

We found the discussions to be uniformly excellent and thought provoking, and are grateful to the discussants for illuminating the problem from interestingly different perspectives. We organize our response into four topics: double use of the data and conditioning, choice of  $T$ , domains of applicability, and use of tail areas.

#### DOUBLE USE OF THE DATA AND CONDITIONING

Evans and Stern both discuss the central issue: What constitutes double use of the data in computing a  $p$  value? Interestingly, they take opposite perspectives on this issue. Evans argues that the only way to avoid double use of the data is to separate the data into two independent subsets (data splitting), using one subset to fit the model and the other to assess the validity of the model. Because the procedures that we recommend do not formally split the data into two parts, Evans suggests that they can be guilty of double use of the data. In contrast, Stern does not feel that it is necessary to worry about double use of the data to the extent that we worry; he does not feel that even the posterior predictive  $p$  value involves double use of the data.

These contrasting positions perfectly highlight what we view to be the main motivation for our paper. We feel that it is very important to avoid a double use of the data, but that separating the data into two independent subsets is a too-drastic solution that can lose considerable power. We consider each of these issues in turn.

Stern asserts that the posterior predictive distribution is a legitimate distribution (at least to a Bayesian), and that there is nothing controversial about using this distribution to compute the probability of events. We certainly agree with the first part of the statement, and also agree with the second, *unless* the event being considered is an event determined from the same data used to compute the posterior predictive distribution. This is precisely what is done with the posterior predictive  $p$  value, with the tail region (the "event") being determined by the same data used to compute the posterior predictive distribution. Put bluntly, we feel that this is a classic "double use of the data." The fact that the posterior predictive  $p$  value seems to be uniformly conservative (in  $\theta$ ) is but one indication of this. (Nothing that is a legitimate Bayesian  $p$  value can be uniformly conservative or uniformly anticonservative; "on average," a legitimate Bayesian  $p$  value must be uniform, as Evans observes.) The fact that the plug-in  $p$  value seems to be closer to uniform than the posterior predictive  $p$  value should also raise alarms; virtually everyone would agree that the plug-in  $p$  value makes double use of the data, yet it appears to typically be superior to the posterior predictive  $p$  value!

In response to Evans, all we can basically do is restate our arguments to the effect that the procedures that we recommend do avoid double use of the data. For the conditional predictive  $p$  value, we feel that the case is particularly clear.

Recall that one of the interpretations that can be given this  $p$  value is that it arises as the observed tail area corresponding to the distribution  $m(t|u)$ , which is a legitimate Bayesian predictive distribution and hence does not entail double use of the data. We did not extensively discuss this feature in the article (primarily because we focused on the easier-to-compute partial posterior predictive  $p$  value), but it is worthy of emphasis. "Pure" Bayesian reasoning suggests that all information for model criticism lies in the prior predictive distribution  $m(x)$ ; the trick is in determining how to appropriately use this distribution. Given that we have decided to use some type of  $p$  value, our conjecture is that the conditional prior predictive distributions,  $m(t|u)$ , form the class of admissible Bayesian distributions with which one can construct  $p$  values (using the word "admissible" in a generic sense and not implying that all of these are necessarily good for model checking). Note that the posterior predictive  $p$  value cannot be written as a  $p$  value arising from one of these conditional distributions.

The fact that the  $p$  values that we recommend appear to be uniform, even from a frequentist perspective, further bolsters the claim that these procedures effectively avoid double use of the data. (The failure in this regard of the procedures such as  $p_{\text{post}}$  and  $p_{\text{plug}}$  is suggestive.) We certainly agree with Evans that if  $T$  and  $U$  are independent, then there clearly is no double use of the data, but we do not feel that this is a necessary condition; the simultaneous Bayesian justification (via  $m(t|u)$ ) and frequentist justification (via uniformity) suffices to establish that our procedures effectively avoid double use of the data.

Example 2 of Evans is interesting in this regard, in that he observes that direct use of the prior predictive  $p$  value is problematical and suggests that the only cure for the problem is data splitting. Note, however, that  $\bar{X}$  is a sufficient statistic for  $\theta$ , so that the natural distribution for model checking is (writing rather loosely)  $m(x|\bar{x})$ , the uniform distribution on all sequences  $x$  that sum to  $n\bar{x}$ . This does not suggest a departure statistic,  $t(x)$ , to use in model checking, but it is a good distribution to use to compute the  $p$  value, once  $T$  has been chosen (and, of course, results in  $p_{\text{sim}}$ ). The message is that double use of the data, and the problems that it causes, can be avoided by appropriate conditioning, so that data splitting is not strictly necessary.

It should be noted that there are considerable advantages in not requiring that the data be separated into two independent parts for model validation. First, such splitting is wasteful, especially when data are scarce; one clearly sacrifices power by such a division (assuming, of course, that the same end can be accomplished by appropriate conditioning). Second, the data are often dependent, and it is

then not clear how to proceed with the "independent subsets" approach. Finally, even if the data are independent, it is far from clear how to determine the relative sizes of the two subsets, so that one must face an unnecessary (to us) complication of the analysis.

Marden's nit-pick about the Fisher exact test is absolutely correct, but reinforces the dilemma facing frequentist statistics. Most frequentists condition when there is an obvious ancillary statistic (e.g., the randomization variable), and we are certainly supportive of this instinct. Indeed, we go considerably further and feel that frequentists should be conditioning all the time, even when there is no ancillary statistic (see, e.g., Berger, Boukai, and Wang 1997; Berger, Brown, and Wolpert 1994). The difficulty is that there are no guiding principles (other than ancillarity) to help frequentists obtain good conditioning statistics. We thus turned to Bayesian reasoning to help sort out the conditioning; we suspect that our recommended procedures have excellent frequentist properties precisely because this reasoning did effectively suggest good frequentist conditioning statistics.

### CHOICE OF $T$

Evans worries that we advocate an informal choice of  $T$ , and does not feel that this is a good idea. We did not mean to give the impression that this is what we advocate; we were simply observing that people often choose  $T$  rather informally for model checking, and we want a procedure that is legitimate for even informal choices. We did not really understand the subsequent comments of Evans about limitations of our choice of  $T$ , because we allowed any  $T$  whatsoever. For instance, Evans advocates computing tail areas for the quantity  $m_T(t(\mathbf{x})|U(\mathbf{x}))/m_T(t(\mathbf{x}))$ . To us, this is just another statistic,  $t^*(\mathbf{x})$  say, that could be used if desired. (Note, however, that Evans would not use the same reference distribution as we would to compute the  $p$  value for this statistic, so that there is certainly a difference between the two methods.)

The choice of  $T$  is undoubtedly highly relevant to having good power to detect model inadequacies, which is why we would not want to minimize the importance of careful choice of  $T$ . Our view has simply been that when presented with a choice of  $T$ , we can say what to do (use  $p_{ppost}$ ). Whether the partial posterior predictive approach can say useful things about the choice of  $T$  is not yet clear to us.

In this regard, it is interesting to consider the "bad" choices of  $T$  in Examples 4 and 5 that led to the "liberal" behavior of  $p_{ppost}$  that Boos and Stern found disturbing. Indeed, Stern suggests that he is much happier with  $p_{post}$  in Example 5, because it returns a  $p$  value of nearly .5 in the considered situation.

Recall first that (conventionally)  $T$  is to be chosen so that large values of  $T$  would cause one to distrust the model. In Example 5, if  $T = n$  (the largest possible value of  $T$ ), then  $p_{ppost} = 1/(n+1)$ , whereas  $p_{post} \cong .5$ . To us, if large values of  $T$  are cause for doubting the model and one observes the largest possible value of  $T$ , then a  $p$  value of  $1/(n+1)$  is more reasonable than .5. Stern's counterargument is

essentially that  $p_{ppost}$  is really just rejecting the prior in this case, which is not appropriate, because the prior probably was just chosen for convenience.

There is another issue here, however—that the given choice of  $T$  in Example 5 is disastrous. (Sufficient statistics are virtually useless for model checking.) The small value of  $p_{ppost}$  is simply saying that something is wrong, in light of the fact that large values of  $T$  were a priori viewed as surprising. Indeed, something is wrong: The (model, prior,  $T$ ) triplet is not compatible with a priori beliefs and thus should be reexamined. Hopefully, the statistician will recognize that the main error was in the choice of  $T$  and will come up with something better. In contrast,  $p_{post} \cong .5$  suggests that nothing is wrong with the analysis. Note that similar comments could be made about the choice  $T = X_{11}$  in Example 4.

We may be belaboring the obvious, but it is worth stressing again that in model-checking, conservatism is dangerous, whereas the aforementioned type of liberalism carries little danger. Conservative  $p$  values are dangerous in model checking in that they can give statisticians confidence in bad models. In contrast, the type of liberalism that can arise with  $p_{ppost}$  in the most extreme cases carries no real danger; it simply points out an inconsistency among the assumptions in the analysis (including the assumption that the specified  $T$  is useful for model criticism) that must be resolved.

Despite our relative comfort with the results from  $p_{ppost}$  here, we are in accordance with Boos in not being convinced that  $p_{ppost}$  is an all-purpose solution for contingency tables (or other discrete problems) for any choice of  $T$ . In particular, as we mentioned in the article, one does have to exert some care in choice of  $T$  in discrete problems to avoid having "near sufficiency" simply because of the discreteness.

### DOMAINS OF APPLICABILITY

Boos mentions that modifications of  $p_{sup}$ , especially when combined with other techniques such as conditioning on sufficient statistics, are often not excessively conservative and hence can have good power. We agree, but feel that  $p_{ppost}$  has considerably wider applicability.

Stern comes at this from the other side, noting that  $p_{ppost}$  can be considerably more difficult to use than  $p_{post}$  for complex models. We admittedly only consider simple models in the article, but that is for pedagogical reasons; one can effectively understand and theoretically evaluate an approach only in simple models. We feel that  $p_{ppost}$  is usable in complex situations, but Stern's basic point is right: computation of  $p_{post}$  typically will be easier than computation of  $p_{ppost}$ .

Furthermore, if good diagnostic statistics or discrepancies are selected, then  $p_{post}$  should be quite effective. As argued by RVV (2000), a diagnostic statistic or discrepancy is "good," for use with  $p_{post}$  if it is approximately "centered." But what happens if it is not approximately centered? Then, all too often, one will find that  $p_{post}$  is not small (when  $p_{ppost}$  would be small) and will (erroneously) conclude that there is no reason to question the model.

Will users naturally find "good" diagnostic or discrepancy statistics? In our very simple Example 2 when  $n = 4$ ,

the choice of  $T = \min\{x_1, x_2, x_3, x_4\}$  does not obviously seem "bad," yet  $p_{\text{post}}$  will never be as small as, say, .05, no matter how strong the evidence against the model. We are not nearly as confident as Stern that in truly complex models, users will naturally find "good" diagnostic statistics, or at least we are not confident that nonexperts will naturally produce such.

We are not arguing that the posterior predictive  $p$  value (and related diagnostics) should be abandoned, but rather that its potentially severe conservatism should be more fully recognized. Obviously, if  $p_{\text{post}}$  turns out to be small, then one has learned that the model should be reexamined. The problem is that a large value of  $p_{\text{post}}$  provides little assurance that the model is okay, unless one knows that  $T$  is appropriately centered. Stern alludes to the growing list of successful uses of the posterior predictive approach, in which bad models were identified; presumably there is also a growing (but unidentified) list in which bad models were not detected because of the conservatism of the posterior predictive approach.

In this sense, a large value of  $p_{\text{ppost}}$  would provide considerably more assurance in a model. But Stern rightly wonders whether it will be routinely possible to compute  $p_{\text{ppost}}$  in practice. All we are really suggesting is that often  $p_{\text{ppost}}$  can be computed and should then be used, but often it will not be easily computable, and then one should then try  $p_{\text{post}}$  or  $p_{\text{plug}}$  as the best available options.

Stern also mentions the diagnostic uses of the posterior predictive approach. Looking at, say, posterior predictive residuals does not cause the issue of potential excessive conservatism to disappear, but the ability to look at such diagnostics is certainly a strength of the posterior predictive approach. It is not clear to us whether the partial posterior predictive approach can be generalized in this direction.

Marden has gone where we would not have dared to go: into an investigation of the application of the idea of partial predictive  $p$  values to a setting based on a bootstrap analysis. We are quite fascinated with his analysis, not so much because his version of the partial predictive  $p$  value seemed to be successful (this is no longer a surprise to us), but because he proposed a quite interesting (though ad hoc) nonparametric Bayesian analysis. It looks intriguing, and we hope to learn more about the justification for this analysis.

#### USE OF TAIL AREAS

Piccinato is concerned with the basic "mistake" of replacing the actual data by a tail area when computing the  $p$  value. He indicates that he would be much more comfortable with measures of surprise based on  $\Pr(t(\mathbf{X}) = t(\mathbf{x}_{\text{obs}}))$  than those based on  $\Pr(t(\mathbf{X}) \geq t(\mathbf{x}_{\text{obs}}))$ . Piccinato discusses

the logical reasons for this conclusion, as well as the practical reason that the two quantities often suggest very different evidentiary statements, and it is the former that is the actual evidence.

When there is an alternative under consideration, so that we know how to utilize  $\Pr(t(\mathbf{X}) = t(\mathbf{x}_{\text{obs}}))$  (through, say, default Bayes factors or conditional frequentist tests), we are in complete agreement with Piccinato and would not consider using the  $p$  value. One might suppose that the "calibration of  $p$  values" that we mention (derived in Sellke, Bayarri, and Berger 1999) solves the problem, but it does not really do so in that it is a "lower bound" calibration that may well be too low, especially for larger sample sizes (although it is always a significant improvement on the raw  $p$  value).

Even when there is no alternative available (our stated situation), we are still highly sympathetic with Piccinato's position. Indeed, we began this particular avenue of research precisely by trying to develop useful measures of surprise based on  $\Pr(t(\mathbf{X}) = t(\mathbf{x}_{\text{obs}}))$ . [Our efforts in this direction, as well as the history of such efforts (which is quite substantial) was given in Bayarri and Berger 1997.] More recent studies in this direction include those by Bayarri and Morales (1999) and Castellanos (1999). Unfortunately, for model checking, we were simply not able to find ways of directly utilizing  $\Pr(t(\mathbf{X}) = t(\mathbf{x}_{\text{obs}}))$  that we were convinced were superior to  $p_{\text{ppost}}$  (together with the calibration). However, we would not be at all unhappy if someone could succeed in finding an effective way to use  $\Pr(t(\mathbf{X}) = t(\mathbf{x}_{\text{obs}}))$ , because this also agrees more directly with our own intuition. Note that even for surprise measures based on this quantity, it typically is necessary to eliminate a nuisance parameter  $\theta$ , and the techniques we advocate in the article are still useful to that end.

Our only real quibble with Marden is the first sentence of his Section 2, in which he states that  $p$  values are decidedly frequentist and suggests that Bayesians thus would be uncomfortable with their use. Our view is that Bayesians are not at all uncomfortable with *optimal* frequentist measures, but, alas, there is nothing truly frequentist (much less optimal) about  $p$  values (e.g., they trivially have no interpretation as actual frequentist error probabilities). This quibble is nit-nit-picking, however, in that Bayesians are indeed very uncomfortable with  $p$  values—but frequentists should be equally uncomfortable!

#### ADDITIONAL REFERENCES

- Bayarri, M. J., and Morales, J. (1999), "Bayesian Measures of Surprise for Outlier Detection," Working Paper 99-38, ISDS, Duke University.  
 Castellanos, M. E. (1999), "Medidas de Sorpresa para Bondad de Ajuste," masters thesis, University of Valencia.