is a point in R^2 whose expectation is $x(\lambda_j)$. Consequently,

$$h(R_\phi^{-1}(y_j - \omega)/\rho) - 1$$

has zero mean for each $j = 1, \ldots, n$, and nonzero mean off the template. These n elementary estimating functions can be combined linearly in an optimal manner to obtain estimates of the parameters of interest without involving the nuisance parameters.

It would be of considerable interest to know whether the preceding method can be extended in useful ways, possibly to nonlinear distortions of templates.

Comment: Alternative Aspects of Conditional Inference

George Casella, Thomas J. DiCiccio and Martin T. Wells

The roles of conditioning in inference are almost too varied to be summarized in one paper. Professor Reid has done a wonderful job of explaining and illustrating some of these roles. We expand on a number of her points, with particular attention to the practical uses and implementation of the methods. We also discuss some overall goals of conditional inference and alternative ways of achieving them.

1. INTRODUCTION

The techniques of conditional inference are a collection of extremely powerful tools. They allow for the construction of procedures with extraordinarily good properties, especially in terms of frequentist asymptotic behavior. In fact, in many cases these procedures are so good that one begins to wonder why they are not more widely used; that is, although statistics methodology journals often contain articles on conditional inference, such techniques have not really found their way into the arsenal of the applied statistician and thus into the subject matter journals. There are, we feel, two reasons for this. One is that, unfortunately, the procedures are fairly complex in their derivation and, hence, in their implementation, and for that reason alone they may not have received thorough consideration. The second reason is somewhat more subtle, but perhaps more important. If an experimenter uses conditional inference techniques, the goal of the analysis (and the exact type of ultimate inference to be made) is not at all clear. In Section 1, Reid recounts four roles of conditional inference that are identified by Cox (1988). However, to a prospective user of these techniques, these goals are vague, and the effort needed to actually implement these solutions can be prohibitive. For example, consider Example 3.3, used to illustrate conditional inference techniques in the estimation of the gamma shape parameter when the scale parameter is unknown. The density given by (3.3) and (3.5), which contain components that are "difficult to calculate," is offered as a conditional inference solution to the problem. This density can be used to test an hypothesis or, with some difficulty, to calculate a confidence interval, but the details of carrying out these procedures are quite complex. Moreover, if one is interested in a point estimate and evaluation of the performance of the estimate, this density will not suffice. Rather, one might use a saddlepoint approximation (Reid, 1988) for the density of the maximum likelihood estimate, yielding a density proportional to

$$\Gamma^n(\hat{\psi})\Gamma^n(\psi)\{\hat{\psi}\Delta'(\hat{\psi}) - 1\}^{1/2} \cdot \exp[n\{(\hat{\psi} - \psi)\Delta(\hat{\psi}) + \hat{\psi} - \psi \ln \hat{\psi}\}],$$

where $\Delta(\cdot)$ is the digamma function. Although the approximation is remarkably accurate, computation of the normalizing constant (which involves integrating this function with respect to $\hat{\psi}$) is quite demanding, limiting the use of the formula. Thus, the "naive" user is shortchanged. Rather than the accurate approximations and, hence, more precise inference, the user gets only halfway there and can be faced with calculations of prohibitive complexity.

George Casella is Professor, Biometrics Unit, and Thomas J. DiCiccio and Martin T. Wells are Associate Professors, Department of Social Statistics, Cornell University, Ithaca, New York 14853.

The failing is that conditional inference has been developed by the *cognoscenti* for their use, and to these experts the problems of this paragraph are easy to surmount.

We are not presented with any unifying idea or goal that is the core of conditional inference. Instead, what is presented are "techniques" of conditional inference, but not a comprehensive "theory" of conditional inference. Such a theory, of course, exists. Indeed, there is a rich theory. But, to the advancement of powerful problem-solving techniques, this development has been neglected. For example, referring to Cox's four roles, we are not guided as to what probability calculations are relevant. How does one measure the lost information in order to know how much has been recovered? How does one measure the influence of the nuisance parameters, so that a reasonable factorization can be decided upon? Last, and most important, what should we do with that extremely accurate density approximation? These comments are not meant to be adversarial or confrontational, but rather they are meant to highlight areas that we believe need to be developed in order for the theory of conditional inference to obtain the widespread use it deserves.

In our discussion we will focus on three main topics, which are all aimed at clarifying understanding and aiding practical applications of theory and methods of conditional inference and allied techniques. We first discuss the elimination of nuisance parameters and different options for obtaining a reasonable density on which to base inference. We then consider some more practical aspects and discuss methods that aid implementation of conditional solutions. Finally, we describe a Bayesian–frequentist synthesis, first illustrating Bayesian methods for computation and inference with conditional techniques and then showing how conditional inference techniques are useful in the construction of Bayes procedures having good frequentist properties.

2. ELIMINATION OF NUISANCE PARAMETERS

It seems to us that a major, and extremely desirable, goal of conditional inference can be stated as the accurate approximation of a likelihood function of the parameter of interest, free of nuisance parameters. Such likelihoods can be obtained through the modified profile likelihood of Barndorff-Nielsen (1983) or the conditional profile likelihood of Cox and Reid (1987).

These likelihoods are often obtained through delicate expansions and substitutions, sometimes resulting in formulas that are extremely difficult to understand and interpret. Moreover, the exact implementation of these methods is not straightforward as there does not seem to be an overall "recipe." For example, the exact degree of nuisance parameter elimination is tied to the type of density factorization possible, such as (3.1), (3.2) or (3.6). Although such factorizations often can be recognized, what concerns us is that the implementation of nuisance parameter elimination is hard to characterize. Example 3.1 uses a factorization that conditions one part of a sufficient statistic on another; Example 3.3 is similar, but uses a different form of the sufficient statistic. Example 3.6 (see also Example 5.3) seems to take advantage of the pivotal structure of the problem, and its implementation is also equivalent to integration of the parameters according to a Haar measure prior. Thus, there is a great opportunity for a naive user to be bewildered about implementation.

To us, this is a perfect illustration of the need for synthesis in statistics. The Bayesian paradigm is perfectly suited for elimination of nuisance parameters, and can leave one with a density (actually a posterior density) that only depends on the parameters of interest. The methodology is straightforward and completely general. One merely specifies a prior for the nuisance parameter and then integrates to get the desired density (similar to Example 3.6, where the location–scale structure naturally suggests a Haar measure prior). The actual choice of the prior, while often of concern in theory, is somewhat less of a concern in practice, as typical "flat" or "default" priors (Berger and Bernardo, 1992; Clarke and Wasserman, 1993) will lead to reasonable frequentist inferences (as illustrated in Strawderman, Casella and Wells, 1995). Indeed, working a little harder on the prior can often yield extremely interesting results.

3. PRACTICAL INFERENCE

We are somewhat disappointed in the examples in Professor Reid's paper, as they tend to be more stylized than practical. This is particularly unfortunate since these methods can be extremely useful in practical situations, and it is important to highlight this point.

Perhaps this stylization of examples reflects what can be perceived as a misplaced emphasis in the development of the conditional inference theory. The development of approximations of densities, distribution functions and likelihoods has been for models derived from statistical theory rather than models derived from the concerns of experimenters. The somewhat related topic of saddlepoint approxima-

tions has been more successful in making headway into practical solutions.

3.1 The Saddlepoint Alternative

Although there is a fundamental difference between the inference from the saddlepoint approach and that from the conditional inference approach, much of the mechanics is so similar that a comparison is almost required. Moreover, in many important exponential family models, the marginal inferences that result from saddlepoint approximations are equivalent to conditional inferences.

There is, however, a distinct difference between the saddlepoint and p^* approaches in their implementation, resulting in an advantage to the saddlepoint approach. A role of the ancillary statistic in the p^*-formula (2.1) is to separate the data into the maximum likelihood estimate and the ancillary, so the p^*-formula can actually be used as a density for the maximum likelihood estimate. (The more fundamental role of the ancillary is to reduce the dimension of the sufficient statistic to that of the parameter of interest, putting the problem more in the form of an exponential family. However, the "separation" role is quite important in the mechanics of the implementation.) For example, getting from (2.1) to (2.2) uses the identity

$$x - \mu = (\hat{\mu} - \mu) + (x - \hat{\mu}) = \hat{\mu} - \mu + a,$$

where $\hat{\mu}$ is the maximum likelihood estimator and a is an ancillary statistic. Such a decomposition is transparent in the location–scale case, but becomes less so in other cases, where implementation of the p^*-formula can necessitate involved calculations of approximate ancillaries.

This problem is not shared by the saddlepoint approximation, and the reason is, perhaps, most evident from the exponential-tilting derivation of the saddlepoint. The auxiliary random variable used to center the approximation is added to the mix, and hence is always separate from the statistics of interest. Thus, no separation or factorization is required to obtain the desired density approximation, making the saddlepoint somewhat more accessible as an approximation technique in complicated problems. Alternatively, and equivalently, the saddlepoint approximation can also start from an estimating equation, derived from the experimenter's model of interest. To illustrate the differences in these approaches, consider the important practical case of logistic regression, in particular, logistic regression with one covariate and an unknown intercept,

$$\begin{aligned} Y_i &\sim \text{Bernoulli}(p_i(\theta)), \\ p_i(\theta) &= [1 + \exp(-(\theta_0 + \theta_1 x_i))]^{-1}. \end{aligned} \tag{1}$$

From a conditional inference point of view, estimation of the coefficient density was considered by Barndorff-Nielsen and Cox (1979) and Davison (1988), who used double-saddlepoint (numerator and denominator) approximations to approximate the conditional density of the regression coefficients. Their delicate approximations eliminate nuisance parameters by conditioning and result in answers that are quite difficult to compute. The logistic model can also be directly attacked with saddlepoints starting from estimating equations (see, e.g., Field and Ronchetti, 1990), as done in Strawderman, Casella and Wells (1995). (Note that we are now invoking the discussant's privilege: talk about your own work.) The cumulant generating function for n independent observations from (1) is

$$\begin{aligned} K_n(t|s, \theta) = \sum_{i=1}^{n} &\log[1 - p_i(\theta) + p_i(\theta) \exp t' z_i] \\ &- t' z_i p_i(s), \end{aligned}$$

where t, s and θ are 2×1 vectors, $z_i = (1, x_i)'$, the two-dimensional saddlepoint is $W_0 = n^{-1}(s - \theta)$ and the approximate density of $\hat{\theta}$ is given by

$$\begin{aligned} g_{\hat{\theta}}(s|\theta) = &\frac{\left|\sum_{i=1}^{n} z_i z_i' p_i(s)(1 - p_i(s))\right|^{1/2}}{2\pi} \\ &\cdot \exp K_n(s - \theta \mid s, \theta), \end{aligned} \tag{2}$$

where $|\cdot|$ denotes the determinant. A similar formula holds for parameter vectors of fixed but arbitrary dimension. Approximation (2) has been shown to be extremely accurate, even for samples as small as $n = 20$. Note that (2) is a marginal, not a conditional, density, which illustrates an essential difference in the saddlepoint versus conditional inference methodologies. However, as we shall see in Section 4.1, we can use some simple computing techniques (such as the Gibbs sampler) to allow us to derive conditional-type inferences from (2).

In the case of exponential families, in which ordinary logistic regression falls, the saddlepoint and conditional inference (double saddlepoint) approaches yield equivalent answers. Moreover, the estimating equation approach is somewhat more general, allowing us to apply approximations such as (2) to more general situations.

3.2 Computing the Conditional Solution

Professor Reid mentions that computing methods, such as Monte Carlo Markov chain, have exploited the fact that conditional solutions are often easier to calculate than marginal densities. However, the conditional solutions from the p^*-formula or a saddlepoint approximation can, themselves, be

extremely complicated. This becomes particularly apparent in the problem of confidence interval construction, which often necessitates many evaluations of the target density (and its constant), with each evaluation at a different parameter point.

This suggests another synthesis. Rather than just note that conditional densities can make techniques such as Monte Carlo Markov chain easier to implement, it is also the case that these computing techniques can eliminate much of the delicate approximation needed to use either the p^*-formula or a saddlepoint approximation. For example, consider the linear models in Examples 3.7 and 5.3, perhaps still the most common models in statistics. They can be analyzed easily and effectively using computational methods.

The special structure of these linear regression examples, which results in the existence of a useful pivot, makes it easy to apply some now-standard computing methods. The general expression for the density of the pivot statistics $\mathbf{t} = (t_1, \ldots, t_p)$ and v, where

$$t_1 = \frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}}, \ldots, \quad t_p = \frac{\hat{\beta}_p - \beta_p}{\hat{\sigma}}, \quad v = \frac{\hat{\sigma}}{\sigma},$$

conditional on the ancillary statistics $a_i(\mathbf{y}) = (y_i - X\hat{\beta}_i)/\hat{\sigma}$, has the specific form (Fraser, 1979, Chapter 6; Fraser, Lee and Reid, 1990)

$$(3) \qquad \frac{f_0[(X\mathbf{t}+\mathbf{a})v]v^{n-1}|X'X|^{1/2}}{\iint f_0[(X\mathbf{t}+\mathbf{a})v]v^{n-1}|X'X|^{1/2}\,d\mathbf{t}\,dv}.$$

For any density f_0, we can draw samples from (3) using a method such as the accept–reject or Metropolis algorithms, or perhaps the Gibbs sampler (see Tanner, 1993, or Robert, 1994). For example, if interest is in the marginal distribution of t_k, we draw m realizations from (3), say, $\{(t_k^{(j)}, \mathbf{t}_{-k}^{(j)}), v^{(j)}\}_{j=1}^m$. Then we can calculate a Monte Carlo approximation to the marginal density of t_k,

$$(4) \qquad \hat{f}(t) = \frac{1}{m}\sum_{j=1}^{m} \frac{\phi(t_k^{(j)}|\mathbf{t}_{-k}^{(j)}, v^{(j)})f(t, \mathbf{t}_{-k}^{(j)}, v^{(j)})}{f(t_k^{(j)}, \mathbf{t}_{-k}^{(j)}, v^{(j)})},$$

where $\phi(\cdot|\cdot)$ is any conditional density. If $f(t)$ is the true marginal, then for any such density $\phi(\cdot|\cdot)$ we have $\hat{f}(t) \to f(t)$ at a geometric rate as $m \to \infty$, and the convergence will be faster the closer ϕ resembles the true conditional density. This type of computing is a quite viable alternative to the marginalization asymptotics of Section 4.3, as expression (4) is exact to any degree of accuracy desired.

Details of the application of this technique, as well as a number of examples, are given in Casella, Wells and Tanner (1994). Bayesian applications of Monte Carlo marginalization are given in Gelfand, Smith and Lee (1991) and Chen (1994). The general theory of Monte Carlo marginalization is a consequence of the conditional Monte Carlo method, nicely explained in Hammersley and Handscomb (1964). Thus, the easy computer implementation that Professor Reid mentions in her concluding remarks may already exist in another guise within a slightly different inferential framework.

4. THE BAYESIAN CONNECTION

4.1 A Bayesian Solution

In Section 2 we discussed the Bayesian solution to marginalization and how well suited it is to general problems. When that is combined with the computational techniques described in the previous section, a powerful tool for calculation of marginal densities emerges. In particular, the Monte Carlo marginalization (4) can be combined with the logistic saddlepoint density (2) to obtain the marginal density of interest. This strategy can also circumvent the problem of inference with nonorthogonal parameters.

Specifically from the saddlepoint density (2), or its higher-dimensional analog, it is straightforward to marginalize to the univariate density of any one coefficient. However, in contrast to a normal approximation, the marginal saddlepoint density for each parameter (obtained by integrating out the remaining variables) depends upon the true values of all of the parameters. For example, if the parameter vector is $(\theta_0, \theta_1, \theta_2, \theta_3)$ and interest is in making inferences about θ_2, the marginal density for $\hat{\theta}_2$ is not immediately useful since it will depend upon θ_2 as well as θ_0, θ_1 and θ_3. Thus, simply using numerical integration to marginalize the saddlepoint density is not recommended unless it is known that there is parameter orthogonality.

Using a Bayesian approach, a marginal posterior distribution (which is similar to a conditional density) can be calculated for the parameter of interest, and this marginal behaves quite nicely under frequentist evaluations. In the above illustration, placing a uniform improper prior on each of θ_0, θ_1, θ_2 and θ_3 yields a posterior density of the parameters, given the data, that is proportional to the saddlepoint density of the MLE's. To obtain the marginal posterior density for each parameter, one can first apply the Gibbs sampler (or other sampling scheme) to obtain observations from the joint posterior density, and use Monte Carlo marginalization to obtain the desired result. Strawderman, Casella and Wells (1995) have had success with this method in the generalized linear model setting and have found that the Bayesian HPD regions maintained reasonable frequentist coverage. Indeed, with the use of

typical flat priors, adequate frequentist performance of Bayes procedures is to be expected. However, one might hope to improve, and this leads naturally to the question of the existence of priors that could yield even better frequentist performance.

4.2 Bayesian–Frequentist Conditional Inference

The fact that the Bayes intervals of the previous section were also good frequentist intervals is not a coincidence. Underlying the theory of conditional inference are structures that are common to both paradigms. As mentioned in Section 6.3 and reviewed in Reid (1995), these similarities have been explored and exploited previously to understand the types of Bayesian priors that are likely to result in good frequentist inference. We would like to explore further the connection between conditional and Bayesian inference, in particular using asymptotic tail probability approximations to help identify noninformative priors.

Suppose that Z is a continuous random variable having probability density function of the form

$$f_Z(z) \propto b(z)\exp\{k(z)\} = \exp\{k(z) + \log b(z)\},$$

where $k(z)$ and its derivatives are of order $O(n)$, the derivatives of $\log b(z)$ are of order $O(1)$ and $k(z)$ is maximized at $\hat{z}$, so that $Z - \hat{z}$ is of order $O_p(n^{-1/2})$. DiCiccio and Martin (1991) showed

$$\text{pr}(Z \geq z) = \Phi(r) + \varphi(r)(r^{-1} - v^{-1}) + O(n^{-3/2}), \tag{5}$$

where $\hat{z} - z$ is assumed to be of order $O(n^{-1/2})$,

$$r = \text{sgn}(\hat{z} - z)\big[2\{k(\hat{z}) - k(z)\}\big]^{1/2},$$
$$v = \frac{k^{(1)}(z)}{\{-k^{(2)}(\hat{z})\}^{1/2}}\frac{b(\hat{z})}{b(z)},$$

and $k^{(j)}(z) = d^j k(z)/dz^j$, $j = 1, 2$. Now consider Bayesian inference for $\theta = (\psi, \lambda)$ based on an observed random vector $Y = (Y_1, \ldots, Y_n)$ and a prior probability density function $\pi(\theta)$. The Tierney, Kass and Kadane (1989) Laplace approximation to the marginal posterior density of ψ is

$$\pi_{\psi|Y}(\psi) \propto \pi(\psi, \hat{\lambda}_\psi)\big|j_{\lambda\lambda}(\psi, \hat{\lambda}_\psi)\big|^{-1/2} \cdot \exp\{l(\psi, \hat{\lambda}_\psi)\}, \tag{6}$$

where $l(\theta)$ is the log-likelihood function for θ based on Y and $j_{\lambda\lambda}(\theta) = -l_{\lambda\lambda}(\theta)$. When approximation (6) is normalized, it has relative error of order $O(n^{-3/2})$. An asymptotic expression for posterior tail probabilities of ψ can be obtained by applying formula (5) to (6). There are two obvious ways to proceed: either choose

$$b(\psi) = \pi(\psi, \hat{\lambda}_\psi)\big|j_{\lambda\lambda}(\psi, \hat{\lambda}_\psi)\big|^{-1/2},$$
$$k(\psi) = l(\psi, \hat{\lambda}_\psi), \tag{7}$$

or else choose

$$b(\psi) = 1,$$
$$k(\psi) = l(\psi, \hat{\lambda}_\psi) - \tfrac{1}{2}\log\big|j_{\lambda\lambda}(\psi, \hat{\lambda}_\psi)\big|^{-1/2} - \log\pi(\psi, \hat{\lambda}_\psi). \tag{8}$$

Choices (7) and (8) produce, respectively, approximations of the "double saddlepoint" and the "sequential saddlepoint" form, which are mentioned by Professor Reid in Section 4.3. Numerical investigations show that (8) generally produces more accurate approximations than (7); moreover, in practice, numerical integration of the Laplace approximation (6) might be feasible and extremely accurate. However, choice (7) is preferable for the purpose of identifying noninformative priors. Using (7) in conjunction with approximation (5) yields, for values ψ_0 such that $\hat{\psi} - \psi_0$ is of order $O(n^{-1/2})$, the tail probability approximation

$$\text{pr}(\psi \geq \psi_0|Y) = \Phi(r_p) + \varphi(r_p)(r_p^{-1} - v_p^{-1}) + O(n^{-3/2}), \tag{9}$$

where $r_p = \text{sgn}(\hat{\psi} - \psi_0)\big[2\{l(\hat{\psi}, \hat{\lambda}) - l(\psi_0, \hat{\lambda}_0)\}\big]^{1/2}$ is the signed root of the likelihood ratio statistic, $\hat{\lambda}_0 = \hat{\lambda}_{\psi_0}$,

$$v_p = l_\psi(\psi_0, \hat{\lambda}_0)\frac{\big|j_{\lambda\lambda}(\psi_0, \hat{\lambda}_0)\big|^{1/2}}{\big|j_{\theta\theta}(\hat{\psi}, \hat{\lambda})\big|^{1/2}}\frac{\pi(\hat{\psi}, \hat{\lambda})}{\pi(\psi_0, \hat{\lambda}_0)},$$

and $l_\psi(\psi, \lambda) = \partial l(\psi, \lambda)/\partial\psi$. Thus, the value of ψ_0 satisfying $\phi(r_p) + \varphi(r_p)(r_p^{-1} + v_p^{-1}) = \alpha$ agrees with the posterior $1 - \alpha$ quantile of ψ to error of order $O(n^{-2})$.

On the other hand, from a frequentist perspective, Barndorff-Nielsen (1986, 1991) has shown that the standard normal approximation to the conditional distribution of $r_p^* = r_p + r_p^{-1}\log(u_p/r_p)$, given an exact or approximate ancillary statistic t, has error of order $O(n^{-3/2})$, where ψ_0 now denotes the true value of the parameter of interest,

$$u_p = \frac{\big|l_{;\hat{\theta}}(\hat{\psi}, \hat{\lambda}) - l_{;\hat{\theta}}(\psi_0, \hat{\lambda}_0) \quad l_{\lambda;\hat{\theta}}(\psi_0, \hat{\lambda}_0)\big|}{\big\{\big|j_{\lambda\lambda}(\psi_0, \hat{\lambda}_0)\big|\,\big|j_{\theta\theta}(\hat{\psi}, \hat{\lambda})\big|\big\}^{1/2}},$$

where $l_{;\hat{\theta}}(\psi, \lambda)$ is the column vector of partial derivatives of $l(\psi, \lambda; \hat{\theta}, t)$ taken with respect to $\hat{\theta}$; $l_{\lambda;\hat{\theta}}(\psi, \lambda)$ is the matrix of second-order partial derivatives of $l(\psi, \lambda; \hat{\theta}, t)$ taken with respect to λ and $\hat{\theta}$; and u_p takes the same sign as r_p. Hence,

the value of ψ_0 that satisfies $\Phi(r_p^*) = \alpha$ is an approximate $1-\alpha$ confidence limit having conditional coverage error of order $O(n^{-3/2})$, as is the value of ψ_0 such that $\Phi(r_p) + \varphi(r_p)(r_p^{-1} - u_p^{-1}) = \alpha$, since it can be shown that

$$\Phi(r_p^*) = \Phi(r_p) + \varphi(r_p)(r_p^{-1} - u_p^{-1}) + O(n^{-3/2}). \tag{10}$$

Approximation (10) generalizes (2.3) to the nuisance parameter case and produces (4.3) for canonical parameters of linear exponential families.

One notion of a noninformative prior is that the posterior $1-\alpha$ quantile of ψ is, under repeated sampling, an approximate upper $1-\alpha$ confidence limit having coverage error of order $O(n^{-1})$. Writing $\theta = (\theta^1, \ldots, \theta^d)$, with $\psi = \theta^1$, Peers (1965) showed that to be noninformative the prior $\pi(\theta)$ must satisfy the equation

$$\sum_{j=1}^{d} i^{1j}(i^{11})^{-1/2}\frac{\partial}{\partial\theta^j}\{\log\pi(\theta)\} + \sum_{j=1}^{d}\frac{\partial}{\partial\theta^j}\{i^{1j}(i^{11})^{-1/2}\} = 0, \tag{11}$$

where $i_{ij} = E\{-\partial^2 l(\theta)/\partial\theta^i\,\partial\theta^j\}$ and (i^{ij}) is the matrix inverse of (i_{ij}). Tibshirani (1989) noted that, when ψ and λ are orthogonal, (11) reduces to

$$(i_{\psi\psi})^{-1/2}\frac{\partial}{\partial\psi}\{\log\pi(\theta)\} + \frac{\partial}{\partial\psi}(i_{\psi\psi})^{-1/2} = 0,$$

which has solutions of the form

$$\pi(\psi, \lambda) \propto \{i_{\psi\psi}(\psi, \lambda)\}^{1/2} g(\lambda), \tag{12}$$

where $g(\lambda)$ is arbitrary. DiCiccio and Martin (1993) showed that if $\pi(\theta)$ is noninformative, then $v_p = u_p + O_p(n^{-1})$ in the repeated sampling sense. Thus, the Bayesian approximate confidence limits agree with the limits from (10) to error of order $O_p(n^{-3/2})$, and it follows that the Bayesian limits have *conditional* coverage error of order $O(n^{-1})$ given exact or approximate ancillary statistics.

Equation (11) does not have a unique solution, and it is natural to ask whether solutions can be identified for which the coverage error of the approximate limits is of order $O(n^{-3/2})$. This improved coverage accuracy holds if the prior is such that $v_p = u_p + O_p(n^{-3/2})$, that is, if $\pi(\psi, \lambda)$ satisfies

$$\frac{\pi(\psi_0, \hat\lambda_0)}{\pi(\hat\psi, \hat\lambda)} = l_\psi(\psi_0, \hat\lambda_0)\frac{|j_{\lambda\lambda}(\psi_0, \hat\lambda_0)|^{1/2}}{|j_{\theta\theta}(\hat\psi, \hat\lambda)|^{1/2}}(u_p)^{-1}, \tag{13}$$

to error of order $O_p(n^{-3/2})$.

The use of (11) can be illustrated in the gamma model discussed in Examples 3.3 and 5.2. Consider a sample $Y_1, \ldots, Y_n$ from the distribution having density

$$f(y; \mu, \nu) = \left\{\frac{(\nu/\mu)^\nu}{\Gamma(\nu)}\right\} y^{\nu-1}\exp\left\{-\left(\frac{\nu}{\mu}\right)y\right\}, \quad y > 0,$$

where μ is the mean, ν is the shape parameter and μ and ν are orthogonal. Suppose that μ is the parameter of interest and ν is the nuisance parameter. Since $i_{\mu\mu} = n\nu/\mu^2$, expression (12) shows that the noninformative priors are of the form

$$\pi(\mu, \nu) \propto \frac{g(\nu)}{\mu}, \tag{14}$$

where $g(\nu)$ is arbitrary. For this problem, Barndorff-Nielsen (1986) showed that

$$u_p = n^{1/2}\frac{(\hat\mu - \mu_0)}{\mu_0}\hat\nu^{1/2}\left\{\frac{\phi^{(1)}(\hat\nu)}{\phi^{(1)}(\hat\nu_0)}\right\}^{1/2},$$

where $\phi(\nu) = d\log\Gamma(\nu)/d\nu - \log\nu$ and $\phi^{(1)}(\nu) = d^2\log\Gamma(\nu)/d\nu^2 - 1/\nu$. In this case, expression (13) becomes

$$\frac{\pi(\mu_0, \hat\nu_0)}{\pi(\hat\mu, \hat\nu)} = \frac{\hat\mu}{\mu_0}\frac{\hat\nu_0}{\hat\nu}\left\{\frac{\phi^{(1)}(\hat\nu_0)}{\phi^{(1)}(\hat\nu)}\right\}^{1/2},$$

which recommends the prior $\pi(\mu, \nu) = \nu\phi^{(1)}(\nu)/\mu$, corresponding to the choice $g(\nu) = \nu\phi^{(1)}(\nu)$ in (11). The situation is not so clear when ν is the parameter of interest and μ is the nuisance parameter. Since $i_{\nu\nu} = n\phi^{(1)}(\nu)$, noninformative priors are of the form

$$\pi(\nu, \mu) \propto g(\mu)\{\phi^{(1)}(\nu)\}^{1/2}, \tag{15}$$

where $g(\mu)$ is arbitrary. In this case,

$$u_p = n^{1/2}(\hat\nu - \nu_0)\left(\frac{\nu_0}{\hat\nu}\right)^{1/2}\{\phi^{(1)}(\hat\nu)\}^{1/2},$$

and by (13), the prior $\pi(\nu, \mu)$ should be chosen so that, to error of order $O_p(n^{-3/2})$,

$$\frac{\pi(\nu_0, \hat\mu_0)}{\pi(\hat\nu, \hat\mu)} = \frac{\phi(\hat\nu) - \phi(\nu_0)}{(\hat\nu - \nu_0)\phi^{(1)}(\hat\nu)} = 1 + \frac{1}{2}(\nu_0 - \hat\nu)\frac{\phi^{(2)}(\hat\nu)}{\phi^{(1)}(\hat\nu)} + \frac{1}{6}(\nu_0 - \hat\nu)^2\frac{\phi^{(3)}(\hat\nu)}{\phi^{(1)}(\hat\nu)}. \tag{16}$$

Since $\hat\mu_0 = \hat\mu$ in this case, there is no loss in restricting attention to priors $\pi(\nu, \mu)$ that are functions of ν alone. An easy Taylor expansion shows that the choice $\pi(\nu, \mu) \propto \{\phi^{(1)}(\nu)\}^{1/2}$ satisfies (16) only to error of order $O(n^{-1})$. Moreover, the reference prior $\pi(\nu, \mu) \propto \mu\{\phi^{(1)}(\nu)\}^{1/2}$ considered by Liseo (1993) also satisfies (16) only to error of order $O_p(n^{-1})$.

Expression (16) suggests that it might be impossible to find a prior density that produces confidence limits having coverage error of order $O(n^{-3/2})$; see DiCiccio, Keller and Martin (1992).

Many of the likelihood adjustments and distributional corrections discussed in the paper can be viewed, at least to error of order $O_p(n^{-1})$, in terms of the quantities z_0 and a that arise in Efron's (1987) BC_a confidence limits. Efron defined $z_0 = \Phi^{-1}\{\mathrm{pr}(\hat{\psi} \le \psi_0)\}$, and a is related to the skewness of the score function; both z_0 and a are of order $O(n^{-1/2})$. In the setting of Section 4.2, DiCiccio and Efron (1992) and Efron (1993) showed that $E(r_p) = -z_0 + O(n^{-1})$ and that $r_p + z_0$ has the standard normal distribution to error of order $O(n^{-1})$. Moreover,

$$E\{l'_p(\psi)\} = (a - z_0)\{-l''_p(\psi)\}^{1/2} + O(n^{-1})$$

and

$$l_c(\psi) = l_p(\psi) - (a - z_0)\{-l''_p(\psi)\}^{1/2} + O(n^{-1}).$$

As many authors have noted, adjustment of the log profile likelihood function $l_p(\psi)$ reduces the bias of the profile score. Also, $E(r_c) = -a + O(n^{-1/2})$, and $r_c + a$ has the standard normal distribution to error of order $O(n^{-1})$. Further details are given in DiCiccio and Efron (1995).

ACKNOWLEDGMENTS

The first author's research is supported by NSF Grant DMS-93-05547. This is paper BU-1283-M in the Biometrics Unit, Cornell University, Ithaca, New York 14853. The third author's research is supported by NIH Grant No. RO1-CA61120.

Comment

A. P. Dawid and C. Goutis

Nancy Reid has presented a clear and valuable overview of the uses of conditioning, and of associated techniques of analysis. We wish to focus on some difficulties which can arise from too uncritical an attitude to conditional inference.

It is implicit in Reid's account, as in most others, that the goal of conditional inference has been achieved when we have identified the appropriate conditional "frame of reference" (Dawid, 1991). From that point on, it is implied, we should be free to use any favourite method of inference within that new frame. However, a more thorough-going analysis casts doubt on this assumption. This doubt may be evidenced in several related ways.

First there is the problem of nonuniqueness of (maximal) ancillary statistics, and the consequent arbitrariness, in general, of the conditional frame of reference. The collected works of Basu (1988), which deal thoroughly with these topics, should be required reading for anyone contemplating conditional inference. For example, if (X_i, Y_i) have a bivariate normal distribution with known variances and unknown correlation ρ, each of $\mathbf{X} = (X_1, \ldots, X_n)$ and $\mathbf{Y} = (Y_1, \ldots, Y_n)$ is ancillary, and inference conditional on either appears equally justified. We cannot, however, condition on both, since $(\mathbf{X}, \mathbf{Y})$ reproduces the whole sample.

Next, there is Birnbaum's (1962) celebrated demonstration that acceptance of both the sufficiency and conditionality principles demands acceptance of the likelihood principle—and is thus incompatible with any method of inference which does not respect that principle. A much weaker version of this argument and conclusion, which nevertheless implies the irrelevance of optional stopping and is hence incompatible with many common forms of inference, is given by Dawid (1986).

Then there is the "conflict between conditioning and power" mentioned in Section 6.2. A concrete example, based on Cox (1958a), is analysed in Dawid (1983, pages 99–100). In a problem with point null and alternative hypotheses, and a simple experimental ancillary, the rule "use the likelihood ratio test with size $\alpha = 0.05$," if applied conditionally on the ancillary, does not agree with any unconditional likelihood ratio test and is thus less powerful than the overall 0.05-level test (which has differing conditional α-levels). However, the Neyman–Pearson lemma, which simply requires use of *some* likelihood ratio test, can nonetheless be ap-

Philip Dawid is Professor of Statistics and Costas Goutis is Lecturer in Statistics, Department of Statistical Science, University College London, Gower Street, London WC1E 6BT, United Kingdom.