DISCUSSION

D. R. Bellhouse, University of Western Ontario

The three papers are very different in terms of presentation and content. My first comment is an attempt to unify them and to put the papers into perspective along with other work in the field. The reason for these papers and their associated packaged programs is in that the major packaged programs such as SPSS, BMDP, and SAS, the statistical analyses depend on the assumption, usually tacit, of simple random sampling. The statistical analyses provided may include: regression analysis, including the analysis of variance; contingency tables, including log-linear models; and multivariate methods such as discriminant and factor analysis. When one of these major packaged programs is used to analyze a survey with a complex sampling design, problems may arise in the analysis and interpretation of the data. Misleading results and conclusions may be obtained. The authors of the current papers and a number of other researchers have responded to the problems created by complex sampling designs. In the area of regression analysis, a large amount of work has been done. This is evidenced by the fact that two packaged programs in regression analysis have been presented. Another program, SUPERCARP, is available. Also, much of Professor Shah's attention in his paper is directed towards working on refinements to the program - development of techniques when there are large numbers of regressors and an inadequate number of degrees of freedom, rather than the initial developments of statistical theory. Fewer results have been obtained in the area of contingency tables but work is progressing rapidly. The one program presented is a computerization of Dr. Fay's theory obtained within the last couple of years. Very little work has been done in the area of multivariate methods. However, new results from the sampling group in Southampton in the United Kingdom may soon be forthcoming.

At the beginning of section 4 of Dr. Fay's paper, there appears the following quote, "Work of Rao and Scott (1981) and others on applying generalized design effects is omitted here, since their methods are not yet linked to specific software." My only point here is that this illustrates an earlier remark that work is progressing very rapidly in the area of contingency tables for complex surveys. Hidiraglou and Rao (1981) have used the generalized design effects methodology. The computations were performed by extending the computer program MINICARP. I believe Rao and Scott are currently extending their theory to multiway tables and Hidiraglou is developing the appropriate software.

In the two papers presented by Dr. Lepkowski and Professor Shah, one parameter of interest to estimate is $\underline{B} = (X^TX)^{-1}X^T\underline{y}$. One logical basis for interest in this parameter is to consider the linear model

$$\underline{y} = X\underline{\beta} + \underline{e}$$

$$E(\underline{e}) = \underline{0},\ E(\underline{e}\underline{e}^T) = \sigma^2 I$$

on the whole set of finite population units. Then $\underline{B}$ is the least squares estimate of $\underline{\beta}$. Professor Shah does this explicitly. Related to this, Dr. Lepkowski, in section 2 of his paper, makes a separation between design based and model based statistical inference. My own thoughts on this issue are that there is a subtle though not formal relationship between the model and the design. For example, consider a complex design with stratification and clustering. If the reasons for stratification are due to differences between strata, then a model should be entertained which allows for differing slopes and differing variances between strata. Likewise, if the cluster is a family or a grouping of closely related units, then the modeller should consider correlated errors within clusters as part of the model. Once a model is formulated on the population units, then the finite populations parameters can be obtained as estimates of the superpopulation parameters as in the previously given regression model.

Replication and jackknife methods for variance estimation and the calculation of test statistics are techniques that are employed in all the papers. Application of these methods is limited in pps designs to sampling with replacement. Professor Shah gets around this problem by providing other variance estimates for a stratified two-stage design. In examples given by Dr. Fay in his paper and by Dr. Lepkowski in his presentation, the strict requirement of with replacement sampling did not hold. The question then arises: how robust are these methods to deviations from with replacement sampling?

Finally, as many discussants do, I have an advertisement for my own work [Bellhouse (1980, 1982)]. I have developed a computer program, still in its experimental stages, which estimates the sampling variances of estimates of means or totals for any stratified multistage cluster sample. The computations are based on the "textbook formulae" for stratified sampling, two-stage sampling, and so on rather than replication or jackknife techniques. The key to the program is recognizing the equivalence of a tree structure to multistage sampling designs. The program is run by making the traversal of a tree equivalent to successive variance calculations in a multistage design. The program includes options for various pps without replacement designs. In terms of computer time, the program is not burdensome. For stratified two-stage cluster sampling with pps sampling in the primaries, the program required 17 CPU seconds to obtain a variance-covariance matrix for 5 variables on 600 cases. The calculations were done on a minicomputer; the time could be reduced by more than a factor of 2 using a newer and larger computer.

REFERENCES

Bellhouse, D. R. (1980). Computation of variance-covariance estimates for general multi-stage designs, in COMPSTAT 1980: Proceedings in Computational Statistics, pp. 57-63 ed. M. M. Barritt and D. Wishart, Physica-Verlag.

Bellhouse, D. R. (1982). Computing methods for variance estimation in complex surveys. Paper presented at the Conference on Data Analysis and Complex Surveys, Jerusalem.

Hidiraglou, M. A. and Rao, J. N. K. (1981). Chi-square tests for the analysis of categorical data from the Canada Health Survey. Presented at the International Statistical Institute Meetings, Buenos Aires, Argentina.