cases, is there any reason that conditional coverage should be desired?

In summary, for the first model, pr($Y \mid S$) is the reference distribution for model checking and pr($S; \phi$) is the reference distribution for inference about the parameter ϕ. For the second model, pr(A), where A represents the ancillary component of $S = Y$, is the reference distribution for model checking and pr($S \mid A; \pi$) is the reference distribution for inference about the parameter π. For the third model, model checking is not possible, and pr($Y; \alpha$) is the reference distribution for inference about any parameter of interest (i.e., any function of α).

Comment

Diana E. Duffy

1. INTRODUCTION

Professor Agresti is to be congratulated for a well-written and timely survey on exact conditional inference for contingency tables. At this point in time, 8 to 10 years after some of the key initial advances in computing strategies (Pagano and Halvorsen, 1981; Mehta and Patel, 1983; Pagano and Tritchler, 1983a, 1983b), it is instructive to take stock of both where the field is presently and where the field may be headed. For practitioners and applied statisticians, Agresti offers a practical introduction to currently available exact methods. For researchers in methodology and in statistical computing and algorithms. Agresti offers directions for possible future research.

Exact conditional inference for contingency tables involves assessing the exact (discrete) sampling distribution of test statistics and parameter estimates of interest after conditioning on the sufficient statistics for nuisance parameters. The sufficient statistics for nuisance parameters correspond directly to certain margins in the corresponding contingency table; as long as one operates within the arena of loglinear models, conditioning on these margins will eliminate the nuisance parameters. The exact sampling distribution of interest is then the distribution over all possible tables that could be observed with certain fixed margins (i.e., those margins fixed by the sampling design plus those margins fixed by the conditioning). I will refer to this set of tables as the conditional reference set. It is worth noting the following correspondence between conditional and unconditional problems: the conditional reference set for a problem with a set S_1 of margins fixed by the sampling design and a set S_2 of margins fixed by conditioning is identical to the sample space for a (different) problem in which margins in both S_1 and S_2 are fixed by the sampling design. For example, the conditional reference set for testing independence in a 2×2 table under product binomial sampling (one-fixed margin) is equivalent to the sample space for a 2×2 table with both margins fixed.

In this commentary, I would like to expand on two areas that offer challenges for future work. Throughout this discussion, I adopt Agresti's notation as described in his Section 1.2 in toto, and I refer to points in his paper by simply giving his name and the section number.

2. BAYES AND RELATED INFERENCES

The existing literature on Bayesian methods for analyzing contingency tables dates at least to Lindley (1964). One way to categorize the proposed methods is through the choice of prior. The simplest methods are those for 2×2 tables under product binomial sampling with beta priors; see Altham (1969, 1971) for examples. Generalizations to full multinomial sampling and to $I \times J$ tables for I or $J > 2$ lead to Dirichlet priors on the cell probabilities. These are discussed in Good (1967, 1975, 1976), Good and Crook (1974), Gunel and Dickey (1974), Crook and Good (1980), and Albert and Gupta (1982, 1983a, 1983b). Normal priors on logarithmic functions of the cell probabilities are discussed in Leonard (1975) and Nazaret (1987). Empirical Bayes analogs of the Dirichlet and normal approaches are described in Albert (1987) and Laird

Diane E. Duffy is Director, Statistics and Data Analysis Research Group, Bellcore, 445 South Street, Morristown, New Jersey 07962-1910.

(1978), respectively. In most of these cases, the emphasis is on noninformative priors (although the work of Albert and Gupta is a notable exception). The issue of the sampling design and which margins are fixed by it is raised from the beginning by Lindley (1964) and figures very prominently in the work of Crook and Good.

In general, Bayesian analysis involves modeling prior knowledge about parameters, updating this knowledge in light of the observations, and making inferences based on the resulting posterior distribution for the parameters. In this context, nuisance parameters are not a nuisance conceptually—one simply integrates them out of the joint posterior distribution to produce a marginal posterior distribution for the parameters of interest. Practically, however, nuisance parameters may indeed be a nuisance if it is difficult to obtain a prior distribution for them or if it is difficult to integrate them out of the posterior. (The first difficulty, i.e., obtaining a prior distribution, may be overcome by choosing a suitable noninformative prior.)

Philosophical issues (completely) aside, it is certainly possible to consider a conditional Bayesian analysis in which nuisance parameters are first dealt with by conditioning and then a prior distribution is chosen for the remaining parameters. Although such an approach does mix apples and oranges to some extent, it may offer an interesting perspective, particularly in the current context due to the aforementioned correspondence between the conditional reference set for a given problem and the sample space for another problem. The following simple example serves to illustrate a possible conditional Bayesian approach to testing a hypothesis in a two-way table.

Under full multinominal sampling, the conditional reference set for a test of the model (X) versus the model (X, Y) in a 2×2 table is based on the distribution of $\{n_{+1}, n_{+2}\}$ given $\{n_{1+}, n_{2+}\}$. This distribution depends on the conditional probabilities of falling in the first column for each of the two rows, that is, on $\pi_1 = p_{11}/p_{1+}$ and $\pi_2 = p_{21}/p_{2+}$, and is given by

$$\begin{aligned}\Pr(n_{+1} = x \mid n_{1+}, n_{2+}, \pi_1, \pi_2) \\ = \sum_{n_{11} \in T(x)} \operatorname{bin}(n_{11} \mid n_{1+}, \pi_1) \\ \cdot \operatorname{bin}(x - n_{11} \mid n_{2+}, \pi_2)\end{aligned} \tag{1}$$

for $0 \le x \le n$, where $\operatorname{bin}(y \mid n, p)$ denotes the usual binomial density and where $T(x) = \{t \text{ integer: } \max(0, x - n_{2+}) \le t \le \min(n_{+1}, x)\}$. The null hypothesis is H_0: $\pi_1 = \pi_2 = 0.5$ and the alternative hypothesis is unrestricted. For a conditional Bayesian test, one starts by modeling prior belief in π_1 and π_2. Given the form of (1), one is tempted to consider a product of independent beta distributions, one for each of π_1 and π_2, because the beta is the natural conjugate prior for the binomial. In this context, however, one cannot simply put beta priors on π_1 and π_2, as these priors put zero mass on the null hypothesis. Instead, a prior probability for H_0 is chosen, and the remaining probability is spread out over values of π_1 and π_2 not satisfying H_0 (see Berger, 1985, pages 149–150). Given an appropriate prior distribution, the test is conducted by calculating the posterior probability of the null hypothesis. In this case, the null hypothesis concerns both π_1 and π_2, so that the posterior probability of the null is calculated directly from the joint posterior distribution. [The reader may note the similarities between the above example and the problem considered by Altham (1969) and described by Agresti, Section 8.3. There are, however, two key differences. First, Altham's null and alternative hypotheses are different and permit the use of simple beta priors. Second, Altham's test is an unconditional test under a product binomial sampling model, whereas the test described above is a conditional test under full multinomial sampling. These two problems both lead to consideration of a 2×2 table with one fixed margin and illustrate the aforementioned correspondence between certain conditional reference sets and sample spaces.]

The conditional Bayesian test previously described can be compared to the usual (unconditional) Bayesian test for the same problem. For this test, the null and alternative hypotheses are given by: H_0: $p_{11} = p_{12}, p_{21} = p_{22}$ and H_1: $p_{ij} = p_{i+}\, p_{+j}$. There are at least two possible approaches to the choice of a prior distribution. One can model prior opinion in terms of the two parameters p_{1+} and p_{+1}; these parameters determine the cell probabilities through the equations $p_{11} = p_{1+}\, p_{+1}$, $p_{12} = p_{1+}(1 - p_{+1})$, $p_{21} = (1 - p_{1+})p_{+1}$ and $p_{22} = (1 - p_{+1})(1 - p_{+1})$ and range (independently) over $[0, 1]$. Again, independent beta distributions for p_{1+} and p_{+1} are conjugate, but give zero mass to the null hypothesis that is equivalent to $p_{+1} = 0.5$ (and, thus, cannot be used outright). Alternately, one can model prior opinion in terms of the parameters in the loglinear representation of the cell probabilities. Let $l_{ij} = \log(p_{ij})$ and recall that the alternative hypothesis is equivalent to $l_{ij} = \lambda + \lambda_i^1 + \lambda_j^2$ with the identifiability constraints $\lambda_2^1 = \lambda_2^2 = 0$. In order to ensure that the cell probabilities sum to 1, it is standard to set $\lambda = \log[1 + \exp(\lambda_1^1) + \exp(\lambda_1^2) + \exp(\lambda_1^1 + \lambda_1^2)]$. This leaves two free parameters, λ_1^1, and λ_1^2, each of which ranges on $\mathbf{R}$. The null hypothesis is equivalent to $\lambda_1^2 = 0$, and any prior giving positive mass to the set of $\{\lambda_1^1, \lambda_1^2 = 0\}$ can be considered. In either case, the null hypothesis

concerns just one of the parameters, so that the relevant marginal posterior distribution is used to compute the null posterior probability.

In the example considered in the previous paragraphs, neither the conditional nor the unconditional Bayesian test admits a simple natural conjugate prior distribution. This is clearly not always the case. For example, Altham's (1969) unconditional test of $\pi_1 = \pi_2$ under product binomial sampling makes use of beta conjugate priors. Altham's test corresponds to the null model (X, Y), the alternative model (XY), and a sampling design that fixes the column totals (i.e., the Y margin). It is interesting to note that a conditional Bayesian test of this same model concerns the odds ratio θ and the noncentral hypergeometric distribution [Agresti, Section 1.3, equation (1.2)]. More specifically, with $g(\theta)$ a prior density for θ (supported on $[1, \infty)$ with atom at $\theta = 1$), the test of H_0: $\theta = 1$ against the one-sided alternative H_1: $\theta > 1$ has posterior probability for H_0 given by

$$(2) \qquad \frac{g(1) f(n_{11} \mid n, n_{1+}, n_{+1}; \theta = 1)}{\int_1^\infty g(\theta) f(n_{11} \mid n, n_{1+}, n_{+1}; \theta)\, d\theta}$$

where the function f is given in Agresti's equation (1.2) and where the integral in the denominator is understood to include the summand corresponding to the atom at $\theta = 1$. Because of the complexity of the noncentral hypergeometric distribution as a function of θ, there is no natural candidate for the prior $g(\theta)$. The distribution on θ induced by independent beta distributions on π_1 and π_2 is one possibility (see Nurminen and Mutanen, 1987, for the form of the distribution).

The previous paragraphs raise two interesting questions. First, some of the Bayesian techniques involve enumeration of a set of tables with certain fixed margins (see, e.g., Good, 1976). Can any of the sophisticated algorithms for calculating and approximating exact p-values be adapted to aid Bayesian computations? Second, how do the conditional Bayesian tests compare to the more standard (unconditional) Bayesian tests and to the frequentist tests, both in general and in the specific case where noninformative priors are chosen? The work of Good (1976) and Crook and Good (1980) indicates that the Bayes factor for the test of independence in a two-way table is not very highly dependent on which margins are fixed when the prior is chosen from the symmetric Dirichlet class. Does this lack of dependence on fixed margins hold true more broadly so that the very issue of conditioning or not is less important in the Bayesian context? (Recall that, in the classical context, conditional and unconditional analyses can lead to vastly different conclusions; see Agresti, Section 8.2.) One might begin to address some of these questions in the simple context of testing independence in two-way tables by comparing the Bayes factors that Good (1976) and Crook and Good (1980) obtain for their Model 3 (both margins fixed by sampling design) to the Bayes factor resulting from (2) with some sort of noninformative prior $g(\theta)$.

3. SENSITIVITY ANALYSIS

As discussed by Dupont (1986) and noted by Agresti in Section 2.1, the p-value for Fisher's exact test can be quite sensitive to small perturbations in the observed data. In Agresti's example, adding one count to a 2×2 table with 200 counts changes a two-sided p-value from 0.073 to 0.050. This could well be a practically significant change given the all-too-common practice of using 5% significance cut-offs. Dupont generates nearly 1,000 tables, with total counts ranging from 60 to 498 that satisfy a certain set of criteria. He compares each table to a table with one additional count (in a specific cell) and shows both that this sensitivity of the usual p-value for Fisher's exact test is widespread, and that the alternate method of doubling the one-sided p-value to obtain a two-sided test is considerably more robust.

One open question is the extent to which the sensitivity noted for Fisher's exact test is also present for other exact p-values. One could expect exact inferences to be highly sensitive to data perturbations whenever the sampling distribution is highly discrete (i.e., it has most of its mass concentrated on a relatively few points). In situations where high sensitivity is a concern, it is imperative that analysts be able to assess the degree of sensitivity present in their data. In a given application, assessing the degree of sensitivity of an inference to small perturbations of the data first requires a defined notion of "small perturbations." For exact inferences in contingency tables, there are two basic levels of perturbation: those that do not change the conditional reference set and those that do. For example, in a test of independence in a 2×2 table, perturbations that do not change the conditional reference set are those that preserve both margins, i.e., they add an amount x (possibly negative) to both cells of one diagonal and subtract x from both cells of the other diagonal. Any other perturbation scheme in the 2×2 case involves changing the conditional reference set. Schemes that do not change the conditional reference set do not have much appeal and are not considered further here.

Among possible perturbation schemes for con-

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.