Calibrated Bayes: A Bayes/Frequentist Roadmap

Roderick LITTLE

The lack of an agreed inferential basis for statistics makes life "interesting" for academic statisticians, but at the price of negative implications for the status of statistics in industry, science, and government. The practice of our discipline will mature only when we can come to a basic agreement about how to apply statistics to real problems. Simple and more general illustrations are given of the negative consequences of the existing schism between frequentists and Bayesians.

An assessment of strengths and weaknesses of the frequentist and Bayes systems of inference suggests that calibrated Bayes—a compromise based on the work of Box, Rubin, and others—captures the strengths of both approaches and provides a roadmap for future advances. The approach asserts that inferences under a particular model should be Bayesian, but model assessment can and should involve frequentist ideas. This article also discusses some implications of this proposed compromise for the teaching and practice of statistics.

KEY WORDS: Bayesian statistics; Frequentist statistics; Likelihood principle; Model checking; Statistical inference.

1. INTRODUCTION

The year 2005 marks the term of the 100th president of the American Statistical Association, in the person of Fritz Scheuren. Such occasions promote reflections on where the ASA stands, and more generally on the state of the discipline of statistics in the world. The 99th ASA President, Brad Efron, was optimistic about the state of statistics in his address to the Association (Efron 2005). He labeled the 19th Century as generally Bayesian, the 20th Century as generally frequentist, and suggested that statistics in the 21st Century will require a combination of Bayesian and frequentist ideas. In the spirit of Efron's call for a synthesis, I advocate here a compromise based on the calibrated Bayesian ideas of George Box, Don Rubin, and others. The topic is very broad, and I limit references to work with which I am most familiar, without meaning to slight the large body of other significant contributions.

In the next section I reflect on past debates of statistical philosophy, and argue that a resolution of philosophical disagreements about how to do statistics would help our profession. Some aspects of the conflict are illustrated with basic and more general examples. In Sections 3 and 4, I provide my personal perspective on strengths and weaknesses of the frequentist and Bayesian approaches to statistics, and in Section 5 I argue that calibrated Bayes is a compromise that capitalizes on the strengths of both systems of inference. The calibrated Bayes compromise provides a useful roadmap, but remains sketchy on the boundary between inference and model selection, suggesting areas of future development. In Section 6, I discuss some implications of the proposed compromise for the future teaching and practice of statistics.

2. THE BAYES/FREQUENTIST SCHISM: DOES IT MATTER?

I was a student of statistics in London in the early 1970s, when debates raged about alternative philosophies of statistics. Elements of the debate included:

1. Birnbaum's (1962) "proof" of the likelihood principle, which if true invalidates frequentist inference—more on this later;

2. books emphasizing issues of comparative inference (Hacking 1965; Edwards 1972; Barnett 1973; Cox and Hinkley 1974);

3. read papers at the Royal Statistical Society that focused on competing systems of inference, with associated lively discussions (e.g. Dawid, Stone, and Zidek 1973; Wilkinson 1977; Bernardo 1979);

4. alleged "counter-examples" to frequentist inference (Robinson 1975);

5. debates on the "foundations of survey inference," focusing on model-based versus design-based inference (Basu 1971; Smith 1976; Hansen, Madow, and Tepping 1983); and

6. the statements and work of outspoken and influential statisticians like Dennis Lindley and Oscar Kempthorne.

I viewed these debates with a mixture of fascination and youthful incomprehension. Many insights emerged, but no clear winners—no agreed inferential philosophy for how to do statistics has emerged. The Bayes/frequentist schism remains unresolved.

At some point people seemed to lose interest in this debate. These days I find very few sessions on statistical philosophy at statistical meetings, particularly where different inferential views are contrasted and argued. Bayesians debate "objective" versus "subjective" approaches to Bayesian inference, but that seems like an argument between siblings in the Bayesian family, largely ignored by frequentists.

Whether or not the inferential debate has receded, it is no longer academic! Thirty years ago, applications of Bayes were limited to smallish problems by the inability to compute the high-dimensional integrations involved in multiparameter models. In-

Roderick Little is Professor and Chair, Department of Biostatistics, School of Public Health, University of Michigan, 1420 Washington Heights, Ann Arbor, MI 48109-2029 (E-mail: *rlittle@umich.edu*). I thank Fritz Scheuren for the opportunity to present this work as the President's Invited Address at the 2005 Joint Statistical Meetings, and for his valued friendship and support. If there are useful ideas in this article, I owe them to my many statistical mentors, including Martin Beale, David Cox, Wilfred Dixon, Maurice Kendall, Paul Meier, Don Rubin, and David Wallace. I also thank my colleagues at the University of Michigan for continual support and advice, and my energetic students, particularly Guangyu Zhang for help with computing.

 DOI: 10.1198/000313006X117837

creased computational power and the development of Monte Carlo approaches to computing posterior distributions has turned this weakness of Bayes into a strength, and Bayesian approaches to large complicated models are now common in the statistical and scientific literatures.

I see current-day statisticians as roughly divided into three main camps: (a) frequentists (F), who abhor the Bayesian approach, or never learned much about it; (b) Bayesians (B), with varying views on the role of frequentist ideas; and (c) pragmatists, who do not have an overarching philosophy and pick and choose what seems to work for the problem at hand.

To be a bit more specific about (a) and (b), I regard a "frequentist" as one who bases inference for an unknown parameter θ on hypothesis tests or confidence intervals, derived from the distribution of statistics in repeated sampling. I regard a "Bayesian" as one who bases inferences about θ on its posterior distribution, under some model for the data and prior distribution for unknown parameters. Included in the latter is "subjective" Bayes, where proper priors are elicited, and "objective" Bayes, where conventional "reference priors" are adopted. In my view these different facets of the Bayesian paradigm both have useful roles, depending on context. I also regard asymptotic maximum likelihood inference as a form of large-sample Bayes, with the interval for θ being interpreted as a posterior credibility interval rather than a confidence interval. This broad view of Bayes provides a large class of practical frequentist methods with a Bayesian interpretation.

Within either Bayes and frequentist paradigms, one might add loss functions and adopt a decision-theoretic perspective; I do not do that here since I lack the expertise to do it justice. It is also my impression that (for better or worse) this perspective is rarely adopted by practicing statisticians, who usually confine attention to testing hypotheses and estimating parameters with associated measures of uncertainty. I also recognize that much of statistics involves describing and summarizing data without any attempt to make a formal statistical inference. My focus on the latter is not intended to devalue the former.

The classification (a) to (c) captures the major approaches to inference in my view, but it is by no means comprehensive; for example, there are "likelihoodists" who attempt to base inferences directly on the likelihood without introducing a prior distribution (Edwards 1972; Royall 1997).

Currently, it seems to me that the pragmatist approach (c) to inference predominates. Pragmatists might argue that good statisticians can get sensible answers under Bayes or frequentist paradigms; indeed maybe two philosophies are better than one, since they provide more tools for the statistician's toolkit! For example, sampling statisticians often use randomization inference for some problems, and models for other problems. I take a pragmatic attitude to applications, but I confess I am discomforted by this "inferential schizophrenia." Since the Bayesian (B) and frequentist (F) philosophies can differ even on simple problems, at some point decisions seem needed as to which is right. I believe our credibility as statisticians is undermined when we cannot agree on the fundamentals of our subject—as Efron (2005) noted:

> The physicists I talked to were really bothered by our 250-year old Bayesian-frequentist argument. Basically, there's only one way of doing physics, but there seems to be at least two way of doing statistics, and they don't always give the same answers.

A prominent Bayesian (Berger 2000) writes in a similar vein:

> Note that I am not arguing for an eclectic attitude toward statistics here; indeed, I think the general refusal in our field to strive for a unified perspective has been the single biggest impediment to its advancement.

Berger also saw the need for synthesis, adding that "any unification that will be achieved will almost certainly have frequentist components to it."

Some examples are offered to support the idea that one system of statistical inference may be better than two. The first two are basic examples where the Bayesian and frequentist approaches lead to fundamentally different results. The other examples are broader, indicating some areas where our philosophical differences create confusion in many areas of statistical application.

Example 1: Tests of independence in a 2 × 2 contingency table. Students with even a cursory exposure to statistics learn the Pearson chi-squared test of independence in a 2×2 table. Yet even in this well-studied problem, deep philosophical questions lurk close to the surface. Consider a one-sided test $H_0 : \pi_A = \pi_B; H_a : \pi_A > \pi_B$ for independent samples assigned two treatments, where π_j is the success rate for treatment $j, j = 1, 2$. Three competitors of the standard Pearson chi-squared test (P) are the Pearson test with Yates's continuity correction (Y), the Fisher exact test (F), or the Bayesian solution, which computes the posterior probability that $\pi_A < \pi_B$, based on some choice of prior distribution for the success rates. We illustrate the Bayesian approach for Jeffreys's reference prior $p(\pi_A, \pi_B) \propto \pi_A^{-1/2}\pi_B^{-1/2}$ (B), while emphasizing that other prior distributions yield different answers. Asymptotically, these approaches yield similar answers, but in small or moderate samples they can differ in important ways. For example, Table 1, taken from Little (1989), yields a one-sided P value of 0.016 for P, 0.030 for F, 0.032 for Y, and a posterior probability that $\pi_A < \pi_B$ of 0.013 for B.

Table 1. (2 × 2) Contingency Table for Example 1

Treatment	Success	Failure
A	170	2
B	162	9

The P values for F and Y tend to be similar for the one-sided problem, and are known to be conservative when one margin of the 2×2 table is fixed (a common feature of many designs); P is better calibrated when one margin is fixed, but is approximate. F is exact if both margins are fixed. So for the frequentist, the choice of P versus F or Y comes down to whether or not we condition on the second margin. There seems to me very little agreement on this question (Yates 1984; Little 1989). If the second margin were ancillary, many frequentists would condition on it. In this case the second margin is not exactly ancillary for the odds ratio, but it is approximately ancillary, in the sense that information in the margin about the odds ratio tends to zero as the sample size increases. So there is no clear frequentist answer for this most basic of problems.

The Bayesian answer avoids ambiguity about conditioning on the second margin; indeed conditioning is never really an issue

with the Bayesian approach, because posterior distributions condition on all the data. On the other hand, there is nothing unique about the Bayesian answer either, since the posterior probability depends on the choice of prior, and the theory of "reference priors" leading to the Jeffreys's prior has its own problems.

My second example is a minor wrinkle on another problem from Statistics 101:

Example 2: Single sample t inference with a bound on precision. Consider an independent normally distributed sample with $n = 7$ observations, with sample mean $\overline{x} = 1$ and standard deviation $s = 1$. If the population standard deviation (say σ) is unknown, the usual t-based 95% interval for the population mean is

$$I^{\rm BRP}_{0.05}(s) = I^{\rm F}_{0.05}(s) = \overline{x} \pm 2.447(s/\sqrt{n}) = 1 \pm 0.92, \quad (1)$$

where a frequentist F interprets this as a 95% confidence interval (CI), and a Bayesian B interprets it as a 95% posterior credibility interval, based on Jeffreys's reference prior (RP) distribution $p(\mu, \sigma) \propto 1/\sigma$. The correspondence of the B and F intervals is well known. Suppose now that we are told that $\sigma = 1.5$, as when σ is the known precision of a measuring instrument. The standard 95% interval is then

$$\begin{aligned} I^{\rm BRP}_{0.05}(\sigma = 1.5) &= I^{\rm F}_{0.05}(\sigma = 1.5) \\ &= \overline{x} \pm 1.96(1.5/\sqrt{n}) = 1 \pm 1.11. \quad (2) \end{aligned}$$

A collaborator hoping for an interval that excludes a null value of zero might prefer (1), but both F and B can agree that (2) is the correct inference, the wider interval reflecting the fact that the sample variance s is underestimating the true variance σ. Now, suppose the experimenter reports that $\sigma > 1.5$, since he remembers some additional unaccounted sources of variability. Three candidate 95% intervals for μ are Equations (1) and (2), or the Bayesian credibility interval with the reference prior modified to incorporate the constraint that $\sigma > 1.5$, namely:

$$I^{\rm BRP}_{0.05}(\sigma > 1.5) = 1 \pm 1.45. \quad (3)$$

Pick your poison:

(a) Equation (1) seems the optimal 95% frequentist confidence interval, given that it is has exact nominal coverage and σ is unknown, but it is counter-intuitive for inference: advocates of this confidence interval will have difficulty explaining how the information that $\sigma > 1.5$ leads to a narrower interval than Equation (2), the standard frequentist interval when $\sigma = 1.5$! A referee scoffs at any notion that (1) is *the* frequentist solution because it ignores the lower bound on σ; but I know of no interval that has exact 95% coverage and takes appropriate account of this information.

(b) Equation (2) is the obvious asymptotic approximation, given that 1.5 is the maximum likelihood estimate of σ. However, for a sample size $n = 7$ the appeal to asymptotics is clearly wrong, and it is not clear with what to replace it. The constraint $\sigma > 1.5$ implies that Equation (2) has less than 95% confidence coverage, since it is based on a known underestimate of σ. Interestingly, this is true even though it contains the exact t interval, Equation (1), for the observed sample; a neat illustration that a CI is not a probability interval, since with a probability interval that clearly cannot happen!

(c) Equation (3) is the Bayes interval subject to the constraint that $\sigma > 1.5$. It is appropriately wider than Equation (2), but it is (as always) dependent on the choice of prior distribution.

These examples illustrate worrying ambiguity in simple settings; I would also argue that differences in Bayes and frequentist solutions undermine the credibility of statisticians in much more complex real-world settings. Here are some broader examples:

Example 3: Penalties for peeking? Clinical trials often have interim analyses to assess whether they should be continued, or stopped because one treatment has a decisive advantage. Should inferences be affected by these multiple looks at the data? A frequentist says yes, tests need to be modified to maintain the nominal alpha-level for tests of the null (spending functions) (e.g., DeMets and Ware 1980). A Bayesian says no, the stopping rule is ignorable, so the posterior distribution is unaffected by prior looks at data (e.g., Lindley 1972, p. 24). This example is cited as a counter-example by both Bayesians and frequentists! If we statisticians can't agree which theory this example is counter to, what is a clinician to make of this debate?

Example 4: Multiple imputation combining rules for missing data. Rubin developed multiple imputation, which imputes more than one draw from the predictive distribution of the missing values (Rubin 1987). Multiple imputation combining rules (MICR) have been developed for interval estimation and hypothesis testing, which can be applied to yield inferences that account for imputation uncertainty (Rubin 1987; Little and Rubin 2002). There is much controversy about these combining rules under potentially misspecified imputation models (Meng 1994; Rubin 1996; Fay 1996; Rao 1996; Robins and Wang 2000). The 300-pound gorilla looming over these debates is to my mind philosophical, namely whether the inference is frequentist or Bayesian (see Figure 1). In particular, MICRs are based on Bayesian principles, whereas the criticisms focus on frequentist issues like unbiased estimation of sampling variance. Without an agreement on the underlying philosophy of inference, it is hard to see how these disputes can be resolved.

Example 5: Survey weights in regression. Substantive analyses of survey data are usually model-based, whereas the survey statisticians who collect the data typically advocate design-based inference. These groups differ on basic issues like how to weight the data: should we use design-based weights proportional to the inverse of the probability of selection, or model-based weights proportional to the inverse of the residual model variance? Bayesian inference in surveys is one variant of the model-based approach, and design-based inference is inherently frequentist, so this division is a further illustration of the Bayes-frequentist divide. I have argued that a unified approach based on models that account for the survey design would reduce friction and speed progress (Little 2004).

Example 6: Modeling the Census undercount. The Census undercount is a very complicated problem, involving an inter-

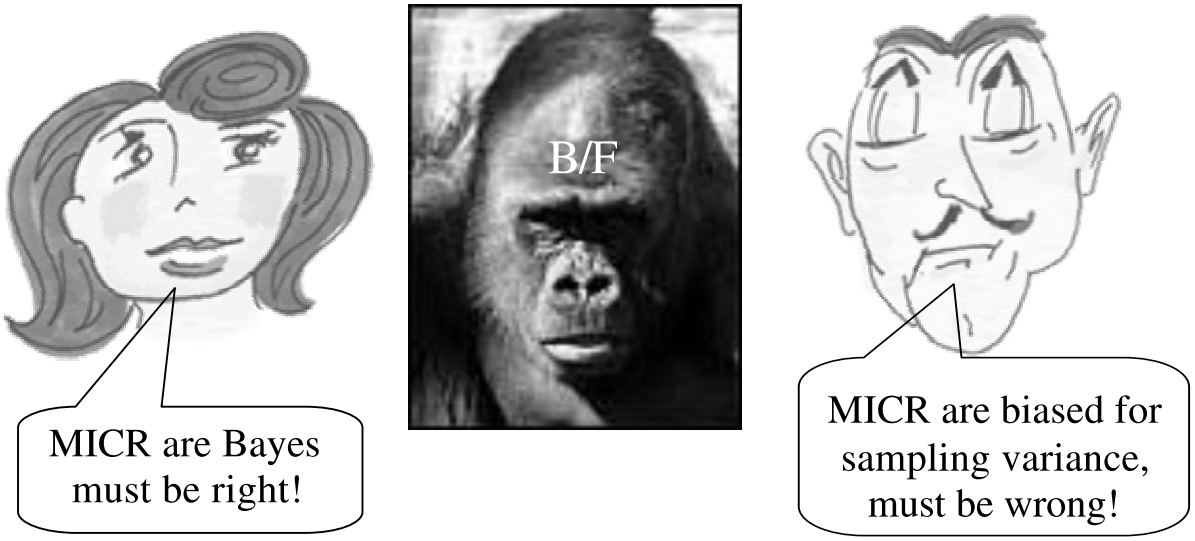

Figure 1. The Bayes/frequentist gorilla lurks behind arguments about multiple imputation combining rules (MICRs).

play between politics, data collection, human behavior, and complex statistical modeling. The design-based tradition at the Census Bureau, antipathy towards subjective elements of models in a data-collection agency, and divisions within the profession of the value of statistical modeling in such settings, have led to limited and hesitant use of statistical models for this problem. The lack of a unified front on the role of statistical modeling within the profession has encouraged politicians to weigh in on sides that favor their self-interest. Thus, it seems at times that the Census Bureau is limited to the kind of basic statistical tools "that a U.S. congressman can understand." One might contrast this with the sophisticated modeling that occurs in other areas of policy interest, like econometric simulation models or infectious disease modeling (e.g. Longini et al. 2005). With a unified front that accepted modeling as the basic tool for inference, we might develop the complex methodology needed to launch this spaceship.

Given these confusions and ambiguities arising from differences in frequentist and Bayesian statistics, is there a compromise that captures the strengths of both approaches? I think there is, and it emerges naturally from an assessment of the relative strengths and weaknesses of the Bayesian and frequentist paradigms. My personal assessment is provided in the next two sections.

3. STRENGTHS AND WEAKNESSES OF FREQUENTIST INFERENCE

The frequentist paradigm avoids the need for a prior distribution, and makes a clear separation of the role of prior information in model formulation and the role of data in estimating parameters. These pieces are treated on a more equal footing in the Bayesian approach, in that the prior density and likelihood multiply to create the posterior distribution. The frequentist approach is flexible, in the sense that full modeling is not necessarily required, and inferences lack the formal structure of Bayes's theorem under a fully specified prior and likelihood. In a sense any method is frequentist, provided its frequentist properties can be studied. The focus on repeated sampling properties tends to assure that frequentist inferences are well calibrated; for example, in the survey sampling setting, design-based inference automatically takes into account survey design features that might be ignored in a model-based approach.

The frequentist paradigm has serious shortcomings, however. To be succinct, the frequentist paradigm is not prescriptive, incomplete, ambiguous, and incoherent. (Apart from that it's a great theory!) Let me elaborate.

Frequentist theory is not prescriptive. Frequentist theory seems to me a set of concepts for assessing properties of inference procedures rather than an inferential system per se.

There is no unified theory for how to generate these procedures. Thus, the principle of least squares generates some useful methods, but is too limited for a general theory. Unbiasedness seems a desirable property, but it has severe limitations as a general principle. Illustrations include the James-Stein results on inadmissibility of unbiased estimates, and (my favorite) Basu's famous elephant example (Basu 1971), where an unbiased estimator based on a misguided underlying model is always useless. Generalized estimating equations provide a very broad class of procedures, but there seems no general prescription for how to choose the equations, and the theory is basically asymptotic.

Although frequentist inference as a whole is not, I feel, prescriptive, some parts of it are: in particular, Efron (1986) suggested that Fisher's theory of maximum likelihood estimation, with measures of uncertainty based on the observed information, is popular because it provides an "automatic" form of frequentist inference. This is not to me an argument against Bayes, because it is the form of frequentist inference closest to Bayes, and it has a large sample Bayes interpretation. Efron contrasted this with Bayesian theory, which "requires a great deal of thought about the given situation to apply sensibly." However, given enough data and some attention to the treacherous rocks represented by improper posterior distributions, one may use Bayes with a reference prior to achieve an inference with same degree of thought as Fisherian theory, and often with better results. In his discus-

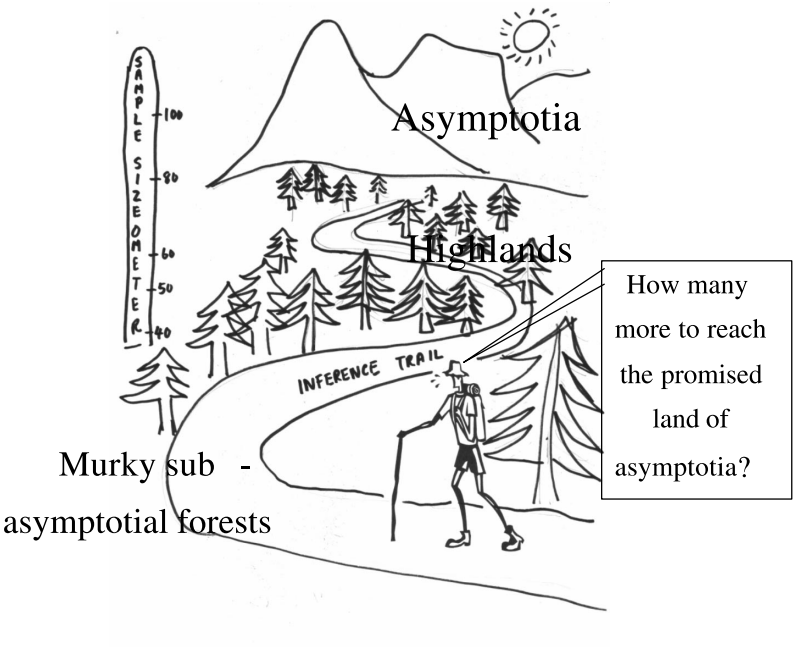

Figure 2. How large a sample is needed to reach the land of asymptotia?

sion of Efron (1986), Lindley (1986) noted the automatic nature of Bayes's theorem for generating the posterior distribution.

Frequentist theory is incomplete. If we define frequentist answers as capturing the information in the data, and yielding answers with exact frequentist properties—for example, a 95% confidence interval covering the true parameter 95% of the time in repeated sampling—then frequentist statistics does not yield enough answers. That is, exact finite-sample frequentist solutions are limited to a narrow class of problems. A famous historical example is the Behrens-Fisher problem, concerning two independent normal samples with means (μ_1, μ_2) and distinct variances (σ_1^2, σ_2^2), where there is no efficient confidence interval for the difference in means with exact confidence coverage (although there are many serviceable approximate solutions).

Perhaps the lack of a satisfactory small-sample theory has led frequentists to focus on asymptotic properties, where the ground is much more fertile. An example is the current enthusiasm for semiparametric efficiency, with asymptotic results driven by a search for robustness without invoking modeling assumptions. This focus generates a lot of elegant mathematics and some useful practical methods as well. However, the emphasis on asymptotic properties has its dangers, as it is often unclear whether we have reached the magic "land of asymptotia" (see Figure 2). Asymptotic results are of limited use for assessing the trade-off between model complexity and the available sample size. Simulations can do this, but they are often treated with disdain by mathematical statisticians, and have limited scope.

Frequentist theory is ambiguous. Specifically, the reference set for determining repeated-sampling properties is often ambiguous, and frequentist theory suffers from nonuniqueness of ancillary statistics (Cox 1971). Example 1 is one of many examples; another is whether to condition on the post-stratum sample counts in post-stratification of sample surveys (Holt and Smith 1979).

Frequentist theory is incoherent, in the sense that it violates the likelihood principle. Birnbaum (1962) claimed to "prove" that under a well-specified model M, S + C = L. That is,

- S = sufficiency—the principle that data can be reduced to sufficient statistics, when combined with
- C = conditionality—the principle of conditioning on ancillary statistics that have distributions unrelated to parameters of interest, which leads to
- L = likelihood principle—models and datasets leading to the same likelihood function should generate the same statistical inferences.

The likelihood principle plays an important role in the inferential debate since it is satisfied by Bayesian inference and violated by frequentist inference. A classic example of the latter is contained in two coin-tossing experiments, where θ is the probability of a head and $1-\theta$ the probability of a tail. Consider two experiments: (a) binomial sampling, where the coin is tossed $n = 12$ times and $X = 3$ of the tosses turn up heads (E1); and (b) negative binomial sampling, where the coin is tossed until a predetermined number $x = 3$ heads are obtained, and $N = 12$ tosses are needed (E2). Both E1 and E2 lead to the same likelihood, namely

$$\mathrm{L} \propto \theta^3(1-\theta)^9.$$

Hence, under the likelihood principle, these two experiment/data combinations should yield the same inferences. The maximum likelihood estimates are the same (3/12), but since the sampling spaces of E1 and E2 are different, the P values from the usual exact tests of $H_0 : \theta = 1/2$ against $H_a : \theta < 1/2$ are different—0.073 for E1 and 0.033 for E2. These repeated sampling inferences (and the confidence intervals obtained by inverting the

tests) violate the likelihood principle, since they differ, despite the fact that the likelihood function L is the same.

Birnbaum's (1962) paper caused quite a stir when it came out. For example, Savage opened his discussion with the words:

> . . . this is really an historic occasion. This paper is a landmark in statistics because it seems to me improbable that many people will be able to read it. . . without coming away with considerable respect for the likelihood principle.

In a recent discussion of these ideas, Robins and Wasserman (2000) argued for reexamination of the likelihood principle in the context of an infinite-dimensional example described by Robins and Ritov (1997).

4. STRENGTHS AND WEAKNESSES OF THE BAYESIAN PARADIGM

The Bayesian paradigm addresses many of the weaknesses of the frequentist approach described above. Specifically it is a prescriptive, complete, unambiguous (for a given choice of model and prior distribution), and coherent. For a given Bayesian model and prior distribution, Bayes's theorem is the simple prescription that supplies the inference. It may be difficult to compute, and checks are needed to ensure that the posterior distribution is proper, but the solution is clear and ambiguous. Bayes's inference is also coherent, in that Bayes's theorem is the correct way to update beliefs (as represented by probability distributions) to incorporate new information (e.g., Savage 1954; de Finetti 1974). Coherency is often argued in the decision-theoretic framework (e.g., Ferguson 1967).

Bayesian inferences (fixed probability intervals for unknown quantities), not frequentist inferences (random intervals for fixed quantities), are arguably what people really want; in particular, confidence intervals lead to the kinds of inconsistencies illustrated in Example 2, or in more elaborate examples such as that of Robinson (1975).

There are, however, difficulties in implementing the Bayes approach that inhibit its adoption. Some often-cited problems with the Bayesian paradigm are not to my mind very compelling. Perhaps the most common is that Bayes is viewed as too subjective for scientific inference, requiring a subjective definition of probability and the selection of a prior distribution. However, subjectivity is to me a matter of degree, and Bayesian models can run the full gamut, from standard regression models with reference priors that mimic frequentist intervals, to more subjective models that bring in proper prior information (Press 1986). My broad view of Bayesian methods includes methods based on noninformative priors that some classify as frequentist (e.g., Samaniego and Reneau 1994). Frequentist methods also vary in subjectivity. For example, a covariate selected out of a regression equation is in effect being given a sharp prior distribution with all the probability at zero. Models with strong assumptions, such as models selectivity bias (e.g., Heckman 1976), are no less subjective because they are analyzed using frequentist methods. Some statisticians worry about the subjective definition of probability that underlies the Bayesian approach, but I am not one of them.

Another criticism of Bayesianism is that it denies the role of randomization for design, since the randomization distribution is not the basis for model-based inferences. Indeed, some Bayesians have fueled this criticism by denying that randomization plays any kind of useful design role. On the contrary, the utility of randomization from the Bayesian perspective becomes clear when the model is expanded to include indicators for the selection of cases or allocation of treatment. Randomization provides a practical way to assure that the selection or allocation mechanisms are ignorable for inference, without making ignorable selection or allocation a questionable assumption. Gelman, Carlin, Stern, and Rubin (2003, chap. 7) provided a clear discussion of this point.

There are, however, some difficulties with the practical implementation of the Bayesian approach that I find more compelling.

Bayes requires and relies on full specification of a model (likelihood and prior). In general Bayes involves a full probability model for the data, including a prior distribution for unknown parameters—Efron (2005) discussed this high degree of specification. Developing a good model is often challenging, particularly in complex problems. Where does this model come from? Is it trustworthy? Bayes is much less prescriptive about how to select models than it is once model and prior distribution are selected.

Bayes yields "too many answers." I complained that the frequentist paradigm does not provide enough exact answers; with Bayes, there is an embarrassment of riches, because once the likelihood is nailed down, every prior distribution leads to a different answer! If forced to pick a prior distribution, the problem is which prior to choose. If the mapping from the prior distribution to the posterior distribution is considered the key, as argued cogently by some Bayesians (e.g., Leamer 1978), there is still a problem with the surfeit of posterior distributions. Sensitivity analysis is often a rational choice, but it is not a choice that appeals much to practitioners who are looking for clear-cut answers.

Models are always wrong, and bad models lead to bad answers. Although the search for procedures with good frequentist properties provides some degree of protection against model misspecification under the frequentist paradigm, there seems no such built-in protection under the strict Bayesian paradigm. Models are always idealizations and hence simplified, and models that are disastrously wrong lead to disastrous answers. This makes the search for good model checks important, but models are also vulnerable to subtle misspecification errors that are not easily picked up by model diagnostics. The following example is influential in survey sampling circles.

Example 7: A nonrobust model for disproportionate stratified sampling. Hansen, Madow, and Tepping (1983) considered estimators of the finite population mean $\overline{Y}$ of a survey variable Y, in the setting of disproportionate stratified sampling with an auxiliary variable X known for all units of the population. They considered the model-based prediction estimator obtained from the ratio model

$$\left[y_i \mid x_i, z_i = j, \beta, \sigma^2\right] \overset{\sim}{\text{iid}} N\left(\beta x_i, \sigma^2 x_i\right), \tag{4}$$

which leads (ignoring finite population corrections) to the simple ratio estimator $\overline{y}_{\mathrm{R}} = (\overline{y}/\overline{x})\overline{X}$, where $\overline{X}$ is the population mean of X and $(\overline{y}, \overline{x})$ are the unweighted sample means of Y and X that ignore the differential sampling weights across

strata. They conducted simulations comparing the performance of $\overline{y}_{\mathrm{R}}$ with the combined ratio estimate $\overline{y}_{\mathrm{CR}} = (\overline{y}_{\mathrm{st}}/\overline{x}_{\mathrm{st}})\overline{X}$, where $(\overline{y}_{\mathrm{st}}, \overline{x}_{\mathrm{st}})$ are stratified means that incorporate the sampling weights. If the ratio model (4) is true, $\overline{y}_{\mathrm{R}}$ is better than $\overline{y}_{\mathrm{CR}}$, so the sampling weights should be ignored. However, Hansen, Madow, and Tepping (1983) showed that the bias of $\overline{y}_{\mathrm{R}}$ can be serious even when diagnostic checks for whether β in the ratio model is constant across the strata suggest that assumption is plausible. Valliant, Dorfman, and Royall (2000) questioned Hansen et al.'s choice of diagnostics, but my view is that under disproportionate stratified sampling, a model like (4) that ignores stratum effects is too vulnerable to misspecification to be a sound basis for inference, unless there are convincing reasons to believe that stratum effects are not present. One setting where (4) is justified is when the strata are created using random numbers, since then stratum is clearly independent of outcomes. However, in practice strata are never created in this way, but rather are based on characteristics likely to be related to the survey outcomes. If the sample size is large, even a slight misspecification in (4) caused by minor differences in the distribution of Y between strata can induce a bias in $\overline{y}_{\mathrm{R}}$ that dominates mean squared error and corrupts confidence coverage.

Bayes is less convincing for model formulation and assessment than for inference under a given model: Bayesian model averaging is intuitive and compelling, but in any given problem there is still the problem of deciding the class of models over which averaging takes place, and how to choose the prior probabilities of models in the class. Bayesian hypothesis testing has the logic of Bayes's theorem in its favor, but comparing models of different dimension is tricky, and sensitive to the choice of priors. Strictly subjective Bayesians claim they can make this work, but the approach is a hard sell for scientific inference (and I suspect applied Bayesians "peek" at the data in practice). It seems unlikely to me that Bayesian model assessment can ever achieve the degree of clarity of Bayesian inference under an agreed model.

5. CALIBRATED BAYES: A POTENTIAL RESOLUTION OF THE BAYES/FREQUENTIST SCHISM

A crude summary of the strengths and weaknesses described above is given in Table 2: Bayesian statistics is strong for inference under an assumed model, but is relatively weak for the development and assessment of models. Frequentist statistics provides a useful tool for model development and assessment, but is a weak tool for inference under an assumed model. If this summary is accepted, then the natural compromise is to use frequentist methods for model development and assessment, and Bayesian methods for inference under a model. This capitalizes on the strengths of both paradigms, and is the essence of the approach known as calibrated Bayes.

Table 2. Summary of Strengths and Weaknesses of Bayes and Frequentist Paradigms

Activity	*Bayes*	*Frequentist*
Inference under assumed model	Strong	Weak
Model formulation/Assessment	Weak	Strong

Many statisticians have advanced the calibrated Bayesian idea; some examples are Peers (1965), Welch (1965), and Dawid (1982). I myself have assessed the frequentist properties of Bayesian procedures in methodological work (e.g., Little 1988). But two seminal papers by leading proponents of this school, Box (1980) and Rubin (1984), are required reading for those interested in this approach. Box (1980) wrote that

> I believe that . . . sampling theory is needed for exploration and ultimate criticism of the entertained model in the light of the current data, while Bayes's theory is needed for estimation of parameters conditional on adequacy of the model.

Box (1980) based his implementation of this idea on the factorization:

$$p(Y, \theta \mid M) = p(Y \mid M)\, p(\theta \mid Y, M),$$

where the second term on the right side is the posterior distribution of the parameter θ given data Y and model M, and is the basis for inference, and the first term on the right side is the marginal distribution of the data Y under the model M, and is used to assess the validity of M, with the aid of frequentist considerations. Specifically, discrepancy functions of the observed data $d(Y_{\mathrm{obs}})$ are assessed from the perspective of realizations from their marginal distribution $p(d(Y) \mid M)$. A questionable feature of this "prior predictive checking" is that checks are sensitive to the choice of prior distribution even when this choice has limited impact on the posterior inference; in particular it leads to problems with assessment of models involving noninformative priors.

Rubin (1984) wrote that

> The applied statistician should be Bayesian in principle and calibrated to the real world in practice—appropriate frequency calculations help to define such a tie . . . frequency calculations are useful for making Bayesian statements scientific, scientific in the sense of capable of being shown wrong by empirical test; here the technique is the calibration of Bayesian probabilities to the frequencies of actual events.

Rubin (1984) and Gelman, Meng, and Stern (1996) advocated model checking based on a different factorization than that of Box (1980), namely:

$$p(Y^*, \theta^*, \theta \mid Y_{\mathrm{obs}}, M) = p(Y^*, \theta^* \mid Y_{\mathrm{obs}}, \theta, M)\, p(\theta \mid Y_{\mathrm{obs}}, M),$$

where (Y^*, θ^*) is the realization of a future data and parameter values based on the posterior predictive distribution given model M and observed data Y_{obs}. This leads to posterior predictive checks (Rubin 1984; Gelman, Meng, and Stern 1996), which extend frequentist checking methods by not limiting attention to checking statistics that have a known distribution under the model. These checks involve an amalgam of Bayesian and frequentist ideas, but are clearly frequentist in spirit in that they concern embedding the observed data within a sequence of unobserved datasets that could have been generated under M, and seeing whether the observed data are "reasonable."

These methods have been criticized for using the data twice and hence not yielding tail-area P values that are uniformly distributed under the posited model (Bayarri and Berger 2000; Robins, van der Vaart, and Ventura 2000), but it seems to me that they nevertheless provide a promising avenue for model-checking.

What are the implications of the calibrated Bayes viewpoint for the examples in this article? Space precludes a detailed discussion, but some brief considerations follow:

Example 1 (continued). Tests of independence in a 2 × 2 contingency table. The standard Bayesian inference adds Beta priors for the success rates in the two groups and computes the posterior probability that $\pi_1 > \pi_2$. Proper prior distributions may be entertained in certain contexts; when there is little prior evidence about the success rates, the choice of "objective prior" has been debated, but Jeffreys's prior is one plausible conventional choice. The Fisher exact test P value corresponds to an odd choice of prior distribution (Altham 1969). Calibrated Bayes methods limit ambiguities in the reference set for frequentist assessments to model evaluation, rather than to model inference under a specified model. In particular, Gelman (2003) argued for posterior predictive checks that condition on the margin of the contingency table fixed by the design.

Example 2 (continued): Single sample t inference with a bound on precision. As a simple example of posterior predictive checking, Figure 3 displays the posterior predictive distribution of the sample variance s^{*2} under the normal model in Example 2, assuming a Jeffreys's prior for the parameters. The posterior probability of observing a sample variance in future datasets as low as the observed sample variance of $s^2 = 1$ is about 0.065, which is low but not exceptional, so the posited model seems not unreasonable.

Example 3 (continued): Penalties for peeking? Since the calibrated Bayes inference is Bayesian, there are no penalties for peeking in the inference—the inference is unaffected by interim analysis. On the other hand, interim analyses and stopping rules do increase sensitivity of the Bayesian inference to the choice of prior distribution (Rosenbaum and Rubin 1984), so they do have subtle implications for the robustness of the inference. Thus, models for datasets subject to interim analyses need to be carefully justified and checked.

Example 4 (continued): Multiple imputation combining rules for missing data. The calibrated Bayes inference is Bayesian, Rubin's (1987) MICRs are valid as approximations to this inference. On the other hand, in situations where multiple imputation is used in public use datasets, the imputation model needs to take into account the fact that the user may be adopting a different model or analysis. Rubin argued that the imputation model should be relatively "weak," in the sense of including rather than excluding covariates, arguing that it is worth sacrificing some efficiency to avoid imposing a strong imputation model on the user of the dataset.

Example 5 (continued): Survey weights in regression. The calibrated Bayes approach implies that survey inference should be model-based rather than design-based—with large samples the likelihood often dominates the prior, and the Bayesian approach yields similar results to the super-population modeling paradigm popular in model-based survey inference (Valliant, Dorfman, and Royall 2000). However, the calibrated Bayes perspective does encourage frequentist assessments of the properties of the Bayesian inferences, and in particular favors models that lead to estimates with good frequentist properties. One such property is design consistency (Brewer 1979; Isaki and Fuller 1982), which holds when an estimate tends to the estimand as the sample size increases, irrespective of the truth of the model. Restriction to models that yield design-consistent estimates (Little 1983, 2004; Firth and Bennett 1998) avoids models that ignore features of the survey design and are vulnerable to misspecification.

Thus, the calibrated Bayes approach leads to regression models for complex surveys that take explicit account of features of the sample design like stratification and weighting. In particular, it can be shown that design-weighted regression estimates, with the weights incorporating factors for nonconstant variance, can be justified as approximate Bayes for models that include effects for strata defined by different probabilities of selection (Little 1991; Little 2004, Example 11).

Example 6 (continued): Modeling the Census undercount. Complex models need to be developed that capture the intricacies of Census data collection. For an initial attempt at such a Bayesian model for combining Census, post-enumeration data and demographic analysis, see Elliott and Little (2005). The model assessment component would be helped by building research pseudo-populations of records from earlier censuses that form the basis for simulation assessments of different model procedures. Subjective elements of the model needed to be made explicit and resolved by consensus of experts, and in some cases, sensitivity analyses may be needed to assess the impact of alternative assumptions on Census answers.

Example 7 (continued) A nonrobust model for disproportionate stratified sampling. The model (4) criticized in Example 7 does not yield to a design-consistent estimate of the population mean, and hence is not appropriate from a calibrated Bayes perspective. A simple modification that achieves design consistency is to allow for differences in the regression coefficient across strata, as in the model

$$\left[y_i \mid x_i, z_i = j, \beta, \sigma^2\right] \overset{\text{iid}}{\sim} N\left(\beta_j, x_i, \sigma_j^2 x_i\right), \qquad (5)$$

which leads to the separate ratio estimator $\overline{y}_{\text{sr}} = \Sigma_j P_j \overline{y}_j \left(\overline{X}_j / \overline{x}_j\right)$, where in stratum j, P_j is the population proportion and $\overline{y}_j, \overline{x}_j, \overline{X}_j$ are, respectively, the sample means of Y and X and the population mean of X. This model is robust to misspecification in large samples. In small samples $\overline{y}_{\text{sr}}$ may be excessively noisy, but smoothing $\overline{y}_{\text{sr}}$ towards $\overline{y}_{\text{R}}$ can be achieved by assuming a prior distribution on the slopes $\{\beta_j\}$ with mean β and variance τ^2. The estimator from this random effects model is also design consistent, and might prove a strong competitor to $\overline{y}_{\text{sr}}$ or the estimator $\overline{y}_{\text{cr}}$ considered by Hansen, Madow, and Tepping (1983). The latter is a prediction estimator under model (5), but ignores the information in the known population means $\{\overline{X}_j\}$.

Fisherian significance tests have a role within the calibrated Bayes paradigm for model checking (Box 1980; Rubin 1984). For example, a global test of whether data are consistent with a null model is allowed without the need to specify the alternative hypothesis. On the other hand, classical hypothesis testing does not have a role for inference about model parameters—not in my view a serious loss. For an interesting recent assessment of

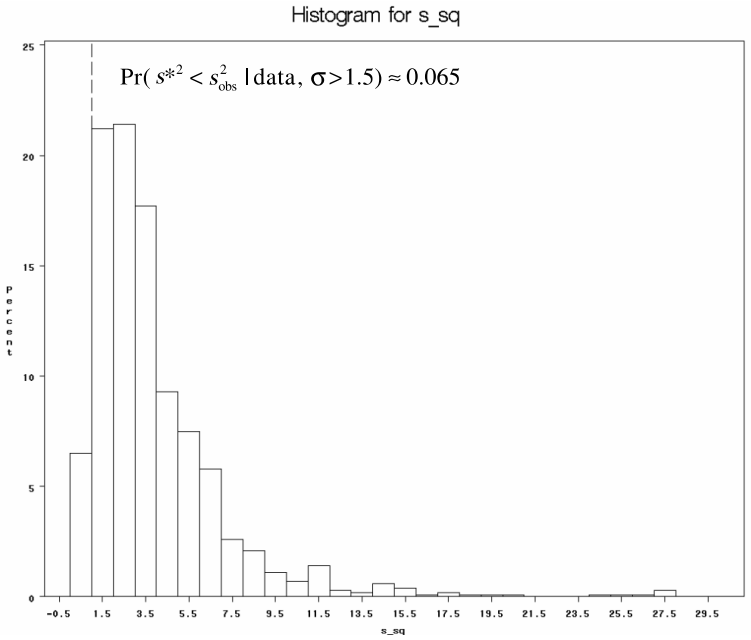

Figure 3. The posterior predictive distribution of the sample variance under the model of Example 2.

the various forms of hypothesis testing that lends support to this position, see Christensen (2005).

This is not to claim that calibrated Bayes solves all the problems of statistical inference. Ambiguities arise at the frontier between model inference and model checking. How much peeking at the data is allowed in developing the model without seriously corrupting the inference? When is model selection appropriate as opposed to model averaging (e.g., Draper 1995)? There remains much to argue about here, but I still think that the calibrated Bayes provides a useful roadmap for many problems of statistical modeling and inference.

6. IMPLICATIONS FOR THE TEACHING AND PRACTICE OF STATISTICS

If calibrated Bayes provides a useful roadmap, what are its implications for the future teaching and practice of statistics? I conclude by offering some thoughts on this issue:

1. Bayesian statistical methods need to be taught! Currently Bayesian statistics is absent or "optional" in many programs for training MS and even Ph.D. statisticians, and Ph.D. statisticians are trained with very little exposure to Bayesian ideas, beyond a few lectures in a theory sequence dominated by frequentist ideas. This is clearly incompatible with my roadmap, and it seems to me unconscionable given the prominence of Bayes in science, as evidenced by the strong representation of modern-day Bayesians in science citations (Science Watch 2002). As a first step, I would argue that a Bayes course should be a required component of any MS or Ph.D. program in statistics.

When it comes to consumers of statistics, Bayes is not a part of most introductory statistics courses, so most think of frequentist statistics as all of statistics, and are not aware that Bayesian inference exists. Defenders of the status quo claim that Bayesian inference is too difficult to teach to students with limited mathematical ability, but my view is that these difficulties are overrated. The basic idea of Bayes's theorem does not require calculus, and Bayesian methods seem to me quite teachable if the emphasis is placed on interpretation of models and results, rather than on the inner workings of Bayesian calculations. Indeed, Bayesian posterior credibility intervals have a much more direct interpretation than confidence intervals, as illustrated in Example 2. Frequentist hypothesis testing is no picnic to teach to consumers of statistics, for that matter!

2. More emphasis on statistical modeling over methods. Since the roadmap advocates model-based inference, it emphasizes statistical models over statistical methods. Formulating useful statistical models for real problems is not simple, and students need more instruction on how to fit models to complicated datasets. We need to elucidate the subtleties of model development. Issues include the following: (a) models with better fits can yield worse predictions than methods that fit the observed data better; (b) all model assumptions are not equal, for example, in regression lack of normality of errors is secondary to misspecification of the error variance, which is in turn secondary to misspecification of the mean structure; and (c) If inferences are to be Bayesian, more attention needs to be paid to the difficulties of picking priors in high-dimensional complex models, objective or subjective.

3. More attention is needed to assessments of model fit. Models are imperfect idealizations, and hence need careful checking; according to the roadmap, this is where frequentist methods have an important role. These methods include Fisherian significance tests of null models, diagnostics that check the model in directions that are important for the target inferences, and model-checking devices like posterior predictive checking and cross-validation. Such diagnostics are well known for regression, but perhaps less developed and taught for other models.

To summarize, Bayes and frequentist ideas are important for good statistical inference, and both sets of ideas need to be developed and taught. The calibrated Bayes compromise capitalizes on strengths of Bayes and frequentist paradigms. It leaves much to argue about, but it is a good roadmap for future advances.

[Received February 2006. Revised March 2006.]

REFERENCES

Altham, P. M. E. (1969), "Exact Bayesian Analysis of a 2 × 2 Contingency Table and Fisher's Exact Significance Test," *Journal of the Royal Statistical Society*, Series B, 31, 261–269.

Barnett, V. (1973), *Comparative Statistical Inference*, London: Wiley.

Basu, D. (1971), "An Essay on the Logical Foundations of Survey Sampling, Part 1," *Foundations of Statistical Inference*, Toronto: Holt, Rinehart and Winston, pp. 203–242.

Bayarri, M. J., and Berger, J. O. (2000), "P Values for Composite Null Models" (with discussion), *Journal of the American Statistical Association*, 95, 1127–1172.

Berger, J. M. (2000), "Bayesian Analysis: A Look at Today and Thoughts for Tomorrow," *Journal of the American Statistical Association*, 95, 1269–1276.

Bernardo (1979), "Reference Posterior Distributions for Bayesian Inference" (with discussion), *Journal of the Royal Statistical Society*, Series B, 41, 113–147.

Birnbaum, A. (1962), "On the Foundations of Statistical Inference" (with discussion), *Journal of the American Statistical Association*, 57, 269–326.

Box, G. E. P. (1980), "Sampling and Bayes Inference in Scientific Modelling and Robustness" (with discussion), *Journal of the Royal Statistical Society*, Series A, 143, 383–430.

Brewer, K. R. W. (1979), "A Class of Robust Sampling Designs for Large-Scale Surveys," *Journal of the American Statistical Association*, 74, 911–915.

Christensen, R. (2005), "Testing Fisher, Neyman, Pearson and Bayes," *The American Statistician*, 59, 121–126.

Cox, D. R. (1971), "The Choice Between Alternative Ancillary Statistics," *Journal of the Royal Statistical Society*, Series B, 33, 251–255.

Cox, D. R., and Hinkley, D. V. (1974), *Theoretical Statistics*, London: Chapman & Hall.

Dawid, A. P. (1982), "The Well-Calibrated Bayesian," *Journal of the American Statistical Association*, 77, 605–610.

Dawid, A. P., Stone, M., and Zidek, J. V. (1973), "Marginalization Paradoxes in Bayesian and Structural Inference" (with discussion), *Journal of the Royal Statistical Society*, Series B, 35, 189–233.

Draper, D. (1995), "Assessment and Propagation of Model Uncertainty" (with discussion), *Journal of the Royal Statistical Society*, Series B, 57, 45–97.

De Finetti, B. (1974), *Theory of Probability*, New York: Wiley.

DeMets, D. L., and Ware, J. H. (1980), "Group Sequential Methods for Clinical Trials with a One-Sided Hypothesis," *Biometrika*, 67, 651–660.

Edwards, A. W. F. (1972), *Likelihood*, Cambridge: Cambridge University Press.

Efron, B. (1986), "Why Isn't Everyone a Bayesian?" (with discussion), *The American Statistician*, 40, 1–11.

——— (2005), "Bayesians, Frequentists and Scientists," *Journal of the American Statistical Association*, 100, 1–5.

Elliott, M., and Little, R. J. (2005), "A Bayesian Approach to Census 2000 Evaluation Using A.C.E. Survey Data and Demographic Analysis," *Journal of the American Statistical Association*, 100, 380–388.

Fay, R. E. (1996), "Alternative Paradigms for the Analysis of Imputed Survey Data," *Journal of the American Statistical Association*, 91, 490–498.

Ferguson, T. S. (1967), *Mathematical Statistics: A Decision-Theoretic Viewpoint*, New York: Wiley.

Firth, D., and Bennett, K. E. (1998), "Robust Models in Probability Sampling," *Journal of the Royal Statistical Society*, Series B, 60, 3–21.

Gelman, A. (2003), "A Bayesian Formulation of Exploratory Data Analysis and Goodness-of-Fit Testing," *International Statistical Review*, 71, 369–382.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2003), *Bayesian Data Analysis* (2nd ed.), New York: CRC Press.

Gelman, A., Meng, X.-L., and Stern, H. (1996), "Posterior Predictive Assessment of Model Fitness Via Realized Discrepancies" (with discussion), *Statistica Sinica*, 6, 733–807.

Hacking, A. (1965), *The Logic of Statistical Inference*, Cambridge: Cambridge University Press.

Hansen, M. H., Madow, W. G., and Tepping, B. J. (1983), "An Evaluation of Model-Dependent and Probability-Sampling Inferences in Sample Surveys" (with discussion), *Journal of the American Statistical Association*, 78, 776–793.

Heckman, J. (1976), "The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables, and a Simple Estimator for Such Models," *Annals of Economic and Social Measurement*, 5, 475–492.

Holt, D., and Smith, T. M. F. (1979), "Poststratification," *Journal of the Royal Statistical Society*, Series A, 142, 33–46.

Isaki, C. T., and Fuller, W. A. (1982), "Survey Design Under the Regression Superpopulation Model," *Journal of the American Statistical Association*, 77, 89–96.

Leamer, E. (1978), *Specification Searches: Ad-Hoc Inferences with Experimental Data*, New York: Wiley.

Lindley, D. (1972), *Bayesian Statistics: A Review*, Philadelphia: SIAM.

——— (1986), Comment on "Why Isn't Everyone a Bayesian?" by B. Efron, *The American Statistician*, 40, 6–7.

Little, R. J. A. (1983), Comment on "An Evaluation of Model Dependent and Probability Sampling Inferences in Sample Surveys," by M. H. Hansen, W. G. Madow, and B. J. Tepping, *Journal of the American Statistical Association*, 78, 797–799.

——— (1988), "Small Sample Inference About Means from Bivariate Normal Data with Missing Values," *Computational Statistics and Data Analysis*, 7, 161–178.

——— (1989), "On Testing the Equality of Two Independent Binomial Proportions," *The American Statistician*, 43, 283–288.

——— (1991), "Inference With Survey Weights," *Journal of Official Statistics*, 7, 405–424.

——— (2004), "To Model or Not to Model? Competing Modes of Inference for Finite Population Sampling," *Journal of the American Statistical Association*, 99, 546–556.

Little, R. J. A., and Rubin, D. B. (2002), *Statistical Analysis with Missing Data* (2nd ed.), New York: Wiley.

Longini, I. M., Nizam, A., Xu, S., Ungchusak, K., Hanshaoworakul, W., Cummings, D., and Halloran, M. E. (2005), "Containing Pandemic Influenza at the Source," *Science*, 309, 1083–1087.

Meng, X.-L. (1994), "Multiple Imputation Inferences with Uncongenial Sources of Input" (with discussion), *Statistical Science*, 9, 538–573.

Peers, H. W. (1965), "On Confidence Points and Bayesian Probability Points in the Case of Several Parameters," *Journal of the Royal Statistical Society Series*, Series B, 27, 9–16.

Press, S. J. (1986), Comment on "Why Isn't Everyone a Bayesian?" by B. Efron, *The American Statistician*, 40, 9–10.

Rao, J. N. K. (1996), "On Variance Estimation with Imputed Data," *Journal of the American Statistical Association*, 91, 499–506.

Robins, J. M., and Ritov, Y. (1997), "Toward a Curse of Dimensionality Appropriate (CODA) Asymptotic Theory for Semi-parametric 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" (with discussion), *Journal of the American Statistical Association*, 95, 1143–1172.

Robins, J. M., and Wang, N. (2000), "Inference for Imputation Estimators," *Biometrika*, 87, 113–124.

Robins, J. M., and Wasserman, L. (2000), "Conditioning, Likelihood, and Coherence: A Review of Some Foundational Concepts," *Journal of the American Statistical Association*, 95, 1340–1346.

Robinson, G. K. (1975), "Some Counter-Examples to the Theory of Confidence Intervals," *Biometrika*, 62, 155–161.

Rosenbaum, P. R., and Rubin, D. B. (1984), "Sensitivity of Bayes Inference with Data-Dependent Stopping Rules," *The American Statistician*, 38, 106–109.

Royall, R. M. (1997), *Statistical Evidence: A Likelihood Paradigm*, Boca Raton: CDC Press.

Rubin, D. B. (1984), "Bayesianly Justifiable and Relevant Frequency Calculations for the Applied Statistician," *The Annals of Statistics*, 12, 1151–1172.

——— (1987), *Multiple Imputation for Nonresponse in Surveys*, New York: Wiley.

——— (1996), "Multiple Imputation After 18+ Years," *Journal of the American*

Statistical Association, 91, 473–489.

Samaniego, F. J., and Reneau, D. M. (1994), "Towards a Reconciliation of the Bayesian and Frequentist Approaches to Point Estimation," *Journal of the American Statistical Association*, 89, 947–957.

Savage, L. J. (1954), *The Foundations of Statistics*, New York: Wiley.

Science Watch (2002), "Vital Statistics on the Numbers Game: Highly Cited Authors in Mathematics, 1991–2001," *Science Watch*, 13, 3, p. 2.

Smith, T. M. F. (1976), "The Foundations of Survey Sampling: A Review" (with discussion), *Journal of the Royal Statistical Society*, Series A, 139, 183–204 .

Valliant, R., Dorfman, A. H., and Royall, R. M. (2000), *Finite Population Sampling and Inference: A Prediction Approach*, New York: Wiley.

Welch, B. L. (1965), "On Comparisons Between Confidence Point Procedures in the Case of a Single Parameter," *Journal of the Royal Statistical Society*, Series B, 27, 1–8.

Wilkinson, G. N. (1977), "On Resolving the Controversy in Statistical Inference" (with discussion), *Journal of the Royal Statistical Society*, Series B, 39, 119–171.

Yates, F. (1984), "Tests of Significance for 2 $\times$ 2 Contingency Tables" (with discussion), *Journal of the Royal Statistical Society*, Series A, 147, 426–463.