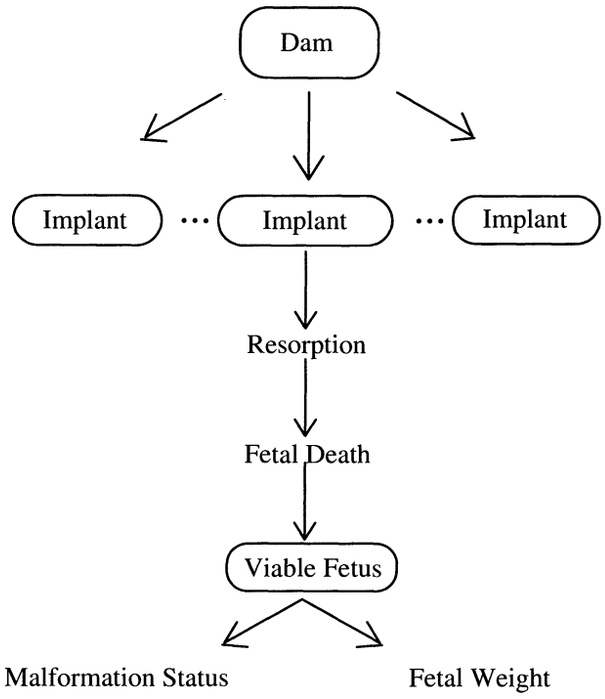

FIG. 3. *Fetal outcomes in developmental toxicity studies*

gested formulating the problem in terms of a trinomial outcome (dead, malformed, normal). Chen, Kodell, Howe and Gaylor (1991) suggest a parametric approach based on a Dirichlet trinomial distribution. Ryan (1992b), Catalano, Scharfstein and Ryan (1994), Zhu, Krewski and Ross (1994) and Krewski and Zhu (1994) use estimating equations. In general, the best approach for the analysis of correlated multinomial data is not well established. There are several different ways to set up either parametric approaches or estimating equations, but there has not been any systematic study or comparison of the various approaches. Finding ways to analyze dose effects on fetal weight and malformation status is another interesting challenge. In general, there are few methods available for the analysis of multivariate data involving a mixture of discrete and continuous outcomes. Methods for clustered data of this kind are virtually nonexistent, although Catalano and Ryan (1992) and Catalano et al. (1993) suggest one approach based on conditional estimating equations. Theoretically, there is no reason why marginal estimating equations could not be constructed for such data. However, there has been little work on this topic.

ACKNOWLEDGMENTS

The author thanks Nick Lange and Paige Williams for help with the figures, as well as Paul Catalano and Ralph Kodell for helpful comments on the manuscript.

Rejoinder

N. Reid

1. INTRODUCTION

Criticisms of non-Bayesian conditional inference fall roughly into one of two categories: foundational and practical. The foundational criticisms of conditioning revolve around whether or not conditioning should be a basic statistical principle, on a similar footing to, for example, sufficiency. Practical criticism of conditioning tends to concentrate more on the fact that models arising in applications tend to be complex and not often readily amenable to a textbook treatment of conditioning or marginalizing. As well, in many practical settings questions about modelling or sampling, such as whether or not observations are independent, are more crucial than questions of whether to use a first-order or higher-order approximation.

A third aspect of the discussion, closely related to these, is the claim that a Bayesian approach addresses both these criticisms, by being logically coherent as well as practically straightforward. In addition, it automatically conditions on all the data; what could be more conditional than that? A related, somewhat more technical, part of this debate is the extent to which Bayesian and non-Bayesian solutions to a problem can be made to agree.

In this paper I tried to emphasize techniques of conditional inference, rather than the philosophy of conditional inference. However, this is a paper on conditional inference in the theory of statistics, not in the practice of statistics. A paper which explored to what extent conditional ideas could be used in "real" applications would have a very different focus. It might perhaps come to a negative

conclusion, although I personally do not think so, but it would be a different paper. The recent asymptotic theory of conditional inference, as a contribution to statistical theory, has developed in my view as an attempt to understand the fundamental role of likelihood in a frequentist theory of inference, and this requires looking at models that are considerably abstracted from statistical practice. This is not at all the same thing as constructing a theory by and for the "cognoscenti" referred to by Casella, DiCiccio and Wells.

2. FOUNDATIONS

I tried to avoid as much as possible a discussion of the foundations of inference, as I find their study confusing. However, it is difficult to delve very deeply into a study of conditional inference without coming up against foundational aspects, and Dawid and Goutis quite rightly raise some troubling issues. The most concrete of these, also mentioned by McCullagh, is the fact that maximal ancillaries are not uniquely defined and conditioning on different maximal ancillaries may lead to different exact solutions. The problem is reviewed in Barndorff-Nielsen and Cox (1994, Chapter 2), which also includes a discussion of Cox's (1971) suggestion for choosing between alternative ancillary statistics, as well as references to more recent work. It would be interesting to see if Cox's method can distinguish between various approximately ancillary statistics as well and, in particular, can identify the ancillary statistic that ensures the validity of p^* to $O(n^{-3/2})$.

Another criticism of conditioning that is foundational, raised by Dawid and Goutis, is the conflict between conditioning and power. This has been much discussed in the literature, and to my mind unconditional measures of power are only rarely relevant. I find the discussion provided by Severini expanding Barnard's (1982) approach very interesting and helpful in clarifying some of the important points.

The more widely quoted foundational issue is that the conditionality principle implies the likelihood principle, which in turn seems to lead directly to a Bayesian approach to inference, because some framework is needed in which probability statements can be made. If the randomness is not to come from repeated sampling, it must come from somewhere else, unless we all agree to assess and compare likelihoods without reference to probability statements. This issue I do not know how to debate at all, beyond relying on what I outlined in the paper: conditioning seems to be a convenient tool for a number of different purposes. Whether it should be enshrined as a principle I do not know. In general most of the principles that statisticians espouse fail in one aspect or another, but luckily a pragmatic approach to problems seems to be fairly effective in applications. Still, it is vaguely unsettling to be unable to pin the discipline down on a principled foundation.

3. THE BAYESIAN SOLUTION

It has been repeatedly pointed out and is again here by several discussants, including McCullagh (to my surprise), Casella, DiCiccio and Wells, and Dawid and Goutis, that conditioning is automatically effected in a Bayesian approach, which is at the same time both conceptually more straightforward and philosophically more coherent.

To which my answer continues to be "Yes, but..."; the but referring to the difficulty of choosing a prior for a high-dimensional nuisance parameter. I simply have not yet seen any convincing solution in theory or in practice. In many practical applications the suggestion of Casella, DiCiccio and Wells is followed: assign flat priors to any parameters deemed unimportant. The example of exponential regression due to Mitchell (1967) and outlined in Cox and Hinkley (1974, Chapter 10) strikes me as a fairly realistic illustration that a combination of flat priors can have surprisingly nonflat effects. This point is also discussed in Kass and Wasserman (1994). At the same time, work on constructing noninformative or conventional or objective or reference priors is proceeding apace, but there does not seem to be a consensus solution yet. (Berger and Ye, 1991, develop reference priors for Mitchell's example.) In my view choosing a prior for a high-dimensional parameter, even subjectively, is as difficult as finding an exact or approximate conditional solution. Given this difficulty, the fact that the Bayesian approach is logically consistent strikes me as somewhat irrelevant.

Casella, DiCiccio and Wells argue in their Section 2 that non-Bayesian elimination of nuisance parameters is bewildering, apparently because in some models nuisance parameters are eliminated by conditioning, and in others by marginalizing. Bewilderment is often a function of familiarity; the large and growing literature on noninformative priors is also difficult to synthesize. While typical flat or default priors may lead to unsurprising inferences, the evidence that they lead to reasonable frequentist inferences is, with some exceptions, fairly anecdotal.

The exceptions are studies of the exact frequentist properties of certain posterior inferences, such as is outlined in Section 4 of Casella, DiCiccio and Wells.

The problem addressed there is that of trying to determine priors for which the resulting Bayesian inferences are also accurate frequentist inference statements. Priors satisfying this are often called noninformative; the phrase "matching prior" is often used as well. As I mentioned above, there is a large and growing literature on this subject. Some of it is reviewed in Kass and Wasserman (1994) and some in Reid (1995). In Casella, DiCiccio and Wells' Section 4, the argument works from an r^* approximation back to a noninformative prior, which is an interesting turn on the usual approach. It is suggested in the gamma example that it may not be possible to match frequentist and Bayesian inferences to $O(n^{-3/2})$ adding to growing evidence (Pierce and Peters, 1994) that this may be the case in general. Connecting up to parametric bootstrap methods also seems to be limited to $O(n^{-1})$ matching.

In my opinion one of the most promising approaches to inference is through an exploration of matching priors, or more generally, compatibility of Bayesian and frequentist inferences. This is also suggested by Dawid and Goutis, and an approach described in Dawid (1991) has been developed in some detail in Sweeting (1995a, b).

4. EXAMPLES

It is certainly the case that the examples considered in my paper are abstracted from the types of models used in practical applications. This is the main (nonfoundational) criticism addressed to the topic of conditional inference, and it was raised by all the referees.

There are two aspects of the difficulty: one inferential and the other computational. The inferential difficulty is, in effect, knowing what to condition on; this is for the most part solved only in the cases of exponential linear models and transformation models. Extensions involving approximate elimination of nuisance parameters, such as Severini has discussed here and in several related papers, have not yet settled on an obvious solution, and may not. (Development of noninformative priors in problems with nuisance parameters is similarly unresolved.) It may be that the "right" solution turns out to be one that is independent of the conditioning event, as suggested by Severini and by Dawid and Goutis.

However, even in cases where the appropriate conditional distribution is straightforward, conditional methods are often not used. Examples include any linear regression model with nonnormal errors, as discussed in Example 5.3, and any generalized linear model with canonical link. This is, I think, because the computational simplicity of the method is not widely appreciated, and I have perhaps also failed to convey this simplicity.

The example of logistic regression given by Casella, DiCiccio and Wells is an example in which nuisance parameters can be exactly eliminated by conditioning, in much the same way as described for exponential regression in my Section 5.3, and as described as well in Davison (1988). Inference for the slope parameter $\psi = \theta_1$, as a programming exercise for an M.Sc. class, goes as follows:

- Given a sample (y_i, x_i), $i = 1, \ldots, n$, and a fixed value of ψ, solve for $\hat{\lambda}_\psi$ the equation

$$y_+ = \sum \frac{\exp(\hat{\lambda}_\psi + \psi x_i)}{1 + \exp(\hat{\lambda}_\psi + \psi x_i)},$$

  where $y_+ = \sum y_i$.
- Compute the function

$$l_c(\psi) = \hat{\lambda}_\psi y_+ + \psi (yx)_+ - \sum \log(1 + \exp(\hat{\lambda}_\psi + \psi x_i)) + \left(\frac{1}{2}\right) \log \sum \frac{\exp(\hat{\lambda}_\psi + \psi x_i)}{(1 + \exp(\hat{\lambda}_\psi + \psi x_i))^2}$$

  where $(yx)_+ = \sum y_i x_i$.
- Find the value of ψ ($\hat{\psi}_c$, say) that maximizes $l_c(\psi)$, and the second derivative of l_c at the maximum.
- Combine the values computed above into the expression given by equation (4.2).
- (Bonus) Find upper and lower 95% confidence bounds for ψ by repeating the steps above for a grid of values for ψ.

As I do not have a Gibbs sampler on my computer yet, this is much more straightforward for me than computing a joint density for both parameters and then marginalizing either numerically or by Casella, DiCiccio and Wells' approximation (4), which requires an importance sampling density as well. The steps involved in the general location–scale model are identical to the above, although of course the form of the joint likelihood is different.

My point is that this type of calculation could be built into the package that fits logistic regression, because it is in the class of models for which exact conditional solutions are available. Until it is, discussions of computational simplicity or complexity are largely subjective.

If we extend this example to multiple covariates, equation (4) of Casella, DiCiccio and Wells, with an appropriate change of notation, still gives the marginal density for the joint distribution of all the parameters, and they suggest that it can be

integrated to find marginal posterior densities for components of interest, using flat priors for the regression coefficients. (It is not clear to me how one can be sure that anomalies of the type that arise in Mitchell's example cited above do not arise in this model.) The conditional approach, carried out approximately, requires an exercise similar to that outlined above to be recomputed for each component parameter of interest. While this is tiresome, it is not particularly complicated.

Casella, DiCiccio and Wells refer in their Section 3.1 to the saddlepoint alternative. I think it is more accurate to refer to it as a marginal alternative, since, as they acknowledge, Field's saddlepoint approximation for M-estimates is an approximation to the marginal sampling distribution of these estimates. It shares the drawback with other marginal solutions that elimination of nuisance parameters is not achieved in models where nuisance parameters are eliminated by conditioning. The saddlepoint method is a technique of approximation, which can be applied to conditional or unconditional models, but is not an inferential methodology.

5. CONCLUSION

The approach of Liang and Zeger is more directly motivated by particular practical applications, and it presents a quite different method for eliminating or minimizing the effect of nuisance parameters. At first glance our two papers seem quite unrelated, but the discussion of Lindsay and Waterman shows that they are more closely related than might appear. In particular, Lindsay and Li give convincing evidence that projection using Bhattacharyya scores can imitate conditioning. It would be interesting to know if this approach reproduces exact results when they are available, and what the connection might be to approximate ancillarity.

Dawid and Goutis refer to some confusion in my use of the terms sufficiency and ancillarity in the presence of nuisance parameters, and they point out that one aspect of this is the use of the phrase "the nuisance parameter," when in fact this parameter is typically not uniquely defined. Severini provides a more careful, and more helpful, definition of S-ancillarity related to this point. Barndorff-Nielsen and Cox (1994, Chapter 8) emphasize the importance of finding procedures which are invariant to interest-respecting reparametrizations. It is possible that, in particular models, a natural form of the nuisance parameter could be constructed using the estimating equations approach; that is, the nuisance parameter in a fully specified parametric model could be chosen to coincide with the nuisance parameter that would arise in a compatible semiparametric estimating equations approach.

In conclusion I find it heartening to see that discussions of theoretical statistics continue to engender lively debate.

ACKNOWLEDGMENTS

I would like to thank the discussants for their contributions and the Editor, Rob Kass, for his encouragement.

Rejoinder

Kung-Yee Liang and Scott L. Zeger

We would like to thank the discussants for their thoughtful comments. We would also like to add our congratulations to Dr. Reid, for her clear exposition on conditional inferences, a tool for reducing the influence of nuisance parameters in a fully parametric setting.

The two papers by Dr. Reid and ourselves address, in part, the common question of how to draw inferences in the presence of nuisance parameters. As pointed out by the discussants, they also both focus on methods most directly applicable to exponential family models. However, the fundamental distinction between these two papers is the degree to which we specify a probability mechanism for the data. Dr. Reid's paper starts with the assumption that a full probability mechanism can be specified. We begin with the assumption that it is not possible nor perhaps desirable to do so.

Peter McCullagh comments on the role of conditional inference when the likelihood is fully specified. He reflects upon the inherent contradiction in the practice of statistics that conditionality and sufficiency are accepted, while the likelihood principle is not. He raises the important point that the like-