$\alpha(a) = 0.05$ 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.