tingency tables are: (i) jackknife-type perturbations that decrease each count by one in turn, (ii) perturbations that involve increasing or decreasing each count by one in turn, (iii) perturbations loosely based on a notion of misclassification that preserves the total sample size but reallocates up to a certain fraction of the observations, and (iv) more general perturbations that need not preserve total sample size and also permit more than one cell count to alter. Schemes such as (i) and (ii) have a certain natural appeal in moderate-to-large size contingency tables—one would like to think that changing just one cell by one count could not seriously effect the p-value. In a context where protection against misclassification is desired, a scheme of type (iii) is appropriate. A misclassification-based scheme may differ depending on what, if any, margins are fixed by the sampling design. For example, in a 2 × 2 table with fixed column margin, one may be primarily concerned with potential errors in row classification. In other words, one could want a perturbation scheme that preserved the column margin.

Working with a set of approximately one dozen real examples of 2 × 2 tables culled from assorted textbooks, the effects of perturbation schemes of types (i), (ii) and (iii) on Fisher's exact test were studied. Denote the actual p-value by P and the minimum and maximum p-values achieved over the set of perturbations by P_L and P_U, respectively. The following tentative conclusions rest on this limited experience; in the interests of space, I will illustrate the points exclusively with respect to the table given by Agresti in Section 2.1 having counts by row (10, 90/20, 80). Scheme (ii) may be preferable to scheme (i), as there are cases where scheme (i) alters the p-value in only one direction so that one of P_L or P_U equals P, while scheme (ii) has $P_L \neq P$ and $P_U \neq P$. For (10, 90/20, 80), $P = 0.073$; under scheme (i), $P_L = 0.043$ and $P_U = P = 0.073$; under scheme (ii), $P_L = 0.043$ and $P_U = 0.082$. Under scheme (ii), it is frequently, but by no means always, the case that P_L and P_U are achieved by increasing and decreasing the same cell. For (10, 90/20, 80), this is the case with P_L arising from the table (9, 90/20, 80) and P_U arising from the table (11, 90/20, 80). Scheme (iii) often leads to a much wider range of p-values. For (10,90/20,80), moving one count under scheme (iii), $P_L = 0.028$ for (9, 90/21, 80) and $P_U = 0.117$ for (11, 89/20, 80). In the event that scheme (iii) is restricted to perturbations that preserve the row margin, $P_L = 0.043$ for (9, 91/20, 80) and $P_U = 0.117$ is unchanged.

It is possible that further work on the sensitivity of exact inference may lead to rough guidelines on, say, the percentage change in p-value corresponding to some set of perturbations for the data (see Dupont, 1986). For the present time, however, it would be helpful if software packages were setup to easily permit sensitivity analysis based on these or other perturbation schemes. Developing techniques to permit efficient sensitivity analyses (i.e., without repeating the computation for each perturbed table) would be a useful area for research.

Comment

Leonardo D. Epstein and Stephen E. Fienberg

We would like to congratulate Professor Agresti for his thorough review of the recent literature on exact inference in contingency tables and for organizing it in a way that allows us to focus on some key statistical issues. Our first observation relates to the work "exact," which has an everyday meaning that may not coincide with its technical meaning in the present context. It is also a value-laden descriptor that suggests that any statistical method that is not exact may not be very good. As the following comments imply, nothing could be further from the truth.

The most widely studied problem involving categorical data, and seemingly the simplest, is that of drawing inferences for the risk ratio and risk difference in 2 × 2 contingency tables. Yet, this simple situation highlights many of the most controversial aspects of statistical methodology and theory. Before discussing these issues, we note that there are few practical statistical problems that come in the simple form of a 2 × 2 table. Most investigations involve a large number of variables, both continu-

Leonardo D. Epstein is an Assistant Professor in the Department of Biostatistics at the Johns Hopkins University, 615 North Wolfe Street, Baltimore, Maryland 21205. Stephen E. Fienberg is a Professor of Statistics and Law, and Vice President of Academic Affairs at York University, 4700 Keele Street, North York, Ontario M3J 1P3, Canada.

ous and categorical, and the 2 × 2 table that investigators claim to be concerned with is actually a collapsed version of the actual data array, and inferential issues in the 2 × 2 table must be looked at as they are imbedded in the broader inference context. Agresti's paper does go beyond the 2 × 2 setting, but in doing so it shows the murkiness of inference issues in the real-world settings that most practicing statisticians face on a day-to-day basis. We now turn to some of the more technical and focused issues of inference.

What do we mean by exact inference? The word "exact" in the context of statistical methodology has come to take on a pair of intertwined but specific meanings, i.e., small (as opposed to large) sample distributional features on the one hand, and conditional (as opposed to unconditional) inference where the role of conditioning on "other" sufficient statistics is justified on various grounds. We believe that these issues need to be separated and that each needs to be addressed as to its relevance.

As Agresti makes abundantly clear, once we get beyond the 2 × 2 table, we almost always resort to approximations, usually of a large sample nature. Agresti notes the relevance of alternative assumptions for large, sparse contingency tables, where the number of parameters increases with the sample size; but he dismisses this alternative a bit too quickly. For example, the results of Haberman (1977) on conditional likelihood ratio tests involving small numbers of degrees of freedom, in this large, sparse asymptotics, suggest that influence for a single parameter might be well handled in this alternative framework. We also come away from our rereading of the literature cited by Agresti on this issue feeling more optimistic about the adequacy of large-sample approximations, especially if we leave aside the conditioning on marginal totals.

Next, we turn to the issue of conditioning. The weaknesses of conditioning on an ancillary statistics, when such a procedure is conceived either as a principle or as a helpful technique, is a controversial matter. Fisher's exact test for independence and other procedures for exact inferences regarding the odds ratio condition on the marginal totals of the 2 × 2 table; however, for the risk ratio and risk difference, the marginal total is not an ancillary statistic, and exact procedures are unavailable (Thomas and Gart, 1978; Santner and Snell, 1980). Bandorff-Nielsen (1976) shows the same result for a related concept of M-ancillarity. This actually points to serious difficulties underlying the implementation of procedures that condition on an ancillary statistic. For example, the existence of an ancillary statistic depends on the parametrization being used. Also, even if an ancillary statistic does exist, it may not be unique, and thus the "theory" does not provide and infallible guide. Furthermore, alternative attempts at providing a "logical justification" of conditional tests for 2 × 2 tables [e.g., by Greenland (1991)] are heuristic at best.

Plackett (1977) addressed an issue assumed by Fisher (1935) when he originally developed the exact test for 2 × 2 tables. As Plackett notes, Fisher did not actually say that one should condition on the marginal totals because they are ancillary and contain no information regarding the odds ratio. Rather, Fisher asked the reader to assume that this were the case. Plackett shows that, for finite samples, there is some information in the margins, in effect because of "boundary effects." As a consequence, the appeal to ancillarity, even when interest is focused on the odds-ratio, does have an element of approximation normally ignored by those who favor conditioning.

In addition to questioning the general applicability of conditioning on ancillary statistics, we believe that questions should be asked about hypothesis tests in general, and p-values in particular, and whether they provide, even in the best of cases, a sensible approximate measure of the adequacy of a null hypothesis vis-à-vis an alternative. These questions carry over to confidence regions derived by inverting exact hypothesis tests. Below, we discuss these points in mode detail.

Tests constructed on the basis of efficient scores are approximate, even if all the involved computations are exact. We follow Cox and Hinkley (1974, page 106). Suppose a random variable Y has density $f(y/\theta)$, where θ is unidimensional. We write the likelihood function for data y as $l_y(\theta)$. The likelihood ratio, $\mathrm{lr}(\theta_1, \theta_0)(y) = l_y(\theta_1)/l_y(\theta_0)$, is used to test the simple null hypothesis H_0: $\theta = \theta_0$ versus the simple alternative H_1: $\theta = \theta_1$. The likelihood ratio critical region is obtained from large values of $\mathrm{lr}(\theta_1, \theta_0)(y)$. Suppose, however, that our interest is focused on testing H_0: $\theta = \theta_0$ versus H_A: $\theta > \theta_0$. Efficient scores are used to approximate the likelihood ratio when the following approach is used for this testing problem. Writing the particular alternative $\theta_A = \theta_0 + \delta$, where $\delta > 0$ is small, we have

$$(1)\quad \log\frac{l_y(\theta_0 + \delta)}{l_y(\theta_0)} = \delta\frac{d}{d\theta}\log l_y(\theta_0) + o(\delta).$$

For sufficiently small $\delta > 0$, an approximate likelihood ratio critical region is obtained from large values of $d/d\theta \log l_y(\theta_0)$. The random variable $U = d/d\theta \log l_y(\theta_0)$ is referred to as the efficient score for Y. Therefore, when in an hypothesis testing problem efficient scores are used instead of the likelihood ratio, the use of the approximation (1) is implicit.

Typical examples of two-sided alternatives are "the odds-ratio is not one" and "β is not zero," where β is a parameter in a logistic-linear model. In such situations, no uniformly most powerful test typically exists (Cox and Hinkley, 1974, page 105). This is the case whether one conditions on ancillary statistics or not. Also, there is no agreement among practictioners on how to compute p-values when the alternative is two-sided. For example, many practitioners would compute the p-value as $P(|T| \geq |t_{obs}| \mid H_0)$, where T is the test statistic. Cox and Hinkley (1974, page 106), on the other hand, suggest a slightly different procedure based on $p^+_{obs} = P(T \geq t_{obs} \mid H_0)$ and $p^-_{obs} = P(T \leq t_{obs} \mid H_0)$. They recommend reporting $p_{obs} = P(M \leq m_{obs} \mid H_0)$ as the p-value, where $m_{obs} = \min(p^+_{obs}, p^-_{obs})$ and $M = \min(P^+, P^-)$. For discrete distributions, this choice leads to a value different from $P(|T| \geq |t_{obs}| \mid H_0)$.

The extent to which exact p-values should or actually do contribute substantially to a statistical analysis is problematic. Our point is probably best summarized in a discussion on significance tests by Cox and Hinkley (1974, page 66), who write, "Further, if strong evidence against H_0 is obtained, the significance test gives no more than a guide to the type and magnitude of the departure."

These difficulties of interpretation are compounded in model building strategies that apply several tests to the same data. A variety of books and articles (e.g., Andersen, 1974) recommend and reinforce building models by applying simultaneously several tests using the same α-level, one test for each parameter present in the current working model. For example, a frequently used strategy is to test successively that the parameters in a model are zero. If the p-value of the least significant parameter is above 0.05, say, then the predictor variable corresponding to that parameter is removed from the current working model. There are several difficulties with the interpretation of p-values when one uses this or any other procedure that applies several tests simultaneously or sequentially. First, if one knew how many tests one would apply and if the test statistics were independent, then one could obtain the resulting overall p-value, which will be larger than the nominal p-value for each of the tests involved. Second, if one knew the number of tests to be applied but if the test statistics are not independent, the situation is much more difficult and most of the time one can only obtain an upper bound for the overall p-value. Thus, the availability of exact p-values for the intermediate tests would not be much of a contribution to the statistical analysis. Finally, and most importantly, one does not know in advance how many tests will be performed in building and specifying a model. The uncertainty about the number of tests to be applied is usually ignored, and the interpretation of the overall p-value is even more obscure. This leaves the relevance of exact inferences for individual tests open to greater question.

We end our discussion with an additional word of caution about the interpretation of statistical tests. Statisticians have been pointing out for some time to the inadequacy of statistical tests to obtain measures of evidence about an hypothesis [for example, see Berger and Sellke (1987) and references there for a critique of p-values as measures of evidence, and Royall (1991) for a general critique of both Neyman–Pearson tests and p-values from an evidential perspective], Berger and Sellke (1987) argue that, in general, p-values overstate the evidence provided by the data against the null hypothesis. Therefore, even if one can compute p-values exactly, they are interpretable as a measure of evidence only in a very approximate sense.

This notion of measures of evidence leads rather naturally to the Bayesian perspective for inference. Many statisticians acknowledge the foundational superiority of Bayesian methods; however, they feel that computational difficulties keep them from applying these methods in practice. It appears to us that the computational difficulties involved in implementing exact tests, although of different nature, are as serious as those involved in implementing Bayesian methods. Today, at least the most common problems involving discrete data have been addressed, in one way or another, from a Bayesian perspective. For a surprising number of these problems, posterior probabilities and moments can be computed exactly or up to a prespecified error bound. Among the relevant recent papers on Bayesian inference for loglinear models not referenced by Agresti are Spiegelhalter and Smith (1982), Raftery (1986), and Epstein and Fienberg (1991).

Agresti's paper reviews and summarizes the recent literature on exact tests and inference. He focuses in large part on technical aspects, especially in the discussions of multidimensional tables and logistic regression. Here, we have restricted our comments to what we see as the more fundamental questions, many of which are also raised by Agresti. After a careful study of Agresti's excellent review, we remain unconvinced by the usual answers to these questions proffered by supporters of exact inference.