Comment

Bruce G. Lindsay and Bing Li

The two papers before us consider the same basic problem: statistical inference for a finite dimensional parameter, possibly in the presence of nuisance parameters. The strikingly different results arise from the differing approaches to making modelling assumptions. Whereas Professors Liang and Zeger would have us make the minimal assumptions necessary to achieve the inference, Professor Reid shows that a completely believed parametric model assumption can be turned into a gold mine of more precise asymptotic approximations. We wish to discuss here some aspects of the middle ground between these two extremes and how it relates to conditional inference.

Perhaps it is useful to make a distinction between the goals that we attempt to achieve by employing conditional inference and the natural consequences to which conditional inference leads. These goals are, for example, (i) to make the assessment of the precision of a statistical method as true to the experiment that actually occurred as possible and (ii) to make the inference about the interest parameter as accurate as possible by minimizing the effect of the estimation of the nuisance parameter. If an appropriate parametric model is applicable, as it is in many important examples, then conditional inference is a powerful means to achieve these purposes. However, we want a statistical procedure to possess these desirable properties whether or not we have suitable ancillary statistics to condition on, and whether or not we have a fully prescribed model under which we can talk about conditional probability in accurate terms. Although the principle of conditioning on ancillary statistics or on sufficient statistics for the nuisance parameter is very clear when we have rigidly prescribed a parametric model, we ask what its statistical meaning might be outside those contexts. We offer here some illustrations of how the idea of projection, as used by Liang and Zeger, can be useful in achieving these goals under such circumstances.

We start with a basic tool, the Bhattacharyya scores. Let $X = (X_1, \ldots, X_n)$ be independent random observations with density $f(x;\, \theta)$. The Bhattacharyya scores B_i, $i = 0, 1, 2, \ldots,$ are defined by

$$(1)\quad B_0 = 1, \quad B_k = \frac{\partial^i f(x;\, \theta)/\partial\theta^i}{f(x;\, \theta)}, \quad i = 1, 2, \ldots.$$

For example, if we write the log-likelihood function as $l(\theta;\, X)$, then B_1 is the score function $\dot{l}$ and B_2 is $\ddot{l} + \dot{l}^2$. We consider $\{B_i\colon\ i = 0, 1, \ldots, k\}$ as vectors in the Hilbert space of square-integrable functions $h(x, \theta)$, with inner product $E\{h_1(X, \theta)h_2(X, \theta);\, \theta\}$. We use $\mathscr{B}$ to denote the subspace spanned some or all Bhattacharyya scores; $P_{\mathscr{B}}$, the orthogonal projection onto $\mathscr{B}$; and $I - P_{\mathscr{B}}$, the orthogonal projection onto the orthogonal complement of $\mathscr{B}$. We wish to discuss how one can use these scores to evaluate or improve conditional-type properties of inference.

Our first illustration is an optimality property of the observed Fisher information. Let $\hat{\theta}$ be the maximum likelihood estimate. Conditional inference suggests that the assessment of the precision of $\hat{\theta}$ be conditioned on any ancillary statistics. The results of Efron and Hinkley (1978) indicate that, for translation families and numerous other cases, the variance of $\hat{\theta}$, conditioned upon an ancillary, is approximated by the inverse of the observed Fisher information. Here, the goal is to make the precision assessment more relevant to the realized experiment, and it is achieved by drawing inference conditioning on ancillary statistics.

It is possible, however, to achieve this goal without specifying an ancillary statistic, either exact or approximate. Consider the following (unconditional) minimization problem: choose a statistic $T(X)$ that minimizes the mean squared error

$$E_{\theta_0}\{(\hat{\theta} - \theta_0)^2 - T(X)\}^2.$$

Lindsay and Li (1995) demonstrated that, among a wide class of statistics, the asymptotically optimal choice of $T(X)$ is once again the inverse of the observed Fisher information. The result relies not on the specification of ancillaries or approximate ancillaries, which may be difficult to obtain under some circumstances, but rather an asymptotic Cramér–Rao–type argument based on the projection of $(\hat{\theta} - \theta_0)^2 - T(X)$ onto $\mathscr{B} = \text{span}\{B_i\colon\ i = 0, 1, 2\}$.

Bruce G. Lindsay is Professor and Bing Li is Assistant Professor, Department of Statistics, Pennsylvania State University, University Park, Pennsylvania 16802.

Our second illustration concerns the construction of an estimating equation for the interest parameter in the presence of the nuisance parameter. A reasonable construction is based on the following consideration:

(a) the consistency of the estimation of λ should not affect the consistency of the estimation of ψ, that is, $E\{g(X, \psi_0, \lambda); \psi_0, \lambda'\} = 0$ for all λ and λ'. This is condition (b) of Section 3.1 of Liang and Zeger's paper;

(b) among all those $h(x, \psi, \lambda)$ that satisfy (a), g has the highest efficiency, or is nearest to the unconditional score; that is,

$$E[\{g(X, \psi_0, \lambda) - U(X, \psi_0, \lambda)\}^2; \psi_0, \lambda] \le E[\{h(X, \psi_0, \lambda) - U(X, \psi_0, \lambda)\}^2; \psi_0, \lambda],$$

where U denotes the unconditional score about ψ evaluated at ψ_0 and λ.

Suppose that the data X can be decomposed into $(T, S(\psi))$ such that the likelihood has the following factorization:

$$(2) \qquad f(x; \psi, \lambda) = f(t \mid s(\psi); \psi) f(t; \psi, \lambda).$$

The first factor, the conditional likelihood of T given $S(\psi)$, is independent of λ only when the value of ψ that indexes the likelihood coincides with that which indexes the the random variable S. Let $\psi = \psi_0$ be the true value of the parameter of interest, as in single versus composite test for ψ. As discussed by Reid, there are various reasons to base inference on the "conditional likelihood" $f(t \mid s(\psi_0), \psi, \lambda)$. For example, the likelihood ratio tests based on this have similar regions and, if the conditional density has monotone likelihood ratio, they are most powerful unbiased tests. The conditional score function should, then, be defined as the derivative of this "conditional likelihood," that is,

$$g(x, \psi_0, \lambda) = \frac{\partial \log f(x; \psi_0, \lambda)}{\partial \psi} - \frac{\partial \log f(s(\psi_0); \psi, \lambda)}{\partial \psi}\bigg|_{\psi=\psi_0}.$$

Under some regularity conditions, the conditional score g is the unique estimating equation that satisfies conditions (a) and (b). See Godambe (1976) and Lindsay (1982).

Again, the conditional procedure leads to an estimating equation that meets our goal, provided that factorization (2) applies. However, requirements (a) and (b) themselves are not conditional. Indeed, they can be satisfied, in theory at least, even without the factorization (2). The optimization problem implicit in (a) and (b) amounts to projecting the true score function about ψ onto the orthogonal space of $\mathscr{B}$, the Hilbert space spanned by all the Bhattacharyya scores about λ [replacing $\partial\theta$ by $\partial\lambda$ in (1)]; see Small and McLeish (1988, 1989). In other words,

$$(I - P_{\mathscr{B}})U(X, \psi_0, \lambda)$$

is always the optimal solution, whether or not there is a sufficient statistic for λ (Lindsay and Waterman, 1992). Perhaps the most important aspect of the Waterman–Lindsay results is that if we write

$$(I - P_{\mathscr{B}})U = (I - P_1)U + (P_1 - P_2)U + \cdots,$$

where P_k is the projection onto span$\{B_i\colon i = 0, \ldots, k\}$, then the stochastic magnitude in n of the correction terms $(P_k - P_{k+1})U$ are declining in k, and in examples dramatically so, so that simply by using $(I - P_2)U$, one often has a score function nearly identical to $(I - P_{\mathscr{B}})U$. Thus, just as in the observed Fisher information problem, the key aspects of conditioning can be obtained by projection onto just the first- and second-order scores.

It is also important that (a) and (b) are well-posed requirements under semiparametric assumptions, such as discussed in Liang and Zeger's paper, under which we cannot talk about conditional probability in accurate terms.

The final issue we wish to raise is the question of suitable asymptotics. Any actual statistical experiment has a finite sample size. It is very likely that any number of asymptotic schemes could be used to approximate the finite sample probability calculations. For example, in the Neyman–Scott problem, two widely different approximations to the properties of the maximum likelihood estimator arise, depending on whether the number of nuisance parameters is fixed as the number of observations increases (standard asymptotics, MLE is consistent and efficient) or the number of nuisance parameters increases proportionally to the number of observations (MLE is possibly inconsistent, or consistent but inefficient; Neyman and Scott, 1948). The second asymptotics tells us that some aspects of the first asymptotics must be rather inaccurate when there are many nuisance parameters. Although standard approaches that improve inferential properties under the first asymptotics are no doubt superior in the Neyman–Scott problem as well, unless they are evaluated specifically for the effect of many nuisance parameters, it is hard to assess the extent to which they are successful.

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.