$\alpha(a) = 0.05$ 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.

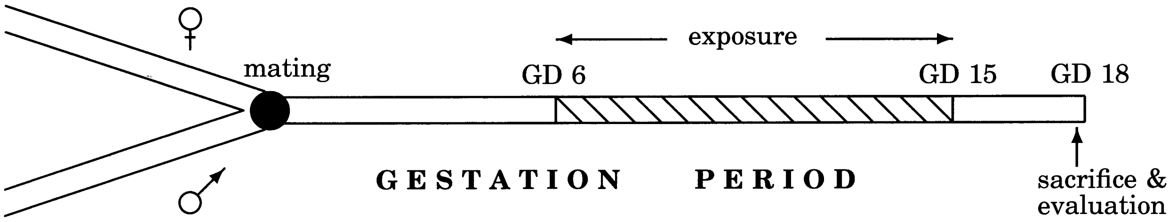

FIG. 1. *Chronological events in a developmental toxicity study.*

tual simplicity and a certain biological appeal. Consider a study with a control and I dose groups, and let y_{ij} denote the number of defects among the n_{ij} offspring born to animal j in dose group i, $i = 0, \ldots, I$, $j = 1, \ldots, N_i$. The beta-binomial distribution is derived by assuming that y_{ij} comes from a binomial(n_{ij}, p_i) distribution, where the response probabilities p_i follow a beta distribution with parameters α_i and β_i that depend on exposure group. The corresponding marginal distribution for Y_{ij} can be obtained by integrating over this beta distribution to obtain

$$\begin{aligned} &\Pr(Y_{ij} = y_{ij} \mid n_{ij}) \\ (1) \qquad &= \binom{n_{ij}}{y_{ij}} \frac{B(\alpha_i + y_{ij}, \beta_i + n_{ij} - y_{ij})}{B(\alpha_i, \beta_i)}, \end{aligned}$$

where $B(\alpha_i, \beta_i) = \Gamma(\alpha_i)\Gamma(\beta_i)/\Gamma(\alpha_i + \beta_i)$. Rather than modelling the α_i and β_i directly, it has been common (Williams, 1975) to reparametrize in terms of the beta mean,

$$E(p_i) = \mu_i = \frac{\alpha_i}{\alpha_i + \beta_i},$$

and an additional parameter $\phi_i = 1/(\alpha_i + \beta_i + 1)$ related to the variance of p_i,

$$\operatorname{Var}(p_i) = \mu_i(1 - \mu_i)\phi_i,$$

and also corresponding to the correlation between two littermates. As will be discussed presently, ϕ_i may or may not change with exposure group. After reparametrizing in this way, one can characterize the exposure effect through a suitable dose–response function on μ_i, for example,

$$(2) \qquad \mu_i = h(\theta_0 + \theta_1 d_i^{\theta_2}),$$

where $h(\cdot)$ is a cdf. The most familiar example of a logistic model corresponds to $h(x) = \exp(x)/(1 + \exp(x))$ and $\theta_2 = 1$. A particularly popular choice (corresponding to the one-hit model used in carcinogenesis) is $h(x) = 1 - \exp(-x)$. Regardless of the choice for $h(\cdot)$, most applications in developmental toxicity require the additional flexibility afforded by the "shape" parameter θ_2. Some authors (see Krewski and Zhu, 1994) favor the inclusion of litter size (n_{ij}) as an additional covariate modifying the expected response rate for a specific litter.

Consider now the topic addressed by Professors Liang and Zeger, namely, estimation of the parameters of interest (θ_0, θ_1 and θ_2) in the presence of the nuisance parameters $\phi_0, \ldots, \phi_I$. As discussed in their paper, likelihood-based estimation is problematic. For one thing, the beta-binomial distribution has a computationally awkward form:

$$\begin{aligned} &\mathbb{P}(Y_{ij} = y_{ij}) \\ &= \binom{n_{ij}}{y_{ij}} \frac{\prod_{k=0}^{y_{ij}-1}(\mu_i + k\eta_i)\prod_{k=0}^{n_{ij}-y_{ij}-1}(1 - \mu_i + k\eta_i)}{\prod_{k=0}^{n_{ij}-1}(1 + k\eta_i)}, \end{aligned}$$

with μ_i defined as above and $\eta_i = \phi_i/(1 + \phi_i)$. Clearly, this distribution does not belong to the exponential family, nor does it yield low-dimensional sufficient statistics for the unknown parameters. Hence, conditional likelihood approaches will not work. Nor are marginal likelihood approaches useful, since it is difficult to find a suitable prior distribution on the nuisance parameters ϕ_i. In spite of these problems, a number of authors have used maximum likelihood based on the beta-binomial distribution to fit dose–response models to developmental toxicity data. Chen and Kodell (1989), for example, fit a three-parameter model of the form (2) with $h(x) = 1 - \exp(-x)$, allowing a separate ϕ_i for each dose group. As cautioned by Kupper, Portier, Hogan and Yamamoto (1986), it is crucial to allow ϕ to change with dose when fitting a beta-binomial distribution to teratology data, since, otherwise, estimators of the parameters of interest may be seriously biased.

In a spirit similar to that conveyed by Liang and Zeger, Williams (1988) suggested in response to Kupper et al. that bias caused by mismodelling the correlation can be largely avoided by using quasi-likelihood rather than maximum likelihood to fit a model with the beta-binomial mean and variance structure. He described a simple iterative algorithm (Williams, 1982) to fit such a model in standard statistical packages (he suggested GLIM). In fact, his proposal corresponds to fitting equation (4.1) from

Liang and Zeger, with a common ϕ estimated from the model deviance, divided by the degrees of freedom remaining after estimating the mean parameters. Quasilikelihood and estimating equations work well for the beta-binomial distribution because the moments take a particularly simple form:

$$\mathbb{E}(Y_{ij}) = n_{ij}\mu_i$$

and

$$\mathrm{Var}(Y_{ij}) = n_{ij}\mu_i(1-\mu_i)[1+\phi_i(n_{ij}-1)].$$

Clearly, the function $g_i = g_i(y_i, \theta) = (y_{ij} - n_{ij}\mu_i)$ has expected value 0. Hence the estimating equation (4.1) from Liang and Zeger takes the form

$$\mathbf{g} = \sum_{i=0}^{I}\sum_{j=1}^{N_i} \frac{\partial\mu}{\partial\boldsymbol{\theta}} \frac{(y_{ij} - n_{ij}\mu_i)}{n_{ij}\mu_i(1-\mu_i)[1+\phi_i(n_{ij}-1)]} = \mathbf{0}.$$

The particular form of $\mathbf{g}$ will depend on the function $h(\cdot)$ linking the mean response rate to exposure and other covariates. For the logistic model, for example,

$$\mathbf{g} = \sum_{i=0}^{I}\sum_{j=1}^{N_i} \begin{pmatrix} 1 \\ d_i^{\theta_2} \\ \theta_1 d_i^{\theta_2}\log(d_i) \end{pmatrix} \frac{(y_{ij} - n_{ij}\mu_i)}{[1+\phi_i(n_{ij}-1)]} = 0.$$

Note that these equations depend on the unknown nuisance parameters ϕ_i, $i = 0, \ldots, I$. Since alternative "pivotal" or "information unbiased" $\mathbf{g}$-functions are not readily apparent, it it useful to appeal to the argument of Liang and Zeger that estimation of the nuisance parameters will have relatively little impact on the solution to $\mathbf{g} = \mathbf{0}$.

To illustrate the impact of estimating nuisance parameters on the estimation of the parameters of interest, maximum likelihood and estimating equations were both used to fit a dose–response model to the same data reported by Chen and Kodell (1989). The data were from a study of an industrial plasticizer, DEHP, and included a control and four dose groups exposed to (0.44, 0.091, 0.191 and 0.292 g/kg). The study contained a total of 131 dams with an average of 12 offspring per litter (range 1 to 19). All models were fitted using the S-PLUS function nlminb. The beta-binomial model can be easily fitted using this function along with a user-defined function to calculate the log-likelihood. Estimating equations can also be easily solved with nlminb along with a user-defined function to calculate the inner product of the estimating equations. An advantage of this approach is that it easily allows for nonlinear mean functions, such as the one specified in (2). Other approaches based on linearizing approximations are possible (see Ryan, 1992a). Figure 2 provides a graphical representation of the data, with each dot corresponding to the response rate for a particular litter, while the crosses show the overall response rate within each dose group.

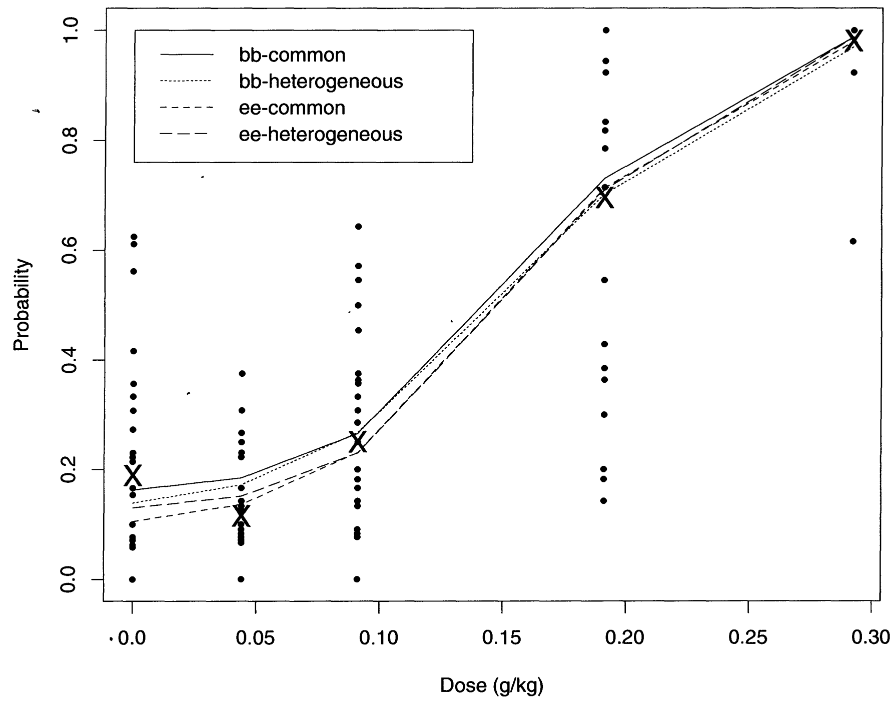

FIG. 2. *Observed and predicted response rates—DEHP data.*

The lines shown in the figure correspond to various fitted values, to be discussed presently. The full data set can be found in Chen and Kodell (1989). Table 1 shows the results of fitting a dose–response model of the form (2), using the logistic function for $h(x)$.

The results in Table 1 show that, under the maximum likelihood approach, the estimated parameters characterizing the mean response rate change substantially, according to whether or not ϕ is allowed to change with dose group. In contrast, estimates based on the estimating equations, while affected to a degree, do not change so dramatically. Further insight can be gained by examining the fitted lines in Figure 2, corresponding to the four model fits summarized in Table 1. This conclusion accords with the results of Liang and Hanfelt (1994), who addressed similar questions for the analysis of teratology data, though without the "power parameter," θ_2, and only for the two group case.

SOME OTHER ISSUES

The question of how to handle correlated binary data is one of the simpler ones that arises in the context of developmental toxicology. The remainder of this discussion focuses on some other more complicated issues where further work is needed.

Testing for and quantifying dose response is but the first step in the complex process of quantitative risk assessment. Another important step is the calculation of a "benchmark dose," defined by Crump (1984) as a lower confidence limit on the dose that corresponds to a specified increased risk above background. Suppose, for example, that the mean response rate has the following functional relationship to dose (d):

$$\mu(d) = 1 - \exp[-(\theta_0 + \theta_1 d^{\theta_2})].$$

It follows that the dose level corresponding to a $q\%$ increased risk above background will be

$$ED_q = \left(\frac{-\log(1-(q/100)\exp(\theta_0))}{\theta_1}\right)^{1/\theta_2}.$$

One could argue that if the ED_q is really the quantity of interest, then the nuisance parameters include not only the ϕ_i's, but also the θ's. Hence, an important and useful question is how to estimate ED_q, while minimizing any bias caused by simultaneous estimation of the remaining nuisance parameters. While the class of models defined at (2) is widely used in practice, it can be difficult to apply and often encounters convergence problems. The main problem is the relatively high correlation between the estimates of θ_1 and θ_2. The results in the table illustrated this point because the estimated values of θ_1 and θ_2 vary considerably, yet the fitted dose–response curves are all reasonably similar. Better methods are clearly needed.

Another problem related to estimation of the ED_q is that regulators are interested in a lower confidence bound on this quantity. In the likelihood-based setting, most people prefer the use of a likelihood-based confidence interval since, in general, these will have better coverage properties (see Chen and Kodell, 1989). Unfortunately, there is no analogue of a likelihood-based confidence interval presently available for estimation based on estimating equations. Development of such approaches would be very useful.

One of the most challenging and interesting aspects of analyzing developmental toxicity data is the complex nature of the outcomes of interest. While our discussion so far has assumed a binary outcome indicating whether or not each fetus is defective, reality is more complex. Figure 3 illustrates some of the outcomes measured in a typical experiment: offspring may die early in gestation and be resorbed; they may die later and be recorded as a fetal death; they may survive and develop any of several different malformation types, or they may have low birth weight. Finding ways to capture the effects of exposure on this multivariate outcome, as well as adjusting for intralitter correlation provides a rich source of interesting statistical problems. To characterize the effects of exposure on death and malformation, several authors have sug-

TABLE 1
Model fitting for DEHP data

	Maximum likelihood				Estimating equation			
	Common ϕ		Dose-specific ϕ		Common ϕ		Dose-specific ϕ	
	Estimator	(se)	Estimator	(se)	Estimator	(se)	Estimator	(se)
θ_0	−1.64	(0.165)	−1.90	(0.179)	−1.83	(0.122)	−2.13	(0.243)
θ_1	66.31	(29.727)	62.87	(28.964)	38.03	(8.764)	42.91	(19.182)
θ_2	1.95	(0.315)	1.88	(0.312)	1.60	(0.164)	1.59	(0.320)
ϕ	0.19		(19, 0.01, 0.09, 0.38, 0.26)		0.19		(0.50, 0, 0.11, 0.41, 0.24)	

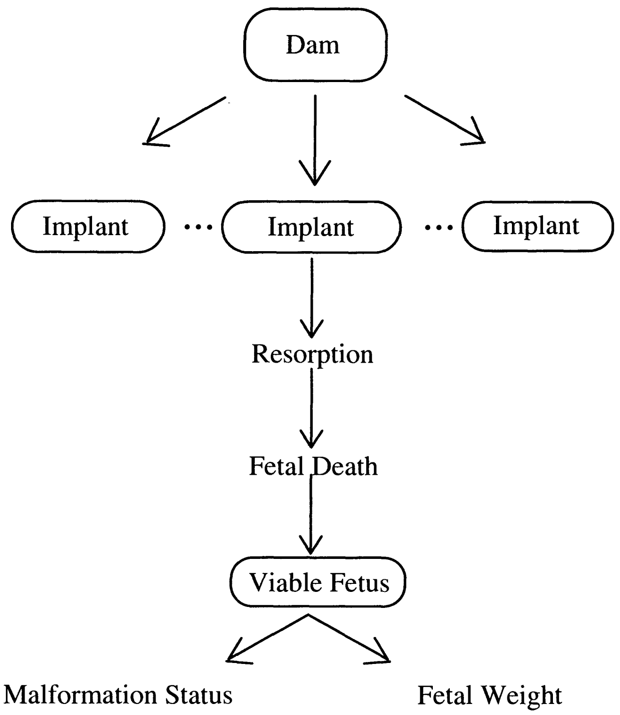

FIG. 3. *Fetal outcomes in developmental toxicity studies*

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