Statistica Sinica **6**(1996), 733-807

POSTERIOR PREDICTIVE ASSESSMENT OF MODEL FITNESS VIA REALIZED DISCREPANCIES

Andrew Gelman, Xiao-Li Meng and Hal Stern

Columbia University, The University of Chicago, and Iowa State University

Abstract: This paper considers Bayesian counterparts of the classical tests for goodness of fit and their use in judging the fit of a single Bayesian model to the observed data. We focus on posterior predictive assessment, in a framework that also includes conditioning on auxiliary statistics. The Bayesian formulation facilitates the construction and calculation of a meaningful reference distribution not only for any (classical) statistic, but also for any parameter-dependent "statistic" or *discrepancy*. The latter allows us to propose the *realized discrepancy assessment* of model fitness, which directly measures the true discrepancy between data and the posited model, for any aspect of the model which we want to explore. The computation required for the realized discrepancy assessment is a straightforward byproduct of the posterior simulation used for the original Bayesian analysis.

We illustrate with three applied examples. The first example, which serves mainly to motivate the work, illustrates the difficulty of classical tests in assessing the fitness of a Poisson model to a positron emission tomography image that is constrained to be nonnegative. The second and third examples illustrate the details of the posterior predictive approach in two problems: estimation in a model with inequality constraints on the parameters, and estimation in a mixture model. In all three examples, standard test statistics (either a χ^2 or a likelihood ratio) are not pivotal: the difficulty is not just how to compute the reference distribution for the test, but that in the classical framework no such distribution exists, independent of the unknown model parameters.

Key words and phrases: Bayesian p-value, χ^2 test, discrepancy, graphical assessment, mixture model, model criticism, posterior predictive p-value, prior predictive p-value, realized discrepancy.

1. Introduction

1.1. Classical and Bayesian model assessment

Assessing the plausibility of a posited model (or of assumptions in general) is always fundamental, especially in Bayesian data analyses. A Bayesian analysis, which conditions on the whole probability model, can be very misleading when the model is far from plausible. Any meaningful Bayesian analysis, therefore, should at least include a check to see if the posited model should be excluded

because it fails to provide a reasonable summary of the data at hand. A standard classical approach for this kind of model-checking is to perform a goodness-of-fit test, which calculates a tail-area probability under the posited model to quantify the extremeness of the observed value of a selected discrepancy (e.g., differences between observations and predictions). The essence of the classical approach, as we see it, lies in comparing the observed value of the discrepancy to its reference (i.e., sampling) distribution derived under the posited model. The tail-area probability, or p-value, is only a computationally convenient way of locating the observed value in the reference distribution, especially in the days when graphical display of the reference distribution could not be done routinely. Plotting the observed value against a reference distribution is more informative than merely reporting a p-value.

The main purpose of this paper is to extend the essence of the classical approach to the Bayesian framework, with the aim of providing pragmatic methods of assessing the fitness of a single model, especially in complex situations where classical frequentist calculation of a reference distribution of an informative statistic is not feasible. Our work is based on several earlier papers, particularly Rubin (1984), §5, on assessing the fitness of a single Bayesian model. We use the term "assessment" instead of "testing" to highlight the fundamental difference between assessing the *discrepancies* between a model and data and testing the *correctness* of a model. We believe that there is a general concern that there has been too much emphasis on the latter problem and that, in the words of Tiao and Xu (1993), there is a need in practice for "... development of diagnostic tools with a greater emphasis on assessing the usefulness of an assumed model for specific purposes at hand rather than on whether the model is true."

For some problems, such as linear models, common goodness-of-fit tests serve as good diagnostics (when used and interpreted appropriately) and are easy to implement, because their reference distributions are known, at least approximately. Useful approximations to distributions of test statistics are possible for some problems (see, for example, Chernoff (1954), concerning extensions of the linear model), but are not always available (see, for example, McCullagh (1985, 1986), concerning the difficulty of finding distributions of classical goodness-of-fit tests in generalized linear models). The classical approach relying on known, or approximately known, reference distributions encounters difficulty in at least three kinds of models: models with severe restrictions on the parameters, such as positivity; models with probabilistic constraints, which may arise from a strong prior distribution; and unusual models that cannot be parameterized as generalized linear models. The following example illustrates the type of problem with which we are concerned, where it is difficult in practice and may be impossible in theory to construct a sensible classical test.

1.2. A motivating example

Gelman (1990, 1992) describes a positron emission tomography experiment whose goal is to estimate the density of a radioactive isotope in a cross-section of the brain. The two-dimensional image is estimated from gamma-ray counts in a ring of detectors around the head. Each count is classified in one of n bins, based on the positions of the detectors when the gamma rays are detected, and a typical experimental run has about 6,000,000 counts. The bin counts, y_i, are modeled as independent Poisson random variables with means θ_i that can be written as a linear function of the unknown image g, $\theta = Ag + r$, where $\theta = (\theta_1, \ldots, \theta_n)$, A is a known linear operator that maps the continuous g to a vector of length n, and r is a known vector of corrections. Both A and r, as well as the image, g, are nonnegative. In practice, g is discretized into "pixels" and becomes a long (length greater than n) nonnegative vector, and A becomes a matrix with nonnegative elements.

Were it not for the nonnegativity constraint, there would be no problem finding an image to fit the data; in fact, an infinite number of images g solve the linear equation, $y = Ag + r$. However, due to the Poisson noise, and perhaps to failures in the model, it often occurs in practice that no exact nonnegative solutions exist, and we must use an estimate (or family of estimates) $\hat{g}$ for which there is some discrepancy between the data, y, and their expectations, $\hat{\theta} = A\hat{g}+r$. The discrepancy between y and $\hat{\theta}$, however, should not be too large; given the truth of the model, it is limited by the variance in the independent Poisson distributions. For example, the χ^2 discrepancy,

$$X^2(y;\hat{\theta}) = \sum_{i=1}^{n} \frac{(y_i - \hat{\theta}_i)^2}{\hat{\theta}_i}, \tag{1}$$

should be no greater than could have arisen from a χ^2 distribution with n degrees of freedom ($y_i > 50$ for almost all the bins i, and so the χ^2 distribution, based on the normal approximation to the Poisson, is essentially exact). In fact, $X^2(y;\hat{\theta})$ should be considerably less, since $\hat{\theta}$ is the best fit to the data.

The posited model was fit to a real dataset, y, with $n = 22{,}464$. We would ultimately like to determine the posterior distribution, $P(\theta|y)$, given a reasonable prior distribution. As a first step, we need to determine if the model is invalidated by the observed data. For this dataset, the best-fit nonnegative image $\hat{g}$ was not an exact fit; the discrepancy between y and $\hat{\theta}$ was $X^2(y;\hat{\theta}) \approx 30{,}000$. This is unquestionably a rejection of the model, unexplainable by the Poisson variation. At this point, the model and data should be examined to find the causes of the lack of fit. Possible failures in the model include error in the specification of A and r, lack of independence or super-Poisson variance in the counts, and error from discretizing the continuous image, g.

There is no difficulty in deciding to reject the model here as long as $X^2(y;\hat{\theta})$ is greater than $n+2\sqrt{2n} \approx 23{,}000$, for we can be almost certain that the model does not fit the data in this case. Suppose, however, that the experimental procedure is carefully examined, the model is made more accurate, and the new model is fit to the data, yielding a χ^2 discrepancy of 22,000, or 20,000. We should probably still be distrustful of the model, since a whole continuous image is being fit to the data. (In fact, as the total number of counts increases, the Poisson variances decrease proportionally, and it becomes increasingly likely that an exact fit image $\hat{g}$ will exist that solves $y = A\hat{g} + r$. Thus, conditional on the truth of the model, $X^2(y;\hat{\theta})$ must be zero, in the limit as the number of counts approaches infinity with a fixed number of bins, n. Due to massive near-collinearity, the positron emission tomography model is not near that asymptotic state even with 6,000,000 total counts.) If k linear parameters were fit, the χ^2 discrepancy defined in (1) would have a χ^2 distribution with $n - k$ degrees of freedom. It is difficult to determine how many degrees of freedom correspond to the fitting of a continuous image.

We have arrived at a practical problem: how to assess the quality of a model's fit to the observed data for complicated models in "close calls" for which the simple χ^2_n bound is too crude. The problem is important and common, because if we take modeling seriously, as we should with any Bayesian analysis, we will gradually improve models that clearly do not fit, and upgrade them into close calls.

1.3. A brief overview

The preceding example illustrates the need for model checking in complex situations. In Bayesian statistics, a model can be checked in at least three ways: (1) examining sensitivity of inferences to reasonable changes in the prior distribution and the likelihood; (2) checking that the posterior inferences are reasonable, given the substantive context of the model; and (3) checking that the model fits the data. We address the third of these concerns using the posterior predictive distribution for a *discrepancy*, an extension of classical test statistics to allow dependence on unknown (nuisance) parameters. Posterior predictive assessment was introduced by Guttman (1967), applied by Rubin (1981), and given a formal Bayesian definition by Rubin (1984). Our new methodological contribution is the adoption of more general discrepancies, which allows more direct assessment of the discrepancy between data and the posited model.

The recent rapid development of Bayesian computation allows us to fit more realistic and sophisticated models than previously possible, and thus there is a corresponding need for general methods to assess the fitness of these models when classical tests are not applicable, as our motivating example demonstrates. The

approach we discuss here appears to be one such method. It is simple, both conceptually and computationally, and connects well to the classical goodness-of-fit methods that most researchers are familiar with. It is also very general, applicable for comparing observations with model predictions in any form. Our own applied work has benefited from the application of this method, as documented here and in many examples in Gelman, Carlin, Stern and Rubin (1995) (also see Belin and Rubin (1995), Gelman and Meng (1996) and Upadhyay and Smith (1993)). Meng (1994) discusses a similar method for testing parameter values within a model rather than for the entire model. West (1986) and Gelfand, Dey, and Chang (1992) also present posterior predictive approaches to model evaluation, in the context of sequential data and cross-validation of the existing data set, respectively, rather than comparing to hypothetical replications of the entire data set.

Our paper is organized as follows. Section 2 consists of definitions, a discussion of computational issues, and a theoretical illustration of the posterior predictive approach. Section 3 presents detailed illustration in two applied contexts. Section 4 provides discussion of various issues, including a brief comparison to the prior predictive approach of Box (1980).

2. Posterior Predictive Assessment of Model Fitness

2.1. Posterior predictive assessment using classical test statistics

We use the notation y for the observed data, H for the assumed model, and θ for the d-dimensional unknown model parameter (d can be infinite). Also, T denotes a test statistic, a function from data space to the real numbers. In addition, we may specify a set of *auxiliary statistics*, $A(y)$—functions of the data that are to be held constant in replications (for example, the sample size). In general, $A(y)$ may be a scalar, a vector, or the empty set.

To avoid confusion with the observed data, y, define y^{rep} as the *replicated* data that could have been observed, or, to think predictively, as the data that would appear if the experiment that produced y today were replicated tomorrow with the same model, H, the same (unknown) value of θ that produced y, and the same value of the auxiliary statistics, $A(y)$. Denote the distribution of this replication by P_A:

$$P_A(y^{\text{rep}}|H,\theta) = P[y^{\text{rep}} \mid H,\theta,A(y^{\text{rep}}) = A(y)]. \tag{2}$$

In this notation, the classical p-value based on T is

$$p_c(y,\theta) = P_A[T(y^{\text{rep}}) \geq T(y) \mid H,\theta], \tag{3}$$

and a p-value very close to 0 indicates that the lack of fit in the direction of the test statistic, $T(y)$, is unlikely to have occurred under the model. In (3), y is

fixed, with all the randomness coming from y^{rep}, and the value of p_c is obtainable only when it is free of the (nuisance) parameter θ. As emphasized in Section 1.1, the key ingredient of the classical approach is not the p-value given in (3), but rather locating $T(y)$ in the distribution of $T(y^{\text{rep}})$ derived from (2).

In the Bayesian framework, the inference for θ is provided by its posterior distribution, $P(\theta|H, y)$, where the model H now also includes the prior distribution, $P(\theta)$. Correspondingly, the reference distribution of the future observation y^{rep}, given $A(y)$, is its *posterior predictive distribution*,

$$P_A(y^{\text{rep}}|H, y) = \int P_A(y^{\text{rep}}|H, \theta)P(\theta|H, y)d\theta. \tag{4}$$

The observed value of T, $T(y)$, is then plotted against the distribution of $T(y^{\text{rep}})$ induced by (4). The corresponding tail-area probability, analogous to (3), is

$$p_b(y) = P_A[T(y^{\text{rep}}) \geq T(y)|H, y] = \int p_c(y, \theta)P(\theta|H, y)d\theta, \tag{5}$$

which is the classical p-value of (3) averaged over the posterior distribution of θ. This is the p-value defined by Rubin (1984), which we term the *posterior predictive p-value* (also see Meng (1994)) to contrast it with the *prior predictive p-value* of Box (1980); see Section 4.1 for discussion.

Clearly, the sampling and posterior predictive reference distributions of $T(y^{\text{rep}})$ are identical when T is a pivotal quantity, conditional on $A(y)$, under the model H. For any posited Bayesian model, the posterior predictive approach has the virtue of constructing a well-defined reference distribution (with a corresponding tail-area probability), which is easy to simulate (see Section 2.3), for any test statistic.

Interestingly, the posterior predictive replication appears to be the replication that the classical approach intends to address. Figure 1a shows the posterior predictive reference set, which corresponds to repeating the experiment tomorrow with the same model, H, and the same (unknown) value of θ that produced today's data y. Because θ is unknown, but assumed to have the same value that generated the current data y, we simulate from its posterior distribution given y. Figure 1a is a conditional independence graph (see, e.g., Wermuth and Lauritzen (1990)) that displays the dependence relations between y, θ, and y^{rep}. Given θ, the data y and the replicate y^{rep} are independent, and both represent possible datasets resulting from the given value of θ and the model H. The fact that θ is displayed only once in the graph is *not* intended to imply that θ is known or "fixed," as in a frequentist analysis. (Figures 1b and 1c will be discussed in Section 4.1.)

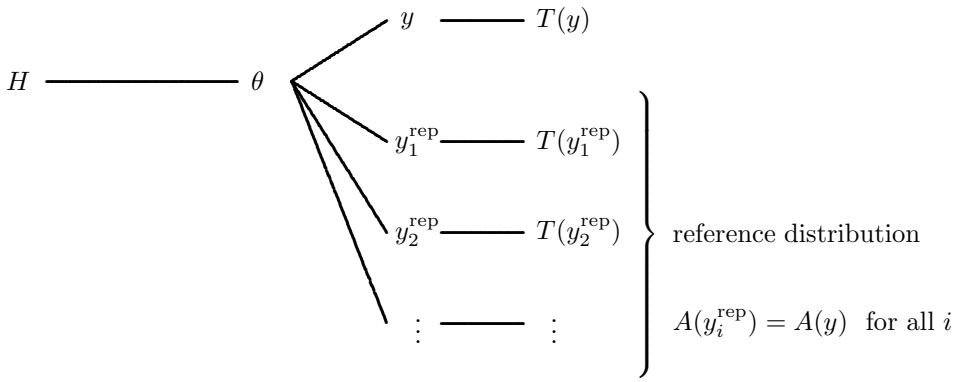

Figure 1a. The posterior predictive distribution

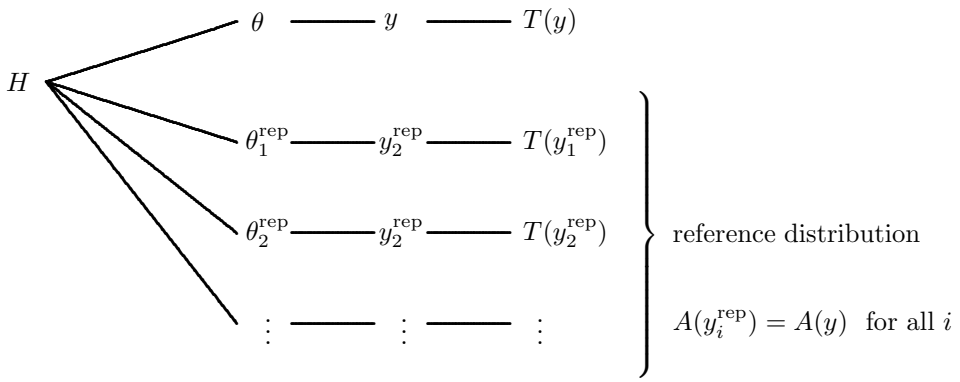

Figure 1b. The prior predictive distribution

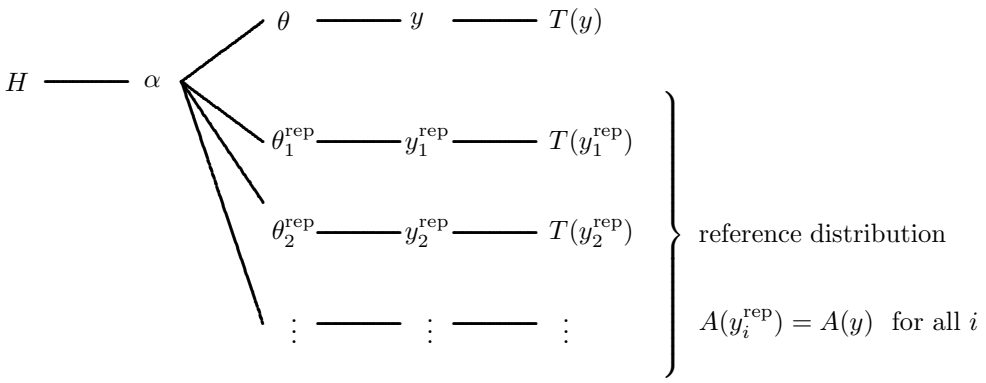

Figure 1c. A mixed predictive distribution

2.2. Posterior predictive assessment using discrepancies

The Bayesian formulation not only handles the problem of nuisance parameters for classical test statistics but also naturally allows the use of test "statistics" that depend on θ. This generalization beyond Rubin's (1984) formulation is important because it allows us to compare directly the discrepancy between the observed data and the posited model, instead of between the data and the best fit of the model. It also, as we shall show in Section 2.3, greatly simplifies the computation of tail-area probabilities. Parameter-dependent test statistics have been suggested before by Tsui and Weerahandi (1989) and Meng (1994) for the case of testing parameters within a model. In particular, Tsui and Weerahandi (1989) termed such a statistic a "test variable" and Meng (1994) called it a "discrepancy variable". We prefer the latter as it emphasizes that the focus here is to measure discrepancies between a model and the data, not to test whether a model is true; we will also drop the term "variable" to avoid possible confusion with model variables (e.g., age).

For a selected discrepancy, $D(y;\theta)$, its reference distribution is derived from the joint posterior distribution of y^{rep} and θ,

$$P_A(y^{\text{rep}},\theta|H,y) = P_A(y^{\text{rep}}|H,\theta)P(\theta|H,y). \tag{6}$$

This replication is a natural extension of (4), which is a marginal distribution of (6). Directly locating the *realized* value, $D(y;\theta)$, within its reference distribution, however, becomes nonfeasible when $D(y;\theta)$ depends on the unknown θ. The classical tail-area approach now becomes one possible technical device for measuring the location. Specifically, we can formally define a tail-area probability of D under its posterior reference distribution:

$$p_b(y) = P_A[D(y^{\text{rep}};\theta) \geq D(y;\theta)|H,y], \tag{7}$$

which includes p_b of (5) as a special case. Interestingly, although the realized value itself is not observable, this posterior predictive p-value is well defined and calculable. The reference set for the discrepancy is the same as that in Figure 1a, except that it is now composed of pairs (y^{rep},θ) instead of just y^{rep}. The term "realized" discrepancy is borrowed from Zellner (1975), where such discrepancies were used for conducting Bayesian analysis of regression error terms (i.e., no predictive replications were involved). Also see Chaloner and Brant (1988) and Chaloner (1991) for related work.

By suitably eliminating its dependence on unknown θ, we can also construct classical test statistics from a discrepancy. For example, as illustrated in Section

2.4 for the χ^2 test, the classical statistics that arise from comparing data with the best fit of the models typically correspond to the *minimum discrepancy*:

$$D_{\min}(y) = \min_{\theta} D(y;\theta).$$

Another possibility is the *average discrepancy* statistic,

$$D_{\text{avg}}(y) = E(D(y;\theta)|H,y) = \int D(y;\theta)P(\theta|H,y)d\theta.$$

The corresponding posterior predictive p-values are defined by (5) with T replaced by $D_{\min}$ and D_{avg}, respectively.

2.3. Computation for posterior predictive assessment

Computation of reference distributions of discrepancies and the corresponding tail-area probabilities can be performed analytically for some simple problems (e.g., Meng (1994)), but in complicated models, such as arise in practical applications, it is more easily accomplished via Monte Carlo simulation. This typically is not an extra computational burden, because simulation is a standard tool for Bayesian analysis with complex models. In other words, the required computation is typically a byproduct of the usual Bayesian simulation that provides a set of draws of θ from the posterior distribution, $P(\theta|H,y)$.

Specifically, consider the computation required for comparing the realized discrepancy $D(y;\theta)$ to its reference distribution under (6). Given a set of (possibly dependent) draws, $\theta^j, j = 1,\ldots,J$, we need only perform the following two steps for each j:

1. Given θ^j, draw a simulated replicated data set, $y^{\text{rep }j}$, from the sampling distribution, $P_A(y^{\text{rep}}|H,\theta^j)$.
2. Calculate $D(y;\theta^j)$ and $D(y^{\text{rep }j};\theta^j)$.

Having obtained $\left\{\left(D(y;\theta^j), D(y^{\text{rep }j};\theta^j)\right), j = 1,\ldots,J\right\}$, we can make a scatterplot (see Figures 7 and 9) to make a graphical assessment, and estimate p_b of (7) by the proportion of the J pairs for which $D(y^{\text{rep }j};\theta^j)$ exceeds $D(y;\theta^j)$.

For many problems, step 1 is easy (step 2 is almost always straightforward). In addition, once the replicates have been drawn, the same draws $\{(\theta^j, y^{\text{rep }j}), j = 1,\ldots,J\}$ can be used for as many realized discrepancies as one wishes. This is particularly convenient if we are interested in measuring the discrepancy of various aspects of the model (e.g., mean-squared errors, quantiles, patterns of residuals, etc.). In cases where the classical p-value based on $D(y;\theta)$ (i.e., treating θ as known) is easy to calculate analytically, one can simulate p_b more efficiently by directly averaging the classical p-values, $p_c(y,\theta^j), j = 1,\ldots,J$. In any case,

we recommend making the scatterplot whenever feasible because it tells us the typical magnitudes of $D(y;\theta)$ and $D(y^{\text{rep}};\theta)$. We also note that "double parametric bootstrap" or various Bayesian bootstrap methods can sometimes be used to obtain approximations to posterior predictive distributions (e.g., Rubin (1987), Ch. 4; Tsay (1992)).

Simulating reference distributions for $D_{\min}$ and D_{avg} is more complicated because one must minimize or average over θ when evaluating their values. To compute $D_{\min}$, one needs to determine, for each j, the value θ for which $D(y^{\text{rep}\,j};\theta)$ is minimized; this may not be an easy computation. The computation for $D_{\text{avg}}(y)$ requires a potentially even more difficult integration. Thus, while the minimum and average discrepancies are interesting theoretically and for comparison to classical methods, they are typically much harder to compute than the realized discrepancy, particular with complicated models.

2.4. A theoretical example: χ^2 discrepancies

As a theoretical illustration, consider the χ^2 discrepancy, by which we simply mean a sum of squares of standardized residuals of the data with respect to their expectations under a posited model. For simplicity, assume that the data are expressed as a vector of n independent observations (not necessarily identically distributed), $y = (y_1, \ldots, y_n)$, given the parameter vector θ, and no auxiliary statistics. The χ^2 discrepancy is then

$$X^2(y;\theta) = \sum_{i=1}^{n} \frac{(y_i - E(y_i|\theta))^2}{\text{Var}\,(y_i|\theta)}. \tag{8}$$

For example, the discrepancy in equation (1) in Section 1.2 is the above formula for the Poisson distribution, evaluated at the estimate $\hat{\theta}$. In this section, we assume that, given θ, expression (8) has an approximate χ^2_n distribution.

Now suppose we are interested in assessing the fitness of a model, H, that constrains θ to lie in a subspace of $\Re^n$, and for the purpose of theoretical comparisons we focus on calculating posterior predictive p-values. Given a prior distribution, $P(\theta)$, on the subspace, we can calculate the posterior predictive p-value based on X^2 as

$$p_b(y) = \int P(\chi^2_n \geq X^2(y;\theta))P(\theta|H,y)d\theta, \tag{9}$$

where χ^2_n represents a chi-squared random variable with n degrees of freedom. This computation is straightforward, once draws from $P(\theta|H,y)$ are obtained, because the tail-area probability function for χ^2_n can be found in any standard statistical software.

Computations for posterior predictive p-values for minimum χ^2, $X^2_{\min}$, and average χ^2, X^2_{avg}, are more complicated in general, as discussed in Section 2.3. However, when H is a linear model (i.e., θ is constrained to lie on a hyperplane of dimension k), the minimum χ^2 discrepancy is essentially equivalent to the classical goodness-of-fit test statistic. (The classical χ^2 test is sometimes evaluated at the maximum likelihood estimate (MLE) and sometimes at the minimum-χ^2 estimate, a distinction of some controversy (see, e.g., Berkson (1980)); we consider the minimum χ^2 in our presentation, but similar results could be obtained using the MLE.) Thus $X^2_{\min}(y)$ is approximately pivotal with a χ^2_{n-k} distribution (e.g., Cochran (1952)). Consequently, the posterior predictive p-value can be approximated by $P(\chi^2_{n-k} \geq X^2_{\min}(y))$.

Furthermore, if θ is given a diffuse uniform prior distribution in the subspace defined by a linear model H, then the posterior predictive distributions of $X^2(y;\theta)$ and $X^2_{\text{avg}}(y)$ are closely related to that of $X^2_{\min}(y)$. With the diffuse prior distribution, the posterior distribution of $X^2(y;\theta) - X^2_{\min}(y)$ is approximately χ^2_k. Then we can decompose the average χ^2 statistic as follows:

$$X^2_{\text{avg}}(y) = E[X^2_{\min}(y) + (X^2(y;\theta) - X^2_{\min}(y))|y] \approx X^2_{\min}(y) + k,$$

and thus the average χ^2 discrepancy is approximately equivalent to the minimum χ^2 discrepancy, just shifted by a constant, k.

For the realized discrepancy, $X^2(y;\theta)$, the same decomposition can be applied to (9), which yields

$$\begin{aligned} p_b(y) &= \int P[\chi^2_n \geq X^2_{\min}(y) + (X^2(y;\theta) - X^2_{\min})]P(\theta|H,y)d\theta \\ &= P(\chi^2_n \geq X^2_{\min}(y) + \chi^2_k) = P(\chi^2_n - \chi^2_k \geq X^2_{\min}(y)), \end{aligned}$$

where χ^2_n and χ^2_k are independent random variables. In other words, assessing a linear model using $X^2(y;\theta)$ is equivalent to using $X^2_{\min}(y)$ but with a different reference distribution: instead of a χ^2_{n-k}, the reference distribution is the difference between two independent χ^2 random variables, $\chi^2_n - \chi^2_k$. This implies that the posterior predictive p-value for the realized discrepancy, $X^2(y;\theta)$, is larger than that from $X^2_{\min}(y)$; the reference distribution of the former has a larger variance, $2(n+k)$ versus $2(n-k)$. Suppose, for example, $n = 250$, $k = 200$, and data y are observed for which $X^2_{\min}(y) = 80$. Under $X^2_{\min}(y)$, this is three standard deviations away from the mean of the χ^2_{50} reference distribution—an indication of lack of fit. The corresponding reference distribution for $X^2(y;\theta)$ is $\chi^2_{250} - \chi^2_{200}$, which has the same mean of 50 but with a larger standard deviation of 30, and thus the data do not appear to be a surprise at all.

How do we interpret this difference? The lack of fit under $X^2_{\min}$ shows that the data are not as close to the best fitting model, in terms of the sum of the

standardized residuals, as would be expected from a model with a large number of parameters. However, it is possible that this lack of fit will not adversely affect practical inferences from the data. In the example considered here, the realized discrepancy indicates that the data are reasonably close to what could be expected in replications under the hypothesized model. The extra 30 by which the minimum discrepancy exceeds its expectation seems large compared to 50 degrees of freedom but small when examined in the context of the 250-dimensional space of y. In the next section, we indicate the possibility of using the difference between the posterior predictive assessments from these two discrepancies to detect whether the lack of fit is due to the likelihood or due to the prior specification.

If prior knowledge of θ is added, as expressed by a nonuniform prior distribution, the posterior predictive p-value for $X^2_{\min}(y)$ is unchanged, since $X^2_{\min}(y)$ is still a pivotal quantity in the linear model case, but the assessments based on $X^2_{\text{avg}}(y)$ and $X^2(y;\theta)$ now change, as they are now measuring discrepancy from the prior model as well as the likelihood. Sensitivity to the prior distribution is discussed in Section 3 in the context of our applied examples, and further discussed in Section 4.2.

3. Illustration with Two Applied Examples

3.1. Fitting an increasing, convex mortality rate function

For a simple real-life example, we reanalyze the data of Broffitt (1988), who presents a problem in the estimation of mortality rates (Carlin (1992), provides another Bayesian analysis of these data). For each age, t, from 35 to 64 years, inclusive, Table 1 gives N_t, the number of people insured under a certain policy and y_t, the number of insured who died. (People who joined or left the policy in the middle of the year are counted as half.) We wish to estimate the mortality rate (probability of death) at each age, under the assumption that the rate is increasing and convex over the observed range. The observed mortality rates are shown in Figure 2 as a solid line. The observed deaths at each age, y_t, are assumed to follow independent binomial distributions, with rates equal to the unknown mortality rates, θ_t, and known population sizes, N_t (equivalently, we could consider the values N_t as random but treat them as auxiliary statistics). Because the population for each age was in the hundreds or thousands, and the rates were so low, we use the Poisson approximation for mathematical convenience: $P(y|\theta) \propto \prod_t \theta_t^{y_t} e^{-N_t\theta_t}$. An optimization routine was used to maximize this likelihood, under the constraint that the mortality rate be increasing and convex. The maximum likelihood fit is displayed as the dotted line in Figure 2. Having obtained an estimate, we would like to check its fit to the data. The obvious possible flaws of the model are the Poisson distribution and the assumed convexity.

Table 1. Mortality rate data from Broffitt (1988)

age, t	number insured, N_t	number of deaths, y_t	age, t	number insured, N_t	number of deaths, y_t
35	1771.5	3	50	1516.0	4
36	2126.5	1	51	1371.5	7
37	2743.5	3	52	1343.0	4
38	2766.0	2	53	1304.0	4
39	2463.0	2	54	1232.5	11
40	2368.0	4	55	1204.5	11
41	2310.0	4	56	1113.5	13
42	2306.5	7	57	1048.0	12
43	2059.5	5	58	1155.0	12
44	1917.0	2	59	1018.5	19
45	1931.0	8	60	945.0	12
46	1746.5	13	61	853.0	16
47	1580.0	8	62	750.0	12
48	1580.0	2	63	693.0	6
49	1467.5	7	64	594.0	10

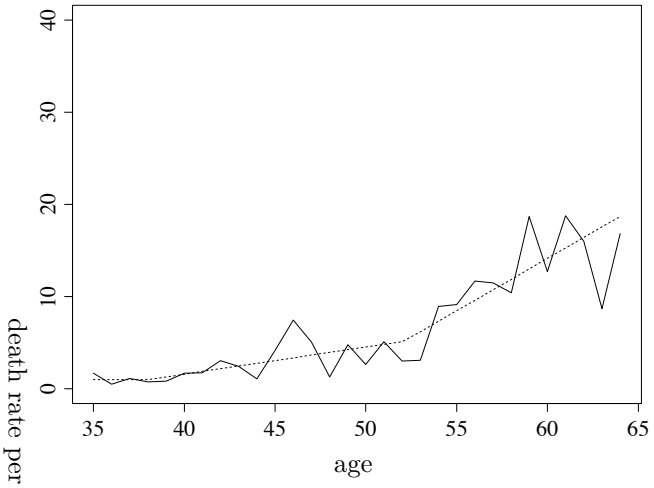

Figure 2. Observed mortality rates and the maximum likelihood estimate of the mortality rate function, under the constraint that it be increasing and convex.

The χ^2 discrepancy between the data and the maximum likelihood estimate is 30.0, and the minimum χ^2 discrepancy (using minimum χ^2 estimates in place of the MLE) is 29.3. These are based on 30 data points, with 30 parameters

being fit. There are obviously, however, less than 30 free parameters, because of the constraints implied by the assumption of increasing, convex rates. In fact, the MLE lies on the boundary of constraint space; the solution is essentially characterized by only four parameters, corresponding to the two endpoints and the two points of inflection of the best-fit increasing, convex curve in Figure 2. So perhaps a χ^2_{26} distribution is a reasonable approximation to the reference distribution for the minimum χ^2 statistic?

As a direct check, we can simulate the sampling distribution of the minimum χ^2 statistic, assuming $\theta = \hat{\theta}$, the MLE. The resulting distribution of $X^2_{\min}(y^{\text{rep}})$ is shown in Figure 3; it has a mean of 23.0 and a variance of 43.4 (by comparison, the mean and variance of a χ^2_{26} distribution are 26 and 52, respectively). The observed value, $X^2_{\min}(y) = 29.3$, is plotted as a vertical line in Figure 3; it corresponds to a tail-area probability of 16%. The distribution of Figure 3 is only an approximation, however, as the value of θ that generates the current data is unknown. In particular, we do not expect the true θ to lie exactly on the boundary of the constrained parameter space. Moving θ into the interior would lead to simulated data that would fit the constraints better, and thus the distribution of Figure 3 provides a conservative p-value for the minimum χ^2 discrepancy.

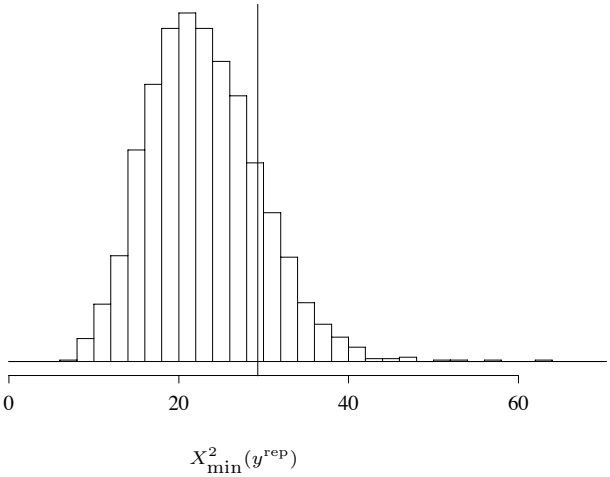

Figure 3. Histogram of 1000 simulations from the reference distribution for $X^2_{\min}(y^{\text{rep}})$, the minimum χ^2 statistic for the mortality rates: classical approximation with θ set to the maximum likelihood estimate. The vertical line represents $X^2_{\min}(y)$, the observed value of the minimum χ^2 statistic.

To conduct a Bayesian inference, we need to define a prior distribution for θ. Since we were willing to use the MLE, we use a uniform prior distribution, under

the constraint of increasing convexity. (The uniform distribution is also chosen here for simplicity; Broffitt (1988) and Carlin (1992) apply various forms of the gamma prior distribution.) Samples from the posterior distribution are generated by simulating a random walk through the space of permissible values of θ, using the Metropolis algorithm. Nine parallel sequences were simulated, three starting at the MLE and three at each of two crude extreme estimates of θ—one a linear function, the other a quadratic, chosen to loosely fit the raw data. Convergence of the simulations was monitored using the method of Gelman and Rubin (1992), with the iterations stopped after the within-sequence and total variances were roughly equal for all components of θ. Nine draws from the posterior distribution for θ, one from each of the simulated sequences, are plotted as dotted lines in Figure 4, with the MLE from Figure 2 displayed as a solid line for comparison.

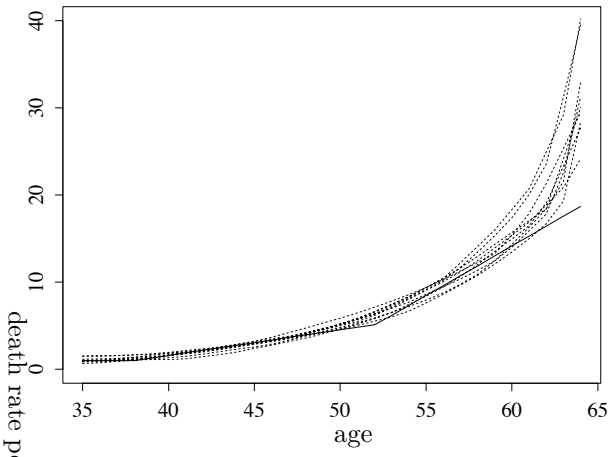

Figure 4. Nine draws from the posterior distribution of increasing, convex mortality rates, with the maximum likelihood estimate (solid line) as a comparison.

3.2. Posterior predictive assessment for the mortality rate example

To make a posterior predictive assessment of the fitness, it is necessary to define a reference set of replications; i.e., a set of "fixed features" in the notion of Rubin (1984). For this dataset, we defined replications in which the (observed) population size and (unobserved) mortality rates at each age stayed the same, with only the number of deaths varying, according to their assumed Poisson distributions. For each draw from the posterior distribution of θ, we simulated a replication; a random sample of nine replicated datasets corresponding to the previous nine draws of θ is plotted as dotted lines in Figure 5, with the observed frequencies from Figure 2 displayed as a solid line for comparison.

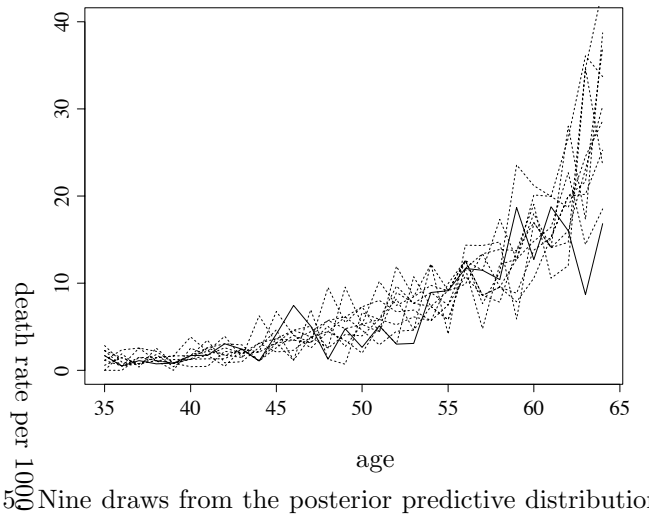

Figure 5. Nine draws from the posterior predictive distribution of mortality rates, corresponding to the nine draws of Figure 4, with the raw data (solid line) as a comparison.

The fitness of the model can be checked directly by examining a graph such as Figure 5—is the solid line an outlier in the forest of dotted lines?—or quantitatively, by defining a statistic such as y_{64}, the number of deaths at age 64, and comparing it to the distribution of simulated values of y_{64}^{rep}. We emphasize that whenever feasible, graphical assessment, including those corresponding to p-value calculations, should be made, because graphs provide the opportunity to directly inspect the magnitude of the various quantities, as well as helping to detect problems that may not be easily "visible" otherwise (e.g., Figure 5 shows that the lack of fit is much more prominent for later ages; we will return to this point shortly).

Checking residuals, especially in the form of the χ^2 discrepancy, is another standard practice for detecting lack of fit, and here we illustrate the minimum and realized χ^2 discrepancies discussed in Section 2.4. Besides illustrating the posterior predictive approach graphically, we also estimate the associated p-value as it provides a useful probability statement (when interpreted correctly) supplementing the graphical assessment. For each simulated replication, y^{rep}, the optimization routine was run to find the minimum χ^2 discrepancy, $X^2_{\min}(y^{\text{rep}})$. A histogram of these minimum χ^2 values—the reference distribution for $X^2_{\min}(y)$—is displayed in Figure 6. With a mean of 21.1 and a variance of 39.6, this posterior predictive reference distribution has lower values than the approximate distribution based on the MLE and displayed in Figure 3. The posterior predictive

p-value of the minimum χ^2 is 10%, which is more extreme than the maximum likelihood approximation, as expected.

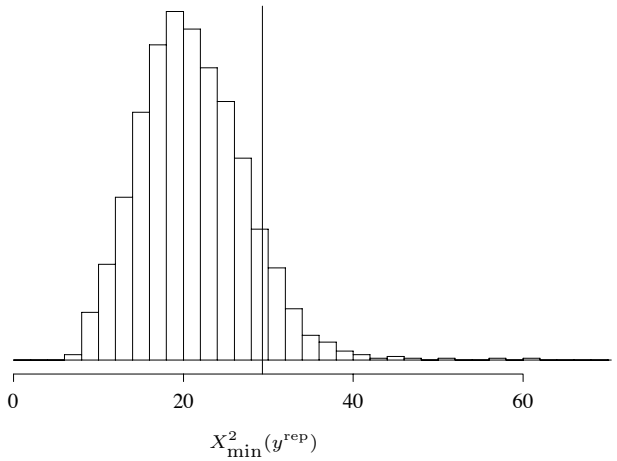

Figure 6. Histogram of 1000 simulations from the reference distribution for $X^2_{\min}(y^{\text{rep}})$, the minimum χ^2 statistic for the mortality rates, using the posterior predictive distribution. The vertical line represents $X^2_{\min}(y)$, the observed value of the minimum χ^2 statistic.

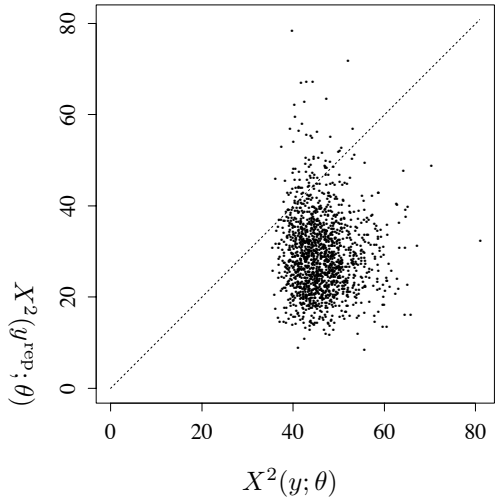

Figure 7. Scatterplot of predictive vs. realized χ^2 discrepancies for the mortality rates, under the joint posterior distribution; the p-value is estimated by the proportion of points above the 45° line.

For the realized discrepancy, Figure 7 shows a scatterplot of the realized discrepancy, $X^2(y;\theta)$ and the predictive discrepancy, $X^2(y^{\text{rep}};\theta)$, in which each point corresponds to a different value of (θ, y^{rep}) drawn from the posterior distribution, as described in Section 2.3. The tail-area probability of the realized discrepancy is just the probability that the predictive discrepancy exceeds the realized discrepancy, which in this case equals 6.3%, the proportion of points above the 45° line in the figure. The realized discrepancy p-value is more extreme than the minimum discrepancy p-value, which perhaps suggests that, given the constraint of increasing convexity, it is the uniform prior distribution, not necessarily the likelihood, that does not fit the data. (The analysis of linear models in Section 2.4 suggests that if the likelihood were the source of the poor fit, the minimum discrepancy assessment would give the more extreme tail-area probability. Assessing sources of lack of fit is an important topic that requires further investigation; related discussion is provided in Section 4.1.)

Even if we do not overhaul the model immediately, it is useful to note that the posterior predictive datasets were mostly higher than the observed data for the later ages (see Figure 5), and to consider this information when reformulating the model or setting a prior distribution for a similar new dataset. Even if the assumption of convex mortality rate is true in the natural population, it is very likely that the insurance company has screened out some high-risk older people, and thus destroys the convexity for later ages. In addition, there are general problems with using the uniform prior distribution for parameters constrained to be increasing, as discussed by Gelman (1996) in the context of this example.

3.3. Assessing fitness of a mixture model in psychology

Stern, Arcus, Kagan, Rubin, and Snidman (1995) fit a latent class model to the data from an infant temperament study. Ninety-three infants were scored on the degree of motor activity and crying to stimuli at 4 months and the degree of fear to unfamiliar stimuli at 14 months. Table 2 gives the data, y, in the form of a $4 \times 3 \times 3$ contingency table. The latent class model specifies that the population of infants is a mixture of relatively homogeneous subpopulations, within which the observed variables are independent of each other. The parameter vector, θ, includes the proportion of the population belonging to each mixture class and the multinomial probabilities that specify the distribution of the observed variables within a class. Psychological and physiological arguments suggest two to four classes for the mixture, with specific predictions about the nature of the infants in each class.

Table 3 summarizes the results of fitting one through four-class models using the EM algorithm. A discrepancy usually associated with contingency tables is the log likelihood ratio (with respect to the saturated model), $D(y;\theta) =$

$2\sum_i y_i \log[y_i/E(y_i|\theta)]$, where the sum is over the cells of the contingency table. The final column in Table 3 gives $D_{\min}(y)$ for each model. The two-class mixture model provides an adequate fit that does not appear to improve with additional classes. The maximum likelihood estimates of the parameters of the two-class model (not shown) indicate that the two classes correspond to two groups: the uninhibited children (low scores on all variables) and the inhibited children (high scores on all variables). It is well known that the usual asymptotic reference distribution for the likelihood ratio test (the χ^2 distribution) is not appropriate for mixture models (e.g., Titterington, Smith, and Makov (1985)).

Table 2. Infant temperament data

motor	cry	fear= 1	fear= 2	fear= 3
1	1	5	4	1
1	2	0	1	2
1	3	2	0	2
2	1	15	4	2
2	2	2	3	1
2	3	4	4	2
3	1	3	3	4
3	2	0	2	3
3	3	1	1	7
4	1	2	1	2
4	2	0	1	3
4	3	0	3	3

Table 3. Comparing latent class models for the data of Table 2

Model Description	Degrees of Freedom	$D_{\min}(y)$
Independence (= 1 class)	28	48.761
2 Latent Classes	20	14.150
3 Latent Classes	12	9.109
4 Latent Classes	4	4.718

At one level, this is a model selection problem (i.e., choosing the number of classes) for which a complete Bayesian analysis, incorporating the uncertainty in the number of classes, could be carried out. However, such an analysis is complicated by the fact that the parameters of the various probability models (e.g., the two- and four-class mixture models) are related, but not in a straightforward

manner. To be more explicit, theory suggests at least two fundamentally different groups of children, the inhibited and the uninhibited. Additional classes, if they are present, represent a bifurcation of one or both of these classes, and the parameters of such classes are related to the parameters of the two-class model in a way that is difficult to model explicitly. Given the small amount of data, we restrict our attention to only assessing the fit of the two-class model.

The prior distribution of the parameters of the latent class model is taken to be a product of independent Dirichlet distributions: one for the class proportions, and one for each set of multinomial parameters within a mixture class. The Dirichlet parameters were chosen so that the multinomial probabilities for a variable (e.g., motor activity) are centered around the values expected by the psychological theory but with large variance. The use of a weak but not uniform prior distribution helps identify the mixture classes (e.g., the first class of the two-class mixture specifies the uninhibited infants). With this prior distribution and the latent class model, draws from the posterior distribution are obtained using the data augmentation algorithm of Tanner and Wong (1987). Ten widely dispersed starting values were selected and the convergence of the simulations was monitored using the method of Gelman and Rubin (1992). The draws from the posterior distribution of the parameters for the two-class model were centered about the MLE. Rubin and Stern (1994) describe the prior distribution and resulting Bayesian analysis more fully. Their analysis of the data includes a posterior predictive evaluation of the likelihood ratio statistic for testing a one-class model versus a two-class model but does not directly address the fit of the model.

3.4. Posterior predictive assessment for the psychology example

To assess the quality of fit of the two-class model, we define replications of the data in which the parameters of the latent class model are the same as those responsible for the available data. These replications may be considered as data sets that would be expected if new samples of infants were to be selected from the same population. For each draw from the posterior distribution, a replicated data set y^{rep} was drawn according to the latent class sampling distribution. The reference distribution of the minimum discrepancy, $D_{\min}(y^{\text{rep}})$, based on 500 replications, is shown in Figure 8 with a vertical line indicating the observed value $D_{\min}(y)$. The mean of this distribution, 23.4, and the variance, 45.3, differ nontrivially from the χ^2_{20} distribution that would be expected if the usual asymptotic results applied. The posterior predictive p-value is 93% based on these replications.

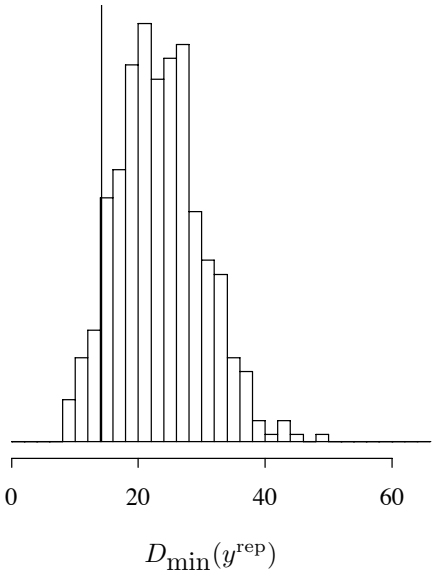

Figure 8. Histogram of 1000 simulations from the reference distribution for $D_{\min}(y^{\text{rep}})$, the log likelihood ratio statistic for the latent class example, using the posterior predictive distribution. The vertical line represents $D_{\min}(y)$, the observed value of the minimum discrepancy.

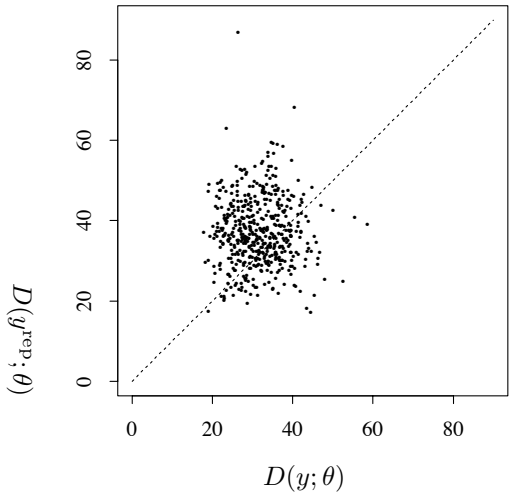

Figure 9. Scatterplot of predictive vs. realized log likelihood ratio discrepancies for the latent class model, under the joint posterior distribution; the p-value is estimated by the proportion of points above the 45° line.

Using $D_{\min}$ with mixture models is often problematic because of multimodality of the likelihood. For the data at hand, two modes of the two-class mixture likelihood were found, and for larger models the situation can be worse. Model assessments based on only a single mode, such as with $D_{\min}$, may ignore important information.

In contrast, the realized discrepancy, $D(y;\theta)$, uses the entire posterior distribution rather than a single mode. In addition, the realized discrepancy requires much less computation, as discussed in Section 2.3. Figure 9 is a scatterplot of the discrepancies for the observed data and for the replications under the two-class model. The p-value for the realized discrepancy is 74% based on 500 Monte Carlo draws. Thus, we have not obtained any evidence suggesting lack of fit of the two-class model. Again, this says nothing about the correctness of the model, only that the values of the discrepancy measures we have examined are reasonable under the posited model.

To indicate the type of result that we might obtain for a clearly inadequate model, we performed the same calculations for the one-class mixture model (equivalent to the independence model for the contingency table), yielding p-values for the minimum and realized discrepancies of 2.4% and 5.8%, respectively. Here the realized discrepancy gives the less extreme p-value, which seems to confirm that the lack of fit is due to the likelihood, not the prior distribution (as mentioned in Section 3.2, this assertion requires validation from further studies). At this point, a person who is using posterior predictive p-values in the Neyman-Pearson fashion might be troubled by the fact that the two p-values are on the opposite sides of the conventional cut-off point 5%. We emphasize again that we are merely using them as probability measures of lack of fit of various aspects of the posited model, which we know is at best a useful inferential approximation to the underlying true model. Nevertheless, empirical and theoretical studies so far do suggest that posterior predictive p-values generally have reasonable long-run frequentist properties. For example, the theoretical results on Type I error presented in Meng (1994) can be easily extended to replications involving auxiliary statistics, as presented in an early version of this paper, Gelman, Meng, and Stern (1992).

4. Making Choices

4.1. Differences among replications and comparison to the prior predictive distribution of Box (1980)

Choosing replications amounts to specifying the joint distribution, $P[y, \theta, y^{\text{rep}} \mid A(y^{\text{rep}}) = A(y)]$, from which all reference distributions can be derived by conditioning on y, with the auxiliary statistics being fixed at their observed values. Three different replications are illustrated in Figure 1. Figure 1a is

the posterior predictive replication, the focus of this paper, corresponding to repeating the experiment tomorrow with *the same* (*unknown*) *value of* θ that produced today's data, y. In contrast, Figure 1b shows the prior predictive replication, advocated by Box (1980), in which *new values of both* θ *and* y are assumed to occur tomorrow. Figure 1c shows a mixed predictive replication that can be useful with hierarchical models, where we may choose to assume that the same value of the hyperparameters α defines the replications but that new values of the parameters θ are assumed to occur. In practice, the choice of a model for y^{rep}, as well as the choice of auxiliary statistics, should depend on which hypothetical replications are of interest.

Unlike the posterior predictive replication, the prior predictive replication is undefined under improper prior distributions, which limits its use in practice. More generally, if the parameters θ are well-estimated from the data, posterior predictive assessments give results similar to classical procedures for reasonable prior distributions. In contrast, results of prior predictive assessments are sensitive to the prior distribution, even in analyses with a large amount of data in which the prior distribution is essentially irrelevant to posterior inferences.

As an illustration, we compare the prior and posterior predictive distributions for the following simple theoretical example. Consider 100 observations, $y_1, \ldots, y_n$, modeled as independent samples from a $N(\theta, 1)$ distribution with a diffuse prior distribution, say, $p(\theta) = \frac{1}{2A}$ for $\theta \in [-A, A]$ with some extremely large value of A, such as 10^5. We wish to check the model using, as a test statistic, $T(y) = \max_i |y_i|$: is the maximum absolute observed value consistent with the normal model? We choose this test statistic as an illustration of the kind of measure that might be used to identify outliers or extreme points. Consider a data set in which $\overline{y} = 5.1$ and $T(y) = 8.1$. To perform the posterior predictive check, we first determine the posterior distribution of θ, $N(5.1, 1/100)$, and then compute the posterior predictive distribution of $T(y^{\text{rep}})$ by simulation: for each posterior simulation of θ, we draw 100 values $y_i^{\text{rep}} \sim N(\theta, 1)$ and compute the maximum of their absolute values. Figure 10 displays a histogram of 1000 values of $T(y^{\text{rep}})$; 149 of the 1000 values are greater than 8.1, giving an estimated p-value of 15%. Thus the observed $T(y)$ is larger than usual under the model, but not surprisingly so.

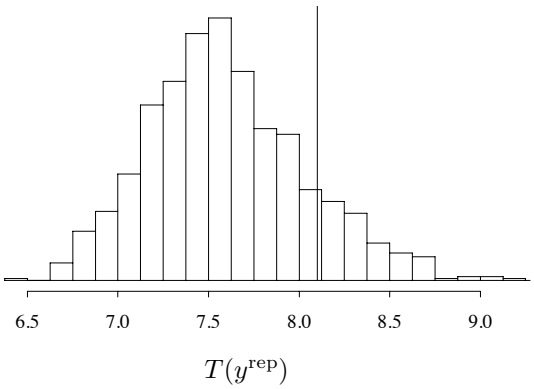

Figure 10. Histogram of 1000 simulations from the posterior predictive distribution for $T(y^{\text{rep}})$, the maximum absolute observed value for the hypothetical normal example. The vertical line represents $T(y)$, the observed value of the test statistic.

In contrast, the prior predictive distribution of $T(y)$, given the diffuse uniform prior distribution on θ, is spread roughly uniformly from 0 to A (with some nonuniformity at the boundaries). For a large value of A, the observed $T(y) = 8.1$ is in the extreme left tail of the distribution, with a p-value close to 1 (or, equivalently, a p-value close to 0 for the test statistic $-T(y)$). The difference between the posterior and prior predictive replications is that the former treats the prior distribution as an outmoded first guess, whereas the latter treats the prior distribution as a true "population distribution." It is commonly useful in Bayesian analysis to use a diffuse prior distribution for parameters in a well-understood model with the understanding that posterior inferences will be reasonable. (For this particular example, the prior and posterior checks could be made more similar by choosing a test statistic that is approximately pivotal, such as $T(y) = \max_i |y_i| - |\overline{y}|$, but our general desire is to be able to check the fit of the model for *any* function of data and parameters, not just pivotal quantities.)

4.2. The impact of prior distributions

As we stated in the previous section, in the posterior predictive framework, the prior distribution for the parameters of the model need not be especially accurate (and it is often not accurate in routine practice), as long as the posterior predictions are "near" the data. This relates to the observation that Bayesian methods based on convenient prior models (normality, uniformity, etc.) can often yield useful inferences in practice.

To check the impact of prior distributions, we calculated posterior predictive p-values for the mixture example using one- and two-class models under a variety of prior distributions. Two properties of the prior distribution were varied. The

center of each class of the prior distribution was chosen either to match the values suggested by the psychological theory, or to represent a uniform distribution over the levels of each multinomial variable. The strength of the prior information was also varied (by changing the scale of the Dirichlet distributions as measured by the sum of the Dirichlet parameters). As long as the prior distributions are not particularly strong, the size of the p-values and the conclusions reached remained essentially unchanged. This was true for p-values based on both minimum and realized discrepancies.

If the prior distribution is strongly informative, however, it affects the tail-area probabilities of different discrepancies in different ways. Realized discrepancies are naturally quite sensitive to such prior distributions. The posterior predictions obtained under strong incorrect prior specifications may be quite far from the observed data. For example, in Section 3.3, a strong prior distribution specifying two mixture classes, but not corresponding to inhibited and uninhibited children, leads to a tail-area probability of essentially zero. By comparison, minimum discrepancies are much less sensitive to the prior distribution, because the original dataset is judged relative to the best-fitting parameter value rather than to the entire posterior distribution. Conversely, a strong prior distribution, if trusted, can be used to more effectively assess the fit of the likelihood part of the model.

4.3. Choosing discrepancies

In the traditional framework of testing the correctness of a null hypothesis, the choices of test statistics are strongly influenced by the consideration of power, which is necessarily governed by the specifications of alternative hypotheses. The reason for preferring a more powerful test is to increase the chance of rejecting a null hypothesis when it is wrong. In the context of checking a whole model (in contrast to testing a few parameters within a model), we know that virtually all models are wrong, and thus a more relevant focus is how the model fits in aspects that are important for our problems at hand. Therefore, the choice of discrepancies should reflect our inferential interests, as well as some standard checks on overall fitness (such as those measured by the χ^2 discrepancy). For example, Gelman, Carlin, Stern, and Rubin (1995), §18.3, check the fit of a model used to estimate a population total from a simple random sample; they use the sample total as a test statistic and find poor fits under some reasonable-looking models (such as the log-normal and power-transformed normal). They find that models that poorly fit the sample total yield poor inferences for the population total, although the same models can perform excellently for estimating the population median. We also note that there is no need of adjusting for multiple comparisons

when we use more than one discrepancy because here we are merely assessing the fitness of a model from various perspectives.

We emphasize that the posterior predictive approach is suitable for assessing the fitness of a *single* model (which could be a super-model that incorporates various alternative models, as in Leamer (1978), Madigan and Raftery (1994), and Draper (1995)) to the available data. As we illustrate here and elsewhere (e.g., Gelman, Carlin, Stern, and Rubin (1995)), it is entirely possible to construct sensible discrepancies to detect the lack of fit of a single model, in the absence of *explicit* alternative models. We disagree with the opinion that one should never reject a model unless there is an available alternative. A consumer surely can refuse to buy a defective product, even if it is the only item available. If the consumer does decide to purchase it despite its deficiencies, he or she surely still would like to know about its defects and their possible consequences. The posterior predictive assessment provides an (imperfect) method for detecting the "defective products" of applied statistics—invalid models. Whenever we choose to work with a model, we should be aware of its deficiencies, always being aware of the difference between practical significance and statistical significance. We should always report any known defects of our chosen models, so as not to mislead others who might be inclined to use the same models, as they may use the model for purposes for which the model defects may be far more damaging. Indeed, Bayesian inference is a powerful tool for learning about model defects, because we have the ability to examine, as a discrepancy measure, any function of data and parameters.

Acknowledgements

We thank Art Dempster, Donald Rubin, George Tiao, and Michael Stein for comments and criticism, Giuseppe Russo for the mortality rate example, Jerome Kagan, Nancy Snidman, and Doreen Arcus for the infant temperament study data, and the National Science Foundation for grants DMS-9204504, DMS-9404305, DMS-9457824, DMS-9404479, and DMS-9505043. We also appreciate suggestions from many reviewers that led to a more concise presentation.

References

Belin, T. R., and Rubin, D. B. (1995). The analysis of repeated-measures data on schizophrenic reaction times using mixture models. *Statistics in Medicine* **14**, 747-768.

Berkson, J. (1980). Minimum chi-square, not maximum likelihood (with discussion). *Ann. Statist.* **8**, 457-487.

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

Broffitt, J. D. (1988). Increasing and increasing convex Bayesian graduation. *Transactions of the Society of Actuaries* **40**, 115-148.

Carlin, B. P. (1992). A simple Monte Carlo approach to Bayesian graduation. *Transactions of the Society of Actuaries* **44**, 55-76.

Chaloner, K. (1991). Bayesian residual analysis in the presence of censoring. *Biometrika* **78**, 637-644.

Chaloner, K. and Brant, R. (1988). A Bayesian approach to outlier detection and residual analysis. *Biometrika* **75**, 651-659.

Chernoff, H. (1954). On the distribution of the likelihood ratio. *Ann. Math. Statist.* **25**, 573-578.

Cochran, W. G. (1952). The χ^2 test of goodness of fit. *Ann. Math. Statist.* **23**, 315-345.

Draper, D. (1995). Assessment and propagation of model uncertainty (with discussion). *J. Roy. Statist. Soc. Ser.B* **57**, 45-97.

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, (Edited by J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), 147-167, Oxford University Press.

Gelman, A. (1990). Topics in image reconstruction for emission tomography. Ph.D. thesis, Dept. of Statistics, Harvard University.

Gelman, A. (1992). Statistical analysis of a medical imaging experiment. Tech. Report #349, Dept. of Statistics, University of California, Berkeley.

Gelman, A. (1996). Bayesian model-building by pure thought: some principles and examples. *Statist. Sinica* **6**, 215-232.

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

Gelman, A. and Meng, X. L. (1996). Model checking and model improvement. In *Practical Markov Chain Monte Carlo*, (Edited by W. Gilks, S. Richardson, and D. Spiegelhalter), 189-201. Chapman and Hall, London.

Gelman, A., Meng. X. L. and Stern, H. S. (1992). Bayesian model checking using tail area probabilities. Tech. Report #355, Dept. of Statistics, University of Chicago.

Gelman, A. and Rubin, D. B. (1992). Inferences from iterative simulation using multiple sequences (with discussion). *Statist. Sci.* **7**, 457-511.

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

Leamer, E. E. (1978). *Specification Searches: Ad Hoc Inference with Nonexperimental Data*. John Wiley, New York.

Madigan, D. and Raftery, A. E. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. *J. Amer. Statist. Assoc.* **89**, 1535-1546.

McCullagh, P. (1985). On the asymptotic distribution of Pearson's statistic in linear exponential-family models. *Internat. Statist. Rev.* **53**, 61-67.

McCullagh, P. (1986). The conditional distribution of goodness-of-fit statistics for discrete data. *J. Amer. Statist. Assoc.* **81**, 104-107.

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

Rubin, D. B. (1981). Estimation in parallel randomized experiments. *J. Educ. Statist.* **6**, 377-400.

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

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

Rubin, D. B. and Stern, H. S. (1994). Testing in latent class models using a posterior predictive check distribution. In *Latent Variables Analysis: Applications for Developmental Research*, (Edited by A. von Eye and C. Clogg), 420-438.

Stern, H. S., Arcus, D., Kagan, J., Rubin, D. B. and Snidman, N. (1995). Using mixture models in temperament research. *Internat. J. Behav. Devel* **18**, 407-423.

Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation (with discussion). *J. Amer. Statist. Assoc.* **82**, 528-550.

Tiao, G. C. and Xu, D. (1993). Robustness of maximum likelihood estimates for multi-step predictions: the exponential smoothing case. *Biometrika* **80**, 623-641.

Titterington, D. M., Smith, A. F. M. and Makov, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*. John Wiley, New York.

Tsay, R. S. (1992). Model checking via parametric bootstraps in time series analysis. *Appl. Statist.* **41**, 1-15.

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

Upadhyay, S. K. and Smith, A. F. M. (1993). A Bayesian approach to model comparison in reliability via predictive simulation. Tech. Report, Dept. of Mathematics, Imperial College, London.

Wermuth, N. and Lauritzen, S. L. (1990). On substantive research hypotheses, conditional independence graphs and graphical chain models. *J. Roy. Statist. Soc. Ser.B* **52**, 21-50.

West, M. (1986). Bayesian model monitoring. *J. Roy. Statist. Soc. Ser.B* **48**, 70-78.

Zellner, A. (1975). Bayesian analysis of regression error terms. *J. Amer. Statist. Assoc.* **70**, 138-144.

Department of Statistics, Columbia University, New York, NY 10027, U.S.A.

Department of Statistics, University of Chicago, Chicago, IL 60637, U.S.A.

Department of Statistics, Iowa State University, Ames, IA 50011, U.S.A.

(Received June 1994; accepted March 1996)

COMMENT : UTILITY, SENSITIVITY ANALYSIS, AND CROSS-VALIDATION IN BAYESIAN MODEL-CHECKING

David Draper

University of Bath

It might be best discussing this interesting paper to step back a bit and see where its message fits into the overall picture of Bayesian (and frequentist) modeling. I find it useful in my own applied work to think like a de-Finetti-style Bayesian when formulating inferences and predictions, and to think like a frequentist when evaluating their quality (by paying attention to calibrative summaries of discrepancies between actual observables and predictive distributions for those observables), and it is in the spirit of this sort of attempted fusion of coherence and calibration that I make these comments.

In the de Finetti approach you begin with observable (and already observed) data y, and wish on the basis of y and assumptions $\mathcal{A}$–about conditional exchangeability relationships, background beliefs about how the world works, and contextual information about how the data were gathered–to assess uncertainty about observables y^* not yet seen, through predictive distributions $p(y^*|y, \mathcal{A})$. Inference about unobservable parameters θ is often expressible in this framework by telling hypothetical stories about quantities that emerge from the conditional exchangeability judgments. In the mortality example of Gelman, Meng and Stern (1996), GMS, for instance, you start with the number N_t of people carrying a particular insurance policy and the number y_t of those people who died in a particular (unspecified) time period, for ages t from 35 to 64. Since mortality generally increases as people get older, it might well be reasonable to model your uncertainty about the implied person-level death indicators not by taking these indicators as unconditionally exchangeable, but instead as conditionally exchangeable given age (as in GMS). This leads through the usual device of de Finetti's theorem for 0/1 outcomes to the positing of an "underlying death rate" θ_t for each age, interpretable as the observed death rate of a hypothetical infinite sample of policy-holders of age t, and the requirement, to achieve at least approximate coherence, that you adopt the model $\theta = (\theta_{35}, \ldots, \theta_{64}) \sim p(\theta)$, $(y_t|\theta_t) \overset{\text{indep}}{\sim} \text{bin}(N_t, \theta_t)$ for some prior distribution $p(\theta)$ which ought to reflect the presumed monotone relationship between mortality and age.

Thus in this problem the task of (initial) model selection–which in general may be regarded as the choice of a subset of the space $\mathcal{M} = \{\text{all possible models for the observables}\}$ on which to place nonzero prior probability–has reduced to the specification of a prior $p(\theta_t, t = 35, \ldots, 64)$ on 30 quantities which must live between 0 and 1 and which ought perhaps to be nondecreasing in t. Once such a prior has been specified, coherence has taken us as far as it can, without addressing the calibrative concern that the resulting model, however reasonable in appearance, may not accord well with the observables. This is where GMS come in: their contribution is to further develop Rubin's (1984) work on posterior predictive assessment, as one possible way of measuring any discrepancies that may exist between the observables and their predictive distributions under a given model specification.

That there is a pressing need to perform such calibrative work with any posited model has become even clearer at a time when Markov Chain Monte Carlo (MCMC) methods permit the realization of de Finetti's modeling approach in much more complicated settings than ever before. (In fact, Bayesians are dwelling so much these days on another category of diagnostics–MCMC convergence-monitoring–that a strong reminder of the value of model-checking is all the more welcome.) The model-specification problem would have been

even worse in GMS's insurance example had the observables y_t of interest been continuous instead of dichotomous: in that case, with the same conditional exchangeability assumptions as before, de Finetti tells us that for coherence you would have to posit the existence of an "underlying cdf" F_t, interpretable as the empirical cdf for the outcome y_t based on a hypothetical infinite sample of policyholders of age t, and adopt the model $F = (F_{35}, \ldots, F_{64}) \sim p(F), (y_t|F_t) \overset{\text{indep}}{\sim} F_t$ for some prior distribution $p(F)$ on $\mathcal{F}^{30}$, where $\mathcal{F}$ is *the space of all possible cdfs on the real line*. Nobody knows how to specify such a prior in a truly rich way. Instead (Draper (1995a)) people typically cheat, by looking at things like probability plots of the data in each age category; if, e.g., the data look roughly lognormal, you might replace the infinite-dimensional prior specification problem on $F_t \in \mathcal{F}$ by a two-dimensional Gaussian specification, as in $(\mu_t, \sigma_t) \sim p(\mu_t, \sigma_t), (\log y_t|\mu_t, \sigma_t) \sim N(\mu_t, \sigma_t^2)$. With the circular, use-the-data-twice character of this approach to model-building, calibrative checks of the quality of a model's predictions on new data not used in the model-selection process become even more important.

However, having made the case for model-checking, it does not necessarily follow that GMS's predictive p-values based on omnibus "discrepancy variables"–in their examples, χ^2–style summaries–should be the way to do it. Here are a few comments on GMS's method and some thoughts on other ways to proceed.

- In their examples GMS do not share with us their approach to the entire model-building process–from initial specification, through modification when deficiencies are identified, to final model acceptance–but I suspect from their writings elsewhere that they agree broadly with the model-expansion strategy exemplified by Box and Tiao (1962), Draper (1995a), and others: start with something reasonable based on context, check it in relevant ways, expand it hierarchically (either continuously or discretely) in directions indicated by any deficiencies found (note that the averaging over competing models implied by this expansion will be guided numerically by Bayes factors, not p-values), and stop when it passes all relevant checks. But how is the word "relevant" in this strategy to be made operational? To put it another way, the point of model-checking is to see if the current front-running model is good enough; but if you buy into GMS's framework, how do you know if a particular predictive p-value is small enough to render the current model unacceptable? More generally and more importantly, the model should indeed be good enough, but good enough for what purpose?

  Fully satisfying answers to these questions must include an attempt to quantify the **utilities** involved in taking the available actions (e.g., simplify the current model, leave it alone, expand it), and I suspect that in many (most?) cases you would not be led to p-values at all if you went straight to utility (e.g.,

the lack of fit for high ages in GMS's Fig. 2 would be fatal for some purposes to which the monotone model might be put but unimportant for others, and how do p-values help you make that distinction?). If the explicit goal is real-world decision-making, then model-checking must be driven by utilities elicited from the structure of the problem; if the short-term goal is scientific summary, Spiegelhalter (1995) has sketched a partial utility solution, by using Bernardo's (1979) results on the relationship between information criteria and scientific utility to suggest when a simple model in an expanded class of models may suffice. This tells us when to simplify, but not when (and how) to expand the current model, and when to stop expanding. Hodges (1987) must certainly be right that the entire modeling process should be thought of as a constrained optimization problem over a set of resources that includes the analyst's time–which if taken seriously will tell you when to stop–but I am not aware of any serious applications of this idea yet.

The punchline, often noted elsewhere (e.g., Lindley (1968)) but more frequently honored in the breach than in the observance, is that utility is difficult but cannot be ignored. It is ironic that GMS bring up the difference between practical and statistical significance in the last three sentences of the paper. Their methods are solely about the latter; if they had addressed the former they would have had to come to grips with utility.

- GMS assert in Section 4.3 that "... it is entirely possible to construct sensible discrepancy variables to detect the lack of fit of a single model, in the absence of *explicit* alternative models," landing them squarely in a kettle of soup that has been brewing since the arguments between Neyman and Fisher in the 1920s over the role of the alternative hypothesis in significance testing. It is true that their method makes no explicit appeal to an alternative model, but how do they recommend that people choose discrepancy measures in practice? For example, why did they focus on y_{64} in Section 3.2? Because they had an alternative in mind in which the underlying mortality rates for policy-holders were not in fact monotone.

  Every choice of discrepancy measure implicitly specifies an alternative to at least some extent (this is even true of omnibus discrepancies like χ^2, because in any given problem you can usually think of two or more omnibus-style measures which may yield different strengths of evidence against the current model). It is not possible to avoid alternative models entirely; the thing to do is to use **sensitivity analysis** to see if variation across the plausible alternatives is modest, on the inferential and predictive scales of practical significance–if so, expand the model hierarchically; if not, report the results of the sensitivity analysis and stop.

- It is notable that GMS's Figs. 2, 4 and 5 provide much more insight in the mortality example than their p-values and the plots that go with them. In thinking about this example I was led to produce full-posterior inferential plots (Figs. 1 and 2) as a first step in assessing model inadequacy, by (1) choosing a prior for $(\theta_{35}, \ldots, \theta_{64})$ in which each component was $U(0, 0.05)$ subject to the constraint that $\theta_{30} \leq \theta_{31} \leq \cdots \leq \theta_{64}$, (2) using Gibbs sampling to simulate from the joint posterior for all 30 components of θ, (3) making perspective and percentile plots of this joint posterior, and (4) superimposing the observed mortality rates with two-standard-error traces either way. (In this modeling I have not enforced a convexity constraint as GMS do, but it is evident from Fig. 2 that convexity does not have much effect in this case. Incidentally, I did not find it necessary to use an MCMC convergence-monitoring strategy that was anywhere near as complicated as the Gelman-Rubin approach used by GMS–a burn-in of 1K–5K, with starting values obtained from a regression of y_t/N_t on $(t - \bar{t})$ and $(t - \bar{t})^2$, followed by a single run of 5K–10K produced results in good agreement with overdispersed strategies.) The uncertainty bands shown in Fig. 2 are inferential (about the underlying rates) rather than predictive (about future observed counts) in character, but even without the addition of predictive uncertainty (see below) it is clear that something funny is going on in the high age groups. Model expansion in this example might involve telling a selection-bias story like the one mentioned by GMS, in which the underlying death rates in the population at large are monotone with age but the insurance company is skimming off the healthier people as policy-holders; in this modeling the N_t would no longer be ancillary (e.g., a comparison of the prevalences by age $(N_t/\sum N_i)$ in this data set with known population prevalence would reveal a systematic tendency of the insurance company to underinsure old people in relation to their prevalence).
- In model-checking it does seem quite reasonable, as GMS advocate, to think predictively about $\mathcal{D} = \{$data sets you could get in the future if your current model is "right" and you repeated the data-gathering activity$\}$ and to ask where the present data set sits in $\mathcal{D}$ as measured by some distance d in dataset-space, thereby inducing a density f of d values and locating d_{obs} in this distribution, but why summarize the unusualness of d_{obs} with a tail-area? This requires a stronger defense than the appeal GMS make to frequentist familiarity. In the absence of a serious attempt to quantify utility, of the type outlined above, it seems more natural to me to calculate a summary more like a Bayes factor or likelihood ratio, e.g., $f_{\max}/f(d_{\text{obs}})$. In the ad-hoc world created by abandoning utility this is arguably just as arbitrary as a tail-area for judging unusualness, but the p-value scale has become ossified at cutpoints that tend to exaggerate the evidence against the null in relation to

Bayes-factor-like summaries (Berger and Delampady (1987)), which makes its perpetuation well worth avoiding.

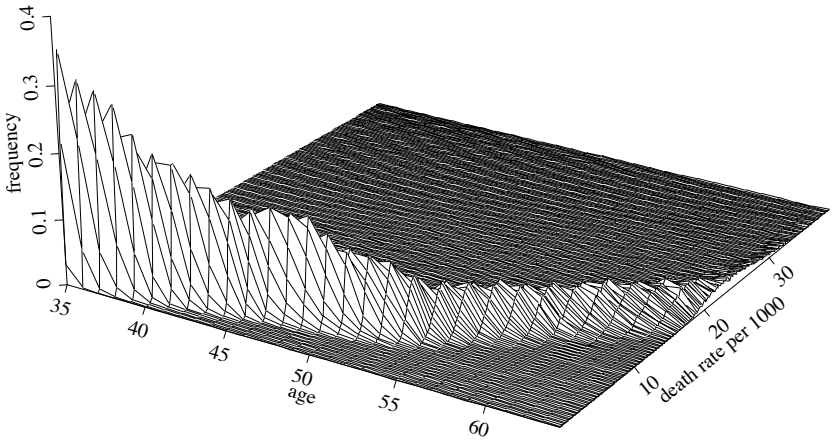

Figure 1. Perspective plot of the marginal posteriors for $(\theta_{35}, \ldots, \theta_{64})$ under the assumption of monotonicity of the θ_t in the mortality example.

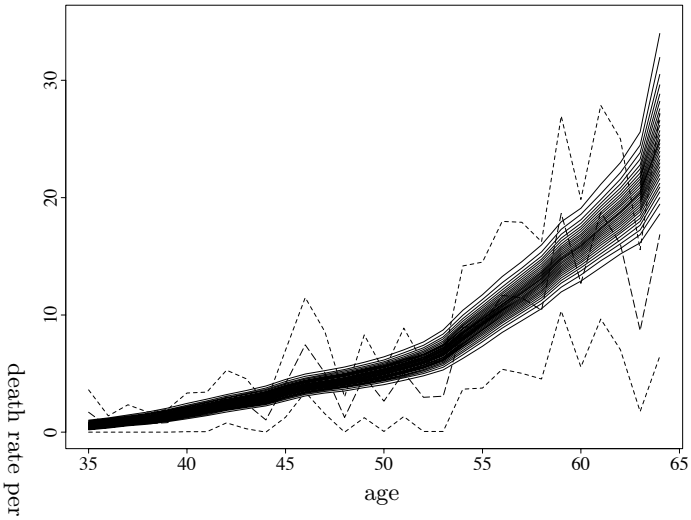

Figure 2. Plot of the $100(1/k, \ldots, (k-1)/k)$-th percentiles (for $k = 30$) of the same joint posterior as in Figure 1, with the observed mortality rates, and inferential uncertainty bands two standard errors either way, superimposed.

- As an alternative to GMS's version of posterior predictive assessment of χ^2

goodness-of-fit measures, it seems more natural to me to use predictive distributions in a somewhat different way, by **cross-validating** the modeling process: setting aside some of your data, fitting the current model to the rest, and then locating the observed values of the set-aside data in their respective predictive distributions given the chosen model (for an example of this using predictive z-scores, see Draper (1995b)). This differs from GMS's prescription in two ways: it uses a non-omnibus discrepancy measure that can better highlight which parts of the data set are badly fit by the current model, and it avoids the diagnostic overfitting problem inherent in, e.g., ordinary least-squares residuals from Gaussian linear models in relation to externally studentized residuals, in which the modeling that produces the ordinary residuals tends to over-accommodate the very observations whose unusualness is at issue.

In the mortality example, for instance, the two categories exhibiting the greatest relative discrepancy with the monotone model in Figure 2 are ages 46 and 63. Setting aside these age groups one by one, fitting the nondecreasing model to the rest of the data, and simulating from the predictive distribution for the set-aside y_t given its observed N_t produces in this case approximate beta-binomial distributions with parameters (α, β, N) of $(47, 13137, 1747)$ and $(37, 1464, 693)$ for $t = 46$ and 63, respectively. The observed values of $f_{\max}/f(d_{\text{obs}})$ from these predictive distributions come out about 19 to 1 at age 46 and 27 to 1 at age 63 (the corresponding one-tailed p-values are 1 in 65 and 1 in 204, consistent with Berger and Delampady's message). If instead I judge the unusualness of y_{63} by comparing it with its predictive distribution based on the entire data set, it looks substantially less unusual, because the fitting process has made substantial use of y_{63} to produce the monotone posterior for the entire θ vector: the approximate beta-binomial now has parameters (61, 2954, 693), and $f_{\max}/f(d_{\text{obs}})$ drops from 27 to 1 down to about 9 to 1.

Without guidance from utility I am at the usual disadvantage of not knowing precisely what ratios like 19 to 1 and 27 to 1 imply behaviorally, but informally both values indicate a lack of fit that would be worth investigating. There is no obvious substantive story to explain the discrepancy for $t = 46$, as there was for $t = 63$, but with 30 age categories in which comparisons of this type may be made, one of them is perhaps entitled to exhibit this level of misbehavior "for no good reason."

- I am not convinced by GMS's examples or intuitive discussion that their approach can diagnostically separate mis-specification of the prior and the likelihood. It would seem better to treat this as part of the overall sensitivity analysis described above–hold the prior (say) constant, change the likelihood

across a broad plausible range, summarize the effects of this change on relevant aspects of the posterior distribution, and then repeat these steps reversing the role of prior and likelihood. It will become clear by doing this which aspects of the current model are most in need of modifying.

School of Mathematical Sciences, University of Bath, Bath BA2 7AY, England.

COMMENT

Bruce M. Hill

University of Michigan

This is an interesting and thoughtful article concerning an important problem, model assessment, that despite its long history, going back at least to Daniel Bernoulli's celebrated analysis of the planetary orbits, is still largely unsolved and controversial. In a certain class of such problems the authors argue that the posterior predictive p-value can make a substantial improvement over both the classical p-value and the prior predictive p-value, and in this I believe they are correct. On the other hand, it can be argued that all the p-values are of limited use for model selection in important decision problems where there are serious alternative models in mind, and especially for prediction of future observations. This is true even if such alternatives are only implicit, as when a test statistic is chosen to achieve power against specific alternatives. In my opinion, only the full-blown Jeffreys-Savage (JS) theory of hypothesis testing, as in Jeffreys (1961), Savage (1962) and Hill (1993), can cope with such more demanding problems.

Like many others, I have come to regard the classical p-value as a useful diagnostic device, particularly in screening large numbers of possibly meaningful treatment comparisons. It is one of many ways quickly to alert oneself to some of the important features of a data set. However, in my opinion it is not particularly suited for careful decision-making in serious problems, or even for hypothesis testing. Its primary function is to alert one to the need for making such a more careful analysis, and perhaps to search for better models. Whether one wishes actually to go beyond the p-value depends upon, among other things, the importance of the problem, whether the quality of the data and information about the model and a priori distributions is sufficiently high for such an analysis to be worthwhile, and ultimately upon the perceived utility of such an analysis.

The main defect of the p-value is that it fails to consider, explicitly, what happens under any alternative model, and thus can at best only suggest in a

crude way that some conventional (or null) model may be faulty. Sometimes the only alternative currently taken seriously is even more faulty, and in this case a small p-value may also suggest the need to look for better models. This is more than a mere possibility, and there are important problems in which this frequently occurs (such as the variance component problems discussed below), where the data fit the conventional model only too well, and where there may be unknown constraints upon the data, including intentional or unintentional cheating.

The predictive p-value, in either its posterior or prior form, can also be of diagnostic value, but with exactly the same qualifications. There are four different procedures that I think should be compared: the classical p-value, the posterior predictive p-value, the prior predictive p-value, and the JS theory of hypothesis testing, in which the posterior probability of the conventional model plays a role similar to that of the various p-values. As I see it the first three p-values can have an important role to play in cases where the JS approach is highly non-robust to the input of prior distributions for the parameters specified by the various models, and particularly for such prior distributions under alternative models. The first two p values can then yield a relatively simple and possibly robust analysis, which is quick and undemanding, and sometimes yield a good approximation to the JS approach.

The posterior predictive p-value is closely related to its classical counterpart. If one wishes to assess the suitability of a model that has unknown parameters, and if it were the case that the parameter value was logically independent of the model, then it would presumably be best to make the test (or analysis) using the true value of the parameter. I like the examples of the authors, particularly the tomography example, but it is relatively complicated, and perhaps a simpler example will help focus on the substantive issues concerning the four procedures. Suppose the question concerns whether errors of measurement for a fixed physically meaningful parameter θ are approximately Gaussian, versus various models involving symmetrical long-tailed distributions (including the Cauchy) for the errors. Suppose that the parameter θ is the center of symmetry for the distribution under each of the models, and that the meaning and value of the parameter is independent of the model being tested. If we knew the true value of θ then it would be preferable to make such a test using the true value, because this would allow the measure of discrepancy from the conventional (Gaussian) model to be uncontaminated by the problem of estimating the parameter. But given the data, if the conventional model were true then the posterior distribution for θ would best represent current knowledge about the parameter under that model, and so for the same reason a test based upon this posterior distribution would appear to be preferable. This appears to be the basic rationale for the posterior predictive p-value. However the matter is substantially more subtle than might first appear,

and in particular, one must also consider the question from the viewpoint of the way in which the model is to be employed.

A major defect of the classical view of hypothesis testing, as emphasized by the authors, is that it attempts to test only whether the model is true. This came out of the tradition in physics, where models such as Newtonian mechanics, the gas laws, fluid dynamics, and so on, come so close to being "true" in the sense of fitting (much of) the data, that one tends to neglect issues about the use of the model. However, in typical statistical problems (especially in the biological and social sciences but not exclusively so) one is almost certain a priori that the model taken literally is false in a non-trivial way, and so one is instead concerned whether the magnitude of discrepancies is sufficiently small so that the model can be employed for some specific purpose. See Hill (1980) for a proposal of this sort in connection with the variance components problem.

Consider the contemplated launching of a space shuttle. Given the conditions at the planned time of launching, what is a reasonable evaluation of the posterior (or post-data) probability for success? This might appear to be quite high, based upon past experience with regard to previous launchings. But one might also have a model relating the ambient conditions (not completely known and perhaps varying rapidly at the current time) to certain other parameters of the model (also not fully known) which in turn might, based partly upon past experience, be known to have negative implications for success. For example, it might be known that a complicated interaction of temperature with other variables does greatly lower the probability of a successful launch. In this case it is necessary to assess, as best one can, the available information about such parameters and covariates and ancillary variables, including models for their effect upon the probability of success. A serious Bayesian does so by means of prior, posterior, and even post-data distributions for the variables that are thought most to influence the chance of success.

One is here then not concerned so much with the truth of the model, as with a careful assessment of the chance of success, including as realistically as possible all the major factors believed to influence this chance. The computations would often, in part, be made on the basis of an assumed model or perhaps models. Of course one uses the best models one knows of, but these may not be good enough, and one may even know that the current models are not likely to be good enough. Data that was indicative that the model was defective in a serious way, or even not firmly established for the current conditions, would suggest the conservative wisdom of delaying the launch until the conditions were those in which it was believed there was a high probability of success. For example, if the model involved an assumption about how certain components would perform in a certain temperature range, that was believed valid based upon past experience, and if the current temperature was outside that range, or even near the borderline

of that range, then the model that said the components would function at the current temperature would be seriously in doubt. This illustrates what I believe is an especially important type of decision problem, and one for which I think the posterior p-value would be useful at best only as a preliminary diagnostic device.

In my example of long-tailed distributions, suppose that some important decision such as whether to lauch a space shuttle, or whether to perform a delicate operation, might hinge upon whether the parameter θ has approximately some specified value θ_0 for which past experience suggests a high probability of success of the undertaking, conditional upon that value. A conservative view would be that in order to launch the shuttle, or perform the operation, the parameter must not only be close to θ_0, but one must essentially know that this is in fact the case. That is, there must be convincing evidence that the parameter is close to the target value. Now the only way I know seriously to discuss this issue is in terms of the posterior distribution, or more generally, the post-data distribution for θ. In Hill (1990) a theory of Bayesian data-analysis was proposed in which a distinction between posterior and post-data distributions was made. The distinction concerns whether or not, on the basis of data-analysis, one alters a pre-specified model, or perhaps even finds a new model. In complicated real-world problems a posterior distribution is often not accessible solely by means of Bayes's theorem, and in this case it is best to refer to probability distributions that critically depend upon the data-analysis as post-data distributions. Posterior distributions in the sense of Bayes are the ideal but may not be available. This would be true if only for computational reasons, but a deeper reason concerns limitations in human ability to deal with a large number of variables.

Suppose then that there is a known function of the data and parameter θ, such as the posterior or post-data distribution of the parameter given the data, upon which the perceived chance of success largely depends. Then under the Gaussian model for the errors, one might have the usual sharp posterior distribution for θ, with a precision that depends primarily upon the sample standard deviation and sample size, as in Lindley (1965, p. 36) and the analysis is quite robust to the a priori distribution, which may therefore be taken to be the uniform distribution, or perhaps the Cauchy distribution recommended by H. Jeffreys for testing purposes if one is concerned with propriety. Let us say that given the conventional Gaussian model, the conclusion of the analysis is that the posterior probability of success on this particular occasion is quite high, and so conditional upon this model one would make the launch. On the other hand, given the long-tailed model (more or less specifically modelled) with say tail-index α, the posterior distribution for θ is such that for the given occasion the appropriate conclusion is that the posterior distribution does not give sufficiently high mass to the near vicinity of θ_0, and so one should not launch. In this example it is

clear that the posterior or post-data distribution under alternative models cannot be ignored, and therefore that any p-value that does not take such alternative models explicitly into account is of limited use in deciding whether to launch.

With regard to the comparison between the prior and posterior predictive p-values, the posterior p-value would appear to be more appropriate for the launch problem if one is only concerned with the chance of success on the particular occasion, that is with the particular realized values of θ and all ancillary statistics and covariates. On the other hand, the prior p-value is more appropriate if one is concerned with usage of the model on a number of such occasions, in which the parameter values vary according to some a priori distribution π, which is based upon past experience. To see a practically important distinction between these two predictive p-values, suppose that in my long-tailed example the parameter $\alpha > 0$ is the usual tail-index for the upper tail of the distribution. It may well be the case that the parameter θ and the parameter α are regarded as dependent, with let us say large values of θ correlated with small values of α, so that when the observations are large, then one tends to have a very long-tailed distribution for the observations. Then it may be that for sufficiently small observations, the predictive p-value suggests that on this particular occasion the long-tailed model is not needed, and there is a high posterior probability for a successful launch. But one could hardly justify the conclusion that the model would be appropriate over a sequence of repetitions, since the a priori distribution for θ might give substantial mass to values for which the long-tailed model would be more likely to be appropriate.

Although I have argued above that the posterior predictive p-value can provide useful diagnostics for decisions such as to employ the conventional model and launch on the particular occasion, there are some further natural questions to ask. Should one be using a p-value at all for the most serious decision problems, such as a space launch or operation, or only for preliminary diagnostic purposes? And how do such predictive p-values relate to the Bayes factor (JS theory) for an assumed alternative model?

I would maintain that even if one is concerned only with success on the particular occasion, then the measure of discrepancy (or test statistic) is not arbitrary, and requires some serious attention to the credible alternative models. It cannot be based merely on conventional procedures such as minimum chi-square, which may have little to do with, for example, the probability of a successful launch. The standard measures of discrepancy are omnibus measures and are not tailored to a particular use of the model, such as whether or not to launch. In the long-tailed problem there is no way to ignore the distribution under the alternative, since unless the data strongly indicate that the conventional model is true, and the posterior distribution for θ is sufficiently concentrated near θ_0 then, if under plausible alternative models, the posterior distribution for the parameter

θ is not so concentrated, or is centered at the wrong place, one should not make the launch. Indeed, it is clear that the selection of a measure of discrepancy must go hand in glove with the intended usage of the model. Just as the predictive p-value derives its improvement over the classical p-value by giving more serious consideration to the a priori distribution under the conventional model, so too similar improvements can be made by giving more serious attention to the a priori distribution under plausible alternative models.

The authors suggest that it is sometimes appropriate to make tests without an alternative in mind. A partly contrary view is implied in the heading of Chapter V of Jeffreys (1961), where the Cat tells Alice that if you don't have any idea where you want to go, it doesn't much matter which way you go. Presumably it is desirable to replace a model that is inadequate in some respects by a better one. The celebrated Michelson-Morley experiment (1887), perhaps the most famous experiment in all of science, led to the rejection of the ether (such rejection not even being contemplated beforehand by the experimenters), and eventually played an important role in the creation of relativity theory. But in physics the models that are used have already been subjected to careful scrutiny, and while not literally true, have been shown to provide excellent approximations in well-defined circumstances. It is a major achievement to modify a particular model so as to cover a wider variety of conditions; and a careful experiment that suggests that a standard model is incorrect in some circumstances often leads to better understanding and improvements. This is not always the case in applications of modern statistics.

In the best uses of the classical p-value, such as by Daniel Bernoulli, or by R. A. Fisher in connection with his exact analysis of a 2 by 2 contingency table, there is an implicit alternative which suggests a direction of discrepancy. In Bernoulli's problem the implicit alternative is any model in which the planetary orbits are nearly coplanar, and in Fisher's problem the implicit alternative is one in which the treatment is highly effective relative to the placebo. There is a non-denumerable infinity of even continuous functions that might be used blindly to provide a test statistic, under the conventional model. It is then nearly certain that a statistician can select such a test statistic, if he wishes, to achieve essentially any desired p-value. As argued in Hill (1985-86) not only such shenanigans, but any use of data-analysis invalidates the non-Bayesian theory of statistics. Of course no intelligent person with integrity would allow such tests to be taken seriously, and we should instead focus on only those test statistics that can be argued to be relevant to some sensible alternative.

Next, nonnegativity considerations arise not only in the tomography example, but also in many other standard problems, such as in variance components estimation, and the use of kernels in the estimation of densities, and pose a serious difficulty for non-Bayesians. It was shown in Hill (1965, 1967, 1980, 1994a) that

in the variance components problem, negative unbiased estimates of the between variance component often imply a flat likelihood function for the true variance component under the model, so little information is provided by the experiment under the usual model. In such situations the posterior predictive p-value would in effect turn out to employ essentially the a priori distribution for the between variance component in its computation. There is more than an analogy here, because as pointed out in Hill (1965, p. 808) certain questions regarding Poisson distributions are mathematically nearly identical with the related questions for variance components. For example, if one has independent Poisson measurements of radioactivity with and without a source present, and the background reading is the larger of the two, then the likelihood function for the magnitude of the source can be flat.

In pursuing such issues I was led to introduce an alternative assumption for the one-way variance components model, namely that there is a constant negative correlation between the errors in a row in Hill (1967). This then led me to perform the full JS analysis comparing the standard model with such an alternative. I found the results of this analysis convincing, and it showed that appropriate data would lead one to reject the standard model in favor of one with such negative correlations. This was a particularly simple such JS analysis, since both null and alternative models were plausible, with the alternative model having only one extra parameter, namely the correlation coefficient. The posterior predictive p-value might also lead to rejection of the standard model, but would not allow a meaningful comparison of the standard model with the alternative model, nor allow for predictions based upon a mixture of the posterior distributions for the two models. In my opinion the inability to obtain useful predictive procedures giving weight to more than one model is a very serious limitation inherent in all approaches other than the JS. Since the optimal weight necessarily involves the posterior probability for the model, one is forced to assess this quantity. The examples of Bayes (1764), Harrison and Stevens (1976), and Hill (1994b) illustrate various aspects of the Bayesian theory of prediction.

In conclusion, the authors should be congratulated for making one of the few serious contributions to an important issue. This comes out of their careful attention to the use of a priori information under some conventional model. Since it has taken some 2.5 centuries to progress from the p-value of Daniel Bernoulli to the posterior predictive p-value, it may take quite a while until we progress to more careful attention to the a priori information under alternative models, and until such time the authors offer a real improvement over the classical p-value.

Department of Statistics, University of Michigan, Ann Arbor, MI 48109, U.S.A.

COMMENT

Robert E. Kass and Larry Wasserman

Carnegie Mellon University

Although, as Good (1971) pointed out, there may be at least as many *kinds* of Bayesians as there are Bayesians, there is nonetheless a recognizable orthodoxy in Bayesian thinking which is derived from the premise that all inferences are gambles, together with the fundamental result that for any state of knowledge optimal gambles are based on expectations. Aside from a few very special or contrived situations there has not been any justification of hypothetical-replication tail areas in terms of optimal gambles. Thus, these tail area calculations play no role in "orthodox" Bayesian inference.

In contrast to the orthodox view, Dempster (1971) argued in favor of such tail area calculations, articulating a Bayesian version of the Fisherian dualism in probabilistic inference: for estimation based on a model, one uses posterior distributions while for model assessment one uses hypothetical tail areas to calibrate surprise. (Dempster described these with the labels *predictive* and *postdictive* reasoning; although he chose his terminology to carry an explicit meaning, the use of a predictive tail area becomes an instance of Dempster's "postdictive" inference.) Box (1980) took up the same theme, proposing prior predictive tail areas as general model assessment tools, and Rubin (1984) recommended, instead, *posterior* predictive tail areas. (See Good (1953) and Evans (1995) for alternative approaches to quantifying surprise.) The authors' contribution is to show that the concept of posterior predictive model assessment can be applied very generally and fits well with modern computational methods.

Whether one finds this methodology attractive depends, to a large extent, on one's philosophy of inference. In our commentary we would like to mention the objections to hypothetical-replication tail areas and make a couple of remarks about Bayes factors; we will then offer our own evaluative opinions.

Objections to Tail Areas

The authors' method is to substitute posterior predictive p-values for frequentist p-values in performing goodness-of-fit tests. Their tests are quite similar in spirit to the analogous frequentist versions. Thus, the authors' approach is open to many of the same criticisms often leveled against the classical approach. Here are three of them:

1. The rationale for the method is, at least, very different from the intuitive probabilistic inference provided by other Bayesian procedures. In commenting

on the oddity of using hypothetical replications of data in place of the data at hand, Jeffreys (1961, p. 385) wrote, "What the use of P implies, therefore, is that a hypothesis that may be true may be rejected because it has not predicted observable results that have not occurred. ... On the face of it the fact that such results have not occurred might more reasonably be taken as evidence for the law, not against it."

One way to avoid these hypothetical replications is to use the *posterior* distribution of a discrepancy measure directly, rather than the posterior predictive distribution. For example, in the multinomial case, we might use the Kullback-Leibler divergence to assess how far the data are from the model. If $\hat{p}$ is the MLE for the unreduced multinomial and $p(\theta)$ is the multinomial according to the entertained model under parameter value θ, then we might examine the posterior distribution of $K(p(\theta), \hat{p})$. This proposal is similar to a suggestion made by Dempster (1975) to examine the posterior distribution of the likelihood ratio, which also avoids the hypothetical replications.

2. How are the posterior predictive p-values to be calibrated? Should we adopt the conventional levels that are widely applied with frequentist significance testing? We found it very surprising that the authors' $p = .058$ would be "the type of result we might obtain for a clearly inadequate model."

   The problem of calibration can be especially difficult when sample sizes are large, since small deviations from models become easy to find. The difficulty is that it is often unclear how small a deviation would have to be in order to be considered inconsequential.

3. Although, strictly speaking, questions about power seem irrelevant both in Bayesian inference and in pure significance testing, there is sometimes a legitimate concern. We agree with the sentiment that judgment of poor fit must depend on the questions being posed. (In Kass, Tierney, and Kadane (1989), the same point was emphasized in assessing case influence and prior sensitivity.) However, there are many instances in which a variety of alternative discrepancy measures may apparently apply to the same roughly-posed questions. How then is the data analyst to weigh the results? Is this necessarily a matter of personal scientific judgment, or might there be some guiding principles? As in other places where personal judgment enters, there is an issue as to how standards might be created for scientific reporting, where summaries are necessarily brief and somewhat conventional. Though less important than the two points above, this third remains intriguing to us.

What does it Mean to Ask Whether a Model is True?

The authors contrast their approach with others partly by saying there is often too much emphasis on asking whether a model is correct. We agree that

the question tends to be overemphasized by practitioners. However, we have a very hard time finding any content to statements saying that models are rarely "perfectly true" (Gelman, Carlin, Stern and Rubin (1995, p. 162). This may be a subtle point that ought to be argued at length, but let us very briefly outline our difficulty.

Simply put, statistical models are used to connect scientific theories with observed data. Of course, strictly speaking, the models are essentially never correct, and the scientific theories themselves are also never true in any absolute sense. However, just as obviously, in scientific work, as in everyday life, we find it tremendously useful to *act as if* our theories (our "models of the world") are true. Operationally, our acting as if a theory were correct is the same thing as believing it to be correct. The qualification to this statement, and the source of the authors' alternative perspective, is that for some purposes one may act as if a theory were correct while for other purposes one may choose not to do so. Thus, it might seem sensible to remove our global notion of acting as if a model were true and replace it with a qualified version, acting instead as if a model were true *for some specified purpose*.

We agree with the authors that it can be useful and important to make such qualifications. However, they often are relatively minor within the scope of our behavior and, therefore, can be something of a distraction. In everyday life, for example, our "mental model" might say that a particular clock keeps time accurately. Of course, we could qualify this according to purpose, but we rarely bother because (i) it suffices to leave our qualifications (specifying mechanical and electrical conditions required to make it run, as well as degree of accuracy) implicit and (ii) we like to rely on that clock. Thus, we all carry around scientific or personal models that we effectively assume to be correct, without need of any explicit qualification.

In this regard, we find the authors' analogy with consumer purchasing behavior to be somewhat inappropriate. They say, "A consumer surely can refuse to buy a defective product, even if it is the only item available. If the consumer does decide to purchase it despite its deficiencies, he or she surely would like to know about its defects and their possible consequences." Yes, we agree, it is very important to be aware of serious deficiencies in models. This is what we like about the methodology the authors describe. However, in many contexts we will end up using a model for all future work until a better one can be found. In the scientific context, Jeffreys (1961, p. 391) made this point by saying, "There has not been a single date in the history of the law of gravitation when a modern significance test would not have rejected all laws and left us with no law."

To summarize, while we agree there is sometimes a misplaced emphasis on methods that ask which of two models is "true", and support the authors in

their quest for methods that consider (albeit rather indirectly) the purpose to which a model might be put, we strongly believe it is frequently useful to weigh alternative models against each other by asking which of them would serve best if we were to act as if it were correct. This, of course, explains our interest in Bayes factors.

So, What About Bayes Factors?

The authors' main interest is in goodness-of-fit of a single model, without consideration of any alternatives. However, they also apply their technology to the weighing of competing hypotheses. Despite their choice to focus on " 'assessment' instead of 'testing' " the title of Rubin and Stern (1994) is "Testing in latent class models using a posterior predictive check distribution." This, we feel, will often remain useful, but can potentially be dangerous: it drifts into cases in which *evidence* is what is desired, and this precisely what they caution users to avoid (Gelman, Carlin, Stern and Rubin (1995, p. 173)).

When evidence is what's needed, Bayes factors are, in principle, the tool to use. Kass and Raftery (1995) have reviewed the methodology, providing examples, discussion, and many references. As emphasized by Kass and Raftery (1995), a serious practical difficulty involves the sensitivity to choice of prior on model parameters. In particular, while improper reference ("noninformative") priors can be used on any parameters that are common to both of two models (the nuisance parameters), they can not be used on the parameters of interest that distinguish between the models. Although we acknowledge this to be a serious problem that limits the applicability of Bayes factor technology to special circumstances, we have discussed a solution that we think often offers a useful procedure (Kass and Wasserman (1995)), and wish to emphasize here that the technology can be pushed and these "special circumstances" can be important.

It is also possible to put the goodness of fit problem in the framework of assessing the evidence of alternative models by expanding the nominal model. A natural way to do this is to start with a family of densities $\{g(u|\psi); \psi \in \Psi\}$ on the unit interval and transform back to the real line under the nominal family $\{f(x|\theta); \theta \in \Theta\}$. This gives an expanded family

$$h(x|\theta, \psi) = f(x|\theta)g(F(x|\theta)|\psi),$$

where $F(x|\theta)$ is the c.d.f corresponding to $f(x|\theta)$.

Neyman (1937) suggested the family

$$\log g(u|\psi) = \sum_{j=1}^{k} \psi_j \phi_j(u) - c(\psi),$$

where $\psi = (\psi_1, \psi_2, \ldots)$, $\phi_1, \phi_2, \ldots$ are orthogonal polynomials on $[0, 1]$ and $c(\psi)$ is a normalizing constant (see Ledwina (1994) for a recent discussion of this family). Bayarri (1985) investigated a particular one parameter family for $g(u|\psi)$ called the α-distribution. (The family was originally introduced by Bernardo (1982).) Verdinelli and Wasserman (1995) use the infinite dimensional version of Neyman's family:

$$\log g(u|\psi) = \sum_{j=1}^{\infty} \psi_j \phi_j(u) - c(\psi).$$

They take $\phi_j \sim N(0, \tau^2/c_j^2)$ and give τ a half normal prior centered at 0 with variance w. Verdinelli and Wasserman (1995) show how to choose the c_j so that the Bayes factor is consistent, i.e., goes to 0 or infinity if the nominal model is false or true. This leaves a single hyperparameter w to choose. They suggest a default value for w based on simulation studies and use MCMC methods to estimate the Bayes factor. The model is nonparametric in the sense that it can be shown to have support over (and be consistent for) a large, infinite dimensional class of densities. Of course, many other alternative models could be developed. The point is that it is possible to construct flexible classes of alternative models with some generality. The success of these methods in practical problems has yet to be determined.

Our Opinions

We applaud the authors' investigation and believe the methodology will, because it is often easy to apply in conjunction with posterior simulation, substantially extend goodness-of-fit assessment. Like the authors, we are not ourselves followers of the orthodox Bayesian viewpoint we alluded to at the outset: we believe Bayesian inference has an important role beyond decision-making. We disagree philosophically with the authors on a couple of points and highlighted them because applied statistical work should pay some attention to foundational issues. On the other hand, we expect these disagreements to be relatively minor, being matters of shades of emphasis. In the end, applied work is likely to be eclectic, acknowledging that all good statistical methods have their strengths and limitations.

Our reservations about posterior predictive p-values are described well in terms of Dempster's (1975) separation of probability to represent belief about unknown quantities of interest (his "predictive" reasoning) and probability to calibrate surprise (his "postdictive" reasoning). We continue to see a role for the representation of belief about hypotheses when one is primarily interested in weighing evidence in favor of competing models; we believe posterior predictive p-values, while remaining a useful supplement, should, at least in principle, be replaced by Bayes factors for this sometimes-important purpose. We also continue

to wonder just what role in statistics should be played by formal expressions of surprise.

Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213, U.S.A.

COMMENT: POSTERIOR PREDICTIVE ASSESSMENT FOR DATA SUBSETS IN HIERARCHICAL MODELS VIA MCMC

Steven M. Lewis and Adrian E. Raftery

University of Washington

1. Introduction

We congratulate Gelman, Meng and Stern – hereafter GMS – on an excellent exposition of the uses of posterior predictive model checking. The methods described are very useful, and we believe that this clear and compelling paper will help to propagate their use.

Model checking is, as the authors point out, always a vital component of model building. The frequentist approach relies on the clever choice of discrepancy measures that are pivotal and whose distribution under the hypothesized model is known, at least approximately. The Bayesian approach described by the authors is more general. The discrepancy measures used do not have to be pivotal or to have a known distribution, and so can be chosen for substantive reasons or to check aspects of the model that cause concern.

In addition, there is always a natural set of discrepancy measures available, namely the log predictive probability of each observation given all the others, or given the previous ones in some ordering. The former generalizes the externally studentized residuals of linear regression, or, more generally, leave-one-out cross-validation (e.g. Gelfand, Dey and Chang (1992)), while the latter corresponds to the prequential model assessment approach of Dawid (1984, 1992). These are available for all models, and so there is less need for cleverness than in the frequentist approach.

We would like to make explicit, and to illustrate, an extension of the idea, namely that posterior predictive assessment can be applied not just to an *overall* discrepancy measure, but also to discrepancy measures calculated for subsets of

the data, such as males and females, blacks and whites, and so on. This could reveal not just *whether* the model is inadequate, but also *in what way*, and so could be useful for improving the model. We illustrate this with a hierarchical event history model developed for World Fertility Survey data and estimated by Markov chain Monte Carlo (MCMC). We also comment on the use of posterior predictive checks for point processes in time and space, and on how the GMS methods fit into the overall model-building process.

2. Posterior Predictive Assessment for Major Subsets of the Data

GMS provide an informative discussion of the use of posterior predictive distributions to check the *overall* fit of a model. They argue that for $D(\boldsymbol{Y};\boldsymbol{\theta})$, a discrepancy measure between sample and population quantities, a tail-area probability of D under the posterior reference distribution is

$$p_b(\boldsymbol{y}) = \int P[D(\boldsymbol{Y};\boldsymbol{\theta}) \geq D(\boldsymbol{y};\boldsymbol{\theta})|M,\boldsymbol{\theta}]P(\boldsymbol{\theta}|M,\boldsymbol{y})d\boldsymbol{\theta},$$

where M is used here to represent the assumed model. In other words $p_b(\boldsymbol{y})$ is the probability of obtaining as unusual a discrepancy measure as the one actually observed under the posterior predictive distribution.

Posterior predictive distributions can also be used to check the fit for substantively important subsets as well. For example, in modeling human fertility the estimated model may fit the probabilities of births for women with less than a secondary education but may not fit probabilities for more educated women or the estimated model might fit women with lower parities but may not fit women with higher parities. In turn, this may well suggest areas to look for improvements in the model.

GMS describe how a sample from the posterior distribution obtained by MCMC can be used to calculate different discrepancy measures for assessing overall fit. Such a sample from the posterior can also be used to obtain a sample from the posterior predictive distribution which in turn can be used to assess the fit of the model to important subsets of the data. Our adaptation of the method of GMS can be written as an algorithm in the following way.

1. Draw $\boldsymbol{\theta}$ from the posterior distribution, $P(\boldsymbol{\theta} \mid M, \boldsymbol{y})$.
2. Draw $\ddot{\boldsymbol{y}}$ from the sampling distribution, $P(\boldsymbol{Y} \mid M, \boldsymbol{\theta})$. We now have a single realization from the joint distribution, $P(\boldsymbol{\theta}, \boldsymbol{Y} \mid M, \boldsymbol{y})$.
3. Repeat steps 1-2 many times.

The set of $\ddot{\boldsymbol{y}}$'s drawn using this procedure constitute a sample from the posterior predictive distribution since

$$P(\boldsymbol{Y} \mid M, \boldsymbol{y}) = \int P(\boldsymbol{Y} \mid M, \boldsymbol{\theta})P(\boldsymbol{\theta} \mid M, \boldsymbol{y})d\boldsymbol{\theta}.$$

GMS report model fitness in terms of a single number, the tail area probability for their choice of discrepancy measure, which they call a p-value. This idea can be expanded by computing such values for several major subsets of the data, to diagnose the *nature* of the model inadequacy, and thus to help suggest the form of a better model. It can also be expanded by not restricting attention to measures that can be viewed as distances, and by considering *both* upper and lower tail area probabilities.

3. Example: A Hierarchical Event History Model from Demography

We used posterior predictive distributions to check the fit of a logistic hierarchical model for fertility data collected in Iran as part of the World Fertility Survey. For a description of the study and the main findings, see Raftery, Lewis and Aghajanian (1995), and for further details of the analysis see Lewis (1994) and Raftery, Lewis, Aghajanian and Kahn (1996).

We fitted a logistic hierarchical model of the form

$$\text{logit}(\pi_{it}) = \beta_0 + \sum_{p=1}^{p} \beta_p x_{pit} + \alpha_i,$$

where x_{pit} was the pth covariate for the ith woman in exposure-year t, β_0, $\beta_1, \ldots, \beta_P$ were unknown regression coefficients and the α_i's were unobserved woman-specific random effects representing unmeasured characteristics that affect fertility, such as fecundability and coital frequency. We assumed that the α_i's were independent random variates from a normal distribution with a mean of 0 and a common variance, σ^2.

Logistic hierarchical models were fitted to data from several different provinces in Iran using MCMC. Boyrahmad province was one of the provinces modeled. An important covariate in the model was the woman's parity in the current exposure-year. (Parity is the number of previous children the woman has had.) In Boyrahmad, parity ranged between 1 and 5, inclusive. We were concerned as to whether or not the model fit adequately at each of the five parities. In other words we wanted to assess model fit for the parity 1 exposure-years, the parity 2 exposure-years and so on.

Table 1. Tabulation of sample from posterior predictive distribution

Parity	Number of Births Predicted									
	0	1	2	3	4	5	6	7	8	9
1	64	120	135	**106**	48	16	8	3		
2	18	57	**104**	102	102	59	40	11	6	1
3	79	140	**134**	89	49	8	1			
4	84	157	**126**	94	31	6	2			
5	180	**173**	104	38	5					

In the example the support of the posterior predictive distribution within each parity level was restricted to the integers between 0 and the number of exposure-years found in the sample at each parity level. Such a tabulation is presented in Table 1. In Table 1 the boldface entries reflect the actual number of births reported in the IFS from Boyrahmad province. For each parity, the actual number of births seems well within the central part of the posterior predictive distribution.

The main points of this display can be shown more concisely using a two-number summary consisting of the proportions of simulated numbers of births above and below the observed numbers, as in Table 2. This confirms that the observed data agree well with the posterior predictive distribution. However, there is a slight suggestion that there were more births than expected at parity 1, and less at parity 2, indicating the incorporation of a parity effect might be worth considering. One approach would be to expand the model this way and then to compare the expanded model with the model without a parity effect using Bayes factors (e.g. Kass and Raftery (1995)).

Table 2. Two-number summaries for location of observed number of births

Parity	Less Than Observed	Greater Than Observed
1	0.64	0.15
2	0.15	0.64
3	0.44	0.29
4	0.48	0.27
5	0.36	0.29

4. Posterior Predictive Distribution of Squared Standardized Pearson Residuals, Summed Over Subsets

In the previous section we used the number of births as a summary statistic. Any number of other summary statistics might have been used instead. For example, the sum of the squared standardized Pearson residuals,

$$T(\boldsymbol{y}) \equiv \sum_i \Big(\frac{y_i - n_i \hat{\pi}_i}{\sqrt{n_i \hat{\pi}_i (1 - \hat{\pi}_i)(1 - h_{ii})}} \Big)^2,$$

where h_{ii} is the ith diagonal element of the hat matrix, is a good alternative summary statistic.

Figure 1 contains plots of the posterior predictive distributions of the sums of the squared standardized Pearson residuals over the parity 1, parity 2, parity 3 and parity 4 observations. The posterior predictive distributions displayed in this figure were produced using the procedure described in Section 2, except that instead of drawing from the sampling distribution of the response variable, draws

were taken from the sampling distribution of the squared standardized Pearson residuals.

Plots of squared standardized Pearson residual posterior predictive distributions

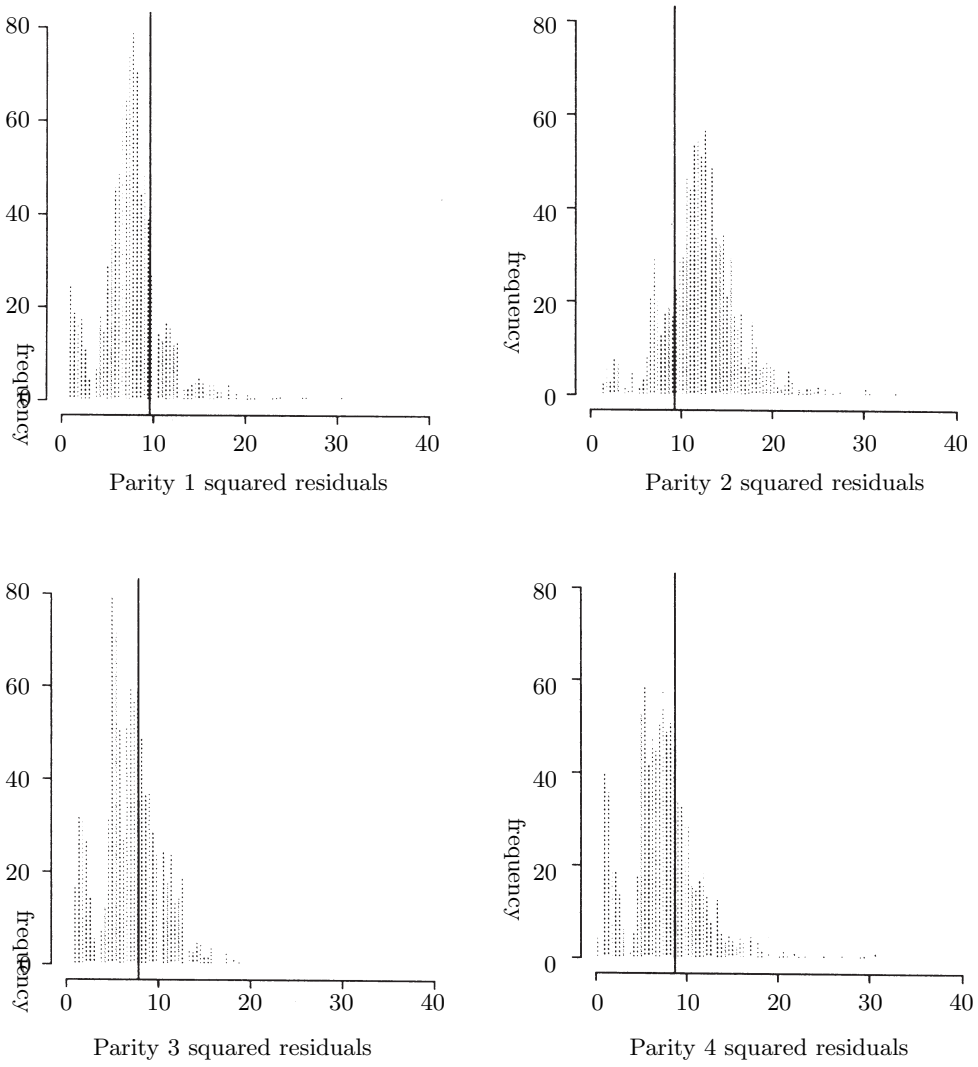

Figure 1. Sums of squared standardized Pearson residuals in reference to their parity determined posterior predictive distributions.

Looking at the plots in Figure 1 we see that the observed sums of the residuals (represented by the solid vertical lines in each plot) are not at all unusual. For the parity 1 observations it looks as if the observed sum might be located in the right tail of the distribution, giving a hint that the model might be improved by

adding an indicator variable for the parity 1 observations. In fact, Raftery, Lewis, Aghajanian and Kahn (1996) came to just such a conclusion on the basis of the data from all of Iran (here we are analyzing the data from only one province for illustrative purposes).

Table 3 shows two-number summaries for each parity level. The poorest fitting sets of observations are those for parities 1 and 2, but they do not fit badly: in each case at least 20% of the posterior predictive distribution is beyond the observed value.

Table 3. Two-number summaries for location of observed sum of squared standardized Pearson residuals

Parity	Less Than Observed	Greater Than Observed
1	0.79	0.21
2	0.20	0.80
3	0.64	0.36
4	0.69	0.31
5	0.61	0.39

5. Application to Point Processes

The adequacy of models for point processes in time or space has traditionally been checked more using graphs of *functions* that describe the process, with pointwise confidence bands around them, than by single-number overall measures of fit. For point processes in time, such functions include $N(t)$, the cumulative number of events up to time t, and the conditional intensity function (Cox and Lewis (1966), p. 69). For spatial point processes, the function $K(t)$ = the average number of events in a ball of radius t around a randomly picked event, and related functions, are often used (Cressie (1993); Diggle (1983); Ripley (1977, 1981)).

The distribution of the graphed function, such as $N(t)$, or $K(t)$, under the hypothesized model usually depends on unknown parameters, $\boldsymbol{\theta}$, that are estimated. The pointwise confidence bands are most often computed by simulation using the plug-in rule, i.e. conditionally on an estimate of $\boldsymbol{\theta}$. This underestimates uncertainty because it takes no account of uncertainty about the model parameters, and so may lead to overstringent model assessment if the discrepancy measure is not pivotal.

The GMS method shows immediately how this difficulty can be overcome:
(1) simulate a value of $\boldsymbol{\theta}$ from its posterior distribution;
(2) simulate (or find analytically) the function to be graphed given $\boldsymbol{\theta}$;
(3) repeat (1) and (2) many times (19 and 99 are conventional numbers of times in the spatial point process literature);

(4) plot the resulting pointwise confidence bands.

This was done for a nonhomogeneous Poisson process model arising in software reliability by Raftery (1988). Results for two data sets are shown in Figure 2. For the first data set, $N(t)$ lies within the posterior predictive confidence bands over the entire range of the data, so there is no indication of model inadequacy. In contrast, the plot shows clearly that the model does not fit the second data set at all well, as the observed function lies outside the confidence bands for over two-thirds of the events. This observation was subsequently used to develop a better fitting model for these data, with very different implications for the main question of interest, i.e. the extent to which the software had been debugged (Raftery (1987)). The idea is essentially the same as that underlying Figures 4 and 5 of GMS, and can be applied in any context where model checking is reasonably based on a function rather than (or in addition to) a single overall discrepancy measure. Another application of posterior predictive checking to point processes was given by Raftery (1989), this time in the context of environmental monitoring.

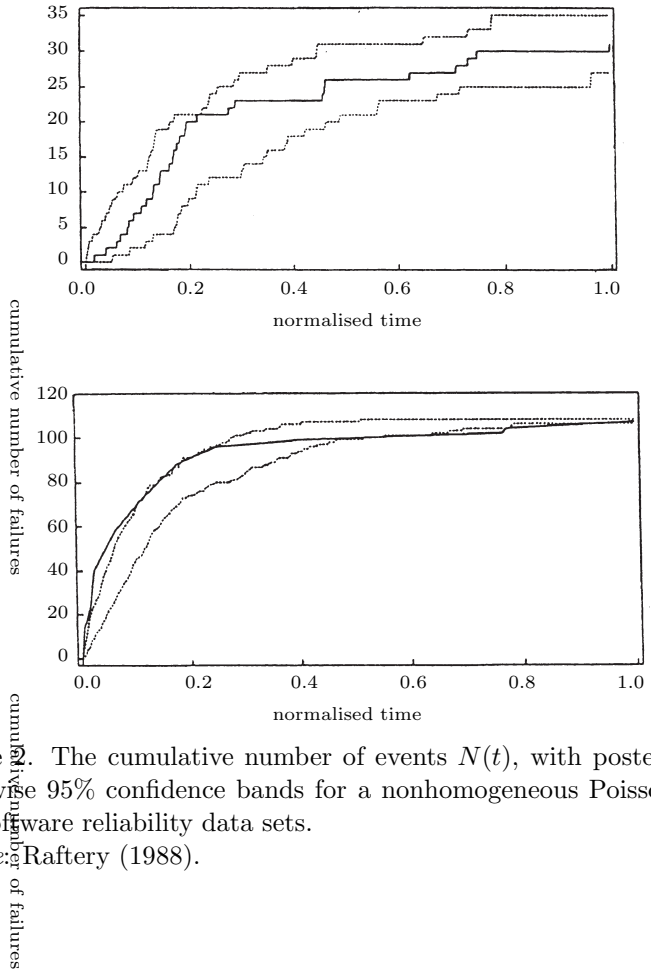

Figure 2. The cumulative number of events $N(t)$, with posterior predictive pointwise 95% confidence bands for a nonhomogeneous Poisson process, for two software reliability data sets.
Source: Raftery (1988).

6. Posterior Predictive Assessment in Model Building

We view posterior predictive assessment as a useful component of the overall model-building process. Any model should be checked using graphical displays and diagnostic checks. Posterior predictive assessment provides a general methodology for calibrating such displays and checks, which should be used to indicate which improvements to the model one should develop and test.

However, we feel that only rarely should posterior predictive assessment be used to reject a model in an absolute sense when no better alternative is available. Rather, it should be used to develop a new model, which is then compared to the current model using Bayes factors. In many applications, the task is to make a decision, and this has to be done on the basis of the best available model, even if it is not a perfect one. Jeffreys (1961, p. 391) pointed out that the best available model gets used even when inadequacies are found, noting that "there has not been a single date in the history of the law of gravitation when a modern significance test would not have rejected all laws and left us with no law."

Thus we feel that posterior predictive assessment should usually be used to point the way to a better model, rather than to reject the current model in an absolute sense. This is compounded by the fact that the methods of GMS are based on tail area probabilities rather than the probability of the data themselves under the model, and so one should be cautious in interpreting their output as evidence against the current model (e.g. Berger and Delampady (1987); Berger and Sellke (1987)). Conversely, one should not necessarily stop trying to improve the model just because it is accepted using posterior predictive assessment methods. For a related debate between Gelman and Raftery (with the roles of author and discussant reversed), see Raftery (1995).

Acknowledgements

This research was supported by National Institutes of Health Grant 5-R01-HD26330 and by ONR Grant N-00014-91-J-1074.

Center for Studies in Demography and Ecology, University of Washington, Box 353340, Seattle, WA 98195-3340, U.S.A.

Department of Statistics, University of Washington, Box 354322, Seattle, WA 98195-4322, U.S.A.

COMMENT : ON POSTERIOR PREDICTIVE p-VALUES

Donald B. Rubin

Harvard University

Gelman, Meng and Stern ((1996); GMS) are to be congratulated for insightfully summarizing and extending the theory of posterior predictive check (ppc) distributions and for illustrating various applications in revealing examples. The extension to include generalized discrepancies $D(y;\theta)$ rather than just statistics $T(y)$, as in Rubin (1984), may be practically quite important because of the potentially tremendous computational advantages when averaging rather than maximizing over θ, especially with multi-modal likelihoods.

As more realistic and complex models are fit to data using increasingly sophisticated software and hardware, statisticians with serious interests in applications will be compelled to employ Bayesian technology, and I'm convinced that ppc's will be an integral part of that technology. Thus, I strongly advocate the central theme of GMS that a scientist can question and reject a posited model– with the implied directive to return to the scientific drawing board, without having any specific alternative model available; for example, quantifying evidence of failings of Newtonian physics must be allowed without having Einsteinium relativity available for the assessment. Those who define "Bayesian statistics" to exclude such model-assessment activities condemn it to be of marginal relevance in science rather than at its core, where I feel it belongs.

My opinion on this topic has remained the same for many years (Rubin (1984, p. 1166)).

> 5. Model monitoring by posterior predictive checks
> ... Given observed data, X_{obs}, what would we expect to see in hypothetical replications of the study that generated X_{obs}? Intuitively, if the model specifications are appropriate, we would expect to see something similar to what we saw this time, at least similar in "relevant ways". This statement, which is essentially a fundamental premise of frequency inference, seems to me so basic that it needs no defense. Unlike the frequentist, the Bayesian, though, will condition on all observed values.

Need for Careful Terminology

For ppc's to achieve the broad acceptance they deserve, I expect that it will be necessary to obtain a better understanding of their operating characteristics in

the traditional frequentist sense. Some issues here appear to me to be relatively subtle and elusive, and so precise and evocative terminology may be especially important. For example, it is important to realize that the function $p_c(y;\theta)$ in equation (3), the GMS "classical p-value" for statistic $T(y)$, is not generally a p-value in the standard sense because of its potential dependence on the unknown parameter θ; perhaps "classical tail-area function" for statistic $T(y)$ would be more descriptive.

Special care is also necessary when reading the symbol "y^{rep}". Although the meaning of y^{rep} in equations (4)-(7) is the same (a draw from its posterior predictive distribution, i.e., given observed data y under the posited model H), the meaning of y^{rep} in (3) comes from (2), a draw form its posited sampling distribution; that is, under H and given θ, in (3), y and y^{rep} are i.i.d. copies of the same random variable, with y^{rep} the random variable and y conditioned upon. Thus with continuous $T(y)$, the distribution of $p_c(y;\theta)$ will be uniform on $(0,1)$ for each θ, and close to uniform for discrete $T(y)$. In general, however, $p_c(y;\theta)$ cannot be calculated from y alone, so the need arises to eliminate θ if an actual p-value is to be obtained.

Interpreting the ppc p-Value from Bayesian and Frequentist Perspectives

Of course, the Bayesian eliminates a nuisance unknown, such as θ in $p_c(y;\theta)$, by integration over its posterior distribution. Thus the Bayesian who is interested in $p_c(y;\theta)$ is naturally led to its posterior expectation, the posterior predictive p-value, p_b in (4) or (7), which is a valid Bayesian posterior probability for the event $D(y;\theta) > D(y^{\text{rep}};\theta)$.

When $p_c(y;\theta)$ itself is the object of interest, however, aspects of its posterior distribution other than its mean are also important, just as the full posterior distribution of the bias of a coin can be important. This possibility was also mentioned in Meng (1994a). The posterior distribution of $p_c(y;\theta)$ can be simulated by obtaining repeated draws of $y^{\text{rep}\,j}$ at *each* drawn value of θ^j, thereby estimating $p_c(y;\theta^j)$, whence we obtain the posterior distribution of $p_c(y;\theta)$ when the θ^j are draws from the posterior distribution of θ. Even with a general discrepancy $D(y;\theta)$ replacing the statistic $T(y)$ in (3), if the posterior distribution of $p_c(y;\theta)$ is tightly concentrated about p_b with little sensitivity to the prior distribution on θ, then p_b should be acceptable to a frequentist as an honest p-value, because it has nearly a uniform sampling distribution under H for all plausible θ; any applied frequentist, even a devoted one, should not care much about operating characteristics under distributions that are irrelevant given observed data.

Moreover, even when the posterior distribution of $p_c(y;\theta)$ is not tightly concentrated about its mean, the ppc p-value, p_b, has a comfortable frequency interpretation. As in Rubin (1984, p. 1160) simply describe the averaging of $p_c(y;\theta)$

over the posterior distribution of θ as a frequency operation involving conditioning on those values of θ that could have led to the observed value of y.

> Suppose we first draw equally likely values of θ from $p(\theta)$, and label these $\theta_1, \ldots, \theta_s$. The θ_j, $j = 1, \ldots, s$, can be thought of as representing the possible populations that might have generated the observed X[y in GMS]. For each θ_j, we now draw an X from $f(X|\theta = \theta_j)$; label these $X_1, \ldots, X_s$. The X_j represent possible values of X that might have been observed under the full model $f(X|\theta)p(\theta)$. Now some of the X will look just like the observed X and many will not; of course, subject to the degree of rounding and the number of possible values of X, s might have to be very large in order to find generated X_j that agree with the observed X, but this creates no problem for our conceptual experiment. Suppose we collect together all X_j that match the observed X, and then all θ_j that correspond to these X_j. This collection of θ_j represents the values of θ that could have generated the observed X; formally, this collection of θ values represents the posterior distribution of θ.

And the average value of $p_c(y;\theta)$ over this collection of θ values is the ppc p-value.

On the Typical Conservatism of ppc p-Values

Despite this direct frequency interpretation of the ppc p-value, it is still of interest to understand its repeated-sampling operating characteristics, that is, its distribution over repeated draws of y given fixed θ under H. Meng (1994a) provides some results when averaging over the prior distribution of θ, but sharper results are desirable, for example concerning the typical conservatism of ppc p-value noted by GMS, Meng (1994a), and Rubin (1996a). This conservatism can be connected to recent discussions of the potential conservatism of multiple imputation in Meng (1994) and Rubin (1996b) stimulated by comments in Fay (1992, 1996). For this discussion, I assume that H is true and that there exists a true value of θ.

Consider a hypothetical infinite complete data set $y_{\text{complete}} = (y, y^{\text{rep}\,1}, y^{\text{rep}\,2}, \ldots)$ where each of the infinite number of $y^{\text{rep}\,j}$ is an actual i.i.d. replication, i.e., drawn from GMS's equation (2). The data analyst's hypothetical complete-data statistic p_{complete} is the infinite sample analogue of $p_c(y;\theta)$:

$$p_{\text{complete}}(y_{\text{complete}}) = [\text{proportion of } T(y^{\text{rep}\,j}) \text{ that are larger than } T(y)].$$

The value of $p_{\text{complete}}(y_{\text{complete}})$ is $p_c(y;\theta)$ evaluated at the true value of θ, and its distribution over repeated draws of y is uniform on $(0, 1)$.

But $p_{\text{complete}}(y_{\text{complete}})$ cannot be calculated from the observed data y because the $y^{\text{rep}\,j}$ are missing. Consequently, in the spirit of multiple imputation (Rubin (1987)) they will be imputed from their posterior predictive distribution (i.e. from GMS's equation (4)) to produce $y^{\text{imputed}}_{\text{complete}}$. The data analyst will then evaluate the complete-data statistic on this hypothetical infinite imputed data set to obtain $p_{\text{complete}}(y^{\text{imputed}}_{\text{complete}})$, which equals $p_b(y)$ of GMS's equation (5). These imputations will typically be "superefficient" for the statistic p_{complete} in the following sense (Rubin (1996b), Section 3.6): first, by the consistency of Bayesian estimates under the correct model in large samples of y, the expectations of $p_{\text{complete}}(y_{\text{complete}})$ and $p_{\text{complete}}(y^{\text{imputed}}_{\text{complete}})$ are the same, both equal to .5, over repeated new draws of data y; and second, because y and the imputations of $y^{\text{rep}\,j}$ are positively correlated, the variance of $p_{\text{complete}}(y^{\text{imputed}}_{\text{complete}}) = p_b(y)$ will tend to be less than the variance of $p_{\text{complete}}(y_{\text{complete}}) = p_c(y;\theta)$. As a result, $p_b(y)$ will be centered at .5 but less variable than a uniform random variable, and thus will have conservative operating characteristics. The extra conservativeness typically found when using a generalized discrepancy rather than a statistic presumably arises from the extra missing information when doing an analysis involving missing θ in addition to missing y^{rep}.

Mathematically rigorous versions of these theoretical assertions could, I believe, be based on a combination of the work by GMS, Meng (1994, 1994a), and Rubin (1996b). Of course, conservatism in failing to reject a true posited model is not a bad result, even if it occurs more frequently than indicated by a nominal level, and doing so is "confidence-valid" in the sense of Neyman (1934) (see Rubin (1996b), Section 1.4 and Rejoinder for some discussion).

When is a Posited Model "Good Enough"?

In this context of failing to reject a posited model, it is important to reinforce one of GMS's main points. Even when a ppc with a very powerful discrepancy measure can establish the inadequacy of posited model, the model still may be perfectly adequate for other purposes as measured by a less powerful, but more relevant discrepancy. For example, consider a long-history time-series data base, with two possible models for predicting weather. A ppc p-value based on 60-day-ahead forecasts might discard one model that satisfies ppc evaluations for 7-day-ahead forecasts, and vice-versa for the other model. Clearly, however, we need not discard both models: each model can be valuable for its intended purpose and need not be rejected because it is unsatisfactory for a different purpose.

For another example, consider a very large data set where all sorts of interactions are clearly present, but a parsimonious model does pragmatically nearly as well as a "full" model (e.g., in a regression context, $R^2 = .300$ with the full model and $R^2 = .299$ with the simple model, and the extra precision of prediction

using the full model is essentially lost in the error variance $(1 - R^2)$ of predictive distributions). The decision whether to use a "rejectable" model must be based on pragmatic and scientific criteria. In some cases, as when the truth of an accepted scientific model is in question, using powerful and diverse discrepancies will be appropriate, whereas in many other cases, as when using a model for obtaining predictive distributions across different settings, less powerful discrepancies, which can accept the adequacy of model for the purpose at hand, will be appropriate.

The Critical Role of "Fixed Features" of the Replications

Critically important to this discussion of when a model is "good enough" is the issue of what is meant by a replication – what are the "fixed features" (Rubin (1984), Section 5.3) underlying the new draws of (θ, y^{rep})? This choice can be as important as the choice of discrepancy for determining the adequacy of a posited model. For example, Rubin and Stern (1994) show that by imbedding both the χ^2 test and Fisher's exact test for a 2×2 table within the ppc framework, the essential difference between the answers generated by the two approaches is not due to the statistics themselves or the prior distribution over the marginal nuisance parameter, but rather simply due to the definition of what constitutes a replication. If the replications keep the margins fixed, so that, for example, the replications of a medical experiment always involve the same numbers of healthy and sick people as in this study, using either χ^2 or the Fisher statistic as the discrepancy gives the Fisher hypergeometric p-value as the ppc p-value, whereas if the replications randomly sample new patients for the replicated experiments, the ppc p-values, using either statistic for the discrepancy and almost any prior distribution on the nuisance parameter, are very nearly equal to the asymptotic χ^2 p-value.

For other interesting examples of the critical role of fixed features of the replications, consider sequential experiments with data-dependent stopping rules (as in Rubin (1984), Section 4.4). Suppose we keep sampling in a normal problem until the t statistic is greater than 2.5, and in our study we stop with the observed t statistic equal to 2.51. Considering t to be the discrepancy for the posited model of zero mean, this result may be just what we would expect to see when the replications correspond to the actual study design, whereas if the "replications" were defined to be simple random samples with the same fixed sample size as the observed sample size, the ppc replications would lead to a more extreme p-value and a possible rejection of the posited model. Study design does influence Bayesian posterior assessment of posited models!

In conclusion, this fine contribution by GMS should encourage more applications of ppc p-values, thereby improving the quality of applications of Bayesian

statistics, and moreover should stimulate new theoretical work concerning frequentist operating characteristics for these diagnostic techniques, thereby providing frequentists with new tools, and improving the bridge between Bayesians and frequentists.

Acknowledgements

Support for work and applications of posterior predictive check distribution has been received from the National Science Foundation, the National Institute of Mental Health, and the U.S. Bureau of the Census.

Department of Statistics, Harvard University, 1 Oxford St., Cambridge, MA 02138, U.S.A.

COMMENT

Sam Weerahandi and Kam-Wah Tsui

Bellcore and University of Wisconsin

Gelman, Meng, and Stern (GMS) provide a useful methodology in a Bayesian framework to assess the goodness of fit of a statistical model to the observed data. Their approach is based on the idea of discrepancy variables and posterior predictive p-values defined in Meng (1994) in a related problem. The treatment in GMS is particularly appealing as there is a class of natural and familiar discrepancy measures to facilitate the approach. Moreover, the GMS treatment does not seem to have an undesirable feature (described below) of the former paper on posterior predictive p-values.

We would like to comment on some aspects of the proposed method and address some of the problems raised in the examples discussed by GMS. As there are other types of applications where their approach has great potential for providing solutions, we also attempt to bring the attention of Bayesians to a class of problems in which the conventional Bayesian approach seems to have failed. In Section 3 below, we describe one such class, which is in a sense similar to the one undertaken in GMS; to stimulate much needed further research in this direction we solve a particular problem by taking the authors' approach.

1. Advantages of the Approach

One major advantage of the Bayesian approach proposed in GMS is its ability to provide relatively simple p-values for problems of testing simple/point null

hypotheses (e.g., the equality of a number of parameters) involving one or more parameters. The conventional Bayesian approach often runs into difficulties or fails in testing such hypotheses unless a more complicated form of prior placing a point mass on the null is assumed.

Another welcome feature of the posterior predictive p-values presented in GMS is that it has avoided a certain undesirable feature in Meng's original paper. In the treatment of Meng (1994), the prior distribution is confined to part of the parameter space as opposed to the whole parameter space as in a true Bayesian treatment. We considered this undesirable because when a prior distribution has been specified one should be able to carry out any kind of inference, not just hypothesis testing. For example, in ANOVA, after the original null hypothesis has been tested, one must carry out multiple comparisons and provide point estimates and interval estimates for parameters of interest.

2. Some Reservations

Although the examples used in GMS to motivate the problem and illustrate the solution are interesting and provide insight into the difficulties of classical treatments, some of them are not quite convincing due to the inherent problems of the assumed models themselves. For instance, in the motivating example on the position emission tomography experiment, one can assess and conclude the inadequacy of the model even before seeing any data. A model should be consistent with not only the observed data and estimated parameters, but also with any other data set that could have been observed. In the current application this means that the assumed model should have the property that each component of θ is a positive quantity. Therefore, it is customary in this type of application to model a transformed value of the parameter, such as $\log(\theta)$ (the logarithm of each component), rather than θ itself. After transforming the parameter space, the positivity constraint discussed in the paper is no longer an issue.

In many instances, the computation of the posterior predictive p-value is based on realizations simulated from the posterior distribution. Therefore the reliability of the simulated result is important and requires much attention. When the posterior distribution cannot be simulated directly, an indirect simulation method (Markov Chain Monte Carlo) such as the Metropolis algorithm is often used, as in the mortality rate example described in Sections 3.1 and 3.2 in GMS. In Figure 6, the histogram is based on 1000 simulations from the reference distribution for the minimum chi-square statistic for the mortality rate. It is not clear in this paper if the 1000's were generated in 1000 parallel sequences using the Metropolis algorithm, similar to the description given in the last paragraph of Section 3.1, or the 1000's were generated in one sequence after the convergence of the simulation is checked. Probably both lead to similar shapes of the histogram.

Among the recent literature on indirect simulation of a posterior distribution, for example, Geyer (1992) argues that simulation based on just one sequence of realization is preferable. In general, some discussion is needed on how sensitive the computed value of the posterior predictive p-value is to the way the posterior distribution is simulated.

3. Solving ANOVA Problems

To our knowledge, currently there are no Bayesian solutions available to problems in ANOVA involving fixed effects models when the underlying error variances are not necessarily equal. The conventional Bayesian treatment seems to have failed in this context. In a sense ANOVA problems have some features similar to the problem of the assessment of model fitness. For example, when we are to test the equality of some treatment means, the question we need to answer is, "are the observed differences in sample means just an artifact of sampling variation or is it due to differences in population means?" The question makes perfect sense and the treatment means may (or may not) indeed be equal because treatments being compared might be nothing but pure water from the same source flavored with banana, papaya, or mango. Given this reality, the idea of trying to compute the posterior probability of the null hypothesis based on a special prior (e.g. prior probability of 1/2 that the null hypothesis is true) is not only naive but also does not address the real underlying problem. Moreover, one should be able to perform other inferences such as multiple comparisons and interval estimation using the specified prior. The approach which attempts to handle ANOVA problems in terms of contrasts also becomes complicated or intractable in higher-way ANOVA problems and under heteroscedasticity.

The GMS approach provides great promise to solving ANOVA problems. In this section we attempt to solve the simple one-way ANOVA problem by taking their approach. We hope that this will stimulate much needed further research in this direction. Recently Weerahandi (1995) obtained a generalized p-value (cf. Tsui and Weerahandi (1989)) for the one-way ANOVA situation. The posterior predictive p-value approach of GMS may be able to produce a p-value equal to or similar to the generalized p-value using an appropriate prior. Below we solve the problem assuming the classical noninformative prior, which can be employed in any type of inference concerning the treatment means including the problem of testing their equality. The results can be easily obtained with natural conjugate priors as well. The approach presented below should prove to be useful in solving many higher-way ANOVA problems.

Consider the problem of comparing the means of k populations with unequal variances. Suppose a random sample of size n_i is available from the ith population, $i = 1, \ldots, k$. Let $X_{ij},\ i = 1, \ldots, k,\ j = 1, \ldots, n_i$ be the random variables

representing the observations taken from the k populations. Denote by $\mathbf{X}$ the vector of all X_{ij}'s. Let

$$\overline{X}_i = \sum_{j=1}^{n_i} X_{ij}/n_i \quad \text{and} \quad S_i^2 = \sum_{j=1}^{n_i} (X_{ij} - \overline{X}_i)^2/n_i, \ i = 1, \ldots, k,$$

be the sample means and the sample variances (MLE's) of the k populations. The observed values of these random variables are denoted as $\overline{x}_i$, s_i^2, $i = 1, \ldots, k$, $j = 1, \ldots, n_i$, respectively.

Let μ_i be the mean of the ith population. With the assumption of normally distributed observations consider the linear model:

$$X_{ij} = \mu_i + \epsilon_{ij}, \qquad \text{with} \quad \epsilon_{ij} \sim N(0, \sigma_i^2), \ i = 1, \ldots, k, \ j = 1, \ldots, n_i. \tag{3.1}$$

Consider the null hypothesis

$$H_0 : \ \mu_1 = \mu_2 = \cdots = \mu_k. \tag{3.2}$$

As in Weerahandi (1995), define the *standardized between group sum of squares*

$$\tilde{S}_b = \tilde{S}_b(\sigma_1^2, \ldots, \sigma_k^2) = \sum_{i=1}^{k} \frac{n_i}{\sigma_i^2} (\bar{X}_i - \bar{X}_w)^2 \Big/ \sum_{i=1}^{k} \frac{n_i}{\sigma_i^2}, \tag{3.3}$$

where $\bar{X}_w = (\sum_{i=1}^k n_i \bar{X}_i/\sigma_i^2)/(\sum_{i=1}^k n_i/\sigma_i^2)$. Let $\tilde{s}_b = \tilde{s}_b(\sigma_1^2, \ldots, \sigma_k^2)$ be the value of the random variable $\tilde{S}_b$ obtained by replacing $\bar{X}_i$'s appearing in (3.3) by their observed values $\bar{x}_i$, $i = 1, \ldots, k$.

Let $\boldsymbol{\sigma}^2$ be the vector of k population variances and let $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$ be the vector of all unknown parameters. In the Bayesian treatment of GMS the usual non-informative prior (as opposed to the prior that Meng (1994) used in the $k = 2$ case) is

$$\pi(\boldsymbol{\theta}) \propto \sigma_1^{-2} \cdots \sigma_k^{-2}.$$

Then the posterior density $f(\boldsymbol{\sigma}^2)$ of $(\sigma_1^2, \ldots, \sigma_k^2)$ given the data $\boldsymbol{x}$ is the product of the densities of σ_i^2, $i = 1, \ldots, k$ with the ith distribution given by

$$Y_i = \sigma_i^{-2} n_i s_i^2 \sim \chi_{n_i - 1}^2. \tag{3.4}$$

Using the notation in Meng (1994)

$$D(\boldsymbol{X}, \boldsymbol{\theta}) = \tilde{S}_b(\sigma_1^2, \ldots, \sigma_k^2)$$

is a discrepancy variable appropriate for the current testing problem. In fact in the $k = 2$ case this reduces to the discrepancy variable used by Meng (1994) and is the obvious direct counterpart in the general case.

We use the notation $\boldsymbol{X}$ in place of $\boldsymbol{x}^{\text{rep}}$ in GMS. Then, under H_0 and given $(\sigma_1^2, \ldots, \sigma_k^2)$, we have the conditional distribution

$$W = D(\boldsymbol{X}, \boldsymbol{\theta}) \sim \chi^2_{k-1}. \tag{3.5}$$

Moreover, since this is free of nuisance parameters, it is in fact independently distributed from the distribution (3.4) of $(\sigma_1^2, \ldots, \sigma_k^2)$. Hence, noting that $D(\boldsymbol{x}, \boldsymbol{\theta}) = \tilde{s}_b(\sigma_1^2, \ldots, \sigma_k^2)$ we can compute the posterior predictive p-value as

$$\begin{aligned} p_b(\boldsymbol{x}) &= \int Pr(W \geq \tilde{s}_b(\sigma_1^2, \ldots, \sigma_k^2) \mid H_0)) f(\boldsymbol{\sigma}^2) d\boldsymbol{\sigma}^2 \\ &= E\Big(Pr(W \geq \tilde{s}_b(\frac{n_1 s_1^2}{Y_1}, \ldots, \frac{n_k s_k^2}{Y_k}))\Big). \end{aligned}$$

Now proceeding with arguments similar to those in Weerahandi (1995), pp. 597-598, we get,

$$p_b = 1 - E\Big(H(\frac{N-k}{k-1} \tilde{s}_b[\frac{n_1 s_1^2}{B_1 B_2 \cdots B_{k-1}}, \frac{n_2 s_2^2}{(1-B_1) B_2 \cdots B_{k-1}}, \ldots, \frac{n_k s_k^2}{(1-B_{k-1})}])\Big),$$

where H is the cdf of the F-distribution with $k-1$ and $N-k$ degrees of freedom and the expectation is taken with respect to the independent beta random variables

$$B_j \sim Beta\Big[\sum_{i=1}^{j} \frac{(n_i - 1)}{2},\ \frac{n_{j+1} - 1}{2}\Big],\ j = 1, 2, \ldots, k-1. \tag{3.6}$$

This is the same expression that Weerahandi (1995) obtained for the generalized p-value for testing the equality of means. Therefore, the two solutions are numerically equivalent. The p-value can be conveniently computed using the XPro software package.

Bellcore, 445 South St., Morristown, NJ 07960-6438, U.S.A.

REJOINDER

Andrew Gelman, Xiao-Li Meng and Hal Stern

The primary purpose of our paper is to encourage routine goodness-of-fit assessment for complex models. This is especially important in view of the increasing utility and popularity of models with which we have limited experience,

partially due to recent advances in computational methods. With sophisticated models there can often be difficulties in applying classical goodness-of-fit tests, and the posterior predictive assessment is useful in such cases, especially given the increased flexibility provided by discrepancies that can be functions of parameters as well as data. We use the plural "models" to emphasize that even when one has fit several models for comparison or as part of a sensitivity analysis, the fit of each model still needs to be assessed before the analysis can be used as a serious component of a scientific investigation.

Comparing model predictions to observed data is such an obvious idea for model checking that we did not expect any statistician would object to it. We did, however, expect some reservations about using tail-area probabilities, largely as a reaction to the abuse of p-values in practice. We strongly condemn the abuse of p-values, but that does not mean we should throw out the baby with the bath water, since p-values, when viewed as appropriate probability statements, can be useful for model checking. We thus are grateful to *Statistica Sinica* and especially its Editor for organizing this collection of excellent discussions, which provides us with an opportunity to further emphasize our key points. We are also very grateful to the discussants for their valuable comments and suggestions, and are particularly encouraged by the consensus on the critical importance of Bayesian model diagnosis. Below, we respond to various points for which we believe some elaboration will be useful for readers; we list the names of relevant discussants in brackets whenever one of our points is relevant to that discussion, using abbreviations for multiple authors (KW = Kass and Wasserman; LR = Lewis and Raftery; WT = Weerahandi and Tsui).

On the General Use of Posterior Predictive Assessments in Practice

Our work was motivated by our belief that posterior predictive assessments, as formalized in Rubin (1984), are the natural Bayesian extension of classical model checking. Simulation-based model checking has a long history in statistics; examples from this century include a comparison of data on dog learning to simulated data from a stochastic model (Bush and Mosteller (1954), §11.2) and comparison of photographic images to a simulation from a stochastic image model (Ripley (1988), Ch. 6). In each of these examples, the parameters of the stochastic model were fit to data, and simulations were computed with the parameters set to their point estimates. The model checking was useful in these examples, as in ours, because the aspects of the data being checked were not the same as those being fit by the model. In our terminology, we would consider these simulations to be approximations to posterior predictive assessments, using a point estimate of θ rather than its posterior distribution. In more complicated problems, such as with hierarchical models, the plug-in rule can seriously understate posterior uncertainty (as noted by LR), and it is thus important to reflect

such uncertainty in model checking. Rubin's (1984) posterior predictive check then comes naturally, and a key component of Rubin's formulation is to set up posterior predictive checks in terms of replicated data. In our notation, the distribution $p(\theta|y, H)$ determines Bayesian inference conditional on the model H, but one typically needs to define the joint distribution $p(y^{\text{rep}}, \theta|y, H)$ in order to check the model. Posterior predictive assessment thus can be viewed as a part of the Bayesian analysis over the larger probability space that includes y^{rep}.

We echo Draper's emphasis that both practical and statistical significance are important. The former is addressed by the magnitude of the discrepancy, in real terms, compared to its reference distribution; the latter is addressed by the p-value. For example, Gelman et al. (1995), §8.4 check the fit of a regression model fit to U.S. election data using, as a test statistic, the proportion of "party switches": elections in which the candidate of the incumbent party loses. The observed proportion of party switches is $T(y) = 0.126$; by comparison, in the posterior predictive simulations, 95% of the values of $T(y^{\text{rep}})$ fall in the range $[0.130, 0.143]$. The observed value of $T(y)$ falls far in the tail of the posterior predictive distribution of $T(y^{\text{rep}})$—the lack of fit is thus statistically significant—but, for the purposes for which the model was being used (estimating the advantage of incumbency), a misfit of 0.01 in the proportion of switches is minor and not practically significant. We see no problem with noting this lack of fit (which may be useful to know about if we seek to expand the model or understand the data better), while still finding the model useful.

Since writing this paper, we have made posterior predictive assessment a standard part of our applied model fitting (examples appear in Gelman et al. (1995), Ch. 6, 8, 13, 16, and 18). In most cases, the plot of the realized discrepancy compared to the reference distribution is far more valuable than the p-value itself. Even with simple methods such as plots of residuals, it can be useful in complicated models to use the posterior predictive distribution as a reference distribution for comparison to the data. LR's examples, including one that predates our paper, further illustrate the usefulness and flexibility of posterior predictive assessments. There have also been posterior predictive assessments in Bayesian modeling stimulated by our paper (the first version of which was completed in 1992; see, for example, Green et al. (1994) and Vounatsou and Smith (1995)). We expect that posterior predictive assessment, like Bayes methods in general, will become increasingly common as users of statistical models realize how general the idea is and how easily it can be applied.

Model Checking Without an Explicit Alternative is Inevitable

At the end of a Bayesian analysis, we summarize our inference by the joint posterior distribution of all unknown quantities. (Sensitivity analysis is a set

of such posterior distributions.) This posterior inference can be the result of a series of model improvements and expansions, of model averaging, or of model selection via Bayes factors or other criteria. To be confident that we have arrived at a reasonable Bayesian inference, we need (at least) to know whether this final model fails to capture important features of the data. That this model survived the model building process does not logically assure us that it will be in good agreement with the data, unless we have included goodness-of-fit assessment as part of the model improvement process that led to the final model. Once we are at the end of the analysis, either because we have exhausted our model improvement effort or because we have reached a time constraint [Draper], we are not able to consider any (further) explicit alternative models. However, we can and should check the fitness of the final model to the data, and report any serious discrepancy we find. We may (as KW suggest) keep this model for our current needs despite some identified model failures. For serious decision problems [Draper; Hill], a final check on the model is particularly important, because the ultimate decision will condition on this final model (or models) being acceptable.

For example, in our mortality example, if we have to stop after fitting the monotone-convexity model, we can still check it against the observed rates, as we did, without formulating any explicit alternatives. We indeed had several implicit alternatives in mind (e.g., non-convexity; alternative prior distributions requiring convexity but not favoring quadratic curves) when we chose the discrepancy variables χ^2 and y_{64} [Draper]. But there is no need, and in fact it is often impossible, to construct a model to quantify exactly an implicit alternative (e.g., "not monotone") and then carry out the calculation of Bayes factors [Draper; Hill; KW]. We agree with Draper that our Figures 4–5, which represent a form of model checking without explicit alternatives, are more informative than aggregate discrepancy measures for detecting the lack of fit and for suggesting directions for improvement [KW; LR; Hill]. We chose such an example precisely because it allows direct validation of the findings of the posterior predictive assessment based on simple discrepancies. For more complex problems, such direct plots may not be feasible (e.g., we may not have a scalar quantity, such as age, to plot against). We certainly do not suggest that the discrepancy approach be used as the only model check. Examining the posterior distribution with respect to subject matter expertise will generally be a first check of model adequacy (and sometimes programming accuracy). The more methods we have, the better, as long as we apply each for the purpose for which it is designed. Hill summarizes well for the posterior predictive assessment: "Its primary function is to alert one to the need of making ... more careful analysis, and perhaps to search for better models."

We certainly agree that flaws discovered during model diagnosis should be used to suggest alternative models whenever feasible [LR; KW; Hill]. As Box

and Tiao (1973) and Gelman et al. (1995) indicate, model building, like scientific investigation in general, is an iterative process. Included in this process is the possibility of finding flaws in a model before a better explicit alternative has appeared to replace it, suggesting more experiments or the need for better understanding of the underlying process from a substantive perspective. A "rejected" model might be useful for many purposes (e.g., Newton's law) while its rejection encourages us to continue searching for an improved model (e.g., Einstein's relativity).

Flexibility is Necessary in Model Checking

The fact that models are posited as approximations to the truth for specific purposes suggests that a sensible model checking technique must be flexible enough to accommodate different needs. For example, the question of how unusual a realized discrepancy must be before one casts doubt about the aspect of the model being checked [KW] must be judged in the practical context depending on the utility [Draper] and the nature of the data. One problem that is brought up repeatedly is that, in practice, any model being fit to real data can be "rejected" by a goodness-of-fit test if the number of data points is large enough. The resolution of this problem is that there is no need to abandon a model if its lack of fit is minor in real terms, but it is still useful to discover problems with the model with an eye to ultimately improving it (as in the previous election example).

An important feature of the method we advocate is the flexibility in choosing discrepancies and replications to diagnosis in what way the posited models are inadequate [LR]. We agree with KW's statement that "it might seem sensible to remove our global notion of acting as if a model were true and replace it with a qualified version, acting instead as if a model were true *for some specified purpose*." This is indeed the purpose of classical goodness-of-fit testing and its generalization, posterior predictive assessment—the purpose of the checking is not simply to "accept" or "reject", but rather to discover what aspects of the data are not being fit well by the model. It is clear that this goal cannot be achieved with a fully automated diagnosis method.

In practice, we strongly suggest the use of more than one discrepancy, some of which might be omnibus measures and others which might be problem-specific (e.g., the χ^2 and y_{64} in the mortality example). Recall that measures need not be distances [LR]. Multiple discrepancy measures allow for the diagnosis of various aspects of the posited models that need improvement [Draper; KW; LR]. We welcome any suggestions that encourage multiple diagnostic checks (e.g, LR's use of subsets and Draper's separation of likelihood and prior density). We emphasis that there is no need to (formally) weight the findings from different discrepancies

to provide a single summary, because they are designed to be sensitive to different aspects of the model [KW].

Rubin's comments on the conservatism of posterior predictive p-values, the related concept of "confidence validity", and the possible effects of varying the definition of replications raise issues that require statisticians to move beyond seeking unique, optimal, and automated procedures. While it is useful to seek general rules that serve some common purposes (e.g., the χ^2 tests in our paper or the general discrepancies mentioned by LR), it is important for statistical methods to retain the flexibility to meet the needs of a wide range of scientific problems [Rubin]. Hill's post-data analysis is also an attempt to make statistical methods flexible enough for complex real-life applications. Also, there is a need for theoretical studies of phenomena that appear to arise from the resulting flexible statistical framework, such as "superefficiency" [Rubin].

Even for the task of model comparison, which KW and, to some extent, LR, suggest should be reserved for Bayes factors, we believe predictive distributions can have a useful role in practice; this topic is pursued further by Bernardo and Smith (1994) and Draper (1995).

Every Method Requires Caution

The essence of a goodness-of-fit checking is the comparison between data (or, more generally, a realized discrepancy) and a reference distribution. This comparison can often be done informally or graphically, as in the examples of Bush and Mosteller (1954) and Ripley (1988) cited earlier, as well as the figures in our paper and in the discussions by Draper and LR. We agree with Draper that other numerical summaries of the plots can be useful, such as LR's upper and lower tail-area probabilities for discrete discrepancies. We welcome Draper's suggestion of using the density ratio in the form $f_{\max}/f(d_{\text{obs}})$, but this approach has its own drawbacks. It is not applicable when the discrepancy depends on both parameter and data, a central theme of our proposal. In addition, unlike the p-value, the density ratio does not have a direct interpretation as a posterior probability.

Draper also suggests cross-validation, which can be a useful diagnostic tool. However, we do not share Draper's view that cross-validation is a more natural approach than the posterior predictive approach and the implication that cross-validation is more suitable for general practice. It is well known that cross-validation works best with large amounts of "unstructured" data (essentially those that are i.i.d., possibly after some grouping), as discussed in Bailey, Harding, and Smith (1989). For highly structured data or small data sets (relative to the number of parameters), the typical setting of the sophisticated Bayesian models for which the posterior predictive assessment is designed, it is generally

difficult to determine suitable "training" and "test" samples that will provide information relevant to the model's fit to the parent sample. For instance, for the image problem, it is very difficult to decide a training image and a test image that are scientifically relevant for the original image. For the mortality example, which is essentially a non-linear regression problem, Draper's analysis demonstrates the difficulty of using cross-validation for regression analysis. Omitting influential cases can greatly alter the estimated regression, and it is well known that such a cross-validation analysis necessarily shows more lack-of-fit than the full-data analysis (e.g., Stone (1974), Bailey et al. (1989)).

Regarding KW's suggestion that the posterior distribution of a discrepancy be used rather than hypothetical replications, we actually explored such a possibility at an early stage of our work, using Dempster's choice of discrepancy, the log likelihood function. Our conclusion was that such an approach is essentially only useful for comparing two models. Indeed, in the response to a question raised by Cox on the implementation of his approach, Dempster (1974) made it clear that if two models are being fit to the same data, then the one that yields the posterior density of -2loglikelihood that is far to the left of the other is regarded as a better fit. Although "what constitutes 'far to the left' is not easy to formalize precisely" (Dempster (1974), p. 353), having two posterior densities provides a comparative scale for making such judgments. In contrast, when we only have a single posterior density of a discrepancy, we find the magnitude of the realized discrepancy given the posited model and the observed data, but we do not know the magnitude of the corresponding realized discrepancy when the data are indeed from the posited model. Inspecting a single posterior density of a discrepancy might be sufficient for detecting a lack of fit when one has a rough idea of the "acceptable" magnitude. But it is obviously more desirable to make this determination in a more formal way, for the sake of both validity and calibration, and the posterior predictive check using realized discrepancy is designed for that purpose. It seems to us that in the context of assessing goodness-of-fit of a model for a given data set, hypothetical replications are inevitable.

Concerning Various Examples and Applications

For the shuttle launching problem, we agree fully with Hill that the only sensible thing to do, as statisticians, is to perform a sensitivity study and to present the result to the decision makers. Neither a posterior predictive assessment nor a Bayes factor nor any other statistical method can detect lack of fit of the model for data outside the range of the existing data (e.g., low temperatures).

Regarding the psychology application in Sections 3.3–3.4 questioned by KW, it is tempting to try to embed the two-class model into a suitable four-class model

but this turned out not to be practical given the complicated relationships between the two-class and four-class models dictated by the substantive theory. It was thus decided that it would be more effective, especially for the intended audience of psychologists and statisticians (see Stern et al. (1995); Rubin and Stern (1994)), to treat the four-class model as an implicit alternative when testing the two-class model. It would be interesting to see alternative statistical methods applied to this data set with the same scientific goals, especially to see if more sophisticated methods can provide a better scientific understanding of infant temperament. This example indicates that the line between implicit and explicit alternatives is not absolute, and it also reinforces the importance of using substantive knowledge in applied statistics.

The importance of scientific knowledge also is evident in the tomographic image problem. Although we agree in general with WT that model constraints can and should be considered before seeing any data (e.g., Gelman (1996)), the log transformation does not remove our difficulties in this example, because the expected data are a linear transformation of the image on the untransformed scale. Taking logs removes the positivity constraint but at the cost of making the problem nonlinear; under either parameterization, the χ^2 or other natural discrepancies do not have invariant distributions.

WT's ANOVA problem, we believe, can be better handled by a hierarchical linear model, which can be solved using a direct Bayesian approach (see, for example, Gelman et al. (1995), Ch. 5). The posterior predictive assessment is useful for checking the posited hierarchical model, but the inference about the group means should be summarized by their joint posterior distribution. The p-value provided by WT could perhaps be used to check whether an analysis under the assumption of equal means provides a useful approximate analysis for those situations when it would be a major investment for the analyst to carry out the more general analysis.

Hill's variance-component model provides another good example of the need for model assessment. There the usual variance component model is fit to a data set with the result that the sum of squares between group means is smaller (perhaps much smaller) than would be implied by the within-group variance. This is often taken as evidence of zero between-group variance while Hill suggests the data may be indicating negative correlation of errors within groups (for problems in which such negative correlation is plausible). Since the negative correlation model is quite a bit more complex, one can easily imagine fitting a variance components model first and assessing its adequacy. A diagnostic measure like the range of group means might point out the lack of fit of the variance components model and the statistician, perhaps with a subject matter expert, would then need to determine if the result is due to chance, negative correlation, or some

other explanation. Glickman and Stern (1995) present a posterior predictive assessment of this sort in analyzing American football scores but conclude that negative correlation is not a plausible alternative in that case. Hill's model is a very useful alternative but beginning with that model in all variance component problems does not seem to be the best modeling strategy. It should also be pointed out that the negative correlation model would need to be checked against the observed data as well.

What Jeffreys (1961) Did (and Said)

The optimality of "orthodox" Bayesian inference [KW], like any optimality result, is necessarily built upon the assumption that the underlying model is true. Hypothetical replications, with or without tail-area calculations, are used to check such an assumption. We thus do not see any conflict between such results and the use of posterior predictive model checks; the situation is very much like the debate on the likelihood principle—adopting the likelihood principle does not require one to blindly adopt a particular likelihood (see McCullagh (1995)). Although it is unnecessary to cite anyone to support this obvious point, we would like to conclude our rejoinder by mentioning what Jeffreys (1961) did in analyzing data, since his work was quoted several times by discussants [Hill; KW; LR]. We hope this citation will show that there is indeed no conflict between being a Bayesian and using hypothetical replications and tail-area probabilities (of course, here our model assumes that the same Jeffreys was the author for all the quoted pieces). In a study assessing whether correction of an astronomical constant was required, after tabulating weighted individual deviations from a weighted mean, Jeffreys (1961, p. 305) wrote "The weighted mean is +0.69 and gives $\chi^2 = 1043/7.7^2 = 16.9$ on 7 degrees of freedom. This is beyond the 2 per cent. point, and is enough to arouse suspicion."

Bailey, R. A., Harding, S. A. and Smith, G. L. (1989). Cross-validation. In *Encyclopedia of Statistical Science. Supplement Volume* (Edited by S. Kotz and N. L. Johnson), 39-44, John Wiley, New York.

Bayarri, M. J. (1985). A Bayesian test for goodness-of-fit. Unpublished manuscript, University of Valencia.

Bayes, T. (1764). An essay towards solving a problem in the doctrine of chances. Reprinted by W. Edwards Deming. Hafner Publishing Co., 1963, New York.

Berger, J. O. and Delampady, M. (1987). Testing precise hypotheses (with discussion). *Statist. Sci.* **2**, 317-352.

Berger, J. O. and Sellke, T. (1987). Testing a point null hypothesis: The irreconcilability of P values and evidence (with discussion). *J. Amer. Statist. Assoc.* **82**, 112-139.

Bernardo, J. M. (1979). Expected information as expected utility. *Ann. Statist.* **7**, 686-690.

Bernardo, J. M. (1982). Bayesian model testing from a Bayesian point of view (in Spanish). *Trab. Estadist.* **32**, 16-30.

Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Theory.* John Wiley, New York.

Box, G. E. P. and Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis.* John Wiley, New York.

Bush, R. R. and Mosteller, F. (1954). *Stochastic Models for Learning.* John Wiley, New York.

Cox, D. R. and Lewis, P. A. W. (1966). *The Statistical Analysis of Series of Events.* Methuen, London.

Cressie, N. A. C. (1993). *Statistics for Spatial Data* (revised edition). John Wiley & Sons, New York.

Dawid, A. P. (1984). Statistical theory. The prequential approach. *J. Roy. Statist. Soc. Ser.A* **147**, 278-292.

Dawid, A. P. (1992). Prequential analysis, stochastic complexity and Bayesian inference (with discussion). In *Bayesian Statistics* 4 (Edited by J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), 109-125, Oxford University Press, Oxford, UK.

Dempster, A. P. (1971). Model searching and estimation in the logic of inference. In *Foundations of Statistical Inference* (Edited by V. P. Godambe and D. A. Sprott, Holt, Rinehart, and Winston: Toronto), 56-81.

Dempster, A. P. (1974). The direct use of likelihood for significance testing (with discussion). In *Proceedings of Conference on Foundational Questions in Statistical Inference* (Edited by O. Barndorff-Nielsen et al.), 335-354, Department of Theoretical Statistics, University of Aarhus, Denmark.

Dempster, A. P. (1975). A subjectivist look at robustness, *Proc. Internat. Statist. Inst.* **40**, 349-374.

Diggle, P. J.(1983). *Statistical Analysis of Spatial Point Patterns.* Academic Press, New York.

Draper, D. (1995a). Assessment and propagation of model uncertainty (with discussion). *J. Roy. Statist. Soc. Ser.B* **57**, 45-97.

Draper, D. (1995b) Discussion of "Model uncertainty, data mining, and statistical inference," by C. Chatfield. *J. Roy. Statist. Soc. Ser.A* **158**, 419-466.

Evans, M. (1995). Bayesian inference procedures derived via the concept of relative surprise.

Fay, R. E. (1992). When are inferences from multiple imputation valid? *Proceedings of the Survey Research Methods Section of the American Statistical Association,* Alexandria, VA, pp.227-232.

Fay, R. E. (1996). Valid inferences from imputed survey data. To appear in *J. Amer. Statist. Assoc.*

Geyer, C. J. (1992). Practical Markov chain Monte Carlo. *Statist. Sci.* **7**, 473-483.

Glickman, M. and Stern, H. (1995). An autoregressive state-space model for National Football League (NFL) scores. Technical Report, Department of Statistics, Iowa State University.

Good, I. J. (1953). The appropriate mathematical tools for describing and measuring uncertainty. In *Uncertainty and Business Decisions*, (Edited by C. F. Carter, G. P. Meredith and G. L. S. Shackle). Liverpool: University Press.

Good, I. J. (1971). 46656 varieties of Bayesians, *Amer. Statist.* **25**, 62-63.

Green, E. J., Roesch, F. A., Smith, A. F. M. and Strawderman, W. E. (1994). Bayesian estimation for the three-parameter Weibull distribution with tree diameter data. *Biometrics* **50**, 254-269.

Harrison, P. J. and Stevens, C. F. (1976). Bayesian forecasting (with discussion). *J. Roy. Statist. Soc. Ser.B* **38**, 205-247.

Hill, B. M. (1965). Inference about variance components in the one-way model. *J. Amer. Statist. Assoc.* **58**, 918-932.

Hill, B. M. (1967). Correlated errors in the random model. *J. Amer. Statist. Assoc.* **62**, 1387-1400.

Hill, B. M. (1980). Robust analysis of the random model and weighted least squares regression. In *Evaluation of Econometric Models* (Edited by J. Kmenta and J. Ramsey), 197-217 Academic Press, New York.

Hill, B. M. (1985-86). Some subjective Bayesian considerations in the selection of models (with discussion). *Econometric Rev.* **4**, 191-288.

Hill, B. M. (1990). A theory of Bayesian data analysis. In *Bayesian and Likelihood Methods in Statistics and Econometrics: Essays in Honor of George A. Barnard* (Edited by S. Geisser, J. S. Hodges, S. J. Press, A. Zellner), 49-73, North-Holland.

Hill, B. M. (1993). Dutch books, the Jeffreys-Savage theory of hypothesis testing, and Bayesian reliability. In *Reliability and Decision Making* (Edited by R. Barlow, C. Clarotti and F. Spizzichino), Chapman & Hall.

Hill, B. M. (1994a). On Steinian shrinkage estimators: The finite/infinite problem and formalism in probability and statistics. In *Aspects of Uncertainty* (Edited by P. Freeman and A. F. M. Smith), 223-260, John Wiley & Sons, Ltd.

Hill, B. M. (1994b). Bayesian forecasting of economic time series. *Econometric Theory* **10**, 483-513, Cambridge University Press.

Hodges, J. S. (1987). Uncertainty, policy analysis, and statistics (with discussion). *Statist. Sci.* **2**, 259-291.

Jeffreys, H. (1961). *Theory of Probability* (3rd edition). Oxford University Press, London.

Jeffreys, H. (1980). Some general points in probability theory. In *Bayesian Analysis in Probability and Statistics* (Edited by A. Zellner), 451-453, North-Holland Publishing Company.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *J. Amer. Statist. Assoc.* **90**, 773-795.

Kass, R. E., Tierney, L. and Kadane, J. B. (1989). Approximate methods for assessing influence and sensitivity in Bayesian analysis. *Biometrika* **76**, 663-675.

Kass, R. E. and Wasserman, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *J. Amer. Statist. Assoc.* **90**, 928-934.

Ledwina, T. (1994). Data-driven version of Neyman's smooth test of fit. *J. Amer. Statist. Assoc.* **89**, 1000-1005.

Lewis, S. M. (1994). Multilevel modeling of discrete event history data using Markov chain Monte Carlo methods. Unpublished Ph.D.dissertation. Department of statistics, University of Washington, Seattle, WA.

Lindley, D. V. (1965). *Probability and Statistics* 2. *Inference*. Cambridge University Press.

Lindley, D. V. (1968). The choice of variables in multiple regression (with discussion). *J. Roy. Statist. Soc. Ser.B* **30**, 31-66.

McCullagh, P. (1995). Discussion of papers by Reid and Zeger and Liang. *Statist. Sci.* **10**, 177-179.

Meng, X. L. (1994a). Multiple imputation inferences with uncongenial sources of input (with discussion). *Statist. Sci.* **9**, 538-574.

Neyman, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *J. Roy. Statist. Soc. Ser.A* **97**, 558-606.

Neyman, J. (1937). "Smooth" test for goodness of fit. *Skandinavisk Aktuarietidskrift* **20**, 150-199.

Raftery, A. E. (1987). Inference and prediction for a general order statistic model with unknown population size. *J. Amer. Statist. Assoc.* **82**, 1163-1168.

Raftery, A. E. (1988). Analysis of a simple debugging model. *Appl. Statist.* **37**, 12-22.

Raftery, A. E. (1989). Comment: Are ozone exceedance rates decreasing? *Statist. Sci.* **4**, 378-381.

Raftery, A. E. (1995). Bayesian model selection in social research (with discussion). In *Sociological Methodology 1995* (Edited by P. V. Marsden), **25**, 111-196, Blackwell, Oxford, UK.

Raftery, A. E., Lewis, S. M. and Aghajanian, A. (1995). Demand or ideation? Evidence from the Iranian marital fertility decline. *Demography* **32**, 159-182.

Raftery, A. E., Lewis, S. M., Aghajanian, A. and Kahn, M. J. (1996). Event history modeling of World Fertility Survey data. To appear in *Mathematical Population Studies.*

Ripley, B. D. (1977). Modelling spatial patterns (with discussion). *J. Roy. Statist. Soc. Ser.B* **39**, 172-212.

Ripley, B. D. (1981). *Spatial Statistics.* John Wiley & Sons, New York.

Ripley, B. D. (1988). *Statistical Inference for Spatial Processes.* Cambridge University Press, New York.

Rubin, D. B. (1996a). More powerful randomization-based p-values in double-blind trials with noncompliance. To appear in *Statist. Medicine.*

Rubin, D. B. (1996b). Multiple imputation after 18+ years (with discussion). *J. Amer. Statist. Assoc.* **91**, 473-520.

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

Spiegelhalter, D. J. (1995). Discussion of "Assessment and propagation of model uncertainty," by D. Draper. *J. Roy. Statist. Soc. Ser.B* **57**, 45-97.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions (with discussion). *J. Roy. Statist. Soc. Ser.B* **45**, 133-150.

Verdinelli, I. and Wasserman, L. (1995). Bayesian goodness of fit testing using infinite dimensional exponential families. Manuscript in progress.

Vounatsou, P. and Smith, A. F. M. (1995). Bayesian analysis of ring-recovery data via Markov chain Monte Carlo simulation. *Biometrics* **51**, 687-708.

Weerahandi, S. (1995). ANOVA under unequal error variances. *Biometrics* **51**, 589-599.

Wrinch, D. and Jeffreys, H. (1921). On certain fundamental principles of scientific inquiry. *Philosophical Magazine* **42**, 369-390.