Bayesian Tests for Goodness of Fit Using Tail Area Probabilities

By

Andrew Gelman
Department of Statistics
University of California
Berkeley, CA 94720

Xiao-Li Meng
Department of Statistics
University of Chicago
Chicago, IL 60637

Hal S. Stern
Department of Statistics
Harvard University
Cambridge, MA 02138

Technical Report No. 372
October 1992

Department of Statistics
University of California
Berkeley, California 94720

Bayesian Tests for Goodness of Fit Using Tail Area Probabilities*

Andrew Gelman
Department of Statistics
University of California
Berkeley, CA 94720

Xiao-Li Meng
Department of Statistics
University of Chicago
Chicago, IL 60637

Hal S. Stern
Department of Statistics
Harvard University
Cambridge, MA 02138

October 1, 1992

Abstract

Classical goodness of fit tests, including exact permutation tests, are well-defined and calculable only if the test statistic is a pivotal quantity. Examples of problems where the classical approach fails include models with constraints, Bayesian models with strong prior distributions, hierarchical models, and missing data problems. Using posterior predictive distributions, we systematically explore the Bayesian counterparts of the classical tests for goodness-of-fit and their use in Bayesian model monitoring in the sense of Rubin (1984). The Bayesian formulation not only allows a tail-area probability (p-value) to be defined and calculated for any statistic, but also allows a test "statistic" to be a function of both data and unknown (nuisance) parameters (Meng, 1992). The latter allows us to propose the ***realized discrepancy test*** of goodness-of-fit, which directly measures the true discrepancy between data and the model, for any aspect of the model. We demonstrate how to compute the tail-area probability for the Bayesian test using simulation, and compare different versions of the test using the χ^2 discrepancy for linear models. A frequency evaluation shows that, if the replication is defined by new parameters and new data, then the Type I error of an α-level Bayesian test is typically less than α, and will never exceed 2α. Also, the posterior predictive test is contrasted with ***prior*** predictive testing (Box, 1980).

Three applied examples are considered. In the first example, which is used to motivate the work, we consider fitting the Poisson model to estimate a positron emission tomography image that is constrained to be all-nonnegative. The classical χ^2 test fails because the constrained model does not have a fixed number of degrees of freedom in the usual sense. Under the Bayesian approach, however, goodness-of-fit can be tested directly. The second and third examples illustrate the details of the Bayesian posterior predictive approach in two problems for which no classical procedure is available: estimation in a model with constraints on the parameters, and determining the number

*We thank Donald Rubin for helpful discussions, 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 financial support.

of components in a mixture model. In all three examples, the classical approach fails because the test statistic is not a pivotal quantity: the difficulty is not just how to compute the reference distribution for the test, but that no such distribution exists, independent of the unknown model parameters.

Keywords: Bayesian inference, Bayesian p-value, χ^2 test, contingency table, discrepancy test, likelihood ratio test, mixture model, model monitoring, posterior predictive test, prior predictive test, p-value, realized discrepancy, significance test.

1 Introduction

1.1 Goodness-of-fit tests

Checking the correctness of an assumed model is important in statistics, especially in Bayesian statistics. Bayesian prior-to-posterior analysis conditions on the whole structure (i.e., not just a few moments) of a probability model, and can yield false inferences when the model is false. A good Bayesian analysis, therefore, should at least include some check of the plausibility of the model and its fit to the data. In the classical setting, model checking is often facilitated by a goodness-of-fit test, which quantifies the extremeness of the observed value of a selected measure of discrepancy (e.g., differences between observations and predictions) by calculating a tail-area probability given that the model under consideration is true. This *tail-area probability* is often called the *p-value* or *significance level*, and we will use these terms interchangeably. This paper attempts to study Bayesian versions of the usual Fisherian goodness-of-fit tests and explore their uses in Bayesian model monitoring in the sense of Rubin (1984).

As pointed out in Rubin (1984), a goodness-of-fit test can be based on any "test statistic" or function of the data, and the choices depend on what aspects or characteristics of the models are considered to be important for the problems under study. For example, if one wishes to directly compare a set of observations with predictions, then a χ^2 test on the residuals might be appropriate. On the other hand, the Kolmogorov-Smirnov test can be useful to test the goodness-of-fit of an estimate of a continuous distribution function.

Because in the classical setting a test statistic cannot depend on any unknown quantities, these comparisons are actually made between the data and the *best-fit distribution* (typically maximum likelihood) within the family of distributions being tested. Our Bayesian formulation will allow a discrepancy measure to be a function of both data and unknown parameters, and thus allow more direct comparisons between the sample and population characteristics.

1.2 Difficulties with non-Bayesian methods

In the classical approach, the test statistic—typically a measure of discrepancy between the best-fit model and the data—is calculated, and the desired p-value measure of goodness-of-fit is determined, based on the sampling distribution of the data under the model. The main technical problem with the classical method is that, in general, the p-value depends on the unknown parameters.

For some problems, such as linear models, the common discrepancy tests have exactly known null distributions, or at least good approximations. For example, if the parameters can vary freely in a hyperplane, the log-likelihood has an asymptotic χ^2_{n-k} distribution, where n is the number of data points and k is the number of parameters being fit. Unfortunately, once we move beyond unrestricted linear models, generalized linear models, and so forth, the handy approximations fail, especially for complicated models with many parameters. Thus the sampling distributions of the test statistics depend crucially on the unknown parameters, even for moderately large sample sizes (see our example in Section 2 with a total of 6,000,000 counts). In other words, as is well known, useful test statistics are typically not pivotal quantities.

The classical approach can fail in at least three kinds of models: severe restrictions on the parameters, such as positivity; probabilistic constraints, which arise from a hierarchical model or just a strong prior distribution; and unusual models that cannot be parameterized as generalized linear models. Useful approximations to the distribution of test statistics are possible for simple extensions of the linear model (see, for example, Chernoff, 1954) but are not useful for more realistic models, especially involving many parameters. In fact, computing the distribution of classical goodness-of-fit tests can be difficult even in standard generalized linear models (see McCullagh, 1985, 1986). Once we move beyond the simplest models and asymptotic approximations, no clearly-defined classical p-values exist for useful test statistic, even in the case that the sampling distribution of the test statistic can be calculated exactly.

1.3 Bayesian remedy

When the standard asymptotic-based methods fail, the Bayesian approach, described in this paper, determines a unique significance level for any goodness-of-fit test. That is, given a set of data, a hypothesized model (including a prior distribution), and a goodness-of-fit measure, a unique p-value can be computed, using the posterior distribution under the model. The Bayesian method reproduces the classical results in simple problems with pivotal quantities. The price a non-Bayesian must pay for this logical precision, of course, is the assignment of a prior distribution to the model parameters. Rejecting a Bayesian model is a rejection of the whole package, and one may suspect that one is rejecting the prior distribution rather than the model constraints and the likelihood. We will discuss this issue when presenting the examples.

Of the many versions of Bayesian model monitoring or hypothesis testing in the literature, what we present here is the posterior calculation of tail-area probabilities, an idea introduced by Guttman (1967), applied by Rubin (1981), and given a formal Bayesian definition by Rubin (1984). We also briefly examine p-values based on the prior distribution, as described by Box (1980). Conceptually, the ideas presented here may be thought of as a Bayesian extension of the classical approach to significance testing. In this paper, we focus on testing a whole model, not just a few parameters of it. See Meng (1992) for a discussion of Bayesian and classical tail-area tests for parameter values within a model.

It may seem surprising that Bayesian tail-area probabilities have not been formally applied to discrepancy tests, considering the long history of the χ^2 test and of Bayesian statistics itself. We believe that the two concepts have escaped combination for so long for two reasons: first, significance testing has for a long time been considered non-Bayesian (see, e.g., Berger and Sellke, 1987). While willing to occasionally compute χ^2 tests and the like, Bayesians have not been quite ready to treat them as respectable methods. For instance, Jeffreys (1939) and Jaynes (1978) both use χ^2 tests to good effect, but are mysteriously silent on the connection between the significance probabilities and the Bayesian probability distributions on parameters. Dempster (1974, p. 233) writes, "some Bayesians may feel comfortable switching over to a significance testing mode to provide checks on their assumed models." This is of course what we are proposing; the advance of Rubin (1984) is to give the significance tests a Bayesian interpretation, which, as we show in this paper, provides a framework for theoretical and computational improvements. Good (1967, 1992) recommends tail-area P-values, but only as approximations to Bayes factors.

Second, until recently, Bayesian methods were applied to simple enough models that the usual χ^2 asymptotics sufficed whenever a goodness-of-fit test was desired. As a result of increased experience and computing power, Bayesian model monitoring has become practical for complex models (for example, Smith and Roberts, 1992, recommend using iterative simulation methods to apply the methods of Rubin, 1984), and thus more general methods for monitoring more realistic models are in demand.

1.4 The role of tail-area testing in applied Bayesian statistics

In Bayesian statistics, a model can be tested 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. Tail-area testing addresses only the third of these concerns: even if a model is not rejected by a significance test, it can still be distrusted or discarded for other reasons.

If a data set has an extremely low tail-area probability, we say it has refuted the model (or else an extremely low-probability event has occurred), and it is desirable to improve the model until it fits. It may sometimes be practical to continue inference using an ill-fitting model, but on such occasions we would still like to know that the data do not jibe with the predictions derived from the posterior distribution.

As usual with goodness-of-fit testing, a large p-value (e.g., 0.5) does not mean the model is "true," but only that the model's predictions are not suggested by that test to disagree with the data at hand. Also, "rejection" should not be the end of a data analysis, but rather a time for examining and revising the model.

Where possible, we try to be fully Bayesian, and consider the widest possible model to fit any dataset, so that all choices of "model selection" occur within a large super-model. *Within* any model, we would just compute posterior probabilities, with no need for p-values. However, in practice, even all-encompassing super-models need to be tested for fit to the data. In addition, it is often much more convenient to test a smaller model using tail-area probabilities than to embed it into a reasonable larger class. While not the end of any Bayesian analysis, tail-area tests are useful intermediate steps that, for a little effort, can tell a lot about the relation between posterior distributions and data.

We do not recommend the use of p-values to *compare* models; when two or more models are being considered for a single dataset, we would just apply the full Bayesian analysis to the model class that includes the candidate models as special cases. If necessary, approximations could be used to restrict the model, as in Madigan and Raftery (1991).

For more detailed discussions of Bayesian tail-area testing and related ideas, see Dempster (1971, 1974), Box (1980), and Rubin (1984).

1.5 Outline of the paper

This paper presents Bayesian goodness-of-fit testing as a method of solving the problem of defining and calculating exact significance probabilities, a problem with serious implications when considering whether to accept a probability model to use for inference. Section 2 presents a motivating example from medical imaging where the classical approximation fails. Section 3 defines the Bayesian versions of discrepancy tests, including the *realized*, *average*, and *minimum* discrepancy tests. Section 4 illustrates these approaches for the χ^2 test, where we imagine they will be most frequently applied in practice. Section 5 presents simulation methods for computing the Bayesian p-values as a byproduct of standard methods of simulating posterior distributions. Two real-data applications are presented in Section 6. Section 7 provides some general results about the frequency properties of Bayesian tail-area probabilities. We conclude in Section 8 with a discussion of the implications of the choice of reference distribution, prior distribution, and test statistic for a Bayesian test, including a comparison to the method of Box (1980).

In this paper we discuss only "pure significance tests," with no specified alternative hypotheses. Of course, the test statistic for a pure significance test may be motivated by a specific alternative, as in the likelihood ratio test, but we will only discuss the use of testing to highlight lack of fit of the null model. In particular, we do not cover posterior odds ratios and Bayes factors (e.g., Jeffreys, 1939; Spiegelhalter and Smith, 1982; Berger and Sellke, 1987; Aitkin, 1991), model selection (e.g., Stone, 1974; Raghunathan, 1984; Raftery, 1986), Q-values (Schaafsma, Tolboom, and Van der Meulen, 1989), or other methods that compare the null model to specified alternative hypotheses. We intend to use p-values to judge the fit of a single model to a dataset, *not* to assess the posterior probability that a model is true, and *not* to obtain a procedure with a specified long-run error probability.

2 Motivating example from medical imaging

The rewards of the Bayesian approach are most clear with real-world examples, all of which are complicated. This section provides some detail about an example in which it is difficult in practice and maybe impossible in theory to define a classical p-value. Gelman (1990, 1992a) 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 $n = 22{,}464$ 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 all-nonnegative. In practice, g is discretized into "pixels" and becomes a long all-nonnegative vector, and A becomes a matrix with all-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 all-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, $\hat{\theta} = A\hat{g} + r$.

However, the discrepancy between y and $\hat{\theta}$ should not be great; given the truth of the model, it is limited by the variance in the independent Poisson distributions. To be precise, 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. In fact, $X^2(y;\hat{\theta})$ should be considerably less, since a whole continuous image is being fit to the data. (As is typical for positron emission tomography, $y_i > 50$ for almost all the bins i, and so the χ^2 distribution, based on the normal approximation to the Poisson, is essentially exact.)

The hypothesized model was fit to a real dataset, y, with n = 22,464. We would ultimately like to estimate θ; i.e., to determine the posterior distribution, $P(\theta|y)$, given a reasonable prior distribution. We would also like to examine the fit of the model to the data. For this dataset, the best-fit nonnegative image $\hat{g}$ was not an exact fit; the discrepancy between y and $\hat{\theta} = A\hat{g} + r$ 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.

Consider, now, the following scenario: the experimental procedure is carefully examined, the model is made more accurate, and the new model is fit to the data, yielding a best-fit image $\hat{g}$. Suppose we now calculate the estimate of the cell expectations, $\hat{\theta} = A\hat{g} + r$, and the χ^2 discrepancy, $X^2(y;\hat{\theta})$. How should we judge the goodness-of-fit of the new model? Certainly if $X^2(y;\hat{\theta})$ is greater than $n + 2\sqrt{2n} \approx$ 23,000, we can be almost certain that the model does not fit the data. Suppose, however, $X^2(y;\hat{\theta})$ were to equal 22,000? We should probably still be distrustful of the model (and thus any Bayesian inferences derived from it), since a whole continuous image is being fit to the data. After all, if k linear parameters were fit, the minimum χ^2 statistic would have a χ^2 distribution with $n - k$ degrees of freedom. For that matter, even $X^2(y;\hat{\theta}) = 20{,}000$ might arouse suspicion of a poor fit, if we judge the fitting of a continuous image as equal to at least 3,000 independent parameters.[1] Unfortunately, the positivity constraints in the estimated image do not correspond to any fixed number of "degrees of freedom" in the sense of a linear model.

We have arrived at a practical problem: how to assess goodness-of-fit for complicated models in "close calls" for which the simple χ^2_n bound is too crude. The problem is important, because if we take modeling seriously, we will gradually improve models that clearly do not fit, and upgrade them into close calls. Ultimately, we would like to perform Bayesian inference using a model that fits the data and incorporates all of our prior knowledge.

The current problem, however, is more serious than merely obtaining an accurate approximation. In the classical framework, the p-value (and rejection regions) depend on the unknown parameters, and quite possibly vary so much with the continuous parameters so as to be useless in making a rejection decision. In the next section, we review the posterior predictive approach to hypothesis testing, which allows an exact significance probability to be defined in any setting.

3 Bayesian tests for goodness-of-fit

3.1 Notation and classical p-values

We will use the notation y for data (possibly multivariate), H for the assumed model, and θ for the unknown model parameters (θ may be multivariate, or even infinite dimensional, in a nonparametric model). A classical goodness-of-fit test comprises a *test statistic*, T, that is a function from data space to the real numbers; its observed value, $T(y)$; and the

[1] 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 that 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.

reference distribution of possible values $T(y)$ that could have been observed under H. The p-value of the test is the tail-area probability corresponding to the observed quantile, $T(y)$, in the reference distribution of possible values.

To avoid confusion with the observed data, y, we define y^{rep} as the *replicated* data that *could have been* observed, or, to think predictively, as the data we *would* see if the experiment that produced y today were replicated with the same model, H, and the same value of θ that produced the observed data. We consider the definition of y^{rep} to be part of our joint probability model. In this notation, the *reference set* is the set of possible values of y^{rep}, and the reference distribution of the hypothesis test is the distribution of $T(y^{\text{rep}})$ under the model. In other words, the classical p-value based on T is

$$p_c(y,\theta) = P(T(y^{\text{rep}}) \geq T(y) \mid H, \theta). \tag{2}$$

Here, y is to be interpreted as fixed, with all the randomness coming from y^{rep}.

It is clear that the classical p-value in (2) is well-defined and calculable only if (a) θ is known, or (b) T is a pivotal quantity, that is, the sampling distribution of T is free of the nuisance parameter. Unfortunately, in most practical situations, neither (a) nor (b) is true. Even if a pivotal, or approximately pivotal, quantity exists, it may be no help for testing a different aspect of the model fit. A common practice in the classical setting for handling this dependence on θ is to insert an estimate (typically the maximum likelihood estimate under H) for θ. This approach makes some sense: if the model is true, then the data tell us something about θ, and we should use that knowledge.[2] But it obviously fails to take into account the uncertainty due to unknown parameters. More sophisticated methods such as finding a range of p-values corresponding to the possible range of values of θ can be thought of approximations to the Bayesian approach in the next section.

[2] One might be wary of a method that tests a model H by using an estimate $\hat{\theta}$ that assumes H is true (typically a necessary assumption for any estimate, certainly if a standard error is included too). This is, however, not a good reason for suspicion; the point of testing a model, or any null hypothesis, is to *assume it is true* and check how surprising the data are under that assumption.

3.2 Bayesian testing using classical test statistics

In the Bayesian framework, the (nuisance) parameters are no longer a problem, because they can be averaged out using their posterior distribution. More specifically, a Bayesian model has, in addition to the unknown parameter θ, a prior distribution, $P(\theta)$. Once the data, y, have been observed, θ is characterized by its *posterior distribution*, $P(\theta|y)$. Under the Bayesian model, the reference distribution of the future observation y^{rep} is averaged over the posterior distribution of θ:

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

The tail-area probability under this reference distribution is then,

$$\begin{aligned} p_b(y) &= P(T(y^{\text{rep}}) \geq T(y)|H,y) & (3) \\ &= \int P(T(y^{\text{rep}}) \geq T(y)|H,\theta)P(\theta|H,y)d\theta \\ &= \int p_c(y,\theta)P(\theta|H,y)d\theta, & (4) \end{aligned}$$

that is, the classical p-value of (2), averaged over the posterior distribution of θ. This is the significance level, based on the posterior predictive distribution, defined by Rubin (1984).

Clearly, the Bayesian and non-Bayesian p-values are identical when T is a pivotal quantity under the model, H. In addition, the common non-Bayesian method of inserting $\hat{\theta}$ is asymptotically equivalent to the Bayesian posterior predictive test, given the usual regularity conditions. For any problem, the Bayesian test has the virtue of defining a unique significance level, not just some estimates or bounds, and can be computed straightforwardly, perhaps with the help of simulation, as demonstrated in Section 5.

The posterior predictive distribution is indeed the replication that the classical approach intends to address, although it cannot be quantified in the classical setting if there are unknown parameters. Figure 1a shows the posterior predictive reference set, which corresponds to repeating the experiment tomorrow with the same model, H, and same (unknown) value of θ that produced today's data, y. Because θ is unknown, its posterior distribution is averaged over. (Figures 1b and 1c are discussed in Sections 8.1 and 8.5, respectively, and can be ignored here.)

3.3 Bayesian testing using generalized test statistics

The Bayesian formulation not only helps solve the problem of nuisance parameters, a problem that the classical approach almost always faces, but also allows us to generalize further by defining test statistics that are a function of both data, y, and the true (but unknown) model parameters, θ. This generalization beyond the Rubin (1984) formulation is important because it allows us directly to compare the discrepancy between the observed data and the true model, instead of between the data and the best fit of the model. It also, as we shall show in Section 5, simplifies the computations of significance levels.

Let $D(y;\theta)$ be a discrepancy measure between sample and population quantities. If we take D as a generalized test statistic and put it in the place of T in (3), we can formally obtain a tail-area probability of D under the posterior reference distribution:

$$\begin{aligned} p_b(y) &= P(D(y^{\mathrm{rep}};\theta)\geq D(y;\theta)|H,y) && (5)\\ &= \int P(D(y^{\mathrm{rep}};\theta)\geq D(y;\theta)|H,\theta)P(\theta|H,y)d\theta. && (6)\end{aligned}$$

This p-value measures how extreme is the *realized* value of the discrepancy measure, D, among all its possible values that could have been realized under H with the same value of θ that generates current y. Interestingly, although the realized value is not observed, the Bayesian p-value, p_b, is well defined and calculable. The reference set for the generalized test statistic is the same as that in Figure 1a, except that it is now composed of pairs $(y^{\mathrm{rep}},\theta)$ instead of just y^{rep}. (The term "realized" discrepancy is taken from Zellner, 1975.)

The formulation of generalized test statistics also provide a general way to construct classical test statistics, that is, test statistics that do not involve any unknown quantities. For example, as illustrated in Section 4 for the χ^2 test, the classical statistics that arise from comparing data with the best fit under the null typically correspond to the *minimum discrepancy*:

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

Another possibility is the *average discrepancy* statistic,

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

The corresponding Bayesian p-values are defined by (3) with T being replaced by D_{min} and D_{avg}, respectively. The comparison between D_{min} and D_{avg}, as well as with D itself, will be made in the next section in the context of the χ^2 test.

4 Bayesian χ^2 tests

4.1 General case

We now consider a specific kind of discrepancy measure, the χ^2 discrepancy, by which we simply mean a sum of squares of standardized residuals of the data with respect to their true, unknown expectations. For simplicity, we 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. The χ^2 discrepancy is then,

$$X^2(y;\theta) = \sum_{i=1}^{n} \frac{(y_i - E(y_i|\theta))^2}{\mathrm{var}(y_i|\theta)}. \tag{7}$$

For example, the discrepancy in equation (1) in Section 2 is just the above formula for the Poisson distribution, evaluated at the estimate $\hat{\theta}$. For this section, we will assume that, given θ, expression (7) has an approximate χ^2_n distribution.

Now suppose that we are interested in testing a model, H, that constrains θ to lie in a subspace of $\Re^n$. Given a prior distribution, $P(\theta)$, on the subspace, we can calculate the Bayesian p-value based on X^2 as,

$$p_b(y) = \int P(\chi^2_n \geq X^2(y;\theta))P(\theta|H,y)d\theta, \tag{8}$$

where χ^2_n represents a chi-squared random variable with n degrees of freedom. The probability inside the integral is derived from the approximate χ^2_n distribution of X^2 given θ.

Similarly, one can use $X^2_{\min}$ or X^2_{avg} as the test statistic in place of X^2 to calculate the corresponding Bayesian p-values. The computations, however, are more complicated in general, because the sampling distributions of $X^2_{\min}$ and X^2_{avg} given θ are generally intractable, and more simulations are needed beyond those for drawing θ from its posterior density; see Section 5 for more discussion of this point. The minimum discrepancy statistic, $X^2_{\min}$, is roughly equivalent to the classical goodness-of-fit test statistic; both are approximately pivotal quantities for linear and loglinear models.[3]

4.2 χ^2 tests with a linear model

Suppose H is a linear model; that is, θ is constrained to lie on a hyperplane of dimension k, an example that is interesting in its own right and is also important as an asymptotic distribution for a large class of statistical models. For the linear model, it is well known that the minimum χ^2 discrepancy, $X^2_{\min}(y)$, is approximately pivotal with a χ^2_{n-k} distribution (Pearson, 1900; Fisher, 1922; Cochran, 1952). The Bayesian p-value in this case is just $P(\chi^2_{n-k} \geq X^2_{\min}(y))$.

If θ is given a noninformative uniform prior distribution in the subspace defined by H, then the tests based on $X^2(y;\theta)$ and $X^2_{\text{avg}}(y)$ are closely related to the test based on $X^2_{\min}(y)$. With the noninformative 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,

$$\begin{aligned} X^2_{\text{avg}}(y) &= E(X^2(y;\theta)|y) \\ &= E(X^2_{\min}(y) + (X^2(y;\theta) - X^2_{\min}(y))|y) \end{aligned}$$

[3] The classical χ^2 test is sometimes evaluated at the maximum likelihood estimate and sometimes at the minimum-χ^2 estimate, a distinction of some controversy (see, e.g., Berkson, 1980); we consider the minimum χ^2 test in our presentation, but similar results could be obtained using the maximum likelihood estimate.

$$\approx \quad X^2_{\min}(y) + k,$$

and thus the average χ^2 test is equivalent to the minimum χ^2 test, with the reference distribution and the test statistic shifted by a constant, k.

For the generalized test statistic, $X^2(y;\theta)$, the same decomposition can be applied to the Bayesian p-value formula (8),

$$\begin{aligned} p_b(y) &= \int P(\chi^2_n \geq X^2(y;\theta))P(\theta|H,y)d\theta \\ &= \int P(\chi^2_n \geq X^2_{\min}(y) + (X^2(y;\theta) - X^2_{\min}))\ P(\theta|H,y)d\theta \\ &= P(\chi^2_n \geq X^2_{\min}(y) + \chi^2_k) \qquad (9) \\ &= P(\chi^2_n - \chi^2_k \geq X^2_{\min}(y)), \end{aligned}$$

where χ^2_n and χ^2_k are independent χ^2 random variables with n and k degrees of freedom, respectively. In other words, testing the linear model using the generalized test statistic is equivalent to using the minimum χ^2 test statistic, but with a different reference distribution. The reference distribution derived from $\chi^2_n - \chi^2_k$ has the same mean but a larger variance than the more familiar χ^2_{n-k} distribution, and so a test using the generalized test statistic is more conservative. For any given data set and a uniform prior distribution, the minimum (or expected) χ^2 test will yield a more extreme p-value than the test based on the χ^2 discrepancy with respect to the unknown θ.

The reference distribution of $X^2_{\min}$ depends only on $n - k$, while the reference distribution for the realized discrepancy, X^2, also depends on the number of bins, n. For any fixed number of degrees of freedom, $n - k$, the difference between the two reference distributions increases with n, with the realized discrepancy test less likely to reject because it is essentially a randomized test, with independent χ^2_k terms in each side of (9). Suppose, for example, $n = 250$, $k = 200$, and data y are observed for which $X^2_{\min}(y) = 80$. Under the minimum discrepancy test, this is three standard deviations away from the mean of the χ^2_{50} reference distribution—a clear rejection. The corresponding reference distribution for the realized discrepancy test is $\chi^2_{250} - \chi^2_{200}$, which has a mean of 50 and standard deviation of 30, and the data do not appear to be a surprise at all.

What is going on here? The rejection under the minimum discrepancy test is real—the model does not fit this aspect of the data. Specifically, the data are not as close to the best fitting model, as measured by $X^2_{\rm min}$, 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. After all, in applied statistics one rarely expects a model to be "truth," and it is often said that a rejection by a χ^2 test should not be taken seriously when the number of bins is large. 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.

If some prior knowledge of θ is available, as expressed by a nonuniform prior distribution, the Bayesian test for $X^2_{\rm min}$ is the same, since $X^2_{\rm min}$ is still a pivotal quantity, but the tests based on $X^2_{\rm avg}$ and X^2 now change, as the tests are now measuring discrepancy from the prior model as well as the likelihood. Sensitivity of the tests to the prior distribution is discussed in Section 8, in the context of our applied examples.

5 Computation of Bayesian p-values

5.1 Computation using posterior simulation

Simulation is often used for applied Bayesian computation: inference about the unknown parameter, θ, is summarized by a set of draws from the posterior distribution, $P(\theta|y)$. As described by Rubin (1984), the posterior predictive distribution of a test statistic, $T(y)$, can be calculated as a byproduct of the usual Bayesian simulation by (1) drawing values of θ from the posterior distribution, (2) simulating $y^{\rm rep}$ from the sampling distribution, given θ, and (3) comparing $T(y)$ to the sample cumulative distribution function of the set of values $T(y^{\mathrm{rep}})$ from the simulated replications. This procedure is immediate as long as the test statistic is easy to calculate from the data. For example, Rubin (1981) tests a one-way random effects model by comparing the largest observed data point to the distribution of largest observations under the posterior predictive distribution of datasets. Using the same approach, Belin and Rubin (1992) use the average, smallest, and largest within-subject variances to test a family of random-effects mixture models.

Here we present algorithms that use draws from the posterior distribution to compute Monte Carlo estimates of posterior tail-area probabilities. We first present methods for computing Bayesian p-values based on the realized discrepancy $D(y;\theta)$. Typical examples of the realized discrepancy functions are the χ^2 discrepancy (7), the likelihood ratio (measured against a fixed alternative, such as a saturated model for a contingency table), and the maximum absolute difference between modeled and empirical distribution functions (the Kolmogorov-Smirnov discrepancy). We also consider the computations for the minimum discrepancy, which is most commonly used in classical tests; and the average discrepancy, where the average is taken over the posterior distribution of θ.

5.2 Computation for the realized discrepancy

There are two ways to simulate a p-value based on the realized discrepancy, $D(y;\theta)$. The first method is to simulate the tail-area directly using the joint distribution of y^{rep} and θ, as follows:

1. Draw θ from the posterior distribution, $P(\theta|H,y)$.

2. Calculate $D(y;\theta)$.

3. Draw y^{rep} from the sampling distribution, $P(y^{\mathrm{rep}}|H,\theta)$. We now have a sample from the joint distribution, $P(\theta,y^{\mathrm{rep}}|y)$.

4. Calculate $D(y^{\text{rep}}; \theta)$.

5. Repeat steps 1–4 many times. The estimated p-value is the proportion of times that $D(y^{\text{rep}}; \theta)$ exceeds $D(y; \theta)$.

The steps above generally require minimal computation beyond the first step of drawing θ, which might be difficult, but will often be performed anyway as part of a good Bayesian analysis. For most problems, the draws from the sampling distribution in step 3 are easy. In addition, once they have been drawn, the samples from (θ, y^{rep}) can be used to obtain the significance probability of any test statistic, that is, they can be used for simulating p-values for several realized discrepancy measures.

Alternatively, if the classical p-value based on $D(y; \theta)$ (i.e., treating θ as known) is easy to calculate analytically, then one can simulate the Bayesian p-value more efficiently by the following steps:

1. Draw θ from the posterior distribution, $P(\theta | H, y)$.

2. Calculate the classical p-value, $p_c(y, \theta) = P(D(y^{\text{rep}}; \theta) \geq D(y; \theta) | \theta)$, where the probability distribution of y^{rep} is taken over its sampling distribution, conditional on θ.

3. Repeat steps 1–2 many times. The estimated p-value is the average of the classical p-values determined in step 2.

Step 2 requires the computation of tail-area probabilities corresponding to quantiles from a reference distribution which, in general, can be a function of θ. For some problems, such as the χ^2 test with moderate sample size, the sampling distribution of the discrepancy statistic, given θ, is independent of θ. In these cases, the cumulative distribution function of the discrepancy may be tabulated "off-line," and step 2 is tractable even if it requires numerical evaluation.

5.3 Computation for the minimum and average discrepancies

Simulating p-values for D_{min} and D_{avg} is more complicated because one has to eliminate θ when evaluating the values of the test statistics. For D_{min}, the following steps are needed:

1. Draw θ from the posterior distribution, $P(\theta|H, y)$.

2. Draw y^{rep} from the sampling distribution, $P(y^{\text{rep}}|H, \theta)$. We now have a sample from the joint distribution, $P(\theta, y^{\text{rep}}|y)$.

3. Determine the value $\hat{\theta}$ for which $D(y^{\text{rep}}; \hat{\theta})$ is minimized. Calculate $D_{\text{min}}(y^{\text{rep}}) = D(y^{\text{rep}}; \hat{\theta})$.

4. Repeat steps 1–3 many times, to yield a set of values of $D_{\text{min}}(y^{\text{rep}})$. This is the reference set for the posterior predictive test; the estimated p-value is the proportion of simulated values of $D_{\text{min}}(y^{\text{rep}})$ that exceed the observed $D_{\text{min}}(y)$.

The main computational drawback to the use of D_{min} is the required computation in step 3, where estimating $\hat{\theta}$ from y^{rep} may itself require iteration.

Simulating the p-value for the average discrepancy statistic, $D_{\text{avg}}(y)$, is somewhat more difficult, since it is, after all, a more complicated statistic than D_{min}. Mimicking steps 1–4 above is generally infeasible, since now step 3 would require, not just a minimization, but a full simulation (or integration) over the posterior distribution of θ, given y^{rep}. In other words, one may need to perform nested simulation.

In addition, if the posterior distribution for θ given y^{rep} can be multimodal, the optimization, simulation, or integration required to compute p-values for D_{min} or D_{avg} can be difficult or slow compared to the realized discrepancy computation, which requires the distribution of θ to be computed only once, conditional on the observed y.

6 Applications

6.1 Fitting an increasing, convex mortality rate function

For a simple real-life (or real-death) 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; due to low sample size, they are not themselves increasing or convex. 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. Because the population for each age was in the hundreds, and the rates were so low, we use the Poisson approximation for mathematical convenience: $P(y|\theta) \propto \prod_t \theta_t^{y_t} e^{-N_t\theta_t}$.

To fit the model, we used a computer optimization routine to maximize the 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; if the data are atypical of the model being assumed, we would suspect the model. The obvious possible flaws of the model are the Poisson distribution and the assumed convexity.

Classical χ^2 tests

The χ^2 discrepancy between the data and the maximum likelihood estimate is 30.0, and the minimum χ^2 fit is 29.3. These are based on 30 data points, with 30 parameters being fit. However, there are not really thirty free parameters, because of the constraints implied by the increasing, convexity assumption. In fact, the maximum likelihood estimate lies on the boundary of constraint space; at that point, the solution is 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 reference distribution for the minimum chi-squared statistic?

As a direct check, we can simulate the sampling distribution of the minimum chi-squared statistic, assuming $\theta = \hat{\theta}$, the maximum likelihood estimate. The resulting distribution of $X^2_{\text{min}}(y^{\text{rep}})$ 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 test statistic, $X^2_{\text{min}}(y) = 29.3$, is plotted as a vertical line in Figure 3; it corresponds to a tail-area probability of 16.5%. The distribution of Figure 3 is only an approximation, however, as the true value of θ 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, inducing lower values of the minimum χ^2 statistic. Thus, the distribution of Figure 3 should lead to a conservative p-value for the minimum χ^2 test.

Bayesian inference under the hypothesized model

To perform Bayesian inference, we need to define a prior distribution for θ. Since we were willing to use the maximum likelihood estimate, 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 algorithm of Metropolis (1953). Due to the convexity constraint, it was not convenient to alter the components of θ one at a time (the Gibbs sampler); instead, jumps were taken by adding various linear and piecewise-linear functions, chosen randomly, to θ. Three parallel sequences were simulated, starting at the maximum likelihood estimate and 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 samples from the posterior distribution for θ, chosen at random from the last halves of the simulated sequences, are plotted as dotted lines in Figure 4, with the maximum likelihood estimate from Figure 2 displayed as a solid line for comparison.

Bayesian tests using classical test statistics

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

It is certainly possible to test the goodness of fit of the model by directly examining a graph like Figure 5, following Rubin (1981). The inspection can be done visually—is the solid line an outlier among the forest of dotted lines?—or quantitatively, by defining a test 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}. Sometimes, however, a formal test is desired to get a numerical feel for the goodness-of-fit or to present the results to others in a standard form; to illustrate the methods presented in previous sections, we work with the χ^2 test.

For each simulated replication, y^{rep}, the computer optimization routine was run to find the minimum χ^2 discrepancy, $X^2_{\text{min}}(y^{\text{rep}})$. A histogram of these minimum χ^2 values—the reference distribution for the Bayesian minimum χ^2 test—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 maximum likelihood estimate and displayed in Figure 3. The Bayesian posterior p-value of the minimum χ^2 is 9.7%, which is lower than the maximum likelihood approximation, as predicted.

Bayesian tests using generalized test statistics

Finally, we compute the posterior p-value of the χ^2 discrepancy itself, which requires much less computation than the distribution of the minimum χ^2 statistic, since no minimization is required. For each pair of vectors, (θ, y^{rep}), simulated as described above, $X^2(y^{\text{rep}};\theta)$ is computed and compared to $X^2(y;\theta)$. Figure 7 shows a scatterplot of the realized and predictive discrepancies, in which each point represents a different value of θ drawn from the posterior distribution. The tail-area probability of the realized discrepancy test is just the probability that the predictive discrepancy exceeds the observed 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 lower than the minimum discrepancy p-value, which perhaps suggests that it is the prior distribution, not the likelihood, that does not fit the data. (The analysis of linear models in Section 4.2 suggests that if the likelihood were rejecting the data, the minimum discrepancy test would give the more extreme tail-area probability.)

We probably would not overhaul the model merely to fix a p-value of 6.3%, but it is reasonable 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—perhaps the uniform prior distribution in the constrained parameter space should be modified to reduce the tendency for the curves to increase so sharply at the end. Gelman (1992b) describes a similar theoretical problem with a multivariate uniform prior distribution for the values of a curve that is constrained to be increasing, but not necessarily convex.

6.2 Testing a finite mixture model in psychology

Stern et al. (1992) 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 component and the multinomial probabilities that specify the distribution of the observed variables within a component. Psychological and physiological arguments suggest two to four components for the mixture, with specific predictions about the nature of the infants in each component. Determining the number of components supported by the data is the initial goal of the analysis.

Standard analysis

For a specified number of mixture components, the maximum likelihood estimates of the latent class are obtained using the EM algorithm (Dempster, Laird, and Rubin, 1977). The results of fitting one through four-component models are summarized in Table 3. The discrepancy measure that is usually associated with contingency tables is the log likelihood ratio (with respect to the saturated model),

$$L(y;\theta) = 2\sum_i y_i \log\left(\frac{y_i}{E(y_i|\theta)}\right),$$

where the sum is over the cells of the contingency table. The final column in Table 3 gives $L_{\text{min}}(y)$ for each model. The two-component mixture model provides an adequate fit that does not appear to improve with additional components. The maximum likelihood estimates of the parameters of the two-component model (not shown) indicate that the two components correspond to two groups: the uninhibited children (low scores on all variables) and the inhibited children (high scores on all variables). Since substantive theory suggests that up to four types of children may be present, some formal goodness-of-fit test is desirable. 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 (Titterington, Smith, and Makov, 1985), although it is common practice to use the χ^2 distribution as a guideline.

Bayesian inference

A complete Bayesian analysis, incorporating the uncertainty in the number of classes, is complicated by the fact that the parameters of the various probability models (e.g., the two and four-component mixture models) are related, but not in a straightforward manner. Instead, a separate Bayesian analysis is carried out for each plausible number of mixture components. The prior distribution of the parameters of the latent class model is taken to be a product of independent Dirichlet distributions: one for the component proportions, and one for each set of multinomial parameters within a component. 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 (Rubin and Stern, 1992). The use of a weak but not uniform prior distribution helps identify the mixture components (e.g., the first component of the two-component 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). Once the number of iterations required for the data augmentation to converge was determined, sequences of this length were used to generate draws from the posterior distribution. The draws from the posterior distribution of the parameters for the two-component model were centered about the maximum likelihood estimate. In models with more than two components, the additional components cannot be estimated with enough accuracy to identify the type of infants corresponding to the additional components.

Bayesian tests using classical test statistics

To formally assess the quality of fit of the two-component model, we define replications of the data in which the parameters of the latent class model are fixed. These replications may be considered 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 average discrepancy, $L_{\text{min}}(y^{\text{rep}})$, based on 500 replications, is shown in Figure 8 with a dashed line indicating the value of the test statistic for the observed sample. The mean of this distribution, 23.4, and the variance, 45.3, are not consistent with the χ^2_{20} distribution that would be expected if the usual asymptotic results applied. The Bayesian p-value is 92.8% based on these replications. If the goodness of fit test is applied to the one-component mixture model (equivalent to the independence model for the contingency table), the simulated Bayesian posterior p-value for the likelihood ratio statistic is 2.4%.

Bayesian tests using generalized test statistics

The p-value obtained from the Bayesian replications provides a fair measure of the evidence contained in the likelihood ratio statistic where classical methods do not. However, mixture models present a difficulty that is not addressed by the classical test statistic: multimodal likelihoods. For the data at hand, two modes of the two-component mixture likelihood were found, and for larger models the situation can be worse. For example, six different modes were obtained at least twice in maximum likelihood calculations for the three-component mixture with 100 random starting values. The likelihoods of the secondary modes range from 20% to 60% of the peak likelihood. This suggests that inferences based on only a single mode, such as the test based on L_{min}, may ignore important information.

The p-value for the realized discrepancy between the observed data and the probability model, $L(y;\theta)$, uses the entire posterior distribution rather than a single mode. In addition, the realized discrepancy requires much less computation, since the costly maximization required for the computation of $L_{\mathrm{min}}(y^{\mathrm{rep}})$ at each step is avoided. For each draw from the posterior distribution of θ and the replicated data set, y^{rep}, the discrepancy of the replicated data set relative to the parameter values is compared to that of the observed data. Figure 9 is a scatterplot of the discrepancies for the observed data and for the replications under the two component model. The p-value of the realized discrepancy test is 74.0% based on 500 trials. If the adequacy of the single component model is tested using the realized discrepancy, then the p-value is 5.8% based on 500 Monte Carlo samples. Here the minimum discrepancy test gives the more extreme p-value, in contrast to the mortality rate example.

The Bayesian goodness-of-fit tests show no evidence in the current data to suggest rejecting the two-component latent class model. Since there is some reason to believe that a four-component model describes the population better, we can ask whether a larger data set (more cases and more variables) would be expected to reject the two-component model. By averaging over a prior distribution on four-component models, and then over hypothetical data sets obtained from that model, we can evaluate the effect of increasing the size of the dataset. These calculations indicate that the current sample size does not provide sufficient power to reject the two-component model when the data come from the larger model. The infant laboratory has recently obtained measurements on five variables for three hundred infants; this data set should provide the required power.

7 Long-run frequency properties

Although we are not seeking procedures that have specified long-run error probabilities, it is often desirable to check such long-run frequency properties "when investigating or recommending (Bayesianly motivated) procedures for general consumption" (Rubin, 1984). In the absence of specified alternatives, as in our setting, classical evaluation focuses on the Type I error, that is, the probability of rejecting a null hypothesis, given that it is true. In an exact classical test, the Type I error rate of an α-level test will never exceed α; that is,

$$\Pr(p_c \leq \alpha | H) \leq \alpha. \tag{10}$$

The left side above may depend on the unknown parameter θ, in which case the replication underlying equation (10) is only conceptual in the sense that it cannot be simulated.

Such frequency evaluation is not a part of traditional Bayesian analysis, but in the current context the frequency calculation is in fact straightforward because the Bayesian formulation provides a natural way to quantify different replications under which the "Type I" error is measured. For example, if we are interested in the error rate under the replication with the same value of θ that produced today's data, then we can use the posterior predictive distribution as the replication. Of course, such a replication may be viewed as too restrictive since data sets from different studies may well be generated from different values of θ even if they come from the same null model. It therefore may be more relevant to measure the Type I error rate of p_b under the ***prior predictive distribution*** under H (Box, 1980),

$$P(y^{\text{rep}}|H) = \int P(y^{\text{rep}}|H, \theta)P(\theta)d\theta, \tag{11}$$

which allows different values of θ for each replication. In order for (11) to be defined, the prior density $P(\theta|H)$ must be proper, as in Box (1980). Other replications, such as fixing some components of θ, can also be of practical interest. More discussion of defining replications is given in Section 8.1 and in Rubin (1984).

The classical result (10) can be derived by comparing the sampling distribution of p_c to a uniform distribution. The following result establishes a general result for the prior predictive distribution of the Bayesian p-value, p_b, as compared to the uniform distribution. Since classical test statistics are special cases of generalized test statistics, we only state the result in terms of generalized test statistics.

Theorem. Suppose the sampling distribution of $D(y;\theta)$ is continuous. Then under the prior predictive distribution (11), p_b is *stochastically less variable* than a uniform distribution but with the same mean. That is, if U is uniformly distributed on $[0,1]$, then

(i) $E(p_b) = E(U) = \frac{1}{2}$

(ii) $E(h(p_b)) \leq E(h(U))$, for all convex functions h on $[0,1]$.

The proof of the theorem is a simple application of Jensen's inequality, noting that $p_b = E(p_c(y,\theta)|H,y)$, where p_c is given by (2) and has a uniform distribution given θ under our assumption. Details can be found in Meng (1992). The above result indicates that, under the prior predictive distribution of (11), p_b is more centered around $\frac{1}{2}$ than uniform since it gives less weight to the extreme values than the uniform distribution. Intuitively, this suggests that there should exist an α_0 small enough such that

$$\Pr(p_b \leq \alpha) \leq \alpha, \qquad \text{for all } \alpha \in [0,\alpha_0]. \tag{12}$$

Of course, the value of α_0 will depend on the underlying model so there is generally no guarantee such as $\alpha_0 \geq 0.05$. Because of (i) and (ii), however, the left side of (12) cannot be too big compared to α. The following inequality, which is a direct consequence of the above theorem, as proved in Meng (1992), gives an upper bound on the Type I error.

Corollary. Let $G(\alpha) = \Pr(p_b \leq \alpha)$ be the cumulative distribution function of p_b under (11). Then

$$G(\alpha) \leq \alpha + \left[\alpha^2 - 2\int_0^\alpha G(t)dt\right]^{\frac{1}{2}} \leq 1, \quad \text{for all } \alpha \in [0,1]. \tag{13}$$

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

One direct consequence of (13) is that

$$G(\alpha) \leq 2\alpha, \qquad \text{for all } \alpha \leq \tfrac{1}{2}, \tag{14}$$

which implies that, under the prior predictive distribution, *the Type I error rate of p_b will never exceed twice the nominal level* (e.g., with $\alpha = 0.05$, $\Pr(p_b \leq \alpha) \leq 0.1$). Although the bound 2α in (14) is achievable in pathological examples, the factor 2 is typically too high for α lying in the range of interest (i.e., $\alpha \leq 0.1$). See Meng (1992) for more discussion.

8 Choosing an appropriate test

A Bayesian goodness-of-fit test requires a reference set to which the observed dataset is compared, a prior distribution for the parameters of the model under consideration, and a test statistic summarizing the data and unknown parameters. We discuss each of these features in turn.

8.1 The reference distribution

Choosing a reference distribution amounts to specifying a joint distribution, $P(y, \theta, y^{\text{rep}})$, from which all tail-area probabilities can be computed by conditioning on y and integrating out θ and y^{rep}.

Both the examples in Section 6 consider replications in which the values of the model parameters are fixed (although unknown), and therefore, draws from the posterior predictive distribution are used to obtain replicated data sets. It is also possible, in the manner of Box

(1980), to define the distribution of the predictive data, y^{rep}, as the marginal distribution of the data under the model:

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

In this manner of thinking, the Bayesian model has no free parameters at all, because the "true value" of any parameter θ can be thought of as a realization of its known prior distribution. This *prior predictive distribution* is exactly known under the model, without reference to θ. We may thus think of the model H as a point null hypothesis, with a tail-area probability that does not depend on θ and, in fact, averages the parameter-dependent significance probability over the *prior* distribution of θ:

$$\begin{aligned} p\text{-value}(y) &= P(T(y^{\text{rep}}) \geq T(y)|H) \\ &= \int P(T(y^{\text{rep}}) \geq T(y)|H,\theta)P(\theta|H)d\theta \\ &= \int p_c(y,\theta)P(\theta|H)d\theta, \end{aligned}$$

as proposed by Box (1980).

The prior and posterior predictive distributions for θ are, in general, different, and their associated significance probabilities can have quite different implications about the fit of the model to the data. Although both are based on Bayesian logic, the two approaches differ in their definitions of the reference distribution for y^{rep}, as is shown in Figure 1.

Figure 1a shows the posterior predictive reference set, which corresponds to repeating the experiment tomorrow with the same (unknown) value of θ that produced today's data, y. Because θ is unknown, its posterior distribution is averaged over.

In contrast, Figure 1b shows the reference set corresponding to the prior predictive distribution, in which new values of both θ and y are assumed to occur tomorrow. Since a new value of θ will appear, the information of today's data about θ is irrelevant (once we know the prior distribution), and the prior distribution of θ should be used.

In practice, the choice of a model for hypothesis testing should depend on which hypothetical replications are of interest. For some problems, an intermediate reference set will be defined, in which some components of θ, and perhaps of y also, will be fixed, and thus drawn according to their posterior distributions, while the others are allowed to vary under their conditional prior distributions. Rubin (1984) discusses these options in detail.

Of course, the various choices affect only the model for the predictive replications; Bayesian *estimation* of θ, which assumes the truth of the model, is identical under what we call the "prior" and "posterior" formulations.

8.2 Comparison with the method of Box (1980)

We illustrate the Bayesian goodness-of-fit test, in its prior and posterior forms, for an elementary example. Consider a single observation, y, from a normal distribution with mean θ and standard deviation 1. We wish to use y to test the above likelihood, along with a normal prior distribution for θ with mean 0 and standard deviation 10. As a test statistic, we will simply use y itself.

Prior predictive distribution

Combining the prior distribution and the likelihood yields the marginal distribution of y:

$$y|H \sim N(0, 101).$$

When considering the Bayesian model as a point null hypothesis, we just use this as the fixed reference distribution for y^{rep}. If the test statistic is y, then the prior predictive p-value is the tail area probability based on this distribution. Thus, for example, an observation of $y = 50$ is nearly five standard deviations away from the mean of the prior predictive distribution, and leads to a clear rejection of the model.

Posterior predictive distribution

To determine the posterior tail-area probability, we must derive the posterior distribution of θ and then the posterior predictive distribution of y^{rep}, both under the hypothesized model H. From standard Bayesian calculations, the posterior distribution of θ is,

$$\theta | y \sim N\left(\frac{y}{\frac{1}{100}+1}, \frac{1}{\frac{1}{100}+1}\right),$$

and the posterior predictive distribution of y^{rep}, given y, is,

$$y^{\mathrm{rep}} | y \sim N\left(\frac{y}{\frac{1}{100}+1}, \frac{1}{\frac{1}{100}+1}+1\right). \tag{15}$$

If we simply let the test statistic be y, then the posterior predictive p-value is just the tail area probability with respect to the normal distribution (15). For example, an observation of $y = 50$ is only 0.35 standard deviations away from the mean under the posterior predictive distribution. The observation, $y = 50$, is thus consistent with the posterior but not the prior predictive distribution.

The difference between the prior and posterior tests is the difference between the questions, "Is the prior distribution *correct*?" and "Is the prior distribution *useful* in that it implies a plausible posterior model?" As the above example shows, it is possible to answer "no" to the first question and "yes" to the second. Different reference sets correspond to different conceptions of the prior distribution. The posterior predictive distribution treats the prior as an outmoded first guess, whereas the prior predictive distribution treats the prior as a true "population distribution."

The comparison becomes clearer if we consider an improper uniform prior distribution, which is well known often to yield reasonable *posterior* inference even though any data point whatsoever appears to be a surprise if, for instance, we use as a test statistic the inverse of the absolute value of y: $T(y) = |y|^{-1}$. With an improper uniform prior distribution, the prior predictive distribution of $|y|^{-1}$ will be concentrated at 0, and any finite data point, y, will have a prior predictive p-value of zero. By comparison, the posterior predictive p-value will be 0.5 in this case.

8.3 The role of the prior distribution

In our posterior testing framework, the prior distribution for the parameters of the model need not be especially accurate, as long as the posterior distribution is "near" the data. This relates to the observation that Bayesian methods based on arbitrary prior models (normality, uniformity, Gaussian random walk (used to derive autoregressive time series models), etc.) can often yield useful inference in practice.

In the latent class example of Section 6.2, p-values for evaluating the fit of the one and two-component models have been calculated under a variety of prior distributions. Two properties of the prior distribution were varied. The center of each component 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 tests based on D_{min} and D.

If the prior distribution is strongly informative, however, it affects the tail-area probabilities of different tests in different ways. Tests based on the realized discrepancy are naturally quite sensitive to such prior distributions. The posterior distribution obtained under strong incorrect prior specifications may be quite far from the data. For example, in Section 6.2, a strong prior distribution specifying two mixture components, but not corresponding to inhibited and uninhibited children, leads to a tail-area probability of essentially zero, and thus the model is rejected. By comparison, the minimum discrepancy test is 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. The sensitivity of different test statistics to the specification of the prior distribution may be important in the selection of the test statistic for a particular application.

8.4 Choosing a test statistic

As in the classical approach, our test statistic appears arbitrary, except in the case of a point null hypothesis and a point alternative, in which case the Neyman-Pearson lemma justifies the likelihood ratio test. In more complicated models, as illustrated here, test statistics are typically measures of residuals between the data and the model, or between the data and the best fit of the model.

If one decides to test based on a discrepancy measure, D, there is still the choice of whether to apply the Bayesian test to D_{min}, D_{avg}, D itself, or some other function of the posterior discrepancy. The minimum discrepancy has the practical advantage of being standard in current statistical practice and easily understandable. In addition, D_{min} is a function only of the data and the constraints on the model, and not of the prior distribution of the model parameters. A disadvantage of the minimum discrepancy is that it measures only the best fit model parameters, and ignores how much of the posterior distribution of the model is actually close to the data. This problem is potentially serious if the posterior distribution is multimodal, as in the example of Section 6.2.

The average discrepancy has the clear advantage of using the whole posterior distribution, not just a single point, and has the related feature of possibly testing the prior distribution as well as the constraints on the parameters. Testing the prior distribution is an advantage for serious Bayesian modelers and perhaps a disadvantage to others who are just using convenient noninformative prior distributions. The evidence collected thus far indicates that weak or noninformative prior distributions are not likely to affect the goodness-of-fit test whereas strong prior information may affect the test. If the prior information is strong, it should certainly be tested along with the rest of the model. The main drawback of the average discrepancy test is computational: a simulation or integration is required inside a larger simulation loop.

The realized discrepancy test shares the above-mentioned virtues (or defects) of the average discrepancy test and, in addition, is easy to compute, especially if simulations from the posterior distribution have already been obtained, as is now becoming almost standard in Bayesian statistics. Although the realized discrepancy cannot be directly observed, testing it is in some ways simpler than any other discrepancy test.

8.5 Recommendations

In general, we recommend using the posterior predictive reference distribution, except in some hierarchical models with local parameters θ and hyperparameters α in which it is reasonable to imagine θ varying in the hypothetical replications, while the hyperparameters α remain fixed—that is, use the posterior predictive distribution for α but a conditional prior predictive distribution (conditional on the α but not y) for θ, as pictured in Figure 1c. The posterior predictive distribution must be used for any parameter that has a noninformative prior distribution.

In choosing the prior distribution (and, for that matter, the likelihood), robustness to model specification is a separate problem from goodness-of-fit, and is not addressed by the methods in this article. As discussed above, however, different discrepancy measures test different aspects of the fully-specified probability model.

Finally, test statistics can be often chosen to address specific substantive predictions of the model, as discussed by Rubin (1981, 1984), or suspicious patterns in the data that do not seem to have been included in the model, as in Belin and Rubin (1992). Often these problem-specific test statistics depend only on the data (i.e., they are not generalized test statistics). When considering test statistics based on discrepancies—possibly because discrepancies such as χ^2 and the likelihood ratio are conventional, or because of a particular substantive discrepancy of interest in the problem at hand—we recommend the realized discrepancy test, because of its direct interpretation and computational simplicity compared to the minimum or average discrepancy tests.

References

Aitkin, M. (1991). Posterior Bayes factors. *Journal of the Royal Statistical Society B* **53**, 111–142.

Belin, T. R., and Rubin, D. B. (1992). The analysis of repeated-measures data on schizophrenic reaction times using mixture models. Technical report.

Berger, J. O., and Sellke, T. (1987). Testing a point null hypothesis: the irreconcilability of P values and evidence. *Journal of the American Statistical Association* **82**, 112–139.

Berkson, J. (1980). Minimum chi-square, not maximum likelihood (with discussion). *Annals of Statistics* **8**, 457–487.

Box, G. E. P. (1980). Sampling and Bayes' inference in scientific modelling and robustness. *Journal of the Royal Statistical Society 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*, to appear.

Chernoff, H. (1954). On the distribution of the likelihood ratio. *Annals of Mathematical Statistics* **25**, 573–578.

Cochran, W. G. (1952). The χ^2 test of goodness of fit. *Annals of Mathematical Statistics* **23**, 315–345.

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

Dempster, A. P. (1974). In *Proceedings of Conference on Foundational Questions in Statistical Inference*, ed. O. Barndorff-Nielsen et al, 335–354. Department of Theoretical Statistics, University of Aarhus, Denmark.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical*

Society B **39**, 1–38.

Fisher, R. A. (1922). On the interpretation of chi square from contingency tables, and the calculation of P. *Journal of the Royal Statistical Society* **85**, 87–94.

Geisser, S., and Eddy, W. F. (1979). A predictive approach to model selection. *Journal of the American Statistical Association* **74**, 153–160.

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

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

Gelman, A. (1992b). Who needs data: restricting image models by pure thought. Technical Report, Department of Statistics, University of California, Berkeley.

Gelman, A., Meng, X. L., Rubin, D. B., and Schafer, J. L. (1992). Bayesian computations for loglinear contingency table models. Technical Report.

Gelman, A., and Rubin, D. B. (1992). Inferences from iterative simulation using multiple sequences (with discussion). *Statistical Science*, to appear.

Good, I. J. (1967). A Bayesian significance test for multinomial distributions (with discussion). *Journal of the Royal Statistical Society B* **29**, 399–431.

Good, I. J. (1992). The Bayes/non-Bayes compromise: a brief review. *Journal of the American Statistical Association* **87**, 597–606.

Guttman, I. (1967). The use of the concept of a future observation in goodness-of-fit problems. *Journal of the Royal Statistical Society B* **29**, 83–100.

Jaynes, E. T. (1978). Where do we stand on maximum entropy? In *The Maximum Entropy Formalism*, ed. R. D. Levine and M. Tribus. Cambridge, Mass.: MIT Press. Also reprinted in Jaynes (1983).

Jaynes, E. T. (1983). *Papers on Probability, Statistics, and Statistical Physics*, ed. R. D. Rosenkrantz. Dordrecht, Holland: Reidel.

Jeffreys, H. (1939). *Theory of Probability.* Oxford University Press.

Madigan, D., and Raftery, A. E. (1991). Model selection and accounting for model uncertainty in graphical models using Occam's window. Technical Report #213, Department of Statistics, University of Washington.

McCullagh, P. (1985). On the asymptotic distribution of Pearson's statistic in linear exponential-family models. *International Statistical Review* **53**, 61–67.

McCullagh, P. (1986). The conditional distribution of goodness-of-fit statistics for discrete data. *Journal of the American Statistical Association* **81**, 104–107.

Meng, X. L. (1992). Bayesian p-value: a different probability measure for testing (precise) hypotheses. Technical Report #341, Department of Statistics, University of Chicago.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *Journal of Chemical Physics* **21**, 1087–1092.

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine, Series 5*, **50**, 157–175.

Raftery, A. E. (1986). Choosing models for cross-classifications. *American Sociological Review* **51**, 145–146.

Raghunathan, T. E. (1984). A new model selection criterion. Research Report S-96, Department of Statistics, Harvard University.

Rubin, D. B. (1981). Estimation in parallel randomized experiments. *Journal of Educational Statistics* **6**, 377–400.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician, §5. *Annals of Statistics* **12**, 1151–1172.

Rubin, D. B., and Stern, H. S. (1992). Testing in latent class models using a posterior predictive check distribution. Technical Report, Department of Statistics, Harvard University.

Schaafsma, W., Tolboom, J., and Van der Meulen, B. (1989). Discussing truth or falsity by computing a Q-value. *Statistical Data Analysis and Inference*, ed. Y. Dodge, 85–100. Amsterdam: North-Holland.

Spiegelhalter, D. J., and Smith, A. F. M. (1982). Bayes factors for linear and log-linear models with vague prior information. *Journal of the Royal Statistical Society B* **44**, 377–387.

Smith, A. F. M., and Roberts, G. O. (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo results (with discussion). *Journal of the Royal Statistical Society* **55**, to appear.

Stern, H. S., Arcus, D., Kagan, J., Rubin, D. B., and Snidman, N. (1992). Statistical choices in temperament research. Technical report, Department of Statistics, Harvard University.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society B*, **36**, 111–147.

Tanner, M. A., and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association* **52**, 528–550.

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

Zellner, A. (1975). Bayesian analysis of regression error terms. *Journal of the American Statistical Association* **70**, 138–144.

Table 1: Mortality rate data from Broffitt (1988)

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

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

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

Figure 1a: The posterior predictive distribution

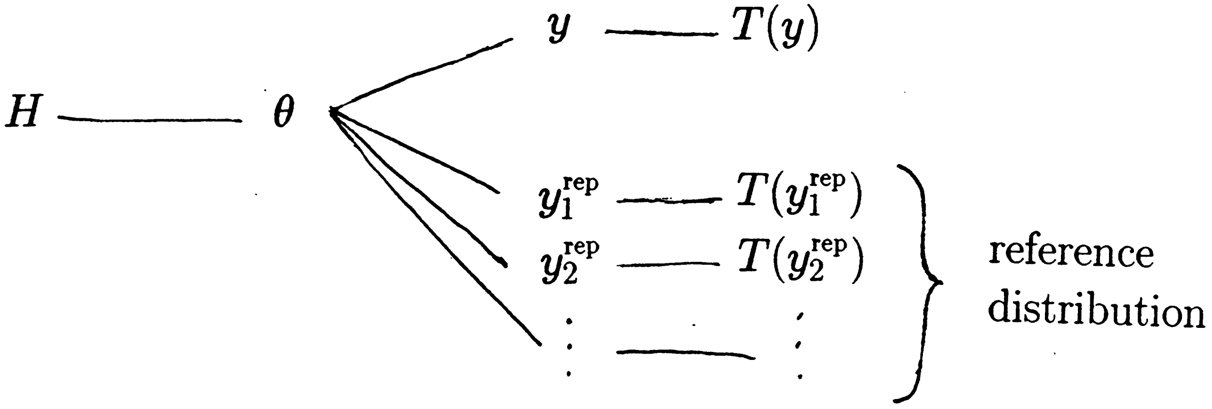


Figure 1b: The prior predictive distribution

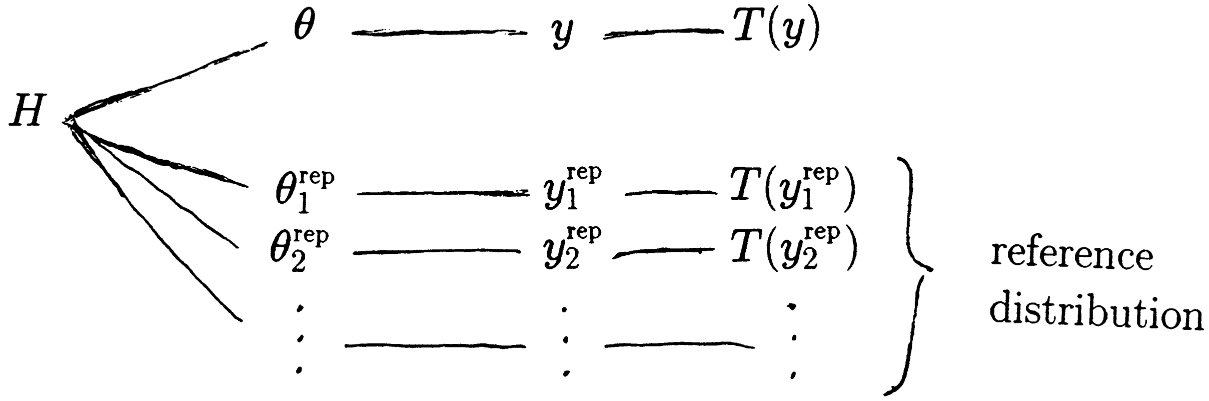


Figure 1c: A mixed predictive distribution

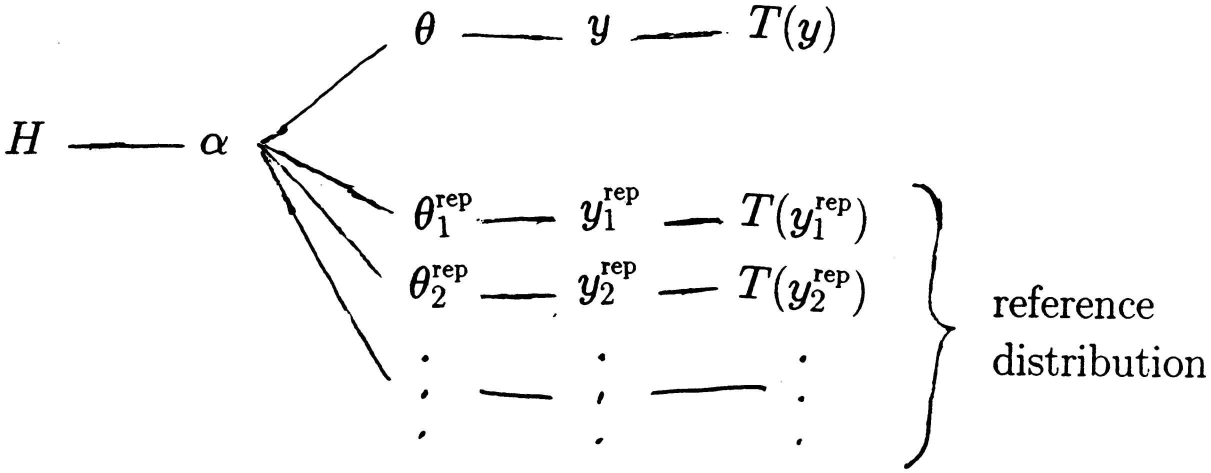


Figure 2: Observed mortality frequencies and the maximum likelihood estimate of the mortality rate function, under the constraint that it be increasing and convex

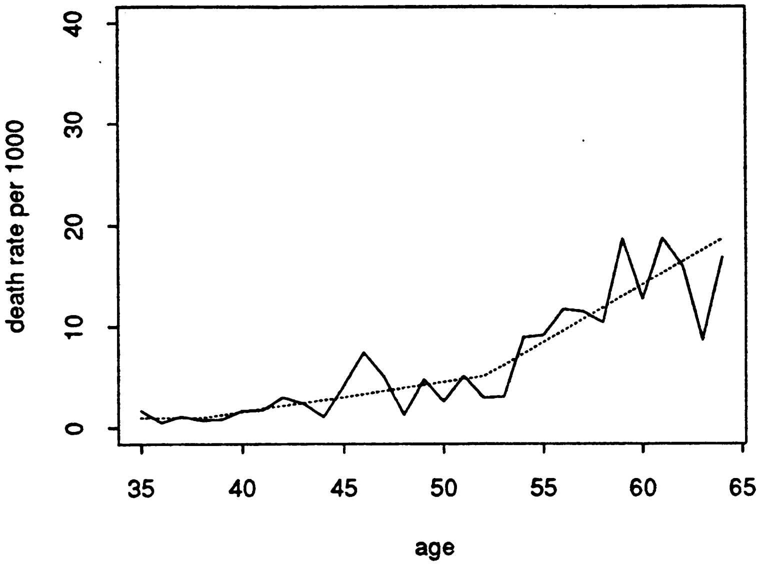


Figure 3: Histogram of simulations from the reference distribution for the minimum χ^2 statistic for the mortality rates: classical approximation with θ set to the maximum likelihood estimate

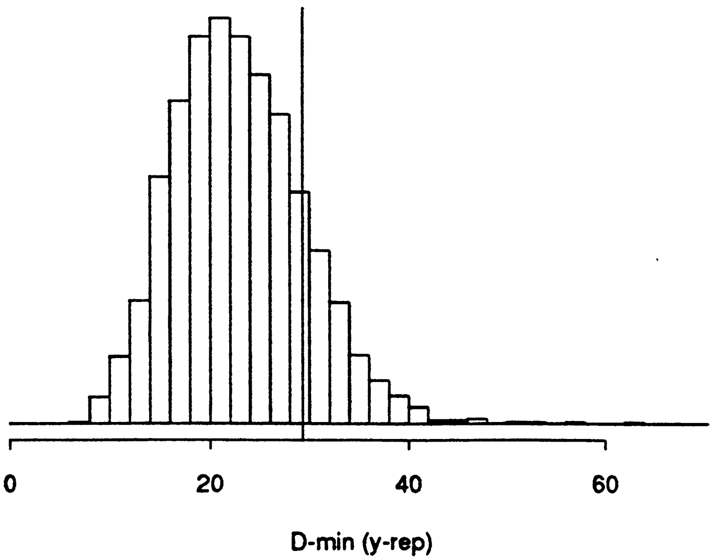


Figure 4: Nine draws from the posterior distribution of increasing, convex mortality rates, with the maximum likelihood estimate as a comparison

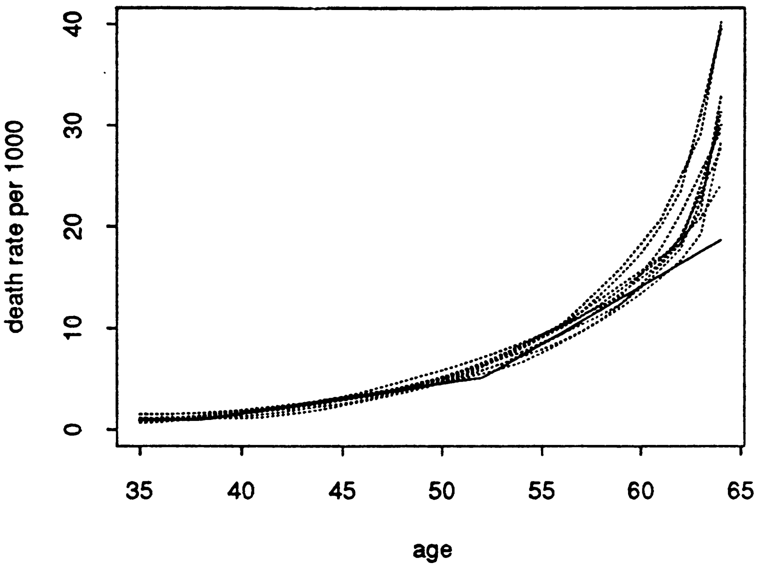


Figure 5: Nine draws from the posterior predictive distribution of mortality frequencies, corresponding to the nine draws of Figure 4, with the raw data as a comparison

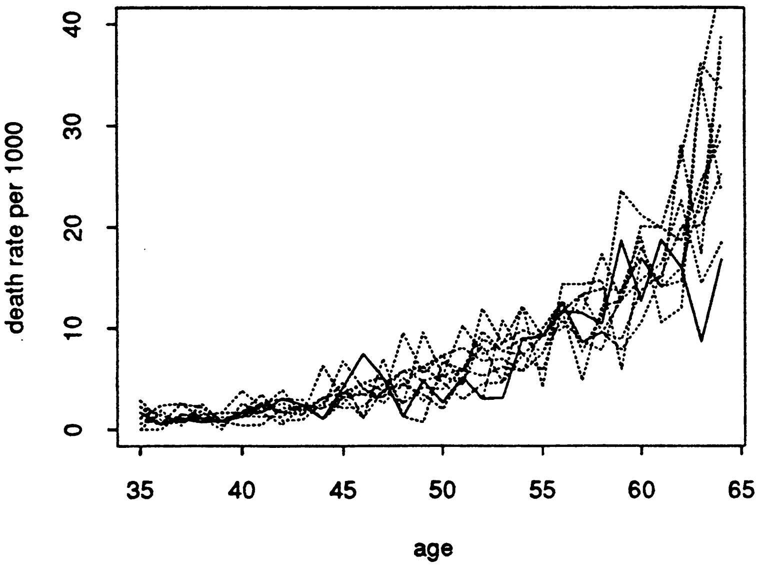


Figure 6: Histogram of simulations from the reference distribution for the minimum χ^2 statistic for the mortality rates, using the Bayesian posterior predictive distribution

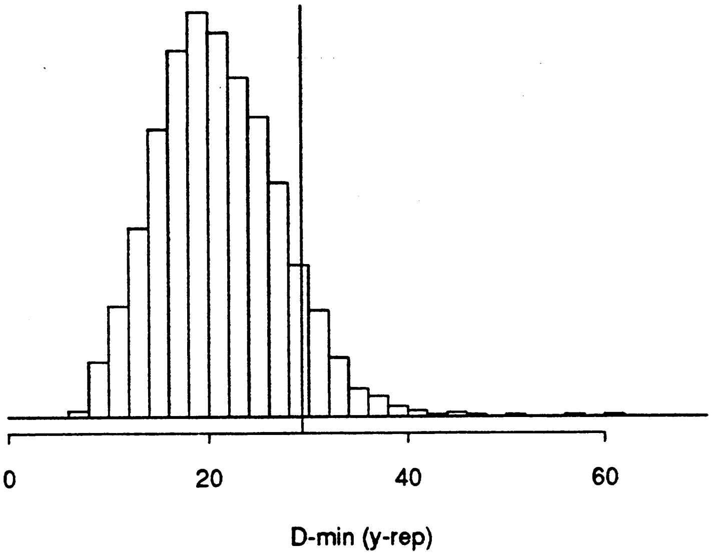


Figure 7: Scatterplot of predictive vs. realized χ^2 discrepancies for the mortality rates, under the Bayesian posterior distribution; the p-value is estimated by the proportion of points above the 45° line.

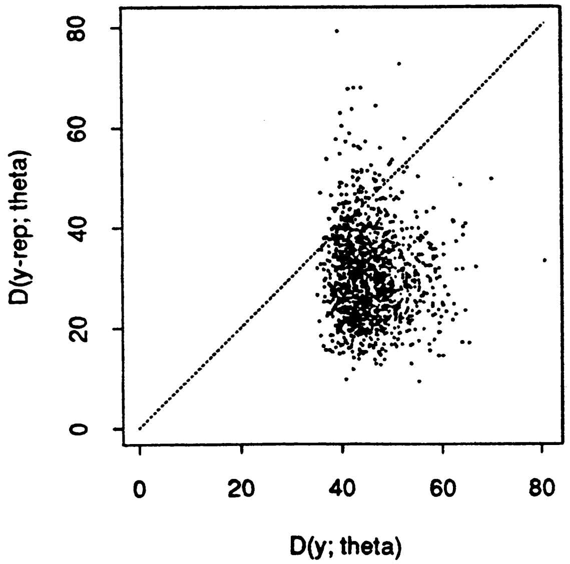


Figure 8: Histogram of simulations from the reference distribution for the log likelihood ratio statistic for the latent class example, using the Bayesian posterior predictive distribution.

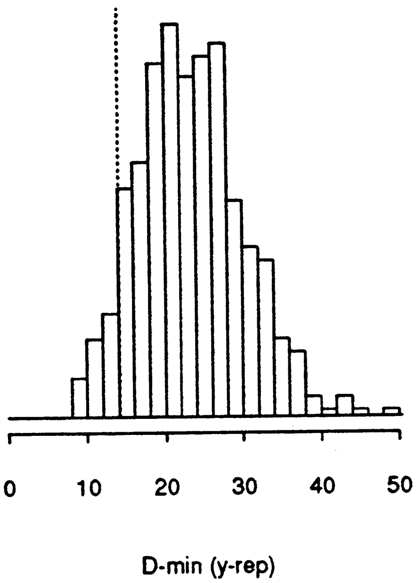
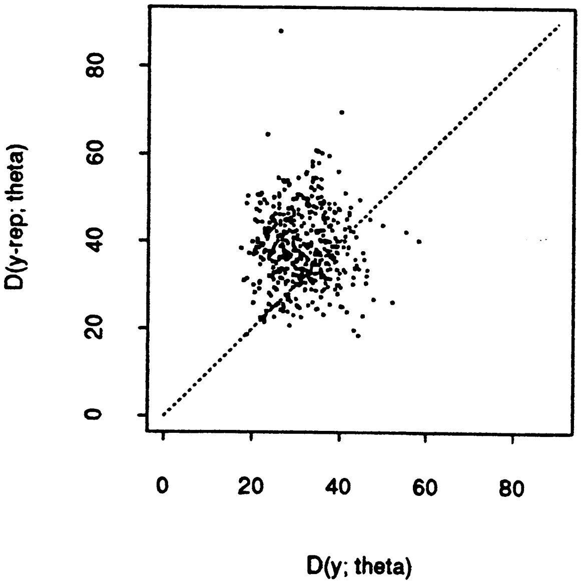


Figure 9: Scatterplot of predictive vs. realized log likelihood ratio discrepancies for the latent class model, under the Bayesian posterior distribution; the p-value is estimated by the proportion of points above the 45° line.


D(y-rep; theta)
D(y; theta)