Electronic Journal of Statistics
ISSN: 1935-7524
arXiv: math.PR/0000000

Understanding posterior p-values

Andrew Gelman*

Department of Statistics, Columbia University, New York

Abstract: Posterior predictive p-values do not in general have uniform distributions under the null hypothesis (except in the special case of ancillary test variables) but instead tend to have distributions more concentrated near 0.5. From different perspectives, such nonuniform distributions have been portrayed as desirable (as reflecting an ability of vague prior distributions to nonetheless yield accurate posterior predictions) or undesirable (as making it more difficult to reject a false model). We explore this tension through two simple normal-distribution examples. In one example, we argue that the low power of the posterior predictive check is desirable from a statistical perspective; in the other, the posterior predictive check seems inappropriate. Our conclusion is that the relevance of the p-value depends on the applied context, a point which (ironically) can be seen even in these two toy examples.

AMS 2000 subject classifications: Primary 62F15, 62C10, 62F03.
Keywords and phrases: Bayesian inference, model checking, posterior predictive check, p-value, u-value.

1. Introduction

Bayesian predictive checking generalizes classical hypothesis testing by averaging over the posterior distribution of the unknown parameter vectors rather than fixing them at some point estimate (Rubin, 1984, Gelman, Meng, and Stern, 1996). Bayesian tests do not rely on the construction of pivotal quantities or on asymptotic results, and are therefore applicable to any probability model. This is not to suggest that the tests are automatic; as with classical testing, the choice of test quantity and appropriate predictive distribution requires careful consideration of the type of inferences required for the problem being considered.

Posterior predictive p-values do not in general have uniform distributions under the null hypothesis (except in the special case of ancillary test variables) but instead tend to have distributions more concentrated near 0.5 (Meng, 1994). This has led to criticism that these predictive probabilities are not calibrated (Bayarri and Berger, 2000). The nonuniformity of the posterior distribution has been attributed to a double use of the data (Bayarri and Castellanos, 2007), although this latter claim has been disputed based on the argument that the predictive p-value is a valid posterior probability whether or not its marginal distribution is uniform (Gelman, 2007). From an applied direction, some researchers have proposed adjusted predictive checks that are calibrated to have

*We thank two reviewers and the Institute of Education Sciences and the National Science Foundation for partial support of this work.

asymptotic uniform null distributions (Robins, Vaart, and Ventura, 2000); others have argued that, in applied examples, posterior predictive checks are directly interpretable without the need for comparison to a reference uniform distribution (Gelman et al., 2003).

In this brief note we address the practical concern about the nonuniformity of the marginal distribution of the posterior predictive p-value: when is it desirable (as reflecting an ability of vague prior distributions to nonetheless yield accurate posterior predictions) and when is it undesirable (as making it more difficult to reject a false model)?

In Sections 3 and 4 of this note, we consider two simple examples, one of which shows the posterior predictive check performing well (and for which a calibrated p-value would miss the point entirely) and one of which shows the posterior predictive check being essentially useless while the calibrated p-value would work fine.

To make our comparisons clear, we design both examples so that there is a large amount of posterior uncertainty so that the distribution of the posterior p-value is highly concentrated around 0.5. Thus we are focusing on the key question of the appropriateness of a model check that, in some settings (when the prior is weak so the predictive distribution is close to the data), will essentially never reject. Our conclusion is that the relevance of the p-value depends on the applied context, a point which (ironically) can be seen even in these two toy examples.

2. Goals of p-values and model diagnostics

For the purposes of this paper, our sole goal in model checking will be to reveal settings where the model's predictions are inconsistent with observed data. Thus, our concern is with future uses of the fitted model. In an Bayesian context, our only concern is obtaining a satisfactory posterior distribution; in a classical setting we want to be assured that the data are coherent with predictive inferences given point or interval estimates.

In particular, there are two goals we are *not* addressing here. First, we are not concerned with statistical power in the usual sense; we have no interest in maximizing the probability of rejecting a model, conditional on it being false. In our work, all our models are false, and we know ahead of time that they could be clearly rejected given sufficient data. We are working within a world in which the purpose of a p-value or diagnostic of fit is to reveal systematic differences between the model and some aspects of the data; if the model is false but its predictions fit the data, we do not want our test to reject.

The second goal we shall not address is the computation or evaluation of Bayes factors or posterior model probabilities. In noninformative or weakly informative Bayesian inference, a model can predict well but with an unrealistic prior distribution that results in a marginal model probability that is at best meaningless and at worst highly misleading. This problem is central to the Jeffreys-Lindley paradox (Lindley, 1957). In this paper, when we say that a

model fits the data well, we are speaking of posterior predictions conditional on the model; we are not considering posterior model probabilities that would be used for Bayesian model averaging.

Thus, the ideas in these paper consider p-values in the context of goodness-of-fit testing for a fitted model, not as tests that would be inverted to obtain confidence intervals, and not as tests of prior distributions. From our perspective, a p-value near zero is relevant because it reveals a systematic misfit of data to some aspect of the fitted predictive distribution.

3. *P*-values and *u*-values

Consider a Bayesian model with parameters θ and continuous data y. These ideas generalize to hierarchical models but for simplicity we will consider θ as a single block with potential replications y^{rep} defined given the same values of θ as produced the observed data. Further suppose the model will be checked based on the p-value of some test statistic $T(y)$.

In the special case that θ is known (or estimated to a very high precision) or in which $T(y)$ is ancillary (or nearly so), the posterior predictive p-value $\Pr(T(y^{\text{rep}} > T(y)|y)$ has a distribution that is uniform (or approximately uniform) if the model is true. Under these conditions, p-values less than 0.1 occur 10% of the time, p-values less than 0.05 occur 5% of the time, and so forth.

More generally, when posterior uncertainty in θ propagates to the distribution of $T(y|\theta)$, the distribution of the p-value, if the model is true, is more concentrated near the middle of the range: the p-value is more likely to be near 0.5 than near 0 or 1. (To be more precise, the sampling distribution of the p-value has been shown to be 'stochastically less variable' than uniform.)

To clarify, a *u-value* is any function of the data y that has a $\text{U}(0,1)$ sampling distribution. A u-value can be averaged over the distribution of θ to give it a Bayesian flavor, but it is fundamentally *not* Bayesian, in that it cannot necessarily be interpreted as a posterior probability (Gelman, 2003). In contrast, the posterior predictive p-value is such a probability statement, conditional on the model and data, about what might be expected in future replications.

The p-value is to the u-value as the posterior interval is to the confidence interval. Just as posterior intervals are not, in general, classical confidence intervals (in the sense of having the stated probability coverage conditional on any value of θ), Bayesian p-values are not generally u-values.

This property has led some to characterize posterior predictive checks as conservative or uncalibrated. We do not think such labeling is helpful; rather, we interpret p-values directly as probabilities. The sample space for a posterior predictive check—the set of all possible events whose probabilities sum to 1—comes from the posterior distribution of y^{rep}. If a posterior predictive p-value is 0.4, say, that means that, if we believe the model, we think there is a 40% chance that tomorrow's value of $T(y^{\text{rep}})$ will exceed today's $T(y)$. If we were able to observe such replications in many settings, and if our models were actually true, we could collect them and check that, indeed, this happens 40% of the

time when the p-value is 0.4, that it happens 30% of the time when the p-value is 0.3, and so forth. These p-values are as calibrated as any other model-based probability, for example a statement such as, 'From a roll of this particular pair of loaded dice, the probability of getting double-sixes is 0.11,' or, 'There is a 50% probability that Barack Obama won more than 52% of the white vote in Michigan in the 2008 election.'

That said, it can sometimes be helpful to compare posterior p-values to their corresponding recalibrated u-values under the prior predictive distribution.

We consider two examples in which the posterior p-value has a sampling distribution that is strongly peaked near 0.5. In the first example, this highly nonuniform posterior predictive distribution is fine, and we do not see the corresponding u-value as making much sense at all; rather, it destroys the useful meaning of the test. In the second example, however, the noise in the predictive distribution makes the raw test (and its posterior predictive p-value) essentially unusable, and the recalibrated u-value seems much closer to what would be desired in practice.

4. Example where the posterior predictive p-value is stuck near 0.5 and this is desirable: Testing the sample mean as fit by a normal distribution

Consider the data model, $y \sim \mathrm{N}(\theta, 1)$ and prior distribution $\theta \sim \mathrm{N}(0, A^2)$, with the prior scale A set to some high value such as 100 (thus, a noninformative prior distribution). We will use as the test statistic the sample mean, y. In this case, y is just a single data point, but that's just for mathematical convenience. The point is that we're using the sample mean or regression coefficient to test the fit of a normal linear model.

The model check based on the sample mean will essentially never reject. In fact, the posterior predictive p-value, $\Pr(y^{\mathrm{rep}} > y|y)$, will be near 0.5 for just about any y that might reasonably come from this model. Here is the math:

$$\begin{aligned} \theta|y &\sim \mathrm{N}\left(\frac{A^2}{A^2+1}y,\, \frac{A^2}{A^2+1}\right) \\ y^{\mathrm{rep}}|y &\sim \mathrm{N}\left(\frac{A^2}{A^2+1}y,\, 1+\frac{A^2}{A^2+1}\right). \end{aligned}$$

The posterior predictive p-value is

$$p\text{-value} = \Phi\left(-\frac{y - \mathrm{E}(y^{\mathrm{rep}}|y)}{\mathrm{sd}(y^{\mathrm{rep}}|y)}\right),$$

and the marginal (prior predictive) distribution of y is $\mathrm{N}(0, A^2+1)$.

Plugging in $A = 100$, the posterior predictive p-value is $\Phi(-y/14{,}000)$, so to get a (one-sided) p-value of 0.025, you need $y > 28{,}000$, an extremely unlikely event if the marginal distribution of y is $\mathrm{N}(0, 100^2)$. Even in the extremely unlikely event of $y = 500$ (that is, five standard deviations away from

the mean, under the prior predictive distribution), the p-value is still only $\Phi(-500/14{,}000) = 0.486$. Thus, in this example, we can be virtually certain that the p-value will fall between 0.48 and 0.52.

If we wanted, we could adjust the p to have a uniform distribution under the prior predictive distribution. The recalibrated p-value (the u-value) is simply the normal distribution function evaluated at $y/\sqrt{A^2+1}$. For example, with $A = 100$ and $y = 500$, this recalibrated p-value would be 3×10^{-7}.

But this is not what we want! Continue with this example in which $A = 100$ and the observed data are $y = 500$. Yes, this is an unexpected value, inconsistent with the prior distribution, and worthy of ringing the alarm in a prior predictive test. But what about the posterior distribution? The inference is good: $p(\theta|y) = \mathrm{N}(\theta|499.95, 0.99995^2)$. In this case the prior distribution is acting noninformatively, despite being contradicted by the data. The posterior distribution is fine, and an extreme p-value would be inappropriate for our goal of checking the fit of the posterior distribution to this aspect of the data. If, however, we were interested in the prior distribution (for the purpose of computing the marginal probability of the data for use in a Bayes factor, for example), then the lack of it of the prior would be important.

To put it another way, consider this example, keeping the data at $y = 500$ and sliding the value of the prior scale, A, from 50 up to 5000. As this prior scale changes, there is essentially no change in the posterior distribution: under all these values, $p(\theta|y)$ is extremely close to $\mathrm{N}(\theta|500, 1)$. And, correspondingly, the posterior predictive p-value for the test statistic $T(y) = y$ is close to 0.5. But the u-value—the p-value recalibrated to have a uniform prior predictive distribution—changes a lot, as does the marginal probability of the data.

We do not like the u-value as a measure of model fit here, as it is sensitive to aspects of the model that have essentially no influence on the posterior distribution. The mean really is well estimated by a normal linear model. This is not 'conservatism' or 'low power' of a statistical test; it reflects the real ability of this model to fit this aspect of the data. We have no particular need for p-values to have a uniform distribution; we can interpret them directly as probability statements about y^{rep}.

5. Example where the posterior predictive *p*-value is stuck near 0.5 and this is undesirable: Testing skewness in the presence of huge amounts of missing data

We next consider an example where the posterior predictive p-value gives a bad answer but a recalibrated u-value succeeds. Consider the normal model with n independent data points, $y_i \sim \mathrm{N}(\theta, 1)$, a flat prior distribution, $\theta \sim \mathrm{N}(0, A^2)$, with a high value of A, and use the sample skewness as a test statistic: $T(y) = \frac{1}{n}\sum_{i=1}^{n}\left((y_i - \bar{y})/s_y\right)^3$. This test variable is ancillary, thus its p-value (which can be calculated using the t distribution function) is also a u-value.

So far so good. Now we muddy the waters by supposing that $n = 1000$ but with only 10 values observed; the other 990 cases are missing. And we still

want to use the sample skewness as a test variable. As pointed out by Meng (1994), The Bayesian p-values averages over the distribution of unknowns and thus can be computed for discrepancy variables that depend on parameters or latent variables.

We could simply compute the sample skewness of the 10 observations; this will yield a p-value (and a u-value) that is just fine. But that would be cheating. We want to do our test on the full, completed data—all 1000 observations—using the posterior distribution to average over the 990 missing values.

This should work fine under recalibration: all the noise introduced by the missing data gets averaged over in the calibration step. But the straight posterior predictive check does not work so well. The problem is that the test statistic will be dominated by the 990 imputed values. These will be imputed under the model—that is what it means to average over the posterior distribution—and so it will be virtually impossible for the test to reveal any problem with the data, even if as a sample of 10 they are clearly skewed.

If we let y^c be the completed dataset (that is, the 10 observed values y_i, along with 990 latent data points y_i^{missing}), then our realized test variable is $T(y^c) = \frac{1}{1000}\sum_{i=1}^{1000}\left((y_i^c - \bar{y^c})/s_{y^c}\right)^3$, and the p-value is $\Pr(T(y^{c\,\text{rep}}) > T(y^c)|y)$, where $y^{c\,\text{rep}}$ is a replicate completed dataset of length 1000 drawn from the model. One would compute the p-value as follows: First draw some number S simulations of θ from the posterior distribution, $p(\theta|y)$. Then do the following for each simulated θ: (1) sample the 990 missing values y_i^{missing} given θ and append them to y to construct y^c; (2) independently sample 1000 new (replicated) observations given θ to create $y^{c\,\text{rep}}$; (3) compute $T(y^c)$ and $T(y^{c\,\text{rep}})$ and record which is larger. Finally, compute the p-value as the proportion of the S cases in which $T(y^{c\,\text{rep}})$ exceeds $T(y^c)$. The difficulty is that 990 of the observations in y^c are being drawn from the very same posterior distribution that is being used to simulate y^{rep}, and thus it will be almost impossible for this realized-discrepancy test to reject—even if the 10 actual observations reveal very poor fit to the normal distribution.

What the posterior predictive test is saying here is that the sample skewness of a new set of 1000 observations will look similar to that of 1000 data points so far (of which only 10 have been observed), if the model is in fact true. This is fine as a mathematical statement but not so relevant for practical evaluation of model fit.

Gelman et al. (2000) found a similar problem when trying to check discrete-data regressions using realized discrepancies based on latent continuous outcomes. The level of noise introduced by imputing the continuous variables overwhelmed the ability of residual plots to reveal model misfit. In that case the checks were graphical rather than numerical and no p-values were involved, but the same general principle held: when the predictive check was defined based on information imputed from the model itself, this degrades the ability of the test to reveal poor model fit.

Finally, we close these two examples by emphasizing that the results from any of these diagnostics depends on the test quantities being considered. For

example, suppose a continuous model is fit to data that take on only a few discrete values. Such a model can be fine for estimating expectations (indeed, there is a literature in econometrics on the validity of linear regression inferences even if applied to discrete data) but will fare poorly when predicting individual data points—and a practitioner who is concerned about such predictions should keep this in mind when performing diagnostics.

6. Discussion

We have now seen two examples where, if the model is true or close to true, the posterior predictive p-value will almost certainly be very close to 0.5. In the first example (checking the sample mean under a normal model), this is just fine and appropriately reflects the ability of the model to capture a particular aspect of the data; in the second example (with 99% missing cases), the posterior p-value does not work so well. This reinforces the general principle that model checks, like other statistical inferences, are defined relative to some external purpose, and it also suggests the difficulties of any general recommendations on the use of p-values or u-values for practical model checking.

Our conclusion might well seem unsatisfactory, as we do not offer clear guidelines on how best to interpret p-values and u-values. Given the subtleties that arise even in simple toy examples, how can we proceed in more complex problems where the stakes are higher but the mathematics is less clear?

We offer two recommendations. First, if the goal of testing and diagnostics is (if as we believe it should be) to discover systematic discrepancies between the data and fitted model, that there is no *general* reason to be bothered by p-value distributions that are far from uniform. It is generally highly implausible to suppose that a model is true (especially in a Bayesian context in which there is a prior distribution to worry about as well!), and thus it is good, not bad, that an imperfect model can provide a good predictive distribution, after being fit to data. Second, we recommend caution in interpreting diagnostics that depend strongly on parameters or latent data. It is a theoretically appealing feature of Bayesian model checking that the test variable can depend on unknown quantities, but statements of fit based on such realized discrepancies can be difficult to interpret. So we recommend avoiding realized discrepancy measures unless the amount of imputed information contained in them is very small.

References

Bayarri, M. J. and Berger, J. (2000). P-values for composite null models. *Journal of the American Statistical Association* **95**, 1127–1142.

Bayarri, M. J., and Castellanos, M. E. (2007). Bayesian checking of the second levels of hierarchical models (with discussion). *Statistical Science* **22**, 322–367.

Gelman, A., Goegebeur, Y., Tuerlinckx, F., and Van Mechelen, I. (2000). Diagnostic checks for discrete-data regression models using posterior predictive simulations. *Applied Statistics* **49**, 247–268.

Gelman, A. (2003). A Bayesian formulation of exploratory data analysis and goodness-of-fit testing. *International Statistical Review* **71**, 369–382.

Gelman, A. (2007). Discussion of 'Bayesian checking of the second levels of hierarchical models,' by M. J. Bayarri and M. E. Castellanos. *Statistical Science* **22**, 349–352.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2003). *Bayesian Data Analysis*, second edition. London: CRC Press.

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

Lindley, D. V. (1957). A statistical paradox. *Biometrika* **44**, 187–192.

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

Robins, J. M., Vaart, A., and Ventura, V. (2000). Asymptotic distribution of p values in composite null models. *Journal of the American Statistical Association* **95**, 1143–1156.

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