from the ancillary statistic A_a (as well as the model) while $\hat{\theta}$, a complimentary portion of the data, is being used to assess the accuracy of these coverage functions. In parametric inference the roles of these statistics are reversed in that the ancillary statistic A_a assesses the accuracy of $\hat{\theta}$ in determining the true model. Practical examples are needed to bear out the sensibility of basing recipe choice on coverages at and near $\hat{\theta}$.

Many practical models such as generalized linear models do not admit exact ancillary A_a upon which to condition. In such instances we must find approximate ancillaries as has been done in Hinkley (1980) and Barndorff-Nielsen (1980, 1983).

I do not agree with Bjørnstad's suggestion that $\text{pr}\{\mathbf{Z} \in I_{.9}(\mathbf{Y}); \theta\}$ as an unconditional probability can be used to meaningfully assess the various recipes. Also measuring the worth of an interval (or its associated recipe) by its guarantee of 90% coverage, $\inf_\theta \text{pr}\{C_\theta(\mathbf{Y}) \geq .9\}$ where $C_\theta(y) = \text{pr}\{\mathbf{Z} \in I_{.9}(y) | y; \theta\}$, amounts to a worst case scenario assessment. This could be a very unrepresentative assessment measure to use as a basis for recipe choice.

ADDITIONAL REFERENCES

BARNARD, G. A. (1985). A coherent view of statistical inference. Technical Report, Dept. Statistics, Univ. Waterloo, Waterloo, Ontario, Canada.

BOX, G. E. P. (1980). Sampling and Bayes' inference in scientific modelling and robustness (with discussion). *J. Roy. Statist. Soc. Ser. A* **143** 383–430.

BURRIDGE, J. (1986). Discussion of "Predictive likelihood inference with applications" by R. W. Butler. *J. Roy. Statist. Soc. Ser. B* **48** 29–30.

COX, D. R. (1986). Discussion of "Predictive likelihood inference with applications" by R. W. Butler. *J. Roy. Statist. Soc. Ser. B* **48** 27.

COX, D. R. and HINKLEY, D. V. (1974). *Theoretical Statistics*. Chapman and Hall, New York.

EFRON, B. and HINKLEY, D. V. (1978). Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information (with discussion). *Biometrika* **65** 457–487.

FISHER, R. A. (1934). Two properties of mathematical likelihood. *Proc. Roy. Soc. Ser. A* **144** 285–307.

FISHER, R. A. (1935). The fiducial argument in statistical inferences. *Ann. Eugen.* **6** 391–398.

FISHER, R. A. (1973). *Statistical Methods and Scientific Inference*, 3rd ed. Hafner, New York.

HINKLEY, D. V. (1980). Likelihood as approximate pivotal distribution. *Biometrika* **67** 287–292.

Comment

Tom Leonard, Kam-Wah Tsui and John S. J. Hsu

Professor Bjørnstad is to be congratulated on an excellent review of an important area. Previous statistical practice largely referred to point predictions and estimated standard errors when predicting future observations from current data. When analyzing time series, contingency tables or nonlinear regression models, it is often thought necessary to refer to asymptotics, even to obtain an approximate standard error. However, methods are now available permitting precise predictions based upon finite samples. Moreover, the applied statistician can refer to an entire predictive likelihood or density or probability mass function, summarizing the information in the data about any future observation. This broadens the type of nonlinear model, with several parameters, which may yield useful predictions. These predictions can now be expressed in terms of probability statements, thus enhancing their interpretability, e.g., for noisy data sets.

Tom Leonard is Associate Professor, Kam-Wah Tsui is Professor and John S. J. Hsu is a graduate student, Department of Statistics, University of Wisconsin-Madison, Madison, Wisconsin 53706.

Let $p(\mathbf{y} \mid \boldsymbol{\theta})$ denote our density (or probability mass function) for an $n \times 1$ vector $\mathbf{y}$ of current observations, given a $p \times 1$ vector $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_p)^T$ of unknown parameters, and $p(\mathbf{z} \mid \boldsymbol{\theta})$ represent the corresponding density for an independent $m \times 1$ vector $\mathbf{z}$ of future observations. If $\pi(\boldsymbol{\theta})$ is the prior density of $\boldsymbol{\theta}$, for $\boldsymbol{\theta}$ lying in the parameter space Θ, then the predictive distribution

$$(1) \qquad p(\mathbf{z} \mid \mathbf{y}) = \int_\Theta p(\mathbf{z} \mid \boldsymbol{\theta}) \pi(\boldsymbol{\theta} \mid \mathbf{y})\, d\boldsymbol{\theta}$$

of $\mathbf{z}$ given $\mathbf{y}$ is also representable in the form

$$(2) \qquad p(\mathbf{z} \mid \mathbf{y}) = \frac{p(\mathbf{z} \mid \boldsymbol{\theta}) \pi(\boldsymbol{\theta} \mid \mathbf{y})}{\pi(\boldsymbol{\theta} \mid \mathbf{y}, \mathbf{z})}, \quad \boldsymbol{\theta} \in \Theta.$$

Here we have

$$(3) \qquad \pi(\boldsymbol{\theta} \mid \mathbf{y}) \propto \pi(\boldsymbol{\theta}) p(\mathbf{y} \mid \boldsymbol{\theta}), \quad \boldsymbol{\theta} \in \Theta,$$

denoting the *posterior density* of $\boldsymbol{\theta}$, given $\mathbf{y}$, and

$$(4) \qquad \pi(\boldsymbol{\theta} \mid \mathbf{y}, \mathbf{z}) \propto p(\mathbf{z} \mid \boldsymbol{\theta}) \pi(\boldsymbol{\theta} \mid \mathbf{y}), \quad \boldsymbol{\theta} \in \Theta,$$

denoting the *postposterior density* of $\boldsymbol{\theta}$, given $\mathbf{y}$ and $\mathbf{z}$.

Leonard (1982) was the first to use a multivariate normal approximation to (4) to justify the Laplacian approximation to (1)

$$p^*(\mathbf{z} \mid \mathbf{y}) \propto |\mathbf{R}_Z|^{-1/2} p(\mathbf{z}, \theta_Z \mid \mathbf{y}), \tag{5}$$

where

$$p(\mathbf{z}, \theta \mid \mathbf{y}) = p(\mathbf{z} \mid \theta)\pi(\theta \mid \mathbf{y}) \tag{6}$$

and

$$\mathbf{R}_Z = \left. \frac{-\partial^2 \log p(\mathbf{z}, \theta \mid \mathbf{y})}{\partial(\theta\theta^T)} \right|_{\theta=\theta_z} \tag{7}$$

is the postposterior information matrix, with θ_Z denoting the conditional maximum of (6) with respect to θ for each fixed $\mathbf{z}$ and the observed $\mathbf{y}$. He approximated the postposterior mean vector and covariance matrix by θ_Z and $\mathbf{R}_Z^{-1}$, respectively, in which case (5) is a consequence of replacing θ in (2) by θ_Z. Recent developments and extensions of (5) are summarized by Hsu, Leonard and Tsui (1988, Section 8) and Leonard, Hsu and Tsui (1989), who also consider the important problem of making a marginal inference about a function of several unknown parameters. The tremendous research effort stimulated by the short 1982 note was perhaps unpredictable at the time. We would now like to resummarize and extend the applications of the original idea. Note that if the prior density $\pi(\theta)$ can be specified then the approximation (5) will be unnecessary in situations where importance sampling (e.g., Rubinstein, 1981) can be used to simulate the exact predictive distribution in (1). Let $\tilde{\theta}$ denote the vector maximizing the posterior density $\pi(\theta \mid \mathbf{y})$ and $\tilde{\mathbf{R}}$ denote the corresponding posterior information matrix. Then, in cases where the posterior distribution of θ is roughly multivariate normal, the exact predictive distribution in (2) can be simulated as precisely as possible by

$$p(\mathbf{z} \mid \mathbf{y}) = \lim_{l \to \infty} \sum_{j=1}^{l} \left[\frac{p(\mathbf{z}, \theta_j \mid \mathbf{y})}{\Psi(\theta_j)} \right] \Big/ \sum_{j=1}^{l} \left[\frac{\pi(\theta_j \mid \mathbf{y})}{\Psi(\theta_j)} \right], \tag{8}$$

where $\Psi(\theta)$ denotes a multivariate normal density for θ, with mean vector $\tilde{\theta}$ and covariance matrix $\tilde{\mathbf{R}}^{-1}$ and $\theta_1, \theta_2, \cdots$ represent simulated realizations of θ from $\Psi(\theta)$.

The representation in (2) can be used to refine the approximation to (1) described in (5). Ladalla (1976) and Johnson and Ladalla (1979) propose very accurate approximations, based upon Edgeworth expansions, to posterior mean vectors and covariance matrices. Hence, a range of superior approximations, $\tilde{\theta}_Z$ and $\tilde{\mathbf{R}}_Z^{-1}$ can be developed for the postposterior mean vector and covariance matrix, and these can be used to replace θ_Z and $\mathbf{R}_Z^{-1}$ in (5). More complicated adjustments to (5) can be developed via Edgeworth expansions to the postposterior density in the denominator of (2). We do not regard any refinement to (5) as sensible unless it can be shown to parallel an approximation to $\pi(\theta \mid \mathbf{y}, \mathbf{z})$.

The second term on the right-hand side of (5) corresponds to the profile predictive likelihood. The following counterexample demonstrates that adjustments to this formulation are indeed necessary (see also Bjørnstad's Example 5 of Section 3).

Suppose that (y_{i1}, y_{i2}) are independent for $i = 1, \cdots, n$ and independent of (z_1, z_2), where all observations are normally distributed with common variance τ^2. Let y_{ij} possess mean θ_i, $i = 1, \cdots, n$, and z_j possess mean θ_{n+1}, for $j = 1, 2$. Consider the prediction of $u = z_1 - z_2$. The profile predictive likelihood of u, given only $d_1 = y_{11} - y_{21}, \cdots, d_n = y_{1n} - y_{2n}$, is

$$l_1(u \mid \mathbf{d}) \propto \left[u^2 + \sum_{i=1}^{n} d_i^2 \right]^{-(1/2)(n+1)} \tag{9}$$

which is proportional to a t-density with n degrees of freedom.

This contradicts the less sensible profile predictive likelihood

$$l_2(u \mid \mathbf{y}) \propto \left[u^2 + \sum_{i=1}^{n} d_i^2 \right]^{-(1/2)(2n+1)}$$

of u, given all the y's, which doubles the degrees of freedom. The determinant adjustment in (5), however, still leads to (9), under a suggestion for the linear model with unknown variance which is described below.

We now discuss the choice of prior $\pi(\theta)$. No prior measure, whether proper or improper represents prior ignorance, i.e., there is no such person as an ignorant Bayesian! For example, an improper uniform distribution on p-dimensional real space provides information that θ is equally likely to lie in either of two regions, if these possess the same hypervolume. Therefore (1) and any "predictive likelihood" based upon (1) must depend upon a specification of prior information. Proponents of predictive likelihood would appear to be misleading themselves if they arrive at a specification based on (5), which is apparently free from prior assumptions about θ.

The practical problem remains as to how an applied statistician could construct an appropriate prior for θ. If there is definite prior information then one may wish to think in terms of interdependencies between the θ's. One of the few available ways of reasonably doing this is to firstly transform the parameters in such a manner that a multivariate normal distribution seems reasonable for the new parameters (e.g., Leonard, 1972, 1973, 1975). The multivariate normal covariance matrix then provides a flexible mode for representing interrelationships between the parameters. Natural transformations such as log or logit, may

often be used and frequently coincide with transformations which will improve the multivariate normality of the postposterior distribution (4), and hence the approximation in (5).

As the multivariate normal prior variances approach infinity, the precision of the prior information decreases, and our specification provides a uniform distribution over R^p for the transformed parameters. Therefore, if a uniform distribution is to be employed, we recommend first transforming the θ's to satisfy our notions of multivariate normality. This leads to the approximation in (5) but with $p(\mathbf{z}, \boldsymbol{\theta} \mid \mathbf{y})$ in (5) and (7) replaced by

$$p(\mathbf{z}, \mathbf{y} \mid \boldsymbol{\theta}) = p(\mathbf{z} \mid \boldsymbol{\theta}) p(\mathbf{y} \mid \boldsymbol{\theta}), \tag{10}$$

thus satisfying Butler's concept of a "predictive likelihood." The dependence on the parametrization, which seems to be of immense concern to Professor Butler, can be greatly reduced by seeking a parametrization which ensures good multivariate normality.

As an example, suppose that $\mathbf{y}$ and $\mathbf{z}$ possess independent multivariate normal distributions with respective mean vectors $\mathbf{X}_1\boldsymbol{\beta}$ and $\mathbf{X}_2\boldsymbol{\beta}$, zero covariances, and all variances equal to τ^2, where $\mathbf{X}_1$ and $\mathbf{X}_2$ are specified $n \times q$ and $m \times q$ matrices. Then, under a uniform prior for the normalizing transformation $\boldsymbol{\theta} = (\boldsymbol{\beta}, \log \tau^2)$, Leonard's approximation gives predictive distributions for the z's which correctly relate to the t-distribution with $n - q$ degrees of freedom, and predictive distributions for sums of squares which correctly relate to the chi-squared distribution with $n - q$ degrees of freedom. None of Butler's differing suggestions match all these degrees of freedom precisely, and his approximations will therefore possess inferior frequency properties in this important case.

However, if p is moderately large, it can be dangerous to use a uniform prior for any parametrization of $\boldsymbol{\theta}$ since this is likely to lead to vastly inferior frequency properties for both point estimates of θ's and point predictions of the z's, owing to the Stein-effect (e.g., Ghosh, Hwang and Tsui, 1983; Morris, 1983). A more promising approach involves an exchangeable prior distribution. Following Leonard (1976), assume that appropriate transformations have been performed on the $\boldsymbol{\theta}$ vector (or the $\boldsymbol{\beta}$ vector in the above linear model example) to ensure "prior white noise," i.e., reasonability of the prior assumption that, given μ and σ^2, the elements of the new $\boldsymbol{\theta}$ are independent and normally distributed with common mean μ and variance σ^2, where μ and σ^2 are independent and uniformly distributed over their ranges of possible values.

In hierarchical models of this type (see also Leonard and Novick, 1986; Albert, 1988), the postposterior distribution of $(\boldsymbol{\theta}, \mu, \sigma^2)$ can seldom be transformed to ensure approximate multivariate normality, owing to tedious dependencies between σ^2 and $\boldsymbol{\theta}$. It is more accurate to condition on σ^2 first, with μ integrated out from the prior. The predictive distribution of $\mathbf{z}$, given $\mathbf{y}$ and σ^2, can be approximated by $p^*(\mathbf{z} \mid \mathbf{y}, \sigma^2)$, which replaces $p(\mathbf{z}, \boldsymbol{\theta} \mid \mathbf{y})$ in (5) and (7) by

$$\begin{aligned} &p^*(\mathbf{z}, \boldsymbol{\theta} \mid \mathbf{y}, \sigma^2) \\ &\quad \propto (\sigma^2)^{-(1/2)(p-1)} p(\mathbf{z} \mid \boldsymbol{\theta}) p(\mathbf{y} \mid \boldsymbol{\theta}) \\ &\quad \times \exp\left\{-\frac{\sum_{i=1}^{p} (\theta_i - \theta)^2}{2\sigma^2}\right\} \end{aligned} \tag{11}$$

with θ denoting the average θ_i.

The posterior distribution of σ^2, given $\mathbf{y}$, may be replaced by the Laplacian approximation

$$\pi^*(\sigma^2 \mid \mathbf{y}) \propto |\mathbf{R}_\sigma|^{-1/2} \sup_{\boldsymbol{\theta}} \pi(\boldsymbol{\theta}, \sigma^2 \mid \mathbf{y}), \tag{12}$$

where

$$\begin{aligned} &\pi(\boldsymbol{\theta}, \sigma^2 \mid \mathbf{y}) \\ &\quad \propto (\sigma^2)^{-(1/2)(p-1)} p(\mathbf{y} \mid \boldsymbol{\theta}) \exp\left\{-\frac{\sum_{i=1}^{p} (\theta_i - \theta)^2}{2\sigma^2}\right\} \end{aligned} \tag{13}$$

and

$$\mathbf{R}_\sigma = \left. \frac{-\partial^2 \log \pi(\boldsymbol{\theta}, \sigma^2 \mid \mathbf{y})}{\partial(\boldsymbol{\theta}\boldsymbol{\theta}^T)} \right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_\sigma} \tag{14}$$

with $\boldsymbol{\theta}_\sigma$ denoting the vector maximizing (11) for fixed σ^2. The unconditional predictive distribution of $\mathbf{Z}$, given $\mathbf{y}$, may now be approximately computed by the one-dimensional numerical integration

$$p^*(\mathbf{z} \mid \mathbf{y}) = \int_0^\infty p^*(\mathbf{z} \mid \mathbf{y}, \sigma^2) \pi^*(\sigma^2 \mid \mathbf{y})\, d\sigma^2. \tag{15}$$

The possibility in (15) provides the main suggestion of this comment. It is likely to lead to more stable and reasonable predictions than using (5) with a uniform prior. Extensions of this idea are available which consider several exchangeable subsets of the $\boldsymbol{\theta}$ vector.

We now address the more tedious problem of approximating the marginal distribution, conditional on $\mathbf{y}$, of a real-valued function

$$t = g(\mathbf{u}) \tag{16}$$

of $\mathbf{u} = (\boldsymbol{\theta}, \mathbf{z})^T$. If t does not depend upon $\boldsymbol{\theta}$, then our problem reduces to approximating the predictive distribution, given $\mathbf{y}$, of this summary statistic of the future observations. We in general refer to

$$p(\mathbf{u} \mid \mathbf{y}) = p(\mathbf{z}, \boldsymbol{\theta} \mid \mathbf{y}) = p(\mathbf{z} \mid \boldsymbol{\theta}) \pi(\boldsymbol{\theta} \mid \mathbf{y}) \tag{17}$$

and the procedure due to Leonard, Hsu and Tsui (1989) for approximating the marginal posterior density of a function of several parameters. This seems to be more accurate in a number of special cases than a rather tentative modification to Leonard's (1982) procedure suggested by Kass, Tierney and Kadane

(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 θ 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,