The Annals of Statistics
1994, Vol. 22, No. 3, 1142–1160

POSTERIOR PREDICTIVE *p*-VALUES[1]

BY XIAO-LI MENG

University of Chicago

This article is dedicated to my mentor, Donald B. Rubin, for his 50th birthday.

Extending work of Rubin, this paper explores a Bayesian counterpart of the classical p-value, namely, a tail-area probability of a "test statistic" under a null hypothesis. The Bayesian formulation, using posterior predictive replications of the data, allows a "test statistic" to depend on both data and unknown (nuisance) parameters and thus permits a direct measure of the discrepancy between sample and population quantities. The tail-area probability for a "test statistic" is then found under the joint posterior distribution of replicate data and the (nuisance) parameters, both conditional on the null hypothesis. This *posterior predictive p-value* can also be viewed as the posterior mean of a classical p-value, averaging over the posterior distribution of (nuisance) parameters under the null hypothesis, and thus it provides one general method for dealing with nuisance parameters. Two classical examples, including the Behrens–Fisher problem, are used to illustrate the posterior predictive p-value and some of its interesting properties, which also reveal a new (Bayesian) interpretation for some classical p-values. An application to multiple-imputation inference is also presented. A frequency evaluation shows that, in general, if the replication is defined by new (nuisance) parameters and new data, then the Type I frequentist error of an α-level posterior predictive test is often close to but less than α and will never exceed 2α.

1. Introduction. There is perhaps no single notion in statistics, other than the p-value, that has been so widely used and yet so seriously criticized for so long. The core of the controversy is whether a p-value provides adequate "evidence" against a null hypothesis. Recommendations range from cautious interpretation to complete dismissal. Even accepting the utility of the p-value as an inferential tool, the issue of how to handle nuisance parameters, as in the Behrens–Fisher problem, has generated much debate. The relevant literature is simply too extensive to list. Several discussion papers, including Cox (1977), Shafer (1982), Berger and Delampady (1987), Berger and Sellke (1987) and Casella and Berger (1987), could, however, be singled out for their wide scope, their prominence in the general debate and their comprehensive references and

Received October 1992; revised October 1993.

[1]This research was supported in part by NSF Grant DMS-92-04504. The manuscript was prepared using computer facilities supported in part by the NSF Grants DMS-89-05292, DMS-87-03942 and DMS-86-01732 awarded to the Department of Statistics at the University of Chicago, by the Fairchild Foundation and by the University of Chicago Block Fund.

AMS 1991 *subject classifications*. Primary 62F03; secondary 62A99.

Key words and phrases. Bayesian p-value, Behrens–Fisher problem, discrepancy, multiple imputation, nuisance parameter, pivot, p-value, significance level, tail-area probability, test variable, Type I error.

review of other important works. An interesting recent study of the p-value by Hwang, Casella, Robert, Wells and Farrell (1992) also contains a brief summary of the controversy about the p-value. By focusing primarily on the Behrens–Fisher problem, Wallace (1980) provides an enlightening review on ways of handling nuisance parameters from different perspectives.

As noted in Hwang, Casella, Robert, Wells and Farrell (1992), most criticism leveled at the p-value has come from the Bayesian school, especially because the calculation of a p-value almost always involves averaging over sample values that have not occurred, a clear violation of the likelihood principle [e.g., Jeffreys (1967), Berger and Wolpert (1984) and Berger and Delampady (1987)]. As a consequence, the notion of "Bayesian p-value" is often viewed as a "paradox." Despite this common attitude, however, several leading Bayesians [e.g., Dempster (1971, 1973), Box (1980) and Rubin (1984)] have argued that the p-value approach, namely, calculating a tail-area probability of a statistic, can be a useful tool even for Bayesian analysts in monitoring a model's adequacy. This has led to several formulations of "Bayesian p-value."

In particular, Rubin (1984) uses the posterior predictive distribution of a statistic to calculate the tail-area probability corresponding to the observed value of the statistic. We will call such a tail-area probability a *posterior predictive p-value* for obvious reasons. [The tail-area probability used by Box (1980) can be called a *prior predictive p-value*.] As Rubin (1984) pointed out, such a frequency calculation is *Bayesianly justifiable,* although not in the usual sense, and is *Bayesianly relevant* because it helps the process of model diagnosis, a fundamental part of any Bayesian analysis. One purpose of this paper is to illustrate the utility of the posterior predictive p-value from a different angle, that is, how the Bayesian formulation leads to a possible solution to the problem of nuisance parameters that the classical p-value approach often faces. We also extend Rubin's formulation by allowing a test statistic to depend on (nuisance) parameters. Such a parameter-dependent test statistic has been called a "(generalized) test variable" by Tsui and Weerahandi (1989), who introduced such a quantity for dealing with nuisance parameters in the context of significance testing with p-values. Their frequentist setting, however, forces them to impose the pivotality requirement on test variables in the sense that the resulting tail-area probabilities (with suitable supremum over primary parameters) are free of nuisance parameters. The Bayesian analogue discussed in this paper is free of such restrictions. Of course, as with any Bayesian approach, the price one pays for such a "freedom" is the specification of prior distributions. As shall be seen, however, the evaluation of a posterior predictive p-value for a given test variable only requires a prior distribution of the free (nuisance) parameters *under the null hypothesis*. Similar *partial* Bayesian formulations (e.g., only assigning priors for nuisance parameters) have been proposed in the literature [e.g., Cox (1975) and McCullagh (1990)] and seem to be more acceptable outside the Bayesian school, especially when noninformative prior distributions are used.

We emphasize that the formulation of posterior predictive p-values fundamentally inherits the classical construction of tail-area probabilities, and thus

it essentially is subject to the same debate, especially when compared to standard Bayesian approaches (e.g., Bayes factor). In particular, one should always bear in mind that a posterior predictive p-value, like any other p-value, is *not* a probability that the null hypothesis is true. A practical interpretation of it, when communicating with general users, is that it is a measure of *discrepancy* between the observed data and the posited assumptions, among which the hypothesis being tested is only a part. (Not being able to address directly the question of interest is an inherent part of the tail-area formulation that has been repeatedly criticized.) Even this practical interpretation has been seriously challenged as it is argued that a tail-area event typically includes sample points that provide much stronger evidence against a null hypothesis than the actual observation does; see, for example, Jeffreys (1967), Berger and Sellke (1987) and Berger and Delampady (1987).

There is no attempt in this paper to further discuss these issues, for it is not the author's intent to add anything new. Rather, following Dempster (1973), "It is accepted in this paper that significance testing has a legitimate place in the collection of inference tools...," our intention here is to illustrate how the Bayesian formulation can help to tackle the problem of nuisance parameters faced by the classical p-value approach without abandoning its main structures, which are simply too common to be ignored in the current statistical practice. For those who have called for complete dismissal of p-values (at least in the context of hypothesis testing), such a partial Bayesian approach is certainly on the "wrong track" as it helps a "devil." This paper, however, takes the opinion that as long as "subject matter journals are flooded with p-values" [Hwang, Casella, Robert, Wells and Farrell (1992)] and various ad hoc methods (e.g., inserting an estimate for the nuisance parameter) are being used, there is some benefit to have a unified approach for constructing traditionally motivated p-values that are always computable. An indoor animal, not necessarily a pet, is relatively easier to control than many wild creatures. Such a partial Bayesian approach also seems to fall into what Good (1992) called the Bayes/non-Bayes compromise, and it perhaps can help to promote the Bayesian ways of conducting statistical inference in general practice, where full Bayesian analyses are yet to be accepted as a standard approach. Minimally, by adjusting priors for nuisance parameters, it provides an operational device for constructing p-values with acceptable frequency properties.

The paper is organized as follows. Section 2 provides a formal definition of a posterior predictive p-value, and two different interpretations of it. Section 3 illustrates the approach of the posterior predictive p-value in two classical hypothesis problems concerning normal distributions, where a new interpretation of some classical p-values is also revealed. Section 4 applies the approach to multiple-imputation inference [Rubin (1987)], where such a partial Bayesian method is important and useful in deriving general procedures for obtaining significance levels. Section 5 presents some theoretical frequency evaluations for posterior predictive p-values.

2. Posterior predictive p-values.

2.1. *Definition.* In a classical setting, given a null hypothesis H_0: $\psi \in \Psi_0$ (the parameter space for ψ is Ψ) and a test statistic $T(X)$, a p-value is typically defined as

$$p = \Pr\{T(X) \geq T(x) \mid H_0\}. \tag{2.1}$$

In (2.1) the probability is taken over the sampling distribution of X under the null hypothesis, $f(X \mid \psi \in \Psi_0)$, with $T(x)$, the observed value of the test statistic, regarded as a constant. It is well-known that p of (2.1) is observable only if the null set Ψ_0 completely specifies the value of ψ (i.e., $\psi = \psi_0$) or, more generally, $T(X)$ is a pivotal quantity in the sense that its sampling distribution is free of any unknown parameter under the null hypothesis. Unfortunately, in many practical situations, this is not the case; in other words, one often faces the well-known problem of nuisance parameters. Various solutions, such as inserting estimates for the nuisance parameters or taking the supremum of p over the null set, have been provided. It should be emphasized, however, that the "p-values" resulting from these methods are not the tail-area *probabilities* that the classical approach had intended.

It perhaps is a bit ironical that the Bayesian formulation seems to be the most general way of defining and evaluating the tail-area probability *underlying* the classical setting: supposing the null hypothesis is true, if x^{rep} denotes a replication of x (i.e., a "future observation"), what is the probability that $T(x^{\text{rep}}) \geq T(x)$? In the Bayesian setting, this probability is

$$p_B = \Pr\{T(x^{\text{rep}}) \geq T(x) \mid x, H_0\}, \tag{2.2}$$

suppressing, but not neglecting, the dependence of p_B on the choice of T. In evaluating (2.2), the probability is taken over the posterior predictive distribution of x^{rep} conditional on H_0, that is,

$$f(x^{\text{rep}} \mid x, H_0) = \int_{\Psi_0} f(x^{\text{rep}} \mid \psi) \Pi_0(d\psi \mid x), \tag{2.3}$$

where $\Pi_0(\psi \mid x)$ is the posterior distribution of ψ under H_0. The posterior prediction in (2.3) simulates the replication under H_0 with the same (unknown) value of ψ that produced the current observed data, a replication that was intended in the classical approach. The fact that all the information about this unknown value of ψ comes from x and from the posited assumptions leads to the average in (2.3) according to $\Pi_0(\psi \mid x)$, the form of which depends on the nature of H_0.

For example, if Ψ_0 is a point hypothesis for a primary parameter $\theta \in \Theta$, that is, $\Psi_0 = \{(\theta_0, \nu)\colon \nu \in A\}$, where A is a subset of R^d, $d \geq 1$, then expression (2.3) becomes

$$f(x^{\text{rep}} \mid x, H_0) = \int_A f(x^{\text{rep}} \mid \theta_0, \nu) \pi_0(\nu \mid x)\, d\nu. \tag{2.4}$$

In (2.4), $\pi_0(\nu \mid x)$ is the posterior density of the nuisance parameter ν *conditional on* H_0:

$$\pi_0(\nu \mid x) = \frac{f(x \mid \theta_0, \nu)\pi(\nu \mid \theta_0)}{\int_A f(x \mid \theta_0, \nu)\pi(\nu \mid \theta_0)\, d\nu}, \qquad \nu \in A, \tag{2.5}$$

with $\pi(\nu \mid \theta_0)$ being a conditional prior density for ν given $\theta = \theta_0$, which can be improper. Similarly, if Ψ_0 is a composite hypothesis for θ, namely, $\Psi_0 = \{(\theta, \nu)\colon \theta \in \Theta_0,\ \nu \in A\}$, then (2.4) and (2.5) should be replaced, respectively, by

$$f(x^{\text{rep}} \mid x, H_0) = \int_{\Theta_0} \int_A f(x^{\text{rep}} \mid \theta, \nu)\pi_0(\theta, \nu \mid x)\, d\nu\, d\theta \tag{2.6}$$

and

$$\pi_0(\theta, \nu \mid x) = \frac{f(x \mid \theta, \nu)\pi(\nu \mid \theta)\pi_0(\theta)}{\int_{\Theta_0} \int_A f(x \mid \theta, \nu)\pi(\nu \mid \theta)\pi_0(\theta)\, d\nu\, d\theta}, \qquad \theta \in \Theta_0,\ \nu \in A, \tag{2.7}$$

where $\pi(\nu \mid \theta)$ is a conditional prior density of ν given θ and $\pi_0(\theta)$ is a prior density on Θ_0; both densities can be improper. Obviously, (2.4) and (2.5) can be viewed as a special case of (2.6) and (2.7), with Θ_0 only containing a single point.

Another advantage of the Bayesian formulation is that it naturally allows the use of a parameter-dependent "test statistic," which we call a *discrepancy variable* to emphasize the practical interpretation that a p-value is a measure of discrepancy. In other words, in defining a posterior predictive p-value, a classical test statistic $T(x)$ can be replaced by a discrepancy variable $D(x, \psi)$ that is a function of both data and parameters:

$$p_B = \Pr\{D(x^{\text{rep}}, \psi) \geq D(x, \psi) \mid x, H_0\}. \tag{2.8}$$

The probability in (2.8) is taken over the joint posterior distribution of (x^{rep}, ψ) given H_0, namely,

$$f(x^{\text{rep}}, \psi \mid x, H_0) = f(x^{\text{rep}} \mid \psi)\pi_0(\psi \mid x), \qquad \psi \in \Psi_0,$$

where $\pi_0(\psi \mid x)$ is the posterior density (probability) of ψ conditional on H_0. This generalization of (2.2) is important as it allows us to measure directly the discrepancy between sample quantities and population quantities when checking the discrepancy between the data and the assumptions. In fact, a classical test statistic $T(x)$ can often be viewed as a discrepancy variable between sample quantities and the "best fit" of the corresponding population quantities under the null hypothesis, that is, $T(x) = D(x, \hat{\psi}_0)$, with $\hat{\psi}_0$ being an efficient estimate (e.g., MLE) of ψ under H_0. This "best fitting" approach is useful since it eliminates the *first-level* dependence on unknown (nuisance) parameters, namely, the dependence of a discrepancy variable on the unknown parameters. However, it does not solve the entire problem because typically there is a *second-level* dependence—the sampling distribution of $T(x)$ will still depend on the unknown (nuisance) parameters. The posterior predictive p-value given in (2.8) takes care of this two-level dependence at once.

2.2. *An alternative interpretation.* An alternative interpretation of p_B is perhaps more appealing to those who are used to the classical setting. For simplicity, take a point hypothesis case [corresponding to (2.4)] as an example. Suppose for the moment that the value of the nuisance parameter ν is given. Then $D(x, \psi)$, regardless of whether it is a function of ν or not, is a test statistic in the classical sense, and its associated p-value is

$$p(\nu) \equiv \Pr\{D(X, \psi) \geq D(x, \psi) \mid \theta_0, \nu\}, \tag{2.9}$$

where the probability is taken over the sampling distribution, $f(X \mid \theta_0, \nu)$. Notice that $p(\nu)$ depends on ν not only through (possibly) the discrepancy variable, but also through (possibly) its sampling distribution. Since the nuisance parameter ν is unknown, taking $p(\nu)$ as a quantity of interest, one naturally desires to estimate it. One obvious approach to a Bayesian is to take the posterior mean of $p(\nu)$ over the posterior distribution of ν conditional on H_0. This leads to $E[p(\nu) \mid x, H_0]$, which is easily seen to be identical with p_B defined in (2.8).

Reflected in this interpretation is a "two-stage" perspective. At the first stage, one deals with a *testing* problem (in the classical sense) for a *conditional hypothesis* where ν is treated as given. At the second stage, one then deals with an *estimation* problem to handle the fact that ν is unknown. This seems quite logical (if we accept the logic underlying the classical approach for the first stage), at least when θ and ν are distinct, since we are never interested in testing ν. The possible a priori dependence of θ and ν can be taken care of by proper specification of the conditional prior density $\pi_0(\nu \mid \theta)$ (in fact, this only needs to be specified for $\theta = \theta_0$). Such a two-stage construction is also implied in the ideal (Bayesian) evidence againts (or for) H_0: $\Pr(H_0 \mid x) = E_{(\nu \mid x)}[\Pr(H_0 \mid x, \nu) \mid x]$, although there, such a conceptual separation is usually unnecessary as both stages follow the same kind of posterior calculations [i.e., *testing* in the Bayesian setting can be treated merely as an *estimation* of an indicator function; see Hwang, Casella, Robert, Wells and Farrell (1992) for more discussion on this perspective]. As shall be illustrated, this two-stage derivation is also very useful in computation as well as in theoretical studies.

There are several interesting by-products from this perspective. First, it leads to a measure of uncertainty in the classical p-value due to the unknown nuisance parameter. This measure is generated by the posterior distribution of $p(\nu)$ conditional on H_0. If $p(\nu)$ does not depend on ν, then this distribution is degenerate at a single value, the (observable) classical p-value. In general, it describes how the *conditional p-value* $p(\nu)$ is distributed as a function of ν. This posterior distribution has a mean p_B, as defined in (2.8), and has a variance $V_p \equiv \mathrm{Var}[p(\nu) \mid x, H_0]$ that provides a measure of spread around p_B. A useful upper bound for V_p is $p_B(1 - p_B)$. As expected, the posterior distribution of $p(\nu)$ provides a more complete picture of the measure of discrepancy under the null hypothesis than the posterior mean p_B alone, as shall be illustrated in Section 3.1.

Second, it is clear that, for any conditional prior density $\pi(\nu \mid \theta_0)$ that leads

to a proper posterior density $\pi_0(\nu \mid x)$,

$$p_B \leq \sup_{\nu} p(\nu),$$

so that p_B is always less conservative than the "worst scenario" approach [e.g., Bahadur (1965)]. Furthermore, since $V_p \to 0$ asymptotically under standard regularity conditions, a posterior predictive p-value for a classical test statistic is asymptotically equivalent to its classical p-value with inserted (efficient) estimate of ν. These observations may be useful in establishing theoretical results for posterior predictive p-values that are analogous to those for classical p-values [e.g., Bahadur (1967a, b), Bahadur and Raghavachari (1972), Bahadur, Chandra and Lambert (1984) and Hwang, Casella, Robert, Wells and Farrell (1992)].

Third, the two-stage perspective also suggests some choices of "test statistics," or discrepancy variables in the terminology of this paper, that have not been generally studied. For example, since at the first stage ν is treated as given, a *conditional likelihood ratio* (CLR) seems appropriate, at least when θ and ν are distinct:

$$D^C(x, \psi) = \frac{\sup_{\theta \notin \Theta_0} f(x \mid \theta, \nu)}{\sup_{\theta \in \Theta_0} f(x \mid \theta, \nu)}, \tag{2.10}$$

where Θ_0 can be either a point or a composite null hypothesis for θ. (The term "conditional likelihood" is used here in the Bayesian sense, corresponding to the term "conditional hypothesis.") If $\theta \in \Theta_1$ is the specified alternative, then the numerator in (2.10) is replaced by the supremum over $\theta \in \Theta_1$. The dependence of D^C on ν is its key difference from the usual *generalized likelihood ratio* (GLR):

$$D^G(x) = \frac{\sup_{\theta \notin \Theta_0} \sup_{\nu} f(x \mid \theta, \nu)}{\sup_{\theta \in \Theta_0} \sup_{\nu} f(x \mid \theta, \nu)}. \tag{2.11}$$

The choice of discrepancy variables is generally a difficult problem, which has been discussed extensively in the literature; for discussions that are most relevant to the current setting, see Box (1980), Rubin (1984) and Gelman, Meng and Stern (1993). Since the choices of discrepancy variables are influenced by the specifications of alternative hypotheses and since this paper is primarily concerned with defining and evaluating a p-value under a null hypothesis with a given discrepancy variable, we will restrict ourselves to CLR (and GLR) in the following illustrations and applications. More study is needed, for example, on general comparisons among posterior predictive p-values from CLR and GLR and from other forms of discrepancy variables (e.g., an appropriately defined Bayes factor). Another problem worthy of further study, as with other Bayesian approaches, is the sensitivity of posterior predictive p-values to the specifications of priors in finite (small) samples. Again, for the purpose of illustration, we will use standard noninformative priors in this paper, especially because, as in other settings, they lead to results that are (nearly) identical with classical solutions.

3. Two theoretical examples.

3.1. *One-sample normal mean.* As a classical example appearing in almost any introductory textbook in statistics, suppose we have a simple random sample of size n from $N(\mu, \sigma^2)$, and we are interested in testing H_0: $\mu = \mu_0$ with σ^2 unspecified. Since a posterior predictive p-value is invariant under any strictly monotone data-free transformation of a discrepancy variable, using CLR of (2.10) and GLR of (2.11) for the current problem is equivalent to using

$$D^C(x, \psi) = \frac{n(\bar{x} - \mu_0)^2}{\sigma^2} \quad \text{and} \quad D^G(x) = \frac{n(\bar{x} - \mu_0)^2}{s^2},$$

respectively, where $\bar{x}$ and s^2 are the sample mean and sample variance of $x = \{x_1, \ldots, x_n\}$, and $\psi = (\mu, \sigma^2)$. The corresponding conditional p-values are

$$\begin{aligned} p^C(\sigma^2) &\equiv \Pr\{D^C(X, \psi) \geq D^C(x, \psi) \mid \mu_0, \sigma^2\} \\ &= \Pr\left\{\chi_1^2 \geq \frac{n(\bar{x} - \mu_0)^2}{\sigma^2}\right\} \end{aligned} \tag{3.1}$$

and

$$\begin{aligned} p^G &= \Pr\{D^G(X) \geq D^G(x) \mid \mu_0, \sigma^2\} \\ &= \Pr\left\{F_{1,n-1} \geq \frac{n(\bar{x} - \mu_0)^2}{s^2}\right\} \equiv \Pr\{F_{1,n-1} \geq T(x)\}, \end{aligned} \tag{3.2}$$

using the conventional notation for chi-squared and F variables. Because $T(X)$ is a pivotal quantity, the conditional p-value (3.2) and thus the posterior predictive p-value based on GLR is identical with the corresponding classical p-value based on the t-test.

The posterior predictive p-value of CLR depends on the choice of the conditional prior $\pi(\sigma^2 \mid \mu_0)$. In cases where a proper prior is not available, an improper prior is often used on the ground of "noninformativeness." A common improper prior for the current setting is $\pi(\sigma^2 \mid \mu_0) \propto \sigma^{-2}$, which is the Jeffreys prior from the null model $N(\mu_0, \sigma^2)$. Under this prior,

$$[\sigma^2 \mid x, H_0] \sim \frac{ns_0^2}{\chi_n^2}, \tag{3.3}$$

where $s_0^2 = (1/n)\sum_{i=1}^n (x_i - \mu_0)^2$ is the MLE of σ^2 under the null model. It follows from (3.1) and (3.3) that

$$\begin{aligned} p_B^C &= E\left[p^C(\sigma^2) \mid x, H_0\right] \\ &= \Pr\left\{F_{1,n} \geq \frac{n(\bar{x} - \mu_0)^2}{s_0^2}\right\} \equiv \Pr\{F_{1,n} \geq T_0(x)\}. \end{aligned} \tag{3.4}$$

The (slight) difference between p_B^C and p_B^G [$= p^G$ of (3.2)] is expected because of the use of the different discrepancy variables; this contrasts with the classical calculations where $T_0(x)$ and $T(x)$ are equivalent test statistics. There is,

however, an interesting connection between CLR and the classical p-value from the t-test (i.e., p_B^G). Suppose that besides $\pi(\sigma^2 \mid \mu) \propto \sigma^{-2}$ we also specify $\pi(\mu) \propto 1$, corresponding to the standard joint noninformative prior, $\pi(\mu, \sigma^2) \propto \sigma^{-2}$, often used for *estimation* [e.g., Box and Tiao (1973), Chapter 2]. Then the marginal posterior density of σ^2, which can also be viewed as the posterior density of σ^2 under the alternative (i.e., $\mu \neq \mu_0$), is

$$[\sigma^2 \mid x] \sim \frac{(n-1)s^2}{\chi^2_{n-1}}.$$

If we use this posterior density to find the average of $p^C(\sigma^2)$ in (3.1), we recover the classical p-value of (3.2). In other words, if the information $\mu = \mu_0$ is ignored in estimating the variance, then p_B^C coincides with the classical p-value from the t-test. This interesting "coincidence" is somewhat unexpected, as it seems more appealing, at least from a classical testing point of view, to condition on all known quantities under the null hypothesis in quantifying the replication under which a tail-area probability is evaluated. This phenomenon is perhaps also worth further study, since the calculation under the normal assumption that leads to it is also a part of many (large-sample) evaluations of p-values based on the Wald type of statistics.

Another interesting point arises when $n = 1$. With only one observation and a noninformative prior on σ^2, there is no information to suggest which values of $p^C(\sigma^2)$ are most likely. In other words, the posterior distribution of $p^C(\sigma^2)$ should be uniform on $[0, 1]$, implying $p_B^C = \frac{1}{2}$. This is indeed the case, as can be verified by noticing that the posterior distribution of $p^C(\sigma^2)$ [under (3.3)] is the same as the distribution of a random variable (given x),

$$u = 1 - F_1\left(T_0(x)\frac{\chi^2_n}{n}\right),$$

where $F_1(t)$ is the cumulative distribution function (c.d.f.) of χ^2_1. When $n = 1$, $T_0(x) \equiv 1$, and thus $u = 1 - F_1(\chi^2_1) = U[0, 1]$. The fact that $p_B^C = \frac{1}{2}$ can also be verified directly from (3.4), but it is less informative than the fact that the whole posterior distribution of $p^C(\sigma^2)$ is uniform. With a proper prior on σ^2, however, even a single observation will provide some evidence against (or in support of) the null hypothesis, although such evidence may be very diffuse. The posterior distribution of $p^C(\sigma^2)$ provides a way of describing the diffuseness in the "evidence" summarized by $p^C(\sigma^2)$. This is certainly not feasible with the classical t-test, which is not defined when $n = 1$ whether or not one has prior information on σ^2.

3.2. *Two-sample normal means.* A classical example of the difficulty with nuisance parameters is the Behrens–Fisher problem, where it is known that no useful pivotal quantities exist. Suppose we have independent simple random samples from two normal populations, $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$, with sample sizes n_1 and n_2, respectively. We are interested in testing $\mu_1 = \mu_2$ with both σ_1^2

and σ_2^2 completely unspecified; in the notation of Section 2, $\Theta_0 = \{(\mu, \mu),\ -\infty < \mu < \infty\} \subset \Theta = \{(\mu_1, \mu_2),\ -\infty < \mu_1, \mu_2 < \infty\}$. It is easy to check that the CLR of (2.10) in this case is a monotone function of

$$D^C(x, \psi) = \frac{(\bar{x}_1 - \bar{x}_2)^2}{(\sigma_1^2/n_1) + (\sigma_2^2/n_2)}, \tag{3.5}$$

where $\bar{x}_1$ and $\bar{x}_2$ are sample means of the first and second sample, respectively, and $\psi = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$.

Given the free parameters under the null hypothesis, namely $(\mu, \sigma_1^2, \sigma_2^2)$, it is easy to see that the conditional p-value for D^C is

$$p^C(\sigma_1^2, \sigma_2^2) = \Pr\left\{\chi_1^2 \geq \frac{(\bar{x}_1 - \bar{x}_2)^2}{(\sigma_1^2/n_1) + (\sigma_2^2/n_2)}\right\}, \tag{3.6}$$

which happens not to depend on μ, the common mean. Choosing the standard noninformative priors $\pi(\sigma_1^2, \sigma_2^2 \mid \mu) \propto \sigma_1^{-2}\sigma_2^{-2}$ and $\pi(\mu) \propto 1$, the joint posterior distribution of σ_1^2 and σ_2^2 given μ is the same as the distribution (given the data) of

$$\sigma_1^2 = \frac{(n_1 - 1)s_1^2 + n_1(\bar{x}_1 - \mu)^2}{\chi_{n_1}^2} \quad \text{and} \quad \sigma_2^2 = \frac{(n_2 - 1)s_2^2 + n_2(\bar{x}_2 - \mu)^2}{\chi_{n_2}^2}, \tag{3.7}$$

where s_1^2 and s_2^2 are the sample variances of the first and second sample, respectively; $\chi_{n_1}^2, \chi_{n_2}^2$ and μ are mutually independent and μ has a "combined t" distribution [Box and Tiao (1973), Chapter 9]

$$\pi_0(\mu \mid x) \propto \left[1 + \frac{1}{n_1 - 1}\left(\frac{\mu - \bar{x}_1}{s_1/\sqrt{n_1}}\right)^2\right]^{-n_1/2} \left[1 + \frac{1}{n_2 - 1}\left(\frac{\mu - \bar{x}_2}{s_2/\sqrt{n_2}}\right)^2\right]^{-n_2/2} \tag{3.8}$$

This distribution is called the combined t because its limit, as n_1 and n_2 become large, is the "combined" normal from $N(\bar{x}_1, s_1^2/n_1)$ and $N(\bar{x}_2, s_2^2/n_2)$:

$$\mu \dot{\sim} N\left(\frac{\bar{x}_1(s_1^2/n_1)^{-1} + \bar{x}_2(s_2^2/n_2)^{-1}}{(s_1^2/n_1)^{-1} + (s_2^2/n_2)^{-1}},\ \left[(s_1^2/n_1)^{-1} + (s_2^2/n_2)^{-1}\right]^{-1}\right). \tag{3.9}$$

The intuition behind (3.8) and (3.9) is clear. Since $\mu_1 = \mu_2$ under the null hypothesis, an intuitive estimate of the common mean μ is the combined sample means weighted by their precisions $n_j/s_j^2, j = 1, 2$. The density (3.8) is also a special case of the so-called poly-t densities [e.g., Broemeling and Abdullah (1984)].

Combining (3.6) with (3.7) and (3.8) and letting $\tilde{s}_j^2 = (1 - n_j^{-1})s_j^2$, $j = 1, 2$, we have

$$p_B^C = \Pr\left\{F_{1, n_1 + n_2} \geq \frac{(\bar{x}_1 - \bar{x}_2)^2(n_1 + n_2)}{\left[\tilde{s}_1^2 + (\bar{x}_1 - \mu)^2\right]B_{n_1, n_2}^{-1} + \left[\tilde{s}_2^2 + (\bar{x}_2 - \mu)^2\right]\left(1 - B_{n_1, n_2}\right)^{-1}}\right\}, \tag{3.10}$$

where the probability is taken over three mutually independent variables: an F variable F_{1,n_1+n_2}, a Beta variable $B_{n_1,n_2} = \text{Beta}(n_1/2, n_2/2)$, and a combined-$t$ variable μ having the density (3.8). The evaluation of p_B^C can be easily accomplished by Monte Carlo simulation, which perhaps is the most practical method for computing posterior predictive (and other) p-values in general; see Gelman, Meng and Stern (1993) for further discussion.

The posterior predictive p-value of (3.10) is slightly different from the following well-known solution:

$$\tilde{p} = \Pr\left\{F_{1,n_1+n_2-2} \geq \frac{(\bar{x}_1 - \bar{x}_2)^2(n_1+n_2-2)}{\tilde{s}_1^2 B_{n_1-1,n_2-1}^{-1} + \tilde{s}_2^2\left(1 - B_{n_1-1,n_2-1}\right)^{-1}}\right\}, \tag{3.11}$$

where F_{1,n_1+n_2-2} and B_{n_1-1,n_2-1} are independent. This $\tilde{p}$ can be viewed as Jeffreys' Bayesian solution [Jeffreys (1967)] or the Behrens–Fisher fiducial solution] (e.g., Wallace (1980)], depending on one's perspective. It can also be obtained by using Tsui and Weerahandi's (1989) generalized p-value approach. Interestingly, as in Section 3.1, this $\tilde{p}$ is also numerically identical to a posterior average of the conditional p-value of (3.6) if the information $\mu_1 = \mu_2$ is ignored in deriving the posterior distribution of (σ_1^2, σ_2^2). Specifically, under the standard joint noninformative prior $\pi(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2) \propto \sigma_1^{-2}\sigma_2^{-2}$ [Box and Tiao (1973), Chapter 2], the marginal posterior of (σ_1^2, σ_2^2) is

$$\left[(\sigma_1^2, \sigma_2^2) \,|\, x_1, x_2\right] \sim \left[\frac{(n_1-1)s_1^2}{\chi_{n_1-1}^2}, \frac{(n_2-1)s_2^2}{\chi_{n_2-1}^2}\right], \tag{3.12}$$

where the two chi-square variables are independent. Taking the posterior mean of $p^C(\sigma_1^2, \sigma_2^2)$ of (3.6) with respect to (3.12) yields the $\tilde{p}$ of (3.11). Although p_B^C and $\tilde{p}$ are essentially identical when the sample sizes are not too small, their difference may be of theoretical interest since there seems no obvious nonrandomized classical test statistic that could reproduce $\tilde{p}$ as a (correct) posterior predictive p-value.

4. An application to multiple-imputation inference.

4.1. *General setting of multiple imputation.* A problem that is closely related to the two-sample problem in Section 3.2 arises in multiple-imputation inference [Rubin (1987)]. Multiple inputation is a general and efficient method for handling incomplete data, especially for handling nonresponse in large-sample surveys to produce public-use databases. Very recent overviews of multiple imputation are given in Meng (1994) and in Rubin (1995). In the multiple imputation framework, one first builds a model, explicitly or implicitly, that describes the predictive distribution (e.g., posterior predictive distribution) of missing data given the observed data and all relevant information. One then draws independently $m \geq 2$ sets of missing values from this distribution as m sets of imputations. Each set of imputed values together with the fixed set of observed data forms a completed-data set, to which one can apply standard

complete-data techniques to compute, say, the MLE of θ, a k-dimensional quantity of interest, and the large-sample variance associated with the MLE.

Let $\hat{\theta}_1, \ldots, \hat{\theta}_m$ be those m estimates with $U_1, \ldots, U_m$ as the associated variances, respectively. The multiple imputation estimate of θ is then [Rubin (1987), Chapter 3]

$$\bar{\theta}_m = \frac{1}{m} \sum_{l=1}^{m} \hat{\theta}_l$$

with associated variance

$$T_m = \bar{U}_m + \left(1 + \frac{1}{m}\right) B_m,$$

where

$$\bar{U}_m = \frac{1}{m} \sum_{l=1}^{m} U_l$$

measures the *within imputation* variability, and

$$B_m = \frac{1}{m-1} \sum_{l=1}^{m} (\hat{\theta}_l - \bar{\theta}_m)(\hat{\theta}_l - \bar{\theta}_m)^T$$

measures the *between imputation* variability.

Assuming the size of the observed data x_{obs} is large and the imputation is *proper* [see Rubin (1987), Chapter 4], the following approximations hold:

$$[\hat{\theta}_1, \ldots, \hat{\theta}_m \mid x_{\text{obs}}] \sim_{\text{i.i.d.}} N(\bar{\theta}_\infty, B_\infty), \tag{4.1}$$

$$[U_1, \ldots, U_m \mid x_{\text{obs}}] \sim_{\text{i.i.d.}} [\bar{U}_\infty, \ll B_\infty], \tag{4.2}$$

where $\bar{\theta}_\infty, B_\infty$ and $\bar{U}_\infty$ are the limits of $\bar{\theta}_m, B_m$ and $\bar{U}_m$, respectively, as $m \to \infty$ (conditional on x_{obs}), and $[A, \ll C]$ means the distribution is centered around A with lower order of variability than C. Furthermore, the large-sample distributions of $\bar{\theta}_\infty, B_\infty$ and $\bar{U}_\infty$, as functions of x_{obs}, can be approximated by

$$[\bar{\theta}_\infty \mid \psi] \sim N(\theta, U_0 + B_0), \tag{4.3}$$

$$[B_\infty \mid \psi] \sim [B_0, \ll U_0 + B_0], \tag{4.4}$$

$$[\bar{U}_\infty \mid \psi] \sim [U_0, \ll U_0 + B_0], \tag{4.5}$$

where $\psi = (\theta, U_0, B_0)$ consists of the primary parameter θ and nuisance parameter $\nu = (U_0, B_0)$. If we ignore the lower order of variability, we can assume from (4.2) and (4.5) that $U_0 = \bar{U}_m$, and from (4.4) that $B_0 = B_\infty$. It follows then that

$$[\bar{\theta}_m \mid \psi] \sim N\left(\theta, U_0 + \left(1 + \frac{1}{m}\right) B_0\right) \tag{4.6}$$

and, independently,

$$[B_m \mid \psi] \sim \frac{1}{m-1} B_0^{1/2} W_{m-1} B_0^{1/2}, \tag{4.7}$$

where W_{m-1} is a Wishart random variable with $m - 1$ degrees of freedom. Notice that in (4.6) and (4.7) the only nuisance parameter is B_0, since $U_0 (= \bar{U}_m)$ is treated as known.

4.2. *Significance levels from multiple imputation.* The estimates $\bar{\theta}_m$ and T_m given in the previous section provide a simple method for combining m sets of completed-data estimates and variances to produce a valid inference. In practice, standard complete-data analyses often also include calculations of significance levels. This suggests a need for corresponding procedures within the multiple-imputation framework for calculating significance levels from multiply imputed data sets. This task is more complicated primarily because the number of multiple imputations is commonly small (e.g., $m \leq 5$) in practice, and thus standard large-sample theory (with respect to m) does not apply. Bayesianly motivated methods for computing significance levels are thus crucial in order to ensure their frequency calibrations in general. The approach of posterior predictive p-value appears to be one such method.

Suppose one is interested in testing $\theta = \theta_0$ from (4.6) and (4.7). For simplicity, assume θ is a scalar quantity, and hence $W_{m-1} = \chi^2_{m-1}$. As a consequence of (4.6), the CLR is a monotone function of

$$D^C(x, \psi) = \frac{(\bar{\theta}_m - \theta_0)^2}{U_0 + (1 + 1/m)B_0}, \tag{4.8}$$

which yields (use $U_0 = \overline{U}_m$)

$$p^C(B_0) = \Pr\left\{\chi^2_1 \geq \frac{(\bar{\theta}_m - \theta_0)^2}{\overline{U}_m + (1 + 1/m)B_0}\right\}. \tag{4.9}$$

To determine p^C_B, one needs the conditional posterior distribution of the nuisance parameter B_0 given $\theta = \theta_0$ and the data x, which consists of the observed data and all imputed data. Under the noninformative prior $\pi(B_0 \mid \theta) \propto B_0^{-1}$ [in fact, $\pi(B_0 \mid \theta_0)$ is all one needs], it is easy to verify that

$$\pi_0(B_0 \mid x) \propto B_0^{-(m+1)/2} \exp\left\{-\frac{(m-1)B_m}{2B_0}\right\} \times \left[\overline{U}_m + \left(1 + \frac{1}{m}\right)B_0\right]^{-1/2} \exp\left\{-\frac{(\theta_0 - \bar{\theta}_m)^2}{2\left[\overline{U}_m + (1 + 1/m)B_0\right]}\right\},$$

that is, it is a product of an inverse chi-square density and a Pearson type V density [e.g., Johnson and Kotz (1970), page 12]. The calculation of p^C_B then follows from averaging $p^C(B_0)$ over B_0 according to $\pi_0(B_0 \mid x)$ above, which can be accomplished by a Monte Carlo simulation.

As before, a measure that is different from p^C_B can be obtained by ignoring $\theta = \theta_0$ in deriving posterior distribution of B_0. Specifically, under the additional prior assignment $\pi(\theta) \propto 1$, the marginal posterior distribution of B_0 is

$$\pi(B_0 \mid x) \sim (m-1)B_m \chi^{-2}_{m-1}.$$

Averaging (4.9) with respect to this posterior density leads to

$$\tilde{p} = \Pr\left\{\chi^2_1 \geq \frac{(\bar{\theta}_m - \theta_0)^2}{\overline{U}_m + (1 + 1/m)(m-1)B_m \chi^{-2}_{m-1}}\right\},$$

where the two chi-square variables are independent. This $\tilde{p}$ is closely related to the $\tilde{p}$ of (3.11) by assigning one of the samples in (3.11) an infinite number of observations. It is also identical to the Bayesian p-value given in Rubin ([1987), Chapter 3]—the posterior probability of all θ whose posterior density does not exceed the posterior density at θ_0. An additional consideration when comparing p_B^C with $\tilde{p}$ in this case is that $\pi_0(B_0 \mid x)$ depends on the approximation $\overline{U}_m = U_0$, whereas $\pi(B_0 \mid x)$ does not. With large m, the difference between p_B^C and $\tilde{p}$ will not be of practical concern, and $\tilde{p}$ might be preferred because its computation is slightly easier. However, in typical applications of multiple imputation, m is small (e.g., $m \leq 5$), in which case the difference between p_B^C and $\tilde{p}$ may be worth further quantitative investigation. Multivariate (i.e., when θ is multidimensional) extensions of p_B^C and $\tilde{p}$ are straightforward and are closely related to the multivariate Behrens–Fisher problem [Li, Meng, Raghunathan and Rubin (1991)]. Various approximations to $\tilde{p}$ and related calculations are presented in Rubin (1987), Raghunathan (1987), Meng (1988, 1990), Li, Raghunathan and Rubin (1991) and Meng and Rubin (1992). Those results may be useful in approximating p_B^C.

We emphasize that since the Bayesian formulation used above for deriving p_B^C and $\tilde{p}$ may not be a part of analyses by users of these procedures, the full Bayesian analysis, although typically more desirable, cannot be recommended as a general procedure. For example, in analyzing multiply imputed data sets, an investigator may perform a standard regression analysis on each completed data set using an existing sortware and then may use p_B^C or $\tilde{p}$ as a *procedure* to calculate the significance level for testing whether a regression slope is zero. The investigator may not want to perform a full Bayesian analysis for various reasons, but may be willing to use a Bayesianly motivated frequentist procedure that has good frequency properties, such as the ones demonstrated in the next section for posterior predictive p-values.

5. Frequency evaluation. In the classical setting, use of p-values is often associated with a preselected nominal level α, typically 0.01, 0.05 or 0.1. A decision to reject the null hypothesis is made if $p \leq \alpha$. It is often stated that such an α controls the Type I error—the probability of wrongly rejecting the null hypothesis based on an α-level test will never exceed α—because, under the null hypothesis, the sampling distribution of p is either uniform or stochastically larger than uniform, depending on whether or not the sampling distribution of the test statistic is continuous, and thus [see, e.g., Bahadur and Bickel (1970)]

$$\Pr\{p \leq \alpha \mid H_0\} \leq \alpha. \tag{5.1}$$

The left-hand side above may depend on unknown nuisance parameters, in which case the replication underlying (5.1) is only conceptual in the sense that it cannot be simulated in reality.

Such a frequency evaluation is not required for Bayesian analyses, especially when an analysis is carried out for a particular data set. As pointed out by Rubin (1984), however, such "frequency calculations that investigate the operating

characteristics of Bayesian procedures are (Bayesianly) relevant and justifiable when investigating or recommending procedures for general consumption." In fact, in the current context, frequency calculations are more straightforward and flexible with the Bayesian formulation because the Bayesian formulation provides a natural way to quantify different replications under which the "Type I" error is measured.

For example, if one is interested in the error rate under the replication with the free parameters under the null hypothesis having the same (unknown) value as in the actual study, one then can use the posterior predictive distribution, given in (2.3), as the replication. Of course, such a replication may be viewed as too restrictive since data sets from different studies may well be generated from different values of the free parameters even if they share the same null hypothesis. It is, therefore, practically more relevant to measure the Type I error rate of p_B under the *prior predictive distribution conditional on* H_0:

$$f(X \mid H_0) = \int_{\Psi_0} f(X \mid \psi)\Pi_0(d\psi), \tag{5.2}$$

which allows different values of ψ being drawn from the conditional (on H_0) prior distribution $\Pi_0(\psi)$ with each replication. Notice that, in order to do so, the conditional prior distribution needs to be proper [i.e., $\int_{\Psi_0} \Pi_0(d\psi) = 1$], as in Box (1980). Other replications (e.g., fixing some of the free parameters) can also be of practical interest; see Rubin (1984) and Gelman, Meng and Stern (1993).

Notice that (5.1) is a comparison of the sampling distribution of a classical p-value with a uniform distribution. The following result establishes a general relationship between the prior predictive distribution of a posterior predictive p-value and a uniform distribution on $[0, 1]$.

THEOREM 1. *Suppose, given* $\psi \in \Psi_0$, *the sampling distribution of a discrepancy measure* $D(X, \psi)$ *is continuous. Then under the prior predictive distribution* (5.2), *the corresponding posterior predictive p-value,* p_B *of* (2.8), *is stochastically less variable than a uniform distribution but with the same mean; that is, if* U *is uniformly distributed on* $[0, 1]$, *then* (i) $E(p_B) = E(U) = \frac{1}{2}$ *and* (ii) $E[h(p_B)] \leq E[h(U)]$ *for all convex functions* h *on* $[0, 1]$, *where the expectations involving* p_B *are with respect to* (5.2).

PROOF. By the two-stage derivation of Section 2.2, p_B can be expressed as

$$p_B = E\Big[p(x \mid \psi) \mid x, H_0\Big], \tag{5.3}$$

where

$$p(x \mid \psi) = \Pr\{D(X, \psi) \geq D(x, \psi) \mid \psi\}, \qquad \psi \in \Psi_0,$$

is the conditional p-value, which contains (2.9) as a special case with Ψ_0 being a point hypothesis for θ. Under our assumption, given $\psi \in \Psi_0$, the distribution

of $p(X \mid \psi)$ is uniform on $[0, 1]$. Thus, for any convex functon h, we have, by (5.3),

$$\begin{aligned} E[h(p_B)] &= E\left\{h\left(E\left[p(X \mid \psi) \mid X, H_0\right]\right)\right\} \\ &\le E\left\{E\left[h\left(p(X \mid \psi)\right) \mid X, H_0\right]\right\} \quad \text{(by Jensen's inequality)} \\ &= E\left\{E\left[h\left(p(X \mid \psi)\right) \mid \psi \in \Psi_0\right]\right\} = E[h(U)]. \end{aligned}$$

Hence, (ii) holds. Identity (i) follows by taking $h(t) = t$ in the above derivation (obviously the equality holds in this case). In fact, (i) is a consequence of (ii) by taking $h(t) = t$ and $h(t) = -t$. □

The above result indicates that, under the prior predictive distribution (5.2), p_B is more closely centered around $\frac{1}{2}$ than a uniform variable since it gives less weight to the extreme values [obviously $\mathrm{Var}(p_B) \le \frac{1}{12}$ by taking $h(t) = t^2$ in (ii)]. Intuitively, this suggests that there exists an α_0 small enough such that

$$\Pr\{p_B \le \alpha\} \le \alpha \quad \text{for all } \alpha \in [0, \alpha_0], \tag{5.4}$$

where the probability is taken over (5.2). Of course, the value of α_0 will depend on the underlying model so there is no general lower bound on α_0. However, the left side of (5.4) cannot be too big compared to α because of (i) and (ii) of Theorem 1. Indeed, we have the following result, which is only slightly coarser than (5.4), but holds in general. Since this result itself may be of independent interest in other contexts, we list it as a lemma, an elementary proof of which is given in the Appendix.

LEMMA 1. *Let $G(\alpha)$ be the c.d.f. of a random variable W on $[0, 1]$. If W is stochastically less variable than $U[0, 1]$ but with the same mean, then $\forall \alpha \in [0, 1]$*

$$\alpha - \left[\alpha^2 - 2\int_0^\alpha G(t)\,dt\right]^{1/2} \le G(\alpha) \le \alpha + \left[\alpha^2 - 2\int_0^\alpha G(t)\,dt\right]^{1/2} \le 1. \tag{5.5}$$

The first or second inequality becomes an equality for all α if and only if $G(\alpha) \equiv \alpha$.

One direct consequence of (5.5) is that

$$G(\alpha) \le 2\alpha \quad \text{for all } \alpha \le \tfrac{1}{2}. \tag{5.6}$$

Applying (5.6) to p_B, it implies that, under the prior predictive distribution, *the Type I error rate of p_B will never exceed twice the nominal level* (e.g., with $\alpha = 0.05, \Pr\{p_B \le \alpha\} \le 0.1$), and this is true for any nominal level. The bound 2α in (5.6) is achievable in the following pathological example at α_0 for a $G_0(\alpha)$

satisfying the conditions of the lemma:

$$G_0(\alpha) = \begin{cases} 0, & 0 \le \alpha < \alpha_0, \\ 2\alpha_0, & \alpha_0 \le \alpha < 2\alpha_0, \\ \alpha, & 2\alpha_0 \le \alpha \le 1, \end{cases}$$

but inequality (5.5) suggests that in most practical situations the factor 2 is simply too large and $G(\alpha)$ should be quite close to α. For example, if $G(\alpha) \ge c\alpha$ on $\alpha \in [0, \alpha^*]$ for some $c < 1$, then (5.5) implies that

$$G(\alpha) \le \left[1 + (1-c)^{1/2}\right]\alpha \quad \text{for } \alpha \in [0, \alpha^*]. \tag{5.7}$$

Inequality (5.7) is interesting because it says that if the error rate of p_B is not *too low* compared to the nominal level, then it cannot be *too high* either. For example, if $\Pr\{p_B \le \alpha\} \ge 0.75\alpha$ for $\alpha \le 0.05$, then (5.7) implies that

$$0.75\alpha \le \Pr\{p_B \le \alpha\} \le 1.5\alpha \quad \text{for } \alpha \le 0.05.$$

Sharper bounds, such as $\Pr\{p_B \le \alpha\} \le \alpha$ when $\alpha < 0.5$, are conjectured for certain (common) models. Also conjectured are bounds on $\Pr\{p_B \le \alpha\}$ when the prior density in (5.2) differs from that of (2.8), a replication that is more relevant in practice.

APPENDIX

PROOF OF LEMMA 1. By Proposition 8.5.1 of Ross [(1983), page 270], W being stochastically less variable than $U[0,1]$ is equivalent to

$$\int_\alpha^1 \left[1 - G(t)\right]dt \le \tfrac{1}{2}(1-\alpha)^2 \quad \text{for all } \alpha \in [0,1]. \tag{A.1}$$

Since $E(W) = 1/2$, inequality (A.1) is equivalent to

$$\int_0^\alpha G(t)\,dt \le \frac{\alpha^2}{2} \quad \text{for all } \alpha \in [0,1]. \tag{A.2}$$

Given $\alpha \in [0,1]$, for any $s \in [0,1]$, we have from (A.2) that

$$\frac{s^2}{2} \ge \int_0^s G(t)\,dt = \int_0^\alpha G(t)\,dt + \int_\alpha^s G(t)\,dt \ge \int_0^\alpha G(t)\,dt + G(\alpha)(s-\alpha),$$

since $G(t)$ is nondecreasing. It follows then that, if $s \in (\alpha, 1]$,

$$G(\alpha) \le \frac{s^2/2 - \int_0^\alpha G(t)\,dt}{s-\alpha} \equiv b_\alpha(s). \tag{A.3}$$

It is easy to verify that $b_\alpha(s)$ attains its minimum at $s_0 \in (\alpha, 1]$ that satisfies $b_\alpha(s_0) = s_0$, solving which yields

$$G(\alpha) \le s_0 = \alpha + \left[\alpha^2 - 2\int_0^\alpha G(t)\,dt\right]^{1/2} \le 1.$$

The first inequality in (5.5) is established by considering $s \in [0, \alpha)$, in which case "$\leq$" in (A.3) becomes "$\geq$" and $b_\alpha(s_0) = s_0$ is solved on $s_0 \in [0, \alpha)$. The rest of the lemma is easy to verify. □

Acknowledgments. The author is grateful to R. R. Bahadur, A. Gelman and D. B. Rubin for inspiring conversations and constructive suggestions, and to referees and, especially, an Associate Editor for a set of exceptionally stimulating comments on an early version of this paper. Thanks also go to T. Belin, S. Pedlow, H. Stern, G. C. Tiao, D. L. Wallace, W. H. Wong, A. M. Zaslavsky and A. Zellner for helpful exchanges.

REFERENCES

BAHADUR, R. R. (1967a). An optimal property of the likelihood ratio statistic. *Proc. Fifth Berkeley Symp. Math. Statist. Probab.* **1** 13–26. Univ. California Press, Berkeley.

BAHADUR, R. R. (1967b). Rates of convergence of estimates an test statistics. *Ann. Math. Statist.* **38** 303–324.

BAHADUR, R. R. and BICKEL, P. J. (1970). On conditional levels in large samples. In *Essays in Probability and Statistics* (R. C. Bose, I. M. Chakravarti, P. C. Mahalanobis, C. R. Rao, K. J. C. Smith, eds.) **3** 25–34. Univ. North Carolina Press, Chapel Hill.

BAHADUR, R. R., CHANDRA, T. K. and LAMBERT, D. (1984). Some further properties of likelihood ratios on general sample spaces. In *Proceedings of the Indian Statistical Institute Golden Jubilee International Conference on Statistics: Applications and New Directions* (J. K. Ghosh and J. Roy, eds.) 1–19. Indian Statist. Inst., Calcutta.

BAHADUR, R. R. and RAGHAVACHARI, M. (1972). Some asymptotic properties of likelihood ratios on general sample spaces. *Proc. Sixth Berkeley Symp. Math. Statist. Probab.* **1** 129–152. Univ. California Press, Berkeley.

BERGER, J. and DELAMPADY, M. (1987). Testing precise hypotheses (with discussion). *Statist. Sci.* **2** 317–352.

BERGER, J. O. and SELLKE, T. (1987). Testing a point null hypothesis: the irreconcilability of p values and evidence (with discussion). *J. Amer. Statist. Assoc.* **82** 112–122.

BERGER, J. O. and WOLPERT, R. L. (1984). *The Likelihood Principle*. IMS, Hayward, CA.

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

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

BROEMELING, L. and ABDULLAH, M. Y. (1984). An approximation to the poly-t distributions. *Comm. Statist. Theory and Methods* **13** 1407–1422.

CASELLA, G. and BERGER, R. L. (1987). Reconciling Bayesian and frequentist evidence in the one-sided testing problems (with discussion). *J. Amer. Statist. Assoc.* **82** 106–111.

COX, D. R. (1975). A note on partially Bayes inference and the linear model. *Biometrika* **62** 651–654.

COX, D. R. (1977). The role of significance tests (with discussion). *Scand. J. Statist.* **4** 49–70.

DEMPSTER, A. P. (1971). Model searching and estimation in the logic of inference (with discussion). In *Foundations of Statistical Inference* (V. P. Godambe and D. A. Sprott, eds.) 56–81. Holt, Rinehart and Winston, Toronto.

DEMPSTER, A. P. (1973). The direct use of likelihood for significance testing (with discussion). In *Proceedings of Conference on Foundational Questions in Statistical Inference* (O. Barndorff-Nielsen, P. Blaeslid and G. Schou, eds.) 335–354. Dept. Theoretical Statistics, Univ. Aarhus, Denmark.

GELMAN, A., MENG, X. L. and STERN, H. (1993). Bayesian model checking using tail area probabilities. Technical report 355, Dept. Statistics, Univ. Chicago. To appear in *Statist. Sinica*.

GOOD, I. J. (1992). The Bayes/non-Bayes compromise: a brief review. *J. Amer. Statist. Assoc.* **87** 597–606.

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

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

JOHNSON, N. L. and KOTZ, S. (1970). *Continuous Univariate Distributions—I.* Wiley, New York.

LI, K. H., MENG, X. L., RAGHUNATHAN, T. E. and RUBIN, D. B. (1991). Significance levels from repeated p-values with multiply-imputed data. *Statist. Sinica* **1** 65–92.

LI, K. H., RAGHUNATHAN, T. E. and RUBIN, D. B. (1991). Large sample significance levels from multiply-imputed data using moment-based statistics and an F reference distribution. *J. Amer. Statist. Assoc.* **86** 1065–1073.

MCCULLAGH, P. (1990). A note on partially Bayes inference for generalized linear models. Technical Report 284, Dept. Statistics, Univ. Chicago.

MENG, X. L. (1988). Significance levels from repeated significance levels in multiple imputation. Ph.D. qualifying paper, Dept. Statistics, Harvard Univ.

MENG, X. L. (1990). Towards complete results for some incomplete-data problems. Ph.D. dissertation, Dept. Statistics, Harvard Univ. (Printed by U.M.I., Ann Arbor, MI.)

MENG, X. L. (1994). Multiple-imputation inference with uncongenial sources of input (with discussion). *Statist. Sci.* **9** (4).

MENG, X. L. and RUBIN, D. B. (1992). Performing likelihood ratio tests with multiply-imputed data sets. *Biometrika* **79** 103–111.

RAGHUNATHAN, T. E. (1987). Large sample significance levels from multiply-imputed data. Ph.D. dissertation, Dept. Statistics, Harvard Univ.

ROSS, S. M (1983). *Stochastic Processes.* Wiley, New York.

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

RUBIN, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys.* Wiley, New York.

RUBIN, D. B. (1995). Multiple imputation after 18 years. *J. Amer. Statist. Assoc.* To appear.

SHAFER, G. (1982). Lindley's paradox (with discussion). *J. Amer. Statist. Assoc.* **77** 325–351.

TSUI, K.-W. and WEERAHANDI, S. (1989). Generalized p-values in significance testing of hypotheses in the presence of nuisance parameters. *J. Amer. Statist. Assoc.* **84** 602–607.

WALLACE, D. L. (1980). The Behrens–Fisher and Feiller–Creasy problems. In *R. A. Fisher: An Appreciation* (S. F. Fienberg and D. V. Hinkley, eds.) 119–147. Springer, New York.

DEPARTMENT OF STATISTICS
UNIVERSITY OF CHICAGO
5734 UNIVERSITY AVENUE
CHICAGO, ILLINOIS 60637
MENG@GALTON.UCHICAGO.EDU