem of medical diagnosis, see Mossman and Berger (2001).

Table 1 gives an indication of the frequentist performance of the confidence intervals developed by these four methods. It is based on a simulation that repeatedly generates data from binomial distributions with sample sizes $n_i = 20$ and the indicated values of the parameters (p_0, p_1, p_2). For each generated triplet of data in the simulation, the 95% confidence interval is computed using the objective Bayesian algorithm or one of the three classical methods. It is then noted whether the computed interval contains the true θ, or misses to the left or right. The entries in the table are the long run proportion of misses to the left or right. Ideally, these proportions should be 0.025 and, at the least, their sum should be 0.05.

Clearly the objective Bayes interval has quite good frequentist performance, better than any of the classically derived confidence intervals. Furthermore, it can be seen that the objective Bayes intervals are, on average, smaller than the classically derived intervals. (See Mossman and Berger, 2001, for these and more extensive computations.) Finally, the objective Bayes confidence intervals were the simplest to derive and will automatically be conditionally appropriate (see Section 3.2.2), because of their Bayesian derivation.

The finding in the above example, that objective Bayesian analysis very easily provides small confidence sets with excellent frequentist coverage, has been repeatedly shown to happen. See Section 3.4 for additional discussion.

3.2.2 *Obtaining good conditional measures of accuracy.* Developing frequentist confidence intervals using the Bayesian approach automatically provides an additional significant benefit: the confidence statement will be conditionally appropriate. Here is a simple artificial example.

EXAMPLE 3.2. Two observations, X_1 and X_2, are to be taken, where

$$X_i = \begin{cases} \theta + 1, & \text{with probability } 1/2, \\ \theta - 1, & \text{with probability } 1/2. \end{cases}$$

Consider the confidence set for the unknown $\theta \in \Re$,

$$C(X_1, X_2) = \begin{cases} \text{the point } \{\frac{1}{2}(X_1 + X_2)\}, & \text{if } X_1 \neq X_2, \\ \text{the point } \{X_1 - 1\}, & \text{if } X_1 = X_2. \end{cases}$$

The frequentist coverage of this confidence set can easily be shown to be $P_\theta(C(X_1, X_2) \text{ contains } \theta) = 0.75$. This is not at all a sensible report, once the data is at hand. To see this, observe that, if $x_1 \neq x_2$, then we know for sure that their average is equal to θ, so that the confidence set is then actually 100% accurate. On the other hand, if $x_1 = x_2$, we do not know if θ is the data's common value plus 1 or their common value minus 1, and each of these possibilities is equally likely to have occurred.

To obtain sensible frequentist answers here, one must define the conditioning statistic $S = |X_1 - X_2|$, which can be thought of as measuring the "strength of evidence" in the data ($S = 2$ reflecting data with maximal evidential content and $S = 0$ being data of minimal

TABLE 1
The probability that the nominal 95% *interval misses the true* θ *on the left and on the right, for the indicated parameter values and when* $n_0 = n_1 = n_2 = 20$

(p_0, p_1, p_2)	O-Bayes	Log odds	Gart–Nam	Delta
$(\frac{1}{4}, \frac{3}{4}, \frac{1}{4})$	0.0286, 0.0271	0.0153, 0.0155	0.0277, 0.0257	0.0268, 0.0245
$(\frac{1}{10}, \frac{9}{10}, \frac{1}{10})$	0.0223, 0.0247	0.0017, 0.0003	0.0158, 0.0214	0.0083, 0.0041
$(\frac{1}{2}, \frac{9}{10}, \frac{1}{10})$	0.0281, 0.0240	0.0004, 0.0440	0.0240, 0.0212	0.0125, 0.0191

evidential content). Then one defines frequentist coverage conditional on the strength of evidence S. For the example, an easy computation shows that this conditional confidence equals

$$P_\theta\big(C(X_1, X_2) \text{ contains } \theta | S = 2\big) = 1,$$
$$P_\theta\big(C(X_1, X_2) \text{ contains } \theta | S = 0\big) = \tfrac{1}{2},$$

for the two distinct cases, which are the intuitively correct answers.

It is important to realize that conditional frequentist measures are fully frequentist and (to most people) clearly better than unconditional frequentist measures. They have the same unconditional property (e.g., in the above example one will report 100% confidence half the time, and 50% confidence half the time, resulting in an "average" of 75% confidence, as must be the case for a frequentist measure), yet give much better indications of the accuracy for the type of data that one has actually encountered.

In the above example, finding the appropriate conditioning statistic was easy but, in more involved situations, it can be a challenging undertaking. Luckily, intervals developed via the Bayesian approach will automatically condition appropriately. For instance, in the above example, the objective Bayesian approach assigns θ the standard objective prior (for a location parameter) $\pi(\theta) = 1$, from which is easy to compute that the posterior probability assigned to the set $C(X_1, X_2)$ is 1 or 0.5 as the observations differ or are the same. (This is essentially Option 1 of the algorithm described at the beginning of the Section 3.2.1, although here the posterior probabilities can be computed analytically.)

General theory about conditional confidence can be found in Kiefer (1977); see also Robinson (1979), Berger (1985b), Berger and Wolpert (1988), Casella (1988) and Lehmann and Casella (1998). In Section 4.1, we will return to this dual theme that (i) it is crucial for frequentists to condition appropriately; (ii) this is technically most easily accomplished by using Bayesian tools.

3.2.3 *Accuracy assessment in hierarchical models.* As mentioned earlier, the utilization of hierarchical or random effects or mixed or multilevel models has increasingly taken a Bayesian flavor in practice, in part driven by the computational advantages of Gibbs sampling and MCMC analysis. Another reason for this greatly increasing utilization of the Bayesian approach to such problems is that practitioners are finding the inferences that arise from the Bayesian approach to be considerably more realistic than those from competitors, such as various versions of maximum likelihood estimation (or empirical Bayes estimation) or (often worse) unbiased estimation.

One of the potentially severe problems with the maximum likelihood or empirical Bayes approach is that maximum likelihood estimates of variances in hierarchical models (or variance component models) can easily be zero, especially when there are numerous variances in the model that are being estimated. (Unbiased estimation will be even worse in such situations; if the mle is zero, the unbiased estimate will be negative.)

EXAMPLE 3.3. Suppose, for $i = 1, \ldots, p$, that $X_i \sim \text{Normal}(\mu_i, 1)$ and $\mu_i \sim \text{Normal}(0, \tau^2)$, all random variables being independent. Then, marginally, $X_i \sim \text{Normal}(0, 1 + \tau^2)$, so that the likelihood function of τ^2 can be written

$$L(\tau^2) \propto \frac{1}{(1+\tau^2)^{p/2}} \exp\left\{-\frac{S^2}{2(1+\tau^2)}\right\}, \tag{3.1}$$

where $S^2 = \sum X_i^2$. The mle for τ^2 is easily calculated to be $\hat{\tau}^2 = \max\{0, \frac{S^2}{p} - 1\}$. Thus, if $S^2 < p$, the mle would be $\hat{\tau}^2 = 0$ (and the unbiased estimate would be negative). While a value of $S^2 < p$ is somewhat unusual here [if, e.g., $p = 4$ and $\tau^2 = 1$, then $\Pr(S^2 < p) = 0.264$], it is quite common in problems with numerous variance components to have at least one mle variance estimate equal to 0.

For $p = 4$ and $S^2 = 4$, the likelihood function in (3.1) is graphed in Figure 5. While $L(\tau^2)$ is decreasing away from 0, it does not decrease particularly

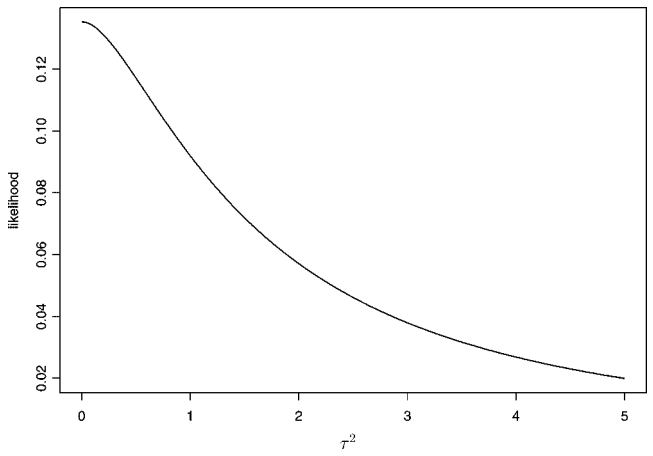

FIG. 5. *Likelihood function of τ^2 when $p = 4$ and $S^2 = 4$ is observed.*

quickly, clearly indicating that there is considerable uncertainty as to the true value of τ^2 even though the mle is 0.

Utilizing an mle of 0 as a variance estimate can be quite dangerous, because it will typically affect the ensuing analysis in an incorrectly aggressive fashion. In the above example, for instance, setting τ^2 to 0 is equivalent to stating that all the μ_i are exactly equal to each other. This is clearly unreasonable in light of the fact that there is actually great uncertainty about τ^2, as reflected in Figure 5. Since the likelihood maximum is occurring at the boundary of the parameter space, it is also very difficult to utilize likelihood or frequentist methods to attempt to incorporate uncertainty about τ^2 into the analysis.

None of these difficulties arises in the Bayesian approach, and the vague nature of the information in the data about such variances will be clearly reflected in the posterior distribution. For instance, if one were to use the constant prior density $\pi(\tau^2) = 1$ in the above example, the posterior density would be proportional to the likelihood in Figure 5, and the significant uncertainty about τ^2 would permeate the analysis.

3.3 Foundations, Minimaxity and Exchangeability

There are numerous ties between frequentist and Bayesian analysis at the foundational level. The foundation of frequentist statistics typically focuses on the class of "optimal" procedures in a given situation, called a *complete class* of procedures. Through the work of Wald (1950) and others, it has long been known that a *complete class* of procedures is identical to the class of Bayes procedures or certain limits thereof. Furthermore, in proving frequentist optimality of a procedure, it is typically necessary to employ Bayesian tools. (See Berger, 1985a; Robert, 2001, for many examples and references.) Hence, at a fundamental level, the frequentist paradigm is intertwined with the Bayesian paradigm.

Interestingly, this fundamental duality has not had a pronounced effect on the Bayesian versus frequentist debate. In part, this is because many frequentists find the search for optimal frequentist procedures to be of limited practical utility (since such searches usually take place in rather limited settings, from the perspective of practice), and hence do not themselves pursue optimality and thereby come into contact with the Bayesian equivalence. Even among frequentists who are significantly concerned with optimality, it is typically perceived that the relationship with Bayesian analysis is a nice mathematical coincidence that can be used to eliminate inferior frequentist procedures, but that Bayesian ideas should not form the basis for choice among acceptable frequentist procedures. Still, the complete class theorems provide a powerful underlying link between frequentist and Bayesian statistics.

One of the most prominent frequentist principles for choosing a statistical procedure is that of *minimaxity*; see Brown (1994, 2000) and Strawderman (2000) for reviews on the important impact of this concept on statistics. Bayesian analysis again provides the most useful tool for deriving minimax procedures: one finds the "least favorable prior distribution," and the minimax procedure is the resulting Bayes rule.

To many Bayesians, the most compelling foundation of statistics is that based on exchangeability, as developed in de Finetti (1970). From the assumption of exchangeability of an infinite sequence, $X_1, X_2, \ldots$, of observations (essentially the assumption that the distribution of the sequence remains the same under permutation of the coordinates), one can sometimes deduce the existence of a particular statistical model, with unknown parameters, and a prior distribution on the parameters. By considering an infinite series of observations, frequentist reasoning—or at least frequentist mathematics—is clearly involved. Reviews of more recent developments and other references can be found in Diaconis (1988b) and Lad (1996).

There are many other foundational arguments that begin with axioms of rational behavior and lead to the conclusion that some type of Bayesian behavior is implied. (See Bernardo and Smith, 1994, for review and references.) Many of these effectively involve simultaneous frequentist–Bayesian evaluations of outcomes, such as Rubin (1987), which is perhaps the weakest set of axioms that implies Bayesian behavior.

3.4 Use of Frequentist Methodology in Prior Development

In principle, a subjective Bayesian need not worry about frequentist ideas—if a prior distribution is elicited and accurately reflects prior beliefs, then Bayes' theorem guarantees that any resulting inference will be optimal. The hitch is that it is not very common to have a prior distribution that accurately reflects all prior beliefs. Suppose, for instance, that the only unknown model parameter is a normal mean θ. Complete assessment of the prior distribution for θ involves an infinite number of judgments [e.g., specification of the probability of the interval $(-\infty, r)$ for any rational number r]. In practice, of course, only a few assessments

are ever made, with the others being made conventionally (e.g., one might specify the first quartile and the median, but then choose a Cauchy density for the prior). Clearly one should worry about the effect of features of the prior that were not elicited.

Even more common in practice is to utilize a default or objective prior distribution, and Bayes's theorem does not then provide any guarantee as to performance. It has proved to be very useful to evaluate partially elicited and objective priors by utilizing frequentist techniques to evaluate their properties in repeated use.

3.4.1 *Information-based developments.* A number of developments of prior distributions utilize information-based arguments that rely on frequentist measures. Consider the *reference prior* theory, for instance, initiated in Bernardo (1979) and refined in Berger and Bernardo (1992). The reference prior is defined to be that distribution which minimizes the asymptotic Kullback–Leibler divergence between the posterior distribution and the prior distribution, thus hopefully obtaining a prior that "minimizes information" in an appropriate sense. This divergence is calculated with respect to a joint frequentist–Bayesian computation since, as in design, it is being computed before any data has been obtained.

The reference prior approach has arguably been the most generally successful method of obtaining Bayes rules that have excellent frequentist performance (see Berger, Philippe and Robert, 1998, as but one example). There are, furthermore, many other features of reference priors that are influenced by frequentist matters. One such feature is that the reference prior typically depends not only on the model, but also on which parameter is the inferential focus. Without such dependence on the "parameter of interest," optimal frequentist performance is typically not attainable by Bayesian methods.

A number of other information-based priors have also been derived. See Soofi (2000) for an overview and references.

3.4.2 *Consistency.* Perhaps the simplest frequentist estimation tool that a Bayesian can usefully employ is consistency: as the sample size grows to ∞, does the estimate being studied converge to the true value (in a suitable sense of convergence). Bayes estimates are virtually always consistent if the parameter space is finite-dimensional (see Schervish, 1995, for a typical result and earlier references), but this need not be true if the parameter space is not finite-dimensional or in irregular cases (see Ghosh, Ghosal and Samanta, 1994). Here is an example of the former.

EXAMPLE 3.4. In numerous models in use today, the number of parameters increases with the amount of data. The classic example of this is the Neyman–Scott problem (Neyman and Scott, 1948), in which one observes

$$X_{ij} \sim \mathcal{N}(\mu_i, \sigma^2), \quad i = 1, \ldots, n,\ j = 1, 2,$$

and is interested in estimating σ^2. Defining $\bar{x}_i = (x_{i1} + x_{i2})/2$, $\bar{\mathbf{x}} = (\bar{x}_1, \ldots, \bar{x}_n)$, $S^2 = \sum_{i=1}^{n} (x_{i1} - x_{i2})^2$ and $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)$, the likelihood function can be written

$$L(\boldsymbol{\mu}, \sigma) \propto \frac{1}{\sigma^{2n}} \exp\left[-\frac{1}{\sigma^2}\left(|\bar{\mathbf{x}} - \boldsymbol{\mu}|^2 + \frac{S^2}{4}\right)\right].$$

Until relatively recently, the most commonly used objective prior was the Jeffreys-rule prior (Jeffreys, 1961), here given by $\pi^J(\boldsymbol{\mu}, \sigma) = 1/\sigma^{n+1}$. The resulting posterior distribution for σ is proportional to the likelihood times the prior, which, after integrating out $\boldsymbol{\mu}$, is

$$\pi(\sigma|\mathbf{x}) \propto \frac{1}{\sigma^{2n+1}} \exp\left[-\frac{S^2}{4\sigma^2}\right].$$

One common Bayesian estimate of σ^2 is the posterior mean, which here is $S^2/[4(n-1)]$. This estimate is inconsistent, as can be seen by applying simple frequentist reasoning to the situation. Indeed, note that $(X_{i1} - X_{i2})^2/(2\sigma^2)$ is a chi-squared random variable with one degree of freedom, and hence that $S^2/(2\sigma^2)$ is chi-squared with n degrees of freedom. It follows by the law of large numbers that $S^2/(2n) \to \sigma^2$, so that the Bayes estimate converges to $\sigma^2/2$, the wrong value. (Any other natural Bayesian estimate, such as the posterior median or posterior mode, can also be seen to be inconsistent.)

The problem in the above example is that the Jeffreys-rule prior is often inappropriate in multidimensional settings, yet it can be difficult or impossible to assess this problem within the Bayesian paradigm itself. Indeed, the inadequacy of the multidimensional Jeffreys-rule prior has led to a search for improved objective priors in multivariable settings. The reference prior approach, mentioned earlier, has been one successful solution. [For the Neyman–Scott problem, the reference prior is $\pi^R(\boldsymbol{\mu}, \sigma) = 1/\sigma$, which results in a consistent posterior mean and, indeed, yields inferences that are numerically equal to the classical inferences for σ^2.] Another approach to developing improved priors is discussed in Section 3.4.3.

3.4.3 *Frequentist performance*: *coverage and admissibility.* Consistency is a rather crude frequentist criterion, and more sophisticated frequentist evaluations of performance of Bayesian procedures are often considered. For instance, one of the most common approaches to evaluation of an objective prior distribution is to see if it yields posterior credible sets that have good frequentist coverage properties. We have already seen examples of this method of evaluation in Examples 2.2 and 3.1.

Evaluation by frequentist coverage has actually been given a formal theoretical definition and is called the *frequentist-matching* approach to developing objective priors. The idea is to look at one-sided Bayesian credible sets for the unknown quantity of interest, and then seek that prior distribution for which the credible sets have optimal frequentist coverage asymptotically. Welch and Peers (1963) developed the first extensive results in this direction, essentially showing that, for one-dimensional continuous parameters, the Jeffreys prior is frequentist-matching. There is an extensive literature devoted to finding frequentist-matching priors in multivariate contexts; see Efron (1993), Rousseau (2000), Ghosh and Kim (2001), Datta, Mukerjee, Ghosh and Sweeting (2000) and Fraser, Reid, Wong and Yi (2003) for some recent results and earlier references.

Other frequentist properties have also been used to help in the choice of an objective prior. For instance, if estimation is the goal, it has long been common to utilize the frequentist concept of admissibility to help in the selection of the prior. The idea behind admissibility is to define a loss function in estimation (e.g., squared error loss), and then see if a proposed estimator can be beaten in terms of frequentist expected loss (e.g., mean squared error). If so, the estimator is said to be *inadmissible*; if it cannot be beaten, it is *admissible*. For instance, in situations having what is known as a group invariance structure, it has long been known that the prior distribution defined by the right-Haar measure will typically yield Bayes estimates that are admissible from a frequentist perspective, while the seemingly more natural (to a Bayesian) left-Haar measure will typically fail to yield admissible estimators. Thus use of the right-Haar priors has become standard. See Berger (1985a) and Robert (2001) for general discussion and many examples of the use of admissibility.

Another situation in which admissibility has played an important role in prior development is in choice of Bayesian priors in hierarchical modeling. In a sense, this topic was initiated in Stein (1956), which effectively showed that the usual constant prior for a multivariate normal mean would result in an inadmissible estimator under quadratic loss (in three or more dimensions). One of the first Bayesian works to address this issue was Hill (1974). To access the huge resulting literature on the role of admissibility in choice of hierarchical priors, see Brown (1971), Berger and Robert (1990), Berger and Strawderman (1996), Robert (2001) and Tang (2001).

Here is an example where initial admissibility considerations led to significant Bayesian developments.

EXAMPLE 3.5. Consider estimation of a covariance matrix $\mathbf{\Sigma}$, based on i.i.d. multivariate normal data $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$, where each column vector $\mathbf{x}_i$ arises from the $N_k(\mathbf{0}, \mathbf{\Sigma})$ density. The sufficient statistic for $\mathbf{\Sigma}$ is $\mathbf{S} = \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i'$. Since Stein (1975), it has been understood that the commonly used estimates of $\mathbf{\Sigma}$, which are various multiples of $\mathbf{S}$ (depending on the loss function considered) are seriously inadmissible. Hence there has been a great effort in the frequentist literature (see Yang and Berger, 1994, for references) to develop better estimators of $\mathbf{\Sigma}$.

The interest in this from the Bayesian perspective is that by far the most commonly used subjective prior for a covariance matrix is the inverse Wishart prior (for subjectively specified a and b)

$$\pi(\mathbf{\Sigma}) \propto |\mathbf{\Sigma}|^{-a/2} \exp\{-\tfrac{1}{2}\operatorname{tr}[b\mathbf{\Sigma}^{-1}]\}. \tag{3.2}$$

A frequently used objective version of this prior is the Jeffreys-rule prior given by choosing $a = k + 1$ and $b = 0$. When one notes that the Bayesian estimates arising from these priors are linear functions of $\mathbf{S}$, which were deemed to be seriously inadequate by the frequentists, there is clear cause for concern in the routine use of these priors.

In this case, it is possible to also indicate the problem with these priors utilizing Bayesian reasoning. Indeed, write $\mathbf{\Sigma} = \mathbf{H}'\mathbf{D}\mathbf{H}$, where $\mathbf{H}$ is an orthogonal matrix and $\mathbf{D}$ is a diagonal matrix with diagonal entries being the eigenvalues of the matrix, $d_1 > d_2 > \cdots > d_k$. A change of variables yields

$$\pi(\mathbf{\Sigma})\, d\mathbf{\Sigma} = |\mathbf{D}|^{-a/2} \exp\{-\tfrac{1}{2}\operatorname{tr}[b\mathbf{D}^{-1}]\} \cdot \prod_{i<j} (d_i - d_j) \cdot I_{[d_1 > \cdots > d_k]}\, d\mathbf{D}\, d\mathbf{H},$$

where $I_{[d_1 > \cdots > d_k]}$ denotes the indicator function on the given set. Since $\prod_{i<j}(d_i - d_j)$ is near zero when any two eigenvalues are close, it follows that the conjugate priors (and the Jeffreys-rule prior) tend to

force apart the eigenvalues of the covariance matrix; the priors give near-zero density to close eigenvalues. This is contrary to typical prior beliefs. Indeed, often in modelling, one is debating between assuming an exchangeable covariance structure (and hence equal eigenvalues) or allowing a more general structure. When one is contemplating whether or not to assume equal eigenvalues, it is clearly inappropriate to use a prior distribution that gives essentially no weight to equal eigenvalues, and instead forces them apart.

As an alternative objective prior here, the reference prior was derived in Yang and Berger (1994) and is given by $\pi^*(\mathbf{D}, \mathbf{H}) = |\mathbf{D}|^{-1} d\mathbf{D}\, d\mathbf{H}$; this clearly eliminates the forcing apart of eigenvalues. Furthermore, it is shown in Yang and Berger (1994) that use of the reference prior often results in improvements in estimating $\mathbf{\Sigma}$ on the order of 50% over use of the Jeffreys prior.

Motivated, in part, by the significant inferiority of the standard inverse Wishart and Jeffreys-rule priors for $\mathbf{\Sigma}$, a large Bayesian literature has developed in recent years that provides alternative prior distributions for a covariance matrix. See Tang (2001) and Daniels and Pourahmadi (2002) for examples and earlier references.

Note that we are not only focusing on objective priors here. Even proper priors that are commonly used by subjectivists can have hidden and highly undesirable features—such as the forcing apart of the eigenvalues for the inverse Wishart priors in the above example—and frequentist (and objective Bayesian) tools can expose these features and allow for development of better subjective priors.

3.4.4 *Robust Bayesian analysis.* Robust Bayesian analysis formally recognizes the impossibility of complete subjective specification of the model and prior distribution; as mentioned earlier, complete specification would involve an infinite number of assessments, even in the simplest situations. It follows that one should, ideally, work with a class of prior distributions Γ with the class reflecting the uncertainty remaining after the (finite) elicitation efforts. (Γ could also reflect the differing judgments of various individuals involved in the decision process.)

While much of robust Bayesian analysis takes place in a purely Bayesian framework (e.g., determining the range of the posterior mean as the prior ranges over Γ), it also has strong connections with the empirical Bayes, gamma minimax and restricted risk Bayes approaches, discussed in Section 2.3. See Berger (1985a, 1994), Delampady et al. (2001) and Ríos, Insua and Ruggeri (2000) for discussion and references.

3.4.5 *Nonparametric Bayesian analysis.* In nonparametric statistical analysis, the unknown quantity in a statistical model is a function or a probability distribution. A Bayesian approach to such problems requires placing a prior distribution on this space of functions or space of probability distributions. Perhaps surprisingly, Bayesian analysis of such problems is computationally quite feasible and is seeing significant practical implementation; cf. Dey, Müller and Sinha (1998).

Function spaces and spaces of probability measures are enormous spaces, and subjective elicitation of a prior on these spaces is not really feasible. Thus, in practice, it is typical to use a convenient form for a nonparametric prior (typically chosen for computational reasons), with perhaps a small number of features of the prior being subjectively specified. Thus, much as in the case of the Neyman–Scott example, one worries that the unspecified features of the prior may overwhelm the data and result in inconsistency or poor frequentist performance. Furthermore, there is evidence (e.g., Freedman, 1999) that Bayesian credible sets and frequentist confidence sets need not agree in nonparametric problems, making it more difficult to judge performance.

There is a long-time literature on such issues, the earlier period going from Freedman (1963) through Diaconis and Freedman (1986). To access the more recent literature, see Barron (1999), Barron, Schervish and Wasserman (1999), Ghosal, Ghosh and van der Vaart (2000), Zhao (2000), Kim and Lee (2001), Belitser and Ghosal (2003) and Ghosh and Ramamoorthi (2003).

3.4.6 *Impropriety and identifiability.* One of the most crucial problems that Bayesians face in dealing with complex modeling situations is that of ensuring that the posterior distribution is proper; use of improper objective priors can result in improper posterior distributions. (Use of "vague proper priors" in such situations will formally result in proper posterior distributions, but these posteriors will essentially be meaningless if the limiting improper objective prior would have resulted in an improper posterior distribution.)

One of the major situations in which impropriety can arise is when there is a problem of parameter identifiability, as in the following example.

EXAMPLE 3.6. Suppose, for $i = 1, \ldots, p$, that $X_i \sim \text{Normal}(\mu_i, \sigma^2)$ and $\mu_i \sim \text{Normal}(0, \tau^2)$, all random variables being independent. Then, marginally, $X_i \sim \text{Normal}(0, \sigma^2 + \tau^2)$, and it is clear that we cannot separately estimate σ^2 and τ^2 (although we can

estimate their sum); in classical language, σ^2 and τ^2 are not *identifiable*. Were a Bayesian to attempt to utilize an improper objective prior here, such as $\pi(\sigma^2, \tau^2) = 1$, the posterior distribution would be improper.

The point here is that frequentist insight and literature about identifiability can be useful to a Bayesian in determining whether there is a problem with posterior propriety. Thus, in the above example, upon recognizing the identifiability problem, the Bayesian will know not to use the improper objective prior and will attempt to elicit a true subjective proper prior for at least one of σ^2 or τ^2. (Of course, more data, such as replications at the first stage of the model, could also be sought.)

3.5 Frequentist Simplifications and Asymptotic Approximations

Situations can occur in which straightforward use of frequentist intuition directly yields sensible answers. In the Neyman–Scott problem, for instance, consideration of the paired differences, $x_{i1} - x_{i2}$, directly yielded a sensible answer. In contrast, a fairly sophisticated objective Bayesian analysis (use of the reference prior) was required for a satisfactory answer.

This is not to say that classical methodology is universally better in such situations. Indeed, Neyman and Scott created this example primarily to show that use of maximum likelihood methodology can be very inadequate; it essentially leads to the same "bad" answer in the example as the Bayesian analysis based on the Jeffreys-rule prior. This points out the dilemma facing Bayesians in use of frequentist simplifications: a frequentist answer might be "simple," but a Bayesian might well feel uneasy in its utilization unless it were felt to approximate a Bayesian answer. (For instance, is the answer conditionally sound, as discussed in Section 3.2.2.) Of course, if only the frequentist answer is available, the issue is moot.

It would be highly useful to catalogue situations in which direct frequentist reasoning is arguably simpler than Bayesian methodology, but we do not attempt to do so. Discussion of this and examples can be found in Robins and Ritov (1997) and Robins and Wasserman (2000).

Outside of standard models (such as the normal linear model), it is unfortunately rather rare to be able to obtain exact frequentist answers for small or moderate sample sizes. Hence much of frequentist methodology relies on asymptotic approximations, based on assuming that the sample size is large.

Asymptotics can also be used to provide an approximation to Bayesian answers for large sample sizes; indeed, Bayesian and frequentist asymptotic answers are often (but not always) the same; see Schervish (1995) for an introduction to Bayesian asymptotics and Le Cam (1986) for a high-level discussion. One might conclude that this is thus another significant potential use of frequentist methodology by Bayesians. It is rather rare for Bayesians to directly use asymptotic answers, however, since Bayesians can typically directly compute exact small sample size answers, often with less effort than derivation of the asymptotic approximation would require.

Still, asymptotic techniques are useful to Bayesians, in a variety of approximations and theoretical developments. For instance, the popular Laplace approximation (cf. Schervish, 1995) and BIC (cf. Schwarz, 1978) are based on an asymptotic arguments. Important Bayesian methodological developments, such as the definition of reference priors, also make considerable use of asymptotic theory, as was mentioned earlier.

4. TESTING, MODEL SELECTION AND MODEL CHECKING

Unlike estimation, frequentist reports and conclusions in testing (and model selection) are often in conflict with their Bayesian counterparts. For a long time it was believed that this was unavoidable—that the two paradigms are essentially irreconcilable for testing. Berger, Brown and Wolpert (1994) showed, however, that this is not necessarily the case; that the main difficulty with frequentist testing was an inappropriate lack of conditioning which could, in a variety of situations, be fixed. This is the focus of the next section, after which we turn to more general issues involving the interaction of frequentist and Bayesian methodology in testing and model selection.

4.1 Conditional Frequentist Testing

Unconditional Neyman–Pearson testing, in which one reports the same error probability regardless of the size of the test statistic (as long as it is in the rejection region), has long been viewed as problematical by most statisticians. To Fisher, this was the main inadequacy of Neyman–Pearson testing, and one of the chief motivations for his championing p-values in testing and model checking. Unfortunately (as Neyman would observe), p-values do not have a frequentist justification in the sense, say, of the frequentist principle in Section 2.2. For more extensive discussion of the perceived inadequacies of these two approaches to testing, see Berger (2003).

The "solution" proposed in Berger, Brown and Wolpert (1994) for testing, following earlier developments in Kiefer (1977), was to use the Neyman–Pearson approach of formally defining frequentist error probabilities of Type I and Type II, but to do so conditional on the observed value of a statistic measuring the "strength of evidence in the data," as was done in Example 3.2. (Other proposed solutions to this problem have been considered in, e.g., Hwang et al., 1992.)

For illustration, suppose that we wish to test that the data $\mathbf{X}$ arises from the simple (i.e., completely specified) hypotheses $H_0: f = f_0$ or $H_1: f = f_1$. The idea is to select a statistic $S = S(\mathbf{X})$ which measures the "strength of the evidence" in $\mathbf{X}$, for or against the hypotheses. Then, conditional error probabilities (CEPs) are computed as

$$\begin{aligned} \alpha(s) &= P(\text{Type I error} \,|S = s) \\ &\equiv P_0\big(\text{reject } H_0 | S(\mathbf{X}) = s\big), \\ \beta(s) &= P(\text{Type II error}|S = s) \\ &\equiv P_1\big(\text{accept } H_0 | S(\mathbf{X}) = s\big), \end{aligned} \tag{4.1}$$

where P_0 and P_1 refer to probability under H_0 and H_1, respectively.

The proposed conditioning statistic S and associated test utilize p-values to measure the strength of the evidence in the data. Specifically (see Wolpert, 1996; Sellke, Bayarri and Berger, 2001), we consider

$$S = \max\{p_0, p_1\},$$

where p_0 is the p-value when testing H_0 versus H_1, and p_1 is the p-value when testing H_1 versus H_0. [Note that the use of p-values in determining evidentiary equivalence is much weaker than their use as an absolute measure of significance; in particular, use of $\psi(p_i)$, where ψ is any strictly increasing function, would determine the same conditioning.] The corresponding conditional frequentist test is then as follows:

$$\begin{aligned} &\text{if } p_0 \le p_1 \text{ reject } H_0 \text{ and} \\ &\qquad\qquad \text{report Type I CEP } \alpha(s); \\ &\text{if } p_0 > p_1 \text{ accept } H_0 \text{ and} \\ &\qquad\qquad \text{report Type II CEP } \beta(s); \end{aligned} \tag{4.2}$$

where the CEPs are given in (4.1).

To this point, there has been no connection with Bayesianism. Conditioning, as above, is completely allowed (and encouraged) within the frequentist paradigm. The Bayesian connection arises because Berger, Brown and Wolpert (1994) show that

$$\alpha(s) = \frac{B(\mathbf{x})}{1 + B(\mathbf{x})} \quad \text{and} \quad \beta(s) = \frac{1}{1 + B(\mathbf{x})}, \tag{4.3}$$

where $B(\mathbf{x})$ is the likelihood ratio (or Bayes factor), and these expressions are precisely the Bayesian posterior probabilities of H_0 and H_1, respectively, assuming the hypotheses have equal prior probabilities of $1/2$. Therefore, a conditional frequentist can simply compute the objective Bayesian posterior probabilities of the hypotheses, and declare that they are the conditional frequentist error probabilities; there is no need to formally derive the conditioning statistic or perform the conditional frequentist computations. (There are some technical details concerning the definition of the rejection region, but these have almost no practical impact; see Berger, 2003, for further discussion.)

The value of having a merging of the frequentist and objective Bayesian answers in testing goes well beyond the technical convenience of computation; statistics as a whole is the big winner because of the unification that results. But we are trying to avoid philosophical issues here and so will simply focus on the methodological advantages that will accrue to frequentism.

Dass and Berger (2003) and Paulo (2002) extend this result to many classical testing scenarios; here is an example from the former.

EXAMPLE 4.1. McDonald, Vance and Gibbons (1995) studied car emission data $\mathbf{X} = (X_1, \ldots, X_n)$, testing whether the i.i.d. X_i follow the Weibull or Lognormal distribution, given, respectively, by

$$H_0 : f_{\mathrm{W}}(x; \beta, \gamma) = \frac{\gamma}{\beta}\left(\frac{x}{\beta}\right)^{\gamma-1} \exp\left[-\left(\frac{x}{\beta}\right)^{\gamma}\right],$$

$$H_1 : f_{\mathrm{L}}(x; \mu, \sigma^2) = \frac{1}{x\sqrt{2\pi\sigma^2}} \exp\left[\frac{-(\ln x - \mu)^2}{2\sigma^2}\right].$$

There are several difficulties with classical analysis of this situation. First, there are no low-dimensional sufficient statistics, and no obvious test statistics; indeed, McDonald, Vance and Gibbons (1995) simply consider a variety of generic tests, such as the likelihood ratio test (MLR), which they eventually recommended as being the most powerful. Second, it is not clear which hypothesis to make the null hypothesis, and the classical conclusion can depend on this choice (although not significantly, in the sense of the choice allowing differing conclusions with low error probabilities). Finally, computation of unconditional error probabilities

requires a rather expensive simulation and, once computed, one is stuck with error probabilities that do not vary with the data.

For comparison, the conditional frequentist test when $n = 16$ (one of the cases considered by McDonald, Vance and Gibbons, 1995) results in the following test:

$$T^C = \begin{cases} \text{if } B(\mathbf{x}) \le 0.94, \\ \quad \text{reject } H_0 \text{ and report Type I CEP} \\ \quad \alpha(\mathbf{x}) = B(\mathbf{x})/(1 + B(\mathbf{x})), \\ \text{if } B(\mathbf{x}) > 0.94, \\ \quad \text{accept } H_0 \text{ and report Type II CEP} \\ \quad \beta(\mathbf{x}) = 1/(1 + B(\mathbf{x})), \end{cases}$$

where

$$B(\mathbf{x}) = \frac{2\Gamma(n)n^{-n/2}}{\Gamma((n-1)/2)\pi^{(n-1)/2}} \cdot \int_0^\infty \left[\frac{y}{n}\sum_{i=1}^n \exp\left(\frac{z_i - \bar{z}}{y s_z}\right)\right]^{-n} dy, \tag{4.4}$$

with $z_i = \ln x_i$, $\bar{z} = \frac{1}{n}\sum_{i=1}^n z_i$ and $s_z^2 = \frac{1}{n}\sum_{i=1}^n (z_i - \bar{z})^2$. In comparison with the situation for the unconditional frequentist test:

- There is a well-defined test statistic $B(\mathbf{x})$.
- If one switches the null hypothesis, the new Bayes factor is simply $B(\mathbf{x})^{-1}$, which will clearly lead to the same CEPs (i.e., the CEPs do not depend on which hypothesis is called the null hypothesis).
- Computation of the CEPs is almost trivial, requiring only a one-dimensional integration.
- Above all, the CEPs vary continuously with the data.

In elaboration of the last point, consider one of the testing situations considered by McDonald, Vance and Gibbons (1995), namely, testing for the distribution of carbon monoxide emission data, based on a sample of size $n = 16$. Data was collected at the four different mileage levels indicated in Table 2, with (b) and (a) indicating "before" or "after" scheduled vehicle maintenance. Note that the decisions for both the MLR and the conditional test would be to accept the lognormal model for the data. McDonald, Vance and Gibbons (1995) did not give the Type II error probability associated with acceptance (perhaps because it would depend on the unknown parameters for many of the test statistics they considered) but, even if Type II error had been provided, note that it would be constant. In contrast, the conditional test has CEPs (here, the conditional Type II errors) that vary fully with the data, usefully indicating the differing certainties in the acceptance decision for the considered mileages. Further analyses and comparisons can be found in Dass and Berger (2003).

TABLE 2
For CO data, the MLR test at level $\alpha = 0.05$, and the conditional test of H_0 : Lognormal versus H_1 : Weibull

Mileage	**0**	**4,000**	**24,000 (b)**	**24,000 (a)**
MLR decision	A	A	A	A
$B(\mathbf{x})$	2.436	9.009	6.211	2.439
T^C decision	A	A	A	A
CEP	0.288	0.099	0.139	0.291

Derivation of the conditional frequentist test. The conditional frequentist analysis here depends on recognizing an important fact: both the Weibull and the lognormal distributions are location–scale distributions (the Weibull after suitable transformation). In this case, the objective Bayesian (and, hence, conditional frequentist) solution to the problem is to utilize the right-Haar prior density for the distributions [$\pi(\mu, \sigma) = \sigma^{-1}$ for the lognormal problem] and compute the resulting Bayes factor. (Unconditional frequentists could, of course, have recognized the invariance of the situation and used this as a test statistic, but they would have faced a much more difficult computational challenge.)

By invariance, the distribution of $B(\mathbf{X})$ under either hypothesis does not depend on model parameters, so that the original testing problem can be reduced to testing two simple hypotheses, namely H_0 : "$B(\mathbf{X})$ has distribution F_{W}^*" versus H_1 : "$B(\mathbf{X})$ has distribution F_{L}^*," where F_{W}^* and F_{L}^* are the distribution functions of $B(\mathbf{X})$ under the Weibull and Lognormal distributions, respectively, with an arbitrary choice of the parameters (e.g., $\beta = \gamma = 1$ for the Weibull, and $\mu = 0, \sigma = 1$ for the Lognormal). Recall that the CEPs happen to equal the objective Bayesian posterior probabilities of the hypotheses.

4.2 Model Selection

Clyde and George (2004) give an excellent review of Bayesian model selection. Frequentists have not typically used Bayesian arguments in model selection, although that may be changing, in part due to the pronounced practical success that Bayesian "model averaging" has achieved. Bayesians often use frequentist arguments to develop approximate model selection methods (such as BIC), to evaluate performance of model selection methods and to develop default priors for model selection. There is a huge list of articles of this type, including many listed in Clyde and

George (2004). Robert (2001), Berger and Pericchi (2001, 2004) and Berger, Ghosh and Mukhopadhyay (2003) also have general discussions and numerous other recent references. The story here is far from settled, in that there is no agreement on the immediate horizon as to even a reasonable method of model selection. It seems highly likely, however, that any such agreement will be based on a mixture of frequentist and Bayesian arguments.

4.3 p-Values for Model Checking

Both classical statisticians and Bayesians routinely use p-values for model checking. We first consider their use by classical statisticians and show the value of Bayesian methodology in the computation of "proper" p-values; then we turn to Bayesian p-values and the importance of frequentist ideas in their evaluation.

4.3.1 *Use of Bayesian methodology in computing classical p-values.* Suppose that a statistical model $H_0 : \mathbf{X} \sim f(\mathbf{x}|\theta)$ is being entertained, data $\mathbf{x}_{\text{obs}}$ is observed and it is desired to check whether the model is adequate, in light of the data. Classical statisticians have long used p-values for this purpose. A strict frequentist would not do so (since p-values do not satisfy the frequentist principle), but most frequentists relax their morals a bit in this situation (i.e., when there is no alternative hypothesis which would allow construction of a Neyman–Pearson test).

The common approach to model checking is to choose a statistic $T = t(\mathbf{X})$, where (without loss of generality) large values of T indicate less compatibility with the model. The p-value is then defined as

$$p = \Pr\big(t(\mathbf{X}) \geq t(\mathbf{x}_{\text{obs}})|\theta\big). \tag{4.5}$$

When θ is known, this probability computation is with respect to $f(\mathbf{x}|\theta)$. The crucial question is what to do when θ is unknown.

For future reference, we note a key property of the p-value when θ is known: considered as a random function of $\mathbf{X}$ in the continuous case, $p(\mathbf{X})$ has a Uniform$(0, 1)$ distribution under H_0. The implication of this property is that p-values then have a common interpretation across statistical problems, making their general use feasible. Indeed, this property, or its asymptotic version for composite H_0, has been used to characterize "proper," well-behaved p-values (see Robins, van der Vaart and Ventura, 2000).

When θ is unknown, computation of a p-value requires some way of "eliminating" θ from (4.5). There are many non-Bayesian ways of doing so, some of which are reviewed in Bayarri and Berger (2000) and Robins, van der Vaart and Ventura (2000). Here we consider only the most common method, which is to replace θ in (4.5) by its mle, $\hat{\theta}$. The resulting p-value will be called the *plug-in p-value* (p_{plug}). Henceforth using a superscript to denote the density with respect to which the p-value in (4.5) is computed, the *plug-in p-value* is thus defined as

$$p_{\text{plug}} = \Pr^{f(\cdot|\hat{\theta})}\big(t(\mathbf{X}) \geq t(\mathbf{x}_{\text{obs}})\big). \tag{4.6}$$

Although very simple to use, there is a worrisome "double use" of the data in p_{plug}, first to estimate θ and then to compute the tail area corresponding to $t(\mathbf{x}_{\text{obs}})$ in that distribution.

Bayarri and Berger (2000) proposed the following alternative way of eliminating θ, based on Bayesian methodology. Begin with an objective prior density $\pi(\theta)$; we recommend, as before, that this be a reference prior (when available), but even a constant prior density will usually work fine. Next, define the *partial posterior density* (Bayesian motivation will be given in the next section)

$$\begin{aligned} \pi(\theta|\mathbf{x}_{\text{obs}} \setminus t_{\text{obs}}) &\propto f(\mathbf{x}_{\text{obs}}|t_{\text{obs}}, \theta)\pi(\theta) \\ &\propto \frac{f(\mathbf{x}_{\text{obs}}|\theta)\pi(\theta)}{f(t_{\text{obs}}|\theta)}, \end{aligned} \tag{4.7}$$

resulting in the *partial posterior predictive density* of T,

$$m(t|\mathbf{x}_{\text{obs}} \setminus t_{\text{obs}}) = \int f(t|\theta)\pi(\theta|\mathbf{x}_{\text{obs}} \setminus t_{\text{obs}})\, d\theta. \tag{4.8}$$

Since this density is free of θ, it can be used in (4.5) to compute the *partial posterior predictive p-value*,

$$p_{\text{ppp}} = \Pr^{m(\cdot|\mathbf{x}_{\text{obs}} \setminus t_{\text{obs}})}(T \geq t_{\text{obs}}). \tag{4.9}$$

Note that p_{ppp} uses only the information in $\mathbf{x}_{\text{obs}}$ that is *not* in $t_{\text{obs}} = t(\mathbf{x}_{\text{obs}})$ to "train" the prior and to then eliminate θ. Intuitively this avoids double use of the data because the contribution of t_{obs} to the posterior is "removed" before θ is eliminated by integration. (The notation $\mathbf{x}_{\text{obs}} \setminus t_{\text{obs}}$ was chosen to indicate this.)

When T is not ancillary (or approximately so), the double use of the data can cause the plug-in p-value to fail dramatically, in the sense of being moderately large even when the null model is clearly wrong. Examples and earlier references are given in Bayarri and Berger (2000). Here is another interesting illustration, taken from Bayarri and Castellanos (2004).

EXAMPLE 4.2. Consider the hierarchical (or random effects) model

$$\begin{aligned} X_{ij}|\mu_i &\sim N(\mu_i, \sigma_i^2) \quad \text{for } i = 1, \ldots, I, \\ &\qquad\qquad\quad j = 1, \ldots, n_i, \\ \mu_i|\nu, \tau &\sim N(\nu, \tau^2) \quad \text{for } i = 1, \ldots, I, \end{aligned} \tag{4.10}$$

where all variables are independent and, for simplicity, we assume that the variances σ_i^2 at the first level are known. Suppose that we are primarily interested in investigating whether the normality assumption for the means μ_i is compatible with the data. We choose the test statistic $T = \max\{\bar{X}_1, \ldots, \bar{X}_I\}$, where $\bar{X}_i$ denotes the usual group sample means, which here are sufficient statistics for the μ_i. When no specific alternative hypothesis is postulated, "optimal" test statistics do not exist, and casual choices (such as this) are not uncommon. The issue under study is whether such (easy) choices of the test statistic can usefully be used for computing p-values.

Since the μ_i are random effects, it is well known that tests should be based on the marginal densities of the sufficient statistics $\bar{X}_i$, with the μ_i integrated out. (Replacing the μ_i by their mle's would result in a vacuous inference here.) The resulting null distribution can be represented

$$\bar{X}_i|\nu, \tau \sim N(\nu, \sigma_i^2 + \tau^2) \quad \text{for } i = 1, \ldots, I. \tag{4.11}$$

Thus p_{plug} is computed with respect to this distribution, with the mle's $\hat{\nu}, \hat{\tau}^2$ [numerically computed from (4.11)] inserted back into (4.11) and (4.10).

To compute the partial posterior predictive p-value, we begin with a common objective prior for (ν, τ^2), namely $\pi(\nu, \tau^2) = 1/\tau$ (not $1/\tau^2$ as is sometimes done, which would result in an improper posterior). The computation of p_{plug} is based on an MCMC, discussed in Bayarri and Castellanos (2004).

Both p-values were computed for a simulated data set, in which one of the groups comes from a distribution with a much larger mean than the other groups. In particular, the data was generated from

$$\begin{aligned} X_{ij}|\mu_i &\sim N(\mu_i, 4) \quad \text{for } i = 1, \ldots, 5, \\ &\qquad\qquad\quad j = 1, \ldots, 8, \\ \mu_i &\sim N(1, 1) \quad \text{for } i = 1, \ldots, 4, \\ \mu_5 &\sim N(5, 1). \end{aligned} \tag{4.12}$$

The resulting sample means were 1.560, 0.641, 1.982, 0.014 and 6.964. Note that the sample mean of the fifth group is 6.65 standard deviations away from the mean of the other four groups. With this data, $p_{\text{plug}} = 0.130$, dramatically failing to clearly indicate that the assumption of i.i.d. normality of the μ_i is wrong, while $p_{\text{ppp}} = 0.010$. Many other similar examples can be found in Castellanos (2002).

A strong indication that a proposed p-value is inappropriate is when it fails to be asymptotically Uniform$(0, 1)$ under the null hypothesis for all values of θ, as a proper p-value should. Robins, van der Vaart and Ventura (2000) prove that the plug-in p-value often fails to be asymptotically proper in this sense, while the partial posterior predictive p-value is asymptotically proper. Furthermore, they show that the latter p-value is uniformly most powerful with respect to Pitman alternatives, lending additional powerful frequentist support to the methodology. Note that this is completely a frequentist evaluation; no Bayesian averaging is involved. Numerical comparisons in the above example of the behavior of p_{plug} and p_{ppp} under the assumed (null) model is deferred to Section 4.3.2.

4.3.2 *Evaluating Bayesian p-values.* Most Bayesian p-values are defined analogously to (4.8) and (4.9), but with alternatives to $\pi(\theta|\mathbf{x}_{\text{obs}} \setminus t_{\text{obs}})$. Subjective Bayesians simply use the prior distribution $\pi(\theta)$ directly to integrate out θ. The resulting p-value, called the *predictive p-value* and popularized by Box (1980), has the property, when considered as a random variable of X, of being uniformly distributed under the null predictive distribution. (See Meng, 1994, for discussion of the importance of this property.) This p-value is thus Uniform$[0, 1]$ in an average sense over θ, which is presumably satisfactory (for consistency of interpretation across problems) to Bayesians who believe in their subjective prior.

Much of model checking, however, takes place in scenarios in which the model is quite tentative and, hence, for which serious subjective prior elicitation (which is invariably highly costly) is not feasible. Hence model checking is more typically done by Bayesians using objective prior distributions; guarantees concerning "average uniformity" of p-values then no longer apply and, indeed, the resulting p-values are not even defined if the objective prior distribution is improper (since the predictive distribution of T will then be improper). The "solution" that has become quite popular is to utilize objective priors, but to use the posterior distribution $\pi(\theta|\mathbf{x}_{\text{obs}})$, instead of the prior, in defining the distribution used to compute a p-value. Formally, this leads to the *posterior predictive p-value*, defined in Guttman (1967) and popularized in Rubin

(1984) and Gelman, Carlin, Stern and Rubin (1995), given by

$$p_{\text{post}} = \Pr^{m(\cdot|\mathbf{x}_{\text{obs}})}(T \geq t_{\text{obs}}),$$
$$m(t|\mathbf{x}_{\text{obs}}) = \int f(t|\theta)\pi(\theta|\mathbf{x}_{\text{obs}})\,d\theta. \tag{4.13}$$

Note that there is also "double use" of the data in p_{post}, first to convert the (possibly improper) prior into a proper distribution for determining the predictive distribution, and then for computing the tail area corresponding to $t(\mathbf{x}_{\text{obs}})$ in that distribution. The detrimental effect of this double use of the data was discussed in Bayarri and Berger (2000) and arises again in Example 4.2. Indeed, computation yields that the posterior predictive p-value for the given data is 0.409, which does not at all suggest a problem with the random effects normality assumption; yet, recall that one of the means was more than six standard deviations away from the others.

The point of this section is to observe that the frequentist property of "uniformity of a p-value under the null hypothesis" provides a useful discriminatory tool for judging the adequacy of Bayesian (and other) p-values. Note that if a p-value is uniform under the null hypothesis in the frequentist sense for any θ, then it has the strong Bayesian property of being marginally Uniform[0, 1] under *any* proper prior distribution. More important, if a proposed p-value is *always* either conservative or anticonservative in a frequentist sense (see Robins, van der Vaart and Ventura, 2000, for definitions), then it is likewise guaranteed to be conservative or anticonservative in a Bayesian sense, no matter what the prior. If the conservatism (or anticonservatism) is severe, then the p-value cannot correspond to any approximate true Bayesian p-value. In this regard, Robins, van der Vaart and Ventura (2000) show that p_{post} is often severely conservative (and, surprisingly, is worse in this regard than is p_{plug}), while p_{ppp} is asymptotically Uniform[0, 1].

It is also of interest to numerically study the nonasymptotic null distribution of the three p-values considered in this section. We thus return to the random effects example.

EXAMPLE 4.3. Consider again the situation described in Example 4.2. We consider $p_{\text{plug}}(\mathbf{X})$, $p_{\text{ppp}}(\mathbf{X})$ and $p_{\text{post}}(\mathbf{X})$ as random variables and simulate their distribution under an instance of the null model. Specifically, we chose the random effects mean and variance to be 0 and 1, respectively, thus simulating X_{ij} as in (4.12), but now generating all five μ_i from the $N(0, 1)$ distribution (so that the null normal hierarchical model is correct). Figure 6 shows the resulting sampling distributions of the three p-values.

Note that the distribution of $p_{\text{ppp}}(\mathbf{X})$ is quite close to uniform, even though only five means were involved. In contrast, the distributions of $p_{\text{plug}}(\mathbf{X})$ and $p_{\text{post}}(\mathbf{X})$ are quite far from uniform, with the latter being the worst. Even for larger numbers of means (e.g., 25), $p_{\text{plug}}(\mathbf{X})$ and $p_{\text{post}}(\mathbf{X})$ remained significantly nonuniform, indicating a serious and inappropriate conservatism.

Of course, Bayesians frequently criticize the direct use of p-values in testing a precise hypothesis, in that p-values then fail to correspond to natural Bayesian measures. There are, however, various *calibrations* of p-values that have been suggested (see Good, 1983; Sellke, Bayarri and Berger, 2001). But these are "higher level" considerations in that, if a purported "Bayesian p-value" fails to adequately criticize a null hypothesis, the higher level concerns are irrelevant.

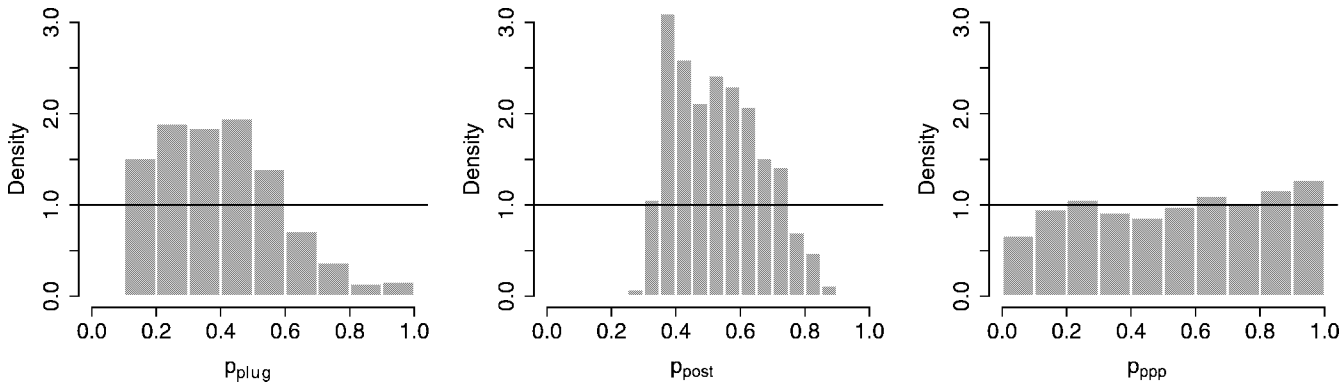

FIG. 6. *Null distribution of* $p_{\text{plug}}(\mathbf{X})$ (left), $p_{\text{post}}(\mathbf{X})$ (center) *and* $p_{\text{ppp}}(\mathbf{X})$ (right) *when the null normal hierarchical model is correct.*

5. AREAS OF CURRENT DISAGREEMENT

It is worth mentioning some aspects of inference in which it seems very difficult to reconcile the frequentist and Bayesian approaches. Of course, as discussed in Section 4, it was similarly believed to be difficult to reconcile frequentist and Bayesian testing until recently, so it may simply be a matter of time until reconciliation also occurs in these other areas.

Multiple comparisons. When doing multiple tests, such as occurs in the variable selection problem in regression, classical statistics performs some type of adjustment (e.g., the Bonferonni adjustment to the significance level) to account for the multiplicity of tests. In contrast, Bayesian analysis does not explicitly adjust for multiplicity of tests, the argument being that a correct adjustment is automatic within the Bayesian paradigm.

Unfortunately, the duality between frequentist and Bayesian methodology in testing, discussed in Section 4, has not been extended to the multiple hypothesis testing framework, so we are left with competing and quite distinct methodologies. Worse, there are multiple competing frequentist methodologies and multiple competing Bayesian methodologies, a situation that is professionally disconcerting, yet common when there is no frequentist–Bayesian consensus. (To access some of these methodologies, see http://www.ba.ttu.edu/isqs/westfall/mcp2002.htm.)

Sequential analysis. The *stopping rule principle* says that once the data have been obtained, the reasons for stopping experimentation should have no bearing on the evidence reported about unknown model parameters. This principle is automatically satisfied by Bayesian analysis, but is viewed as crazy by many frequentists. Indeed frequentist practice in clinical trials is to "spend α" for looks at the data; that is, if there are to be interim analyses during the clinical trial, with the option of stopping the trial early should the data look convincing, frequentists feel that it is then mandatory to adjust the allowed error probability (down) to account for the multiple analyses.

This issue is extensively discussed in Berger and Berry (1988), which has many earlier references. That it is a controversial and difficult issue is admirably expressed by Savage (1962): "I learned the stopping rule principle from Professor Barnard, in conversation in the summer of 1952. Frankly, I then thought it a scandal that anyone in the profession could advance an idea so patently wrong, even as today I can scarcely believe that people resist an idea so patently right."

An interesting recent development is the finding, in Berger, Boukai and Wang (1999), that "optimal" conditional frequentist testing also essentially obeys the stopping rule principle. This suggests the admittedly controversial speculation that optimal (conditional) frequentist procedures will eventually be found to essentially satisfy the stopping rule principle; and that we currently have a controversy only because suboptimal (unconditional) frequentist procedures are being used.

It is, however, also worth noting that in prior development Bayesians sometimes utilize the stopping rule to help define an objective prior, for example, the Jeffreys or reference prior (see Ye, 1993; Sun and Berger, 2003, for examples and motivation). Since the statistical inference will then depend on the stopping rule, objective Bayesian analysis can involve a (probably slight) violation of the stopping rule principle. (See Sweeting, 2001, for related discussion concerning the extent of violation of the likelihood principle by objective Bayesian methods.)

Finite population sampling. In this central area of statistics, classical inference is primarily based on frequentist averaging with respect to the sampling probabilities by which units of the population are selected for inclusion in the sample. In contrast, Bayesian analysis asserts that these sampling probabilities are irrelevant for inference, once the data is at hand. (See Ghosh, 1988, which contains excellent essays by Basu on this subject.) Hence we, again, have a fundamental philosophical and practical conflict.

There have been several arguments (e.g., Rubin, 1984; Robins and Ritov, 1997) to the effect that there are situations in which Bayesians do need to take into account the sampling probabilities, to save themselves from a too-difficult (and potentially nonrobust) prior development. Conversely, there has been a very significant growth in use of Bayesian methodology in finite population contexts, such as in the use of "small area estimation methods" (see, e.g., Rao, 2003). Small area estimation methods actually occur in the broader context of the model-based approach to finite population sampling (which mostly ignores the sampling probabilities), and this model-based approach also has frequentist and Bayesian versions (the differences, however, being much smaller than the differences arising from use, or not, of the sampling probabilities).

6. CONCLUSIONS

It seems quite clear that both Bayesian and frequentist methodology are here to stay, and that we should

not expect either to disappear in the future. This is not to say that *all* Bayesian or *all* frequentist methodology is fine and will survive. To the contrary, there are many areas of frequentist methodology that should be replaced by (existing) Bayesian methodology that provides superior answers, and the verdict is still out on those Bayesian methodologies that have been exposed as having potentially serious frequentist problems.

Philosophical unification of the Bayesian and frequentist positions is not likely, nor desirable, since each illuminates a different aspect of statistical inference. We can hope, however, that we will eventually have a general methodological unification, with both Bayesians and frequentists agreeing on a body of standard statistical procedures for general use.

ACKNOWLEDGMENTS

Research supported by Spanish Ministry of Science and Technology Grant SAF2001-2931 and by NSF Grants DMS-01-03265 and DMS-01-12069. Part of the work was done while the first author was visiting the Statistical and Applied Mathematical Sciences Institute and ISDS, Duke University.

REFERENCES

BARNETT, V. (1982). *Comparative Statistical Inference*, 2nd ed. Wiley, New York.

BARRON, A. (1999). Information-theoretic characterization of Bayes performance and the choice of priors in parametric and nonparametric problems. In *Bayesian Statistics 6* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) 27–52. Oxford Univ. Press.

BARRON, A., SCHERVISH, M. J. and WASSERMAN, L. (1999). The consistency of posterior distributions in nonparametric problems. *Ann. Statist.* **27** 536–561.

BAYARRI, M. J. and BERGER, J. (2000). P-values for composite null models (with discussion). *J. Amer. Statist. Assoc.* **95** 1127–1170.

BAYARRI, M. J. and CASTELLANOS, M. E. (2004). Bayesian checking of hierarchical models. Technical report, Univ. Valencia.

BELITSER, E. and GHOSAL, S. (2003). Adaptive Bayesian inference on the mean of an infinite-dimensional normal distribution. *Ann. Statist.* **31** 536–559.

BERGER, J. (1985a), *Statistical Decision Theory and Bayesian Analysis*, 2nd ed. Springer, New York.

BERGER, J. (1985b). The frequentist viewpoint and conditioning. In *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer* (L. Le Cam and R. Olshen, eds.) 15–44. Wadsworth, Monterey, CA.

BERGER, J. (1994). An overview of robust Bayesian analysis (with discussion). *Test* **3** 5–124.

BERGER, J. (2003). Could Fisher, Jeffreys and Neyman have agreed on testing (with discussion)? *Statist. Sci.* **18** 1–32.

BERGER, J. and BERNARDO, J. (1992). On the development of reference priors. In *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) 35–60. Oxford Univ. Press.

BERGER, J. and BERRY, D. (1988). The relevance of stopping rules in statistical inference (with discussion). In *Statistical Decision Theory and Related Topics IV* (S. Gupta and J. Berger, eds.) **1** 29–72. Springer, New York.

BERGER, J., BOUKAI, B. and WANG, Y. (1999). Simultaneous Bayesian–frequentist sequential testing of nested hypotheses. *Biometrika* **86** 79–92.

BERGER, J., BROWN, L. D. and WOLPERT, R. (1994). A unified conditional frequentist and Bayesian test for fixed and sequential simple hypothesis testing. *Ann. Statist.* **22** 1787–1807.

BERGER, J., GHOSH, J. K. and MUKHOPADHYAY, N. (2003). Approximations and consistency of Bayes factors as model dimension grows. *J. Statist. Plann. Inference* **112** 241–258.

BERGER, J. and PERICCHI, L. (2001). Objective Bayesian methods for model selection: Introduction and comparison (with discussion). In *Model Selection* (P. Lahiri, ed.) 135–207. IMS, Beachwood, OH.

BERGER, J. and PERICCHI, L. (2004). Training samples in objective Bayesian model selection. *Ann. Statist.* **32** 841–869.

BERGER, J., PHILIPPE, A. and ROBERT, C. (1998). Estimation of quadratic functions: Noninformative priors for non-centrality parameters. *Statist. Sinica* **8** 359–376.

BERGER, J. and ROBERT, C. (1990). Subjective hierarchical Bayes estimation of a multivariate normal mean: On the frequentist interface. *Ann. Statist.* **18** 617–651.

BERGER, J. and STRAWDERMAN, W. (1996). Choice of hierarchical priors: Admissibility in estimation of normal means. *Ann. Statist.* **24** 931–951.

BERGER, J. and WOLPERT, R. L. (1988). *The Likelihood Principle*: *A Review*, *Generalizations*, *and Statistical Implications*, 2nd ed. IMS, Hayward, CA. (With discussion.)

BERNARDO, J. M. (1979). Reference posterior distributions for Bayesian inference (with discussion). *J. Roy. Statist. Soc. Ser. B* **41** 113–147.

BERNARDO, J. M. and SMITH, A. F. M. (1994). *Bayesian Theory*. Wiley, New York.

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

BROWN, L. D. (1971). Admissible estimators, recurrent diffusions, and insoluble boundary value problems. *Ann. Math. Statist.* **42** 855–903.

BROWN, L. D. (1994). Minimaxity, more or less. In *Statistical Decision Theory and Related Topics V* (S. Gupta and J. Berger, eds.) 1–18. Springer, New York.

BROWN, L. D. (2000). An essay on statistical decision theory. *J. Amer. Statist. Assoc.* **95** 1277–1281.

BROWN, L. D., CAI, T. T. and DASGUPTA, A. (2001). Interval estimation for a binomial proportion (with discussion). *Statist. Sci.* **16** 101–133.

BROWN, L. D., CAI, T. T. and DASGUPTA, A. (2002). Confidence intervals for a binomial proportion and asymptotic expansions. *Ann. Statist.* **30** 160–201.

CARLIN, B. P. and LOUIS, T. A. (2000). *Bayes and Empirical Bayes Methods for Data Analysis*, 2nd ed. Chapman and Hall, London.

CASELLA, G. (1988). Conditionally acceptable frequentist solutions (with discussion). In *Statistical Decision Theory and Related Topics IV* (S. Gupta and J. Berger, eds.) **1** 73–117. Springer, New York.

CASTELLANOS, M. E. (2002). Diagnóstico Bayesiano de modelos. Ph.D. dissertation, Univ. Miguel Hernández, Spain.

CHALONER, K. and VERDINELLI, I. (1995). Bayesian experimental design: A review. *Statist. Sci.* **10** 273–304.

CLYDE, M. and GEORGE, E. (2004). Model uncertainty. *Statist. Sci.* **19** 81–94.

DANIELS, M. and POURAHMADI, M. (2002). Bayesian analysis of covariance matrices and dynamic models for longitudinal data. *Biometrika* **89** 553–566.

DASS, S. and BERGER, J. (2003). Unified conditional frequentist and Bayesian testing of composite hypotheses. *Scand. J. Statist.* **30** 193–210.

DATTA, G. S., MUKERJEE, R., GHOSH, M. and SWEETING, T. J. (2000). Bayesian prediction with approximate frequentist validity. *Ann. Statist.* **28** 1414–1426.

DAWID, A. P. and SEBASTIANI, P. (1999). Coherent dispersion criteria for optimal experimental design. *Ann. Statist.* **27** 65–81.

DAWID, A. P. and VOVK, V. G. (1999). Prequential probability: Principles and properties. *Bernoulli* **5** 125–162.

DE FINETTI, B. (1970). *Teoria delle Probabilità* **1**, **2**. Einaudi, Torino. [English translations published (1974, 1975) as *Theory of Probability* **1**, **2**. Wiley, New York.]

DELAMPADY, M., DASGUPTA, A., CASELLA, G., RUBIN, H. and STRAWDERMAN, W. E. (2001). A new approach to default priors and robust Bayes methodology. *Canad. J. Statist.* **29** 437–450.

DEY, D., MÜLLER, P. and SINHA, D., eds. (1998). *Practical Nonparametric and Semiparametric Bayesian Statistics*. *Lecture Notes in Statist.* **133**. Springer, New York.

DIACONIS, P. (1988a). Bayesian numerical analysis. In *Statistical Decision Theory and Related Topics IV* (S. Gupta and J. Berger, eds.) **1** 163–175. Springer, New York.

DIACONIS, P. (1988b). Recent progress on de Finetti's notion of exchangeability. In *Bayesian Statistics 3* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.) 111–125. Oxford Univ. Press.

DIACONIS, P. and FREEDMAN, D. (1986). On the consistency of Bayes estimates (with discussion). *Ann. Statist.* **14** 1–67.

EATON, M. L. (1989). *Group Invariance Applications in Statistics*. IMS, Hayward, CA.

EFRON, B. (1993). Bayes and likelihood calculations from confidence intervals. *Biometrika* **80** 3–26.

FRASER, D. A. S., REID, N., WONG, A. and YI, G. Y. (2003). Direct Bayes for interest parameters. In *Bayesian Statistics 7* (J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West, eds.) 529–534. Oxford Univ. Press.

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

FREEDMAN, D. A. (1999). On the Bernstein–von Mises theorem with infinite-dimensional parameters. *Ann. Statist.* **27** 1119–1140.

GART, J. J. and NAM, J. (1988). Approximate interval estimation of the ratio of binomial parameters: A review and corrections for skewness. *Biometrics* **44** 323–338.

GELMAN, A., CARLIN, J. B., STERN, H. and RUBIN, D. B. (1995). *Bayesian Data Analysis*. Chapman and Hall, London.

GHOSAL, S., GHOSH, J. K. and VAN DER VAART, A. W. (2000). Convergence rates of posterior distributions. *Ann. Statist.* **28** 500–531.

GHOSH, J. K., ed. (1988). *Statistical Information and Likelihood. A Collection of Critical Essays*. *Lecture Notes in Statist.* **45**. Springer, New York.

GHOSH, J. K., GHOSAL, S. and SAMANTA, T. (1994). Stability and convergence of the posterior in non-regular problems. In *Statistical Decision Theory and Related Topics V* (S. S. Gupta and J. Berger, eds.) 183–199. Springer, New York.

GHOSH, J. K. and RAMAMOORTHI, R. V. (2003). *Bayesian Nonparametrics.* Springer, New York.

GHOSH, M. and KIM, Y.-H. (2001). The Behrens–Fisher problem revisited: A Bayes–frequentist synthesis. *Canad. J. Statist.* **29** 5–17.

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

GUTTMAN, I. (1967). The use of the concept of a future observation in goodness-of-fit problems. *J. Roy. Statist. Soc. Ser. B* **29** 83–100.

HILL, B. (1974). On coherence, inadmissibility and inference about many parameters in the theory of least squares. In *Studies in Bayesian Econometrics and Statistics* (S. Fienberg and A. Zellner, eds.) 555–584. North-Holland, Amsterdam.

HOBERT, J. (2000). Hierarchical models: A current computational perspective. *J. Amer. Statist. Assoc.* **95** 1312–1316.

HWANG, J. T., CASELLA, G., ROBERT, C., WELLS, M. T. and FARRELL, R. (1992). Estimation of accuracy in testing. *Ann. Statist.* **20** 490–509.

JEFFREYS, H. (1961). *Theory of Probability*, 3rd ed. Oxford Univ. Press.

KIEFER, J. (1977). Conditional confidence statements and confidence estimators (with discussion). *J. Amer. Statist. Assoc.* **72** 789–827.

KIM, Y. and LEE, J. (2001). On posterior consistency of survival models. *Ann. Statist.* **29** 666–686.

LAD, F. (1996). *Operational Subjective Statistical Methods*: *A Mathematical, Philosophical and Historical Introduction*. Wiley, New York.

LE CAM, L. (1986). *Asymptotic Methods in Statistical Decision Theory*. Springer, New York.

LEHMANN, E. L. and CASELLA, G. (1998). *Theory of Point Estimation*, 2nd ed. Springer, New York.

MCDONALD, G. C., VANCE, L. C. and GIBBONS, D. I. (1995). Some tests for discriminating between lognormal and Weibull distributions—an application to emissions data. In *Recent Advances in Life-Testing and Reliability—A Volume in Honor of Alonzo Clifford Cohen, Jr.* (N. Balakrishnan, ed.) Chapter 25. CRC Press, Boca Raton, FL.

MENG, X.-L. (1994). Posterior predictive p-values. *Ann. Statist.* **22** 1142–1160.

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

MOSSMAN, D. and BERGER, J. (2001). Intervals for post-test probabilities: A comparison of five methods. *Medical Decision Making* **21** 498–507.

NEYMAN, J. (1977). Frequentist probability and frequentist statistics. *Synthèse* **36** 97–131.

NEYMAN, J. and SCOTT, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica* **16** 1–32.

O'HAGAN, A. (1992). Some Bayesian numerical analysis. In *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) 345–363. Oxford Univ. Press.

PAULO, R. (2002). Problems on the Bayesian/frequentist interface. Ph.D. dissertation, Duke Univ.

PRATT, J. W. (1965). Bayesian interpretation of standard inference statements (with discussion). *J. Roy. Statist. Soc. Ser. B* **27** 169–203.

RAO, J. N. K. (2003). *Small Area Estimation*. Wiley, New York.

REID, N. (2000). Likelihood. *J. Amer. Statist. Assoc.* **95** 1335–1340.

RÍOS INSUA, D. and RUGGERI, F., eds. (2000). *Robust Bayesian Analysis*. *Lecture Notes in Statist.* **152**. Springer, New York.

ROBBINS, H. (1955). An empirical Bayes approach to statistics. *Proc. Third Berkeley Symp. Math. Statist. Probab.* **1** 157–164. Univ. California Press, Berkeley.

ROBERT, C. P. (2001). *The Bayesian Choice*, 2nd ed. Springer, New York.

ROBERT, C. P. and CASELLA, G. (1999). *Monte Carlo Statistical Methods*. Springer, New York.

ROBINS, J. M. and RITOV, Y. (1997). Toward a curse of dimensionality appropriate (CODA) asymptotic theory for semiparametric models. *Statistics in Medicine* **16** 285–319.

ROBINS, J. M., VAN DER VAART, A. and VENTURA, V. (2000). Asymptotic distribution of p-values in composite null models. *J. Amer. Statist. Assoc.* **95** 1143–1156.

ROBINS, J. and WASSERMAN, L. (2000). Conditioning, likelihood and coherence: A review of some foundational concepts. *J. Amer. Statist. Assoc.* **95** 1340–1346.

ROBINSON, G. K. (1979). Conditional properties of statistical procedures. *Ann. Statist.* **7** 742–755.

ROUSSEAU, J. (2000). Coverage properties of one-sided intervals in the discrete case and applications to matching priors. *Ann. Inst. Statist. Math.* **52** 28–42.

RUBIN, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Ann. Statist.* **12** 1151–1172.

RUBIN, H. (1987). A weak system of axioms for "rational" behavior and the non-separability of utility from prior. *Statist. Decisions* **5** 47–58.

SAVAGE, L. J. (1962). *The Foundations of Statistical Inference.* Methuen, London.

SCHERVISH, M. (1995). *Theory of Statistics*. Springer, New York.

SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464.

SELLKE, T., BAYARRI, M. J. and BERGER, J. (2001). Calibration of p-values for testing precise null hypotheses. *Amer. Statist.* **55** 62–71.

SOOFI, E. (2000). Principal information theoretic approaches. *J. Amer. Statist. Assoc.* **95** 1349–1353.

STEIN, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Proc. Third Berkeley Symp. Math. Statist. Probab* **1** 197–206. Univ. California Press, Berkeley.

STEIN, C. (1975). Estimation of a covariance matrix. Reitz Lecture, IMS–ASA Annual Meeting. (Also unpublished lecture notes.)

STRAWDERMAN, W. (2000). Minimaxity. *J. Amer. Statist. Assoc.* **95** 1364–1368.

SUN, D. and BERGER, J. (2003). Objective priors under sequential experimentation. Technical report, Univ. Missouri.

SWEETING, T. J. (2001). Coverage probability bias, objective Bayes and the likelihood principle. *Biometrika* **88** 657–675.

TANG, D. (2001). Choice of priors for hierarchical models: Admissibility and computation. Ph.D. dissertation, Purdue Univ.

VIDAKOVIC, B. (2000). Gamma-minimax: A paradigm for conservative robust Bayesians. In *Robust Bayesian Analysis. Lecture Notes in Statist.* **152** 241–259. Springer, New York.

WALD, A. (1950). *Statistical Decision Functions*. Wiley, New York.

WELCH, B. and PEERS, H. (1963). On formulae for confidence points based on integrals of weighted likelihoods. *J. Roy. Statist. Soc. Ser. B* **25** 318–329.

WOLPERT, R. L. (1996). Testing simple hypotheses. In *Data Analysis and Information Systems*: *Statistical and Conceptual Approaches* (H.-H. Bock and W. Polasek, eds.) 289–297. Springer, Berlin.

WOODROOFE, M. (1986). Very weak expansions for sequential confidence levels. *Ann. Statist.* **14** 1049–1067.

YANG, R. and BERGER, J. (1994). Estimation of a covariance matrix using the reference prior. *Ann. Statist.* **22** 1195–1211.

YE, K. (1993). Reference priors when the stopping rule depends on the parameter of interest. *J. Amer. Statist. Assoc.* **88** 360–363.

ZHAO, L. (2000). Bayesian aspects of some nonparametric problems. *Ann. Statist.* **28** 532–552.