As one step in this direction, Waterman and Lindsay (1995) have considered an intermediate asymptotics for the Neyman–Scott problem in which the number of parameters goes to infinity, but as the square root of the number of observations. In this setting, the maximum likelihood estimator is asymptotically biased, but the bias can be removed by using projection onto the second-order Bhattacharyya scores, and the resulting estimators attain the asymptotic Cramér–Rao lower bound.

Comment

Peter McCullagh

The modern theory of conditional inference is an attempt to develop a sensible theory of confidence intervals, that is to say, inferential statements about parameters in the absence of prior information or with the explicit declaration of prior ignorance. In that sense, the impetus for recent developments in this area is the same force that motivated Fisher over a period of three decades to develop a solid foundation for his theory of fiducial inference. Although the terminology and formal mathematical theory are due to Neyman (1937), the essential idea and repeated sampling properties of confidence intervals were first spelled out clearly by Fisher (1930). Any ordinary mortal would have been delighted by the enthusiasm with which his ideas on likelihood and interval estimation were espoused, mathematized and extended by Neyman, Pearson and others. For various reasons, Fisher subsequently disowned, and even condemned with characteristic polemic, the idea of confidence interval as an inferential statement. The principal objections raised by Bartlett and Fisher to confidence statements concern their sometimes poor conditional properties and the necessity to specify in advance a particular error rate. While the second of these objections can be overcome to some extent by constructing a set of confidence intervals and presenting the result in the form of a confidence distribution, the first objection is more difficult to surmount. Fisher's effort, though admirable in its goal and skillfully argued, was ultimately unsuccessful.

Neo-Fisherians set themselves a more modest goal. The conditionality principle in some form is accepted, but its consequence, the likelihood principle, is not. If reasonably firm prior information is available, it must be used in Bayes' theorem. This is uncontroversial. If no prior information is available, neither personal opinion nor "objective ignorance prior" is regarded as a satisfactory substitute. Inferential statements must then be constructed without recourse to Bayes' theorem, and such statements must have acceptable conditional properties, at least in large samples. One cannot expect good agreement among statisticians on the basis of small samples because prior information and/or choice of sample space necessarily plays a nonnegligible role. The best that one can hope for is good agreement in large samples. Reid's paper provides a timely opportunity to review the extent to which a satisfactory large-sample frequency theory of inference has been developed in the past two decades.

Before delving into details, it seems pertinent to ask how it is proposed to construct a satisfactory theory based on a mathematical contradiction. Conditionality and sufficiency are accepted, but the likelihood principle is not, in apparent contradiction of Birnbaum's theorem. This prima facie indefensible position cries out for an explanation. The thinking on this issue seems to run as follows:

(i) Many applied statisticians find significance tests very useful in practice.
(ii) Any tool that has proved to be so useful over such a long period cannot be all bad.
(iii) Any statistical principle that denies a role for significance tests cannot be a good principle.

One need only examine the literature on the convergence of the Gibbs sampler or Markov-chain simulation methods to see that even avowed Bayesians find significance tests useful. The indirectness of the interpretation of p-values, a point of sharp criticism in all discussion of principles, does not seem to present a serious obstacle to use. A strong reluc-

Peter McCullagh is Professor, Department of Statistics, University of Chicago, 5734 University Avenue, Chicago, Illinois 60637.

tance to abandon significance tests is not surprising. Consequently, many have refused to accept the likelihood principle, or have sought to evade Birnbaum's equivalence theorem using arguments such as that given by Durbin (1970). My own view is that the likelihood principle is fundamentally sound within the confines of its own arena, that is, provided that the model is judged adequate. However, few statistical models can safely be assumed adequate. Consequently, acceptance of the likelihood principle for estimation conditional on the model does not, in itself, deny a role for significance tests as a way to judge model adequacy (Box, 1980). Confidence intervals, derived provisionally on the model, are a different matter. It is fortunate that most confidence intervals have a direct interpretation coinciding precisely with the interpretation that many elementary textbooks strive in vain to deny. This is essentially the same interpretation that Fisher would like to have established on a firmer footing through his concept of fiducial probability.

Coming back to Reid's article, it is agreed that conditioning on ancillary statistics is essential to avoid misleading conclusions and to make probability statements that are relevant to the data at hand. It is hard to disagree with the claim that conditional confidence intervals and significance tests are preferable to unconditional intervals and tests. However, the paradoxes and ambiguities are not entirely eliminated by conditioning. The general topic of conditional inference is a vigorous research area that has led to numerous improved approximations and has generated many interesting ideas and novel results in the past two decades. But nothing that I have seen in recent years suggests that a totally satisfactory frequency-based theory of inference is likely to emerge. Even for a simple location–scale model where the p^* formula is exact, there may be no unique maximal ancillary (McCullagh, 1992). The various conditional distributions lead to different, and apparently contradictory, inferential statements, all of which are exact. At present, I see no way to avoid these and other irritating contradictions without adopting the likelihood principle and all that it entails. So, while I admire Reid's optimism regarding future developments, I cannot entirely share it.

However, Reid is quite right to emphasize that conditioning is used for several distinct purposes, one of which is the elimination of nuisance parameters. Even though the Bayesian paradigm handles this operation automatically by integration, any device that relieves the statistician of the task of specifying a prior on the nuisance parameters must be attractive. It is not the likelihood principle that many statisticians find unappealing in the orthodox Bayesian paradigm, but the necessity to specify in advance, in a broadly convincing way, the entire set of models to be considered, and a prior distribution on that set. To do this in a way that is neither arbitrary nor capricious is an enormously challenging task. In cases where the option is available, conditioning may provide sensible relief from the megalomania of orthodoxy, though possibly at the cost of sacrifice of principle and some loss of efficiency. The avoidance of unnecessary modeling, while not a matter of principle, remains a sound guideline for statistical practice even if it conflicts with orthodoxy.

Although the two papers under discussion appear unrelated, there are a number of intriguing links. Exponential families play a central role in both papers. The avoidance of unnecessary modeling lies at the heart of recent developments in the theory and practice of estimating functions. The aim is to base all inferences on assumptions that can easily be checked, thereby achieving a degree of robustness to distributional assumptions.

Certain types of transformation models have the property that estimating functions can be constructed for the parameters of interest alone, effectively finessing the nuisance parameters. To give one illustration, consider a simplified version of the problem of shape matching and estimation in which we begin with a standard template defined in two equivalent ways,

$$\{x(t) \in R^2\colon\ 0 \le t < 2\pi\} = \{x\colon\ h(x) = 1\},$$

either parametrically by the periodic function x: $(0, 2\pi) \mapsto R^2$ or by the level sets of h: $R^2 \mapsto R$. The function h is assumed to satisfy the homogeneity condition $h(\rho x) = \rho h(x)$ for $\rho > 0$, so that the level sets of h are scaled versions of the template, which must therefore be star-shaped. Suppose that the actual observations are a centered, scaled and rotated version of the template, observed with error specified by the model

$$y_j = \omega + \varepsilon_j R_\phi x(\lambda_j),$$

in which $\omega \in R^2$ is the center, R_ϕ is a rotation matrix and ε_j is a positive scalar random variable with mean ρ. The parameters of interest are those occurring in the template specification function $x(\cdot)$ together with (ω, ρ, ϕ), and $(\lambda_1, \ldots, \lambda_n)$ are either nuisance parameters or random variables. In the latter case λ and ε are assumed to be independent. Then,

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

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.