DISCUSSION

J.N.K. Rao, Carleton University, Ottawa Canada

All six papers in this session make important contributions to the analysis of categorical data from complex surveys. The authors suggest methods which take proper account of the survey design in the analysis of data. As demonstrated recently by Altham (1976), Brier (1978), Cohen (1976), Fellegi (1980),Rao and Scott (1980), and others, the effect of clustering in the survey design could have substantial impact on the significance level of multinomial-based chisquare tests, e.g., corresponding to a nominal level of 5% the achieved significance level can be as high as 40% or higher, as shown by Rao and Scott (1980) for some data from the 1971 General Household Survey of the U.K. The current popularity of loglinear models together with multinomial-based test statistics makes it all the more important to study the impact of survey design and suggest alternatives which are either asymptotically valid (like the Wald statistic, when a consistent estimator of the covariance matrix of ultimate cell estimates is available) or provide simple corrections to multinomial-based test statistics which minimize the distortion in the achieved significance level.

1. G.G. Koch, M.E. Stokes and D. Brock: Koch and his associates developed asymptotically valid methods (based on weighted regression and the Wald statistic) for the analysis of cross-classified data, including "domain" means, and provided extensive applications of their techniques to large-scale survey data, in particular to data from health surveys. The present paper is a welcome contribution as it clearly illustrates their methods by providing annotated computer output. Inspite of the availability of these applications and computer programs, the effect of survey design is still often ignored even when the necessary data for getting consistent estimators of the covariance matrix were available.

2. R.E. Fay III: In a previous paper presented at the ASA meetings in Washington, D.C., 1979, Fay suggested a jackknife chisquare statistic for designs where sample estimate of a cell can be expressed as sum of independent estimates, i.e. the sample is composed of independent replications. Fay argued that the jackknife statistic is preferable over the Wald statistic when the number of ultimate cells is large since the inversion of cell covariance matrix $\underline{V}$ may become unstable. In the present paper the jackknife statistic is extended to more useful replication methods like BRR (balanced repeated replication).

Fay's point on the instability of $\underline{V}^{-1}$ is well-taken, but the size of the matrix involved in Wald's statistic depends only on the dimension, f, of the hypothesis H: $\underline{h}(\underline{p})=0$ and is given by $\underline{V}_h^{-1} = [\underline{H}(\underline{p})\ \underline{V}\underline{H}(p)']^{-1}$ where $\underline{H}(\underline{p})=(\partial h_i(\underline{p})/\partial p_j)$. For certain hypotheses, the size of $\underline{V}_h$ could be much smaller than that of $\underline{V}$. Moreover, a consistent estimator of $\underline{V}_h$ can be obtained directly from BRR or the jackknife, thus avoiding the calculation of derivative matrix $\underline{H}(\underline{p})$.

As an alternative to Fay's jackknife statistic, a simple correction to Pearson X^2 might be adequate for most purposes, which also avoids the inversion of $\underline{V}_h$. The corrected statistic is given by $X^2/\bar{\delta}$ where $f\bar{\delta}$ equals the estimated asymptotic expectation of $X^2 = n\Sigma(\hat{p}_i-p_i(\hat{\theta}))^2/p_i(\hat{\theta})$ under H, i.e.

$$f\bar{\delta} = n\Sigma\ \hat{Var}(\hat{p}_i-p_i(\hat{\theta}))/p_i(\hat{\theta}). \quad (1)$$

For example, in the case of testing independence in an r x c table: $p_{ij} = p_i\ p_j$, $i = 1, \ldots, r-1$; $j = 1, \ldots, c-1$,

$$f\bar{\delta} = n\Sigma\Sigma\ \hat{v}_{ij}(\underline{h})/[\hat{p}_i\hat{p}_j]$$

where $\hat{v}_{ij}(\underline{h})$ is the jack-knife or BRR variance estimator of $h_{ij}(\hat{\underline{p}}) = \hat{p}_{ij}-\hat{p}_i\hat{p}_j$, and $f=(r-1)(c-1)$. In the general case (1), one could get a jackknife or BRR estimate $\hat{var}(\hat{p}_i-p_i(\hat{\theta})$ by computing the estimates of θ from each pseudo-replication, as in the case of Fay's jackknife statistic. A Ph.D. student, G. Roberts, is presently investigating the properties of $X^2/\bar{\delta}$ relative to Fay's statistic.

3. V. Richards and D.H. Freeman, Jr.: Direct replication and BRR methods of variance estimation are compared for the analysis of contingency tables, using data from the Conn. High Blood Pressure (CHBP) Survey. Direct replication gave smaller standard errors and led to a log-linear model different from the model chosen via BRR. However, the sample design in the non-certainty strata does not seem to permit replication even with independent subsampling in sampled psu's, since only one psu is selected from a stratum. Suppose $\hat{Y}_{i1}, \ldots, \hat{Y}_{it}$ denote the estimates based on independent samples of segments in the i-th sampled psu, then conditionally given the psu, $E(\hat{Y}_{ij}|i) = \hat{Y}_i$ and $Cov(\hat{Y}_{ij}, \hat{Y}_{ik}|i) = 0$ but the unconditional covariance $Cov(\hat{Y}_{ij},\hat{Y}_{ik}) = Cov(\hat{Y}_i,\hat{Y}_i) = V(\hat{Y}_i) \neq 0$. Hence $\hat{Y}_{i1}, \ldots, \hat{Y}_{it}$ are not independent.

4. P.B. Imrey, M.E. Francis and E. Sobel: The authors provide the covariance matrice of cell estimates for two-stage sampling (srs at both stages), and express it as $V = [1+(M-1)\underline{R}]\underline{V}_{srs}$ where $\underline{R}$ is the "intracluster correlation matrix",

a natural extension of the well-known intra-cluster correlation coefficient ρ. Summary statistics based on the eigenvalues of $\underline{RV}_{srs}$ are proposed to measure multivariate design effect of cluster sampling, but the natural matrix for this purpose is $\underline{R}$ rather than $\underline{RV}_{srs}$ since $\underline{R}$ reduces to the usual measure ρ, proposed by Kish in the univariate case. Similar results are also given in Rao and Scott (1980) for the commonly used two-stage design with pps sampling (with replacement) at the first stage. Rao and Scott (1980) called the eigenvalues ρ_i's of $\underline{R}$ "generalized measures of homogeneity" analogous to the measure of homogeneity ρ.

5. J.M. Lepkowski and J.R. Landis: Design effects for cell differences would be useful in arriving at an estimate of the covariance matrix of cell estimates in a contingency table. The attenuation model proposed by Lepkowski and Landis looks promising since it uses a "portable" attenuation factor α as a measure of cross-homogeneity. One can obtain a corrected chi-square statistic, $X^2/\bar{\delta}$, as mentioned earlier, using the estimated covariance matrix obtained from design effects for cell differences under the attenuation model. It would be interesting to study the distortion in the achieved significance level of this corrected statistic from actual survey data.

The authors remark that Fellegi's (1980) correction to X^2 is more conservative than the one proposed by Rao and Scott (1980). This requires further clarification. Felligi proposed $\bar{d} = (d_1+\ldots+d_r)/r$ for a contingency table with r cells, where d_i is the design effect for i-th cell. Rao and Scott suggested $\bar{\delta}$ which reduces to $[d_1(1-p_1)+ \ldots +d_r(1-p_r)]/(r-1)$ in the case of testing a simple hypothesis H: $p_i = p_{0i}$. Lepkowski and Landis compared $d^* = (d_1+ \ldots + d_{r-1})/(r-1)$ with $\delta^* = [d_1(1-p_1)+ \ldots + d_{r-1}(1-p_{r-1})]/(r-1)$ and concluded that $d^* \geq \delta^*$ since $1-p_i \leq 1$. On the other hand, $\bar{d} \geq \bar{\delta}$ for a simple hypothesis if and only if $Cov(p_j,\delta_j) \geq 0$.

6. T.J. Tomberlin: The author proposes random effects logistic models for analysing count data from multistage clustered samples. These models are promising and we need further work on estimating variance components associated with random effects. However, one should be careful in using these models with polytomous responses since such a model would lead to constant design effects for all individual cells and all cell differences and hence restrictive (Rao and Scott, 1980). Even in the case of dichotomous response, it is often difficult to formulate realistic models appropriate for multistage clustered designs. Tomberlin argues that if one of the major purposes of a survey is to provide data for complex statistical analyses such as tests in a multi-way contingency table then the sample design should be so chosen to facilitate simpler analyses. He further suggests the use of designs which can be treated as "ignorable", but it is not clear to me how one gets such a design which will be consistent with operational and cost considerations of a large-scale survey.

REFERENCES

[1] Altham, P.A.E. (1976). "Discrete Variable analysis for individuals grouped into families", Biometrika, 63, 263-269.

[2] Brier, S.S. (1978). "Discrete data models with random effects", Technical Report, University of Minnesota, School of Statistics.

[3] Cohen, J.E. (1976). "The distribution of the chi-squared statistic under clustered sampling from contingency tables", J. Amer. Statist. Assoc., 71, 665-670.

[4] Fellegi, I.P. (1980). "Approximate tests of independence and goodness of fit based on stratified multistage samples", J. Amer. Statist. Assoc., 75, 261-268.

[5] Rao, J.N.K. and Scott, A.J. (1980). "The analysis of categorical data from complex sample surveys: chi-square tests for goodness-of-fit and independence in two-way tables", J. Amer. Statist. Assoc. (revised version submitted in June, 1980).