P Values for Composite Null Models

Author(s): M. J. Bayarri and James O. Berger

Source: *Journal of the American Statistical Association*, Dec., 2000, Vol. 95, No. 452 (Dec., 2000), pp. 1127-1142

Published by: Taylor & Francis, Ltd. on behalf of the American Statistical Association

Stable URL: https://www.jstor.org/stable/2669749

P Values for Composite Null Models

M. J. BAYARRI and James O. BERGER

The problem of investigating compatibility of an assumed model with the data is investigated in the situation when the assumed model has unknown parameters. The most frequently used measures of compatibility are p values, based on statistics T for which large values are deemed to indicate incompatibility of the data and the model. When the null model has unknown parameters, p values are not uniquely defined. The proposals for computing a p value in such a situation include the plug-in and similar p values on the frequentist side, and the predictive and posterior predictive p values on the Bayesian side. We propose two alternatives, the conditional predictive p value and the partial posterior predictive p value, and indicate their advantages from both Bayesian and frequentist perspectives.

KEY WORDS: Bayes factors; Bayesian p values; Conditioning; Model checking; Predictive distributions.

1. INTRODUCTION

1.1 Background

In parametric statistical analysis of data $\mathbf{X}$, one is frequently working at a given moment with an entertained model or hypothesis $H_0 : \mathbf{X} \sim f(\mathbf{x}; \theta)$. We will call this the null model or null hypothesis, even though no alternative is explicitly formulated. We assume that $f(\mathbf{x}; \theta)$ is either a discrete density or a continuous density (with respect to Lebesgue measure). A statistic $T = t(\mathbf{X})$ is chosen to investigate compatibility of the model with the observed data, $\mathbf{x}_{\text{obs}}$. We assume that T has been expressed in such a way that large values of T indicate less compatibility with the model. The most commonly used measure of compatibility is the p value, defined as

$$p = \Pr(t(\mathbf{X}) \geq t(\mathbf{x}_{\text{obs}})). \tag{1}$$

When θ is known, the probability computation in (1) is with respect to $f(\mathbf{x}; \theta)$. The focus in this article is on the choice of the probability distribution used to compute (1) when θ is unknown. In Section 2 we present two new types of p values, which we argue are superior to existing choices. The rest of this section describes the most common of the existing choices. We abuse notation by using $f(t; \theta)$ and $f(t|u; \theta)$ to denote the marginal density of $t(\mathbf{X})$ and the conditional density of $t(\mathbf{X})$ given $u(\mathbf{X}) = u$.

The most obvious way to deal with an unknown θ in computation of the p value is to replace θ in (1) by some estimate, $\hat{\theta}$. In this article we consider only the usual choice for $\hat{\theta}$, namely the maximum likelihood estimator (MLE). We call the resulting p value the *plug-in* p value (p_{plug}). Using a superscript to denote the density with respect to which the p value in (1) is computed, the plug-in p value is thus defined as

$$p_{\text{plug}} = \Pr^{f(\cdot;\hat{\theta})}(t(\mathbf{X}) \geq t(\mathbf{x}_{\text{obs}})). \tag{2}$$

The main strengths of p_{plug} are its simplicity and intuitive appeal. Its main weakness appears to be a failure to account for uncertainty in the estimation of θ, although as we show, this issue is rather involved.

Another natural device for eliminating the unknown θ is to condition on a sufficient statistic, U, for θ. Then $f(\mathbf{x}|u_{\text{obs}}; \theta)$ does not depend on θ, and computations in (1) can be carried out using the completely specified $f(\mathbf{x}|u_{\text{obs}})$. [In fact, U need only be sufficient for θ with respect to $f(t; \theta)$.] We call these p values *similar p values,* a term borrowed from the related notion of similar tests and confidence regions. A similar p value is thus defined as

$$p_{\text{sim}} = \Pr^{f(\cdot|u_{\text{obs}})}(t(\mathbf{X}) \geq t(\mathbf{x}_{\text{obs}})). \tag{3}$$

The main strength of p_{sim} is that it is based on a proper probability computation, which imbues the end result with various desirable properties (discussed later). Its main weaknesses are that the computation can be burdensome and that a suitable sufficient U typically does not exist.

Bayesians have a natural way to eliminate nuisance parameters: integrate them out. Thus if $\pi(\theta)$ is a prior distribution for θ, then the marginal or (prior) *predictive distribution* is

$$m(\mathbf{x}) = \int f(\mathbf{x}; \theta)\pi(\theta)\, d\theta. \tag{4}$$

Because this is free of θ, it can be used to compute a p value, leading to the prior predictive p value, given by

$$p_{\text{prior}} = \Pr^{m(\cdot)}(t(\mathbf{X}) \geq t(\mathbf{x}_{\text{obs}})). \tag{5}$$

The main strengths of p_{prior} are that it is also based on a proper probability computation (at least if $\pi(\theta)$ is proper), and that it suggests a natural and simple T, namely $t(\mathbf{x}) = 1/m(\mathbf{x})$. The main weakness of p_{prior} for pure model checking is its dependence on the prior $\pi(\theta)$; in essence, $m(\mathbf{x})$ measures the likelihood of $\mathbf{x}$ relative to both the model and

M. J. Bayarri is Professor of Statistics and O.R., Department of Statistics and O.R., University of Valencia, 46100 Burjassot, Valencia, Spain (E-mail: *susie.bayarri@uv.es*). James O. Berger is Arts and Sciences Professor of Statistics, Institute of Statistics and Decision Sciences, Duke University, Durham, NC 27708 (E-mail: *berger@stat.duke.edu*). This work was supported in part by National Science Foundation grants DMS-9303556 and DMS-9802261, and by Ministry of Education and Culture and the Generalitat Valenciana (Spain) grants PB96-0776 and POST99-01-7. The authors thank George Casella and Martin Tanner for organizing these contributions to JASA, and the associate editor and three referees for numerous helpful comments that greatly improved the article.

the prior, and an excellent model could come under suspicion if a poor prior distribution were used. For this reason, and because model checking is often considered at early stages of an analysis before careful prior elicitation is performed (and/or because a nonsubjective analysis might be desired from the beginning), it is attractive to attempt to use noninformative priors. Unfortunately, noninformative priors are typically improper, in which case the prior predictive $m(\mathbf{x})$ would also be improper, precluding computation of (5). Box (1980) popularized the use of p_{prior}.

The concerns mentioned in the preceding paragraph have led many Bayesians, beginning with Guttman (1967) and Rubin (1984), to eliminate θ from $f(\mathbf{x};\theta)$ by integrating with respect to the posterior distribution, $\pi(\theta|\mathbf{x}_{\text{obs}})$, instead of the prior, before computing a p value. The posterior predictive p value is thus defined as

$$p_{\text{post}} = \Pr^{m_{\text{post}}(\cdot|\mathbf{x}_{\text{obs}})}(t(\mathbf{X}) \geq t(\mathbf{x}_{\text{obs}})), \tag{6}$$

where

$$m_{\text{post}}(\mathbf{x}|\mathbf{x}_{\text{obs}}) = \int f(\mathbf{x};\theta)\pi(\theta|\mathbf{x}_{\text{obs}})\, d\theta. \tag{7}$$

The main strengths of p_{post} are as follows:

a. Improper noninformative priors can readily be used (since $\pi(\theta|\mathbf{x}_{\text{obs}})$ will typically be proper).
b. $m_{\text{post}}(\mathbf{x}|\mathbf{x}_{\text{obs}})$ typically will be much more heavily influenced by the model than by the prior; indeed, as the sample size goes to infinity, the posterior distribution will essentially concentrate at $\hat{\theta}$, so that p_{post} will (for large n) be very close to p_{plug}.
c. It typically is very easy to compute using output from modern Markov chain Monte Carlo (MCMC) Bayesian analyses.

Its main weakness is that there is an apparent "double use" of the data in (6), first to convert the (possibly improper) prior $\pi(\theta)$ into a proper distribution $\pi(\theta|\mathbf{x}_{\text{obs}})$ for determining the reference distribution $m_{\text{post}}(\mathbf{x}|\mathbf{x}_{\text{obs}})$, and then to compute the tail area corresponding to $t(\mathbf{x}_{\text{obs}})$. This double use of the data can induce unnatural behavior. From a Bayesian perspective, defenders of the prior predictive also point out that the posterior predictive lacks a pure Bayesian interpretation; although this was our original motivation for the developments herein, the arguments in the article are not directly based on such reasoning.

Generalizations of (6) were considered by Meng (1994), Gelman, Carlin, Stern, and Rubin (1995), Gelman, Meng, and Stern (1996), and references therein; in particular, $t(\mathbf{X})$ could be replaced by a function $t(\mathbf{X},\theta)$, and $f(\mathbf{x};\theta)$ in (7) could be replaced by $f(\mathbf{x}|\theta,A)$, where A is some other statistic. We do not discuss such generalizations in this article.

There are also many other related works. Aitkin (1991) used the posterior distribution to compute actual Bayes factors, instead of p values. Evans (1997) introduced a related concept for model checking based on the ratio of the posterior and prior predictive densities.

Other approaches that have been suggested for dealing with the nuisance parameter, θ, in computing (1) include those of Tsui and Weerahandi (1989) (primarily for one-sided testing) and Berger and Boos (1994). The latter authors sought to provide a practical implementation of the conservative frequentist approach that deals with unknown θ by maximization:

$$p_{\text{sup}} = \sup_{\theta} \Pr^{f(\cdot;\theta)}(t(\mathbf{X}) \geq t(\mathbf{x}_{\text{obs}})).$$

This p value is of rather limited usefulness, because the supremum is often too large to provide useful criticism of the model. For instance, in the examples of Sections 2 and 3, p_{sup} can easily be seen to equal 1. Berger and Boos (1994) overcame this difficulty by restricting the supremum to θ in a confidence set for θ (with the noncoverage probability being added to the p value). Although potentially useful to frequentists in formal testing situations, in which conservatism is typically deemed desirable, the approach is less appropriate for model checking in which conservatism would mean that one is often not alerted to the fact that the model is inadequate.

1.2 Evaluation of p Values

What do we want in a p value? For a frequentist, one appealing property would be for p, considered as a random variable, to be uniform[0, 1] under the null, $f(\mathbf{x};\theta)$, for all θ. In some sense, being $U[0,1]$ defines a proper p value, allowing for its common interpretation across problems. Statistical measures that lack a common interpretation across problems are simply not very useful. (For more extensive discussion of this point, see the companion article Robins, van der Vaart, and Ventura 2000, which we henceforth denote by RVV 2000; earlier articles that refer to and/or discuss this "defining" property of a p value include De la Horra and Rodríguez-Bernal 1997, Meng 1994, Robins 1999, Rubin 1996, and Thompson 1997.)

For most problems, exact uniformity under the null for all θ cannot be attained for any p value. Thus one must weaken the requirement to some extent. A natural weaker requirement is that a p value be $U[0,1]$ under the null in an asymptotic sense; this is the subject of RVV (2000). Here we focus on studying the degree to which the various p values deviate from uniformity in finite-sample scenarios.

It is not obvious that Bayesians should be concerned with establishing that a p value is uniform under the null for all θ. For instance, the prior predictive p value is $U[0,1]$ under $m(\mathbf{x})$ (if the prior is proper), which means that it is $U[0,1]$ in an average sense over θ. If the prior distribution is chosen subjectively, then a Bayesian could well argue that this is sufficient; indeed, Meng (1994) suggested that uniformity under $m(\mathbf{x})$ is a useful criterion for the evaluation of any proposed p value. (The more basic issue that a p value is a tail area, and not compatible with true Bayesian measures, is discussed briefly in the next section.)

As mentioned earlier, however, preliminary model checking is most typically done (by Bayesians) with noninformative priors, and if these are improper, there is no "average over θ" that can be used. (We later give an example with

a proper noninformative prior, in which a Bayesian—or non-Bayesian—might settle for "average" uniformity.) Of course, if a p value is uniform under the null in the frequentist sense, then it has the strong Bayesian property of being marginally $U[0, 1]$ under *any* proper prior distribution. This explains why Bayesians should, at least, be highly satisfied if the frequentist requirement obtains. Perhaps more to the point, if a proposed p value is *always* either conservative or anticonservative in a frequentist sense (see RVV 2000 for definitions), then it is likewise guaranteed to be conservative or anticonservative in a Bayesian sense, no matter what the prior. A similar conclusion would hold for large sample sizes (under mild conditions) if a proposed p value were always conservative or anticonservative in a frequentist asymptotic sense. (Interesting related discussions concerning the posterior predictive p value have been given by Gelman et al. 1996, Meng 1994, and Rubin 1996.)

Actually, Bayesians might well go further, not only requiring unconditional uniformity for p values, but also seeking reasonable conditional performance. In this article we limit discussion of this issue to presentation of some examples in which it is clear that study of conditional performance is of value in comparing p values; we do not, however, attempt to present general results in this direction.

There is a vast literature on other methods of evaluating p values. Much of the literature is concerned with power comparisons against alternatives. There is also a significant literature concerned with decision-theoretic evaluations of p values (e.g., Blyth and Staudte 1995; Hwang, Casella, Robert, Wells, and Farrell 1992; Hwang and Pemantle 1997; Hwang and Yang 1997; Schaafsma, Tolboom, and Van Der Meulen 1989; Thompson 1997). Neither of these evaluation techniques is within the scope of this article, because we are specifically concerned with the situation in which no alternative is present (see the next section). From a non-Bayesian perspective, however, evaluation of the new p values by these criteria might well prove very illuminating; see RVV (2000) for interesting results in this direction.

1.3 To Be and Not To Be

This article has five sections. In Section 2 we consider two new p values introduced by Bayarri and Berger (1999), the *partial posterior predictive p value* (p_{ppost}) and the *conditional predictive p value* (p_{cpred}), and compare them with previous p values in specific examples. Some results are also given in Section 2 concerning equality of various p values; of particular interest is a result (Theorem 2) that allows ready computation of p_{sim} in certain situations. In Section 3 we compare the various p values in the normal linear model, where exact computation is possible; this section was directly motivated by RVV (2000). In Section 4 we discuss the situation of discrete sample spaces, with emphasis on analysis of contingency tables; this has long been a highly problematic area, with the discreteness of the sample space causing many p values to be very conservative. We present conclusions in Section 5.

A number of relevant issues are not considered in this article. First, we do not explicitly discuss the choice of T, in part because this is a contentious issue and is not directly related to our development; one of the strengths of the methodology that we propose is that it can be applied to essentially any choice of T. Second, our primary focus is on model checking at initial, exploratory stages of the statistical analysis, and consideration of a wide variety of intuitive T is often useful at that stage. If one has a clearly formulated alternative to H_0, then we would not recommend using *any* p value to perform the test, and would instead use either Bayes factors or conditional frequentist tests (Bayarri and Berger 1999; Berger, Boukai, and Wang 1997; Berger, Brown, and Wolpert 1994; Berger and Delampady 1987; Berger and Sellke 1987; Delampady and Berger 1990; Edwards, Lindman, and Savage 1963). The decision to even formulate an alternative to H_0, however, is often undertaken on the basis of an analysis designed to indicate incompatibility of the model with the data, based on intuitive departure statistics $T = t(\mathbf{x})$. If determination of T required hard work, then we would suggest spending the time instead on actual formulation of the alternative.

The other major issue that we mostly avoid is discussion of measures other than p values of data compatibility with the model. (For a review of a number of other measures that have been proposed, see Bayarri and Berger 1997.) The primary reason for considering only p values is their ubiquitous presence in statistics, together with the fact that they do have some desirable properties (such as invariance to transformations of $\mathbf{X}$). Balanced against this is the near ubiquitous misinterpretation of p values as either frequentist error probabilities or (worse) as the probability of H_0. Luckily, a rather simple calibration is available that allows p values to be given an intuitive interpretation: compute $B(p) = -ep\log(p)$, when $p < e^{-1}$, and interpret this as the odds (or *Bayes factor*) of H_0 to H_1, where H_1 denotes the (unspecified) alternative to H_0. For those who prefer to think in terms of a frequentist error probability α (in rejecting H_0), the calibration is $\alpha(p) = (1 + [-ep\log(p)]^{-1})^{-1}$. As an example, $p = .05$ translates into odds $B(.05) = .41$ (roughly 1–2.5) of H_0 to H_1, and frequentist error probability $\alpha(.05) = .29$ in rejecting H_0.

These calibrations were developed and motivated from a various perspectives by Sellke, Bayarri, and Berger (1999). On the Bayesian side, they arise from robust Bayesian arguments, as lower bounds on Bayes factors for testing H_0. [$B(p)$ arises exactly in testing against a general nonparametric alternative, and arises approximately in parametric analyses.)] On the frequentist side, $\alpha(p)$ arises as a lower bound on the type I error probability, over a large class of conditional frequentist tests, where one conditions on the "strength of evidence" in the data. It is of interest that the calibrations are based in part on starting with a proper p value; that is, a p value that is $U[0, 1]$ in some sense.

Further discussion of some of the philosophical issues surrounding the use of p values has been given by Bayarri and Berger (1999). From now on, we ignore these issues and simply assume that (possibly calibrated) p values are useful,

for whatever reasons, and focus on the issue of which p values are most satisfactory.

2. CONDITIONAL PREDICTIVE p VALUES

Section 2.1 introduces the two new p values that we consider and illustrates their definitions (and those of the other p values) in a standard example; some interesting features of the various p values are also observed. Section 2.2 presents the motivations for the new p values from both Bayesian and frequentist perspectives. Section 2.3 addresses computational issues.

2.1 Methodology

Consider first the partial posterior predictive p value, defined for a prior $\pi(\theta)$ (typically noninformative) as

$$p_{\text{ppost}} = \Pr^{m(\cdot|\mathbf{x}_{\text{obs}} \setminus t_{\text{obs}})}(T \geq t_{\text{obs}}); \tag{8}$$

here $T = t(\mathbf{X}), t_{\text{obs}} = t(\mathbf{x}_{\text{obs}})$, and $m(\cdot|\mathbf{x}_{\text{obs}} \setminus t_{\text{obs}})$ and the (assumed proper) partial posterior $\pi(\cdot|\mathbf{x}_{\text{obs}} \setminus t_{\text{obs}})$ are given by

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

and

$$\pi(\theta|\mathbf{x}_{\text{obs}} \setminus t_{\text{obs}}) \propto f(\mathbf{x}_{\text{obs}}|t_{\text{obs}};\theta)\pi(\theta) \propto \frac{f(\mathbf{x}_{\text{obs}};\theta)\pi(\theta)}{f(t_{\text{obs}};\theta)}. \tag{9}$$

Intuitively, this avoids the double use of the data that occurs in the posterior predictive p value, because the contribution of t_{obs} to the posterior is "removed" before θ is eliminated by integration. (The notation $\mathbf{x}_{\text{obs}} \setminus t_{\text{obs}}$ was chosen to indicate this.)

The second p value that we propose is a specific case of what can be termed a *U-conditional predictive p value,* defined, for some conditioning statistic $U = u(\mathbf{X})$, as

$$p_{\text{cpred}(u)} = \Pr^{m(\cdot|u_{\text{obs}})}(T \geq t_{\text{obs}}); \tag{10}$$

here $u_{\text{obs}} = u(\mathbf{x}_{\text{obs}})$ and (formally)

$$m(t|u) = \int f(t|u;\theta)\pi(\theta|u)\, d\theta, \tag{11}$$

assuming that

$$\pi(\theta|u) = \frac{f(u;\theta)\pi(\theta)}{\int f(u;\theta)\pi(\theta)\, d\theta} \tag{12}$$

is proper. [Recall that $f(t|u;\theta)$ and $f(u;\theta)$ are defined as the conditional and marginal densities of T and U under H_0.]

The specific proposal that we recommend, for the case of continuous data, is obtained by choosing U in (10) to be the conditional MLE of θ, given $t(\mathbf{x}) = t$, defined as

$$\hat{\theta}_{\text{cMLE}}(\mathbf{x}) = \arg\max f(\mathbf{x}|t,\theta) = \arg\max \frac{f(\mathbf{x};\theta)}{f(t;\theta)}. \tag{13}$$

We suppress $\hat{\theta}_{\text{cMLE}}$ and call the resulting p value simply the *conditional predictive p value,* denoted by $p_{\text{cpred}} = p_{\text{cpred}(\hat{\theta}_{\text{cMLE}})}$. Note that $m(t|u)$ is unaffected by one-to-one transformations of $u(\mathbf{x})$, so that any one-to-one transformation of (13) is satisfactory as the choice of $\hat{\theta}_{\text{cMLE}}$.

Note that when T is conditionally independent of $\hat{\theta}_{\text{cMLE}}$ and $(T, \hat{\theta}_{\text{cMLE}})$ are jointly sufficient, both of the foregoing proposals for p values agree; that is, $p_{\text{ppost}} = p_{\text{cpred}}$. This occurs in the following example from Meng (1994), which we use to exhibit the various p values that have been defined so far.

Example 1. Assume that under the null, the X_i are iid $\text{N}(0, \sigma^2)$, with σ^2 unknown. The statistic $t(\mathbf{X}) = |\bar{X}|$ is chosen to measure departure from the model (which would be natural for detecting a discrepancy in the mean of the model). The various p values are given by

$$p = \Pr\{|\bar{X}| > |\bar{x}_{\text{obs}}|\}, \tag{14}$$

with different distributions used to compute the probability. For the Bayesian p values, we utilize the usual noninformative prior for σ^2 : $\pi(\sigma^2) \propto 1/\sigma^2$. Finally, define $s^2 = \sum(x_i - \bar{x})^2/n$.

p_{plug}: Because $\bar{X} \sim \text{N}(0, \sigma^2/n)$ and $\hat{\sigma}^2 = 1/n\sum_{i=1}^n x_i^2 = s^2 + \bar{x}^2$ is the MLE, it follows from (2) and (14) that

$$p_{\text{plug}} = 2\left[1 - \Phi\left(\frac{\sqrt{n}|\bar{x}_{\text{obs}}|}{\sqrt{s_{\text{obs}}^2 + \bar{x}_{\text{obs}}^2}}\right)\right]. \tag{15}$$

One obvious inadequacy with this p value is that $p_{\text{plug}} \to 2[1 - \Phi(\sqrt{n})]$, a positive constant, as $|\bar{x}_{\text{obs}}|/s_{\text{obs}} \to \infty$. Thus, even with arbitrarily strong evidence against the null model, the p value will not go to 0 (for fixed n). For large n, this limiting constant will of course be small, so that it would not pose a practical problem. In practice, however, the number of observations is often not large in comparison to the number of parameters, so that concerns of this type can be relevant. In any case, such behavior is indicative of a fundamental flaw in the procedure. (In this example, one could achieve results that are more satisfactory by plugging in s_{obs} rather than the MLE; indeed, this is related to the "conditional plug-in" p value, which, however, is shown in RVV 2000 to also have deficiencies.)

p_{sim}: A sufficient statistic for σ^2 is $V = \sum_{i=1}^n X_i^2 = \|\mathbf{X}\|^2$. The distribution of $\mathbf{X}$, given $v_{\text{obs}} = \|\mathbf{x}_{\text{obs}}\|^2$, is uniform on $\{\mathbf{x} : \|\mathbf{x}\|^2 = \|\mathbf{x}_{\text{obs}}\|^2\}$, so that (3) and (14) yield

$$\begin{aligned} p_{\text{sim}} &= \Pr\left(\frac{|\bar{X}|}{\|\mathbf{x}_{\text{obs}}\|} > \frac{|\bar{x}_{\text{obs}}|}{\|\mathbf{x}_{\text{obs}}\|}\right) \\ &= \Pr\left(|\bar{Z}| > \frac{|\bar{x}_{\text{obs}}|}{\|\mathbf{x}_{\text{obs}}\|}\right), \end{aligned} \tag{16}$$

where $\mathbf{Z}$ has a uniform distribution on $\{\mathbf{z} : \|\mathbf{z}\|^2 = 1\}$. Although this might appear to be difficult to compute, it is shown later (using Theorem 2) that p_{sim} is exactly equal to p_{ppost} and p_{cpred}, which in turn are equal to the classical p value for the problem given in (18); we found this result surprising.

p_{prior}: The prior predictive p value cannot be computed for this example, because the prior distribution is improper.

p_{post}: The posterior density, $\pi(\sigma^2|\mathbf{x}_{\text{obs}})$, is $\text{Ga}^{-1}(n/2, n(s^2+\bar{x}^2)/2)$, and the posterior predictive distribution of $\bar{X}$ is $m_{\text{post}}(\bar{x}|\mathbf{x}_{\text{obs}}) = t_n(\bar{x}|0, 1/n(s^2_{\text{obs}}+\bar{x}^2_{\text{obs}}))$; here Ga^{-1} and t_n denote the inverse gamma distribution and the t distribution with n degrees of freedom. From (6) and (14), it follows that

$$p_{\text{post}} = 2\left[1 - \Upsilon_n\left(\frac{\sqrt{n}\bar{x}_{\text{obs}}}{\sqrt{s^2_{\text{obs}}+\bar{x}^2_{\text{obs}}}}\right)\right], \quad (17)$$

where Υ_n represents the distribution function of the t distribution with n degrees of freedom. As could be expected, (17) is very similar to p_{plug}, given in (15). Indeed, it has the similar inappropriate behavior that $p_{\text{post}} \to 2[1 - \Upsilon_n(\sqrt{n})]$, a positive constant, as $|\bar{x}_{\text{obs}}|/s_{\text{obs}} \to \infty$. For instance, when $n = 4$ this constant is .12, and the posterior predictive p value never drops below this constant, no matter how many standard deviations $\bar{x}_{\text{obs}}$ is from 0. The inadequacy of p_{post} here (or of p_{plug}) can be directly traced to the double use of the data, in particular to the fact that $\bar{x}_{\text{obs}}$ is involved in computing both the posterior (or the MLE) and the tail area. Interestingly, the problem with p_{plug} is less severe than that with p_{post} in that the limiting constant is smaller (.046 when $n = 4$, for instance).

p_{cpred}: Computation shows that

$$f(\mathbf{x}|t;\sigma^2) \propto \frac{f(\mathbf{x};\sigma^2)}{f(t;\sigma^2)} \propto (\sigma^2)^{[(n-1)/2]} \exp\left\{-\frac{ns^2}{2\sigma^2}\right\},$$

which is maximized at $\hat{\sigma}^2_{\text{cMLE}} = ns^2/(n-1)$. As observed earlier, it is equivalent to take S^2 as the conditioning statistic. It is then easy to show that $\pi(\sigma^2|s^2)$ is $\text{Ga}^{-1}((n-1)/2, ns^2/2)$ and that $m(\bar{x}|s^2_{\text{obs}}) = t_{n-1}(\bar{x}|0, [1/(n-1)]s^2_{\text{obs}})$. The resulting conditional predictive p value is

$$p_{\text{cpred}} = 2\left[1 - \Upsilon_{n-1}\left(\frac{\sqrt{n-1}\bar{x}_{\text{obs}}}{s_{\text{obs}}}\right)\right]. \quad (18)$$

This is perfectly satisfactory and indeed equals the usual classical p value for the problem based on the usual one-sample t statistic, which is known to be uniform under the null.

p_{ppost}: In this case $T = \bar{X}$ is independent of $\hat{\sigma}^2_{\text{cMLE}} \propto S^2$ (and they are clearly jointly sufficient), so that the partial posterior predictive p value equals the conditional predictive p value, p_{cpred}, in (18).

It should be noted that Meng (1994) also considered use of the departure statistic $t(\mathbf{x}) = |\bar{x}|/s_{\text{obs}}$ in the foregoing example, and with this statistic, the posterior predictive p value and the plug-in p value perform fine (being then equal to the other p values). Note, however, that in more complex problems, it may be quite difficult to find "appropriate" departure statistics for use with the posterior predictive or plug-in p values (RVV 2000).

2.2 Motivation and Comparison

2.2.1 Bayesian Motivations for p_{ppost} and p_{cpred}. The U-conditional posterior predictive p values appear to combine the positive features of both the prior predictive and the posterior predictive p values. First, they are based on the prior predictive $m(\mathbf{x})$, which has natural Bayesian meaning; indeed, when $\pi(\theta)$ is proper, $m(t|u)$ is simply the conditional distribution of T given U arising from the prior predictive $m(\mathbf{x})$. Second, with appropriate choice of U, (10) can be made to primarily reflect surprise in the model, with the prior playing only a secondary role. Third, noninformative priors can be used, as long as $\pi(\theta|u)$ is proper. Finally, there is no double use of the data, because only part of the data (u_{obs}) is used to produce the posterior to eliminate θ, whereas another part (t_{obs}) is used when computing the tail area.

Of course, the key to the U-conditional predictive p value is a suitable choice of the conditioning statistic U. Different possible choices of U have been explored by Bayarri and Berger (1997). (See also Evans 1997, where the conditional predictive distribution was used to develop alternate measures of surprise, with U and T chosen to be separate subsamples of the data. A rather different possibility was given by the cross-validatory predictive distribution as described in Gelfand, Dey, and Chang 1992; see Carlin 1999 and the rejoinder in Bayarri and Berger 1999 for more discussion.) The intuition behind suitable choice of U is that one wants U to contain as much information about θ as possible, so that $\pi(\theta|u_{\text{obs}})$ will effectively eliminate θ (via integration), subject to the constraint that U should not involve T, as this could lead to a reduction in discriminatory power of the procedure. In Example 1, for instance, $\sum x_i^2/n$ would contain all information about σ^2 (being a sufficient statistic under the presumed model), but does involve $t(\mathbf{x}) = |\bar{x}|$. The obvious solution (used in Example 1) is to define $u(\mathbf{x}) = s^2 = \sum(x_i - \bar{x})^2/n$, because this contains the information about σ^2 that is independent of $t(\mathbf{X})$.

Investigations by Bayarri and Berger (1997) also suggest that $u(\mathbf{x})$ should have the same dimension as θ. The simplest general algorithm that achieves these various aims, for the case of continuous data, is to define U to be the conditional MLE of θ, given $t(\mathbf{x}) = t$, as defined in (13). [The situation of discrete data is considerably more difficult; whereas $\hat{\theta}_{\text{cMLE}}$ in (13) is still typically well defined, it will not be suitable as a conditioning statistic if the resulting conditional sample space contains too few values.]

While logically appealing, the conditional predictive p value, with the conditioning statistic $\hat{\theta}_{\text{cMLE}}$ chosen as in (13), can be difficult to compute. An attractive alternative is to directly use $f(\mathbf{x}|t;\theta)$ [see (13)] to integrate out θ, rather than simply using it to define $\hat{\theta}_{\text{cMLE}}$. This leads to the partial posterior predictive p value, defined in (8) and (9), which is typically much easier to work with. Furthermore, the parallel with (13) suggests that the partial posterior predictive p value will be very similar to the conditional predictive

p value. This is shown in our (continuous) examples and is further reinforced by RVV (2000), who show that p_{cpred} and p_{ppost} are asymptotically equivalent.

The foregoing Bayesian motivations may appear rather "loose," but history in other areas of statistics has shown that when sound Bayesian reasoning and noninformative priors are used to develop procedures, these procedures typically also have very desirable non-Bayesian properties. That this is so for p_{cpred} and p_{ppost} is discussed in the next section.

2.2.2 Frequentist Motivations and Comparisons. RVV (2000) show that p_{cpred} and p_{ppost} are asymptotic frequentist p values; that is, their asymptotic distribution is $U[0,1]$ for all θ. In this section we study whether this is so for small samples; we say that a p value is a *frequentist p value* for all θ if it is $U[0,1]$ for all θ. We present an illustrative example and two relevant theorems, the first of which follows.

Theorem 1. Let $p(\mathbf{X})$ be any U-conditional predictive p value for a proper $\pi(\theta)$, and consider it as a random variable with respect to the distribution $f(\mathbf{x};\theta)$. Assume that the distribution of $p(\mathbf{X})$ does not depend on θ, and that the conditional distribution of T, given U, is absolutely continuous. Then $p(\mathbf{X})$ is a frequentist p value for all θ. The conclusion also holds for improper $\pi(\theta)$ under condition (A.1) in the Appendix, which is in particular satisfied if U has a location or scale-parameter distribution and $\pi(\theta)$ is the reference prior.

Proof. Suppose that $\pi(\theta)$ is proper. Then both the conditional predictive distribution for $T, m(t|u) = \int f(t|u;\theta)\pi(\theta|u)\,d\theta$, and the prior predictive for $U, m(u) = \int f(u;\theta)\pi(\theta)\,d\theta$, are proper. Also, because $p(\mathbf{X})$ is by definition a proper p value with respect to $m(t|u)$, it follows that

$$\begin{aligned}\Pr^{m(\cdot)}(p(\mathbf{X}) \le \alpha) &= E^{m(u)}E^{m(t|u)}[p(\mathbf{X}) \le \alpha] \\ &= E^{m(u)}[\alpha] = \alpha.\end{aligned}$$

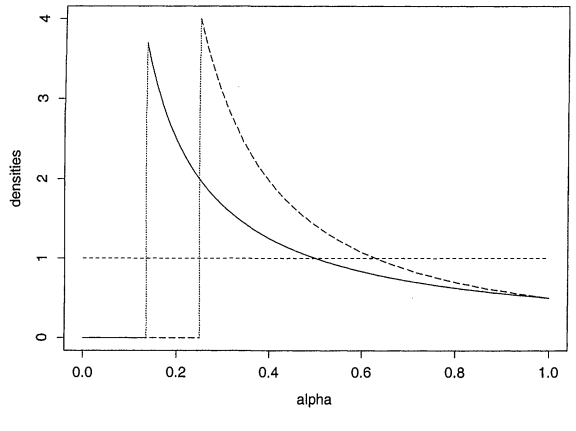

Figure 1. Densities of $p_{plug}(\mathbf{X})$ (——) and $p_{post}(\mathbf{X})$ (- - - -) in Example 2, when n = 2. The uniform density is plotted for reference.

But because

$$\Pr^{m(\cdot)}(p(\mathbf{X}) \le \alpha) = E^{\pi(\theta)}(E^{f(\mathbf{x};\theta)}[p(\mathbf{X}) \le \alpha]),$$

it follows that if $p(\mathbf{X})$ has a distribution that does not depend on θ, then

$$\Pr^{m(\cdot)}(p(\mathbf{X}) \le \alpha) = E^{\pi(\theta)}[c(\alpha)] = c(\alpha),$$

where $c(\alpha)$ is some function of α. It is immediate that $c(\alpha) = \alpha$ and hence that $p(\mathbf{X})$ is an exact p value. The proof for the improper case is given in the Appendix.

An obvious situation in which Theorem 1 applies is when U can be taken to be a sufficient statistic for θ. In that case $m(t|u) = f(t|u)$, and the U-conditional predictive p value equals the frequentist similar p value. Another application of Theorem 1 is to p_{cpred} (and p_{ppost}) in Example 1. From (18), it is clear that their distributions do not depend on σ^2, because the distribution of $\sqrt{(n-1)}\bar{X}/S$ is independent of σ^2. Also, $\hat{\sigma}^2_{\text{cMLE}}$ has a scale-parameter distribution, so it can be immediately concluded that p_{cpred} and p_{ppost} are frequentist p values for all σ^2.

In Example 1, p_{plug} and p_{post} will not be frequentist p values, but the extent to which they deviate from uniformity must be studied numerically. We thus turn to a simpler situation where exact computations can be performed.

Example 2. Assume that $X_1, X_2, \ldots, X_n$ is a random sample from the exponential(λ) distribution. Let $T = X_{(1)}$ (which could be used to investigate the lower tail of the null distribution) and assume that the usual noninformative prior $\pi(\lambda) = 1/\lambda$ is to be used. In the following, $X_{(1)} < X_{(2)} < \cdots < X_{(n)}$ denote the order statistics for the observations. Also, define $S = \sum_{i=1}^n X_i$ and let s_{obs} be the sum of the observed x_i. We derive the different p values and investigate their properties. The following fact, established in the Appendix, as used repeatedly:

$$\Pr\left(\frac{T}{S} \le c\right) = 1 - (1-nc)^{n-1}. \tag{19}$$

p_{plug}: Clearly, $\hat{\lambda} = n/S$ and $T \sim Ex(n\lambda)$, so that

$$p_{\text{plug}} = e^{-n^2 t_{\text{obs}}/s_{\text{obs}}}. \tag{20}$$

That this is conditionally unsatisfactory can be seen by taking $nt_{\text{obs}}/s_{\text{obs}} \to 1$, in which case the model would clearly be contraindicated, yet $p_{\text{plug}} \to e^{-n}$, a nonzero constant. To investigate whether p_{plug} is a frequentist p value for all λ, an easy computation using (20) and (19) yields, for $\alpha > e^{-n}$,

$$\begin{aligned}\Pr(p_{\text{plug}}(\mathbf{X}) \le \alpha) &= \Pr\left(\frac{T}{S} \ge -\frac{\log\alpha}{n^2}\right) \\ &= \left(1 + \frac{\log\alpha}{n}\right)^{n-1}. \end{aligned} \tag{21}$$

Thus $p_{\text{plug}}(\mathbf{X})$ does not have a $U[0,1]$ distribution and is not a frequentist p value. Figure 1 graphs the density corresponding to (21) when $n = 2$, to show the

substantial deviation from uniformity that can occur. Note, however, that $p_{\text{plug}}(\mathbf{X})$ is an asymptotic frequentist p value. Indeed, for large n, (21) is approximately given by

$$\Pr(p_{\text{plug}}(\mathbf{X}) \le \alpha) \approx \alpha\left[1 - \frac{1}{n}\left(\log\alpha + \frac{1}{2}\log^2\alpha\right)\right],$$

which does go to α as $n \to \infty$.

p_{sim}: Because S is sufficient, the distribution of $X_1, X_2, \ldots, X_n$ given s is uniform on the set $\{\mathbf{X} : \sum_{i=1}^n X_i = s\}$, and so

$$p_{\text{sim}} = \Pr(T > t_{\text{obs}} | s_{\text{obs}}) = \Pr\left(W_{(1)} > \frac{nt_{\text{obs}}}{s_{\text{obs}}}\right) = \left(1 - \frac{nt_{\text{obs}}}{s_{\text{obs}}}\right)^{(n-1)},$$

where $W_{(1)} = \min(W_1, W_2, \ldots, W_n)$ and the W_i are iid $U(0,1)$. This will be seen to be equal to p_{ppost}.

p_{prior}: The prior predictive p value cannot be computed for this example, because the prior distribution is improper.

p_{post}: The posterior distribution of λ is easily seen to be $\text{Ga}(n, s_{\text{obs}})$, and the posterior predictive density of T is $(n^2/s_{\text{obs}})[(s_{\text{obs}}/(nt + s_{\text{obs}}))]^{n+1}$. The posterior predictive p value can then be computed as

$$p_{\text{post}} = \Pr^{m_{\text{post}}(t|\mathbf{x}_{\text{obs}})}(T > t_{\text{obs}}) = \left(1 + \frac{nt_{\text{obs}}}{s_{\text{obs}}}\right)^{-n}.$$

It can be seen that $p_{\text{post}} \to 2^{-n}$, a nonzero constant, as $nt_{\text{obs}}/s_{\text{obs}} \to 1$, which is not appropriate behavior. Moreover, the distribution of p_{post} is not $U[0,1]$. Indeed, for $\alpha > 2^{-n}$,

$$\Pr(p_{\text{post}}(\mathbf{X}) \le \alpha) = \Pr\left[\frac{T}{S} \ge -\frac{1}{n}\left(\frac{1}{\alpha^{1/n}}\right)\right] = (2 - \alpha^{-1/n})^{n-1}. \quad (22)$$

The corresponding density function is graphed in Figure 1 when $n = 2$, and is even further from uniformity than is the density corresponding to p_{plug}. Again, however, p_{post} can be shown to be asymptotically $U[0,1]$.

p_{ppost}: An easy computation shows that

$$f(\mathbf{x}|t;\lambda) \propto \lambda^{n-1}\exp\left\{-\lambda\left(\sum x_i - nt\right)\right\}, \quad (23)$$

so that the partial posterior for λ is

$$\pi(\lambda|\mathbf{x}_{\text{obs}} \setminus t_{\text{obs}}) = \frac{\lambda^{n-2}e^{-\lambda(s_{\text{obs}} - nt_{\text{obs}})}}{\Gamma(n-1)(s_{\text{obs}} - nt_{\text{obs}})^{-(n-1)}}$$

and the partial posterior predictive density is

$$m(t|\mathbf{x}_{\text{obs}} \setminus t_{\text{obs}}) = \frac{n(n-1)(s_{\text{obs}} - nt_{\text{obs}})^{n-1}}{(nt + s_{\text{obs}} - nt_{\text{obs}})^n}.$$

The partial posterior predictive p value can then be computed as

$$p_{\text{ppost}} = \Pr^{m(t|\mathbf{x}_{\text{obs}} \setminus t_{\text{obs}})}(T > t_{\text{obs}}) = \left(1 + \frac{nt_{\text{obs}}}{s_{\text{obs}} - nt_{\text{obs}}}\right)^{-(n-1)} = \left(1 - \frac{nt_{\text{obs}}}{s_{\text{obs}}}\right)^{n-1},$$

which is identical to the similar p value. It can be shown that $p_{\text{ppost}} \to 0$ as $nt_{\text{obs}}/s_{\text{obs}} \to 1$, so that there are no apparent conditional difficulties with the partial posterior predictive p value. It will also be seen that p_{ppost} is a frequentist p value for all n.

p_{cpred}: Maximization of (23) over λ yields $\hat{\lambda}_{\text{cMLE}} \propto \sum_{i=1}^n X_i - nX_1 = S - nT$. It can be seen that $X_{(i)} - X_{(1)}$ is independent of $X_{(1)}$, from which it follows directly that $\hat{\lambda}_{\text{cMLE}}$ is independent of T. As discussed in the paragraph preceeding Example 1, it follows that $p_{\text{cpred}} = p_{\text{ppost}}$. Note that the derivation of p_{ppost} was considerably simpler than that of p_{cpred}. Finally, as in the argument leading to (19), it can be shown that $\Pr(p_{\text{ppost}}(\mathbf{X}) \le \alpha)$ does not depend on λ. Also, $\hat{\lambda}_{\text{cMLE}}$ has a scale-parameter distribution, and so, by Theorem 1, p_{cpred} (and hence also p_{ppost} and p_{sim}) is a frequentist p value. Notice that Theorem 1 cannot be directly applied to p_{ppost}.

It is something of a curiosity that in both Examples 1 and 2, p_{sim} coincides with p_{cpred} and p_{ppost}, especially because p_{sim} and p_{cpred} are determined from distributions on completely different (conditional) spaces. This is useful methodologically for those who wish to use p_{cpred} or p_{sim}, because it is typically much easier to derive p_{ppost} than either of the other two p values. The following theorem gives more general conditions under which this equivalence holds. (It is easy to see that Examples 1 and 2 both satisfy the conditions of the theorem.)

Theorem 2. Suppose that $f(\mathbf{x};\theta)$ is a continuous density from the natural scale exponential family and that statistics $T > 0$ and $U > 0$ exist such that $S = T + U$ is sufficient and

$$f(t,u;\theta) = k\theta^\alpha t^\gamma u^{\alpha-\gamma-2}\exp\{-\theta(t+u)\},$$

for some constants $k, \gamma > -1$, and $\gamma < \alpha - 1$. Under the usual noninformative prior, $\pi(\theta) = 1/\theta$, it will be the case that $p_{\text{cpred}}, p_{\text{ppost}}$, and p_{sim} are all equal.

Proof. That p_{cpred} and p_{ppost} are equal follows from direct calculation. To show their equality with p_{sim}, first integrate $f(t,u;\theta)$ with respect to $\pi(\theta) = 1/\theta$, obtaining

$$m(t,u) \propto \frac{t^\gamma u^{\alpha-\gamma-2}}{(t+u)^\alpha}.$$

Thus $m(t|u_{\text{obs}}) = ct^{\gamma}(t+u_{\text{obs}})^{-\alpha}$, where c is the appropriate normalizing constant, and hence

$$p_{\text{cpred}} = \int_{t_{\text{obs}}}^{\infty} m(t|u_{\text{obs}})\,dt = c\int_{t_{\text{obs}}}^{\infty} \frac{t^{\gamma}}{(t+u_{\text{obs}})^{\alpha}}\,dt. \quad (24)$$

An easy computation shows that the conditional density of T given S is

$$f^*(t|s) = c^*\frac{t^{\gamma}(s-t)^{\alpha-\gamma-2}}{s^{\alpha}}, \quad \text{for} \quad 0 < t < s,$$

where c^* is the appropriate normalizing constant. Hence p_{sim} is given by

$$p_{\text{sim}} = \int_{t_{\text{obs}}}^{\infty} f(t|s_{\text{obs}})\,dt = c^*\int_{t_{\text{obs}}}^{s_{\text{obs}}} t^{\gamma}(s_{\text{obs}}-t)^{\alpha-\gamma-2}\,dt.$$

Changing variables to $t = (s_{\text{obs}}w)/(w+u_{\text{obs}})$, the latter integral reduces to that in (24), and the theorem follows.

2.3 Computation

Simulation methods are typically needed to compute the partial posterior predictive p value. These simulations will typically be only modestly more difficult than those involved in computation of either the prior predictive p value or the posterior predictive p value, providing that $f(t_{\text{obs}};\theta)$ is available in closed form.

Noting that the partial posterior predictive p value can be rewritten as

$$p_{\text{ppost}} = \int \Pr(T \geq t_{\text{obs}};\theta)\pi(\theta|\mathbf{x}_{\text{obs}} \setminus t_{\text{obs}})\,d\theta,$$

an obvious strategy is to repeatedly generate θ from $\pi(\theta|\mathbf{x}_{\text{obs}} \setminus t_{\text{obs}})$ and then T from $f(t;\theta)$ [which could of course be done by simply generating $\mathbf{X}$ from $f(\mathbf{x};\theta)$ and computing $t(\mathbf{X})$], estimating p_{ppost} by the fraction of generated T that are greater than t_{obs}.

There are various possibilities for generating from $\pi(\theta|\mathbf{x}_{\text{obs}} \setminus t_{\text{obs}})$. If generation from the full posterior $\pi(\theta|\mathbf{x}_{\text{obs}})$ is easy, then a natural possibility is to use the following simple Metropolis chain: use $\pi(\theta|\mathbf{x}_{\text{obs}})$ as the probing distribution to obtain a candidate θ^*, and then move from the current θ to the candidate with probability $\min\{1, f(t_{\text{obs}};\theta)/f(t_{\text{obs}};\theta^*)\}$.

In some sense, a "bad" discrepancy statistic T is one for which $f(t_{\text{obs}};\theta)$ is highly variable in θ. (Casually chosen T for model checking will often have this property.) Whereas such T will not yield good p values by standard methods, the results in this article (and in RVV 2000) indicate that the new conditional p values will still be quite satisfactory. The price to be paid, however, is that computation of the new p values can be more difficult with such T, because $\pi(\theta|\mathbf{x}_{\text{obs}})$ may no longer be a good probing distribution. A slight modification of the foregoing Metropolis chain can then be considerably more efficient: generate U, a uniform random variable on (0, 1); generate θ' from $\pi(\theta|\mathbf{x}_{\text{obs}})$; and choose the candidate $\theta^* = \theta' + U(\hat{\theta} - \hat{\theta}_{\text{cMLE}})$, where $\hat{\theta}$ and $\hat{\theta}_{\text{cMLE}}$ are the MLE and the conditional MLE [see (13)] of θ. Then, move from the current $\theta = \theta^o + U^o(\hat{\theta} - \hat{\theta}_{\text{cMLE}})$ to the candidate θ^* with probability

$$\min\left\{1, \frac{f(t_{\text{obs}};\theta)}{f(t_{\text{obs}};\theta^*)}\,\frac{\pi(\theta^*)}{\pi(\theta)}\,\frac{\pi(\theta^o)}{\pi(\theta')}\,\frac{f(\mathbf{x}_{\text{obs}};\theta^*)}{f(\mathbf{x}_{\text{obs}};\theta)}\,\frac{f(\mathbf{x}_{\text{obs}};\theta^o)}{f(\mathbf{x}_{\text{obs}};\theta')}\right\}.$$

Alternatives to such direct Monte Carlo computation of p_{ppost} include importance sampling schemes. For instance, if a (possibly dependent) sample $\{\theta_j, j = 1,\ldots,m\}$ from $\pi(\theta|\mathbf{x}_{\text{obs}})$ were available, then one could estimate p_{ppost} by

$$\hat{p}_{\text{ppost}} = \frac{\sum_{j=1}^{m}\Pr(T \geq t_{\text{obs}};\theta_j)/f(t_{\text{obs}};\theta_j)}{\sum_{j=1}^{m} 1/f(t_{\text{obs}};\theta_j)}.$$

[This would work because $f(x_{\text{obs}};\theta)$ is typically considerably more concentrated than $f(t_{\text{obs}};\theta)$.]

When $f(t_{\text{obs}};\theta)$ is not available in closed form, it must be estimated, possibly through some type of kernel estimate; note that T typically is a one-dimensional statistic, and estimation of a one-dimensional density at a point usually is not excessively difficult. Of course, this estimation must be done in conjunction with the Metropolis or importance sampling schemes mentioned earlier, and efficiency might improve if one keeps only widely spaced (i.e., approximately independent) θ.

Computing $p_{\text{cpred}(u)}$ is usually considerably more difficult, unless the densities on the left side or the right side of (11) are available in closed form. Various Gibbs and Metropolis–Hasting schemes for its computation were given by Bayarri and Berger (1999) and are not repeated here. The computational difficulty of $p_{\text{cpred}(u)}$ (see also Pauler 1999) is the main reason why we recommend p_{ppost} for routine use.

3. COMPARISONS IN THE NORMAL LINEAR MODEL

In this section we derive the various p values for the normal linear model and give characterizations of their degree of uniformity (frequentist sense). This section was motivated by RVV (2000), who derive corresponding results under assumptions that yield asymptotic normality. Seeing the results in the finite-sample setting (under normality) should help alleviate any concerns about "asymptopia." Note that in this section we do not attempt to distinguish between the MLE and its value at the observed data; both are denoted by $\hat{\boldsymbol{\theta}}$.

Let $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)^t$ be the $n \times 1$ vector of response variables, let $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)^t$ be the $k \times 1$ vector of regression coefficients, let $\mathbf{V}$ be a full-rank $n \times k$ matrix of covariables, and let $\boldsymbol{\varepsilon}$ be an $n \times 1$ vector of errors. Assume that we are testing

$$H_0 : \mathbf{Y} = \mathbf{V}\boldsymbol{\theta} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim \mathrm{N}_n(\mathbf{0}, \sigma^2\mathbf{1}), \quad \sigma^2 \text{ known.} \quad (25)$$

Consider a linear departure statistic $T = \mathbf{w}^t\mathbf{Y}$, with given $\mathbf{w} = (w_1, w_2, \ldots, w_n)^t$. It follows from (25) that

$$T|\boldsymbol{\theta} \sim \mathrm{N}(\mathbf{w}^t\mathbf{V}\boldsymbol{\theta}, \sigma^2\|\mathbf{w}\|^2). \quad (26)$$

Also, with the usual noninformative prior, $\pi(\boldsymbol{\theta}) = 1$, the posterior distribution, $\pi(\boldsymbol{\theta}|\mathbf{y})$, is $\mathrm{N}_k(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}, \sigma^2(\mathbf{V}^t\mathbf{V})^{-1})$, where $\hat{\boldsymbol{\theta}} = (\mathbf{V}^t\mathbf{V})^{-1}\mathbf{V}^t\mathbf{y}$ is the usual least squares estimate.

3.1 Plug-In *p* Value

It follows from (26) that p_{plug} is given by

$$p_{\text{plug}} = \Pr^{f(t;\hat{\theta})}(T > t_{\text{obs}}) = 1 - \Phi\left(\frac{t_{\text{obs}} - \mathbf{w}^t\mathbf{V}\hat{\boldsymbol{\theta}}}{\sigma\|\mathbf{w}\|}\right).$$

To study the distribution of $p_{\text{plug}}(\mathbf{Y})$ (to assess its frequentist uniformity), note that

$$T - \mathbf{w}^t\mathbf{V}\hat{\boldsymbol{\theta}} \sim \mathrm{N}(\mathbf{w}^t\mathbf{B}\mathbf{V}\boldsymbol{\theta}, \sigma^2\mathbf{w}^t\mathbf{B}\mathbf{B}^t\mathbf{w}), \tag{27}$$

where $\mathbf{B} = \mathbf{I} - \mathbf{V}(\mathbf{V}^t\mathbf{V})^{-1}\mathbf{V}^t$. Because $\mathbf{B}\mathbf{V} = \mathbf{0}$ and $\mathbf{B}\mathbf{B}^t = \mathbf{B}$, it follows that

$$p_{\text{plug}}(\mathbf{Y}) = 1 - \Phi\left(\sqrt{\frac{\mathbf{w}^t\mathbf{B}\mathbf{w}}{\|\mathbf{w}\|^2}}Z\right), \tag{28}$$

where $Z \sim \mathrm{N}(0,1)$. Thus $p_{\text{plug}}(\mathbf{Y})$ will have a $U[0,1]$ distribution only if $\mathbf{w}^t\mathbf{B}\mathbf{w}/\|\mathbf{w}\|^2 = 1$, which in turn can happen only if $\mathbf{V}^t\mathbf{w} = 0$. Although the latter will be satisfied by common choices of T, such as a linear function of the vector of residuals, it clearly need not hold in general. When it does not hold, $\mathbf{w}^t\mathbf{B}\mathbf{w}/\|\mathbf{w}\|^2$ will be smaller than 1, so that p_{plug} will be conservative.

3.2 Posterior Predictive *p* Value

The posterior predictive distribution of T, given $\mathbf{y}_{\text{obs}}$ is $\mathrm{N}(\mathbf{w}^t\mathbf{V}\hat{\boldsymbol{\theta}}, \sigma^2\mathbf{w}^t\mathbf{C}\mathbf{w})$, where $\mathbf{C} = \mathbf{I} + \mathbf{V}(\mathbf{V}^t\mathbf{V})^{-1}\mathbf{V}^t$. It follows that the posterior predictive p value is given by

$$p_{\text{post}} = \Pr^{m_{\text{post}}(t|\mathbf{y}_{\text{obs}})}(T > t_{\text{obs}}) = 1 - \Phi\left(\frac{t_{\text{obs}} - \mathbf{w}^t\mathbf{V}\hat{\boldsymbol{\theta}}}{\sigma\sqrt{\mathbf{w}^t\mathbf{C}\mathbf{w}}}\right).$$

When considered as a random p value and using (27), p_{post} can be expressed as

$$p_{\text{post}}(\mathbf{Y}) = 1 - \Phi\left(\sqrt{\frac{\mathbf{w}^t\mathbf{B}\mathbf{w}}{\mathbf{w}^t\mathbf{C}\mathbf{w}}}Z\right). \tag{29}$$

Again, this will be $U[0,1]$ only if $\mathbf{V}^t\mathbf{w} = \mathbf{0}$. Otherwise, $\mathbf{w}^t\mathbf{C}\mathbf{w}$ will be larger than $\|\mathbf{w}\|^2$ and, comparing (28) and (29), p_{post} will then be even more conservative than p_{plug}. This observation was first made in the asymptotic setting by Robins (1999) and RVV (2000).

3.3 Partial Posterior Predictive *p* Value

Calculation yields

$$f(\mathbf{y}|t_{\text{obs}};\boldsymbol{\theta}) \propto \exp\left\{-\frac{1}{2\sigma^2}\left[(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^t\mathbf{V}^t\mathbf{V}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) - (\mathbf{w}^t\mathbf{V}\boldsymbol{\theta} - T)^t(\|\mathbf{w}\|^2)^{-1}(\mathbf{w}^t\mathbf{V}\boldsymbol{\theta} - T)\right]\right\}. \tag{30}$$

Because the partial posterior distribution $\pi(\boldsymbol{\theta}|\mathbf{y}_{\text{obs}} \setminus t_{\text{obs}})$ is proportional to (30), expanding the quadratic forms in (30) and rearranging terms yields

$$\pi(\boldsymbol{\theta}|\mathbf{y}_{\text{obs}} \setminus t_{\text{obs}}) = \mathrm{N}_k(\boldsymbol{\theta}|\mathbf{u}_{\text{obs}}, \sigma^2\boldsymbol{\Sigma}), \tag{31}$$

where $\mathbf{U} = (\mathbf{V}^t\mathbf{H}\mathbf{V})^{-1}\mathbf{V}^t\mathbf{H}\mathbf{Y}, \boldsymbol{\Sigma} = (\mathbf{V}^t\mathbf{H}\mathbf{V})^{-1}, \mathbf{H} = [\mathbf{I} - (\mathbf{w}\mathbf{w}^t/\|\mathbf{w}\|^2)]$, and the right side of (31) denotes the k-variate normal density in $\boldsymbol{\theta}$ with the given mean and covariance matrix. From (26) and (31), it follows that the partial predictive distribution of T is given by

$$T|\mathbf{y}_{\text{obs}} \setminus t_{\text{obs}} \sim \mathrm{N}(\mathbf{w}^t\mathbf{V}\mathbf{u}_{\text{obs}}, \sigma^2\mathbf{w}^t[\mathbf{I} + \mathbf{V}\boldsymbol{\Sigma}\mathbf{V}^t]\mathbf{w}),$$

so that the partial posterior predictive p value is

$$p_{\text{ppost}} = \Pr^{m(t|\mathbf{y}_{\text{obs}}\setminus t_{\text{obs}})}(T > t_{\text{obs}}) = 1 - \Phi\left(\frac{t_{\text{obs}} - \mathbf{w}^t\mathbf{V}\mathbf{u}_{\text{obs}}}{\sigma\sqrt{\mathbf{w}^t[\mathbf{I} + \mathbf{V}\boldsymbol{\Sigma}\mathbf{V}^t]\mathbf{w}}}\right).$$

To study the distribution of $p_{\text{ppost}}(\mathbf{Y})$, note that

$$\mathbf{T} - \mathbf{w}^t\mathbf{V}\mathbf{U}|\boldsymbol{\theta} \sim \mathrm{N}(\mathbf{w}^t\mathbf{D}\mathbf{V}\boldsymbol{\theta}, \sigma^2\mathbf{w}^t\mathbf{D}\mathbf{D}^t\mathbf{w}),$$

where $\mathbf{D} = \mathbf{I} - \mathbf{V}(\mathbf{V}^t\mathbf{H}\mathbf{V})^{-1}\mathbf{V}^t\mathbf{H}$ and $\mathbf{H}$ is as in (31). Algebra shows that $\mathbf{w}^t\mathbf{D}\mathbf{V} = \mathbf{0}$ and $\mathbf{w}^t\mathbf{D}\mathbf{D}^t\mathbf{w} = \mathbf{w}^t[\mathbf{I} + \mathbf{V}\boldsymbol{\Sigma}\mathbf{V}^t]\mathbf{w}$, so that $p_{\text{ppost}}(\mathbf{Y}) = 1 - \Phi(Z)$, where $Z \sim \mathrm{N}(0,1)$. Thus p_{ppost} is a valid frequentist p value.

3.4 Conditional Predictive *p* Value

It can easily be seen from (30) and (31) that the $\mathbf{U}$ maximizing (30) is precisely the statistic $\mathbf{U}$ given in (31); that is, $\mathbf{U} = (\mathbf{V}^t\mathbf{H}\mathbf{V})^{-1}\mathbf{V}^t\mathbf{H}\mathbf{Y}$. Because T and $\mathbf{U}$ have a joint $(k+1)$-variate normal distribution such that $\mathrm{cov}(T, \mathbf{U}) = \mathbf{w}^t\mathbf{H}\mathbf{V}(\mathbf{V}^t\mathbf{H}\mathbf{V})^{-1} = \mathbf{0}$, they are independent. It follows that p_{cpred} equals p_{ppost}, and hence it is also a valid frequentist p value.

4. DISCRETE SAMPLE SPACES

For discrete sample spaces, the most common classical approach to defining p values is to condition on a statistic U such that $f(x|u;\theta)$ does not depend on θ. The Fisher exact test is the prototypical example that we consider here. Note that conditioning on any U can severely constrain the sample space, resulting in serious conservatism of the resulting p value (because there may then be very few possible observations in the "tail" of the departure statistic, T). We see that p_{ppost} can substantially overcome this difficulty. [We do not consider p_{cpred}, because the choice of the conditioning statistic in (13) typically will not work in discrete problems, and because any conditional p value can fall prey to the same difficulty mentioned earlier.]

We specifically consider the problem of testing homogeneity and independence in 2×2 contingency tables, comparing the similar p value (which is the *Fisher exact test*) and the partial posterior predictive p value. Many other p values for contingency tables have been proposed (a nice survey of proposed tests was given in Agresti 1992; see also Hwang and Yang 1997), and many of these perform considerably better than the Fisher exact test. Our attitude here is not that of seeking to determine an optimal p value for these situations, but rather to see if straightforward implementation of p_{ppost} can offer significant gains. (Recall that we hope to see p_{ppost} used in situations of considerable complexity, in which there is little hope of determining optimal p values; in judging the effectiveness of p_{ppost}, however, it is useful

to consider moderately difficult situations, such as this, to see whether an easy implementation works.)

Consider the following 2×2 contingency table:

	A_1	A_2	Totals
B_1	X_{11}	X_{12}	X_{1+}
B_2	X_{21}	X_{22}	X_{2+}
Totals	X_{+1}	X_{+2}	n

We analyze two common scenarios involving such tables.

Case 1. One of the margins, say $X_{+1} = n_1, X_{+2} = n_2$, is fixed by the design, so that X_{11} and X_{12} can be viewed as independent binomial random variables. We want to study the null model of homogeneity, that the two binomial distributions have the same success probability, θ.

Case 2. The design fixes only the overall sample size, n. A common null model is that classification by A and B is independent, so that the probability of each cell is the product of the corresponding marginal probabilities.

For convenience of notation in this section, we denote an observed value of $X_{\cdot\cdot}$ by a superscript "o"; that is, $x^o_{\cdot\cdot}$.

4.1 Case 1: Test of Homogeneity

Here the null model is

$$f(x_{11}, x_{12}; \theta) = \binom{n_1}{x_{11}} \binom{n_2}{x_{12}} \theta^{x_{11}+x_{12}} (1-\theta)^{n-x_{11}-x_{12}},$$

$$x_{11} = 0, \ldots, n_1, \qquad x_{12} = 0, \ldots, n_2. \quad (32)$$

The Fisher exact test conditions on the other marginal total, say X_{1+}. It is common in textbooks to then take the test statistic (in the conditional problem) to be $T = X_{11}$. An easy computation shows that the (one-tailed) p value corresponding to the Fisher exact test (p_{fet}) is given by

$$p_{\text{fet}} = \sum_{j=t_{\text{obs}}}^{\min\{x^o_{1+}, n_1\}} f(j|x^o_{1+})$$

$$= \sum_{j=t_{\text{obs}}}^{\min\{x^o_{1+}, n_1\}} \binom{n_1}{j} \binom{n_2}{x^o_{1+} - j} \Big/ \binom{n}{x^o_{1+}}.$$

Unconditionally, $T = X_{11}$ is not a particularly sensible statistic for measuring departure from homogeneity. Indeed, Suissa and Shuster (1985) proposed a sensible unconditional T for this particular problem. In illustrating p_{ppost}, however, we first study what happens if one naively "follows the textbooks" and chooses $T = X_{11}$, even though this is not sensible unconditionally. (Our point is to show that p_{ppost} behaves admirably even with a simple, but rather inappropriate, choice of T.) Then we consider a choice of T that is more reasonable from an unconditional perspective.

Choosing the constant prior $\pi(\theta) = 1$, an easy computation shows that the partial posterior distribution is beta$(x^o_{12} + 1, n_2 - x^o_{12} + 1)$ and

$$p_{\text{ppost}} = \sum_{j=t_{\text{obs}}}^{n_1} m(t|\mathbf{x}_{\text{obs}} \setminus t_{\text{obs}})$$

$$= \sum_{j=t_{\text{obs}}}^{n_1} \frac{n_2+1}{n+1} \binom{n_1}{j} \binom{n_2}{x^o_{12}} \Big/ \binom{n}{x^o_{12}+j}.$$

Incidentally, it can be shown in this problem (for the given choice of T) that $p_{\text{cpred}} = p_{\text{ppost}}$; this is thus a discrete situation in which conditioning as in (13) does not unduly restrict the sample space.

A more sensible unconditional choice of the discrepancy statistic is $T = [(1/n_1)X_{11} - (1/n_2)X_{22}]$ (because the null model is that the two binomial populations have the same success probability). The partial posterior predictive p value for this choice does not admit a simple closed-form expression but can be readily computed numerically.

Example 3. As a rather extreme test case, consider $n_1 = n_2 = 3$. Here conditioning on x_{1+} severely restricts the support of the distribution of p_{fet}, which reduces to {.05, .2, .5, .8, .95, 1}. The supports of the distributions of p_{ppost}, for either choice of T, are considerably richer and include more values closer to 0 and 1.

Figure 2 gives the distribution functions of $p_{\text{fet}}(\mathbf{X})$ and $p_{\text{ppost}}(\mathbf{X})$ (for both choices of T) at two different values of θ. Recall that the goal is to have p values with close to uniform distributions, and p_{ppost} clearly fares much better in this regard (the straight dotted lines being the unattainable uniform ideal). As expected, the Fisher exact test is very conservative, which translates into a severe lack of discriminatory power.

Note that p_{ppost} seems to perform somewhat better with the "sensible" discrepancy statistic $T = [(1/n_1)X_{11} - (1/n_2)X_{22}]$ than with $T = X_{11}$, in the sense that it then has a distribution somewhat closer to uniform in the most interesting region of small values of p. However, p_{ppost} seems to be quite satisfactory (and much better than the Fisher exact test), even when the intuitively unsuitable $T = X_{11}$ is used.

Various other θ were also considered. The distribution of p_{ppost} for the "sensible" choice of T is remarkably stable and performs very well for all values of θ. The other two p values (p_{fet} and p_{ppost} with $T = X_{11}$) were excessively conservative for small θ, although p_{ppost} began to perform noticeably better even for values of θ as small as .2. For large values of θ, p_{fet} was again very conservative, whereas p_{ppost} performed remarkably well, unless θ was very large; we discuss this latter situation more fully later.

4.2 Case 2: Test of Independence

Referring to the contingency table with fixed n and defining $\theta = \Pr(A_1)$ and $\xi = \Pr(B_1)$, the null model under independence of classification can be expressed as

$$f(\mathbf{x}; \theta, \xi) = \left(\frac{n!}{x_{11}!x_{12}!x_{21}!x_{22}!}\right) \theta^{x_{+1}} (1-\theta)^{x_{+2}} \xi^{x_{1+}} (1-\xi)^{x_{2+}}.$$

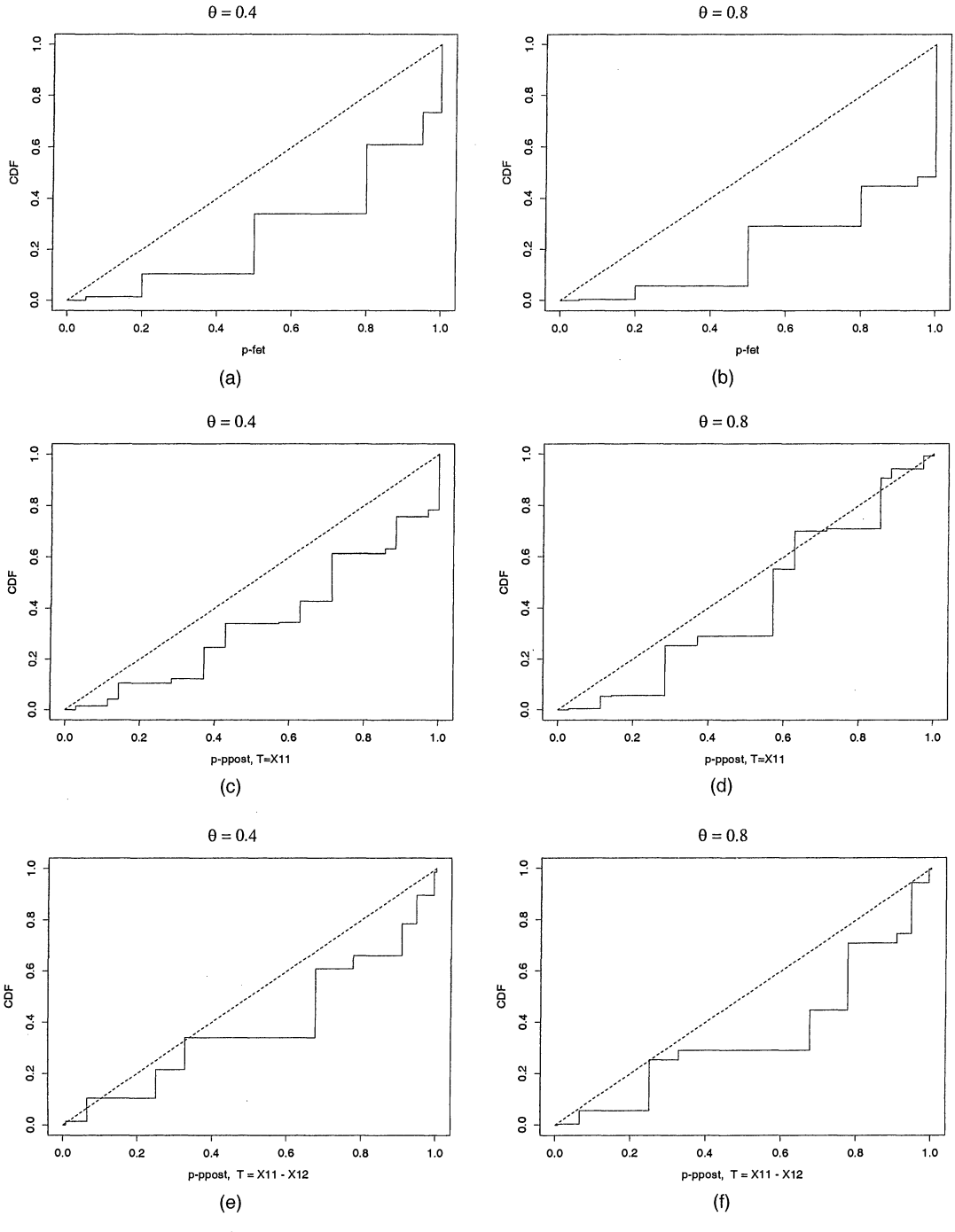

Figure 2. Distributions of $p_{fet}(\mathbf{X})$ [(a) and (b)]; $p_{ppost}(\mathbf{X})$ with $T = X_{11}$ [(c) and (d)] and $T = X_{11} - X_{12}$ [(e) and (f)] for $\theta = .4$ [(a), (c), and (e)] and $\theta = .8$ [(b), (d), and (f)] in Example 3.

Here the Fisher exact test conditions on both margins, and again the "textbook" conditional departure statistic is typically chosen to be $T = X_{11}$. The ensuing conditional density of T is

$$f(t|x^o_{1+}, x^o_{+1}) = \binom{x^o_{1+}}{t}\binom{n - x^o_{1+}}{x^o_{+1} - t} \Big/ \binom{n}{x^o_{+1}},$$

which produces the same p value as in the test for homogeneity. (Note that here, x^o_{1+} plays the role of n_1 in Case 1.)

In deriving p_{ppost}, we restrict attention to the statistic $T = X_{11}$, even though this is not particularly sensible from an unconditional perspective. We do this in part so that it cannot be argued that we obtain better results than p_{fet} by choice of a better T and, in part, to indicate the quality of p_{ppost} even with an inferior choice of T.

Using uniform independent priors for θ and ξ, the partial posterior, $\pi(\theta, \xi|\mathbf{x}_{\text{obs}} \backslash t_{\text{obs}})$, is proportional to $f(\mathbf{x}|t_{\text{obs}}; \theta, \xi)$, and p_{ppost} can most conveniently be expressed as

$$p_{\text{ppost}} = \int_0^1 \int_0^1 \pi(\theta, \xi|\mathbf{x}_{\text{obs}} \setminus t_{\text{obs}}) \sum_{t=t_{\text{obs}}}^{n} \text{binomial}(t|n, \theta\xi)\, d\theta\, d\xi, \quad (33)$$

where

$$\pi(\theta, \xi | \mathbf{x}_{\text{obs}} \setminus t_{\text{obs}}) \propto \theta^{x^o_{21}}(1-\theta)^{x^o_{+2}}\xi^{x^o_{12}}(1-\xi)^{x^o_{2+}}(1-\theta\xi)^{-(n-t_{\text{obs}})}. \quad (34)$$

(Note that for this case, p_{ppost} and p_{cpred} differ, and indeed the latter can be problematical because of possibly restrictive conditioning.)

To compute p_{ppost}, we use importance sampling based on the importance function

$$\frac{1}{2}U(\theta|0,1)\text{beta}(\xi|x^o_{12}+1, x^o_{22}+1) + \frac{1}{2}\text{beta}(\theta|x^o_{21}+1, x^o_{22}+1)U(\xi|0,1). \quad (35)$$

Not only is this an easy importance function to use in terms of random variable generation, but it also is highly efficient computationally, for even very large n. The reasons for this are given in the Appendix, which also presents other computational details.

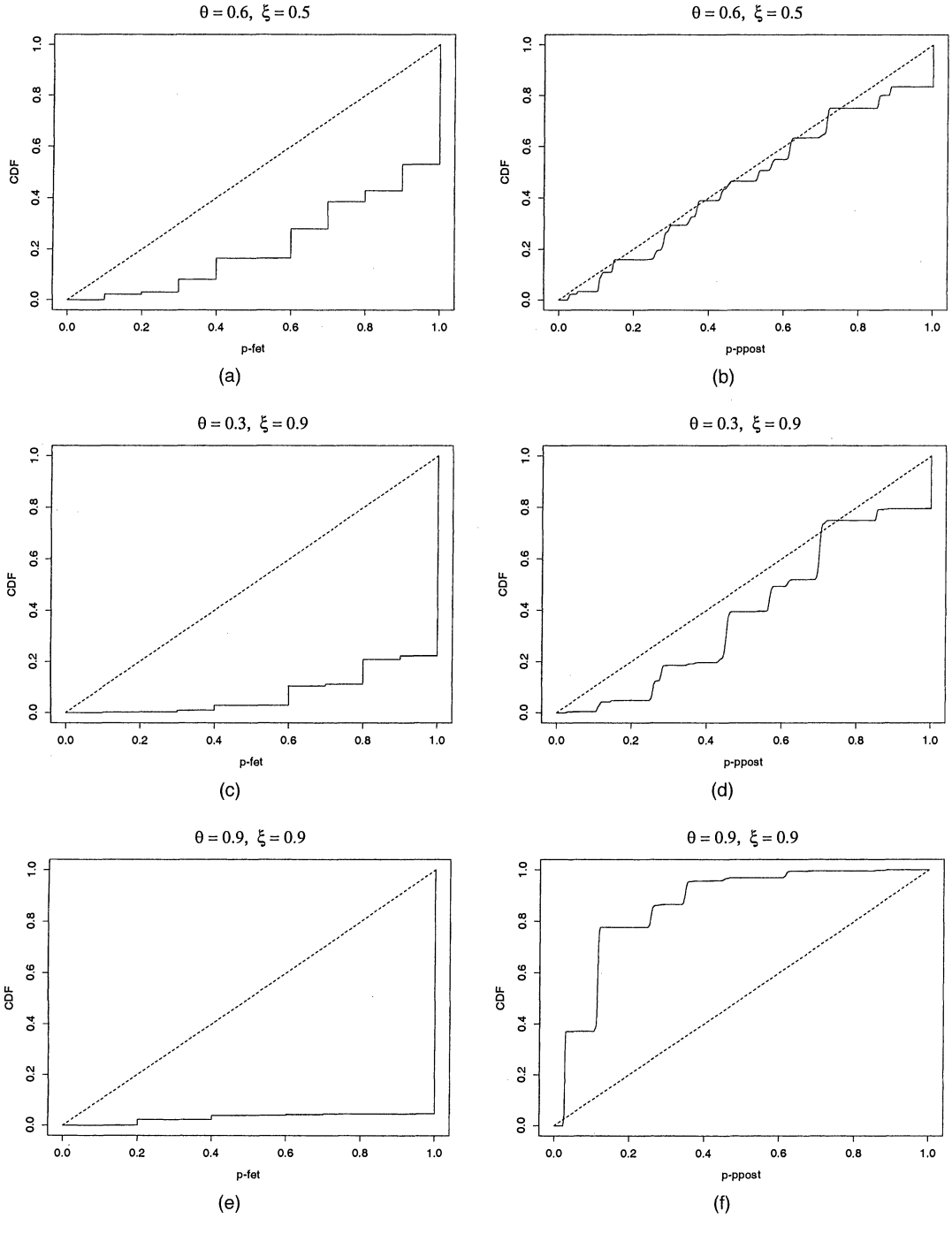

Figure 3. Distributions of $p_{fet}(\mathbf{X})$ [(a), (c), and (e)] and $p_{ppost}(\mathbf{X})$ [(b), (d), and (f)] in Example 4 for $(\theta;\xi) = (.6, .5)$ [(a) and (b)] $(\theta;\xi) = (.3, .9)$ [(c) and (d)] and $(\theta;\xi) = (.9, .9)$ [(e) and (f)].

Example 4. We again consider a rather extreme case, namely $n = 5$. It can be shown that the support of $p_{\text{fet}}(\mathbf{X})$ is limited to {.1, .2, .3, .4, .6, .7, .8, .9}, whereas the support of $p_{\text{ppost}}(\mathbf{X})$ is noticeably richer. The distribution functions of these two p values are given in Figure 3 for various values of the parameters. For all but very large values of the parameters, p_{ppost} seems considerably more uniform than p_{fet}.

Further investigations revealed that when both θ and ξ are small, either p value is quite conservative, with p_{ppost} the less conservative. Both p values perform at their best for $\theta, \xi \approx .5$, with p_{ppost} performing much better. When one of θ or ξ is small and the other is large, p_{fet} is again very conservative, whereas p_{ppost} performs remarkably well.

Both p_{fet} and p_{ppost} are probably asymptotic frequentist p values, and it is of interest to ask how large n must be for their distributions to be approximately uniform. This is especially interesting for smaller values of p, which are typically of most interest when n is large. An illustrative such feature is the sample size needed for the distribution function of a p value at .05 to be within 20% of .05 (the value under the desired uniformity). When $(\theta, \xi) = (.6, .5)$, this obtains for p_{fet} only when $n \approx 500$; in contrast, this occurs for p_{ppost} when $n \approx 10$. When $(\theta, \xi) = (.3, .9), p_{\text{fet}}$ requires $n \approx 1200$, whereas p_{ppost} needs only $n \approx 110$.

The apparent breakdown of both p_{fet} and p_{ppost} for large values of (θ, ξ) [such as (.9, .9) in Fig. 3] deserves special discussion. First, note that p_{fet} becomes almost hopelessly conservative, never stating that the data is incompatible with the model. In contrast, p_{ppost} is markedly anticonservative for this situation. At a very intuitive level, the behavior of p_{ppost} seems more sensible. After all, we declared large values of T to be evidence against the null model, and when (θ, ξ) are both large, the values of $T = X_{11}$ clearly will typically be very large; p_{ppost} reacts to this with ready "rejection" of the null model, whereas p_{fet} ignores all but incredibly large T. This anticonservative behavior of p_{ppost} arises because a very large value of $T = X_{11}$ contains a great deal of information about the parameters, but relatively little information about deviance from the model. This is one negative consequence of using an inferior choice of T.

The most extreme example of an inappropriate choice of T is a sufficient statistic for the parameter; such a statistic is nearly useless for model checking. We examine this further in a very simple example, so as to better understand the nature of p_{ppost} in such a situation.

Example 5. Assume that the null model is $X_i \sim \text{Bernoulli}(\theta), i = 1, \ldots, n$, and that $T = \sum X_i$, a sufficient statistic. Here $m(t|u) = m(t|\mathbf{x}_{\text{obs}} \setminus t_{\text{obs}}) = m(t) = 1/(n+1)$ for $t = 0, \ldots, n$, and $p_{\text{ppost}} = 1 - t_{\text{obs}}/(n+1)$. For large n, the distribution of p_{ppost} is approximately $\text{N}(1 - \theta, \theta(1-\theta)/n)$, which concentrates tightly around $1 - \theta$. Thus when θ is large, the distribution function of p_{ppost} jumps immediately, giving rise to anticonservative p values; in contrast, for small θ, the situation reverses, and p_{ppost} is conservative. Figure 4 shows the resulting distribution functions for three values of θ when $n = 100$.

Of course, this behavior of p_{ppost} is entirely natural according to Bayesian intuition; large values of T are essentially equivalent to large values of θ, and as such are declared to be "surprising." As another argument, note that p_{ppost} is equivalent to p_{prior} here, and choosing T to be sufficient is effectively stating that we will also allow its presence in the tail of the prior to discredit the model.

In contrast, the distributions of both p_{plug} and p_{post} can be seen to concentrate tightly about $1/2$ when n is large, for any value of θ. (That p_{post}, when it differs from uniformity, does so by concentrating closer to $1/2$ was discussed in Meng 1994; see also Rubin 1996.) This is illustrated in Figure 4 for p_{plug} when $n = 100$. Thus, unlike p_{ppost} which provides some kind of information, p_{plug} and p_{post} provide completely useless answers here. (Of course, non-Bayesians may argue that it is better to infer nothing than to in effect base a conclusion on the prior; but recall that in the context we are considering, this means essentially refusing to consider alternatives to the null model, at least when T is chosen poorly.)

As a final comment concerning this issue, recall that requiring uniformity of p values for all values of θ might

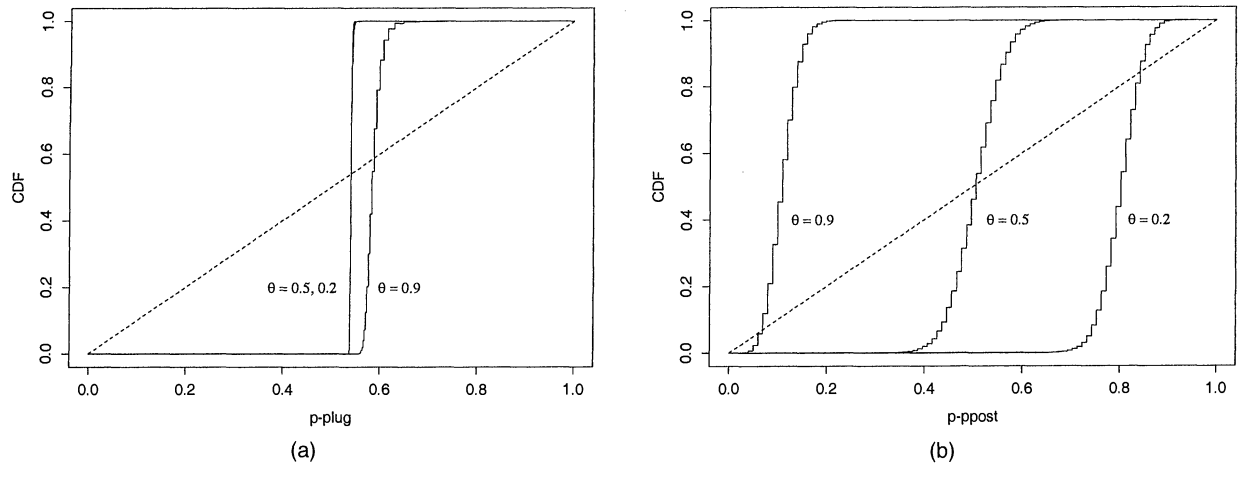

Figure 4. Distributions of $p_{plug}(\mathbf{X})$ (a) and $p_{ppost}(\mathbf{X})$ (b) in Example 5 for θ = .2, .5, .9 and n = 100.

well be too restrictive for a Bayesian (and also possibly for a frequentist). The natural (Bayesian) requirement (see Meng 1994) is to require a p value to be uniform under the prior predictive distribution. Because in this example, the partial posterior predictive reduces to the prior predictive, it follows that p_{ppost} is indeed a p value for a Bayesian. (The "average" of all the distribution functions of p_{ppost} in Fig. 4 is uniform.) In contrast, no Bayesian averages of the distribution functions of p_{plug} (or p_{post}) can be uniform. Of course, the Bayesian reasoning in this example is facilitated by the fact that the noninformative prior is actually proper, but related arguments involving averages with respect to classes of priors probably can be made for the improper case; we do not pursue this here.

When considering the choice of T in discrete situations, this issue of sufficiency can arise in subtle ways. In Example 3, for instance, consider $T = [(1/n_1)X_{11} - (1/n_2)X_{22}]$ when n_1 and n_2 are different prime numbers. From a given nonzero value of T, it is then clear that X_{11} and X_{22} can be completely reconstructed. In this situation, T thus is nearly sufficient, and its use would encounter the aforementioned difficulties. (In contrast, when $n_1 = n_2$, this problem with T does not arise.) Such "technical" near-sufficiency of T in discrete settings can be eliminated by the simple device of replacing T by a binned version, T^*, with the bin size chosen so that each value of T^* corresponds to several sample points.

5. CONCLUSIONS

Our comparisons have not included p_{prior}, because this cannot typically be used with noninformative priors. Also, p_{sim} is just a version of a conditional predictive p value, obtained by choosing the conditioning statistic U to be a sufficient statistic (when available). Indeed, in all of our continuous examples it happened that p_{sim} was equal to p_{cpred}, although this certainly is not true in general. For those wishing to use p_{sim}, this equality is a fortunate occurrence when it obtains (see Theorem 3), because p_{cpred} is typically much easier to compute directly in those situations than p_{sim}. The following discussion is thus limited to the other four p values.

A surprising observation in our examples (first discussed in Robins 1999) is that p_{plug} seems superior to p_{post}, in the sense that it is closer to being a frequentist p value; in particular, it is less conservative. This would seem to contradict the common Bayesian intuition that it is better to account for parameter uncertainty by using a posterior than by simply replacing θ by $\hat{\theta}$. The explanation is that p_{post} does not account for parameter uncertainty in a legitimate Bayesian way, because it involves a double use of the data. (Indeed, the original motivation for p_{cpred} and p_{ppost} was precisely to account for parameter uncertainty in a legitimate Bayesian fashion.) We have considered only a few situations here, but, together with the similar asymptotic conclusion of RVV (2000), it would seem that p_{plug} should be preferred to p_{post} in practice. This is especially so because p_{plug} is typically easier to compute than p_{post}. (In some situations in which Bayesian analysis is being performed via MCMC, a posterior sample of θ's might be more readily available than an MLE, but one could then plug in, say, a posterior mean for θ rather than the MLE.) At the very least, our observations here and those of RVV (2000) indicate that it cannot simply be assumed that p_{post} is better than p_{plug}, as has typically been the case in the literature. It should be noted, however, that posterior predictive p values are also commonly used today with discrepancy statistics that depend on θ, as well as on $\mathbf{x}$, and there are currently no alternatives to their use in such situations (although see RVV 2000).

In all our continuous examples, p_{plug} performed worse in the frequentist sense than either p_{ppost} or p_{cpred}. This again supports the asymptotic conclusions of RVV (2000) and suggests that the latter p values, if available, are to be preferred in practice. Computation is clearly an issue, however, in that p_{plug} is typically easier to compute than the new p values, especially p_{cpred} (see also Pauler 1999). Computation of p_{ppost} is usually not difficult if $f(t;\theta)$ is available in closed form, and we would definitely recommend its use in that case.

The (asymptotic) superiority of p_{ppost} and p_{cpred} arises when the departure statistic T is not appropriately "centered," as discussed by RVV (2000). In a sense, the new p values can be viewed as automatically "centering" a departure statistic T, which can be a considerable simplification in practice, avoiding the need for asymptotics or clever statistical intuition. Indeed, in model checking one often wishes to try a series of rather generic possible discrepancy statistics T, and having an automatic centering mechanism is a considerable simplification.

On a more speculative note, it is quite plausible that use of p_{ppost} and p_{cpred} can result in an improvement (over, say, p_{plug}) with even "centered" choices of T (as long as the distribution of T still depends on θ to some extent). This could be improvement in finite-sample performance or in higher-order asymptotic terms.

The situation involving discrete distributions is more complex, but the gains through use of the new p values, especially p_{ppost} can be quite dramatic. Discreteness of the sample space can cause common p values, such as those from the Fisher exact test, to be very conservative in small samples, whereas the partial posterior p value is rather remarkably uniform, especially if a reasonable discrepancy statistic T is used.

APPENDIX: TECHNICAL AND COMPUTATIONAL DETAILS

Details for Theorem 1

Suppose that $\pi(\theta)$ is improper but that there exists a sequence of increasing compact sets $\Theta_k \subset \Theta$ such that $\cup_{k\geq 1}\Theta_k = \Theta, 0 < m_k = \int_{\Theta_k} \pi(\theta)\,d\theta < \infty, 0 < m(u) = \int_{\Theta} f(u;\theta)\pi(\theta)\,d\theta < \infty$, and

$$\lim_{k\to\infty} m_k \int \frac{(m_k(u))^2}{m(u)}\,du = 1, \tag{A.1}$$

where $m_k(u) = (\int_{\Theta_k} f(u;\theta)\pi(\theta)\,d\theta)/m_k$. Then the conclusion of Theorem 1 holds.

Proof. Define $h(u,\theta) = \Pr(p(\mathbf{X}) \le \alpha | u;\theta)$. From the definition of $p_{\text{cpred}(u)}$, it follows that

$$\int h(u,\theta)\pi(\theta|u)\,d\theta = \alpha. \tag{A.2}$$

By the assumption that $p(\mathbf{X})$ has a distribution that does not depend on θ, it follows that for some constant c,

$$\int h(u,\theta)f(u;\theta)\,du = E[\Pr(p(\mathbf{X}) \le \alpha);\theta] = c. \tag{A.3}$$

It is immediate from (A.2) and (A.3) that

$$\int h(u,\theta)\pi(\theta|u)m_k(u)\,du\,d\theta = \alpha \tag{A.4}$$

and

$$\frac{1}{m_k}\int h(u;\theta)f(u;\theta)\pi(\theta)1_{\Theta_k}\,d\theta\,du = c. \tag{A.5}$$

To prove that $\alpha = c$, completing the proof, we need only show that the difference of the left sides of (A.4) and (A.5) goes to 0 as $k \to \infty$. Breaking the left side of (A.4) into integrals over Θ_k^C and Θ_k, and using the fact that

$$m_k(u) = \frac{1}{m_k}\left[m(u) - \int_{\Theta_k^C} f(u;\theta^*)\pi(\theta^*)\,d\theta^*\right]$$

in the second of these integrals, the difference of the left sides of (A.4) and (A.5) can be written

$$\int_{\Theta_k^C} h(u,\theta)\pi(\theta|u)m_k(u)\,d\theta\,du - \frac{1}{m_k}\int_{\Theta_k} h(u,\theta)\pi(\theta|u) \times \left[\int_{\Theta_k^C} f(u;\theta^*)\pi(\theta^*)\,d\theta^*\right] d\theta\,du.$$

Because $h(u,\theta) \le 1$, algebra shows that each of these terms is bounded in absolute value by

$$\begin{aligned}\int_{\Theta_k^C} \pi(\theta|u)m_k(u)\,d\theta\,du &= 1 - \int_{\Theta_k} \pi(\theta|u)m_k(u)\,d\theta\,du \\ &= 1 - m_k\int \frac{(m_k(u))^2}{m(u)}du,\end{aligned}$$

which goes to 0 by (A.1) and completes the proof.

Verification of (A.1) When U has a Location or Scale Distribution. For convenience, we assume that U has a location distribution with range $\Re$ and that $\Theta = \Re$. Other cases can be handled similarly. Write $f(u;\theta) = g(u-\theta)$, let $G(\cdot)$ denote the cdf corresponding to $g(\cdot)$, and choose $\Theta_k = (-k,k)$. Then $m(u) = \int g(u-\theta)\,d\theta = 1, m_k = \int_{-k}^{k}(1)\,d\theta = 2k, m_k(u) = (1/2k)\int_{-k}^{k} g(u-\theta)\,d\theta = (1/2k)[G(u+k) - G(u-k)]$, and (A.1) becomes

$$\lim_{k\to\infty} 2k\int (m_k(u))^2\,du = 1. \tag{A.6}$$

Note first that $2km_k(u) = [G(u+k) - G(u-k)] \le 1$, so that (A.6) is trivially bounded above by 1. To establish a suitable lower bound, note that $[G(\log k) - G(-\log k)] > (1-\varepsilon)$ for any given $\varepsilon > 0$ and sufficiently large k, so that

$$2k\int (m_k(u))^2\,du \ge \frac{1}{2k}\int_{-k+\log k}^{k-\log k}[G(u+k) - G(u-k)]^2\,du > (1-\varepsilon)^2\frac{2(k-\log k)}{2k}.$$

Because ε was arbitrary, (A.6) is clearly satisfied.

Verification of (19)

Lemma A.1. Let $\mathbf{W} = (W_1, W_2, \ldots, W_n)$ be a random vector with uniform distribution on the simplex $\sum_{i=1}^n W_i = 1$, and let $W_{(1)} = \min\{W_i\}$. Then $\Pr(W_{(1)} \le c) = 1 - (1-nc)^{n-1}$.

Proof. We give a geometric argument. The probability to be computed is

$$\Pr(W_{(1)} \le c) = 1 - \Pr(\text{all } W_i > c) = 1 - q. \tag{A.7}$$

Note that the conditional distribution of $\mathbf{W}$ on the set $\{\mathbf{W} : W_i > c \text{ for all } i\}$ is also uniform, and that this set is itself a simplex of the same shape as the original simplex but with "corners" $(c,\ldots,c,1-(n-1)c), (c,\ldots,1-(n-1)c,c),\ldots(1-(n-1)c,\ldots,c,c)$. The edges of the original simplex have length $\sqrt{2}$, whereas those of the smaller simplex have length $\sqrt{2}(1-nc)$. It follows that q in (A.7) is given by

$$\left(\frac{\sqrt{2}(1-nc)}{\sqrt{2}}\right)^{n-1} = (1-nc)^{n-1},$$

and the lemma follows.

To establish (19), note that

$$\Pr\left(\frac{T}{S} \le c\right) = E^{f(s;\lambda)}\Pr^{f(t|s;\lambda)}\left(\frac{T}{S} \le c\right).$$

But given s, the distribution of $X_1, X_2, \ldots, X_n$ is uniform on the set $\{\mathbf{X} : \sum_{i=1}^n X_i = s\}$. Defining $W_i = X_i/s$, the conditions of Lemma A.1 clearly apply, with $W_{(1)} = T/s$, and the result follows.

Computation of p_{ppost} for Independence in Contingency Tables

From (33) and (34), it is clear that a Monte Carlo importance sampling approximation to p_{ppost} in (33) is given by

$$p_{\text{ppost}} \approx \frac{\sum_{i=1}^L [1 - \mathcal{B}(t_{\text{obs}} - 1; n, \theta_i\xi_i)] \times \pi(\theta_i,\xi_i|\mathbf{x}_{\text{obs}} \setminus t_{\text{obs}})/h(\theta_i,\xi_i)}{\sum_{i=1}^L \pi(\theta_i,\xi_i|\mathbf{x}_{\text{obs}} \setminus t_{\text{obs}})/h(\theta_i,\xi_i)},$$

where $\mathcal{B}(x;n,\varphi)$ is the distribution function at x of the $\text{Bi}(n,\varphi)$ distribution, $h(\theta,\xi)$ is some importance function, and $(\theta_1,\xi_1),(\theta_2,\xi_2),\ldots,(\theta_L,\xi_L)$ are L random draws from $h(\cdot)$.

Importance functions that have a bounded importance ratio, $\pi(\theta_i,\xi_i|\mathbf{x}_{\text{obs}} \setminus t_{\text{obs}})/h(\theta_i,\xi_i)$, and that reasonably approximate the desired distribution are useful for several reasons. First, convergence is typically rapid. Second, an explicit formula for the Monte Carlo variance is then available. Third, the scheme can be readily adapted, via acceptance-rejection, to generate an actual sample from the partial posterior, if desired. The importance function in (35) can be seen to have these properties. In particular, the importance ratio can be computed to be

$$\begin{aligned}&\frac{\pi(\theta,\xi|\mathbf{x}_{\text{obs}} \setminus t_{\text{obs}})}{h(\theta,\xi)} \\ &= \left\{\frac{c_1}{2}\,\frac{(1-\theta\xi)^{n-x_{11}^o}}{\theta^{x_{21}^o}(1-\theta)^{n-x_{11}^o-x_{21}^o}(1-\xi)^{x_{21}^o}} + \frac{c_2}{2}\,\frac{(1-\theta\xi)^{n-x_{11}^o}}{(1-\theta)^{x_{12}^o}\xi^{x_{12}^o}(1-\xi)^{n-x_{11}^o-x_{12}^o}}\right\}^{-1},\end{aligned}$$

where

$$c_1 = \frac{\Gamma(n - x_{11}^o - x_{21}^o + 2)}{\Gamma(x_{12}^o+1)\Gamma(x_{22}^o+1)}$$

and

$$c_2 = \frac{\Gamma(n - x_{11}^o - x_{12}^o + 2)}{\Gamma(x_{21}^o + 1)\Gamma(x_{22}^o + 1)}.$$

It is straightforward to show that this is bounded by $2/(c_1 + c_2)$, using the inequalities $\theta(1-\xi) \leq (1-\theta\xi)$ and $(1-\theta) \leq (1-\theta\xi)$ judiciously. This importance function works well even for very large values of n and extreme values of θ and ξ.

[Received December 1998. Revised November 1999.]

REFERENCES

Agresti, A. (1992), "A Survey of Exact Inference for Contingency Tables" (with discussion), *Statistical Science*, 7, 131–177.

Aitkin, M. (1991), "Posterior Bayes Factors" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 53, 111–142.

Bayarri, M. J., and Berger, J. O. (1997), "Measures of Surprise in Bayesian Analysis," ISDS Discussion Paper 97-46, Duke University.

——— (1999), "Quantifying Surprise in the Data and Model Verification," in *Bayesian Statistics 6*, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, London: Oxford University Press, pp. 53–82.

Berger, J. O., Boukai, B., and Wang, W. (1997), "Unified Frequentist and Bayesian Testing of Precise Hypotheses," *Statistical Science*, 12, 133–160.

Berger, J. O., Brown, L. D., and Wolpert, R. L. (1994), "A Unified Conditional Frequentist and Bayesian Test for Fixed and Sequential Simple Hypothesis Testing," *The Annals of Statistics*, 22, 1787–1807.

Berger, J. O., and Delampady, M. (1987), "Testing Precise Hypotheses" (with discussion), *Statistical Science*, 2, 317–352.

Berger, J. O., and Sellke, T. (1987), "Testing a Point Null Hypothesis: The Irreconciability of p Values and Evidence," *Journal of the American Statistical Association*, 82, 112–122.

Berger, R. L., and Boos, D. D. (1994), "P Values Maximized Over a Confidence Set for the Nuisance Parameter," *Journal of the American Statistical Association*, 89, 1012–1016.

Blyth, C. R., and Staudte, R. G. (1995), "Estimating Statistical Hypotheses," *Probability and Statistics Letters*, 23, 45–52.

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

Carlin, B. P. (1999), Discussion of "Quantifying Surprise in the Data and Model Verification," by M. J. Bayarri and J. O. Berger, in *Bayesian Statistics 6*, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, London: Oxford University Press, pp. 73–74.

De la Horra, J., and Rodríguez-Bernal, M. T. (1997), "Asymptotic Behaviour of the Posterior Predictive P-Value," *Communications in Statistics, Part A—Theory and Methods*, 26, 2689–2699.

Delampady, M., and Berger, J. O. (1990), "Lower Bounds on Bayes Factors for Multinomial Distributions, With Applications to Chi-Squared Tests of Fit," *The Annals of Statistics*, 18, 1295–1316.

Edwards, W., Lindman, H., and Savage, L. J. (1963), "Bayesian Statistical Inference for Psychological Research," *Psychological Review*, 70, 193–242.

Evans, M. (1997), "Bayesian Inference Procedures Derived via the Concept of Relative Surprise," *Communications in Statistics, Part A—Theory and Methods*, 26, 1125–1143.

Gelfand, A. E., Dey, D. K., and Chang, H. (1992), "Model Determination Using Predictive Distributions With Implementation via Sampling-Based Methods," in *Bayesian Statistics 4*, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, London: Oxford University Press, pp. 147–167.

Gelman, A., Carlin, J. B., Stern, H., and Rubin, D. B. (1995), *Bayesian Data Analysis*, London: Chapman and Hall.

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.

Guttman, I. (1967), "The Use of the Concept of a Future Observation in Goodness-of-Fit Problems," *Journal of the Royal Statistical Society*, Ser. B, 29, 83–100.

Hwang, J. T., Casella, G., Robert, C., Wells, M., and Farrell, R. (1992), "Estimation of Accuracy of Testing," *The Annals of Statistics*, 20, 490–509.

Hwang, J. T., and Pemantle, R. (1997), "Estimating the Truth Indicator Function of a Statistical Hypothesis Under a Class of Proper Loss Functions," *Statistics and Decisions*, 15, 103–128.

Hwang, J. T., and Yang, M.-C. (1997), "Evaluate the P-Values for Testing the Independence in 2×2 Contingency Tables Using the Estimated Truth Approach—One Way to Resolve the Controversy Relating to Fisher's Exact Test," technical report, Cornell University, Dept. of Statistical Science.

Meng, X. L. (1994), "Posterior Predictive P-Values," *The Annals of Statistics*, 22, 1142–1160.

Pauler, D. K. (1999), Discussion of "Quantifying Surprise in the Data and Model Verification," by M. J. Bayarri and J. O. Berger, in *Bayesian Statistics 6*, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, London: Oxford University Press, pp. 70–72.

Robins, J. M. (1999), Discussion of "Quantifying Surprise in the Data and Model Verification," by M. J. Bayarri and J. O. Berger, in *Bayesian Statistics 6*, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, London: Oxford University Press, pp. 67–70.

Robins, J. M., van der Vaart, A., and Ventura, V. (2000), "The Asymptotic Distribution of p Values in Composite Null Models," *Journal of the American Statistical Association*, 95, 1143–1156.

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

——— (1996), Discussion of "Posterior Predictive Assessment of Model Fitness via Realized Discrepancies," by A. Gelman, X. Meng, and H. S. Stern, *Statistica Sinica*, 6, 787–792.

Schaafsma, W., Tolboom, J., and Van Der Meulen, E. A. (1989), "Discussing Truth or Falsity by Computing a Q-Value," in *Statistical Data Analysis and Inference*, eds. Y. Dodge, Amsterdam: North-Holland, pp. 85–100.

Sellke, T., Bayarri, M. J., and Berger, J. O. (1999), "Calibration of P Values for Precise Null Hypotheses," ISDS Discussion Paper 99-13, Duke University.

Suissa, S., and Shuster, J. J. (1985), "Exact Unconditional Sample Sizes for the 2×2 Binomial Trial," *Journal of the Royal Statistical Society*, Ser. A, 148, 317–327.

Thompson, P. (1997), "Bayes P-Values," *Statistics and Probability Letters*, 31, 267–271.

Tsui, K.-W., and Weerahandi, S. (1989), "Generalized P Values in Significance Testing of Hypothesis in the Presence of Nuisance Parameters," *Journal of the American Statistical Association*, 84, 602–607.