Comment

Thomas A. Severini

The papers by Reid and by Liang and Zeger cover several important areas of statistics. I will confine my comments to three topics considered in Professor Reid's excellent review of conditional inference.

1. APPROXIMATE SUFFICIENCY IN THE PRESENCE OF A NUISANCE PARAMETER

In Section 5 Reid suggests that it may be possible to approximately eliminate a nuisance parameter by conditioning on a statistic that is approximately sufficient for fixed values of the parameter of interest. I now present some recent results regarding this issue; further details and additional results are available in Severini (1993, 1994a).

Consider a model parameterized by a scalar parameter of interest ψ and a nuisance parameter λ; for simplicity, take λ also to be a scalar although the results hold more generally. A natural approach to consider, given its optimality in exponential family models, is to take $S_1 = \hat{\psi}$, the MLE of ψ, and $S_2 = \hat{\lambda}_\psi$, the MLE of λ for fixed ψ. In exponential family models in which ψ is a linear function of the canonical parameter this leads to exact methods of inference with well-known optimality properties.

In general, under standard regularity conditions,

$$\begin{aligned}&\Pr(S_1 \geq s_1 | S_2 = s_2;\ \psi, \lambda_0 + \varepsilon n^{-1/2}) \\ &\quad - \Pr(S_1 \geq s_1 | S_2 = s_2; \psi, \lambda_0) \\ &\quad = \frac{i_\theta}{2} \gamma \varepsilon^2 n^{-1/2} + O(n^{-1}),\end{aligned}$$

where i_θ is the asymptotic variance of $\sqrt{n}(\hat{\psi} - \psi)$,

$$\gamma = \frac{1}{n} E\left[\left(\frac{\partial l(\psi, \lambda_0)}{\partial \psi} - \beta \frac{\partial l(\psi, \lambda_0)}{\partial \lambda}\right) \frac{\partial^2 l(\psi, \lambda_0)}{\partial \lambda^2}; \psi, \lambda_0\right],$$

$$\beta = \frac{E[\partial^2 l(\psi, \lambda_0)/\partial \psi\, \partial \lambda; \psi, \lambda_0]}{E[\partial^2 l(\psi, \lambda_0)/\partial \lambda^2; \psi, \lambda_0]}$$

and l is the log-likelihood function. Hence, this approach is successful in approximately eliminating the nuisance parameter, to the order considered, provided that $\gamma = O(n^{-1/2})$. In particular, if the distribution of the data for fixed ψ forms a full exponential family model, then $\gamma = 0$; in general, γ is a measure of how close the distribution is to full exponential family form.

Consider an exponential family model with log-likelihood function of the form

$$l(\psi, \lambda) = g(\psi, \lambda) t_1 + \lambda t_2 - k(\psi, \lambda),$$

where (t_1, t_2) is the sufficient statistic and is of order $O_p(n)$ and $\partial g(\psi, \lambda)/\partial \psi \neq 0$. Then $\gamma = O(n^{-1/2})$ requires that

$$\frac{\partial^2 g(\psi, \lambda_0)}{\partial \lambda^2} = 0. \tag{1}$$

Under this condition, $g(\psi, \lambda)$ is linear in λ for each fixed ψ so that there exists a one-dimensional sufficient statistic for λ for each fixed ψ and the conditional distribution of the data given $\hat{\lambda}_\psi$ is exactly free of λ. When (1) does not hold, we could attempt to eliminate λ by conditioning on $S_2 = (\hat{\lambda}_\psi, A)$ for some statistic A, but it is clear that such a statistic S_2 would have to be equivalent to the sufficient statistic in the original, unrestricted, model. Hence, in full exponential family models when the exact theory of conditional inference fails, an approximate theory fails as well, at least using the approach considered here.

This suggests that, for inference in the presence of a nuisance parameter, methods based on the marginal distribution of some quantity, such as the modified likelihood ratio statistic r^*_ψ, are likely to be more generally applicable than methods based on a theory of approximate conditional inference.

2. THE RELATIONSHIP BETWEEN BAYESIAN INFERENCE AND CONDITIONAL INFERENCE

One advantage of Bayesian inference over non-Bayesian methods of inference is in the treatment of problems involving a nuisance parameter. In Bayesian inference, any nuisance parameter can be eliminated by integrating it out, at least in principle. Given the formal appeal of Bayesian methods, as well as some well-known optimality properties, it is of interest to determine when a non-Bayesian method of eliminating a nuisance parameter, such as conditional inference, corresponds to Bayesian inference with respect to some prior distribution.

Thomas Severini is Associate Professor, Department of Statistics, Northwestern University, 2006 Sheridan Road, Evanston, Illinois 60208-4070.

Consider inference about a scalar parameter of interest ψ in the presence of a scalar nuisance parameter λ based on a statistic $S = (S_1, S_2)$. Suppose that the density of S factors in the following manner:

$$f(s; \psi, \lambda) = f(s_1|s_2; \psi)f(s_2; \psi, \lambda).$$

Non-Bayesian inference about ψ can then be based on $f(s_1|s_2; \psi)$. Let $\pi(\lambda|\psi)$ denote a conditional prior density of λ given ψ. If

$$\int f(s_2; \psi, \lambda)\pi(\lambda|\psi)\, d\lambda \tag{2}$$

does not depend on ψ, then the posterior density of ψ based on the density $f(s; \psi, \lambda)$ is the same as the posterior density of ψ based on $f(s_1|s_2; \psi)$. In this case, elimination of λ by using $f(s_1|s_2; \psi)$ does correspond to elimination of λ using Bayesian inference. Hence, we want to determine the conditions under which (2) does not depend on ψ with respect to some prior density.

For instance, if the marginal distribution of S_2 does not depend on ψ, then it is clear that (2) does not depend on ψ for any prior distribution on (ψ, λ) such that ψ and λ are independent. More generally, suppose that S_2 is an S-ancillary statistic (see, e.g., Barndorff-Nielsen, 1978); that is, suppose that the family of density functions

$$\{f(s_2; \psi, \lambda):\ \lambda \in \Lambda\} \tag{3}$$

is the same for each ψ; here Λ denotes the space of possible λ. It is easy to show that $f(s_2; \psi, \lambda)$ depends on (ψ, λ) only through a real-valued parameter $\phi \equiv \phi(\psi, \lambda)$ and, hence, (2) holds for any prior density on (ψ, λ) such that ψ and $\phi(\psi, \lambda)$ are independent.

A similar result holds if for each ψ the family of probability distributions corresponding to (3) forms a transformation model with respect to a transitive group of transformations isomorphic to Λ. In this case, (2) holds with $\pi(\lambda|\psi) \equiv \pi(\lambda)$ taken to be the density of the right-invariant measure on Λ.

However, in many models in which conditional inference is used to eliminate a nuisance parameter, neither of these conditions is satisfied and in some cases it can be shown that (2) does not hold for any prior density, subject to weak regularity conditions (Severini, 1994b). This suggests that, in these cases, it may be desirable to relax the requirement of exact similarity in order to use the information in S_2 for inference about ψ. On the other hand, the results of Davison (1988) indicate that, for exponential family models in which ψ is a linear function of the canonical parameter, elimination of ψ by conditioning on the sufficient statistic for the nuisance parameter does approximately correspond to Bayesian inference. Hence, the additional information available in the distribution of S_2 may be negligible.

3. CONDITIONAL TESTS AND POWER

Although it is often stated that conditional tests are less powerful than unconditional ones, comparisons of this type depend on whether or not there exists an optimal unconditional test, as well as on how the significance level of the conditional test depends on the value of the conditioning statistic. Many of the points raised in the following discussion are also discussed in Barnard (1982).

In the case of testing a simple null hypothesis versus a simple alternative, the unconditional test with level α based on the likelihood ratio statistic has, of course, optimal unconditional properties. Consider a conditional test given a statistic A based on the likelihood ratio statistic, and let $\alpha(a)$ denote the conditional significance level of the test given $A = a$. If it is required that $\alpha(a) = \alpha$ for each a, then the conditional likelihood ratio (CLR) test typically has less power than the unconditional likelihood ratio (ULR) test; throughout this discussion the term power will refer to unconditional power. However, suppose that $\alpha(a)$ is allowed to vary with a; this may be quite natural if A is a measure of the precision of the experiment, such as an effective sample size. In this case, the CLR test may be as powerful as the ULR test with the same unconditional level. This is easy to see by simply taking $\alpha(a)$ to be the conditional significance level of the ULR test given $A = a$. In this case, the conditional test and unconditional test are essentially the same, except that the results of the conditional test are interpreted conditionally on the observed value of A. Similar conclusions hold in those cases in which a uniformly most powerful unconditional test exists. When an optimal unconditional test does not exist, either the conditional test or the unconditional test may have higher power under a given alternative.

These ideas are easily illustrated on the example of two weighing machines with different precisions, considered by Cox (1958a). Let A denote a random variable taking values 0 and 1 each with probability 1/2. Given $A = a$, let X denote a normally distributed random variable with unknown mean μ and standard deviation σ_a, where $\sigma_1 > \sigma_0$. Here we will consider $\sigma_0 = 1$ and $\sigma_1 = 3$. Consider testing $\mu = 0$ versus the alternative $\mu = 3$. The ULR test with level $\alpha = 0.05$ depends on the alternative $\mu = 3$ but is easily determined and can be shown to have power 0.593. The CLR test with

$\alpha(a) = 0.05$ for $a = 0, 1$ can be shown to have power 0.586. However, the power of the test depends heavily on the value of A; when $A = 0$, the power 0.912 as opposed to a power of 0.259 when $A = 1$. Hence, it may be desirable to decrease $\alpha(0)$ and increase $\alpha(1)$. Since the ULR test has conditional level 0.033 when $A = 0$ and 0.067 when $A = 1$, the power of the CLR test is maximized by taking $\alpha(0) = 0.033$ and $\alpha(1) = 0.067$; under these choices the unconditional and conditional tests are identical.

Now consider a test of the null hypothesis $\mu = 0$ versus $\mu > 0$. In this case there does not exist a uniformly most powerful unconditional test. A reasonable choice for a test statistic may be X, the MLE of μ. The test with level 0.05 that rejects the null hypothesis for large values of X has power 0.294, 0.763 and 0.926 at alternatives $\mu = 3, 5$ and 7, respectively. The conditional test described previously with $\alpha(0) = \alpha(1) = 0.05$ also rejects $\mu = 0$ for large X and is uniformly most powerful among conditional tests; this test has power 0.586, 0.754 and 0.877 at $\mu = 3, 5$ and 7, respectively. Hence, which test is more powerful depends on the alternative under consideration. If the unconditional test had been based on the statistic X/σ_A, then the conditional and unconditional tests would be identical; of course, there would still exist unconditional tests with higher power for some alternatives.

The point of this discussion is that there is nothing inherently inefficient about conditional inference even when the properties are assessed unconditionally, although I agree with Reid that such comparisons are typically not directly relevant.

ACKNOWLEDGMENT

This work was supported by a grant from the National Science Foundation.

Comment

Louise M. Ryan

Professors Liang and Zeger deserve congratulations for yet another excellent contribution to the statistical literature. My discussion will first elaborate on their Example 1.3, the analysis of teratology (developmental toxicity) data, then outline some needed extensions and further applications.

Teratology is a fascinating research area, not only because it is such an important public health concern, but also because the statistical problems that arise in this context are so interesting. Due to the limited availability of reliable epidemiological data, controlled experiments in laboratory animals play a critical role in the safety assessment and regulation of substances with potential danger to the developing human fetus. In a typical study (depicted in Figure 1), pregnant dams (usually mice, rats or sometimes rabbits) are randomized to a control group or one of three or four exposed groups. Dams are exposed to the test substance during the period of major organogenesis when the developing offspring are likely to be most sensitive to insult. Just prior to normal delivery, the dams are sacrificed and the uterine contents examined for defects. A typical study might have 20 to 30 dams per group, with anywhere from 1 to 20 offspring per litter.

Anyone familiar with the developmental toxicity literature will be aware of the longstanding debate over how to handle the so-called litter effect (or the tendency of littermates to respond more similarly than nonlittermates). The debate started in the early 1970's with papers in the toxicology journals asking questions like "what are the sampling units" in a teratology study. The paper cited by Professors Liang and Zeger (Weil, 1970) inspired an editorial in the journal *Teratology* by Kalter (1974), complaining that "statistics here has exceeded its role as handmaiden" and suggesting that such considerations are best left to the biologists! In response to this editorial, Staples and Hasemen (1974) emphasized that a proper statistical analysis should use all the fetus-specific information, but must allow for possible correlation between littermates. Since then, much attention has focussed on the development of suitable statistical methods. Earlier suggestions (e.g., Williams, 1975) recommended use of a beta-binomial distribution, mainly because of its concep-

Louise Ryan is Professor of Biostatistics, Harvard School of Public Health and Dana Farber Cancer Institute, 44 Binney Street, Boston, Massachusetts 02115.