

James M. ROBINS, Aad VAN DER VAART, and Valérie VENTURA

Bayarri and Berger shared their rejoinder with us before we began writing ours. We agree with almost everything they say, including their appreciation of the quality of the discussions. Here we add some observations of our own that we think might help clarify the issues.

Stern. In his Section 4, Stern acknowledges that using the statistic \bar{X} for testing a $N(0, \sigma^2)$ model results in a conservative posterior predictive p value. But he suggests that this reflects Bayarri and Berger's poor choice of test statistic and that he, in contrast, has been able to construct intuitive discrepancy measures that provide "great insight" into the performance of the model. Stern does not say what he means by "great insight." If he means that his intuitive discrepancies successfully avoid the problem of very conservative discrepancy p values, then he is incorrect, as we showed in our Example C, which was taken from his book coauthored with Gelman, Carlin, and Rubin. Specifically, Stern and his coauthors proposed the discrepancy $t(\mathbf{X}, \theta) = |\hat{Z}_{.9} - \mu| - |\mu - \hat{Z}_{.1}|$ to check whether the observed degree of skewness, as measured by $t(\mathbf{x}_{\text{obs}}, \theta)$, was compatible with a $N(\mu, c^2)$ distribution, where \hat{Z}_q is the empirical q th quantile of the data. A discrepancy p value (p_{dis}) of .05 corresponds to a true (i.e., adjusted or calibrated) p value of .0004, based on the calculations leading to Figure 4(a). That is, the actual asymptotic level of the nominal .05-level test based on p_{dis} is .0004, indicating that p_{dis} is extremely conservative.

By comparing $p_{\text{dis}}(\mathbf{X})$ with the asymptotically uniform modified discrepancy p value $[\hat{p}_{\text{dis}}(\mathbf{X})]$ of Section 4.2 based on the discrepancy $\hat{t}(\mathbf{X}, \theta) = t(\mathbf{X}, \theta) - \hat{v}_\theta(\theta)' \mathbf{I}_{\theta\theta}^{-1}(\theta) n^{-1} \mathbf{S}_\theta(\theta) = |\hat{Z}_{.9} - \mu| - |\mu - \hat{Z}_{.1}| - 2(\bar{X} - u)$, we can demonstrate that discrepancy p values, besides being conservative, cannot be interpreted as quantitative measures of model adequacy on any scale. Our argument rests on the following observation. For any p value p , define $\Phi^{-1}(1-p)$ to be the associated z score. Then, by our Theorems 3 and 4, in large samples, (a) \hat{p}_{dis} converges to a deterministic monotone increasing function of p_{dis} , (b) the associated z scores have a correlation of 1, and (c) if, in a given dataset, $p_{\text{dis}} = .05$, then, with probability approaching 1, \hat{p}_{dis} will equal .0004; that is they will differ by a factor of 125.

In his discussion, Stern concludes that discrepancy p values are "easily interpretable" as measures of the compatibility of model and data. We believe the previous paragraph refutes his conclusion for the following reason. In large samples, the two discrepancy p values are monotone functions of one another with perfectly correlated z scores. This implies that the two p values test for deviation from the model in precisely the same direction. Hence if discrepancy p values were quantitative measures of compatibility, then p_{dis} and \hat{p}_{dis} would have to be equal or nearly equal. But they differ by two orders of magnitude. In fact, it would appear that the only use to which the discrepancy (or other conservative) p values can be put is as an (often easily com-

puted) initial screen for model incompatibility, because if a conservative p value is small, then it will be even smaller when calibrated. Stern rightly emphasizes this use; however, he fails to sufficiently emphasize that a large conservative p value is not meaningful evidence of model adequacy and that further model checking using an asymptotically uniform p value is required.

We summarize by once again revisiting the weather analogy. As Stern correctly remarks, if we observe snow on the ground (i.e., a small plug-in, posterior predictive, or discrepancy p value), then we can decide to reject swimming (i.e., the null model) without the "expense" of checking the temperature (i.e., of computing an asymptotically uniform p value). But if there is no snow, then we cannot decide whether to go swimming (i.e., accept the null model as adequate) without checking the temperature.

On a different note, Stern makes the correct point at the end of his first section that the effectiveness of the posterior predictive approach depends crucially on the diagnostic statistics or discrepancies selected. He claims that we do not address this point in any depth. We disagree. We showed that (a) the posterior predictive p value is conservative unless the test statistic $t(\mathbf{X})$ has an asymptotic mean that does not depend on θ , (b) the discrepancy p value is conservative unless the discrepancy $t(\mathbf{X}; \theta)$ is uncorrelated with the score $S_\theta(\theta)$ for θ , and (c) if one uses the efficient score for the parameter ψ in an expanded (i.e., alternative) model as a discrepancy, then the resulting discrepancy p value is locally most powerful against that alternative. Indeed, an important motivation for our article was to satisfy the unmet need for a rigorous treatment of the dependence of the posterior predictive and discrepancy p values on the choice of test statistic or discrepancy.

Boos. Boos notes that if one adjusts α to α^* so that all tests have actual asymptotic level α , then all tests that we consider have the same asymptotic power. We made an analogous point in our discussion of the adjusted analytic p value ($p_{\text{adj,anal}}$) in Section 4.2. However, that such an adjustment could be made is of little use to the reader of a statistical report if in fact the adjustment was not carried out and only an unadjusted discrepancy or posterior predictive p value was reported; to perform the adjustment, the reader must calculate the variance function $\tau(\theta)$. Sometimes this will be impossible. For instance, in our example a, $\tau(\theta)$ depends on the correlation between the regressors, which may not be included in the report. Furthermore, to obtain $\tau(\theta)$ requires analytic calculations that are beyond the expertise of most readers. In contrast, $\tau(\theta)$ is always 1 for the partial posterior and conditional predictive p values, and no adjustment is necessary.

Boos is correct in saying that for purposes of model checking, somewhat liberal p values would be preferred by us to conservative p values. However, we wish to emphasize that the purpose of our article was to find asymptotically uniform p values; that is, p values that are (asymptotically) neither liberal nor conservative.

Marden. We only wish to add that in the null model considered by Marden, resampling from the empirical distribution with equal mass at the $y_i = x_i - \bar{x}_{\text{obs}}$ differs from resampling from the MLE (i.e., from the maximum empirical likelihood estimator of the density f) under the null mean 0 model (Owen 1988). It is the latter procedure that would appear to be the natural generalization of the parametric plug-in procedure discussed by both Bayarri and Berger and us. Nonetheless, both Marden's version and the empirical likelihood version of the plug-in yield asymptotically uniform p values (Beran 1986; Owen 1988).

Discrete Data. We would like to consider the examples of discretely distributed observations considered by Bayarri and Berger in their Section 4 from our asymptotic point of view. In their Example 5, they condition a sample of Bernoulli variables on their sum T , the null model being indexed by the unknown common success probability of the Bernoulli variables. As the sum is a sufficient statistic, the conditional law of the observations given T does not depend on the unknown parameter. In this extreme situation, our condition (13), which is motivated by the expectation that the conditional model behaves like a standard parametric model, fails. In fact, the partial posterior in this example is fixed and equal to the prior, taken to be uniform by Bayarri and Berger, and hence does not approach a normal distribution. To the remarks made by Bayarri and Berger it could be added that in this case the partial posterior and the prior predictive p values coincide, and hence both can be criticized under the general criticism given by Bayarri and Berger in their introduction. That is, the p value will depend on the prior. Even for large samples, this p value will not behave as a true (i.e., uniformly distributed) p value. Therefore, the results of our article will not support Bayarri and Berger's suggestion to use the partial posterior p value in this example. We believe that this is in agreement with the results presented by Bayarri and Berger, as none of the p values appear to yield sensible results. This is in contrast to Examples 3 and 4 considered by Bayarri and Berger in their Section 4. These examples come under the general setup of our results, and hence Bayarri and Berger's new p values behave asymptotically as true p values. Bayarri and Berger found that the distribution of the partial posterior p value (p_{ppost}) could be far from uniform when both the sample size was small and the parameter values were extreme. We expect, however, that the asymptotics would kick in at fairly small sample sizes for values of the parameters that are not

too extreme, as the asymptotics are based on the normal approximation to the binomial distribution.

Recommendations. In summary, if one wished to check for model adequacy using p values that are based either on a discrepancy measure $t(\mathbf{X}, \theta)$ or on an uncentered statistic $t(\mathbf{X})$, then we would recommend the following approach. One can begin by computing a discrepancy p -value or a plug-in p -value. The plug-in p -value can use either the MLE or the posterior mean of θ as the plug-in, as these two choices yield asymptotically equivalent p -values that are always less conservative than the posterior predictive p -value. If the discrepancy or plug-in p -value is small then one can reject the model. If not, then one cannot conclude that the data and the model are compatible in the "direction" tested by the statistic or discrepancy, and further model checking using an asymptotically uniform p value is required. As we discussed in Section 4, there are essentially three approaches to obtaining an asymptotically uniform p value. Which of the approaches one should adopt may depend on whether one is comfortable writing computer code, comfortable with mathematical analysis, or comfortable with neither. If comfortable with writing computer code, then one can obtain the partial predictive p value based on $t(\mathbf{X})$ using the methods proposed by Bayarri and Berger. If comfortable with analysis, then one can calculate one of the adjusted analytical p values $p_{\text{plug,adj,anal}} = 1 - \Phi[z_{1-p_{\text{plug}}}/\tau_{\text{plug}}(\theta)]$, $p_{\text{post,adj,anal}} = 1 - \Phi[z_{1-p_{\text{post}}}/\tau_{\text{post}}(\theta)]$ or $p_{\text{dis,adj,anal}} = 1 - \Phi[z_{1-p_{\text{dis}}}/\tau_{\text{dis}}(\theta)]$, the modified statistic $\hat{t}(\mathbf{X}) = t(\mathbf{X}) - \nu_n(\hat{\theta})$, the asymptotically pivotal statistic $\hat{t}(\mathbf{X})/c(\hat{\theta})$, or the modified discrepancy $\hat{t}(\mathbf{X}, \theta) = t(\mathbf{X}, \theta) - \dot{\nu}_\theta(\theta)' \mathbf{i}_{\theta\theta}^{-1}(\theta) n^{-1} \mathbf{S}_\theta(\theta)$. The plug-in, and posterior predictive p value based on $\hat{t}(\mathbf{X})$ and discrepancy p value based on $\hat{t}(\mathbf{X}, \theta)$ are asymptotically uniform, because $\hat{t}(\mathbf{X})$ has a constant asymptotic mean and because $\hat{t}(\mathbf{X})$ and $\hat{t}(\mathbf{X}, \theta)$ are uncorrelated with the score $\mathbf{S}_\theta(\theta)$ for θ . With complex models, the calculation of $\tau(\theta)$, $c(\hat{\theta})$, $\nu_n(\theta)$, $\dot{\nu}_\theta(\theta)$, $\mathbf{i}_{\theta\theta}(\theta)$ and/or $\mathbf{S}_\theta(\theta)$ may require the numerical assessment of high-dimensional integrals, which may limit the usefulness of this approach. Finally, when both coding and analytical calculation are too difficult, one can calculate the adjusted p values $p_{\text{plug,adj}}$, $p_{\text{post,adj}}$ or $p_{\text{dis,adj}}$ using the "double bootstrap." Double-bootstrap calculations are computationally intensive, although the computational burden may potentially be lessened by bootstrap recycling.

When all three approaches can be implemented, the choice among them should be based on their (as yet unstudied) second-order asymptotic and small-sample behavior.

ADDITIONAL REFERENCES

- Beran, R. (1986), "Simulated Power Functions," *The Annals of Statistics*, 14, 151-173.
 Owen, A. B. (1988), "Empirical Likelihood Ratio Confidence Interval for a Single Functional," *Biometrika*, 75, 237-249.