P-values for Composite Null Models

by

M.J. Bayarri and James O. Berger

University of Valencia and Duke University

Duke University

Institute of Statistics and Decision Sciences
Durham, North Carolina 27708, USA

Abstract

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 and phrases. 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 assume $f(\mathbf{x}; \theta)$ 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, $\mathbf{x}_{obs}$, with large values of T indicating less compatibility. The most commonly used measure of compatibility is the p-value, defined as

$$p = \Pr(t(\mathbf{X}) \geq t(\mathbf{x}_{obs})). \quad (1.1)$$

When θ is known, the probability computation in (1.1) is with respect to $f(\mathbf{x};\theta)$. The focus in this article is on the choice of the probability distribution used to compute (1.1) when θ is unknown. In Section 2 we present two new types of p-values, which we subsequently argue to be superior to existing choices. The rest of this section describes the most common of the existing choices. We will be abusing 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 $u(\mathbf{X}) = u$, respectively.

The most obvious way to deal with an unknown θ in the computation of the p-value is to replace θ in (1.1) by some estimate, $\hat{\theta}$; in this paper we will consider only the usual choice for $\hat{\theta}$, namely the m.l.e. The resulting p-value will be called the *plug-in p-value* (p_{plug}). Using a superscript to denote the density with respect to which the p-value in (1.1) is computed, the *plug-in p-value* is thus defined as:

$$p_{plug} = \Pr^{f(\cdot;\hat{\theta})}(t(\mathbf{X}) \geq t(\mathbf{x}_{obs})). \quad (1.2)$$

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

Another natural device for eliminating the unknown θ is to condition on a sufficient statistic, U, for θ. Then, $f(\mathbf{x}|u_{obs};\theta)$ does not depend on θ and computations in (1.1) can be carried out using the completely specified $f(\mathbf{x}|u_{obs})$. (In fact, U need only be sufficient for θ with respect to $f(t;\theta)$). We shall call these p-values *similar p-values*, the term borrowed from the related notion of similar tests and confidence regions. A *similar p-value* is thus defined as

$$p_{sim} = \Pr^{f(\cdot|u_{obs})}(t(\mathbf{X}) \geq t(\mathbf{x}_{obs})). \quad (1.3)$$

The main strength of p_{sim} is that it is based on a proper probability computation, which will imbue the end result with various desirable properties to be discussed later. Its main weaknesses are: *(i)* computation can be burdensome; *(ii)* a suitable sufficient U often does not exist; and *(iii)* after conditioning upon U, the choice of the compatibility statistic T may be forced on the user, and this may not be a desirable choice.

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

$$m(\mathbf{x}) = \int f(\mathbf{x};\theta)\pi(\theta)d\theta. \quad (1.4)$$

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

$$p_{prior} = \Pr^{m(\cdot)}(t(\mathbf{X}) \geq t(\mathbf{x}_{obs})). \tag{1.5}$$

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 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 non-subjective analysis might be desired from the beginning), it is attractive to attempt to utilize noninformative priors. Unfortunately, noninformative priors are typically improper, in which case the prior predictive $m(\mathbf{x})$ would also be improper, precluding computation of (1.5). Box (1980) popularized use of p_{prior}.

The concerns mentioned in the last 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, $\pi(\theta|\mathbf{x}_{obs})$, instead of the prior, before computing a p-value. The *posterior predictive p-value* is thus defined as

$$p_{post} = \Pr^{m_{post}(\cdot|\mathbf{x}_{obs})}(t(\mathbf{X}) \geq t(\mathbf{x}_{obs})), \tag{1.6}$$

where

$$m_{post}(\mathbf{x}|\mathbf{x}_{obs}) = \int f(\mathbf{x};\theta)\pi(\theta|\mathbf{x}_{obs})d\theta. \tag{1.7}$$

The main strengths of p_{post} are: *(i)* improper noninformative priors can readily be used (since $\pi(\theta|\mathbf{x}_{obs})$ will typically be proper); *(ii)* $m_{post}(\mathbf{x}|\mathbf{x}_{obs})$ will typically 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}; and *(iii)* it is typically very easy to compute using output from modern MCMC Bayesian analyses. Its main weakness is that there is an apparent "double use" of the data in (1.6), first to convert the (possibly improper) prior $\pi(\theta)$ into a proper distribution $\pi(\theta|\mathbf{x}_{obs})$ for determining the reference distribution $m_{post}(\mathbf{x}|\mathbf{x}_{obs})$, and then for computing the tail area corresponding to $t(\mathbf{x}_{obs})$. We will see examples of the unnatural behavior that this double use of the data can induce. From a Bayesian perspective, defenders of the prior predictive also point out that the posterior predictive lacks a pure Bayesian interpretation; while this was our original motivation for the developments herein, the arguments in the paper are not directly based on such reasoning.

Generalizations of (1.6) were considered in 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 (1.7) could be replaced by $f(\mathbf{x}|\theta, A)$, where A is some other statistic. We will not discuss such generalizations in this paper.

There are also many other related works. Aitkin (1991) used the posterior distribution to compute actual Bayes factors, instead of p-values. Other approaches that have been suggested for dealing with the nuisance parameter, θ, in computing (1.1) include Tsui and Weerahandi (1989) and Berger and Boos (1994).

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 function of the data, 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 (1999), which we henceforth denote RVV (1999); earlier articles that refer to and/or discuss this 'defining' property of a p-value include Meng(1994), Rubin (1996), De la Horra and Rodríguez-Bernal (1997), Thompson (1997), and Robins (1999).

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 the companion RVV (1999). 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 it is $U[0,1]$ in an average sense over θ. If the prior distribution is chosen subjectively, 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 briefly discussed 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 utilized. (We will later see 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 anti-conservative in a frequentist sense (see RVV 1999 for definitions), then it is likewise guaranteed to be conservative or anti-conservative 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 anti-conservative in a frequentist asymptotic sense. Interesting related discussion concerning the posterior predictive p-value can be found in Meng (1994), Gelman, Meng and Stern (1996), and Rubin (1996).

Actually, Bayesians might well go further, requiring not just unconditional uniformity for p-values, but also seeking reasonable conditional performance. In this paper 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., Schaafsma, Tolboom and Van Der Meulen 1989, Blyth and Staudte 1995, Hwang, Casella, Robert, Wells and Farrell 1992, Hwang and Pemantle 1997, Hwang and Yang 1997, and Thompson 1997). Neither of these evaluation techniques is within the scope of this paper because we are specifically concerned with the situation in which no alternative is present (see the next section). From a non-Bayesian perspective, however, evalution of the new p-values by these criteria might well prove very illuminating; see the companion paper RVV(1999) for some results in this direction.

1.3. To Be and Not To Be

This paper has five sections. In Section 2, we consider two new p-values introduced in 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 3) which allows ready computation of p_{sim} in certain situations. In Section 3, the various p-values are compared in the normal linear model, where exact computation is possible; this section was directly motivated by RVV (1999). Section 4 discusses 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. Conclusions are presented in Section 5.

There are a number of relevant issues that will not be considered in this article. First, we do not explicitly discuss the choice of T, feeling that it will be chosen on a casual and intuitive basis. Our perspective here is that of model checking, rather than formal hypothesis testing. If one has a clearly formulated alternative to H_0, we would not recommend use of *any* p-value to perform the test, and would instead use either Bayes factors or conditional frequentist tests (cf. Edwards, Lindman and Savage 1963, Berger and Sellke 1987, Berger and Delampady 1987, Delampady and Berger 1990, Berger, Brown and Wolpert 1994, Berger, Boukai and Wang 1997, Bayarri and Berger 1999). 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 were to require hard work, we would suggest spending the time, instead, on actual formulation of the alternative.

The other major issue we will mostly avoid is discussion of measures, other than p-values, of compatibility of the data 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 $B(p) = -e\,p\log(p)$, 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 $\alpha(p) = (1 + [-e\,p\log(p)]^{-1})^{-1}$. As an example, $p = 0.05$ translates into odds $B(0.05) = 0.41$ (roughly 1 to 2.5) of H_0 to H_1, and frequentist error probability $\alpha = 0.29$ in rejecting H_0. (Strictly speaking, $B(p)$ is a lower bound on the odds, and $\alpha(p)$ is a conditional frequentist error probability.) These calibrations are developed and motivated from a variety of perspectives in Sellke, Bayarri, and Berger (1999). It is of interest that the calibrations are, in part, themselves based on having a proper p-value, i.e., 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 can be found in Bayarri and Berger (1999). From now on we will 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

Subsection 2.1 introduces the two new p-values that will be considered 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. Subsection 2.2 presents the motivations for the new p-values, from both Bayesian and frequentist perspectives. Computational issues are addressed in Subsection 2.3.

2.1. Methodology

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

$$p_{ppost} = \Pr{}^{m(\cdot\,|\mathbf{x}_{obs}\backslash t_{obs})}(T \geq t_{obs}) \ ; \tag{2.1}$$

here $m(\cdot\,|\mathbf{x}_{obs}\backslash t_{obs})$ and the (assumed proper) *partial posterior* $\pi(\cdot\,|\mathbf{x}_{obs}\backslash t_{obs})$ are given by

$$\begin{aligned} m(t\,|\mathbf{x}_{obs}\backslash t_{obs}) &= \int f(t|\theta)\pi(\theta|\mathbf{x}_{obs}\backslash t_{obs})\, d\theta, \\ \pi(\theta\,|\mathbf{x}_{obs}\backslash t_{obs}) &\propto f(\mathbf{x}_{obs}|t_{obs};\theta)\pi(\theta) \propto \frac{f(\mathbf{x}_{obs};\theta)\pi(\theta)}{f(t_{obs};\theta)}. \end{aligned} \tag{2.2}$$

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 $\mathbf{x}_{obs}\backslash t_{obs}$ was chosen to indicate this.)

The second p-value 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

$$p_{cpred(u)} = \Pr{}^{m(\cdot|u_{obs})}(T \geq t_{obs}); \tag{2.3}$$

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

$$m(t|u) = \int f(t|u;\theta)\pi(\theta|u)d\theta, \tag{2.4}$$

assuming that

$$\pi(\theta|u) = \frac{f(u;\theta)\pi(\theta)}{\int f(u;\theta)\pi(\theta)d\theta} \tag{2.5}$$

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 (2.3) to be the conditional m.l.e. of θ, given $t(\mathbf{x}) = t$, defined as

$$\hat{\theta}_{cMLE}(\mathbf{x}) = \arg\max f(\mathbf{x}|t,\theta) = \arg\max \frac{f(\mathbf{x};\theta)}{f(t;\theta)} \quad . \tag{2.6}$$

We will suppress $\hat{\theta}_{cMLE}$ and call the resulting p-value simply the *conditional predictive p-value*, and denote it by $p_{cpred} = p_{cpred(\hat{\theta}_{cMLE})}$. Note that $m(t|u)$ is unaffected by one-to-one transformations of $u(\mathbf{x})$, so that any one-to-one transformation of (2.6) is satisfactory as the choice of $\hat{\theta}_{cMLE}$.

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

Example 2.1 Assume that, under the null, the X_i are iid $N(0,\sigma^2)$, with σ^2 unknown. The statistic $t(\mathbf{X}) = |\overline{X}|$ 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

$$p = \Pr\{|\overline{X}| > |\overline{x}_{obs}|\}, \tag{2.7}$$

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

$\mathbf{p_{plug}}$ **:** Since $\overline{X} \sim N(0, \sigma^2/n)$ and $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n x_i^2 = s^2 + \overline{x}^2$ is the m.l.e., it follows from (1.2) and (2.7) that

$$p_{plug} = 2\left[1 - \Phi\left(\frac{\sqrt{n}\,|\overline{x}_{obs}|}{\sqrt{s_{obs}^2 + \overline{x}_{obs}^2}}\right)\right] \quad . \tag{2.8}$$

One obvious inadequacy with this p-value is that $p_{plug} \longrightarrow 2[1 - \Phi(\sqrt{n})]$, a positive constant, as $|\overline{x}_{obs}|/s_{obs} \longrightarrow \infty$. Thus, even with arbitrarily strong evidence against the null model, the p-value will not go to zero (for fixed n).

$\mathbf{p_{sim}}$ **:** A sufficient statistic for σ^2 is $V = \sum_{i=1}^n X_i^2 = ||\mathbf{X}||^2$. The distribution of $\mathbf{X}$, given $v_{obs} = ||\mathbf{x}_{obs}||^2$, is uniform on $\{\mathbf{x} : \; ||\mathbf{x}||^2 = ||\mathbf{x}_{obs}||^2\}$, so that (1.3) and (2.7) yield

$$p_{sim} = Pr\left(\frac{|\overline{X}|}{||\mathbf{x}_{obs}||} > \frac{|\overline{x}_{obs}|}{||\mathbf{x}_{obs}||}\right) = Pr\left(|\overline{Z}| > \frac{|\overline{x}_{obs}|}{||\mathbf{x}_{obs}||}\right) \quad , \tag{2.9}$$

where $\mathbf{Z}$ has a uniform distribution on $\{\mathbf{z} : \; ||\mathbf{z}||^2 = 1\}$. Although this might appear to be difficult to compute, it will be seen 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 (2.11); we found this result surprising.

$\mathbf{p_{prior}}$ **:** The prior predictive p-value cannot be computed for this example, since the prior distribution is improper.

$\mathbf{p_{post}}$ **:** The posterior density, $\pi(\sigma^2|\mathbf{x}_{obs})$, is $Ga^{-1}(n/2, n(s^2+\overline{x}^2)/2)$, and the posterior predictive distribution of $\overline{X}$ is $m_{post}(\overline{x}|\mathbf{x}_{obs}) = t_n(\overline{x} \mid 0, \frac{1}{n}(s^2_{obs} + \overline{x}^2_{obs}))$; here Ga^{-1} and t_n denote, respectively, the Inverse Gamma distribution and the t-distribution with n degrees of freedom. From (1.6) and (2.7), it follows that

$$p_{post} = 2\left[1 - \Upsilon_n\left(\frac{\sqrt{n}\;\overline{x}_{obs}}{\sqrt{s^2_{obs} + \overline{x}^2_{obs}}}\right)\right] \quad , \tag{2.10}$$

where Υ_n stands for the distribution function of the t-distribution with n degrees of freedom. As could be expected, (2.10) is very similar to p_{plug}, given in (2.8). Indeed, it has the similar inappropriate behavior that $p_{post} \longrightarrow 2[1 - \Upsilon_n(\sqrt{n})]$, a positive constant, as $|\overline{x}_{obs}|/s_{obs} \longrightarrow \infty$. For instance, when $n = 4$ this constant is 0.12, and the posterior predictive p-value never drops below this constant, no matter how many standard deviations $\overline{x}_{obs}$ is from zero. 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 $\overline{x}_{obs}$ is involved in computing both the posterior (or the m.l.e.) and the tail area. Interestingly, the problem with p_{plug} is less severe than with p_{post}, in that the limiting constant is smaller (0.046 when $n = 4$, for instance).

$\mathbf{p_{cpred}}$ **:** Computation shows that

$$f(\mathbf{x}|t;\sigma^2) \propto \frac{f(\mathbf{x};\sigma^2)}{f(t;\sigma^2)} \propto (\sigma^2)^{-\frac{n-1}{2}} \exp\{-\frac{ns^2}{2\sigma^2}\} \quad ,$$

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

$$p_{cpred} = 2\left[1 - \Upsilon_{n-1}\left(\frac{\sqrt{n-1}\;\overline{x}_{obs}}{s_{obs}}\right)\right] \quad . \tag{2.11}$$

This is perfectly satisfactory and, indeed, equals the usual classical p-value for the problem, which is known to be a true frequentist p-value.

$\mathbf{p_{ppost}}$ **:** In this case, $T = \overline{X}$ is independent of $\hat{\sigma}^2_{cMLE} \propto S^2$ (and they are clearly jointly sufficient) so that the partial posterior predictive p-value equals the conditional predictive p-value, p_{cpred}, in (2.11).

$\square$

It should be noted that Meng (1994) also considered use of the departure statistic $t(\mathbf{x}) = |\overline{x}|/s_{obs}$ in the above 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.

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, (2.3) 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, since only part of the data (u_{obs}) is used to produce the posterior to eliminate θ, while another part (t_{obs}) is used when computing the tail area.

Of course, the key to the U-conditional predictive p-value is suitable choice of the conditioning statistic U. Different possible choices of U are explored in Bayarri and Berger (1997). (See also Evans, 1997, where the conditional predictive distribution is used to develop alternate measures of surprise, with U and T being chosen to be separate subsamples of the data. A rather different possibility is 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 additional discussion.) The intuition behind suitable choice of U is that one wants U to contain as much information about θ as possible, so that $\pi(\theta|u_{obs})$ will effectively eliminate θ (via integration), subject to the constraint that U should not involve T, as this would entail a 'double use' of part of the data. In Example 2.1, 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 $t(\mathbf{x}) = |\overline{x}|$. The obvious solution (used in Example 2.1) is to define $u(\mathbf{x}) = s^2 = \sum(x_i - \overline{x})^2/n$, since this contains the information about σ^2 that is independent of $t(\mathbf{X})$.

Investigations in 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 m.l.e. of θ, given $t(\mathbf{x}) = t$, as given in

(2.6). (The situation of discrete data is considerably more difficult; while $\hat{\theta}_{cMLE}$ in (2.6) is still typically well defined, it will usually not be suitable as a conditioning statistic.)

While logically appealing, the conditional predictive p-value with the conditioning statistic $\hat{\theta}_{cMLE}$ chosen as in (2.6) can be difficult to compute. An attractive alternative I s to directly use $f(\mathbf{x}|t;\theta)$ (see (2.6)) to integrate out θ, rather than simply using it to define $\hat{\theta}_{cMLE}$. This leads to the *partial posterior predictive p-value*, defined in (2.1) and (2.2), which is typically much easier to work with. Furthermore, the parallel with (2.6) suggests that the partial predictive p-value will be very similar to the conditional predictive p-value. This will be seen in our (continuous) examples, and is further reinforced in RVV (1999), which shows that p_{cpred} and p_{ppost} are asymptotically equivalent.

The above Bayesian motivations may appear to be rather 'loose', but history in other areas of statistics has shown that, when sound Bayesian reasoning and noninformative priors are used to develop procedures, then 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 subsection.

2.2.2. Frequentist Motivations and Comparisons

RVV (1999) shows that p_{cpred} and p_{ppost} are asymptotic frequentist p-values; i.e., their asymptotic distribution is $U[0,1]$ for all θ. In this subsection we study if this is so for small samples. The subsection presents 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 $f(\mathbf{x};\theta)$. If the distribution of $p(\mathbf{X})$ does not depend on θ, 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 $\pi(\theta)$ 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, since $p(\mathbf{X})$ is, by definition, a proper p-value with respect to $m(t|u)$, it follows that

$$\mathrm{P}r^{m(\cdot)}(p(\mathbf{X}) \leq \alpha) = \mathrm{E}^{m(u)}\mathrm{E}^{m(t|u)}\,[\; p(\mathbf{X}) \leq \alpha \;] = \mathrm{E}^{m(u)}[\; \alpha \;] = \;\alpha \;\; .$$

But, since

$$\mathrm{P}r^{m(\cdot)}(p(\mathbf{X}) \leq \alpha) = \mathrm{E}^{\pi(\theta)}\left(\mathrm{E}^{f(\mathbf{x};\theta)}\,[\; p(\mathbf{X}) \leq \alpha \;]\right) \;,$$

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

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

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 2.1. From (2.11), it is clear that their distributions do not depend on σ^2, since the distribution of $\sqrt{(n-1)}\overline{X}/S$ is independent of σ^2. Also, $\hat{\sigma}^2_{cMLE}$ has a scale-parameter distribution, so it can be immediately concluded that p_{cpred} and p_{ppost} are frequentist p-values.

In Example 2.1, p_{plug} and p_{post} will not be frequentist p-values, but the extent to which they deviate from uniformity would have to be studied numerically. We thus turn to a simpler situation where exact computations can be performed.

Example 2.2 Assume that $X_1, X_2, \ldots, X_n$ is a random sample from the $Exponential(\lambda)$ distribution. Let $T = X_{(1)}$ (which could be used to investigate the lower tail of the null distribution) and assume the usual noninformative prior $\pi(\lambda) = 1/\lambda$ is to be employed. In the following, $X_{(1)} < X_{(2)} < \ldots < X_{(n)}$ 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, will be used repeatedly:

$$\Pr\left(\frac{T}{S} \leq c\right) = 1 - (1 - nc)^{n-1}\ . \tag{2.12}$$

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

$$p_{plug} = e^{-n^2 t_{obs}/s_{obs}}. \tag{2.13}$$

That this is conditionally unsatisfactory can be seen by taking $nt_{obs}/s_{obs} \to 1$, in which case the model would clearly be contraindicated, yet $p_{plug} \to e^{-n}$, a non-zero constant. To investigate whether p_{plug} is a frequentist p-value, an easy computation using (2.13) and (2.12) yields, for $\alpha > e^{-n}$,

$$\Pr(p_{plug}(\mathbf{X}) \leq \alpha) = \Pr\left(\frac{T}{S} \geq -\frac{\log \alpha}{n^2}\right) = \left(1 + \frac{\log \alpha}{n}\right)^{n-1}\ . \tag{2.14}$$

Thus $p_{plug}(\mathbf{X})$ does not have a $U[0,1]$ distribution, and is not a frequentist p-value. Figure

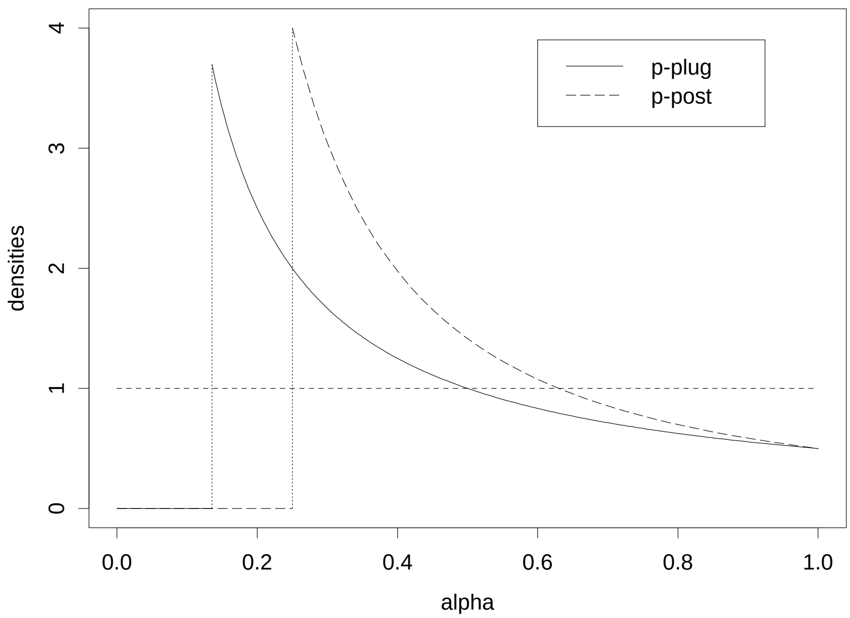

Figure 1: Densities of $p_{plug}(\mathbf{X})$ and $p_{post}(\mathbf{X})$ in Example 2.2, when $n = 2$. The Uniform density is plotted for reference.

1 graphs the *density* corresponding to (2.14) when $n = 2$, in order to show the substantial deviation from uniformity that can occur. Note, however, that $p_{plug}(\mathbf{X})$ is an asymptotic frequentist p-value. Indeed, for large n, (2.14) is approximately given by

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

which does go to α as $n \to \infty$.

$\mathbf{p_{sim}}$ **:** Since 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

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

this will be seen to be equal to p_{ppost}.

$\mathbf{p_{prior}}$ **:** The prior predictive p-value cannot be computed for this example, since the prior distribution is improper.

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

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

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

$$\Pr(p_{post}(\mathbf{X}) \leq \alpha) = \Pr\left[\frac{T}{S} \geq -\frac{1}{n}\ \left(\frac{1}{\alpha^{1/n}}\right)\right] = \left(2 - \alpha^{-1/n}\right)^{n-1} \quad . \tag{2.15}$$

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 $U[0,1]$.

$\mathbf{p_{ppost}}$: An easy computation shows that

$$f(\mathbf{x}|t;\lambda) \propto \lambda^{n-1} \exp\{-\lambda\ (\textstyle\sum x_i - nt)\}, \tag{2.16}$$

so that the partial posterior for λ is

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

and the partial posterior predictive density is

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

The partial posterior p-value can then be computed to be

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

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

$\mathbf{p_{cpred}}$: Maximization of (2.16) over λ yields $\hat{\lambda}_{cMLE} \propto \sum_{i=1}^{n} X_i - nX_1 = S - nT$. It can be seen that $X_{(i)} - X_{(1)}$ is independent of $X_{(1)}$, from which it follows directly that $\hat{\lambda}_{cMLE}$ is independent of T. Therefore, $p_{cpred} = p_{ppost}$. Note that the derivation of p_{ppost} was considerably simpler than that of p_{cpred}. Finally, as in the argument leading to (2.12), it can be shown that $\Pr(p_{ppost}(\mathbf{X}) \leq \alpha)$ does not depend on λ. Also, $\hat{\lambda}_{cMLE}$ 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 can not be directly applied to p_{ppost}.

It is something of a curiosity that, in both Examples 2.1 and 2.2, p_{sim} coincided 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}, since 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 2.1 and 2.2 both satisfy the conditions of the theorem.)

THEOREM 2. *Suppose $f(\mathbf{x};\theta)$ 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, $\alpha>-1$, and $\gamma<\alpha-1$. Under the usual noninformative prior, $\pi(\theta)=1/\theta$, it will be the case that p_{cpred}, p_{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 $m(t|u_{obs})=ct^{\gamma}u^{\alpha-\gamma-2}(t+u)^{-\alpha}$, where c is the appropriate normalizing constant, and hence

$$p_{cpred}=\int_{t_{obs}}^{\infty}m(t|u_{obs})dt=c\int_{t_{obs}}^{\infty}\frac{t^{\gamma}}{(t+u_{obs})^{\alpha}}\,. \tag{2.17}$$

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

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

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

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

Changing variables to $t=(s_{obs}w)/(w+u_{obs})$ in the latter integral reduces to that in (2.17) and the theorem follows.

□

2.3. Computation

Simulation methods will typically be required 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 $f(t_{obs};\theta)$ is available in closed form.

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

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

an obvious strategy is to repeatedly generate θ from $\pi(\theta|\mathbf{x}_{obs}\backslash t_{obs})$ 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 $\pi(\theta|\mathbf{x}_{obs}\backslash t_{obs})$. If generation from the full posterior $\pi(\theta|\mathbf{x}_{obs})$ is easy, then a simple Metropolis chain will do the job: use $\pi(\theta|\mathbf{x}_{obs})$ as the probing distribution to obtain a candidate θ^*, and then move from the current θ to the candidate with probability $\min\{1, f(t_{obs};\theta)/f(t_{obs};\theta^*)\}$. 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 $\pi(\theta|\mathbf{x}_{obs})$ is available, then one could estimate p_{ppost} by

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

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

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

Computing $p_{cpred(u)}$ is usually considerably more difficult, unless the densities on the left hand side or the right hand side of (2.4) are available in closed form. Various Gibbs and Metropolis-Hasting schemes for its computation are given in Bayarri and Berger (1999), and will not be repeated here. The computational difficulty of $p_{cpred(u)}$ (see, also, Pauler 1999) is the main reason 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 (1999), which derives corresponding results under assumptions that yield asymptotic normality. Seeing the results in the finite sample setting (under normality) should help to alleviate any concerns about 'asymptopia.' Note that, in this section, we shall not attempt to distinguish between the maximum likelihood estimator and its value at the observed data; both will be 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, $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)^t$ the $k \times 1$ vector of regression coefficients, $\mathbf{V}$ a full rank $n \times k$ matrix of covariables, and $\boldsymbol{\epsilon}$ an $n \times 1$ vector of errors. Assume

$$\mathbf{Y} = \mathbf{V}\boldsymbol{\theta} + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \sigma^{\mathbf{2}}\mathbf{I}), \quad \sigma^2 \text{ known.} \tag{3.1}$$

Consider a linear departure statistic $T = \mathbf{w}^{\mathbf{t}}\mathbf{Y}$, with given $\mathbf{w} = (w_1, w_2, \ldots, w_n)^t$. It follows from (3.1) that

$$T|\boldsymbol{\theta} \sim N(\mathbf{w}^{\mathbf{t}}\mathbf{V}\boldsymbol{\theta}, \sigma^2||\mathbf{w}||^2). \tag{3.2}$$

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

Plug-in p-value: It follows from (3.2) that p_{plug} is given by

$$p_{plug} = \Pr{}^{f(t;\hat{\theta})}(T > t_{obs}) = 1 - \Phi\left(\frac{t_{obs} - \mathbf{w}^{\mathbf{t}}\mathbf{V}\hat{\boldsymbol{\theta}}}{\sigma\sqrt{||\mathbf{w}||^2}}\right).$$

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

$$T - \mathbf{w}^{\mathbf{t}}\mathbf{V}\hat{\boldsymbol{\theta}} \sim N(\mathbf{w}^{\mathbf{t}}\mathbf{B}\mathbf{V}\boldsymbol{\theta}, \sigma^2\mathbf{w}^{\mathbf{t}}\mathbf{B}\mathbf{B}^{\mathbf{t}}\mathbf{w}), \tag{3.3}$$

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

$$p_{plug}(\mathbf{Y}) = 1 - \Phi\left(\sqrt{\frac{\mathbf{w}^{\mathbf{t}}\mathbf{B}\mathbf{w}}{||\mathbf{w}||^2}}\; Z\right), \tag{3.4}$$

where $Z \sim N(0,1)$. Thus $p_{plug}(\mathbf{Y})$ will have a $U[0,1]$ distribution only if $\mathbf{w}^{\mathbf{t}}\mathbf{B}\mathbf{w}/||\mathbf{w}||^2 = 1$, which in turn can happen only if $\mathbf{V}^{\mathbf{t}}\mathbf{w} = 0$. While 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, $\mathbf{w^t B w}/||\mathbf{w}||^2$ will be smaller than 1, so that p_{plug} will be conservative.

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

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

When considered as a random p-value and using (3.3), p_{post} can be expressed as

$$p_{post}(\mathbf{Y}) = 1 - \Phi\left(\sqrt{\frac{\mathbf{w^t B w}}{\mathbf{w^t C w}}}\; Z\right). \tag{3.5}$$

Again, this will be $U[0,1]$ only if $\mathbf{V^t w} = \mathbf{0}$. Otherwise, $\mathbf{w^t C w}$ will be larger than $||\mathbf{w}||^2$ and, comparing (3.4)and (3.5), p_{post} will then be even more conservative than p_{plug}. This observation was first made in the asymptotic setting in BVV (1999).

Partial posterior predictive p-value: Calculation yields

$$f(\mathbf{y}|t_{obs};\boldsymbol{\theta}) \propto \exp\left\{-\frac{1}{2\sigma^2}\left[(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^t\mathbf{V^t V}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}) - (\mathbf{w^t V}\boldsymbol{\theta} - T)^t(||\mathbf{w}||^2)^{-1}(\mathbf{w^t V}\boldsymbol{\theta} - T)\right]\right\}. \tag{3.6}$$

Since the partial posterior distribution $\pi(\boldsymbol{\theta}|\mathbf{x}_{obs}\backslash t_{obs})$ is proportional to (3.6), expanding the quadratic forms in (3.6) and rearranging terms yields

$$\pi(\boldsymbol{\theta}|\mathbf{x}_{obs}\backslash t_{obs}) = N_k\left(\boldsymbol{\theta}|\ \mathbf{u}_{obs}, \sigma^2\boldsymbol{\Sigma}\right), \tag{3.7}$$

where $\mathbf{U} = \left(\mathbf{V^t H V}\right)^{-1}\mathbf{V^t H Y}$, $\boldsymbol{\Sigma} = \left(\mathbf{V^t H V}\right)^{-1}$, $\mathbf{H} = \left[\mathbf{I} - \frac{\mathbf{w}\mathbf{w}^t}{||\mathbf{w}||^2}\right]$. From (3.2) and (3.7), it follows that the partial predictive distribution of T is given by

$$T\,|\,\mathbf{x}_{obs}\backslash t_{obs} \sim N\left(\mathbf{w^t V u}_{obs}, \sigma^2\mathbf{w^t}\left[\mathbf{I} + \mathbf{V}\boldsymbol{\Sigma}\mathbf{V^t}\right]\mathbf{w}\right),$$

so that the partial posterior predictive p-value is

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

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

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

where $\mathbf{D} = \mathbf{I} - \mathbf{V}(\mathbf{V^t H V})^{-1}\mathbf{V^t H}$ and $\mathbf{H}$ is given below (3.7). Algebra shows that $\mathbf{w^t D V} = \mathbf{0}$ and $\mathbf{w^t D D^t w} = \mathbf{w^t}\left[\mathbf{I} + \mathbf{V}\boldsymbol{\Sigma}\mathbf{V^t}\right]\mathbf{w}$, so that $p_{ppost}(\mathbf{Y}) = 1 - \Phi(Z)$, where $Z \sim N(0,1)$. Thus p_{ppost} is a valid frequentist p-value.

Conditional predictive p-value: It can easily be seen from (3.6) and (3.7) that the U maximizing (3.6) is precisely the statistic U given in (3.7), that is, $\mathbf{U} = \left(\mathbf{V^t H V}\right)^{-1}\mathbf{V^t H Y}$. Since T and U have a joint $(k+1)$-variate Normal distribution such that $\mathbf{Cov}(\mathbf{T},\mathbf{U}) = \mathbf{w^t H V}\left(\mathbf{V^t H V}\right)^{-1} = \mathbf{0}$, 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 shall consider here. Several difficulties are often encountered with this approach. First, conditioning on any U can severely constrain the sample space, resulting in serious conservatism of the resulting p-value (since there may then be very few possible observations in the 'tail' of the departure statistic, T). Second, the choice of T is often essentially 'forced' on the user, which can also be undesirable in terms of power. Finally, what can be termed 'conditional issues' arise in extreme parts of the sample space. We will see that p_{ppost} seems to substantially overcome these difficulties. (We do not consider p_{cpred} because the choice of the conditioning statistic in (2.6) will not typically work in discrete problems, and because *any* conditional p-value can fall prey to the first difficulty mentioned above.)

We study in some detail 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 predictive posterior p-value. There are many other p-values for contingency tables that have been proposed (a nice survey of conditional tests can be found in Agresti 1992; see, also, Hwang and Yang 1997), but most are dependent on the presence of an alternative which, again, we are not considering.

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. It is desired 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 shall denote an observed value of $X_{..}$ by an 'o' superscript, i.e., $x^o_{..}$.

Case 1. Test of homogeneity

Here the null model is

$$f(x_{11}, x_{12}\ ;\theta) = \binom{n_1}{x_{11}}\binom{n_2}{x_{12}}\theta^{x_{11}+x_{12}}(1-\theta)^{n-x_{11}-x_{12}}, \qquad x_{11} = 0, \ldots, n_1,\ x_{12} = 0, \ldots, n_2. \tag{4.1}$$

The Fisher exact test conditions on the other marginal total, say X_{1+}. This conditioning then forces the choice of T, which can be taken to be $T = X_{11}$. (As noted earlier, this may not be a particularly sensible statistic, in general, for measuring departure from homogeneity). An easy computation shows that the (one-tailed) p-value corresponding to this Fisher exact test (p_{fet}) is given by

$$p_{fet} = \sum_{j=t_{obs}}^{\min\{x^o_{1+}, n_1\}} f(t|x^o_{1+}) = \sum_{j=t_{obs}}^{\min\{x^o_{1+}, n_1\}} \binom{n_1}{j}\binom{n_2}{x^o_{1+}-j} \Big/ \binom{n}{x^o_{1+}} \ .$$

We also select $T = X_{11}$ as the discrepancy statistic for p_{ppost}, although this is by no means necessary. Indeed, it seems much more natural to choose a statistic such as $T = (\frac{1}{n_1}X_{11} - \frac{1}{n_2}X_{22})$; intuitively, this should have more power against reasonable alternatives. We are not, however, delving into power comparisons in this article. Furthermore, it is of interest to separate the issue of choice of T from the other facets of dealing with discrete problems; thus, for comparison of

p_{ppost} and p_{fet}, we consider the same T as in the Fisher exact test.

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

$$p_{ppost} = \sum_{j=t_{obs}}^{n_1} m(t\,|\mathbf{x}_{obs}\backslash t_{obs}) = \sum_{j=t_{obs}}^{n_1} \frac{n_2+1}{n_1+1} \binom{n_1}{j} \binom{n_2}{x^o_{12}} \Big/ \binom{n}{x^o_{12}+j} \ .$$

Incidentally, it can be shown in this problem (for the given choice of T) that $p_{cpred} = p_{ppost}$; this is thus a discrete situation in which conditioning as in (2.6) does not unduly restrict the sample space.

Example 4.1 As a rather extreme test case, consider $n_1 = 3$ and $n_2 = 2$. Here, conditioning on x_{1+} causes a dramatic reduction in the sample space for T. Indeed, this space reduces to just one point when $x^o_{1+} = 0$ or $x^o_{1+} = 5$, to two points when $x^o_{1+} = 1, 4$ and to three three points when $x^o_{1+} = 2, 3$. On the other hand, for p_{ppost} the sample space of T is always $\{0, 1, 2, 3\}$. This has the immediate effect of allowing p_{ppost} to take values closer to 0 and 1.

Figure 2 gives the distribution functions of $p_{fet}(\mathbf{X})$ and $p_{ppost}(\mathbf{X})$ for two different values of θ. Recall that the goal is to have p-values with close to uniform distributions, and p_{post} 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.

A variety of other θ were also considered. For very small θ, both p-values were comparable and very conservative, but p_{ppost} began to perform noticeably better even for values of θ as small as 0.2. Both distributions are closest to uniformity for $\theta \approx 0.5$. For large values of θ, p_{fet} was again very conservative, while p_{ppost} performed remarkably well, unless θ was very large; this later situation is discussed more fully in a later scenario.

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

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

Here, the Fisher exact test conditions on both marginals and again results in the (forced) departure statistic $T = X_{11}$. The ensuing conditional density of T is

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

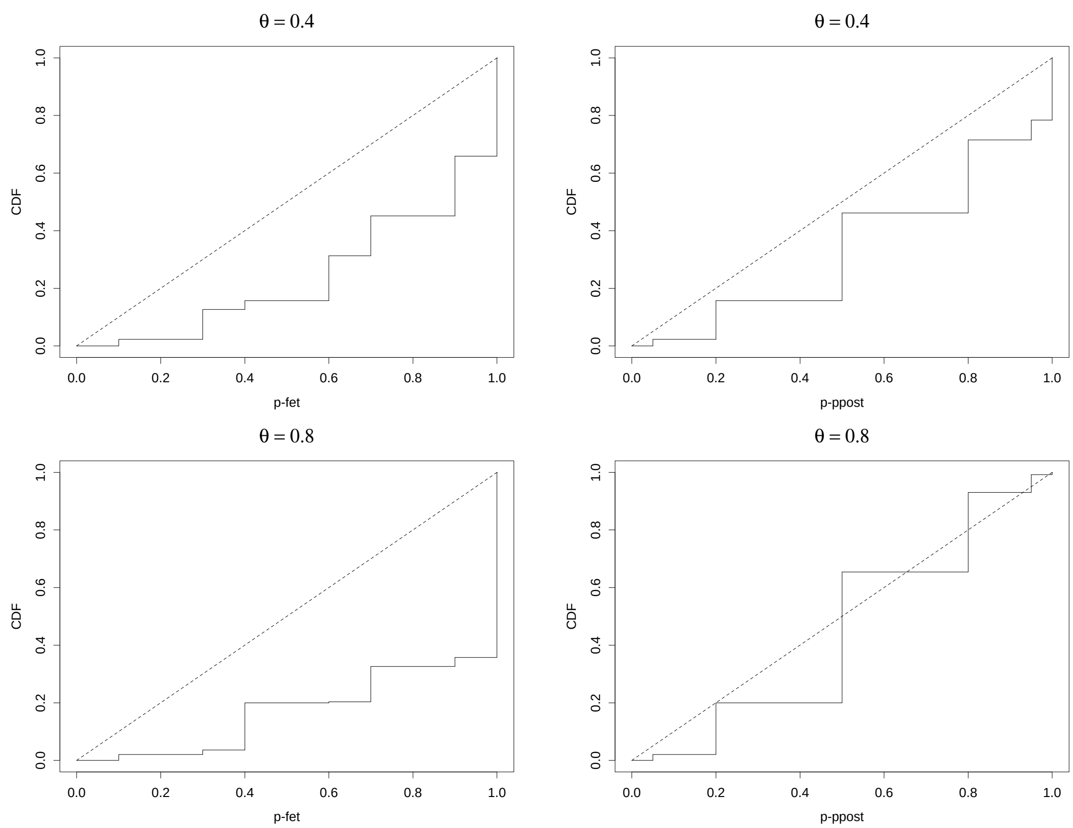

Figure 2: Distributions of $p_{fet}(\mathbf{X})$ (left) and $p_{ppost}(\mathbf{X})$ (right) in Example 4.1 for $\theta = 0.4$ (first row) and $\theta = 0.8$ (second row) .

which produces the same p-value as in the test for homogeneity.

In deriving p_{ppost}, we shall again use the same T as in the Fisher exact test, for purposes of comparison. Using uniform independent priors for θ, ξ, the partial posterior, $\pi(\theta, \xi \,|\mathbf{x}_{obs} \backslash t_{obs})$, is proportional to $f(\mathbf{x}|t_{obs}; \theta, \xi)$ and p_{ppost} can most conveniently be expressed as

$$p_{ppost} = \int_0^1 \int_0^1 \pi(\theta, \xi \,|\mathbf{x}_{obs} \backslash t_{obs}) \sum_{t=t_{obs}}^{n} Bi(t|n, \theta\xi) \, d\theta \, d\xi, \tag{4.2}$$

where

$$\pi(\theta, \xi \,|\mathbf{x}_{obs} \backslash t_{obs}) \propto \theta^{x^o_{21}} (1-\theta)^{x^o_{+2}} \xi^{x^o_{12}} (1-\xi)^{x^o_{2+}} (1-\theta\xi)^{-(n-t_{obs})} \,. \tag{4.3}$$

(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_{post}, we employ importance sampling based on the importance function

$$\frac{1}{2} Un(\theta|0,1) \, Be(\xi|x^o_{12}+1, x^o_{22}+1) + \frac{1}{2} Be(\theta|x^o_{21}+1, x^o_{22}+1) \, Un(\xi|0,1). \tag{4.4}$$

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

Example 4.2 To maintain similarity with the previous example, we consider $n = 5$. It can be shown that the support of $p_{fet}(\mathbf{X})$ is limited to $\{0.1, 0.2, 0.3, 0.4, 0.6, 0.7, 0.8, 0.9\}$, while the support of $p_{ppost}(\mathbf{X})$ 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 (the third row of the table that will be addressed shortly), p_{ppost} seems to be 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} being the less conservative. Both p-values perform at their best for $\theta, \xi \approx 0.5$, with p_{ppost} performing much better. When one of θ, ξ is small and the other large, p_{fet} is again very conservative, while 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 needs to 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 0.05 to be within 20% of 0.05 (the value under the desired uniformity). When $(\theta, \xi) = (0.6, 0.5)$, this obtains for p_{fet} only when $n \approx 500$; in contrast, this occurs for p_{ppost} when $n \approx 10$. When $(\theta, \xi) = (0.3, 0.9)$, p_{fet} requires $n \approx 1200$, while p_{ppost} needs only $n \approx 110$.

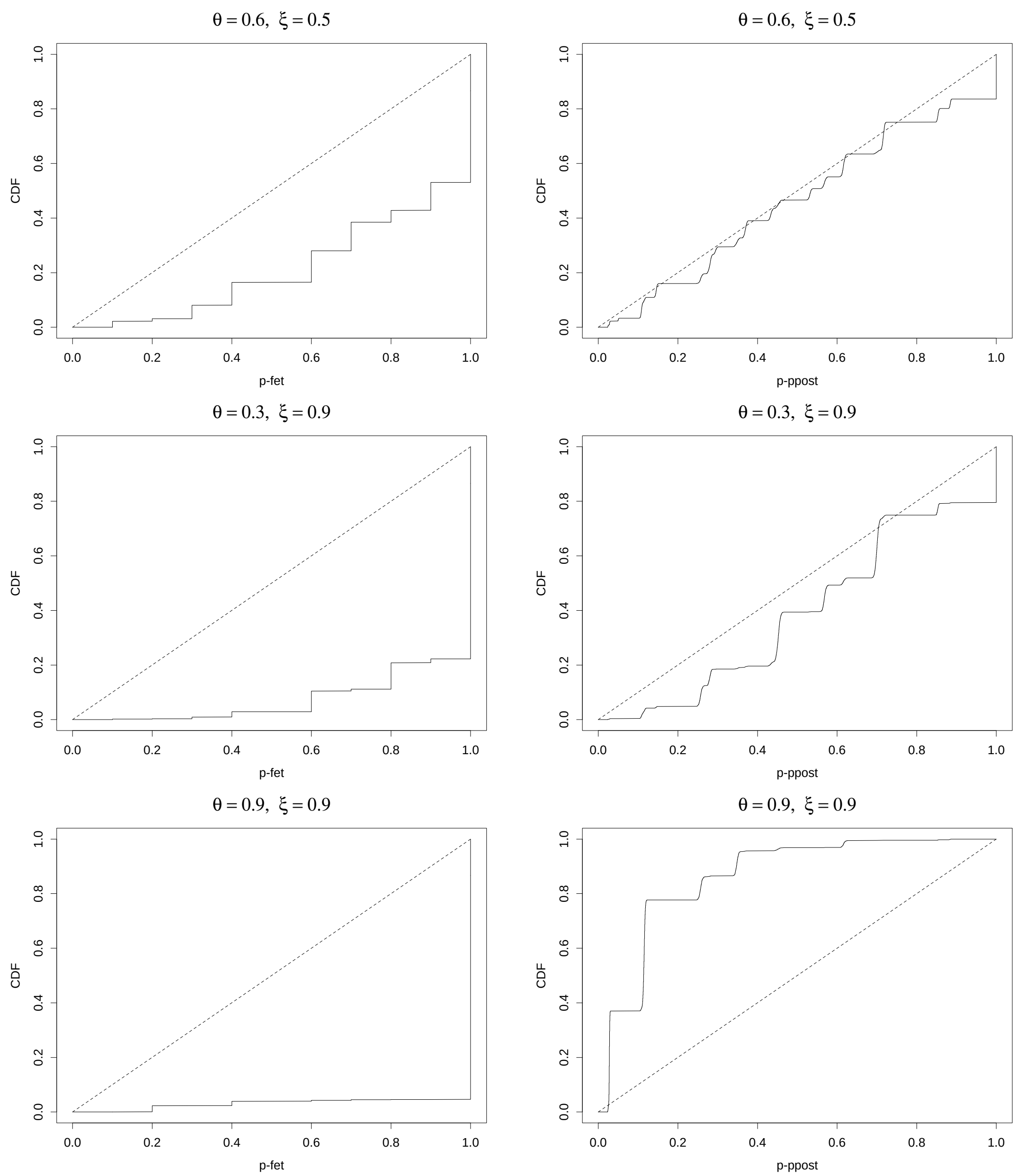

Figure 3: Distributions of $p_{fet}(\mathbf{X})$ (left) and $p_{ppost}(\mathbf{X})$ (right) in Example 4.2 for $(\theta, \xi) = (0.6, 0.5)$ (first row) $(\theta, \xi) = (0.3, 0.9)$ (second row), and $(\theta, \xi) = (0.9, 0.9)$ (third row) .

The apparent breakdown of both p_{fet} and p_{ppost} for large values of (θ, ξ) (such as $(0.9, 0.9)$ in Figure 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 anti-conservative 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 $T = X_{11}$ will clearly typically be very large; p_{ppost} reacts to this with ready 'rejection' of the null model, while p_{fet} ignores all but incredibly large T.

The anti-conservative behavior of p_{ppost} arises because a very large T contains a great deal of information about the parameters, but relatively little information about deviance from the model. The most extreme example of this behavior is when T is chosen to be 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 4.3 Assume $X_i \sim Bernoulli(\theta)$, $i = 1, \ldots, n$, and that $T = \sum X_i$, a sufficient statistic. Here, $m(t|u) = m(t\,|\mathbf{x}_{obs} \backslash t_{obs}) = m(t) = 1/(n+1)$ for $t = 0, \ldots, n$, and $p_{ppost} = 1 - t_{obs}/(n+1)$. For large n, the distribution of p_{ppost} is approximately $N(1-\theta, \theta(1-\theta)/n)$, which tightly concentrates around $1-\theta$. Thus, when θ is large, the distribution function of p_{ppost} jumps immediately, giving rise to anti-conservative 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 such are being 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 being 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 well be too restrictive for a Bayesian. The natural (Bayesian) requirement (see Meng 1994) is to require a p-value to be uniform under the prior predictive distribution. Since, 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 Figure 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

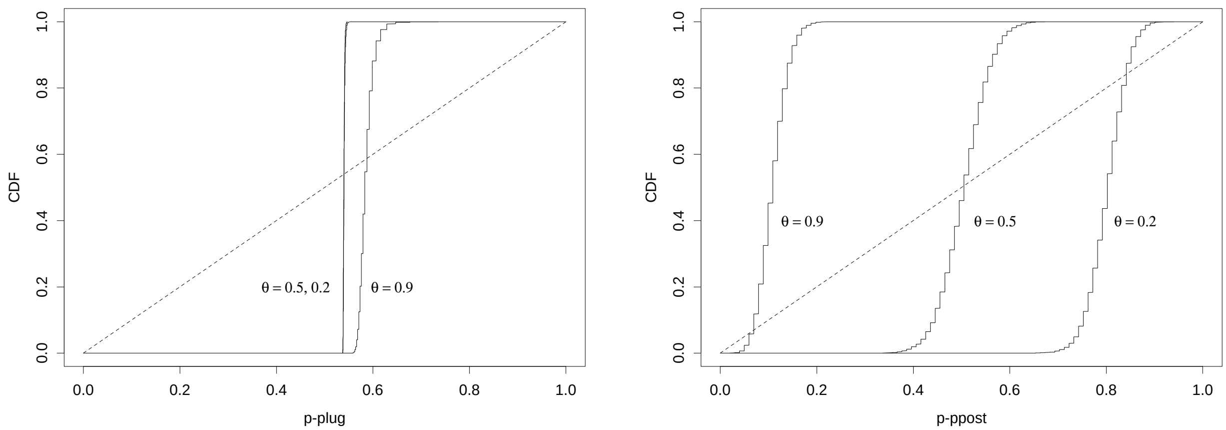

Figure 4: Distributions of $p_{plug}(\mathbf{X})$ (left) and $p_{ppost}(\mathbf{X})$ (right) in Example 4.3 for $\theta = 0.2, 0.5, 0.9$ and $n = 100$.

fact that the noninformative prior is actually proper, but related arguments involving averages with respect to classes of priors can probably be made for the improper case; we do not pursue this here.

5. Conclusions

Our comparisons have not included p_{prior}, since this can not 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 our continuous examples it happened that p_{sim} was equal to p_{cpred}, although this is certainly not true in general. For those wishing to use p_{sim}, this equality is a fortunate occurrence when it obtains (see Theorem 3) since p_{cpred} is typically much easier to compute directly in those situations than is 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 to be 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 to simply replace θ by $\hat{\theta}$. The explanation is that p_{post} does not account for parameter uncertainty in a legitimate Bayesian way, since it involves a double use of the data. (Indeed, the original motivation for p_{cpred} and

p_{post} was precisely to account for parameter uncertainty in a legitimate Bayesian fashion.) We have only considered a few situations here but, together with the similar asymptotic conclusion in RVV (1999), 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 m.l.e., but one could then plug in, say, a posterior mean for θ, rather than the m.l.e.) At the very least, the observations here and in RVV (1999) 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 1999).

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 (1999), and suggests that the latter p-values are to be preferred in practice, if available. 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 in RVV (1999). In a sense, the new p-values can be viewed as automatically 'centering' a departure statistic T, a considerable simplification in practice, especially if T is being chosen on a casual basis for model-checking.

The situation involving discrete distributions is more complex, but the gains through use of the new p-values, especially p_{ppost}, seem quite dramatic. Discreteness of the sample space causes 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. Because the sample space for p_{ppost} is typically considerably richer in discrete problems than the sample space arising from any conditional predictive p-value, we would recommend its use for discrete problems.

Acknowledgements

This work was supported in part by the National Science Foundation (U.S.A.) under Grants DMS-9303556 and DMS-9802261, and by the Ministry of Education and Culture (Spain) under Grant PB96-0776.

Appendix

Details for Theorem 1:

Suppose $\pi(\theta)$ is improper but that there exists a sequence of increasing compact sets $\Theta_k \subset \Theta$ such that $\bigcup_{k\geq 1} \Theta_k = \Theta$, $0 < m_k = \int_{\Theta_k} \pi(\theta)d\theta < \infty$, $0 < m(u) = \int_\Theta f(u;\theta)\pi(\theta)d\theta < \infty$, 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 $h(u,\theta) = \Pr(p(\mathbf{X}) \leq \alpha | u;\theta)$. From the definition of $p_{cpred(u)}$, 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,

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

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

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

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

To prove that $\alpha = c$, completing the proof, we need only show that the difference of the left hand sides of (A.4) and (A.5) goes to 0 as $k \to \infty$. Breaking the left hand 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 hand sides of (A.4) and (A.5) can be written

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

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

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

which goes to zero by (A.1) and completes the proof. $\square$

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 similarly handled. Write $f(u;\theta) = g(u-\theta)$, let $G(\cdot)$ denote the c.d.f. correponding to $g(\cdot)$, and choose $\Theta_k = (-k,k)$. Then $m(u) = \int g(u-\theta) d\theta = 1$, $m_k = \int_{-k}^{k} (1) d\theta = 2k$, $m_k(u) = \frac{1}{2k} \int_{-k}^{k} g(u-\theta) d\theta = \frac{1}{2k} \left[G(u+k) - G(u-k)\right]$, and (A.1) becomes

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

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

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

Since ϵ was arbitrary, it is clear that (A.6) is satisfied.

Verification of (2.12):

Lemma 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 $\{\mathbf{W} : W_i > c \;\; \forall i\}$ 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 $\sqrt{2}$, while 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. $\square$

To establish (2.12), note that

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

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 1 clearly apply, with $W_{(1)} = T/s$, and the result follows.

Computation of p_{ppost} for independence in contingency tables:

From (4.2) and (4.3), it is clear that a Monte Carlo importance sampling approximation to p_{ppost} in (4.2) is given by

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

where $\mathcal{B}(x; n, \varphi)$ is the distribution function at x of the $Bi(n, \varphi)$ 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, $\pi(\theta_i, \xi_i \,|\mathbf{x}_{obs}\backslash t_{obs})\,/h(\theta_i, \xi_i)$, 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 (4.4) can be seen to have these properties. In particular, the importance ratio can be computed to be

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

where

$$c_1 = \frac{\Gamma(n - x_{11}^o - x_{21}^o + 2)}{\Gamma(x_{12}^o+1)\Gamma(x_{22}^o+1)}, \qquad c_2 = \frac{\Gamma(n - x_{11}^o - x_{12}^o + 2)}{\Gamma(x_{21}^o+1)\Gamma(x_{22}^o+1)} .$$

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

References

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

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

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

[4] Bayarri, M. J., and Berger, J. O. (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 (in press).

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

[6] Berger, J. O., Brwon, 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.

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

[8] 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.

[9] 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*, ??, 1012–1016.

[10] Blyth, C. R., and Staudte, R. G. (1995), "Estimating Estatistical Hypothesis," *Probability and Satistics Letters* **23**, 45–52.

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

[12] 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 (in press).

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

[14] 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.

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

[16] Evans, M. (1997), "Bayesian Inference Procedures Derived Via the Concept of Relative Surprise," *Communications in Statistics*, **26**, 1125–1143.

[17] 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.

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

[19] 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.

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

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

[22] 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*, textbf15, 103–128.

[23] 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, Department of Statistical Science.

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

[25] 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 (in press).

[26] 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 (in press).

[27] Robins, J. M., van der Vaart, A., and Ventura, V. (1999), "The Asymptotic Distribution of P-values in Composite Null Models," *Technical Report*, Department of Epidemiology, Harvard School of Public Health.

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

[29] Rubin, D. B. (1996), Discussion of "Posterior Predictive Assessment of Model Fitness Via Realized Discrepancies," *Statistica Sinica*, **6**, 787–792.

[30] Schaafsma, W., Tolboom, J., and Van Der Meulen, E. A. (1989), "Discussing Truth or Falsity by Computing a Q-value," *Statistical Data Analysis and Inference*, 85–100.

[31] Sellke, T., Bayarri, M. J., and Berger, J. O. (1999), "Calibration of P-values for Precise Null Hypotheses," *ISDS Discussion Paper,* Duke University.

[32] Thompson, P. (1997), "Bayes P-values, " *Statistics and Probability Letters*, **31**, 267–271.

[33] 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.