(1988) for our current more difficult problem. We next develop our approximation in the current context using the procedure of Leonard, Hsu and Tsui (1989).

Let $\mathbf{u}_t$ denote the vector conditionally maximizing (17), subject to $g(\mathbf{u}) = t$. Then, expanding $\log p(\mathbf{u} \mid \mathbf{y})$ in a Taylor series about $\mathbf{u} = \mathbf{u}_t$ and neglecting cubic and higher terms in the expansion yields an approximation $p_t^*(\mathbf{u} \mid \mathbf{y})$ to $p(\mathbf{u} \mid \mathbf{y})$ in a neighborhood of $\mathbf{u} = \mathbf{u}_t$. Based upon $p_t^*(\mathbf{u} \mid \mathbf{y})$, the required marginalization can be performed without further approximation, yielding

$$p^*(t \mid \mathbf{y}) \propto p(\mathbf{u}_t \mid \mathbf{y}) \mid \mathbf{R}_t \mid^{-1/2} \times \exp\{\tfrac{1}{2}\mathbf{l}_t^T \mathbf{R}_t^{-1} \mathbf{l}_t\} f(t \mid \mathbf{u}_t^*, \mathbf{R}_t^{-1}), \tag{18}$$

where

$$\mathbf{l}_t = \left. \frac{\partial \log p(\mathbf{u} \mid \mathbf{y})}{\partial \mathbf{u}} \right|_{u=u_t}, \tag{19}$$

$$\mathbf{R}_t = \left. \frac{-\partial^2 \log p(\mathbf{u} \mid \mathbf{y})}{\partial (\mathbf{u}\mathbf{u}^T)} \right|_{u=u_t} \tag{20}$$

and

$$\mathbf{u}_t^* = \mathbf{u}_t + \mathbf{R}_t^{-1} \mathbf{l}_t \tag{21}$$

with $f(t \mid \mathbf{u}, \mathbf{C})$ denoting the density of $t = g(\mathbf{u})$ when $\mathbf{u}$ possesses a multivariate normal distribution with mean vector $\mathbf{u}$ and covariance matrix $\mathbf{C}$.

In the above derivation, we assume that $\mathbf{z}$ and $\boldsymbol{\theta}$ have already been suitably transformed to permit approximate multivariate normality, conditional on $\mathbf{y}$, of the $\mathbf{u}$ vector. When applying (18), it is necessary to either know f or to use a further approximation for this important f component.

We hope that our suggestions will again help to catalyze the literature on this interesting topic. We would like to thank Professor Bjørnstad for highlighting the importance of predictive inference.

ACKNOWLEDGMENTS

The authors would like to thank Ella Mae Matsumura and Ron Butler for helpful comments, and George Tiao for encouraging the development of (5) at the University of Wisconsin in 1980.

ADDITIONAL REFERENCES

ALBERT, J. H. (1988). Computational methods using a Bayesian hierarchical generalized linear model. *J. Amer. Statist. Assoc.* **83** 1037–1044.

GHOSH, M., HWANG, J. T. and TSUI, K. W. (1983). Construction of improved estimators in multiparameter estimation for discrete exponential families (with discussion). *Ann. Statist.* **11** 351–376.

HSU, J. S. J., LEONARD, T. and TSUI, K. W. (1988). Bayesian inference with applications to contingency table analysis. Technical Report No. 825, Univ. Wisconsin-Madison.

JOHNSON, R. A. and LADALLA, J. N. (1979). The large sample behavior of posterior distributions which sample from multiparameter exponential family models, and allied results. *Sankyhā Ser. B* **41** 196–215.

KASS, R. E., TIERNEY, L. and KADANE, J. B. (1988). The validity of posterior expansions based on Laplace's method. Technical Report No. 396, Dept. Statistics, Carnegie Mellon Univ.

LADALLA, J. N. (1976). The large sample behavior of posterior distributions when sampling from multiparameter exponential family models, and allied results. Ph.D. dissertation. Dept. Statistics, Univ. Wisconsin-Madison.

LEONARD, T. (1972). Bayesian methods for binomial data. *Biometrika* **59** 581–589.

LEONARD, T. (1973). A Bayesian method for histograms. *Biometrika* **60** 297–308.

LEONARD, T. (1975). A Bayesian approach to the linear model with unequal variances. *Technometrics* **17** 95–102.

LEONARD, T. (1976). Some alternative approaches to multiparameter estimation. *Biometrika* **63** 69–76.

LEONARD, T. and NOVICK, M. R. (1986). Bayesian full rank marginalization for two-way contingency tables. *J. Educ. Statist.* **11** 33–56.

LEONARD, T., HSU, J. S. J. and TSUI, K. W. (1989). Bayesian marginal inference. *J. Amer. Statist. Assoc.* **84** 1051–1058.

MORRIS, C. (1983). Parametric empirical Bayes inference: Theory and applications (with discussion). *J. Amer. Statist. Assoc.* **78** 47–65.

RUBINSTEIN, R. Y. (1981). *Simulation and the Monte Carlo Method.* Wiley, New York.

Rejoinder

Jan F. Bjørnstad

I would like to thank the discussants for their comments which have extended and illuminated the ideas of predictive likelihood in the review. In this rejoinder, I will expand on some of the issues raised by them, and also take up the issue of additivity for predictive likelihoods.

The comment of Leonard, Tsui and Hsu is concerned mainly with the approximation of Bayesian predictive distributions, and as such deals not with the likelihood approach. A predictive likelihood is based on the joint likelihood $l_y(z, \theta) = f_\theta(y, z)$, not on the Bayes posterior density $f(z \mid y)$. Hence, Leonard,

Tsui and Hsu are correct in pointing out that approximations of the "flat-prior" Bayes posterior density are really Bayesian approaches. However, these approximations can also be viewed as adjustments to the profile predictive likelihood, L_p, to account for the number of unknown parameters in the model. The paired comparisons example illustrates well how inappropriate L_p can be if the number of unknown parameters is large, while the adjustments based on approximations to the Bayes posterior density will correct this deficiency. In this context, it should be noted that none of Butler's suggestions are constructed as approximations to the Bayes predictive density. The new approximation suggested by Leonard, Tsui and Hsu when one wants to predict a real-valued $Z = g(Y')$ seems to be particularly useful when $f(z \mid \theta)$ cannot be obtained in closed analytic form. It therefore widens the area of prediction problems that can be handled.

The main issue raised in the comment of Professor Butler is an important one; namely, how to evaluate predictive likelihoods by considering the predictive intervals they generate. Butler claims that the usual unconditional confidence level, $\mathrm{Cl}(\theta)$, is neither relevant nor useful, and that one should instead use the conditional confidence level given the appropriate ancillary statistic A_a for the data. Let us call this the partial conditional level and denote it by $C_\theta(A_a)$. Butler then judges the quality of a predictive likelihood by the closeness of $C_\theta(A_a)$ to preset nominal values like .90 or .95. When constructing an interval in the usual pivotal frequentist way, it is clearly wise to condition on an ancillary statistic, and in that respect $C_\theta(A_a)$ is an interesting feature of a prediction interval. However, this certainly does not mean that the usual confidence level is useless or irrelevant as a measure of the degree of confidence we have in a predictive interval. For one thing it is, of course, the expectation of $C_\theta(A_a)$. More importantly, $\mathrm{Cl}(\theta)$ should be required to be approximately equal to or larger than the nominal level. Otherwise, *repeated use of the predictive likelihood* cannot be recommended (as seen in the assessment of L_e in Section 4). Hence, the unconditional level $\mathrm{Cl}(\theta)$ plays, in fact, an important role in the assessment of predictive likelihoods, and I have to disagree strongly with Butler's claim that $\mathrm{Cl}(\theta)$ is irrelevant and useless. As noted by Butler, also Cox (1986) mentions the need for assessing predictive likelihoods in terms of hypothetical long-run properties.

Professor Butler claims that I have introduced new material on predictive likelihood assessment. That may be true with regard to considering the conditional coverage given the data, $C_\theta(y)$. But I cannot take credit for suggesting $\mathrm{Cl}(\theta)$. For example, Lejeune and Faulkenberry (1982) assess L_p in the cases of Poisson and binomial models by computing $\mathrm{Cl}(\theta)$ for the predictive intervals generated by L_p. Also, using $\mathrm{Cl}(\theta)$ to assess prediction methods in general is certainly not new, see, e.g., Aitchison and Dunsmore (1975). In fact, it seems to be the most common way of evaluating statistical interval-conclusions generally in prediction problems.

In the normal case in Section 4, the ancillary and sufficient statistics are independent, so in this example $\mathrm{Cl}(\theta)$ and $C_\theta(A_a)$ are identically the same. When $C_\theta(A_a)$ does differ from $\mathrm{Cl}(\theta)$, $C_\theta(A_a)$ certainly gives interesting new insight into the performance of the prediction method. However, I do not think it is enough to consider $\mathrm{Cl}(\theta)$ and $C_\theta(A_a)$. One should also look at the fully conditional confidence level given the data, $C_\theta(y)$. To me, $C_\theta(y)$ is the most relevant characteristic on which to base a criterion for evaluating prediction intervals. One such criterion is $P_\theta(C_\theta(Y) \geq 1 - \alpha)$, where $1 - \alpha$ is the nominal level desired. Also, an important feature of $C_\theta(Y)$ is its expected value $\mathrm{Cl}(\theta)$.

In the normal case of Section 4 and in all the examples considered by Butler, $P_\theta\{C_\theta(Y) \geq 1 - \alpha\}$ is independent of θ. Hence, the issue raised by Butler of whether it is sensible to consider $\inf_\theta P_\theta\{C_\theta(Y) \geq 1 - \alpha\}$ as a criterion does not arise in these cases. If $P_\theta\{C_\theta(Y) \geq 1 - \alpha\}$ depends on θ, an alternative to taking infimum is to consider $P_\theta\{C_\theta(Y) \geq 1 - \alpha\}$ as a function of θ, for example restricted to a neighborhood of $\hat{\theta}$ similar to what Butler considers for $C_\theta(A_a)$.

With a nominal level of, say, 90% an alternative to the usual goal of $\mathrm{Cl}(\theta)$ (or $C_\theta(A_a)$) = .9 is to require that $P_\theta\{C_\theta(Y) \geq .9) = .9$. This means that we want at least 90% conditional coverage 90% of the times. Typically this implies that $\mathrm{Cl}(\theta) > .9$.

There is a way to combine the fully conditional evaluation, based on $C_\theta(y)$, and the partially conditional approach. Say we want 90% confidence *given the data* for the predictive interval. It is then important whether or not the event $E = \{C_\theta(Y) \geq .9\}$ occurs. The probability of E should be calculated in the right frame of reference which, according to the likelihood principle for prediction, is conditional on A_a. This amounts to considering $P_\theta(E \mid A_a) = P_\theta\{C_\theta(Y) \geq .9 \mid A_a\}$ as a way to assess a predictive likelihood.

In light of the above discussion on predictive likelihood assessment, let us now consider Butler's four examples. In Example 1, no ancillary statistic exists and Butler uses $\mathrm{Cl}(\theta)$. The 90%-interval based on the predictive pivotal statistic $A_p = \frac{1}{2}(Z - Y)$, $Y \pm 2.311$, obtains $P(C_\theta(Y) \geq .9) = .756$. The 90% nominal predictive interval based on L_{a2} and L_{a3}, $Y \pm 2.459$, has $P(C_\theta(Y) \geq .9) = .809$, while the interval based on L_p and L_{pc}, $Y \pm 2.045$, gets $P\{C_\theta(Y) \geq .9\} = .619$. To me it is therefore not at all clear that the predictive pivot-approach is preferable to L_{a2} and L_{a3} here. To achieve $P\{C_\theta(Y) \geq .9\} = .9$, the prediction interval

turns out to be $Y \pm 2.836$, which has $\mathrm{Cl}(\theta) \equiv .96$. Hence, in order to have 90% confidence in reaching the nominal level, the unconditional confidence level must be 96%.

Example 2, the $U(\theta - \frac{1}{2}, \theta + \frac{1}{2})$ case, is of special interest. As in the parametric case when constructing confidence interval for θ, it illustrates convincingly the need to condition on the ancillary statistic $A_a = X_2 - X_1$, when constructing a prediction interval using the pivotal statistic $A_p = Z - \bar{X}$. The 90% prediction interval based on the predictive pivot $f(A_p \mid A_a)$ is given by:

$$I = \begin{cases} \bar{x} \pm .45 & \text{if } |A_a| \geq .9, \\ \bar{x} \pm (1 - |A_a| - \sqrt{(1 - |A_a|)/10} & \text{if } |A_a| < .9. \end{cases}$$

It is readily computed that $P(C_\theta(Y) \geq .9) = .543$, which seems rather low. Butler claims that none of the predictive likelihoods are applicable to this problem. This is clearly not the case. They may not utilize all predictive information about Z in the best possible way, but certainly some of them will work reasonably well. For L_c, Butler illustrates clearly when conditioning on R will work properly. Here R contains an ancillary part and hence some information about Z, and L_c will not use all predictive information available in the data. Still, it is possible to derive L_c when $|A_a| > 0$. The nominal 90% predictive interval based on L_c is given by:

$$I_c = \begin{cases} \bar{x} \pm (.45)(1 - 2\log|A_a|) & \text{if } |A_a| \geq e^{-1/18}, \\ \bar{x} \pm (|A_a|^{0.1} e^{-.05} - \frac{1}{2}|A_a|) & \text{if } |A_a| < e^{-1/18}. \end{cases}$$

When $x_1 = .06$ and $x_2 = .98$, $I_c = \bar{x} \pm .4833 = (.037, 1.003)$ which is vastly better than the interval based on $f(A_p)$. The conditional level, given $A_a = .92$, equals .960. However, $P(C_\theta(Y) \geq .9) = .812$ which is a big improvement over the predictive pivot approach. Even though L_c throws away some predictive information about z, and hence will give an interval larger than one would think necessary, the 90% predictive interval based on L_c has a fully conditional level, compared to I, that better reflects how a predictive interval should behave.

Even though the profile predictive likelihood, L_p, does not work well here, it can be derived. Defining $L_p(z \mid y)$ does not require $\hat{\theta}_z$ to be unique, being defined as $\sup_\theta f_\theta(y, z)$. Here, $L_p(z \mid y)$ is such that $Z \sim U(x_{(2)} - 1, x_{(1)} + 1)$, where $x_{(1)} = \min(x_1, x_2)$ and $x_{(2)} = \max(x_1, x_2)$. The 90% predictive interval is then:

$$I_p = \bar{x} \pm (.9)(1 - \tfrac{1}{2}|A_a|).$$

With $x_1 = .06$ and $x_2 = .98$ this becomes (.034, 1.006), practically the same as I_c. Usually, however, I_c will be quite a bit shorter than I_p. $\mathrm{Cl}(\theta) \equiv .988$ for I_p, and it can be shown that $C_\theta(y) \geq .9$ for all (θ, y) such that $P(C_\theta(Y) \geq .9) = 1$. In this case, L_p will lead to unnecessarily large predictive intervals.

Let us for this example also consider the partial guarantee, $P(C_\theta(Y) \geq .9 \mid A_a)$, as a function of $|A_a|$. For L_p this guarantee is 1, of course. For I_c and I, the partial guarantee is given in Figure 1. Figure 1 reveals a weakness in the predictive pivot approach of constructing conditional 90% intervals. It seems that even

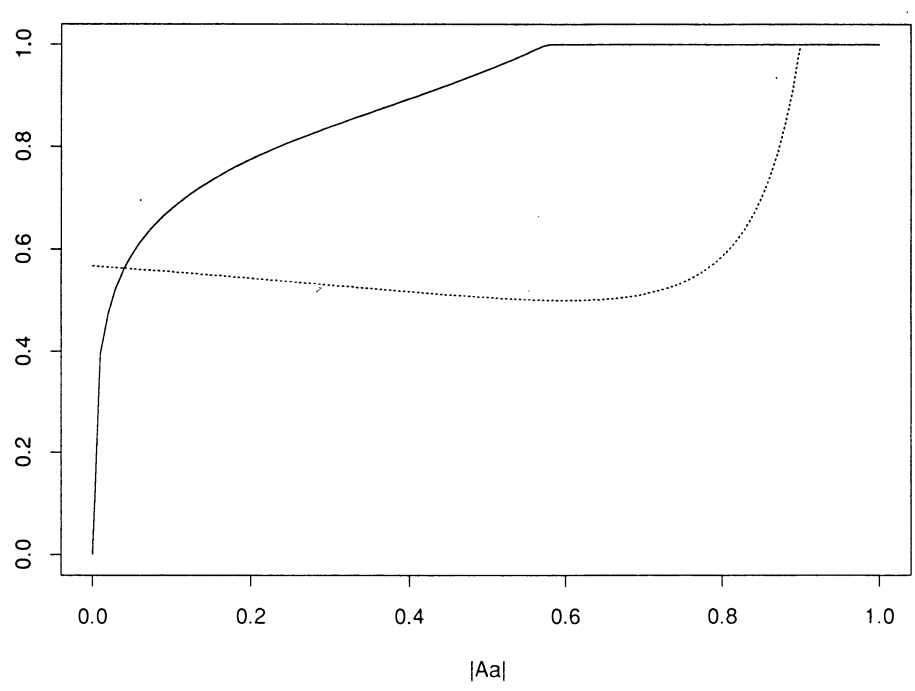

FIG. 1. $P\{C_\theta(Y) \geq .9 \mid A_a\}$ *for* I_c (——) *and* I (- - - -).

the "not fully informative" conditional predictive likelihood does better.

Also in Example 3, I think, one should consider $P(C_\theta(Y) \geq 1 - \alpha)$ before one tries to make a preferential ordering of the predictive likelihoods. In the $U(0, \theta)$ case of Example 4 (also Example 4 in the review), the 90%-interval based on L_p, $(0, 1.826\hat{\theta})$, obtains $P(C_\theta(Y) \geq .9) = .756$, while the 90%-interval based on L_c, $(0, 5\hat{\theta})$, has $P(C_\theta(Y) \geq .9) = .968$, confirming that L_c's interval is too long.

To conclude this discussion on predictive likelihood assessment, I do not think it is enough to judge the value of predictive likelihood by considering only $\mathrm{CI}(\theta)$ or the partial conditional cover $C_\theta(A_a)$. Rather, I regard the fully conditional cover $C_\theta(Y)$ as the basic characteristic of a predictive interval, and contend that $P(C_\theta(Y) \geq \text{nominal level} \mid A_a)$ and $P(C_\theta(Y) \geq \text{nominal level})$ are more relevant criteria for evaluating predictive likelihoods.

The other issue considered by Butler is the motivation behind predictive likelihoods based on conditioning on sufficient statistics. I would like to thank Professor Butler for an interesting and illuminating discussion of the likelihood perspective on assessing the compatibility of z with y and how it motivates the use of L_c in the discrete case. For the continuous case, although the additional Jacobian factor makes the conditional predictive likelihood independent of the choice of minimal sufficient statistic, it does have the unfortunate consequence of making the predictive likelihood not invariant under scale changes of z, which to me seems to be a rather serious defect not shared by L_c.

I would like to conclude this rejoinder by bringing up an apparent deficiency of predictive likelihoods that has not been commented on by the discussants. As mentioned in Section 1, predictive likelihoods are typically not additive, e.g., in the discrete case we usually have $L(z \in B \mid y) \neq \sum_{z \in B} L(z \mid y)$. However, it is possible to modify any predictive likelihood to be additive and such that $L(z \mid y)$ is left unchanged for all z. Moreover, it is the additive version of L that is of interest in practice. Any $L(z \mid y)$ is defined by a functional G on $l_y(z, \theta)$. Consider discrete Y, Z. Then $l_y(z \in B, \theta) = P_\theta(Y = y \cap Z \in B) = \sum_{z \in B} l_y(z, \theta)$. Hence the joint likelihood is additive in z. However, unless G is linear,

$$L(z \in B \mid y) = G\{l_y(z \in B, \theta)\} \neq \sum_{z \in B} G(l_y(z, \theta)).$$

The quantity G may not even be defined on $l_y(z \in B, \theta)$ (e.g. L_c, L_I). Whatever the case may be, we can choose to *define* a modified $L^{(m)}(z \in B \mid y)$, for a given predictive likelihood $G\{l_y(z, \theta)\} = L(z \mid y)$, by

$$L^{(m)}(z \in B \mid y) = \sum_{z \in B} L(z \mid y), \quad \text{if } Z \text{ is discrete}$$

$$= \int_B L(z \mid y)\, dz, \quad \text{if } Z \text{ is continuous.}$$

Since $l_y(z, \theta)$ is additive in z for any given θ, it is certainly a natural requirement that a predictive likelihood also is additive. More importantly, when constructing predictive regions P_y, given by (1.1) in Section 1, we are using directly the additive version $L^{(m)}(z \in B \mid y)$ and not $L(z \in B \mid y)$. Hence, it should be recognized that the predictive likelihood used in constructing confidence regions for z is, in fact, additive and the apparent distinction between Bayes prediction and likelihood prediction in this regard does not exist.

ADDITIONAL REFERENCE

Cox, D. R. (1986). Comment on "Predictive likelihood inference with applications" by R. W. Butler. *J. Roy. Statist. Soc. Ser. B* **48** 1–38.