THE SURVEY OF INCOME AND PROGRAM PARTICIPATION

MULTIVARIATE ANALYSIS BY USERS OF SIPP MICRO-DATA FILES

No. 92

R. P. Chakrabarty
Bureau of the Census

U. S. Department of Commerce BUREAU OF THE CENSUS

Acknowledgments

This paper was prepared by Rameswar P. Chakrabarty, Statistical Research Division, Bureau of the Census. The author would like to thank Lawrence Ernst, Rajendra Singh and Daniel Kasprzyk for their helpful comments on earlier drafts of this paper. Clerical assistance was provided by Pat Ellis. The views expressed are attributable to the author and do not necessarily reflect those of the Census Bureau.

Suggested Citation

Chakrabarty, Rameswar P. "Multivariate Analysis by Users of SIPP Micro-Data Files", SIPP Working Paper Series No. 8915. Washington, D.C.; U.S. Bureau of the Census, 1989.

TABLE OF CONTENTS

MULTIVARIATE ANALYSIS BY USERS OF SIPP MICRO-DATA FILES

1. INTRODUCTION

The survey of Income and Program Participation (SIPP) will undoubtedly become a major source of data on many socioeconomic aspects of the nation's households, families, and individuals. Therefore, it is anticipated that many researchers in areas of sociology, economics and public policy will be interested in analytical studies using SIPP data to enhance our understanding of complex socioeconomic phenomena. These analytical studies will usually be done by various multivariate analyses and modeling such as multivariate regression analysis, factor analysis, logistic regression analysis, discriminant analysis, and survival analysis and hazards modeling. Standard statistical packages like BMDP, SAS and SPSS have programs to carry out such analyses. These packages use statistical methods strictly applicable to samples from infinite populations and thus are not generally suitable for analysis of data from complex survey samples. These packages, however, have often been used by researchers for analysis of survey data mainly because appropriate statistical packages are not available. This practice of using a standard statistical package for analysis of survey data raises the following basic questions. What effects can the sample design have on methods of multivariate analysis? How should such effects be taken into account? In recent years, many researchers have attempted to answer these questions by demonstrating the adverse effects of survey designs on standard statistical methods and developing procedures that take the survey design into account. General methods of variance estimation for complex surveys and procedures that may be appropriate for variance computation by users of SIPP micro-data files will be discussed in another report. This report deals with multivariate analysis of data from complex surveys. At this time, methods of multivariate analysis that take account of the survey design, and computer software to implement such methods are at an early stage of development. Significant progress seems to have been made only for the analysis of categorical data from sample surveys. A review of research done for multivariate analysis and modeling and analysis of categorical data from complex surveys follows. We also indicate appropriate methods that users of SIPP micro-data files can use for data analysis and modeling.

2. GENERAL MULTIVARIATE ANALYSES AND MODELING:

In regression analysis, Nathan and Holt (1980) showed how sample selection dependent on an auxiliary variable can lead to inconsistent estimators. Bebbington and Smith (1977) demonstrated by simulation how disproportionate.stratification can lead to inconsistency and how cluster sampling can inflate variances in principal component analysis. Tortora (1980) applied principal component analysis to the selection of variables in a multiple regression model and, using dummy variables for stratum effects, found that disproportionate allocation affected the selection of variables. Nathan and Holt (1930) and Holt, Smith and Winter (1980) discussed the use of maximum likelihood and probability - weighted procedures in regression analysis to adjust for the effect of selection dependent on an auxiliary variable. Skinner, Holmes and Smith (1986) show that in principal component analysis estimators based on the assumption of simple random sampling can be severely biased for non-self-weighting sample designs. Differential nonresponse could have a similar effect. They show that, under a multivariate normal model, maximum likelihood estimation acts to correct for the conditional model bias of the simple random sample estimators. Probability weighted estimation, although asymptotically design unbiased, displays a conditional model bias.

Shah, Holt and Folsom (1977) evaluated inferences about regression models using "SURREGR" program.[1] Data from the National Longitudinal Survey of High School Class of 1972 and the Health Examination Survey were used in this study. The results from the ordinary least squares (OLS) assuming simple random samples were poor, in general, and suggested that the probability of rejecting the null hypothesis even when it is true, is extremely high. The Taylor series linearization (TSL), though not perfect, produces "good" conservative inferences for regression coefficients (i.e. the probability of rejecting the null hypothesis when it is true is smaller than the nominal value when the number of strata is large **($\geq$24)**). In fact, the transformed Hotelling's $\mathbf{T^2}$ type statistic, using the TSL variance-covaiance matrix, provides fairly robust multivariate inference about regression coefficients. In view of these empirical results, users may want to try to use the "SURREGR" program for regression analysis until a similar program geared to the SIPP design is available. The "SUPER CARP" program [2] also uses the TSL method for regression analysis developed by Fuller (1975) but no evaluation of its robustness is at present available.

No theoretical or empirical evaluation of the effects of survey design on .other multivariate methods like factor analysis, discriminant analysis, canonical correlation analysis, survival analysis and hazards modeling, etc is at present available in the literature. All these multivariate analyses, however, involve the estimation of the covariance matrix like the multivariate regression analysis. Therefore, use of a standard statistical package for such analyses of SIPR or any other complex survey data is likely to lead to misleading inferences. Thus, appropriate software packages for these multivariate analyses of survey data needs to be developed.

In the next section, we discuss the analysis of categorical data from complex surveys which, unlike the general multivariate methods, has been researched extensively.

---

[1]"SURREGR" program for regression analysis is a part of the "SUDAAN" (survey data analysis) system developed by B.V. Shah at the Research Triangle Institute, Research Triangle Park, North Carolina 27709.

[2]"SUPER CARP" (Cluster Analysis and Regression Program) is developed by W. Fuller, M. Hidiroglou, and R. Hickman at the Department of Statistics, Iowa State University, Ames, Iowa 50010.

3. THE ANALYSIS OF CATEGORICAL DATA

Methods for analysis of categorical data have been extensively developed under the assumption of multinorminal or product multinominal sampling. The Pearson chi-squared test statistic for testing goodness of fit and independence in two-way contingency tables are well known. The log-likelihood ratio and the Wald statistic which are asymptotically equivalent to the Pearson statistic, are also used. Recent extension of these methods to multidimensional contingency tables, using log-linear or linear models, have attracted considerable attention because of their close similarity to analysis of variance in systematically providing test statistics of various hypotheses. The theory associated with the Pearson and log-likelihood ratio statistics is given in Bishop, Fienberg and Holland (1975). Grizzle, Starmer, and Koch (1969) provide the results connected with the Wald statistic. Researchers in subject matter areas (e.g., social and health sciences) have long been using these multinomial - based methods to analyze survey data even though these methods may lead to misleading results. Therefore, in recent years considerable research has been done to study the effects of survey designs on test procedures and to develop new test procedures taking the survey design into account. We summarize below major developments in this area.

The Wald statistic, unlike Pearson and log-likelihood ratio statistics. generalizes readily to a complex sampling design, provided a consistent estimator of the covariance matrix of cell estimates is available. An excellent illustration of the use of the generalized Wald statistic is given by Koch, Freeman, and Freeman (1975). This approach provides a valid test statistic for survey data but has some drawbacks in application.
Typical estimators of covariance of the observed cross-classification for complex designs generally yield far less precision than the multinominal analogues, and this reduced precision has a serious effect on the matrix inversion required in the computation of Wald statistic. This instability in the estimated inverse in turn often produces erratic results for tests of hypotheses for complex samples (Fay, 1985). Monte Carlo results (Thomas and Rao, 1984) also indicate that the Wald statistic, although asymptotically valid, often fails to control the type I error rate satisfactorily.

Alternative test statistics have been proposed by Nathan (1975), Shuster and Downing (1976), and Fellegi (1980). These tests require estimation of the fall covariance matrix of the cell estimates, followed by an inversion producing the same potential difficulties as those encountered with the Wald statistic. A comparison of many of these alternatives, including the behavior of a statistic (t') of this form for the test of independence in a two-way classification is given by Fellegi (1980). Cohen (1976) and Altham (1976) proposed a simple model for clustering and showed that the generalized Wald statistic for goodness of fit is a multiple of chi-square (X^2) when the model holds. Brier (1980) considered a similar model, but studied general hypotheses on cell probabilities, and proved that the multiple of the corresponding Pearson statistic is asymptotically distributed as a X 2 random variable when the model holds. Rao and Scott (1981) provide some extensions of these model-based approaches. The disadvantage of these methods, however, is the need for additional assumptions on the covariance of the estimates for the procedures to work acceptably well. Fellegi (1980) proposed a statistic (t") that attempts to correct the chi-squared

statistic under more general assumptions. Rao and Scott (1981) reviewed these methods and examined the properties of the chi- squared test for two-way tables under several common complex sample designs. These two authors, along with their colleagues (Holt et al., 1980; Scott & Rao, 1981; Hidiroglou and Rao, 1981, 1983; Rao and Scott, 1979, 1981, 1984, 1987), studied the distribution of the chi-squared tests under more general log-linear models and developed modified test statistics that can be computed for some models if design effects (deffs) for ultimate cells are available. In a simulation study using household survey data they show that the modified test gives good results in all cases in the study but the ordinary chi-squared test with a desired level of 5 percent, the estimated (asymptotic) significance level can be as high as 41 percent, which is obviously unacceptable. Thus, one can draw very misleading inferences if one naively uses ordinary chi-squared tests for survey data. Hidiroglou and Rao (1987a, 1987b) provide a user's guide to modified chi-squared tests that take account of the survey design. Test procedures are illustrated with data from the Canada Health Survey (1978-1979). Experimental software for analysis of data was developed by modifying the MIMI CARP program (Hidiroglou et. al., 1980). These programs are available but they have not been fully documented for general use.
Logistic regression models are extensively used for the analysis of variation in the estimated proportions associated with a binary response variable; see, for example, Cox (1970). Roberts, Rao and Kumar (1987) showed that the standard logistic regression analysis of survey data could lead to misleading inferences. They adjusted the standard chi-square or likelihood ratio test statistics for logistic regression analysis to take account of the survey design. These adjustments were based on certain generalized design effects. Logistic regression diagnostics to detect any outlying cell proportions in the table and influential points in the factor space were also developed, taking account of the survey design. The authors applied their methods to analyze some data from the October 1980 Canadian Labor Force Survey.

Fay (1985) describes a jackknifed chi-squared statistic X 2/j that is closely related to X 2/6 (where 6 is the generalized design effect) proposed by Rao and Scott (1981). Fay also provides a test statistic for testing the difference of two chi-squared tests under nested models. The jackknifed tests have been incorporated in the computer program CPLX (Contingency Table Analysis for Complex Sample Designs) developed by Fay (1982). The documentation for this program by Fay (1983) includes comments on strategies to formulate replication approaches for common complex designs, and provides two examples of applications of these test statistics. The properties and the range of application of the jackknifed tests-from Fay (1985) are summarized below:

(1) The Monte Carlo results confirmed earlier findings that when data arise from a multinomial distribution, the Pearson test is superior to the likelihood - ratio chi-square (Larntz, 1978) for tests of goodness of fit, and the likelihood - ratio test is superior to the difference of Pearson chi-squares (Haberman, 1977) for tests of parameters in most situations, such as tests of parameters in a logistic model. Monte Carlo results favored the same choices for the jackknifed tests.

(2) Almost any form of complex design has a potentially severe effect on the chi-squared tests. Although the jackknifed tests require somewhat larger sample sizes for the asymptotic theory to approximate actual performance, they otherwise offer

protection against the effects of a complex design

.

(3) The Monte Carlo results indicated that the use of 50 replicates made only slight improvement over 20 replicates for purposes of constructing the simple jackknife. The number of replicates for the stratified jackknife would, however, be larger than the number of strata. (4) For tests of the contribution of specific sets of parameters to a model (testing the relative fit of two nested models), the jackknifed likelihood - ratio test essentially requires only that the corresponding marginal tables be reasonably filled out. Thus, there is a wide potential range of application to logistic and other causal modeling even when the table itself is sparse. Comparison of nested models is often more essential to the development of models for large cross-classifications than the testing of goodness of fit.

(5) Note that the asymptotic theory for the jackknifed tests can be applied to a wider range of problems, such as latent structure models described by Goodman (1974) and Haberman (1974a), log-linear models with structural zeros, and the non-log-linear models proposed by Haberman (1974b), Goodman (1979, 1981), and Clogg (1982) for tables that include variables with ordered categories.

(6) Replication methods can be adopted to a number of complex sample designs, including multi-stage sampling with complex estimation procedures, and finite population corrections. Thus, this procedure can certainly be applied to SIPP data.

(7) One desirable feature of X 2/6 is that it can be computed in some cases without access to the original survey data. When X 2/j can be computed, however, it offers greater protection against rejection of the null hypothesis at a rate appreciably higher than the nominal level.

(8) Like X 2/j another test statistic, X 2/s proposed by Rao and Scott using the Satterwaite approximation (1946), also requires access to the original data unless an estimated covariance matrix for the cells of the complete cross-classification is available. Asymptotically, X 2/s probably gives actual rejection rates over a range both narrower and closer to the nominal level than X 2/j. However, under some practical conditions, the actual (as opposed to the asymptotic) performance of X 2/s is more conservative than X 2/j (Fay, 1983b). Thus, neither test is uniformly superior to the other, but both appear to be better than other alternatives.

(9) The formulation of X 2/s also extend to other models beyond the log-linear model, as does X 2/j. Researchers, however, may find X 2/j easier to implement than X 2/s , since x 2/j only requires algorithms to compute estimates under the model.

These impressive advances in developing valid test statistics for survey data lead us to the following recommendations for analysis of multiway contingency tables:

(1) If the original data are not available, use modified chi-squared tests (Rao and Scott, 1981) to analyze multiway contingency tables, provided some information about the design effects (deffs) for marginal totals (the sort of information given in Kish, Groves, and Krotki (1976), for example) are available in published reports.

(2) If the original data are available, use the jackknifed chi-squared tests (Fay, 1985) given in the CPLX program. Both the program and documentation are in public domain.

(3) SIPP micro-data files users should consider the use of the CPLX program to analyze multiway contingency tables.

4. COMPARISON OF TWO CONTINGENCY TABLES

Sometimes, an analyst may be interested in knowing whether the classification variables achieve the association in a similar way in two contingency tables from two populations. Populations may be different regions or different domains. The traditional approach to compare two populations cross-classified by the same variables when the classifications exhibit some degree of association has been to measure the association in each table by means of a summarizing quantity and to compare the values so obtained. The researcher usually selects a coefficient of association available in a statistical package, and typical selections are the contingency coefficient, based on Pearson's chi-squared statistic, or possibly one of the coefficients with operational meaning discussed by Goodman & Kruskal (1954, 1959).

But most measures of association are not designed to compare the inherent association in two cross-classifications, but rather to measure the association in a single table. Thus, when interest lies in knowing whether both populations achieve the association in a similar way, the comparison of the values of a coefficient for each table may be meaningless, as the values may be achieved in many ways.

Yule (1912), Edwards (1963), Plackett (1965) and Mosteller (1968), among others, have argued that for comparative purposes measures of association should be functions of the cross-product ratios of the tables alone, a requirement which is not satisfied by the chi-squared based measures, nor by most of Goodman & Kruskal's measures with operational meaning. Altham (1970) has shown that the comparison of associations should be based on vectors of differences between contrasts of log frequencies and that it is sensible to look at sums of squares of contrasts as the test statistic. Molina and Smith (1986) use Altham's measure and provide test statistics for comparison of associations which are asymptotically distributed as central chi-squared variables under complex sampling situations. The technique requires the computation of certain dispersion matrices appropriate for complex sampling designs. These computations require micro-data. A corrected test statistic that can be computed if cell variances or design effects are known is also given by Molina and Smith (1986) following the methods of Rao and Scott (1981).

5. RECOMMENDATIONS FOR MULTIVARIATE ANALYSIS

I. General Multivariate Analyses and Modeling

(a) Users should be aware that the use of a standard statistical package, such as BMDP, SAS or SPSS, for multivariate analysis of SIPP or any other complex survey data could lead to misleading inferences.

(b) For multivariate regression analysis, users may 'SURREGR" program developed by the Research Triangle Institute until a similar program geared to the SIPP design is available.

(c) For other multivariate analyses of software package is available development of appropriate software package is available at present. Thus, there is a need for development of appropriate software packages for various multivariate methods for analysis of SIPP data.

II. Analysis of Categorical Data

(a) Users should consider the use of the CPLX program developed by Fay at the Bureau of the Census, to analyze multiway contingency tables.

(b) For comparison of two contigency tables, the use of the test statistic developed by Molina and Smith (1986) is preferable to simple comparison of the value of a coefficient of association.

REFERENCES

Altham, P.A.E. (1976), Discrete Variable Analysis for Individuals Grouped into Families, Biometrika, 63, 263-269.

Bebbington, A.C. and Smith, T.M.F. (1977), The Effect of Survey Design on Multivariate Analysis. The Analysis of Survey Data (Vol. 2), eds C.A. O'Muircheartaigh and C.D. Payne, John Wiley and Sons, New York.

Bishop, Y.M.M., Fienberg, S.E., Holland, P.W. (1975), Discrete MultivariateAnalysis: Theory and Practice, Cambridge, Mass: MIT Press.

Brier, S.S. (1980), Analysis of Contingency Tables Under Cluster Sampling, Biometrika, 67, 591-596.

Clogg, Clifford C. (1982), Some Models for Analysis of Association in Multiway Cross-Classification Having Ordered Categories, Journal of the American Statistical Association, 77, 803-815.

Cohen, J.E. (1976), The Distribution of the Chi-Square Statistic Under Cluster Sampling From Contingency Tables, Journal of the American Statistical Association, 71, 665-670.

Cox, D.R. (1970), Analysis of Binary Data. Chapman and Hall, London.

Edwards, A.W.F. (1963), The Measure of Association in a 2x2 Tables, Journal of the Royal Statistical Society, Series A, 126, 109-114.

Fay, Robert E. (1982), Contingency Table Analysis for Complex Designs: CPLX, Proceedings of the Section on Survey Research Methods, American Statistical Association, pp. 44-53.

Fay, Robert E. (1983a), CPLX-Contingency Table Analysisfor Complex Sample Designs, Program Documentation, unpublished report, U.S. Bureau of the Census.

Fay, Robert E. (1983b), Replication Approaches to the Log-Linear Analysis of Data From Complex Samples, paper presented at the seminar "Recent Developments in the Analysis of Large Scale Data Sets," Statistical Office of the European Communities, Luxembourg, November.

Fellegi, I.P. (1980), Approximate Tests for Independence and Goodness of Fit Based on Stratified Samples, Journal of the American Statistical Association, 75, 261-268.

Fuller, W.A. (1975), Regression Analysis for Sample Surveys, Sankhya, Series C, 37, 117-132.

Goodman, L.A. (1970), The Multivariate Analysis of Quantitative Data: Interactions Among Multiple Classifications, Journal of the American Statistical Association, 65, 226-256.

Goodman, L.A. (1974), The Analysis of Systems of Quantitative Variables When Some of the Variables are Unobservable (Part I), A Modified Latent Structure Approach, American Journal of Sociology, 79, 1179-1254.

Goodman, L.A. (1979), Simple Models for the Analysis of Association in Cross-Classifications Having Ordered Categories, Journal of American Statistical Association, 74, 537-552.

Goodman, L.A. (1981), Association Models and Cannonical Correlation in the Analysis of Cross-Classifications Having Ordered Categories, Journal of the American Statistical Association, 76, 320-334.

Goodman, L.A. and Kruskal, W.H. (1954), Measures of Association for Cross-Classifications, Journal of the American Statistical Association, 49, 732-764.

Goodman, L.A. and Kruskal, W.H. (1959), Measures of Association for Cross-Classifications II: Further Discussion and References, Journal of the American Statistical Association, 54, 123-163.

Grizzle, J.E., Starmer, C.F., and Koch, G.G. (1969), Analysisof Categorical Data by Linear Models, Biometrics, 25, 489-504.

Haberman, Shelby J. (1974a), Log-Linear Models for Frequency Tables Derived by Indirect Observation: Maximum Likelihood Equations, Annals of Statistics, 2, 911-924.

Haberman, Shelby J. (1974b), Log-Linear Models for Frequency Tables with Ordered Classifications, Biometrics, 30, 589-600.

Hidiroglou, M.A., Fuller, W.A., and Hickman, P.D. (1980), MINI CARP: A Program for Estimating Simple Descriptive Statistics and their Variances from Multi-Stage Stratified Designs, Iowa State University, Ames, Iowa.

Hidiroglou, M.A., and Rao, J.N.K. (1981), Chi-Square Tests for the Analysis of Categorical Data from the Canada Health Survey, paper presented at the International Statistical Institute Meetings, Buenos Aires, December 5.

Hidiroglou, M.A. and Rao, J.N.K. (1983), Chi-Square Tests for the Analysis of Three-Way Contingency Tables from the Canada Health Survey, unpublished Technical Report, Statistics Canada.

Hidiroglou, M.A. and Rao, J.N.K. (1987a), Chi-Squared Tests with Categorical Data from Complex Surveys: Part I - Simple Goodness-of-Fit, Homogeneneity and Independence in a Two-Way Table with Applications to the Canada Health Survey (1978-1979). Journal of Official Statistics, 3, 117-132.

Hidiroglou, M.A. and Rao, J.N.K. (1987b), Chi-Squared Tests with Categorical Data from Complex Surveys: Part II - Independence in a Three-Way Table with Applications to the Canada Health Survey (1978-1979). Journal of Official Statistics, 3, 133-140.

Holt, D., Scott, A.J., and Ewings, P.O. (1980), Chi-Squared Tests with Survey Data, Journal of the Royal Statistical Society, Ser. A, 143, 302-320.

Holt, D., Smith, T.M.F. and Winter, P.D. (1980), Regression Analysis of Data from Complex Surveys. Journal of the Royal Statistical Society, Series A., 143, 474-487.

Kish, L., Groves, R.M., and Krotki, K.P. (1976), Sampling Errors for Fertility Survey, Occasional Paper No. 17, London: World Fertility Survey.

Koch, G.G.., Freeman, D.H., Jr., and Freeman, J.L. (1975) Strategies in the Multivariate Analysis of Data from Complex Surveys, International Statistical Review, 43, 59-78.

Larntz, Kinley (1978), Small-Sample Comparisons of Exact Levels for Chi-Squared Goodness-of-Fit Statistics, Journal of the American Statistical Association, 73, 253-263.

Molina, E.A.C. and Smith, T.M.F. (1986), The Effect of Sample Design on the Comparison of Association, Biometrika, 73, 23-33.

Mosteller, F. (1968), Association and Estimation in Contingency Tables, Journal of the American Statistical Association, 63, 1-28.

Nathan, Gad (1975), Tests of Independence in Contingency Tables from Stratified Proportional Samples, Sankhya, Series C, 37, 77-87.

Nathan, G. and Holt, D. (1980), The Effect of Survey Design on Regression Analysis. Journal of the Royal Statistical Society, Series B, 42, 377-386.

Plackett, R.L. (1965), A Class of Bivariate Distributions, Journal, of the American Statistical Association, 60, 516-522.

Rao, J.N.K. and Scott, A.J. (1979), Chi-Squared Tests for Analysis of Categorical Data from Complex Surveys. Proceedings of the American Statistical Association, Survey Research Methods Section. 58-66.

Rao, J.N.K. and Scott, A.J. (1981), The Analysis of Categorical Data from Complex Sample Surveys: Chi-Squared Tests for Goodness of Fit and Independence in Two-Way Tables, Journal of the American Statistical Association, 76, 221-230.

Rao, J.N.K. and Scott, A.J. (1984), On Chi-Squared Tests for Multiway Contingency Tables with Cell Proportions Estimated from Survey Data, Annals of Statistics, 12, 46-60.

Rao, J.N.K. and Scott, A.J. (1987), On the Simple Adjustments of Chi-Squared Tests with Sample Survey Data. Annals of Statistics, 15.

Roberts, G., Rao, J.N.K. and Kumar, S. (1987), Logistic Regression Analysis of Sample Survey Data. Biometrika, 74, 1-12.

Satterwaite, F.E. (1946), An Approximate Distribution of Estimates of Variance Components, Biometrics, *2*, *110-114*.

Scott, A.J., and Rao, J.N.K. (1981), Chi-Squared Tests for Contingency Tables with Pronortions Estimated from Survey Data, in Current Topics in Survey Sampling, eds. D. Krewski, R. Platek, and J.N.K. Rao, New York: Academic Press.

Shah, B.V., Holt, M.14. and Folsom, R.E. (1977), Inference about Regression Models from Sample Survey Data. Bulletin of International Statistical Institute, 47, 43-57.

Shuster, J.J., and Downing, D.J. (1976), Two-way, Contingency Tables for Complex Sampling Schemes, Biometrika, 63, 271-276.

Skinner, C.J., Holmes, D.J. and Smith, T.M.F. (1986), The Effect of Sample Design on Principal Component Analysis. Journal of the American Statistical Association, 81, 789-798.

Thomas, D.R. and Rao, J.N.K. (1984), A Monte Carlo Study of Exact Levels of Goodness-of-Fit Statistics Under Cluster Sampling. Proceedings of the American Statistics Association, Survey Research Methods Section, 207-211.

Tortora, R.D. (1980), The Effect of Disproportionate Stratified Design on Principal Component Analysis Used for Variable Elimination. Proceedings of the American Statistical Association, Survey Research Methods Section, 746-750.

Yule, G.L. (1912), On the Methods of Measuring Associations Between Two Attributes, Journal of the Royal Statistical Society, 75, 579-642.