

Hal S. STERN

The question of how to assess the fit of a composite null model has attracted much attention in recent years, particularly among Bayesian practitioners with alternative ideas about how to account for the parameters defining the composite model. These two complementary articles offer new ideas and new results: Bayarri and Berger (BB) offer two new methods for computing  $p$  values for judging the fit of the null model and compare the performance of these methods with that of existing methods on a number of examples. Robins, van der Vaart, and Ventura (RVV) provide results concerning the asymptotic sampling distributions of the  $p$  value proposals of BB and others. This comment is focused on model assessment using the posterior predictive distribution, as proposed by Gelman, Meng, and Stern (GMS, 1996), Guttman (1967), and Rubin (1981, 1984), and addresses the criticisms of the posterior predictive  $p$  value offered by BB and RVV. I have found posterior predictive model checks to be of great help in my applied work and will continue to use them even after reading these two articles. Herein I explain why.

#### 1. WHAT IS POSTERIOR PREDICTIVE MODEL ASSESSMENT?

A first step is to review the posterior predictive approach to model assessment using the notation of the two articles and review the motivation for the approach. Let  $f(\mathbf{x}; \theta)$  be the model for the data  $\mathbf{X}$  being assessed, or in the words used by BB and RVV, the null model. The posterior predictive approach to model assessment compares the observed data  $\mathbf{x}_{\text{obs}}$  to replicate data that could be observed if the data collection effort or experiment were repeated. The appropriate reference distribution is then the posterior predictive distribution,  $m_{\text{post}}(\mathbf{x}|\mathbf{x}_{\text{obs}})$ , which is defined in BB's (7) and RVV's Table 1. Of course, comparing distributions of possibly large datasets  $\mathbf{X}$  is unwieldy, so instead we are likely to compare the observed value of one or more statistics  $T = t(\mathbf{X})$  to their reference distributions; that is, their distributions under  $m_{\text{post}}(\cdot|\mathbf{x}_{\text{obs}})$ . GMS introduced the use of discrepancies  $t(\mathbf{x}; \theta)$  as a possible alternative to formal test statistics; in that case, the posterior distribution of  $t(\mathbf{x}_{\text{obs}}; \theta)$  (recall that  $\theta$  is a random variable under the Bayesian approach) is compared to the distribution of  $t(\mathbf{x}; \theta)$  under  $m_{\text{dis}}(\mathbf{x}, \theta|\mathbf{x}_{\text{obs}})$ , which is defined in RVV's Table 1. These comparisons may be summarized by calculating tail probabilities or  $p$  values.

I would like to make five important points about posterior predictive model assessment as described in the previous paragraph.

1. The posterior predictive distribution is a legitimate probability distribution; it is, as the name implies, the predictive distribution that a Bayesian naturally would use to talk about future observations from the same process that generated the data  $\mathbf{x}_{\text{obs}}$  (see, e.g., Jeffreys 1961, pp. 142-143). BB refer to a "double use" of the data in their discussion of the posterior predictive approach. It seems as if they are suggesting an inappropriate use of the data, as occurs, for example, when the same data are used to identify a prior distribution and then again in the likelihood function. Put bluntly, there is nothing inappropriate in the posterior predictive approach; BB might prefer to use the posterior predictive distribution to calculate the probability of another event or might prefer to use another distribution in place of the posterior predictive distribution, but this does not make the posterior predictive approach invalid.

2. The posterior predictive approach is especially useful when the model at hand is a serious attempt to describe the scientific phenomenon of interest. The goal then is to determine if there are violations serious enough to warrant developing a new model or perhaps modifying the present one. This is slightly different than the perspective taken by BB, whose "primary focus is on model checking, at initial, exploratory stages of the statistical analysis" (BB, Sec. 1.3). In this regard the examples discussed by BB seem overly simplistic; for example, would anyone seriously contemplate using the iid  $N(0, \sigma^2)$  model rather than the bigger, but still easy to use,  $N(\mu, \sigma^2)$  model? Checking the hypothesis that  $\mu = 0$  would be straightforward in the second model.

3. If a data analysis is carried out using simulation-based inference, as is common for Bayesian data analyses these days, then posterior predictive model assessment is extremely easy to perform. Simulations from the posterior distribution of  $\theta$  are available and simulations from  $f(\mathbf{x}; \theta)$  are typically straightforward (especially for models constructed hierarchically). Once again, this contrasts with BB's proposals.

4. Posterior predictive model checks, as they are sometimes called, can be displayed graphically and provide a great deal of information in the spirit of residual plots in a regression analysis. The posterior predictive  $p$  value is a useful summary, but it is not the fundamental idea of posterior predictive model checking. My co-authors and I have no doubt contributed to the overemphasis on  $p$  values (see, e.g., GMS), but it makes more sense to think of assessing the fit of a null model as an exercise in diagnostics rather than as an exercise in decision theory.

Hal S. Stern is Professor, Department of Statistics, Iowa State University, Ames, IA 50011. The author thanks Donald Rubin and Andrew Gelman for helpful discussions. This work was partially supported by National Institutes of Health grant CA78169.

5. The effectiveness of the posterior predictive approach depends crucially on the diagnostic statistics or discrepancies selected. Neither BB or RVV address this point in any depth.

## 2. *P* VALUES, PROBABILITIES, AND THE UNIFORM DISTRIBUTION

Most of the criticisms of posterior predictive methods in these two articles stem from results demonstrating that posterior predictive *p* values, when viewed as a function of  $\mathbf{x}_{\text{obs}}$ , do not have a uniform sampling distribution under the null hypothesis. They are often conservative (more closely concentrated about .5) in finite samples (BB) and asymptotically (RVV). This conservatism had been noted in previous discussions of the posterior predictive approach (e.g., by GMS and in subsequent discussion in Rubin 1996). Both RVV and BB argue that frequentists and Bayesians alike should demand an asymptotic uniform null sampling distribution as a requirement for any *p* value. In my opinion, this misses a crucial point. *P* values are the probability that an event occurs—specifically, the event that  $t(\mathbf{X})$  greater than or equal to  $t(\mathbf{x}_{\text{obs}})$  is observed. (To be even more precise, posterior predictive *p* values are the conditional probability of this event given the observed data.) As a probability, the posterior predictive *p* value provides useful model checking information with or without an asymptotic null sampling distribution. Extreme probabilities clearly indicate that the observed data are inconsistent with the model.

The previous paragraph is not an attempt to minimize the technical results obtained by BB and RVV. The asymptotic null uniform sampling distribution achieved by BB's partial posterior predictive and conditional predictive *p* values can definitely be a positive aspect of those model assessment techniques in some situations (e.g., with large samples). In addition, RVV provide a service to all by showing how the nonuniform *p* values that result from the plug-in and posterior predictive approaches to model assessment may be modified to achieve the asymptotic null uniform sampling distribution. Posterior predictive checks based on discrepancy measures admit a fix requiring less work than posterior predictive checks based on statistics, so this might be one plausible approach to use. It remains to be seen whether the advantage of having an asymptotic null uniform sampling distribution compensates for the additional analytical or computational work required to compute BB's *p* values.

The focus on model assessment through *p* values in these two articles risks sending the false message that tests are the only way to validate a model. Residual plots in regression analyses are a prime example of a diagnostic approach that does not rely on or need *p* values to inform us about the fit of a hypothesized model. In that context, concerns about a model (e.g., whether a quadratic term is required to improve the fit of a linear model) are often first raised by a diagnostic plot rather than by a formal test. This is the spirit of the posterior predictive model check.

## 3. POSTERIOR PREDICTIVE MODEL CHECKING EXAMPLE

In this section I briefly review an instance in which posterior predictive model checking made a difference in my own work. Glickman and Stern (1998) presented a Gaussian state-space model for analyzing results of professional football games. But this model was not the first model fit to these data; an earlier model was fit in Glickman's (1993) doctoral dissertation. We begin by considering the earlier model.

The state-space model takes football outcomes at a given time to be normally distributed with a mean depending on team rating parameters at the given time and a parameter measuring the advantage of the team playing at home. The team rating parameters are assumed to vary over time (according to a Gaussian model), but the home-field advantage is assumed to be the same for every team and every year. The assumptions about the homefield advantage seem quite restrictive, but there is a great deal of "conventional wisdom" that support them. To assess the validity of the assumptions, we constructed "site-effect" residuals as the difference between the observed game outcomes and the outcomes predicted by the team rating parameters. Note that these residuals adjust for quality of teams but do not adjust for the game site. We computed the average residual for each of the 28 teams and defined a model checking discrepancy as the difference between the largest average residual and the smallest average residual. Note that this is a discrepancy (rather than a test statistic), because the residuals depend on the model parameters.

Figure 1(a) presents a scatterplot showing the joint distribution of the observed discrepancy  $t(\mathbf{x}_{\text{obs}}; \theta)$  and the simulated replicate discrepancy  $t(\mathbf{x}; \theta)$ . The scatterplot clearly shows that the observed values of the discrepancy are generally greater than the values of the discrepancy for data simulated under the model. This indicates that the average site-effect residuals varied significantly more from team to team than was expected under the model (the tail probability or *p* value is .05). This failure of the model suggested a modification that resulted in the final model presented by

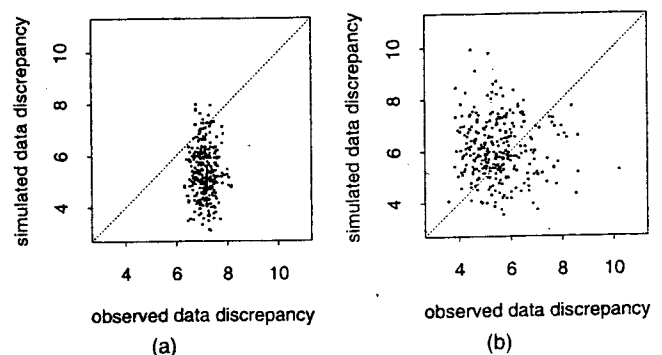


Figure 1. Estimated Bivariate Distribution for Site-Effect Discrepancy Under Two Models. The scatterplots show the empirical joint posterior distributions of the discrepancy evaluated at the observed data and the discrepancy evaluated at simulated replicate data under two models. (a) The model including a single home-field advantage parameter; (b) the model including multiple home-field advantage parameters. There are 300 simulations from the posterior predictive distribution in each plot.

Glickman and Stern (1998) in which each team has a separate home-field advantage parameter and these parameters are assumed to come from a common normal population distribution. Figure 1(b) presents the scatterplot for the new model; the fit seems much better ( $p$  value equal to .68).

This is but a single illustration. The number of examples in which posterior predictive model checks have proven useful continues to grow. Examples have been given by Gelman, Carlin, Stern, and Rubin (1995), Gelman, Goegbeur, Tuerlinckx, and Van Mechelen (2000), GMS (1996), Gelman, Van Mechelen, Verbecke, Heitjan, and Meulders (2000), Rubin (1984), and Rubin and Wu (1997).

#### 4. CHOICE OF EXAMPLES IN BB

BB's proposals are more difficult to compute than the plug-in  $p$  value and the posterior predictive  $p$  value. A consequence is that the comparisons in their article are based on relatively simple examples: an iid  $N(0, \sigma^2)$  model, an exponential model, a linear model, and hypotheses concerning a  $2 \times 2$  table.

I consider the first of these in more detail. There is no disputing the mathematical results provided; using  $t(\mathbf{X}) = \bar{X}$  leads to conservative  $p$  values under the posterior predictive approach: There is, however, surely room to dispute the relevance of this result. The posterior predictive  $p$  value is, as stated earlier, the probability of an event. Here the knowledge built into BB's conditional predictive distribution (conditioning on the sufficient statistic for  $\sigma^2$ ) could easily be integrated into a better choice of test statistic; in fact, BB mention this near the end of Section 2.1. Yet at the very end of Section 2.1, BB go on to say that "in more complex problems, it may be quite difficult to find 'appropriate' departure statistics for use with the posterior predictive or plug-in  $p$  values." I disagree. My own experience in complex models is that I have been able to see how to construct intuitive (often residual-like) discrepancy measures for posterior predictive checks that provide great insight into the performance of a model (as in the Glickman and Stern 1998 example described earlier). On the other hand, it is not at all obvious in complex problems how to proceed with the calculations required to use the BB proposals.

In addition, I disagree with BB's interpretation of the results in their final example (Example 5), where a sufficient statistic is used as a test measure. There, using the sample proportion of successes to assess the fit of a simple iid Bernoulli model yields extreme partial posterior predictive  $p$  values if the sample proportion is very large or very small. BB find this intuitive because, for example, large-sample proportions imply large  $\theta$ , which is "unusual" under the uniform prior distribution. The "full" posterior predictive  $p$  value in this case tends to concentrate closely around .5, suggesting no problems with the model. This difference brings to mind GMS's comparison of the Box (1980) prior predictive  $p$  value and the posterior predictive  $p$  value. Do we really want to reject the null Bernoulli model because

of a small partial posterior predictive  $p$  value that is a result of a flat prior chosen for convenience? The posterior predictive answer in this case says exactly what needs to be said—there is no basis for rejecting the posterior distribution of  $\theta$  under this model based on the test measure in question. In fact, this is the main point of Rubin's (1984) discussion of posterior predictive model checks.

#### 5. CONCLUSIONS

A primary goal of the statistics community today, especially the Bayesian community, must be to provide relevant practical methods for assessing the fit of complex models, including in scientifically interesting situations with irregular asymptotic posterior distributions (e.g., Rubin and Wu 1997). The ease of computation, easy interpretability, and ability to provide graphical evidence (as well as  $p$  values) for suggesting weaknesses of a hypothesized model make posterior predictive methods a natural choice. The failure of posterior predictive  $p$  values to obtain a uniform sampling distribution under the null can make calibration difficult in borderline cases, but because it is the probability of an event the posterior predictive  $p$  value remains a fairly natural concept for data analysts to consider.

RVV talk often in their article about the difficulty in deciding whether to ski or swim when told the temperature is 30 degrees but no scale is provided. A simple diagnostic like looking out the window could easily invalidate an incorrect model (one probably would not ski if there was no snow on the ground) without the "expense" of checking the temperature at all. This is even more true in the case of checking on a statistical model where a straightforward posterior predictive check might invalidate a model (and suggest a better one) without requiring the difficult computations that would provide a  $p$  value with a uniform null sampling distribution.

I firmly believe that posterior predictive  $p$  values remain a valuable diagnostic tool when faced with assessing the fit of a serious, plausible model in a scientific application.

#### ADDITIONAL REFERENCES

- Gelman, A., Goegbeur, 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., Van Mechelen, I., Verbecke, G., Heitjan, D. F., and Meulders, M. (2000), "Bayesian Model Checking for Missing and Latent Data Problems Using Posterior Predictive Simulations," technical report, Columbia University, Dept. of Statistics.
- Glickman, M. E. (1993), "Paired Comparison Models With Time-Varying Parameters," unpublished Ph.D. dissertation, Harvard University, Dept. of Statistics.
- Glickman, M. E., and Stern, H. S. (1998), "A State-Space Model for National Football League (NFL) Scores," *Journal of the American Statistical Association*, 93, 25–35.
- Jeffreys, H. (1961), *Theory of Probability* (3rd ed.), London: Oxford University Press.
- Rubin, D. B. (1981), "Estimation in Parallel Randomized Experiments," *Journal of Educational Statistics*, 6, 377–401.
- Rubin, D. B., and Wu, Y. N. (1997), "Modeling Schizophrenic Behavior Using General Mixture Components," *Biometrics*, 53, 243–261.

1. INTRODUCTION

These two interesting and significant articles are concerned with an important topic in statistics. Virtually all statistical analyses are based on assumptions of some sort. The frequentist assumes a model  $\{f_\theta : \theta \in \Omega\}$  for the generation of the observed data  $x$ , and the Bayesian adds to this a prior probability distribution  $\pi$  for the unknown true value of  $\theta \in \Omega$ . If inferences drawn from these ingredients are to have any validity in a particular application, then it seems clear that these assumptions must be checked for their reasonableness in light of the data obtained; this process is known as model checking.

Typically, a somewhat informal approach is taken toward model checking. As these articles demonstrate, however, this is not appropriate, because the checking process can be flawed. So one aspect of my discussion is to make the point that model checking deserves a more formal treatment, involving examining the questions as to what model checking should be and how it should be approached. In discussing this point, I also make comments about the articles by Bayarri and Berger (BB) and Robins, van der Vaart, and Ventura (RVV) where appropriate. I confine my comments to the Bayesian context, although I believe that some are also appropriate to the frequentist formulation.

2. MODEL CHECKING VERSUS INFERENCE

My first comment is that it seems necessary to distinguish sharply between model checking and inference about model unknowns. Perhaps it is useful to define *formal inference* as the process of applying a theory to a model and data combination to make statements about an unknown of a model or of an unknown of a related model (e.g., a model with the same parameter space and true value of the parameter, as in a prediction problem). In formal inference there is a set of possible values for an unknown, as prescribed by the model, and one is required to choose among these or decide on the plausibility of particular choices, based on the data. In this framework, the model is assumed to be true.

Formal inference is different than model checking, as there is only one possible model. One approach to model checking has been to enlarge the situation to include alternative models and then apply formal inference procedures. This is somewhat unsuitable, however, as it simply leaves another supermodel to be checked if one's inferences are to be validated.

The reason for distinguishing clearly between model checking and formal inference is because I believe that different criteria must be developed to characterize suit-

able model checking procedures. For example, some standard criteria used in statistics to determine appropriate inferences, such as efficiency and maximizing power, can be misleading if strictly applied in the model checking framework. The erroneous "double use of the data" discussed in these articles is an example of applying conventional ideas about efficient use of data to model checking.

3. WHAT ARE SUITABLE CRITERIA FOR MODEL CHECKING PROCEDURES TO SATISFY?

Without attempting to delineate the appropriate criteria for developing model checking procedures, at least one is regularly used in science. In particular, if we are considering the validity of a proposed theory or model, then we do not simply accept the model based on its plausibility or its fit to the data used to develop it, but require that the *fitted* model predict completely *new* data. As we see the fitted model perform adequately with respect to prediction in many different, unrelated contexts, we become more and more confident that the model is indeed appropriate. This seems to me to be a scientific principle whose application in model checking transcends in importance the various formal inference statistical criteria that we might try to use to determine an appropriate procedure for this problem.

In the statistical context we have only the data  $x$  at hand, so the aforementioned principle leads to some kind of split,  $(T(x), U(x)) \leftrightarrow x$ , of the data, where we fit the model using the value  $U(x)$  somehow and then use this fitted model to predict the value of  $T$  and compare this with the observed  $T(x)$ . In Bayesian contexts, fitting the model means constructing the posterior predictive distribution of  $T$  as given by its density  $m_T(\cdot|U(x))$ . The 1-1 nature of the split seems necessary if we are to make full use of all the information in the data about the validity of the model.

As there are typically many different possible choices for the splitting functions  $T$  and  $U$ , we need criteria to choose among them. First of all, we need to be able to fit the model. This implies that the marginal model of  $U$  be indexed by  $\theta$ . Second, we want to make sure that the data  $T(x)$  that we are predicting is truly unrelated, at least as far as the model prescribes, to the data  $U(x)$  used to fit the model. This entails that  $T$  and  $U$  be statistically independent for each  $\theta \in \Omega$ . Essentially, this prescription for avoiding double use of the data was given by Evans (1997). The  $T$  and  $U$  functions prescribed in BB are not required to satisfy this property and so, at least in the finite-sample context, do not avoid the double use of the data. Probably we want the  $T$  and  $U$  functions to satisfy further criteria, such as the

Michael Evans is Professor, Department of Statistics, University of Toronto, Toronto M5S 3G3, Canada.

marginal model for  $T$  also being indexed by  $\theta$ , so that we can perform a full predictive test of the model, but I do not pursue this issue further here. For some situations we may not be able to find such functions except in some asymptotic sense, but at least we should be clear about what avoiding double use of the data really means and strive to attain this ideal when possible. As I explain, there are many contexts where this is entirely feasible. I found both of the articles somewhat vague on this issue. Furthermore, BB seem to advocate a somewhat informal approach to the choice of  $T$  (e.g.,  $T$  is some discrepancy measure), but as they demonstrate, taking such an informal approach to model checking is not a good idea.

Suppose then that we have argued for, or selected in some fashion, a particular split  $(T, U)$ . BB then suggest that the tail probability

$$M_T(t \geq T(x)|U(x)) \tag{1}$$

be computed, where  $M_T(\cdot|U(x))$  is the measure induced by  $m_T(\cdot|U(x))$ , to assess the validity of the model. In general this will be appropriate only when  $T$  is real valued and when the occurrence of  $T(x)$  in low-probability regions for  $M_T(\cdot|U(x))$  corresponds to the right tail of this distribution. If  $T(x)$  occurs in a low-probability region of the left tail or near an extremely low antinode, then this also is evidence against the model. For example, if  $M_T(\cdot|U(x))$  is the chi-squared (50) probability measure and one observes  $T(x) = 10$ , then this is strong evidence against the model being correct, as there is virtually no chance of this occurring when the model is correct [the distribution function of the chi-squared (50) gives the value .0000 at  $T(x) = 10$  and 20] irrespective of the interpretation of the discrepancy measure. The natural way to correct for this defect in (1) is to compute

$$M_T(m_T(t|U(x)) \geq m_T(T(x)|U(x))|U(x)); \tag{2}$$

that is, the posterior probability of observing a value of  $T$  with posterior density no smaller than at the observed value, with values of (2) near 1 being interpreted as evidence that the model is not correct. The quantity (2) also suffers from a defect, at least in continuous models, as it is not invariant under smooth relabellings of  $T$  and in some situations, any value between 0 and 1 can be obtained via an appropriate relabelling. A solution to this problem, as proposed by Evans (1997), is to compute the *observed cross-validated relative surprise* (OCVRS), defined as

$$M_T \left( \frac{m_T(t|U(x))}{m_T(t)} \geq \frac{m_T(T(x)|U(x))}{m_T(T(x))} \middle| U(x) \right), \tag{3}$$

as this is invariant under transformations. Here the ratio  $m_T(t|U(x))/m_T(t)$  is the relative change in belief about the value  $t$  from prior to observing  $U(x)$  to after having done so, and (3) is the posterior probability of having a change in belief no smaller than that observed at  $T(x)$ . If (3) is close to 1, then this provides evidence that the data indicate that  $T(x)$  is a surprising value for  $T$ , and so the

model is rejected. Examples of the application of (3) and formal inferences for model unknowns via relative surprise were discussed by Evans (1997).

Still the question remains as to how we should choose  $T$  and  $U$ . In sampling situations [i.e.,  $x = (x_1, \dots, x_n)$  and the  $x_i$  are iid], there are natural choices. One can simply choose  $[T(x), U(x)]$  as some split of the sample determined prior to observing the data; for example, take  $T(x)$  to be the first  $k$  sample values and  $U(x)$  to be the last  $n - k$ . Clearly we want  $n - k$  large enough so that the fitted model  $m_T(\cdot|U(x))$  is reasonably stable, but we also want to choose  $k$  large enough so that the computation of (3) is a rigorous test of the model. The more data that one is required to predict, the more rigorous the test is. For example, choosing  $k = 1$  or 2 does not seem very rigorous. A reasonable ratio seems to be  $k/n = 25\%$ , and of course the effectiveness of this will also depend on the size of  $n$ ; that is, there is no cure for too little data. With such choices, we have that (3) completely avoids double use of the data and has the important property of being invariant. Further, it is more strongly data driven than traditional posterior approaches, as it is based on how beliefs change from a priori to a posteriori rather than just on properties of the posterior distribution itself.

One objection that can be raised against (3) as I have implemented it in the iid sampling case is that, even when  $k$  is specified there are still  $\binom{n}{k}$  possible choices, and presumably there is information about the fit of the model in all of these. A logical consequence of this line of reasoning is that we should really look at the distribution of (3) under all possible splits. The form of this distribution gives us the full information about the lack of fit of the model based on such a splitting of the data. Of course, there is the combinatorial problem of evaluating this distribution exactly, but in fact it is simple to generate large samples from this distribution by randomly generating splits, evaluating (3) for each of these, and looking at the empirical distribution. Furthermore, we can calibrate what this distribution is telling us about the lack of fit by simulating some datasets from the model and seeing how the distributions vary. I consider an example.

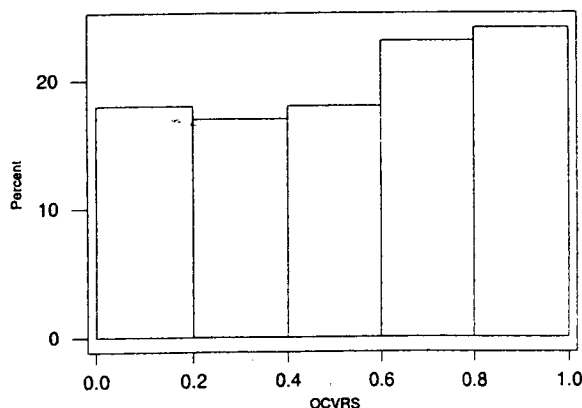


Figure 1. Histogram of OCVRS Values from 1,000 Random Splits When the Model in Example 1 is Correct.

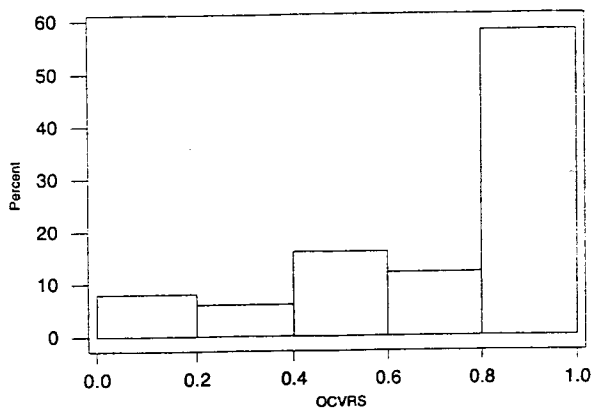


Figure 2. Histogram of OCVRs Values From 1,000 Random Splits When the Model in Example 1 is Incorrect.

*Example 1* (Liu 1999). Suppose that  $x = (x_1, \dots, x_{100})$  is supposed to be a sample from a  $N(\mu, \sigma^2)$  distribution with  $\mu|\sigma \sim N(\mu_0, \tau\sigma^2)$  and  $\sigma^{-2} \sim \beta$  chi-squared (2) and the hyperparameters  $\tau$  and  $\beta$  are chosen to be large so that the prior on the parameter  $(\mu, \sigma^{-2})$  is diffuse. Figure 1 presents a histogram of the results of evaluating (3) for a sample of 1,000 different splits with  $k = 25$  when  $x$  was actually generated from a normal distribution. Figure 2 presents a histogram of the results of evaluating (3) for a sample of 1,000 different splits with  $k = 25$  when  $x$  was actually generated from a Student (2) distribution. We can see from Figure 1 that for some splits, the model looks wrong even though it is correct, as about 25% have (3) greater than .8, but that overall the distribution supports the model. In Figure 2, however, the distribution is much more concentrated about 1, with about 60% of the splits having (3) greater than .8, and this gives clear evidence of the incorrectness of the model.

The model checking procedures that I am advocating are perhaps somewhat unconventional, although cross-validation seems quite natural and has a long history in statistics (see e.g., Geisser 1975; Stone 1974). I cannot claim to have shown that (3) is the right approach, but it does satisfy some natural requirements for model checking, such as being based on the ability of the model to predict new data and avoiding double use of the data when assessing fit and the invariance under relabellings.

I note that one reason for considering that double use of the data is bad is the belief that an overly complicated model will accurately predict the data through overfitting. Approaches such as (3) based on splitting will naturally avoid this problem, because the model will overfit to the data  $U(x)$  and then provide a poor prediction of  $T(x)$ . It is essential that  $T$  and  $U$  have no statistical relationship if this is to hold. Of course, underfitting will also be detected.

#### 4. THE UNIFORM DISTRIBUTION OF THE $P$ VALUE AS A REQUIREMENT FOR A MODEL CHECKING PROCEDURE

RVV assert that under sampling from the model, the  $p$  value should be uniformly distributed. In part, the appro-

priateness of this criterion depends on which model is being used to assess this. For example, Box (1980) proposed using

$$M(m(X) \leq m(x)), \quad (4)$$

where  $m$  is the prior predictive density of the data and  $M$  is the prior predictive measure, as the  $p$  value to use for model checking. Given that the prior  $\pi$  is proper, the Bayesian model says that a priori the appropriate distribution to use to predict a future  $x$  is  $M$ . When  $x \sim M$ , then in fact (4) does have the appropriate uniformity property, at least in the continuous case. That it may not have this property under sampling from  $f_\theta$  when  $\theta$  is true seems irrelevant to me. It is clear also that (4) does not make double use of the data. It is interesting then to contemplate why we would not use (4) for model checking in the case of proper priors. One strike against (4), at least in the continuous case, is that it suffers from a lack of invariance under relabellings of  $x$ . But what about discrete contexts? The fact that it is not defined in the case of improper priors is only a criticism if we accept the use of improper priors and that seems highly controversial to me. Consider, however, the following example.

*Example 2.* Suppose that  $x = (x_1, \dots, x_n)$  is a sample from a Bernoulli( $\theta$ ) distribution with  $\theta \sim U(0, 1)$ . Then we have that  $m(x) = \binom{n}{\bar{x}}(n+1)^{-1}$  and this is a U-shaped function of  $\bar{x}$ . It is easy to see that values of  $\bar{x}$  near the center of its range will lead to small values of (4) even when the model is correct, and this seems unnatural.

In effect, the problem with (4) seems to be that it does not assess the model by constructing the predictive based on observed data and then assessing its predictive power on new data, as discussed in Section 3; that is, it violates a scientific principle of how I feel model checking should be carried out. In essence, the model must be given a chance to construct its predictions using data before the predictive power of the model is assessed. Others, such as Guttman (1967), have recognized this and replaced the prior predictive by the posterior predictive. These measures also will have the uniformity properties when the appropriate model is used to assess this. But in simulations reported by Evans (1997) in the context of the Bernoulli model of Example 2, it is seen that they can suffer from never rejecting the model even with extreme datasets. This is due to the violation of the model checking principle espoused in Section 3 through double use of the data. In other words, I do not believe that the lack of uniformity under sampling from  $f_\theta$  is needed to reject these measures. The procedures based on (3) correct this defect through data splitting.

Cases in which one might want the uniform property under sampling from the true  $f_\theta$  arise, however, whenever one uses improper priors or limiting inferences as priors become increasingly diffuse. It is perhaps natural to want inferences to have appropriate repeated sampling properties in this case. As this is a rather common occurrence in practice, the results of RVV are certainly highly relevant. On the other hand, I am not sure that it is a good idea to assess the validity of a procedure on the basis of how it performs

when improper priors are used. First, one should determine what methods are appropriate in the ideal situation of a Bayesian formulation with a proper prior. What one does when feeling compelled to use an improper prior, or even discard a prior altogether, should be guided by the ideal context, where things can be expected to behave sensibly, and not the other way around.

## ADDITIONAL REFERENCES

- Geisser, S. (1975), "The Predictive Sample Reuse Method With Applications," *Journal of the American Statistical Association*, 70, 320–328.
- Liu, S. (1999), "An Analysis of Some Inference Procedures Derived via Relative Surprise," Ph.D. thesis, University of Toronto, Dept. of Statistics.
- Stone, M. (1974), "Cross-Validatory Choice and Assessment of Statistical Predictions," *Journal of the Royal Statistical Society, Ser. B*, 39, 44–47.

## Comment

Dennis D. BOOS

These are two very interesting articles on  $p$  values for composite null hypotheses. The authors have given us new ideas and some hard mathematics. I congratulate them for these important contributions to the definition and understanding of  $p$  values.

The emphasis of the articles is on Bayesian  $p$  values, those that involve priors (typically noninformative ones), and on checking model adequacy. Because of this, the authors are not concerned about validity in the usual frequentist sense; that is, that decision rules like "reject the null hypothesis if  $p \leq \alpha$ " have probabilities less than or equal to  $\alpha$ . In fact, given a preference, the authors would prefer to have liberal rather than conservative  $p$  values. I point this out to alert the reader that this viewpoint is in contrast to much of standard hypothesis testing where having strictly valid alpha levels is considered important.

Bayarri and Berger (BB) make this point clear at the end of their Section 1.1. I feel, however, that BB are a bit too pessimistic when they claim that  $p_{\text{sup}}$  "is of rather limited usefulness, because the supremum is often too large to provide useful criticism of the model." First, I cite the recent article by Freidlin and Gastwirth (1999) that presents a number of examples that illustrate good power for a modification of  $p_{\text{sup}}$ . Moreover, I point out that the use of  $p_{\text{sim}}$  as a test statistic coupled with  $p_{\text{sup}}$  as a  $p$  value leads to test procedures with good power properties (see Berger 1994).

An interesting result from Robins, van der Vaart, and Ventura (RVV) is that  $p_{\text{plug}}$  is asymptotically uniform or conservative. This intuitively makes sense, because estimating  $\theta$  from the data makes  $f(x; \hat{\theta})$  resemble the data more closely than the true model  $f(x; \theta_0)$  does. Of course, this result is asymptotic and may not meet frequentist standards in small samples. BB's Example 1 suggests that  $p_{\text{plug}}$  will also be conservative in small samples. I would like to point out the sensitivity in Example 1 to the choice of estimator,  $\hat{\sigma}^2 = s^2 + \bar{x}^2$ . If instead, we let

$\hat{\sigma}^2 = s^2$  (a possibility that the authors briefly mention), then

$$p_{\text{plug}} = 2 \left[ 1 - \Phi \left( \frac{\sqrt{n} |\bar{x}_{\text{obs}}|}{s_{\text{obs}}} \right) \right],$$

and under the null hypothesis, this  $p_{\text{plug}}$  is always liberal, because

$$P(p_{\text{plug}} \leq \hat{\alpha}) = P(|t_{n-1}| \geq \sqrt{(n-1)/n} \Phi^{-1}(1-\hat{\alpha}/2)) \geq \hat{\alpha}.$$

Of course, this version of  $p_{\text{plug}}$  is also asymptotically uniform.

BB's discrete examples are not as encouraging as their continuous examples. In particular, the extremely liberal behavior of  $p_{\text{ppost}}$  for large  $\theta$  in Examples 3 and 4 is disturbing. As BB note,  $p_{\text{ppost}}$  performs better for more sensible  $T$ 's, but the point of the examples is supposed to be that  $p_{\text{ppost}}$  performs well even without the best  $T$ 's. Moreover, when  $\theta$  is large, sample sizes may have to be very large before asymptotic approximations are useful. Thus, for contingency tables, I am not convinced that  $p_{\text{ppost}}$  is an all-purpose solution.

On a technical note, RVV define quantities ARP and ARE based on Pitman alternative power calculations. Because the authors do not force the tests to have asymptotic level  $\alpha$  (as is the case with Pitman ARE), it appears that the quantities ARP and ARE merely reflect the asymptotic level differences and provide no new insight into comparison of the different  $p$  values. For example, if one adjusts  $\alpha$  to  $\alpha^*$  so that the tests have actual asymptotic level—that is, find  $\alpha^*$  such that

$$\alpha = 1 - \Phi[z_{1-\alpha^*} \tau^{-1}(\theta)],$$

and determine  $k_{\text{ppost}}$  so that  $p_{\text{ppost}}$  has asymptotic power  $\beta$ —then all of the tests have asymptotic power  $\beta$ .

The most practically useful of the new  $p$  values seems to be  $p_{\text{ppost}}$  because of the computational schemes that BB suggest in their Section 2.3. I envision any of the new  $p$  values being the most useful in complex applications where analytic computations would be virtually impossible. Thus, MCMC approaches to calculating  $p_{\text{ppost}}$  suggest that it has the most potential for actual use.

## ADDITIONAL REFERENCES

- Berger, R. L. (1994), Letter to the Editor on "Exact Power of Conditional and Unconditional Tests: Going Beyond the  $2 \times 2$  Table" by Mehta and Hilton, *The American Statistician*, 48, 175.
- Freidlin, B., and Gastwirth, J. L. (1999), "Unconditional Versions of Several Tests Commonly Used in the Analysis of Contingency Tables," *Biometrics*, 55, 264-267.

## Comment

John I. MARDEN

### 1. INTRODUCTION

What is great about being asked to comment is that one can first thank the distinguished authors for their fine work, then start blustering and harumphing about how wrong-headed they are. I certainly do the former. Unfortunately, I must pass on the latter, because I find these two articles extremely right-headed.

In the next section I mention some controversial points in the articles, with which I agree. In Section 3 I try to take apply the spirit of the proposed  $p$  values to a simple bootstrap situation. I do a little harumphing in Section 4.

### 2. AGREEMENT

*Third Way.* I imagine that pure Bayesians would be uncomfortable with the focus on  $p$  values, because  $p$  values are decidedly frequentist. Bayarri and Berger make a good case for the use of  $p$  values in goodness-of-fit situations, where Bayes factors are typically not available. They include an interesting Bayesian calibration of  $p$  values. Both articles argue for the frequentist notion that  $p$  values should be (approximately) uniform for *each* value of the nuisance parameter. Bayesian considerations are used to develop some of the  $p$  values.

The statistics profession now is secure enough to utilize both Bayesian and frequentist notions in the same analysis. These articles are excellent examples of this third way.

*Nonuniformity.* Strict frequentists may prefer to define the  $p$  value as the supremum (over the nuisance parameter) of the probability of the statistic exceeding its observed value. This notion may be necessary for developing confidence intervals, but for goodness-of-fit it is pointlessly conservative.

### 3. AN EXPERIMENT

The two studies home in on the key factors for obtaining a good reference distribution for the test statistic. The distribution should

- be approximately independent of the statistic

- not be prior-heavy
- take into account variability in its estimation.

The first factor is violated when the regular posterior distribution of the statistic is used. The second is violated when the marginal distribution is used. One might consider the prior to be part of the model, as did Box (1980). But the authors argue persuasively that when assessing goodness of fit, the prior should not be part of the model. The third requirement is violated by the plug-in method, as can be seen when testing that the mean of a normal is 0. The plug-in uses the  $t$  statistic but treats it as a standard normal.

The improved  $p$  values presented in these articles behave impressively. The development is based on parametric models. I was curious as to whether the notions could easily extend to nonparametric bootstrap situations, so I challenged myself to come up with an improved  $p$  value procedure in the following simple problem:

Let  $X_1, \dots, X_n$  be iid, with continuous density  $f$ . Assume that the mean of  $f$  exists. The objective is to find a bootstrap  $p$  value for testing that the mean is 0, versus that it is greater than 0. Thus the nuisance parameter is  $f$ , among the set of densities with mean 0. The statistic is the sample mean  $\bar{X}_n$ .

I allowed myself only a couple of days (because these comments were due soon), so there was not much time for tinkering or tweaking. The procedure had to be simple, because I do not know much about density estimation. The straw man is the plug-in  $p$  value.

*Plug-In.* The data are  $\{x_1, \dots, x_n\}$ , and the sample mean is  $\bar{x}_{\text{obs}}$ . We wish to resample from the data, but with mean 0, so we let  $Y_1^*, \dots, Y_n^*$  be sampled independently from the batch of values  $\{y_1, \dots, y_n\}$ , where  $y_i = x_i - \bar{x}_{\text{obs}}$ . Then the plug-in bootstrap  $p$  value is

$$p_{\text{plug}} = P(\bar{Y}_n^* \geq \bar{x}_{\text{obs}}). \quad (1)$$

The potential improved  $p$  value is close to the *U*-conditional predictive  $p$  value of Bayarri and Berger.

*Challenger.* Take the statistic to be  $(Y_1, \dots, Y_n)$ , where  $Y_i = X_i - \bar{X}_n$ . It is uncorrelated with the statistic  $\bar{X}_n$ ,

though not independent. We need the posterior distribution of  $f$  given  $(Y_1, \dots, Y_n)$ , or at least something that mimics it. My ad hoc solution is to break the real line into bins,  $(c_k, c_{k+1}]$  for integers  $k$ , and let  $n_k = \#\{y_i \in (c_k, c_{k+1}]\}$ . Let  $K = \{k | n_k > 0\}$ . A random  $f, \hat{f}$ , is generated by first generating  $\underline{w}$  from a Dirichlet distribution with parameters  $(n_k | k \in K)$ . Then

$$\hat{f}(z) = \sum_{k \in K} w_k I(c_k - a < z \leq c_{k+1} - a),$$

where  $I(A)$  is the indicator for the set  $A$  and  $a$  is the constant that ensures the mean of  $\hat{f}$  is 0; that is,

$$a = \sum_{k \in K} w_k \frac{c_k + c_{k+1}}{2}.$$

The  $p$  value to challenge the plug-in is then

$$p_{\text{challenger}} = P(\bar{Z}_n \geq \bar{x}_{\text{obs}}), \tag{2}$$

where  $Z_1, \dots, Z_n$  are drawn independently from the  $\hat{f}$  density.

**Results.** I first experimented with the original  $X_i$ 's being normal, but even the plug-in  $p$  value worked well. Instead, I use  $X_i \sim \text{exponential}(1) - 1$ , subtracting "1" so that the mean is 0. The  $c_k = 2k$ .

Simulations were used for each of  $n = 5$  and  $n = 10$ . One thousand samples of  $n$   $X_i$ 's were generated, and the corresponding  $\bar{x}_{\text{obs}}, Y_i^*$ 's for the plug-in  $p$  value (1) and  $\hat{f}$  for the challenger (2), were calculated. For each sample, the  $p$  value was estimated by taking a bootstrap sample of 500 from the  $Y_i^*$ 's or  $\hat{f}$ .

Figure 1 illustrates the difference between the uniform distribution function and the empirical distribution function  $\hat{F}$  of the  $p$  values:  $\text{cdf} - \text{uniform} \equiv u - \hat{F}(u)$ , for  $0 < u < 1$ . Figure 1(a) has the two  $p$  values for  $n = 5$ . Note that both  $p$  values are not especially close to the uniform for larger values of the  $p$  value, but the challenger does improve substantially on the plug-in. The plot for  $n = 10$  also shows that the challenger is better for larger  $p$  values.

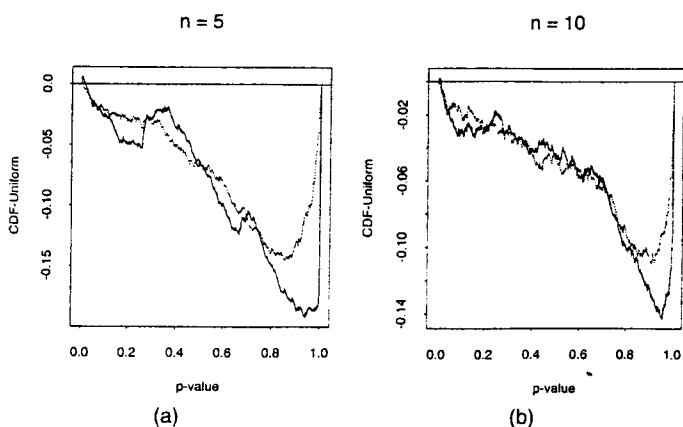


Figure 1. Comparing cdf's of  $p$  Values to Uniform. (a)  $n = 5$ ; (b)  $n = 10$ . (—, plug-in; - - - challenger).

One could argue that large values of the  $p$  value are not very important, but then the alternative could be changed to the mean being less than 0, in which case the  $p$  values become  $(1 - p)$  values.

I think that the experiment was a success. This first try at an improvement to the plug-in  $p$  value showed modest but positive results, and encourages examination of more sophisticated alternatives.

#### 4. A NIT

I have one nit to pick, concerning the Fisher exact test in Bayarri and Berger's Section 4. Consider Case 1, testing the equality of two binomial  $p$ 's. The statistic is  $T = X_{11}$ , with conditioning on the total number of successes,  $X_{1+}$ . If the observed value  $x_{1+}^o$  of  $X_{1+}$  is too small or too large, then the  $p$  value does not take on many values, and the distribution of the  $p$  value is quite conservative. This conservatism is due to using " $\geq$ " instead of " $>$ " in the definition:

$$p_{\text{fet}}(t_{\text{obs}}) = P(X_{11} \geq t_{\text{obs}} | X_{1+} = x_{1+}^o).$$

(Here "fet" denotes Fisher's exact test.)

As an alternative, consider the following *randomized*  $p$  value. Let  $U \sim \text{uniform}(0, 1)$ , independent of the data, and let  $R = X_{11} + U$ . If  $u_{\text{obs}}$  is the observed value of  $U$  and  $t_{\text{obs}}$  is the observed value of  $X_{11}$ , then the  $p$  value based on  $R$  is

$$p_{\text{rand}} = P(X_{11} > t_{\text{obs}} | X_{1+} = x_{1+}^o) \times u_{\text{obs}} + P(X_{11} \geq t_{\text{obs}} | X_{1+} = x_{1+}^o) \times (1 - u_{\text{obs}}). \tag{3}$$

This  $p$  value is exactly uniform(0, 1) under the null. Furthermore, it has optimal power characteristics, because the corresponding hypothesis test is uniformly most powerful among similar tests.

Before you get out the tar and feathers to run me out of the ASA, know that I am not actually recommending the randomized procedure. An inference should not depend on an ancillary statistic, in this case  $U$ . But Little (1989) made a compelling case that  $X_{1+}$  is *approximately* ancillary, and hence any inference should be made conditional on  $X_{1+}$ ; that is, one should use Fisher's exact test. The proposed  $p$  values, other than  $p_{\text{fet}}$ , are "almost" randomized, where the randomization arises from variation in the almost ancillary  $X_{1+}$  rather than the ancillary  $U$ . That is, they are politically correct (almost) randomized tests.

If one accepts that conditioning on  $X_{1+}$  is the right approach, what alternative is there? The problem is that there is not much information concerning the difference in binomial  $p$ 's when  $X_{1+}$  is near 0 or the total sample size, so the best solution might be to give the range of possible values for  $p_{\text{rand}}$  as  $u_{\text{obs}}$  varies:

$$(P(X_{11} > t_{\text{obs}} | X_{1+} = x_{1+}^o), P(X_{11} \geq t_{\text{obs}} | X_{1+} = x_{1+}^o)). \tag{4}$$

It is nonrandomized and contains an implicit admission that the data are not very informative. Alternatively, one could take the average; that is, set  $u_{\text{obs}} = 1/2$  in (3). The resulting

$p$  value will be closer to uniformity than  $p_{\text{fet}}$ , although can be conservative or anticonservative, depending on the data.

## 5. SUMMARY

The articles are great. The ideas look like they will extend

to more complicated situations. The discrete case remains somewhat problematic.

## ADDITIONAL REFERENCE

Little, R. J. A. (1989), "Testing the Equality of Two Independent Binomial Proportions," *The American Statistician*, 43, 283–288.

# Comment

Ludovico PICCINATO

My comments address the problem of model assessment and the practice of integrating over the sample space; I mainly discuss the article by Bayarri and Berger. My first reaction on reading this article was to notice that Bayarri and Berger have been very clever in overcoming so many obstacles in dealing with  $p$  values, such as the double use of data, the impossibility of using improper priors, and the danger of attributing too much importance to the prior. In principle, however, I do not agree on the practice of using  $p$  values in the Bayesian framework.

Anscombe (1963) commented on the fact that model checking was "something of an embarrassment" both for the Bayesian and for the Neyman–Pearson–Wald approaches. He then proposed turning to a version of Fisher's  $p$  values. Much work in this area has been done since then by Bayesian statisticians (as testified by the earlier excellent critical review in Bayarri and Berger 1997), but I believe that their present article, although introducing remarkable technical improvements from a Bayesian point of view, does not change the substance of the solution.

One standard criticism to checking a given model is that models can only be compared and cannot be assessed singly. Known methods are then available, like Bayes factors in their many versions, or complete probabilistic evaluations. But I agree with the authors on the fact that there is often an initial stage when a model is tentatively assumed; after having seen the data, the statistician must decide whether to adopt or to modify the model. This exploratory stage of statistical analysis does not live in a formal setup; formalization is indeed the ultimate goal of the process, and it is natural to use various procedures that might have only intuitive justifications. I have doubts as to the possibility of finding formal general rules for dealing with problems defined in a framework that is only partially formalized, but any effort in this direction is welcome.

## 1. CHECKING ONE MODEL

Bayarri and Berger introduce a general procedure for model checking. By modifying previous proposals, they suggest examining suitable  $p$  values; that is, quantities  $p =$

$\Pr\{t(\mathbf{X}) \geq t(\mathbf{x}_{\text{obs}})\}$ , where  $t$  is a statistic that takes larger values when the data are less compatible with a model  $M$ .

Bayarri and Berger assume that when the model reduces to a single density function  $f$ , with no unknown parameters, this density is used to calculate  $p$ . Their central issue is how to eliminate the nuisance parameter  $\theta$  when instead the model is  $M = \{f(\cdot; \theta), \theta \in \Theta\}$ . The article proves important results with respect to this point, but I first discuss the more basic issue of using  $p$  values. I must admit, however, that dogmatic attitudes are dangerous, because sometimes non-Bayesian methods can be put to a relevant Bayesian use. A surprising and important example, in a close area, is the use of calibrated  $p$  values for testing precise hypotheses in a robust Bayesian framework (Bayarri and Berger 1999; Sellke, Bayarri, and Berger 1999).

I do not object to summarizing the data with a statistic,  $t(\cdot)$ , measuring compatibility between data and model. The difficulty lies in defining a convincing standard for comparisons across different situations for judging when the compatibility is too low to try to improve the model. In the following, I explain why I do not expect that  $p$  values can be useful in this context.

In a complete probabilistic approach (using a discrete notation), one would like to obtain the probability  $\Pr(M|\mathbf{X} = \mathbf{x}_{\text{obs}})$ . In the context of model checking, however, it is sensible to assume that one might not be able or willing to directly elicit  $\Pr(M|\mathbf{X} = \mathbf{x}_{\text{obs}})$ . Moreover, because  $\Pr(M|\mathbf{X} = \mathbf{x}_{\text{obs}}) = \Pr(\mathbf{X} = \mathbf{x}_{\text{obs}}|M) \Pr(M)/\Pr(\mathbf{X} = \mathbf{x}_{\text{obs}})$ , and it is likely that neither  $\Pr(M)$  or  $\Pr(\mathbf{X} = \mathbf{x}_{\text{obs}})$  can be elicited, one is forced to concentrate on  $\Pr(\mathbf{X} = \mathbf{x}_{\text{obs}}|M)$ . If one is limited to a statistic  $t(\mathbf{x})$ , then one deals simply with  $\Pr\{t(\mathbf{X}) = t(\mathbf{x}_{\text{obs}})|M\}$ ; this can imply a loss of information but not a logical contradiction. My concept of "surprise" is based roughly on how small the probability  $\Pr\{t(\mathbf{X}) = t(\mathbf{x}_{\text{obs}})|M\}$  is for suitable  $t$ 's. Note that it is not easy to deal with this probability when  $M$  is composite, unless a proper probability distribution over  $\Theta$  conditional on  $M$  is also elicited.

Stating " $M$  true" means that one of the distributions  $f(\cdot; \theta)$  governs the system, at least with a reasonable ap-

Ludovico Piccinato is Professor, University of Rome, Italy (E-mail: ludpic@pow2.sta.uniroma1.it).

proximation. A way to assess the "surprise" in the data could be to consider the quantity  $\sup_{\theta} f(\mathbf{x}_{\text{obs}}; \theta)$ , because its smallness would also cast doubts on the model  $M$ . I think that most of the usual exploratory analyses (e.g., about the distribution of the empirical frequencies) informally follow this logical scheme.

The core of my argument is that because the values  $\Pr(\mathbf{X} = \mathbf{x} | M)$  for  $\mathbf{x} \neq \mathbf{x}_{\text{obs}}$  are not involved, logically the  $p$  values also should not enter the analysis. Gelman, Meng, and Stern (1996), in a rejoinder to discussion, remarked that these procedures are aimed at "routine goodness-of-fit assessment" in situations where recent advances in computational methods allow the use of complex and untested models. I believe that in such cases, a lack of theoretical justification would be especially dangerous.

Coming to more direct arguments, consider a model that contains only one discrete probability law, say  $f$ . Consider now another model  $f'$  such that  $f'(\mathbf{x}_1) = f(\mathbf{x}_2)$ ,  $f'(\mathbf{x}_2) = f(\mathbf{x}_1)$  and  $f'(\mathbf{x}) = f(\mathbf{x})$  otherwise, where  $t(\mathbf{x}_1) < t(\mathbf{x}_{\text{obs}}) < t(\mathbf{x}_2)$  and  $f(\mathbf{x}_1) \neq f(\mathbf{x}_2)$ . The two models only differ in the probabilities of results that did not occur, but the new  $p$  value  $p'$  is different from  $p$ . I find this illogical, exactly as I find any procedure that violates the likelihood principle illogical. Indeed, this principle cannot be invoked here, because the space of alternatives is not fully explicated. Nonetheless, the criterion of "not involving irrelevant aspects" that lies at the heart of the likelihood principle, in my opinion always holds.

Note that this example is almost a special case of example 3.2 of Bayarri and Berger (1999), where  $t(\mathbf{x}) = 1/m(\mathbf{x})$ . Bayarri and Berger's criticism regards the lack of invariance of the procedure by Box (1980), based on the prior predictive  $p$  value, with respect to transformations of the data. In fact, when the model is degenerate, the functions  $m$  and  $f$  coincide, and changing from  $f$  to  $f'$  is just a data-dependent transformation of the data. My point is that the problem is with the  $p$  value itself, not with the specific use of the prior predictive distribution.

## 2. USING $P$ VALUES WITH COMPOSITE MODELS

Criticisms concerning the use of the  $p$  value as a tool for testing precise hypotheses are traditional in the Bayesian literature, from Edwards, Lindman, and Savage (1963) to Berger and Sellke (1987) and Berger and Delampady (1987), where several of the inconsistencies of  $p$  values have been made evident with various techniques. According to one of my favorite books (Berger and Wolpert 1988, p. 107), "Questionable logic could perhaps be overlooked if it made little difference in practice, but here the averaging over other observations will virtually *always* have a profound effect."

These criticisms cannot be transferred without a justification to Bayarri and Berger's article. Bayarri and Berger in fact state very clearly that when the goal is comparing models, the appropriate tools are Bayes factors, not  $p$  values (see Sec. 1.4). In their Example 1, for instance, they assume

that under the null model, the observations are iid  $N(0, \sigma^2)$  with  $\sigma^2$  unknown, no alternative model is specified, and the statistic chosen to measure departure of  $\mathbf{x}$  from the model is  $t(\mathbf{x}) = |\bar{x}|$ . As I said before, I agree that this compatibility problem is realistic and sensible; however, one must acknowledge that this problem comes very close to testing  $\mu = 0$  under a model with unspecified  $\mu$  and  $\sigma^2$ .

Look what happens when testing  $\mu = 0$  under a  $N(\mu, \sigma^2)$  model for the iid observations, with  $\mu$  and  $\sigma^2$  both unknown, using sampling distributions conditional on  $s_{\text{obs}}^2$ . This conditioning can be interpreted as sacrificing some experimental information to eliminate the nuisance parameter  $\sigma^2$ . Under the model  $N(\mu, \sigma^2)$ , the density  $m(\bar{x} | s_{\text{obs}}^2)$  is  $t_{n-1}(n-1, \mu, s_{\text{obs}}^2/(n-1))$ ; hence, given the outcome  $\mathbf{X} = \mathbf{x}_{\text{obs}}$ , the normalized conditional likelihood for  $\mu$  is  $L(\mu) = (1 + (\bar{x}_{\text{obs}} - \mu)^2/s_{\text{obs}}^2)^{-n/2}$ , so that  $L(0) = (1 + \bar{x}_{\text{obs}}^2/s_{\text{obs}}^2)^{-n/2}$ . This has a clear meaning, even in non-Bayesian terms (see Royall 1997), as a measure of evidential support for  $\mu = 0$ . If we consider as outcome  $|\bar{X}| \geq |\bar{x}_{\text{obs}}|$  instead of  $\mathbf{X} = \mathbf{x}_{\text{obs}}$ , then the corresponding conditional likelihood for  $\mu$  is directly  $p_{\text{cpred}}$ ; this likelihood is already normalized, because

$$\sup_{\mu} \Pr(|\bar{X}| \geq |\bar{x}_{\text{obs}}| | S^2 = s_{\text{obs}}^2) = 1.$$

As a numerical example, when  $n$  and  $\bar{x}_{\text{obs}}/s_{\text{obs}}$  are such that  $p_{\text{cpred}} = .01$ ,  $L(0)$  increases from .021 to .032 as  $n$  varies from 10 to 50. In these cases, one could say that  $p_{\text{cpred}}$  overestimates the surprise for the null model, in accord with the intuition that substituting the actual result  $\mathbf{X} = \mathbf{x}_{\text{obs}}$  with the tail  $|\bar{X}| \geq |\bar{x}_{\text{obs}}|$  diminishes the compatibility of the data with the model. Generally, the relation between the relative likelihoods and the  $p$  values is quite complex, and it varies depending on the situation. Here it depends also on  $n$  and, with small values of  $n$ ,  $L(0)$  could be even smaller than  $p_{\text{cpred}}$ . Hence  $p_{\text{cpred}}$  shows an irregular behavior in a conditional framework for a problem that is close to the original one and has reasonable solutions. Indeed, I cannot exclude the possibility that in this case as well, a suitable calibration might improve the interpretability of  $p_{\text{cpred}}$ .

I am aware that raising these objections might be unfair to Bayarri and Berger, because testing hypotheses is not the field designed for  $p_{\text{cpred}}$ , and one could say that introducing a  $N(\mu, \sigma^2)$  model as an encompassing model is a further ad hoc device. Even the choice of  $L(0)$  as a reference value in the example could be questioned, although its use could be also justified in terms of robustness. In any case, I think that integrating over the sample space after knowing the data will always introduce too much noise; after all,  $p$  values are still  $p$  values.

## ADDITIONAL REFERENCES

- Anscombe, F. J. (1963), "Tests of Goodness of Fit," *Journal of the Royal Statistical Society*, Ser. B, 25, 81-94.  
 Berger, J. O., and Wolpert, R. L. (1988), *The Likelihood Principle* (2nd ed.), Hayward, CA: Institute of Mathematical Statistics.  
 Royall, R. M. (1997), *Statistical Evidence. A Likelihood Paradigm*. London: Chapman and Hall.