Computational Statistics (2000) 15:337–354

© Physica-Verlag 2000

Posterior predictive checks: Principles and discussion

Johannes Berkhof [1] and Iven van Mechelen

Department of Psychology, University of Leuven, Tiensestraat 102, B-3000 Leuven, Belgium

Herbert Hoijtink

Department of Methodology and Statistics, Utrecht University, P.O.Box 80140, 3508 TC Utrecht, The Netherlands

Summary

In this paper, we give a description of posterior predictive checking (introduced by Rubin, 1984) for detecting departures between the data and the posited model and illustrate how the posterior predictive check can be used in practice. We further discuss interpretability, frequency properties, and prior sensitivity of the posterior predictive p-value.

Keywords: Posterior Predictive Check, Discrepancy Measure, Posterior Predictive p-value, Frequency Properties, Prior Sensitivity.

[1] Address correspondence concerning this paper to Johannes Berkhof. The work reported in this paper was supported in part by Grant OT/96/10 of the Research Council of the Katholieke Universiteit Leuven.

1 Introduction

Almost any statistical modeling endeavor implies important questions on model assessment such as: Given a model and observed data, how plausible are inferences with respect to model characteristics that are of major substantive interest? Classical approaches to deal with these questions rely on various goodness-of-fit tests, resulting in p-values. More specifically, if we denote the observations by y^{obs} and a goodness-of-fit test statistic by $T(y)$, the classical p-value is defined as

$$p_C = Pr\{T(y) \geq T(y^{\text{obs}})|H_0\}\ , \tag{1}$$

where the probability in (1) is taken over the sampling distribution of $T(y)$ under the null hypothesis H_0 that the target model M_0 with null set of parameters Θ_0 is true. One may note that the p-value in equation (1) is defined only if the null hypothesis completely specifies the parameter vector, that is, if Θ_0 is a singleton, $\{\theta_0\}$, or if the sampling distribution of $T(y)$ under the null hypothesis does not depend on nuisance parameters that are not fully specified in Θ_0, that is, if $T(y)$ is a pivotal quantity. The property of being at least approximately pivotal is therefore a key consideration in the choice of classical goodness-of-fit statistics.

A well-known instance of a classical goodness-of-fit test statistic is the generalized log likelihood ratio, which can be used to compare two nested models. Under the null hypothesis that the more restrictive model is true, and under general regularity conditions, the latter is known to be a pivotal quantity if the sample size is large enough; it can be shown to be asymptotically χ^2-distributed and to result in asymptotically uniformly most powerful tests.

For many models, however, a standard goodness-of-fit testing approach appears to be troublesome. There are at least three different reasons for this: (1) Several recently developed models are complex to such an extent that it may be very hard if not impossible to derive the asymptotic distribution of relevant goodness-of-fit test statistics. As an example, one may think of models with severe parameter restrictions (e.g., positivity constraints) and models outside the generalized linear model family (Gelman, Meng, and Stern, 1996). The latter class of models includes mixture models, such as multivariate normal mixtures or mixtures of multivariate independent Bernoulli distributions (i.e., latent class models). To test, for instance, a mixture model with k+1 components against a model with k components, one may wish to use a likelihood ratio test. Unfortunately, however, the latter does not have an asymptotic χ^2-distribution as the regularity condition that the null hypothesis has to lie in the interior of the parameter space is not met (Chernoff, 1954). (2) For a number of models, standard goodness-of-fit statistics do have a known asymptotic distribution, but the asymptotic distribution may be insufficiently accurate in realistic sample situations. As an example, one may think of various loglinear psychometric models for binary person by item data: In realistic samples of person by item data, a large number of response patterns will not be observed, which casts serious doubt on the accuracy of the asymptotic χ^2-distribution of likelihood ratio statistics. (3) When monitoring the quality of a statistical model, one is often not primarily interested in assessing the overall quality of the model; rather, one would like to know whether certain inferences with respect to model characteristics are plausible, in particular with respect to those characteristics that are crucial from a substantive viewpoint. Therefore, one may wish to inspect more specific goodness-of-fit statistics than overall measures such as a generalized likelihood ratio. The sampling distribution of such specific statistics, even asymptotic, however, will often be unknown and hard (if not impossible) to derive analytically.

The classical approach to model assessment has been extended within a Bayesian framework, making use of so-called *posterior predictive checks*. This Bayesian extension is based on work by Rubin (1984) and has further been elaborated by Meng (1994) and Gelman, Meng and Stern (1996). Like the classical approach, it results in p-values, albeit based now on a Bayesianly defined reference distribution. The use of posterior predictive checks (*ppc*'s) may be considered either from a frequentist or from a Bayesian viewpoint. Firstly, the *ppc*-procedure may be considered as a frequentist statistical decision procedure, with associated (generalized) p-values. This procedure overcomes the difficulties of classical approaches to goodness-of-fit testing as outlined above: It does not rely on asymptotic properties or approximations, and, in principle, it can be used with all types of models (even the most complex ones) and with all types of goodness-of-fit statistics. Secondly, in Bayesian data analysis, the use of posterior predictive checks can be argued to be the natural way to take for model monitoring. In this viewpoint, the p-values resulting from the procedure are given a different interpretation, namely as measures of deviance between the model and the data, rather than as the expected outcome of a statistical decision procedure.

In the remainder of this paper we will first outline the basic principles of the posterior predictive check approach (Section 2). Section 3 will present an illustrative application. In Section 4 we will discuss the interpretation of the posterior predictive p-value. In Sections 5 and 6, the frequency properties and prior sensitivity of the posterior predictive p-value will be discussed. Section 7 contains some concluding remarks.

2 Principles of posterior predictive checks

In this section we will first outline the basic principles of the posterior predictive check approach for classical test statistics $T(y)$, that is, functions of the data only. Next we will consider a generalization that makes use of discrepancy variables, that is, functions $D(y, \theta)$ of both the data and the parameters and we will explain how the posterior predictive p-value based on a discrep-

ancy variable can be computed in practice. We conclude with a discussion of the major choices to be made in any posterior predictive check application.

2.1 Posterior predictive checks using test statistics

Assume that a study was designed that yielded a set of observed data y^{obs}. Assume further that the data have been generated from a model $p(y|\theta^*)$ (with some fixed but unknown parameter vector θ^*) and that a test statistic $T(y)$ has been chosen. We would like to compare the value the test statistic takes for the actually observed data, with the values it would take if the study were replicated with the same model and the same value θ^* of θ that produced y^{obs}. The replicated data, which we further denote by y^{rep}, can be considered to be drawn from the *posterior predictive distribution*, that is,

$$p(y^{\text{rep}}|y^{\text{obs}}) = \int p(y^{\text{rep}}|\theta)p(\theta|y^{\text{obs}})d\theta \ . \tag{2}$$

One may note that equation (2) involves an integration with respect to parameter vector θ because not all elements of θ are fixed by the null hypothesis and therefore posterior uncertainty is to be taken into account.

The posterior predictive p-value, denoted by p_T, may then be defined as follows:

$$\begin{aligned} p_T &= Pr\{T(y^{\text{rep}}) \geq T(y^{\text{obs}})|y^{\text{obs}}, H_0\} \\ &= \iint I_{T(y^{\text{rep}})\geq T(y^{\text{obs}})} p(y^{\text{rep}}|\theta)p(\theta|y^{\text{obs}})dy^{\text{rep}}d\theta \ . \end{aligned} \tag{3}$$

2.2 Posterior predictive checks using discrepancy measures

In a number of cases, an adequate assesment of the deviations between a model and the data may naturally lead to the formulation of *discrepancy measures*, that is, functions of both the data and the parameters, $D(y, \theta)$. As an example one may think of data vectors y the elements, y_i, of which one wishes to compare with their expectations under the model, $E(y_i|\theta)$. A natural discrepancy measure in that case is the χ^2-discrepancy as defined by Gelman, Meng and Stern (1996):

$$\chi^2(y, \theta) = \sum_{i=1}^{n} \frac{(y_i - E(y_i|\theta))^2}{Var(y_i|\theta)} \ . \tag{4}$$

In case of a discrepancy measure, the posterior predictive value, denoted by p_D may be defined as:

$$\begin{aligned} p_D &= Pr\{D(y^{\text{rep}}, \theta) \geq D(y^{\text{obs}}, \theta)|y^{\text{obs}}, H_0\} \\ &= \iint I_{D(y^{\text{rep}},\theta)\geq D(y^{\text{obs}},\theta)} p(y^{\text{rep}}|\theta)p(\theta|y^{\text{obs}})dy^{\text{rep}}d\theta \ . \end{aligned} \tag{5}$$

2.3 Practical estimation of the posterior predictive p-value

Because a test statistic $T(y)$ is a special type of discrepancy measure $D(y, \theta)$, we will confine ourselves to the practical estimation of p_D, which goes as follows:

1. Choose a discrepancy measure $D(y, \theta)$.
2. Draw a vector $\theta_{(k)}$ from the posterior distribution $p(\theta|y^{\text{obs}}) \propto p(y^{\text{obs}}|\theta)p(\theta)$.
3. Calculate $D(y^{\text{obs}}, \theta_{(k)})$.
4. Simulate a replicated data set $y^{\text{rep}}_{(k)}$ from $p(y|\theta_{(k)})$.
5. Calculate $D(y^{\text{rep}}_{(k)}, \theta_{(k)})$.
6. Repeat steps 2-5 K times (where K denotes a prespecified number of replications).
7. Count the proportion of replicated data sets $y^{\text{rep}}_{(k)}$ for which $D(y^{\text{rep}}_{(k)}, \theta_{(k)}) \geq D(y^{\text{obs}}, \theta_{(k)})$.

Note that in case of a test statistic, step 3 needs to be performed only once. Although the emphasis here is on the computation of the actual p-value, the pairwise comparison of corresponding $D(y^{\text{obs}}, \theta_{(k)})$ and $D(y^{\text{rep}}_{(k)}, \theta_{(k)})$ values can also be graphically displayed by plotting the corresponding discrepancy values against each other. Likewise, the test statistic value $T(y^{\text{obs}})$ can be compared with $T(y^{\text{rep}}_{(1)}), \ldots, T(y^{\text{rep}}_{(K)})$ by localizing $T(y^{\text{obs}})$ in the frequency distribution of $T(y^{\text{rep}})$.

2.4 Choices to be made in posterior predictive checks

Any posterior predictive check involves at least three choices: (1) the choice of a test statistic or discrepancy measure, (2) the choice of the prior distribution, and (3) the definition of the desired type of replications. We will now successively discuss each of these three briefly.

(1) In principle, the posterior predictive check approach can be applied with any test statistic or discrepancy measure. So there is no need to limit the scope to test quantities with known (asymptotic) reference distributions or even to pivotal quantities. Rather, the choice of test quantities can be primarily based on substantive considerations. That is, one may focus on quantities with a straightforward substantive interpretation or quantities that directly capture key model assumptions that are of clear substantive relevance (see, e.g., Gelman et al., 1995; Hoijtink and Molenaar, 1997). A particularly interesting class of test quantities is the one that makes use of various graphical devices (see, e.g., Gelman et al. 1995, in press). Taken together, a careful

choice of test statistics and discrepancy measures will allow the user to depict a profile of strengths and weaknesses of the model under study.

(2) Whereas, from a Bayesian viewpoint, the choice of a prior distribution can be considered part of the actual statistical modeling, from a frequentist viewpoint it can be considered an additional step implied by posterior predictive model checking. Both Bayesians and frequentists may further wonder to what extent the posterior predictive check results may vary as a function of different (plausible) choices for the prior distribution, which is called the *sensitivity issue*.

(2) Defining the nature of replications concerns the selection of aspects of the data to be kept constant in the replications. One may think of fixing the sample size or a subset of parameters that generated the data. As another example, in prediction problems, one may wish (or not wish) to keep constant the values of the predictor variables (while replicating the value of the criteria). In hierarchical problems (e.g., involving pupils nested in classes nested in schools), one may wish to simulate new pupils from the same classes, or new pupils from new classes etc. Interestingly, one may note that, in some cases, a model can be specified in different ways that imply the same value for the likelihood but different definitions of replications, and therefore possibly different *ppc*-results (for an example, see Gelman et al., in press). Besides fixing aspects of the data, one may also wish to fix the values of a subset of parameters. Note that then the replications are not drawn anymore from the posterior predictive distribution. Ultimately, one may even draw an entirely new vector of parameters from the prior distribution for each replication, resulting in a *prior predictive check* (Box, 1980); the latter is most suitable to measure inconsistencies between the data and the prior distribution.

3 Example

The example that will be given in this section uses data collected by Kass (see, Kass and Raftery, 1995, application 2). The data consist of the number of field goals y_i and the number of goal attempts n_i in $i = 1, \ldots, 380$ consecutive basketball games played by Vinnie Johnson for the Detroit Pistons (during the 1985-1989 season). An impression of the data can be obtained from Figure 1 (note that the number of diamonds does not add to 380 because some of the diamonds represent two or more games with the same y_i and n_i). As can be seen, and not surprisingly, the number of points and the number of attempts are positively correlated.

Kass and Raftery (1995) describe a study executed by Hsiao (1994). The goal of this study was to determine if Vinnie Johnson has a *hot hand* (Gilovich, Vallone and Tversky, 1985) or not, that is, if there are games in which this player is significantly better than expected on the basis of the player's overall record. Two models were compared. In the null model, it was assumed that the probability of scoring θ is constant over the games (Vinnie Johnson does

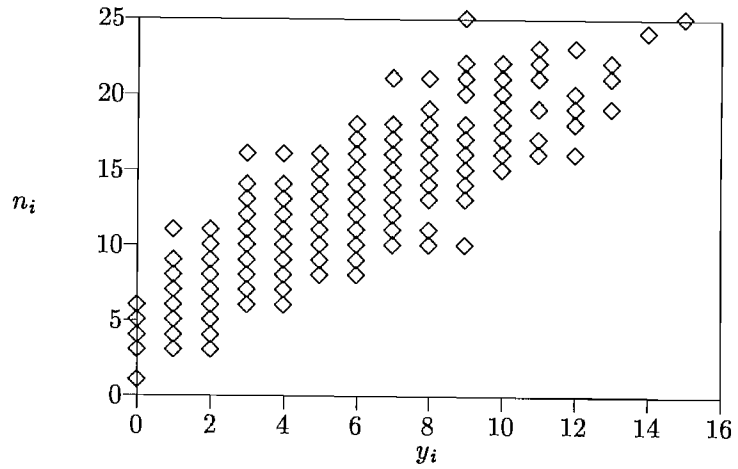


Figure 1: Number of goal attempts (n_i) against number of field goals (y_i) by Vinnie Johnson in 380 consecutive basketball games.

not have a hot hand): $y_i \sim Bin(n_i, \theta)$ for $i = 1, \ldots, 380$. In the alternative model, it was assumed that θ is not constant over games, $y_i \sim Bin(n_i, \theta_i)$ and that $\theta_i \sim Beta(\alpha, \beta)$ (both for $i = 1, \ldots, 380$). This model implies that θ is not constant over games, but has a beta distribution with parameters α and β. Using Bayes factors, to compare the null and alternative model, it was concluded that Vinnie Johnson does not have a hot hand in the sense that there was no substantial game-to-game extra-binomial variability in the probability of scoring.

Besides testing for extra-binomial variability, there are other possibilities to investigate the existence of a hot hand, or, perhaps more adequately, phenomena related to the existence of a hot hand. One might, for example, investigate if some of the y_i/n_i are exceptionally high (i.e. outliers) given the assumption that $y_i \sim Bin(n_i, \theta)$. Or, one might investigate if the number of series of consecutive games in which the player performs better than on average, is larger than expected given the assumption that $y_i \sim Bin(n_i, \theta)$. This would imply that the player has 'hot' series, which consist of (relatively) hot games. Or, finally, one might also investigate if there is a relation between y_i/n_i and n_i. The latter could lead to the conclusion that the hand gets 'hotter' if the number of attempts at a field goal increases. Note that the latter is not consistent with the binomial model, where y_i/n_i and n_i are independent. All possibilities mentioned above will be investigated using posterior predictive checks as introduced in the previous section. Our example will show that posterior predictive inference can be based on test statistics $T(y)$ and discrepancy measures $D(y, \theta)$ (with $y = (y_1, \ldots, y_{380})'$) that are immediately relevant in a substantive (and not necessarily a statistical) sense. Further-

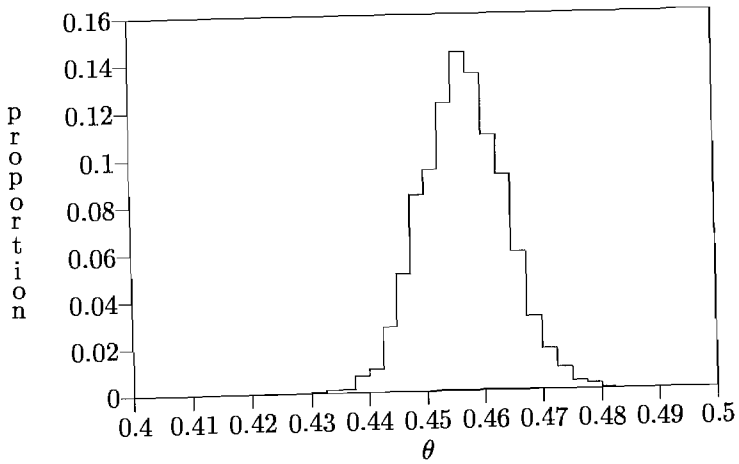


Figure 2: The simulated posterior distribution of θ.

more, our example will illustrate that posterior predictive inference can be useful in the construction of an alternative model. Note that the same view has been expressed by Lewis and Raftery in the discussion to Gelman, Meng, and Stern (1996).

In the sequel, four posterior predictive checks will be carried out to investigate if Vinnie Johnson has something like a hot hand or not. The first posterior predictive check detects whether there are one or more games in which Vinnie Johnson is exceptionally good compared to his performance in the rest of the games. This would be indicated by the presence of outliers i.e. games in which y_i/n_i is larger than expected under the binomial model. A simple but adequate and interpretable *statistic* with which this can be investigated is $T(y) = \max_i\{y_i/n_i\}$ i.e. the largest value of y_i/n_i observed for $i = 1, \ldots, 380$. The question if this number is large given the binomial model can be investigated using the posterior predictive distribution of this statistic. This distribution can be obtained via the following four steps:

1. Draw $\theta_{(k)}$ from its posterior distribution. Using a uniform prior for θ, this is a beta distribution with $\alpha = 2194 + 1$ and $\beta = 4801 - 2194 + 1$, given that over 380 games Vinnie Johnson made 4801 attempts at a field goal, and was succesful in 2194 of these attempts.

2. For $i = 1, \ldots, 380$, simulate $y^{\text{rep}}_{(k)i}$ from $Bin(\theta_{(k)}, n_i)$.

3. Calculate $T(y^{\text{rep}}_{(k)})$.

4. Repeat step 1-3 K=2000 times.

The above steps correspond to steps 2, 4, 5, and 6 of the scheme in Section 2.3. The resulting posterior sample $\theta_{(1)}, \ldots, \theta_{(2000)}$ is displayed in Figure 2. The

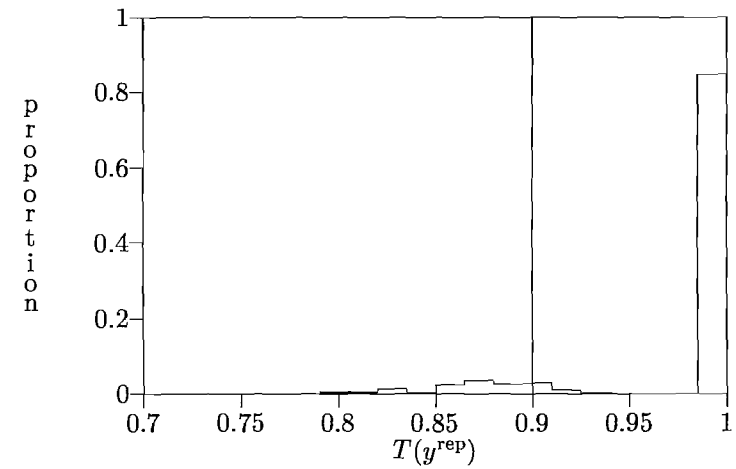


Figure 3: The maximum value of y_i/n_i over 380 consecutive games. The position of $T(y^{\text{obs}})$ is denoted by the vertical line.

resulting sample from the posterior predictive distribution of the test statistic is displayed in Figure 3. After having simulated the posterior predictive distribution, the predictive p-value is computed as the proportion of the 2000 draws for which $T(y^{\text{rep}}_{(k)}) \geq T(y^{\text{obs}})$. As can be seen, the distribution is far from smooth, and is probably quite difficult to trace using analytical methods instead of simulation. The observed value $\max_i\{y_i^{\text{obs}}/n_i\}$ was equal to .90. About 89% of the replicated value were larger, indicating that .90 is not unusual or outlying.

The second statistic $T(y)$ is the number of 'consecutive series of four or more games' in which $y_i/n_i < .50$ for each game. This statistic can be used to investigate if Vinnie Johnson has more periods in which his hand is relatively 'cold' than is to be expected given the binomial model. If the latter would be the case, a coach might be tempted not to play Vinnie Johnson in the fourth game after three relatively 'cold' games. For Vinnie Johnson, during 380 games 28 relatively cold periods were observed. As can be seen in Figure 4, the number of series of relatively cold games expected under the binomial model is roughly between 15 and 35. The observed number is 28 (indicated by a vertical line) and has a posterior predictive p-value of .28, that is, it is not an exceptionally large number. It appears that the number of relatively 'cold' series of Vinnie Johnson is in agreement with the number expected under the binomial model.

The question of interest could also have been if the number of 'consecutive series of four or more games' in which $y_i/n_i > \theta$ is larger than expected under the binomial model. If the latter is the case, a coach might be tempted to play Vinnie Johnson always in a fourth game following a relatively 'hot' series

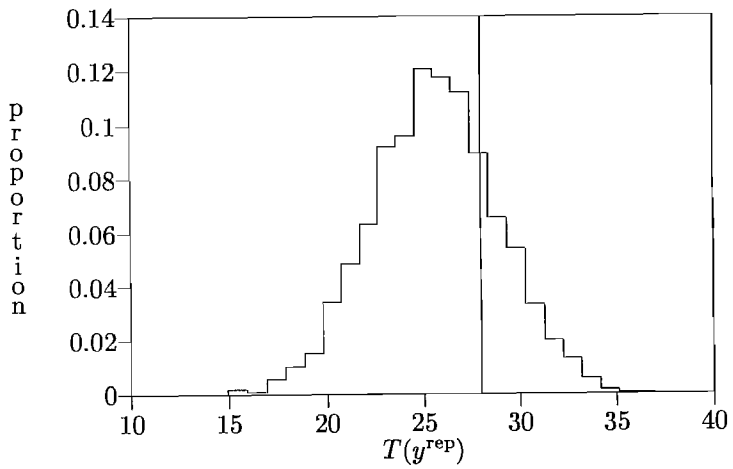


Figure 4: Number of series of at least four consecutive games with $y_i/n_i < .50$. The position of $T(y^{obs})$ is denoted by the vertical line.

of three games. At first sight, this question appears to be rather similar to the question asked before. However, the previous question was independent of θ, and the current question is not. Consequently, it will be answered using a discrepancy measure (which is a function of both θ, and, observed data y^{obs} or replicated data y^{rep}). The posterior predictive p-value is the proportion of draws for which the replicated discrepancy $D(y^{rep}, \theta_{(k)})$ is larger than the observed discrepancy $D(y^{obs}, \theta_{(k)})$. A value of .69 indicates that the number of relatively 'hot' series is not larger than expected under the binomial model.

One may conjecture that having a 'hot' hand depends on the number of attempts at a field goal. The latter would imply that y_i/n_i and n_i are positively correlated. In Figure 5, the posterior predictive distribution of Spearman's correlation coefficient (a statistic) and its observed value are displayed. As can be seen, the observed value of .44 (posterior predictive p-value .00) is highly unusual under the binomial model. It appears that there is a significant and relevant correlation between the probability of a field goal, and, the number of attempts at a field goal. Note that Spearman's correlation could have been computed and evaluated without the posterior predictive machinery, using more 'classical' methods.

A possible explanation for the significant correlation is that Vinnie Johnson plays better if he gets a lot of playing time, for instance because the coach notices that Vinnie Johnson has a 'hot' hand. A second possible explanation is that if Vinnie Johnson plays well, he gets more passes from the other team members and consequently, his number of field attempts will be higher. The plausibility of these explanations could be examined by a posterior predictive check using a discrepancy measure which depends on total playing time per

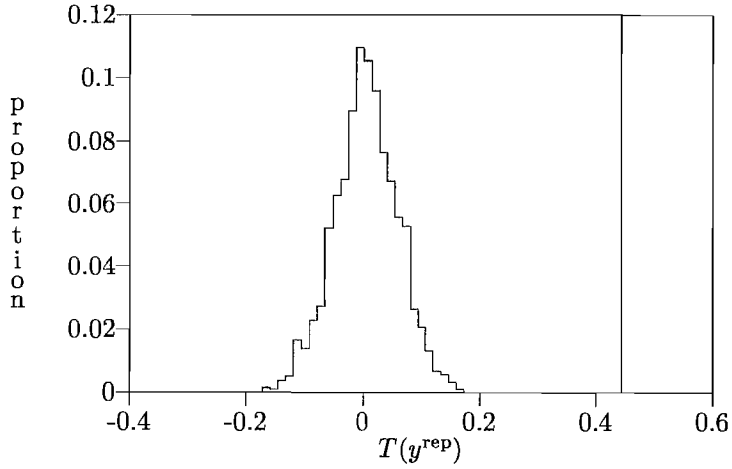


Figure 5: Spearman's correlation coefficient. The position of $T(y^{obs})$ is denoted by the vertical line.

game.

From a model building perspective, posterior predictive inference has been used to obtain ideas for the construction of an alternative model. Given the above, this model should incorporate the dependence between the probability of scoring a field goal and the number of attempts at a field goals. One could, for example, use $y_i \sim Bin(n_i, \theta_i)$ with the specification $\theta_i = \frac{exp(\gamma_0+\gamma_1 n_i)}{1+exp(\gamma_0+\gamma_1 n_i)}$.

4 Interpretation of p-values

The posterior predictive p-value and the classical p-value have a similar interpretation since they coincide if the classical p-value is defined. Therefore, in this section we only discuss the interpretability of the classical p-value.

In Bayesian literature, classical hypothesis testing is often compared with Bayesian approaches to the selection of a model from a discrete set of alternatives. The standard Bayesian solution is to compare the alternatives by means of the Bayes factor (Kass and Raftery, 1995) or a variant of the Bayes factor (Gelfand and Dey, 1994).

The Bayes factor for hypothesis H_0 against H_1 is equal to the posterior odds over the prior odds:

$$BF = \frac{Pr\{H_0|y^{obs}\}}{Pr\{H_1|y^{obs}\}} / \frac{Pr\{H_0\}}{Pr\{H_1\}} . \quad (6)$$

From the above definition, it follows that if both hypotheses have equal prior probability of being true, then BF is larger than one if H_0 has a higher posterior probability of being true. This means that the Bayes factor is a measure

of the relative evidence in favor of H_0. The p-value cannot be interpreted as measures of evidence in favor of an alternative. At best, one may interpret the p-value as evidence against H_0. However, a Bayesian objection against the latter interpretation is that the p-value tends to overstate the correct evidence against H_0, i.e. $Pr\{H_0|y^{\text{obs}}\}$. (The evidence against H_0 may also be measured by $Pr\{H_1|y^{\text{obs}}\}$ but the latter probability is not directly comparable to the p-value). Edwards, Lindsey, and Savage (1963) and Berger and Sellke (1987) showed this for a two-sided z-test where the mean parameter θ is equal to 0 under H_0 and has a prior distribution under H_1. The main result of Edwards, Lindsey, and Savage (1963) and Berger and Sellke (1987) is that with a p-value equal to .05, $Pr\{H_0|y^{\text{obs}}\}$ has a lower bound of .065 over all possible prior distributions specified under H_1. This means that the p-value overstates the evidence against H_0 by at least .065 - .05 = .015, and hence the p-value can impossibly be a correct measure of evidence. If, instead of all priors, only a set of reasonable priors (like for instance all symmetric priors) is considered under H_1, then the smallest possible value of $Pr\{H_0|y^{\text{obs}}\}$ may be even much larger.

According to Berger and Sellke (1987), the reason why the p-value overstates the evidence against H_0 is the following. When computing the Bayes factor, one only conditions on the observed data y^{obs}. The p-value, on the other hand, is the probability of observing a value of y in $\mathcal{A} = \{y : |T(y)| > |T(y^{\text{obs}})|\}$ with $T(y)$ the statistic of the two-sided z-test. As pointed out by Berger and Sellke (1987), "to calculate a p-value, one effectively replaces y^{obs} by the 'knowledge' that y is in $\mathcal{A}$". To motivate this point further, Berger and Sellke (1987) argued that if instead of observing y^{obs}, we only knew that $y^{\text{obs}} \in \mathcal{A}$, then the p-value does not have to overstate the evidence against H_0. The minimum value of $Pr\{H_0|\ y^{\text{obs}} \in \mathcal{A}\}$ over all possible priors is $p/(1+p)$ which is smaller than p.

Notwithstanding the arguments above, the p-value could still be used to detect violations of H_0. In this respect, a p-value should not be regarded as a formal measure of evidence, rather, a p-value of for instance .001 should at least cast doubts about H_0. The p-value can therefore be used in a diagnostic model checking procedure: Small p-values suggest that the null model does not fit the data, in which case it may be worthwhile to extend the model such that a better fit is obtained.

5 Frequency properties of the posterior predictive p-value

5.1 The purpose of frequency calculations

The performance of a classical (frequentist) test is usually characterized by the following two quantities: (1) Type 1 error rate $Pr\{p_C \leq \alpha|H_0\}$ which should not be larger than the nominal significance level α, and (2) Power

$Pr\{p_C \leq \alpha|H_1\}$ which should be as large as possible. Note that these quantities are viewed as relative frequencies in a long sequence of tests and therefore are theoretically distinct from the p-value itself.

Although the posterior predictive p-value is closely related to the classical frequentist p-value, it may be asked whether it is useful to evaluate a posterior predictive test in terms of long term relative frequencies. Bayesian inference is conditional on the observed data set, which seems to make frequency calculations unnecessary. Rubin (1984), however, pointed out that some test quantities are used in more than one data analysis. In order to evaluate whether such a general use of a test quantity is recommendable, it seems worth knowing about the frequency performance of the p-value across different (hypothetically) observed data sets.

5.2 Drawing new data sets under H_0

For studying the Type I error rate of the posterior predictive p-value, we draw S new data sets $y^{(1)}, \ldots, y^{(S)}$ from a population model $p(y|\theta)g_0(\theta)$. The distribution $g_0(\theta)$ only has to be specified if the null set of parameters (i.e. Θ_0) contains "nuisance" parameters which are not fixed by H_0. In that case, the new data sets are obtained by a two step procedure: In the first step, parameter vectors $\theta^{(1)}, \ldots, \theta^{(S)}$ are drawn from $g_0(\theta)$ and in the second step, $y^{(1)}, \ldots, y^{(S)}$ are drawn from $p(y|\theta)$.

It is important to realize that $g_0(\theta)$ is usually different from the prior $p(\theta)$. The two distributions have in common that they become point distributions with mass at θ_0 if Θ_0 is a singleton. However, whereas $p(\theta)$ might for instance be a vague prior with mass at all possible values of θ, $g_0(\theta)$ should yield values $\theta^{(1)}, \ldots, \theta^{(S)}$ which are plausible for the types of problems under study. To illustrate this, suppose we want to develop a test for use in regression models with length as the dependent variable. If we do now want to confine the applicability of the test to a particular species, we might want to generate new lengths between 0 to 40 metres. If, however, the test will only be applied in bird studies, values above 2 metres are irrelevant. In that case, we might set $g_0(\theta)$ equal to a uniform distribution on the interval $(0, 2)$. For the prior $p(\theta)$, we might choose the same distribution but choosing an improper uniform distribution on $(-\infty, \infty)$ seems defensible as well.

5.3 Conditions for uniformity of the p-value under H_0

As mentioned in Section 5.1, it is desirable that the Type 1 error rate $Pr\{p \leq \alpha|H_0\} \leq \alpha$. If the p-value is a continuous function on interval $[0, 1]$, it is best that $Pr\{p < \alpha|H_0\} = \alpha$ in which case p is uniformly distributed on $[0, 1]$. Then the test is neither conservative nor liberal.

Suppose that we perform a posterior predictive test using test statistic

$T(y)$. The posterior predictive p-value, defined in (4), can be written as

$$p_T = Pr\{T(y^{\text{rep}}) \geq T(y^{\text{obs}})|y^{\text{obs}}, H_0\} . \quad (7)$$

A necessary and sufficient condition for uniformity of p_T under H_0 is that $T(y^{\text{obs}})$ and the replications $T(y^{\text{rep}}_{(1)}), \ldots, T(y^{\text{rep}}_{(K)})$ are drawn from the same distribution. To proof this, apply for instance Theorem 5.12 in Mood, Graybill, and Boes (1974, p. 202) to (7). Two cases in which p_T is uniformly distributed, are:

1. The posterior predictive test is based on a pivotal test statistic $T(y)$.
2. The posterior distribution (under H_0) is a point distribution with mass at θ^*.

Regarding case 1, one might be tempted to extend this case to tests based on a discrepancy measure $D(y, \theta)$ of which the density function $p(D(y, \theta)|\theta)$ is free of θ. Yet, although such a discrepancy measure is pivotal, the corresponding posterior predictive p-value is usually nonuniform. Example 1 in Meng (1994) confirms this.

Case 2 also holds for a disrepancy measure $D(y, \theta)$. One may further note that under standard assumptions, the posterior converges to a point distribution with mass at θ^* if the sample size goes to infinity. Then if the posterior predictive distribution of $T(y)$ (or $D(y, \theta)$) converges to the asymptotic distribution of $T(y)$ (or $D(y, \theta^*)$) given θ^*, the null distribution of the p-value converges to a uniform distribution. Within this context, one may also notice the difference between the parametric bootstrap test and the posterior predictive test. The former one is obtained by substituting $p(y|\hat{\theta})$ with $\hat{\theta}$ the maximum likelihood estimate of θ for the posterior predictive distribution $p(y|y^{\text{obs}})$. Asymptotically, $\hat{\theta}$ equals θ^* and in that case the parametric bootstrap test and the posterior predictive test yield the same result.

5.4 Nonuniformity of the p-value under H_0

There are only few results on the repeated sampling performances of the posterior predictive p-values p_D and p_T as defined in Section 2. The most important ones have been derived by Meng (1994). The results of Meng (1994) are valid if we set $g_0(\theta)$ equal to the prior $p(\theta)$. Then the new datasets are drawn from the prior predictive distribution defined by Box (1980):

$$p(y) = \int p(y|\theta)p(\theta)d\theta . \quad (8)$$

Meng (1994) showed that $E(p_D) = \frac{1}{2}$ and

$$E\{h(p_D)\} \leq E\{h(U)\} , \quad (9)$$

where $h(.)$ is a convex function and U is a uniformly distributed variable on $[0, 1]$. Note that (9) implies that $\text{var}(p_D) \leq \text{var}(U)$. Meng (1994) further showed that (9) implies that the Type I error rate of p_D will never exceed twice the nominal value α under any prior predictive distribution:

$$Pr\{p_D \leq \alpha|H_0\} \leq 2\alpha .$$

In practice, however, the density function of p_D will have smaller tails than the uniform distribution and hence the Type I error rate tends to be smaller than the nominal value. Because a test statistic is a special type of discrepancy measure, the above results also hold for p_T. The same results do not necessarily hold if a test quantity is used that is a function of both y and y^{obs}. An example of such a quantity is the posterior mean of a discrepancy measure:

$$\int D(y, \theta)p(\theta|y^{\text{obs}})d\theta . \quad (10)$$

It is further interesting to note that if the new datasets are drawn from the prior predictive distribution, then the prior predictive p-value is uniformly distributed under H_0. There is no analogy for posterior predictive checks; if new datasets are drawn from the posterior predictive distribution, then p_D does not have to be uniformly distributed under H_0.

5.5 Power of the posterior predictive test

The power can be studied in an analogous way as the Type I error. With regard to the simulation of new data sets, the only difference is that we replace $g_0(\theta)$ by $g_1(\theta)$, where $g_1(\theta)$ is a distribution over all parameters which are not fixed by the parameter space under H_1 (i.e. Θ_1).

Although little is known about the power of the posterior predictive test, the posterior predictive test can be criticized for lack of power due to using the data twice: Once to obtain the posterior and once to assess whether the posterior is consistent with the observed data (Carlin and Louis, 1996, p. 57). Consequently, the posterior is pulled towards the observed data and model violations might be camouflaged. In order to minimize the effect of using the data twice, it is advisable to choose the test quantity such that the posterior predictive distribution of the test quantity depends on the posterior as few as possible. Notably, using the data twice does not affect the distribution of a test statistic if it is pivotal.

6 Prior sensitivity

So far, we have drawn a clear distinction between formulating a null hypothesis $H_0 : \theta = \theta_0$ and formulating the prior distribution. This distinction seems a bit artificial since the difference between the null model and an alternative model can be formulated in terms of the prior distribution. In the null

model, the parameters which are fully specified in Θ_0 may considered as having a prior point distribution. In the alternative model, the same parameters have a more diffuse prior. There is, however, an important difference between point priors and other priors: The effect of a point prior on the posterior does not vanish when the sample size tends to infinity. Therefore, we think that there is statistical ground for distinguishing between the formulation of a null hypothesis $H_0 : \theta = \theta_0$ and the specification of a diffuse prior distribution.

In case the sample size is small, the posterior predictive p-value may become prior sensitive. On the one hand, this sensitivity may be undesirable if one wishes to check the fit of the likelihood part of the model $p(y|\theta)$ only; on the other hand, this sensitivity may enable us to formulate tests which aim to detect prior-related inconsistencies between the posterior and the data.

To illustrate the issue of the prior sensitivity, suppose a positive integer y^{obs} is observed and the following likelihood is assumed: $y \sim Bin(n, \theta)$. For the success probability θ, a $Beta(\alpha, \alpha)$ prior is chosen. The replicated data sets $y^{\text{rep}}_{(1)}, \ldots, y^{\text{rep}}_{(K)}$ are then to be drawn from the $Betabin(n, \alpha + y^{\text{obs}}, \alpha + n - y^{\text{obs}})$ distribution. Let us consider a posterior predictive check with test statistic $T(y) = y$. In Table 1, the test results are presented that are obtained when y^{obs} is equal to 0 and n is equal to 25. The hyperparameter α is set equal to .1, .5, 1, 2, and 10. Note that if $\alpha < 1$, the prior is bimodal with modes at 0 and 1. If $\alpha = 1$, the prior is uniform, and if $\alpha > 1$, the prior is unimodal with mode at $\frac{1}{2}$.

From the table, we can read that by changing the prior, p-values ranging from 0 to 1 can be obtained. For instance, the p-value is near 0 if α is equal to 10; the reason for this is that observing $y^{\text{obs}} = 0$ is unlikely when the prior density for the success probability is $Beta(10, 10)$, which is a sharply peaked unimodal function with mode at .5. The p-value is near 1 if α equals .1; this can be explained as follows: A $Beta(.1, .1)$ density is a sharply peaked bimodal function with modes at 0 and 1; the posterior of θ then is a peaked unimodal density with mode near 0. Therefore, p_T (i.e. $Pr\{y \leq 0|y^{\text{obs}}\}$) is near 1. Note that the latter extreme p-value is somewhat puzzling as a $Beta(.1, .1)$ prior can be hardly seen as inconsistent with $y^{\text{obs}} = 0$.

A general advice is to check the sensitivity of the posterior predictive p-value with regard to changes in the prior distribution. If the prior sensitivity is nonnegligible, it seems better to report a range of p-values obtained under different reasonable priors than one specific p-value.

Table 1: p_T for $y^{\text{obs}} = 0$; $y \sim Bin(25, \theta)$; $\theta \sim Beta(\alpha, \alpha)$.

α	.1	.5	1	2	10
p_T	.93	.71	.51	.27	.007

7 Concluding remarks

This paper discusses the posterior predictive check for detecting model violations. An attractive feature of the posterior predictive check is that it can rely on discrepancy measures that may be any function of the data and model parameters; this enables us to define discrepancy measures that address the scientific questions of interest. Another argument for performing posterior predictive checks is that they are easy to perform and do not require large computational effort (once we have draws from the posterior): Only draws are needed from the posterior predictive distribution of the posited model.

Notwithstanding the advantages of posterior predictive checking, the posterior predictive p-value can be discussed on several points. First, like the classical p-value, the posterior predictive p-value may be difficult to interpret. When model selection instead of model fit is of primary concern, one might consider to use quantities like the Bayes factor rather than the *ppc* p-value. Second, the frequentist properties of the posterior predictive p-value are often unclear especially if the sample size is moderate or small. Third, the posterior predictive check may be prior sensitive if asymptotics do not hold. Then special care has to be taken when interpreting the p-value. More research is required to figure out under which prior(s) the p-value best reflects inconsistencies between the data and the null hypothesis.

References

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*, 112-122.

Box, G. E. P. (1980). Sampling and Bayes inference in scientific modelling and robustness. *Journal of the Royal Statistical Society, series A, 143*, 383-430.

Carlin, B.P., and Louis, T.A. (1996). *Bayes and Empirical Bayes Methods for Data Analysis.* London: Chapman & Hall.

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

Edwards, W., Lindman, H., and Savage, L.J. (1963). Bayesian statistical inference for psychological research. *Psychological Review, 70*, 193-242.

Gelfand, A.E., and Dey, D.K. (1994). Bayesian model choice: asymptotics and exact calculations. *Journal of the Royal Statistical Society, series B, 56*, 501-514.

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

Gelman, A., Goegebeur, Y., Tuerlinckx, F., and Van Mechelen, I. (in press). Diagnostic checks for discrete-data regression models using posterior predictive simulations. *Applied Statistics.*

Gelman, A., Meng, X.L. and Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica, 6,* 733-807.

Gilovich, T., Vallone, R. and Tversky, A. (1985). The hot hand in basketball: On the misperception of random sequences. *Cognitive Psychology, 17,* 295-314.

Hoijtink, H. and Molenaar, I.W. (1997). A multidimensional item response model: constrained latent class analysis using the Gibbs sampler and posterior predictive checks. *Psychometrika, 62,* 171-190.

Hsiao, K. (1994). Bayesian tests of extra-binomial variability with emphasis on the boundary case. Ph.D. dissertation, Carnegie Mellon University, Department of Statistics.

Kass, R.E. and Raftery, A.E. (1995). Bayes Factors. *Journal of the American Statistical Association, 90,* 773-795.

Lewis, S.M., and Raftery, A.E. (1996). Comment on "Posterior predictive assessment of model fitness via realized discrepancies" by Gelman, A., Meng, X.L., and Stern, H.S. *Statistica Sinica, 6,* 779-786.

Meng, X.L. (1994). Posterior predictive p-values. *The Annals of Statistics, 22,* 1142-1160.

Mood, A.M., Graybill, F.A., Boes, D.C. (1974). *Introduction to the Theory of Statistics*, 3rd edition. Singapore: McGraw-Hill.

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