Statistics Canada Statistique Canada

SURVEY METHODOLOGY

A JOURNAL
OF
STATISTICS CANADA

VOLUME 15, NUMBER 2
DECEMBER 1989

Canada

SURVEY METHODOLOGY

A JOURNAL OF STATISTICS CANADA

DECEMBER 1989

Published under the authority of
the Minister of Regional Industrial Expansion

March 1990

Price: Canada, $30.00 a year
Other Countries, $35.00 a year

Payment to be made in Canadian funds or equivalent

Catalogue 12-001, Vol. 15, No. 2

ISSN 0714-0045

Ottawa

SURVEY METHODOLOGY

A Journal of Statistics Canada

The Survey Methodology Journal is abstracted in The Survey Statistician and Statistical Theory and Methods Abstracts and is referenced in the Current Index to Statistics, and Journal Contents in Qualitative Methods.

EDITORIAL POLICY

The Survey Methodology Journal publishes articles dealing with various aspects of statistical development relevant to a statistical agency, such as design issues in the context of practical constraints, use of different data sources and collection techniques, total survey error, survey evaluation, research in survey methodology, time series analysis, seasonal adjustment, demographic studies, data integration, estimation and data analysis methods, and general survey systems development. The emphasis is placed on the development and evaluation of specific methodologies as applied to data collection or the data themselves. All papers will be refereed. However, the authors retain full responsibility for the contents of their papers and opinions expressed are not necessarily those of the Editorial Board or of Statistics Canada.

Submission of Manuscripts

The Survey Methodology Journal is published twice a year. Authors are invited to submit their manuscripts in either of the two official languages, English or French to the Editor, Dr. M.P. Singh, Social Survey Methods Division, Statistics Canada, 4th Floor, Jean Talon Building, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6. Two nonreturnable copies of each manuscript prepared following the guidelines given in the Journal are requested.

Subscription Rates

The price of the Survey Methodology Journal (Catalogue No. 12-001) is $30.00 per year in Canada, $35.00 per year for other countries (payment to be made in Canadian funds or equivalent). Subscription order should be sent to: Publication Sales, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6. A reduced price of US $16.00 ($20.00 Can.) is available to members of the American Statistical Association, the International Association of Survey Statisticians, and the Statistical Society of Canada. Please subscribe through your organization.

SURVEY METHODOLOGY

A Journal of Statistics Canada
Volume 15, Number 2, December 1989

CONTENTS

Special Section – Data Analysis

Survey Methodology, December 1989
Vol. 15, No. 2, p. 159
Statistics Canada

In This Issue

The risks involved in using standard statistical methods for the analysis of data from surveys with complex designs are becoming well-known. The special topic section in this issue contains three papers which provide guidance for the analysis of categorical data from such surveys. Tim Holt's efforts were instrumental in putting this section together.

The paper by Rao, Kumar and Roberts, which is the first discussion paper published in Survey Methodology, reviews developments in the analysis of cross-classified categorical data, extends them, and applies them to data from two large, complex surveys. The authors also briefly discuss computational issues. Comments by Fay, Skinner and Molina and a reply by Rao, *et al.* follow the paper.

Thomas describes a Monte Carlo study used to investigate several methods of obtaining simultaneous confidence intervals for proportions under a two-stage clustered design. He shows that some methods behave poorly, with actual coverage rates quite different from the nominal ones. Thomas concludes with guidelines on the choice of methods to use in practice.

The final paper in the section on data analysis for complex surveys, by Morel, deals with logistic regression. Using the results of a Monte Carlo study, he shows that for small samples, a modified Taylorization method for estimating a covariance matrix results in smaller biases than the usual delta method.

The bibliography by Nathan on randomized response which appeared in the previous issue of Survey Methodology attests to the large amount of research which has been devoted to the subject. In this issue, Franklin develops another approach to the randomized response model for sampling from dichotomous populations. The model is general in that it permits the use of randomization from a continuous distribution and multiple trials per respondent. Special attention is given to the case of randomization using the normal distribution function.

MacGibbon and Tomberlin examine the problem of small area estimation with complex survey designs. Their empirical Bayes estimator is a compromise between the highly variable but unbiased classical estimator and the more stable but potentially highly biased synthetic estimator.

A method of updating a PPSWOR sample which attempts to retain the same sample of primary sampling units is presented by Sunter. The method differs from earlier ones proposed by Kish and Scott (1971) and Fellegi (1963) in that it is valid for any sample size and does not require enumeration of all possible samples. The method is of particular importance for multistage survey samples which must be updated, but for which the cost of introducing new PSUs may be high.

Revenue Canada tax files and Family Allowance files are used in Canada to provide population estimates for provinces in non-census years. Verma and Raby examine the consistency of the estimates derived from these two sources. A comparison with the 1986 Census counts is also made.

Swanson presents a method of obtaining confidence intervals for post-censal population estimates. He shows that a Wilcoxon test can be used to determine if a change in model, due to post-censal structural changes, is required. Using empirical data, Swanson shows that ignoring such a change leads to confidence intervals whose coverage is lower than expected.

The Editor

Survey Methodology, December 1989
Vol. 15, No. 2, pp. 161–186
Statistics Canada

Analysis of Sample Survey Data Involving Categorical Response Variables: Methods and Software

J.N.K. RAO, S. KUMAR, and G. ROBERTS[1]

ABSTRACT

During the past 10 years or so, rapid progress has been made in the development of statistical methods of analysing survey data that take account of the complexity of survey design. This progress has been particularly evident in the analysis of cross-classified count data. Developments in this area have included weighted least squares estimation of generalized linear models and associated Wald tests of goodness of fit and subhypotheses, corrections to standard chi-squared or likelihood ratio tests under loglinear models or logistic regression models involving a binary response variable, and jackknifed chisquared tests. This paper illustrates the use of various extensions of these methods on data from complex surveys. The method of Scott, Rao and Thomas (1989) for weighted regression involving singular covariance matrices is applied to data from the Canada Health Survey (1978-79). Methods for logistic regression models are extended to Box-Cox models involving power transformations of cell odds ratios, and their use is illustrated on data from the Canadian Labour Force Survey. Methods for testing equality of parameters in two logistic regression models, corresponding to two time points, are applied to data from the Canadian Labour Force Survey. Finally, a general class of polytomous response models is studied, and corrected chi-squared tests are applied to data from the Canada Health Survey (1978-79). Software to implement these methods using the SAS facilities on a main frame computer is briefly described.

KEY WORDS: Corrections to chi-squared tests; Logistic regression; Power transformations; Wald tests; Weighted least squares.

1. INTRODUCTION

Standard statistical methods, based on the assumption of independent identically distributed observations, are being used extensively by researchers in the social and health sciences, and in other subject matter areas. These methods have also been implemented in standard statistical packages, including SPSSX, BMDP, SAS and GLIM. In practice, however, much data are obtained from complex sample surveys involving clustering and stratification, so that the application of standard methods to these data without some adjustment for survey design can lead to erroneous inferences. In particular, standard errors of parameter estimates and associated confidence intervals can be seriously understated if the complexity of the sample design is ignored in the analysis of data. Moreover, the actual type I error rates of tests of hypotheses can be much bigger than the nominal levels. Standard exploratory data analyses, *e.g.*, residual analysis to detect model deviations, are also affected. Kish and Frankel (1974) and others drew attention to some of these problems with standard methods, and emphasized the need for new methods that take proper account of the complexity of survey design. During the past 10 years or so, rapid progress has been made in the development of such methods, particularly for analysing cross-classified count data. This paper will focus on the analysis of

[1] J.N.K. Rao, Department of Mathematics and Statistics, Carleton University, Ottawa, Ontario; S. Kumar and G. Roberts, Social Surveys Methods Division, Statistics Canada, Ottawa, Ontario.

count data, but it should be noted that important results on other types of analyses have also been obtained: Regression analysis (Fuller 1975; Nathan and Holt 1980; Pfefferman and Nathan 1981; Scott and Holt 1982), principal component analysis (Skinner, Holmes and Smith 1986), factor analysis (Fuller 1986), logistic regression involving continuous covariates (Binder 1983).

Rao and Scott (1984) have made a systematic study of the impact of survey design on standard Pearson chi-squared or likelihood ratio tests for multiway tables of counts, under hierarchical log-linear models. They have also obtained simple first order corrections to standard tests which can be computed from published tables that include "design effects" for cell estimates and marginal totals, thus facilitating secondary analyses from published reports (see also Gross 1984; Bedrick 1983; Rao and Scott 1987). These first order corrections take account of the design in the sense that the actual type I error rates of tests based on the corrected statistics are closer to nominal levels, compared to the standard tests which could have greatly inflated type I error rates. More accurate second order corrections, based on the Satterthwaite approximation to a weighted sum of independent χ^2 variables, were also developed by Rao and Scott (1984), but these tests require the knowledge of a full estimated covariance matrix of cell estimates. Alternative methods that take account of the survey design include the Wald statistics based on weighted least squares (Koch, Freeman and Freeman 1975), and the jackknifed chi-squared tests (Fay 1985), all requiring either the full estimated covariance matrix or access to cluster-level data. Fay (1985) and Thomas and Rao (1987) have shown that the Wald statistic, although asymptotically correct, can become highly unstable as the number of cells in the multiway table increases and the number of sample clusters decreases, leading to unacceptably high type I error rates compared to the nominal level. On the other hand, Fay's jackknife tests and the Rao-Scott corrections have performed well under quite general conditions. In some cases, the instability in the Wald statistic may be remedied by collapsing the table according to eigenvectors associated with the nonnegligible eigenvalues of the estimated covariance matrix adjusted for singularities caused by linear constraints on the probabilities, as proposed by Singh (1985); see also Singh and Kumar (1986).

Roberts, Rao and Kumar (1987) assumed a logistic regression model for the cell (domain) proportions associated with a binary response variable, and obtained first order corrections to standard chi-squared and likelihood ratio tests of goodness-of-fit and nested hypotheses. Simple upper bounds to first order corrections, depending only on the design effects of cell response proportions, were also obtained to facilitate secondary analyses from published tables. Scott (1986) proposed an alternative method which uses standard tests on transformed data derived from the original data and the cell design effects. Roberts, Rao and Kumar (1987) also provided second order corrections to standard tests, but these require access to a full estimated covariance matrix of cell response proportions. Diagnostics for detecting outliers and influential points were developed as well, again taking the survey design into account.

The primary purpose of this paper is to present various extensions of the previous methods and illustrate their use on data from large-scale surveys, including the Canada Health Survey (1978-1979) and the Canadian Labour Force Survey. It is assumed, throughout the paper, that the user has access to a full estimated covariance matrix of cell estimates. In Section 2, weighted least squares (WLS) estimators of the parameters of generalized linear models having singular covariance matrices, caused by linear constraints on the probabilities (or proportions), are presented. Associated Wald tests of goodness-of-fit and of subhypotheses are also provided. A smoothed version of the WLS estimators, and associated Wald tests of subhypotheses are given as well. These methods should be used only when the number of cells in a table is small and/or the number of sample clusters in the survey design is relatively large.

The methods for logistic regression models are extended, in Section 3, to Box-Cox models involving power transformations of cell odds ratios. These models, which include the logistic regression model as a special case, could provide significantly better fits than the logistic regression models, as demonstrated by Guerrero and Johnson (1982) in the context of binomial proportions.

Methods for testing equality of parameters in two logit models, corresponding to two different time periods, are given in Section 4. If the hypothesis of equality is accepted, one could obtain "smoothed" estimates of cell proportions for the current period that are more efficient than the corresponding smoothed estimates based only on the current period data.

Section 5 gives an extension of the type of results obtained for logistic regression models to a general class of polytomous response models. The special case of McCullagh's (1980) ordered response model is studied in detail.

Finally, an account of the software for implementing the above methodology is given in Section 6.

2. WEIGHTED LEAST SQUARES ESTIMATORS AND WALD TESTS

The approach of Koch, Freeman and Freeman (1975) is designed to estimate the parameters of generalized linear models of the form $g^*(p) = X^*\beta^*$, using a sample estimate, $\hat{p}$, of the population cell probabilities denoted by a T-vector p, and a consistent estimate of $\operatorname{cov}(\hat{p}) = V_p$ (say). In this method, the asymptotic covariance matrix of the u-vector $g^*(p)$ is assumed to be nonsingular ($u < T$); however, many models, including the traditional loglinear model, are of the form $g(p) = X\beta$, where $g(p)$ is a T-vector with a singular asymptotic covariance matrix, and X is a $T \times r$ full rank matrix of known constants. It is possible to reduce the latter models to the nonsingular form $g^*(p) = X^*\beta^*$, as done by Grizzle and Williams (1972) for the loglinear model, but Scott, Rao and Thomas (1989) have developed the following unified approach for singular models, by appealing to the optimal theory for linear models having singular covariance matrices.

The cell probabilities p and $\hat{p}$ are subject to linear constraints of the form $K'p = \pi$ and $K'\hat{p} = \pi$, where K is a $T \times L$ full rank matrix of known constants and π is an L-vector of known constants π_i ($L < T$). As a result, the covariance matrix of $\hat{p}$ will be singular. For example, in the case of stratified sampling with complex sample designs within strata, we can write $K = I_L \otimes 1_m$, $\pi_i = n_i/n$ ($i = 1, \ldots, L$) and $p = (p_{11}\ldots p_{1m}; \ldots; p_{L1}\ldots p_{Lm})'$ with $p_{ij} = (n_i/n)\tilde{p}_{ij}$, where $\tilde{p}_{ij}$ is the j-th category probability within the i-th stratum ($\sum_j \tilde{p}_{ij} = 1$; $i = 1, \ldots, L; j = 1, \ldots, m$), n_i is the sample size from the i-th stratum, $\sum n_i = n$, 1_m is a m-vector of 1's, I_L is the identity matrix of order L and $\otimes$ denotes the Kronecker product.

Assume that $X\beta$ can be written as $X_0\beta_0 + X_1\beta_1$, where X_0 is a $T \times L$ matrix such that $K'H^{-1}X_0$ is nonsingular and where $H = (\partial g/\partial p)'$ is the $T \times T$ matrix of partial derivatives of $g(p)$. In particular, X_0 can be taken as K if the constraint matrix K is included in X, as frequently assumed. Since restrictions on p imply constraints on the parameters β, β_0 can be determined exactly from the constraints, for a given β_1.

Weighted least squares estimators

The model may be written as

$$\hat{g} = g(\hat{p}) = X\beta + \delta \tag{2.1}$$

where δ is the error vector with $P\lim \delta = 0$, and $\hat{g}$ has a singular asymptotic covariance matrix $V_g = HV_pH'$ which is consistently estimated as $\hat{V}_g = \hat{H}\hat{V}_p\hat{H}'$, assuming that $\hat{V}_p$ is a consistent estimator of V_p. Here $\hat{H} = H(\hat{p})$. Scott, Rao and Thomas (1989) derived an asymptotically best linear unbiased estimator (ABLUE) of β_1 as

$$\hat{\beta}_1 = (\tilde{X}_1'\hat{M}\tilde{X}_1)^{-1}\tilde{X}_1'\hat{M}\hat{g}, \tag{2.2}$$

where

$$\hat{M} = (\hat{V}_g + X_0X_0')^{-1} \tag{2.3}$$

is a nonsingular generalized inverse of $\hat{V}_g$, and

$$\tilde{X}_1 = [I - X_0X_0'\hat{M}]X_1. \tag{2.4}$$

A consistent estimator of the asymptotic covariance matrix of $\hat{\beta}_1$ is given by

$$\text{est cov}(\hat{\beta}_1) = (\tilde{X}_1'\hat{M}\tilde{X}_1)^{-1}. \tag{2.5}$$

Wald tests

Letting $\hat{\beta} = (X'\hat{M}X)^{-1}X'\hat{M}\hat{g} = (\hat{\beta}_0',\hat{\beta}_1')'$, a Wald test of goodness of fit of the model (2.1) is given by

$$W = (\hat{g} - X\hat{\beta})'\hat{M}(\hat{g} - X\hat{\beta}) \tag{2.6}$$

which is distributed asymptotically as a χ^2 variable with $T - r$ degrees of freedom (d.f.). The model is considered tenable at the α-level if $W > \chi^2_{T-r}(\alpha)$, the upper α-point of χ^2 with $T - r$ d.f..

Given the model (2.1), tests of linear hypotheses on the model parameters β_1 can also be obtained. A Wald test of the linear hypothesis $C_1\beta_1 = c_1$ is given by

$$W_1 = (C_1\hat{\beta}_1 - c_1)'[C_1\text{est cov}(\hat{\beta}_1)C_1']^{-1}(C_1\hat{\beta}_1 - c_1) \tag{2.7}$$

which is distributed asymptotically as a χ^2 variable with h d.f., where C_1 is a $h \times (r - L)$ full rank matrix of known constants $(h < r - L)$, and c_1 is a h-vector of known constants. The hypothesis is rejected at the α-level if $W_1 > \chi^2_h(\alpha)$, the upper α-point of χ^2 with h d.f. Note that β_0 should not be included in the linear hypothesis since it is fixed by the design constraints $K'p = K'g^{-1}(X\beta) = \pi$.

Smoothed version of ABLUE and associated Wald tests

We can also obtain a smoothed version of ABLUE of β_1, say β_1^*, using iteration, as follows:

$$\breve{\beta}_{t+1} = \breve{\beta}_t + (X'M_tX)^{-1}X'M_tH_t(\hat{p} - p_t),\ t = 0,1,2,\ldots \tag{2.8}$$

with starting values $M_0 = \hat{M}$, $\breve{\beta}_0 = (X'\hat{M}X)^{-1}X'\hat{M}\hat{g} = \hat{\beta}$, $H_0 = H(\hat{\beta})$ and $p_0 = p(\hat{\beta})$. Further, $M_t = (\hat{V}_{gt} + X_0X_0')^{-1}$ with $\hat{V}_{gt} = H_t\hat{V}_pH_t'$, $H_t = H(\breve{\beta}_t)$ and $p_t = p(\breve{\beta}_t)$, $t \geq 1$. At convergence, we get $\beta^* = (\beta_0^{*\prime},\beta_1^{*\prime})'$ as the solution of the following equations:

$$X'M(\beta)H(\beta)(\hat{p} - p(\beta)) = 0. \tag{2.9}$$

Equations (2.9) reduce to quasilikelihood equations (McCullagh 1983) when V_p is proportional to $V(p)$, a known function of p. Here, the dependence on β is made explicit by writing $p = p(\beta)$, $H = H(\beta)$ and $M = V_g + X_0X_0' = M(\beta)$. The smoothed estimate β^* also satisfies the constraints $K'p = K'g^{-1}(X\beta) = \pi$, unlike $\hat{\beta}$. The asymptotic covariance matrices of β_1^* and $\hat{\beta}_1$ are identical, but β_1^* might perform better in small samples.

Given the model (2.1), an alternate Wald test of the hypothesis $C_1\beta_1 = c_1$ is given by

$$W_1^* = (C_1\beta_1^* - c_1)'[C_1 \text{ est cov } (\beta_1^*)C_1']^{-1}(C_1\beta_1^* - c_1) \tag{2.10}$$

which is distributed asymptotically as a χ^2 with h d.f., where

$$\text{est cov}(\beta_1^*) = (X_1^{*\prime}M^*X_1^*)^{-1}, \tag{2.11}$$

and $X_1^* = [I - X_0X_0'M^*]X_1, M^* = (V_g^* + X_0X_0')^{-1}$ with $V_g^* = H^*\hat{V}_pH^{*\prime}$ and $H^* = H(\beta^*)$.

Example

The previous results were applied to a two-way table from the Canada Health Survey (1978-79). This survey was designed to provide reliable information on the health of Canadians. The information collected was made up of an interview component for the whole sample and a physical measures component for a subsample. A complex multistage design involving stratification and clustering was employed, and the estimates of cell totals or proportions were subjected to post-stratification on age-sex, to improve their efficiency. The reader is referred to Hidiroglou and Rao (1987) for a description of the survey and the procedures used for estimating cell counts, proportions, and their estimated variances and covariances. For the physical measures component, a collapsed stratum technique for variance estimation was employed since a single primary sampling unit was selected in some of the strata.

Table 1 gives the estimated proportions, $\hat{p}_{ij}$, derived from the physical measures component in a cross-classification of fitness level (recommended = 1, minimal acceptable = 2, below acceptable or screened out = 3) and type of cigarette smoker (regular = 1, occasional = 2, never = 3). The estimated covariance matrix of the $\hat{p}_{ij}$, $\hat{V}_p$, can be obtained from the authors.

Since both the variables in Table 1 are ordinal, we considered the following loglinear model with linear × linear interaction:

$$\log p_{ij} = \tilde{u} + u_{1(i)} + u_{2(j)} + \gamma(v_i - \bar{v})(w_j - \bar{w}), \quad i = 1,2,3 \quad j = 1,2,3 \tag{2.12}$$

Table 1

Estimated Cell Proportions in a 3 × 3 Table (Canada Level):
Type of Cigarette Smoker × Fitness Level (Sample Size n = 2505)
Ages 15-64

Type of cigarette smokers	Fitness Level		
	1	2	3
1	0.22005	0.14951	0.16998
2	0.02301	0.00962	0.01146
3	0.20329	0.09933	0.11374

subject to side constraints $\sum_i u_{1(i)} = \sum_j u_{2(j)} = 0$, where v_i and w_j are known scores with means $\bar{v}$ and $\bar{w}$ respectively. For simplicity, equidistant scores were taken: $u_i = 1,2,3$; $v_j = 1,2,3$. The model (2.12) is of the form $g(p) = X_0\beta_0 + X_1\beta_1$ with $g_{ij}(p) = \log p_{ij}$, $X_0 = K = 1_9$, a 9×1 vector of 1's, $\beta_0 = \tilde{u}$, $\beta_1 = (u_{1(1)}, u_{1(2)}, u_{2(1)}, u_{2(2)}, \gamma)'$, and

$$X_1' = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 & -1 & -1 & -1 \\ 0 & 0 & 0 & 1 & 1 & 1 & -1 & -1 & -1 \\ 1 & 0 & -1 & 1 & 0 & -1 & 1 & 0 & -1 \\ 0 & 1 & -1 & 0 & 1 & -1 & 0 & 1 & -1 \\ 1 & 0 & -1 & 0 & 0 & 0 & -1 & 0 & 1 \end{bmatrix}$$

Noting that $\hat{H} = \operatorname{diag}(\hat{p}_{ij}^{-1}, i = 1,2,3; j = 1,2,3)$, the Wald test of goodness-of-fit of the model (2.12) can be computed from (2.6), using the proportions $\hat{p}_{ij}$ in Table 1 and the estimated covariance matrix, $\hat{V}_p$. We obtain

$$W = 3.59$$

which is not significant at the 5% level compared to $\chi^2_{T-r}(0.05) = \chi^2_3(0.05) = 7.81$ (note that $T = 9, r = 6$). The Wald statistic W is likely to be stable in this example since the number of cells $T(= 9)$ is small relative to the number of sample clusters $(= 50)$.

We can also conduct a test of independence, *i.e.* $\gamma = 0$, given the model (2.12), using W_1 given by (2.7) or W_1^*, based on the smoothed estimates β_1^*, given by (2.10). Noting that $C_1 = (0,\ldots,0,1)$, $c_1 = 0$, we obtain

$$W_1 = 8.23, \quad W_1^* = 8.75,$$

both larger than $\chi^2_1(0.01) = 6.63$, the upper 1% point of χ^2 with 1 d.f. The nested hypothesis of independence is therefore not tenable.

Accepting the model (2.12), we obtain the following values of weighted least squares estimates, $\hat{\beta}_1$, and smoothed estimates, β^*:

$$\hat{\beta}_1 = (0.912, -1.550, 0.339, -0.255, -0.086)'$$

$$\beta_0^* = -2.665, \quad \beta_1^* = (0.917, -1.568, 0.344, -0.262, 0.087)'.$$

The estimate β^* can also be used to produce smoothed estimates of the p_{ij}, $p_{ij}^* = p_{ij}(\beta^*)$, which satisfy the constraint $\sum\sum p_{ij}(\beta^*) = 1$.

3. BOX-COX TRANSFORMATION MODELS

Logistic regression models are extensively used for the analysis of variation in the estimated proportions associated with a binary response variable. Suppose that the population of interest is partitioned into I cells according to the levels of one or more factors. Let P_i be the population response proportion in the i-th cell. Then a logistic regression model for the proportions $P_i = F_i(\beta)$ is given by

$$\log\{F_i/(1 - F_i)\} = x_i'\beta, \quad i = 1, \ldots, I, \tag{3.1}$$

where $x_i = (x_{1i}, \ldots, x_{si})'$ is an s-vector of known constants derived from the factor levels with $x_{1i} = 1$, and β is an s-vector of unknown parameters.

Guerrero and Johnson (1982) extended the applicability of logistic regression models by introducing an additional parameter, λ, through a Box-Cox power transformation of the odds ratios $F_i/(1 - F_i)$. Their model is given by

$$v_i(\lambda) = \{F_i/(1 - F_i)\}^{(\lambda)} = x_i'\beta, \quad i = 1, \ldots, I, \tag{3.2}$$

where β and x_i are as in (3.1) and

$$\{F_i/(1 - F_i)\}^{(\lambda)} = \begin{cases} \log\{F_i/(1 - F_i)\} & \text{if } \lambda = 0 \\ \lambda^{-1}[\{F_i/(1 - F_i)\}^{\lambda} - 1] & \text{if } \lambda \neq 0. \end{cases}$$

The model (3.2) includes as a special case ($\lambda = 0$) the logistic regression model (3.1). Guerrero and Johnson (1982) applied this model to data from the National Survey of Household Income and Expenditures in Mexico to explain the variation in female participation in the Mexican labour force. They found that a value of $\lambda = -6.63$ provided a significantly better fit than the logit model ($\lambda = 0$), the values of the standard chi-squared statistic being 4.8 (7 d.f.) and 12.8 (8 d.f.) respectively. However, they applied standard methods for binomial proportions, ignoring the survey design.

Pseudo MLE

In this section, the methods of Roberts, Rao and Kumar (1987) for the logistic regression model are extended to the power transformation model (3.2). Due to difficulties in obtaining appropriate likelihood functions for general sample designs, we use "pseudo" maximum likelihood estimates, $\hat\beta$ and $\hat\lambda$, obtained from the product binomial likelihood equations for β and λ by replacing the simple response proportion r_i/n_i with the corresponding survey estimate $\hat P_i$ of P_i, and n_i/n with the corresponding survey estimate $\hat W_i$ of the domain proportion W_i. Here r_i is the number of "successes" in a sample of size n_i from the i-th cell, and $n = \sum n_i$. See Guerrero and Johnson (1982), for the product binomial likelihood equations. The pseudo maximum likelihood estimates (m.l.e.), $\hat\theta' = (\hat\beta', \hat\lambda)$, can be obtained iteratively by a quasi-Newton procedure, as in Guerrero and Johnson (1982). The fitted response proportions are given by $\hat F = F_i(\hat\theta)$.

Let $\hat V_P$ be the estimated covariance matrix of the survey estimates,$\hat P = (\hat P_1, \ldots, \hat P_I)'$, and let

$$B = D(\hat F)^{-1} D(1 - \hat F)^{-1} (\partial F/\partial\hat\theta)'. \tag{3.3}$$

Here $D(\hat F) = \text{diag}(\hat F_i, i = 1, \ldots, I)$, $D(1 - \hat F) = \text{diag}(1 - \hat F_i, i = 1, \ldots, I)$ and $(\partial F/\partial\hat\theta)'$ is the $I \times (s + 1)$ matrix of partial derivatives $\partial F_i/\partial\beta_j$ and $\partial F_i/\partial\lambda$ evaluated at $\hat\theta$:

$$\partial F_i/\partial\beta_j = x_{ji} F_i^2 (1/Q_i)^{1+1/\lambda}$$

$$\partial F_i/\partial\lambda = F_i^2 (Q_i \log Q_i - Q_i + 1)\lambda^{-2}(1/Q_i)^{1+1/\lambda}, \tag{3.4}$$

where $Q_i = 1 + \lambda\sum_j x_{ji}\beta_j$. The estimated asymptotic covariance matrix of $\hat\theta$, taking account of the survey design, is then given by (see Roberts 1985)

$$\text{est cov}(\hat\theta) = (B'\hat\Delta B)^{-1}(B'D(\hat W)\hat V_P D(\hat W)B)(B'\hat\Delta B)^{-1}, \tag{3.5}$$

where $\hat{\Delta} = \mathrm{diag}(\hat{W}_i\hat{F}_i(1 - \hat{F}_i); i = 1,\ldots,I)$ and $D(\hat{W}) = \mathrm{diag}(\hat{W}_i, i = 1,\ldots,I)$.

It is also of interest to find the standard errors of the residuals $\hat{R}_i = \hat{P}_i - \hat{F}_i$ since the standardized residuals $\hat{R}_i/\mathrm{s.e.}(\hat{R}_i)$ can be used to detect any outlying cell proportions. The estimated asymptotic covariance matrix of the vector of residuals $\hat{R} = (\hat{R}_1, \ldots, R_I)'$ is given by

$$\mathrm{est\,cov}(\hat{R}) = A\,\mathrm{est\,cov}(\hat{\theta})A' = \hat{V}_R, \tag{3.6}$$

where

$$A = I - D(\hat{F})D(1 - \hat{F})B(B'\hat{\Delta}B)^{-1}B'D(\hat{W}).$$

The square root of the diagonal elements, $\hat{V}_{ii,R}$, of (3.6) provide the estimated standard errors of the $\hat{R}_i, i = 1, \ldots, I$.

Corrections to Standard Tests

The standard chi-squared and likelihood ratio tests of goodness-of-fit of the model (3.2) are given by

$$X^2 = n \sum_{i=1}^{I} (\hat{P}_i - \hat{F}_i)^2 \hat{W}_i / \{\hat{F}_i(1 - \hat{F}_i)\} \tag{3.7}$$

and

$$G^2 = 2n \sum_{i=1}^{I} \hat{W}_i\big[\hat{P}_i \log(\hat{P}_i/\hat{F}_i) + (1 - \hat{P}_i)\log\{(1 - \hat{P}_i)/(1 - \hat{F}_i)\}\big], \tag{3.8}$$

respectively, where the term in [ ] of (3.8) equals $-\log(1 - \hat{F}_i)$ at $\hat{P}_i = 0$ and $-\log\hat{F}_i$ at $\hat{P}_i = 1$.

Under product binomial sampling, it is well-known that both X^2 and G^2 are asymptotically identically distributed as a χ^2 variable with $I - s - 1$ d.f., but for general sample designs this result is no longer valid. In fact, X^2 (or G^2) is asymptotically distributed as a weighted sum, $\sum\delta_k W_k$, of independent χ^2 variables, W_k, each with 1 d.f., where the weights δ_k $(k = 1, \ldots, I - s - 1)$ can be interpreted as "generalized design effects" (see Roberts 1985). Under product binomial sampling, $\delta_k = 1$ for all k, and $\sum\delta_k W_k$ reduces to χ^2 with $I - s - 1$ d.f.

A first-order correction to X^2 (or G^2) is obtained by treating $X_c^2 = X^2/\hat{\delta}.$ or $G_c^2 = G^2/\hat{\delta}.$ as χ^2 with $I - s - 1$ d.f., where

$$(I - s - 1)\hat{\delta}. = \sum \hat{\delta}_k = n \sum_{i=1}^{I} \hat{V}_{ii,R}\hat{W}_i / \{\hat{F}_i(1 - \hat{F}_i)\} \tag{3.9}$$

and $\hat{V}_{ii,R}$ is the estimated variance of the i-th residual $\hat{R}_i$.

A more accurate, second order correction to X^2 (or G^2), based on the Satterthwaite approximation to $\sum\delta_k W_k$, is obtained by treating

$$X_S^2 = \frac{X_c^2}{1 + \hat{a}^2} \text{ or } G_S^2 = \frac{G_c^2}{1 + \hat{a}^2} \text{ as } \chi^2 \text{ with } (I - s - 1)/(1 + \hat{a}^2) \text{ d.f.} \tag{3.10}$$

Here $\hat{a}^2 = \sum(\hat{\delta}_k - \hat{\delta}.)^2/\{(I - s - 1)\hat{\delta}.^2\}$ is the squared coefficient of variation of the $\hat{\delta}_i$ which can be computed, without evaluating the individual weights $\hat{\delta}_i$, from (3.9) and from

$$\sum \hat{\delta}_k^2 = \sum_{i=1}^{I} \sum_{l=1}^{I} \hat{V}_{il,R}^2 (n\hat{W}_i)(n\hat{W}_l)/\{\hat{f}_i\hat{f}_l(1 - \hat{f}_i)(1 - \hat{f}_l)\}, \qquad (3.11)$$

where $\hat{V}_{il,R}$ is the (i,l)-th element of $\hat{V}_R$ given by (3.6).

Nested hypotheses, given the model (3.2), can also be tested by correcting the standard tests for nested hypotheses, but we omit this topic for simplicity (see Roberts 1985 and Kumar and Rao 1985 for details). It is simpler, however, to use Wald tests based on the estimates $\hat{\beta}$ and the associated estimated asymptotic covariance matrix.

Example

The previous method was applied to data from the monthly Canadian Labour Force Survey (October, 1980). The Labour Force Survey design employs multi-stage cluster sampling with two stages in the self-representing urban areas and three or four stages in the non-self-representing areas in each province. A detailed description of the sample design and associated estimation procedures for the Labour Force Survey is given in Statistics Canada (1977).

The sample from the Labour Force Survey, for the present example, consisted of males aged 15-64 who were in the labour force and not full-time students. Two factors, age and education, were chosen to explain the unemployment rates via a Box-Cox transformation model. Age-group levels were formed by dividing the interval $[15,64]$ into ten groups with the j-th age group being the interval $[10 + 5j, 14 + 5j]$ for $j = 1, \ldots, 10$ and then using the mid-point of each interval, $A_j = 12 + 5j$, as the value of age for all persons in that age group. Similarly, the levels of education, E_k, were formed by assigning to each person a value based on the median years of school resulting in the following six levels: 7, 10, 12, 13, 14 and 16. The resultant age by education cross-classification provides a two-way table of $I = 60$ survey estimates, $\hat{P}_{jk}$, of employment rates P_{jk}. The estimated covariance matrix $\hat{V}_P$ was based on more than 450 sample clusters.

We considered the following transformation model for $P_{jk} = F_{jk}(\theta)$ involving linear and quadratic age effects and linear education effect:

$$\begin{aligned} \nu_{jk}(\lambda) &= \{F_{jk}/(1 - F_{jk}\}^{(\lambda)} \\ &= \beta_0 + \beta_1 A_j + \beta_2 A_j^2 + \beta_3 E_k, \, j = 1, \ldots, 10, \, k = 1, \ldots, 6. \end{aligned} \qquad (3.12)$$

Table 2 contains the pseudo m.l.e. of $\theta = (\beta_0, \beta_1, \beta_2, \beta_3, \lambda)'$ and associated standard errors, and the test statistics X^2, G^2, X_S^2 and G_S^2 for testing the goodness-of-fit of the model (3.12). The corresponding values under the logistic regression model ($\lambda = 0$) are also given for comparison.

It is clear from Table 2 that the value of X^2 (or G^2) is essentially equal to the corresponding value under the logistic regression model. Thus in the present example the transformation model provides no improvement in the fit over the logistic regression model. This is also clear from the value of $\hat{\lambda}$ ($= 0.016$) which is not significantly different from $\lambda = 0$ when compared to its standard error ($= 0.085$). The estimates of regression coefficients are essentially equal under the two models, but the standard errors of the $\hat{\beta}_i$ under the Box-Cox model are much larger than the corresponding standard errors under the logistic regression model, due to the large standard error associated with $\hat{\lambda}$ and the fact that the $\hat{\beta}_i$ depend on $\hat{\lambda}$.

Table 2

Pseudo MLE of the Parameters $(\hat{\beta}', \lambda)$, their Standard Errors and Test Statistics Under the Transformation Model and under the Corresponding Logistic Regression Model ($\lambda = 0$)

	Transformation Model		Logistic Regression Model	
	estimate	s.e.	estimate	s.e.
$\hat{\beta}_0$	−3.28	0.975	−3.10	0.247
$\hat{\beta}_1$	0.219	0.0468	0.211	0.013
$\hat{\beta}_2$	−0.00227	0.00049	−0.00218	0.00017
$\hat{\beta}_3$	0.1579	0.0385	0.1509	0.0115
$\hat{\lambda}$	0.016	0.085	—	—
	Test Statistics			
	value	d.f.	value	d.f.
X^2	99.6	55	99.8	56
G_2	102.6	56	102.5	56
X_S^2	40.7	39.2	23.4	24.2
G_S^2	42.0	39.2	23.9	24.2
$X_S^2(0.05)$	54.6	55	47.7	56
$G_S^2(0.05)$	56.4	55	48.9	56

If the survey design is ignored and the value of X^2 (or G^2) is referred to $\chi^2_{0.05}(55) = 73.3$, the upper 5% point of χ^2 with $I - s - 1 = 55$ d.f., we would reject the model (3.12). On the other hand, the value of X_S^2 (or G_S^2) when adjusted to refer to $\chi^2_{0.05}(55)$, denoted as X_S^2 (0.05) (or $G_S^2(0.05)$) in Table 2, is not significant at the 5% level, indicating that the model provides a good fit to the data, $\hat{P}_{jk}$.

Box and Cox (1982) and Hinkley and Runger (1984) argued that statistical inference about β should proceed with the scale determined by the estimate $\hat{\lambda}$ regarded as fixed. Thus, the estimated covariance matrix of $\hat{\beta}$ is determined from (3.5) by replacing $\partial F/\partial\hat{\theta}$ by $\partial F/\partial\hat{\beta}$ in the expression for B (equation (3.3)). For our example, this argument would suggest that we can take $\hat{\lambda} = 0$ and use the estimates of β and associated standard errors (or estimated covariance matrix) under the logistic regression model, given in Table 2.

4. TESTING EQUALITY OF LOGISTIC REGRESSION MODELS

Structural changes between two time periods may be detected through tests of equality of parameters in the corresponding models. Such tests for standard linear regression models have been developed extensively in the econometric literature (see *e.g.*, Amemiya 1985, Sec. 1.5.3).

In this section, corrected chi-squared and likelihood ratio tests of equality of parameters in two logistic regression models, corresponding to two specified time periods, are obtained. If the hypothesis of equality is tenable, then "smoothed" (*i.e.*, fitted) estimates of cell proportions for the current period can be obtained by combining the data for the two periods.

These estimates are more efficient than the corresponding smoothed estimate based only on the current period data. The methodology is applied to data from the October 1980 and October 1981 Canadian Labour Force Survey, to study year-to-year structural changes. Note that the data for October 1980 has already been used, in Section 3, to illustrate the fitting of Box-Cox power transformation models, and it was found that a logistic regression model involving linear and quadratic age effects and linear education effect provides a good fit to the data.

Let P_{ti} be the population response proportion in the i-th cell for the period $t(=1,2)$. Then a logistic regression model for the proportions $P_{ti} = F_i(\beta_t) = F_{ti}$ is given by

$$\log\{F_{ti}/(1 - F_{ti})\} = x_i'\beta_t, \quad i = 1,\ldots,I;\ t = 1,2 \tag{4.1}$$

where x_i is an s-vector of known constants derived from the factor levels, as in (3.1), and β_t is an s-vector of unknown parameters for period t. We are interested in testing the composite hypothesis $\beta_1 = \beta_2(=\beta)$ to study structural changes between the two time periods. If the hypothesis is accepted, "smoothed" estimates of the proportions P_{2i} for the current period $(t = 2)$ can be obtained as $F_i(\hat{\hat{\beta}})$ where $\hat{\hat{\beta}}$ is the pseudo m.l.e. of the common parameter β.

Pseudo MLE

Let $\hat{P}_{1i}$ and $\hat{P}_{2i}$ $(i = 1,\ldots,I)$ be the survey estimates based on sample sizes n_1 and n_2 respectively. Extending the notation in Section 3, "pseudo" maximum likelihood estimates, $\hat{\beta}_t$, are obtained from the product binomial likelihood equations for β_t by replacing the simple response proportions r_{ti}/n_{ti} with the corresponding survey estimates $\hat{P}_{ti}$ of P_{ti} and n_{ti}/n_t with the corresponding survey estimates $\hat{W}_{ti}$ of the domain proportions W_{ti}, thus yielding

$$X'D(\hat{W}_t)\hat{F}_t = X'D(\hat{W}_t)\hat{P}_t, \quad t = 1,2 \tag{4.2}$$

where $\hat{F}_t = F(\hat{\beta}_t)$ is the vector of fitted response proportions for period t, $D(\hat{W}_t) = \text{diag}\,(\hat{W}_{ti}, i = 1,\ldots,I)$, and $X' = (x_1,\ldots,x_I)$. The estimates $\hat{\beta}_t$ are obtained iteratively by a quasi-Newton procedure.

Under the hypothesis $\beta_1 = \beta_2(=\beta)$, the pseudo maximum likelihood estimates, $\hat{\hat{\beta}}$, are obtained by iteration from the following pseudo likelihood equations:

$$X'D(\hat{W}_c)\hat{\hat{F}} = (n_1/n)X'D(\hat{W}_1)\hat{P}_1 + (n_2/n)X'D(\hat{W}_2)\hat{P}_2, \tag{4.3}$$

where $D(\hat{W}_c) = (n_1/n)D(\hat{W}_1) + (n_2/n)D(\hat{W}_2)$, $\hat{\hat{F}} = F(\hat{\hat{\beta}})$ is the vector of fitted response proportions or smoothed estimates of cell proportions for the current period, and $n_1 + n_2 = n$.

Let $\hat{V}_P$ be the estimated covariance matrix of $(\hat{P}_1', \hat{P}_2')'$ partitioned as

$$\hat{V}_P = \begin{bmatrix} \hat{V}_{11P} & \hat{V}_{12P} \\ \hat{V}_{21P} & \hat{V}_{22P} \end{bmatrix}.$$

Then the estimated covariance matrix of smoothed estimates $\hat{\hat{F}}$ is given by

$$\text{est cov}(\hat{\hat{F}}) = B\hat{V}_P B', \tag{4.4}$$

where

$$B = D(\hat{W}_c)^{-1}\hat{\hat{\Delta}}X(X'\hat{\hat{\Delta}}X)^{-1}X'\left[(n_1/n)D(\hat{W}_1),(n_2/n)D(\hat{W}_2)\right] \tag{4.5}$$

and

$$\hat{\hat{\Delta}} = \operatorname{diag}(\hat{W}_c\hat{\hat{F}}_i(1 - \hat{\hat{F}}_i)), i = 1,\ldots,I.$$

If the residuals are defined as $\hat{\hat{R}}_t = \hat{F}_t - \hat{\hat{F}}$, then the estimated covariance matrix of $(\hat{\hat{R}}_1', \hat{\hat{R}}_2')'$ is given by

$$\hat{\hat{V}}_R = \begin{bmatrix} \hat{\hat{V}}_{11R} & \hat{\hat{V}}_{12R} \\ \hat{\hat{V}}_{21R} & \hat{\hat{V}}_{22R} \end{bmatrix} = A\hat{V}_pA'. \tag{4.6}$$

Here

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$$

with

$$A_{t1} = D(\hat{W}_c)^{-1}\hat{\hat{\Delta}}X\left[(X'\hat{\hat{\Delta}}_tX)^{-1}X'D(\hat{W}_t) - \frac{n_t}{n}(X'\hat{\hat{\Delta}}X)^{-1}X'D(\hat{W}_t)\right],$$

and

$$A_{t2} = -D(\hat{W}_c)^{-1}\hat{\hat{\Delta}}X(X'\hat{\hat{\Delta}}X)^{-1}X'\left\{D(\hat{W}) - \frac{n_t}{n}D(\hat{W}_t)\right\},\ t = 1,2,$$

where

$$\hat{\hat{\Delta}}_t = \operatorname{diag}(\hat{W}_{ti}\hat{\hat{F}}_i(1 - \hat{\hat{F}}_i), i = 1,\ldots,I).$$

Corrections to Standard Tests

The standard chi-squared and likelihood ratio tests of the nested hypothesis $\beta_1 = \beta_2$, given the model (4.1), are given by

$$X^2 = X_1^2 + X_2^2 \tag{4.8}$$

and

$$G^2 = G_1^2 + G_2^2, \tag{4.9}$$

where

$$X_t^2 = n_t\sum_{i=1}^{I}(\hat{F}_{ti} - \hat{\hat{F}}_i)^2\hat{W}_{ti}\big/\{\hat{\hat{F}}_i(1 - \hat{\hat{F}}_i)\},\ t = 1,2 \tag{4.10}$$

and

$$G_t^2 = 2n_t\sum_{i=1}^{I}\hat{W}_{ti}\left[\hat{F}_{ti}\log(\hat{F}_{ti}/\hat{\hat{F}}_i) + (1 - \hat{F}_{ti})\log\{(1 - \hat{F}_{ti})/(1 - \hat{\hat{F}}_i)\}\right],\ t = 1,2. \tag{4.11}$$

A first order correction to X^2 (or G^2) is obtained by treating $X_c^2 = X^2/\hat{\delta}_\cdot$ or $G_c^2 = G^2/\hat{\delta}_\cdot$ as χ^2 with s d.f., where

$$s\hat{\delta}_\cdot = n_1 \sum_{i=1}^{I} \hat{V}_{11R}(ii)\hat{W}_{1i}/\{\hat{\hat{F}}_i(1-\hat{\hat{F}}_i)\} + n_2 \sum_{i=1}^{I} \hat{V}_{22R}(ii)\hat{W}_{2i}/\{\hat{\hat{F}}_i(1-\hat{\hat{F}}_i)\} \qquad (4.12)$$

and $\hat{V}_{ttR}(ij)$ is the (i,j)th element of $\hat{V}_{ttR}$. A more accurate, second order correction to X^2 (or G^2), based on the Satterthwaite approximation, is obtained by treating

$$X_S^2 = \frac{X_c^2}{1+\hat{a}^2} \quad \text{or} \quad G_S^2 = \frac{G_c^2}{1+\hat{a}^2} \quad \text{as} \quad \chi^2 \quad \text{with} \quad s/(1+\hat{a}^2) \quad \text{d.f.} \qquad (4.13)$$

Here $\hat{a}^2 = (\sum_{k=1}^{s}\hat{\delta}_k^2 - s\hat{\delta}_\cdot^2)/s\hat{\delta}_\cdot^2$ which can be computed from (4.12) and the following formula for $\sum\hat{\delta}_k^2$:

$$\begin{aligned}\sum_{k=1}^{s}\hat{\delta}_k^2 = {} & n_1^2 \sum_{i=1}^{I}\sum_{j=1}^{I} \frac{\hat{V}_{11R}^2(ij)\hat{W}_{1i}\hat{W}_{1j}}{\hat{\hat{F}}_i\hat{\hat{F}}_j(1-\hat{\hat{F}}_i)(1-\hat{\hat{F}}_j)} \\ & + n_2^2 \sum_{i=1}^{I}\sum_{j=1}^{I} \frac{\hat{V}_{22R}^2(ij)\hat{W}_{2i}\hat{W}_{2j}}{\hat{\hat{F}}_i\hat{\hat{F}}_j(1-\hat{\hat{F}}_i)(1-\hat{\hat{F}}_j)} \\ & + 2n_1n_2 \sum_{i=1}^{I}\sum_{j=1}^{I} \frac{\hat{V}_{12R}^2(ij)\hat{W}_{1i}\hat{W}_{2j}}{\hat{\hat{F}}_i\hat{\hat{F}}_j(1-\hat{\hat{F}}_i)(1-\hat{\hat{F}}_j)}, \end{aligned} \qquad (4.14)$$

where $\hat{V}_{12R}(ij)$ is the (i,j)-th element of $\hat{V}_{12R}$.

Example

The previous method was applied to data from the October 1980 and October 1981 Canadian Labour Survey, to study year-to-year structural changes.

The logistic regression model involving linear and quadratic age effects and linear education effect provided a good fit to data from both periods with the following estimates of β_t:

$$\hat{\beta}_1: \{-3.08, 0.211, -0.00218, 0.1505\}$$

$$\hat{\beta}_2: \{-3.05, 0.179, -0.00169, 0.1707\},$$

where $\log\{\hat{F}_{tjk}/(1-\hat{F}_{tjk})\} = \hat{\beta}_{t0} + \hat{\beta}_{t1}A_j + \hat{\beta}_{t2}A_j^2 + \hat{\beta}_{t3}E_k, j = 1,\ldots,10; k = 1,\ldots,6$ and $\hat{F}_{tjk}$ is the fitted employment rate in the (j,k)-th cell for period t. One cell was omitted in the fitting since the domain sample size n_{2i} is zero for the current period.

Turning to the test of the hypothesis $\beta_1 = \beta_2$, given the logistic regression models, we obtained the following values of X^2, G^2, X_c^2, G_c^2 and X_S^2, G_S^2:

$$X^2 = 42.1 \qquad X_c^2 = 24.6 \qquad X_S^2 = 24.4$$

$$G^2 = 42.2 \qquad G_c^2 = 24.6 \qquad G_S^2 = 24.4.$$

Also $s/(1 + \hat{a}^2) = 4/(1.0089) = 3.965 \doteq 4$. By referring X_S^2 or G_S^2 to $\chi^2_{0.05}(4) = 9.49$, the upper 5% point of χ^2 with 4 d.f., we reject the hypothesis $\beta_1 = \beta_2$ at the 5% level, indicating significant year-to-year structural changes for the month of October. The data for the two time periods, therefore, should not be pooled to get smoothed estimates of unemployment rates, $1 - \hat{\hat{F}}_{jk}$, for the current period.

5. POLYTOMOUS RESPONSE MODELS

A variety of models has been suggested in the literature when the response variable is polytomous. The variety of models reflects, in part, the different scales of measurement possible for polytomous response variables, unlike binary response variables. In the main, there are nominal responses where any permutation of the response categories is equally valid, and ordinal responses where there is a natural ordering of the response categories.

Suppose that the population of interest is partitioned into I cells (or domains) according to the levels of one or more factors. Let $P_{j(i)}$ be the population proportion in the i^{th} cell having the j^{th} response $(j = 1, \ldots, J + 1)$ so that $\sum_{j=1}^{J+1} P_j(i) = 1$ $(i = 1, \ldots, I)$. Then a general polytomous response model for the proportions $P_j(i)$ is given by

$$P_j(i) = F_{ij}(\theta), \; i = 1, \ldots, I; \; j = 1, \ldots, J, \tag{5.1}$$

where θ is an r-vector of unknown parameters $(r \leq IJ)$ and $F_{ij}(\theta)$ is a function of known form. In the nominal case, Haberman (1982) and others proposed the following model: the "multinomial logits" $\log P_j(i) - \sum_{j'=1}^{J+1} \log P_{j'(i)} (J + 1)^{-1}$ are assumed to be unknown linear functions of x_i, the s-vector of known constants derived from the factor levels, *i.e.*,

$$F_{ij}(\theta) = \exp(x_i'\beta_j) \Big/ \sum_{k=1}^{J+1} \exp(x_i'\beta_k), \; i = 1, \ldots, I; \; j = 1, \ldots, J + 1 \tag{5.2}$$

with $\sum \beta_k = 0$. Because of the latter constraint on the β_k, (5.2) may be expressed as

$$F_{ij}(\theta) = \exp(x_i'\beta_j) \Big/ \left[\sum_{k=1}^{J} \exp(x_i'\beta_k) + \prod_{k=1}^{J} \exp(-x_i'\beta_k) \right],$$

$$i = 1, \ldots, I; j = 1, \ldots, J. \tag{5.3}$$

Note that (5.3) reduces to the usual logistic regression model in the special case of binary response.

In the ordinal case, a simple model which also has the feature of being invariant under the grouping of response categories is given by (McCullagh 1980)

$$\log\{C_{j(i)}/(1 - C_{j(i)})\} = \nu_j - x_i'\beta, \; j = 1, \ldots, J; \; i = 1, \ldots, I \tag{5.4}$$

where $C_{j(i)} = \sum_{k=1}^{j} P_{k(i)}$ denotes the j^{th} cumulative probability in the i^{th} domain, and $\theta' = (\nu_1, \ldots, \nu_J, \beta')$. To express (5.4) in the form (5.1), we note that $P_i = L^{-1}C_i$, where $P_i = (P_{1(i)}, \ldots, P_{J(i)})'$, $C_i = (C_{1(i)}, \ldots, C_{J(i)})'$ and L^{-1} is a $J \times J$ nonsingular matrix with 1 in the diagonal, -1 in the $(i + 1, i)^{th}$ position $(i < J)$ and 0 elsewhere.

Pseudo MLE

As before, we use pseudo m.l.e., $\hat{\theta}$ obtained from the product multinomial likelihood equations for θ by replacing the simple response proportions n_{ij}/n_i with the corresponding survey estimates $\hat{P}_{j(i)}$, and n_i/n with the corresponding survey estimate $\hat{W}_i$ of the domain proportion W_i. Here n_{ij} is the number of units with the j^{th} response in a sample of size n_i from the i^{th} domain and $n = \sum n_i$. The fitted response proportions are then given by $\hat{F} = F(\hat{\theta}) = (\hat{F}_1', \ldots, \hat{F}_I')'$, where $\hat{F}_i = (\hat{F}_{i1}, \ldots, \hat{F}_{iJ})'$ and $\hat{F}_{ij} = F_{ij}(\hat{\theta})$.

Let $\hat{V}_P$ be the estimated covariance matrix of the survey estimates $\hat{P} = (\hat{P}_{1(1)}, \ldots, \hat{P}_{J(1)}, \ldots, \hat{P}_{1(I)}, \ldots, \hat{P}_{J(I)})'$, and $\hat{M} = (\partial F/\partial\hat{\theta})'$, the $IJ \times r$ matrix of partial derivatives $\partial F_{ij}/\partial\theta_k$ calculated at $\hat{\theta}$. Also, let $\hat{Q}_i = \text{diag}(\hat{F}_i) - \hat{F}_i\hat{F}_i'$ and $\hat{Q} = \text{diag}(\hat{Q}_i, i = 1, \ldots, I)$. The expressions for the partial derivatives $\partial F_{ij}/\partial\theta_k$ for the models (5.3) and (5.4) are given in Roberts (1985). The estimated asymptotic covariance matrix of $\hat{\theta}$, taking account of the survey design, is then given by (see Roberts 1985).

$$\text{estcov}(\hat{\theta}) = (\hat{M}'\hat{\nabla}\hat{M})^{-1}(\hat{M}'\hat{\nabla}\hat{V}_P\hat{\nabla}'\hat{M})(\hat{M}'\hat{\nabla}\hat{M})^{-1}, \tag{5.5}$$

where $\hat{\nabla} = (D(\hat{W}) \otimes I)\hat{Q}^{-1}$ and $D(\hat{W}) = \text{diag}(\hat{W}_i, i = 1, \ldots, I)$. In the special case of product multinomial sampling, $\hat{V}_P = \hat{\nabla}^{-1}/n$ and (5.5) reduces to $(\hat{M}'\hat{\nabla}\hat{M})^{-1}/n$.

The vector of residuals, $\hat{R} = \hat{P} - \hat{F}$, is also of interest, since it may be useful in detecting model deviations. The estimated asymptotic covariance matrix of $\hat{R}$ is given by

$$\text{estcov}(\hat{R}) = \hat{G}\hat{V}_P\hat{G}' \tag{5.6}$$

where $\hat{G} = I - \hat{M}(\hat{M}'\hat{\nabla}\hat{M})^{-1}\hat{M}'\hat{\nabla}$.

Corrections to standard tests

For simplicity, we consider only the Pearson chi-squared test of goodness-of-fit of the model (5.1). It is given by

$$X^2 = n \sum_{i=1}^{I} \hat{W}_i \sum_{j=1}^{J+1} (\hat{P}_{j(i)} - \hat{F}_{ij})^2/\hat{F}_{ij}. \tag{5.7}$$

Under independent multinomial sampling in each of the domains, it is well-known that X^2 is asymptotically distributed as a χ^2 variable with $IJ - r$ d.f.

To test the nested hypothesis $\theta_2 = 0$, given the model (5.1), let $\hat{\hat{\theta}}_1$ be the pseudo m.l.e. of θ_1 and $\hat{\hat{F}}$ be the corresponding vector of fitted response proportions, where $\theta' = (\theta_1', \theta_2')$, θ_1 is $q \times 1$ and θ_2 is $u \times 1$ $(q + u = r)$. The Pearson chi-squared test of the nested hypothesis is then given by

$$X^2(2|1) = n \sum_{i=1}^{I} \hat{W}_i \sum_{j=1}^{J+1} (\hat{F}_{ij} - \hat{\hat{F}}_{ij})^2/\hat{\hat{F}}_{ij} \tag{5.8}$$

which is asymptotically distributed as χ^2 with u d.f. under independent multinomial sampling in each of the domains. However, for a general sample design, X^2 and $X^2(2|1)$ are both asymptotically distributed as weighted sums of independent χ^2 variables, each with 1 d.f., where the weights can be interpreted as "generalized design effects" of particular linear transformations of $\hat{P}$ (Roberts 1985).

A first-order correction to $X^2(2|1)$ is obtained by treating

$$X_c^2(2|1) = X^2(2|1)/\hat{\delta}.(2|1) \text{ as } \chi^2 \text{ with } u \text{ d.f.}, \tag{5.9}$$

where $\hat{\delta}.(2|1)$ is obtained by replacing θ' by $(\hat{\hat{\theta}}_1',0')$ and V_P by $\hat{V}_P$ in the following definition for $\delta.(2|1)$:

$$u\delta.(2|1) = \sum_{i=1}^{u} \delta_i(2|1) = \text{tr } D(2|1). \tag{5.10}$$

Here, tr denotes the trace operator and $D(2|1)$ is a generalized design effects matrix given by

$$D(2|1) = (H_2'\nabla H_2)^{-1}(H_2'\nabla V_P \nabla' H_2), \tag{5.11}$$

where V_P is the covariance matrix of $\hat{P}$, $\nabla = (D(W) \otimes I)Q^{-1}$, Q is the block diagonal matrix with $Q_i = \text{diag}(F_i) - F_i F_i'$, $i = 1, \ldots, I$, $F_i = F_i(\theta)$, and $H_2 = [I - M_1 \; (M_1'\nabla M_1)^{-1} \, M_1'\nabla]M_2$, where $M_1 = (\partial F/\partial\theta_1)'$ and $M_2 = (\partial F/\partial\theta_2)'$.

A more accurate, second order correction to $X^2(2|1)$, based on the Satterthwaite approximation, is obtained by treating

$$X_S^2(2|1) = X_c^2(2|1)/[1 + \hat{a}(2|1)^2] \text{ as } \chi^2 \text{ with } u/[1 + \hat{a}(2|1)^2] \text{ d.f.} \tag{5.12}$$

Here $\hat{a}(2|1)^2$ is obtained by replacing θ by $(\hat{\hat{\theta}}_1',0')$ in the following definition of $a(2|1)^2$:

$$a(2|1)^2 = \left\{ \sum_{i=1}^{u} \delta_i(2|1)^2 - u\delta.(2|1)^2 \right\} \Big/ u\delta.(2|1)^2, \tag{5.13}$$

where

$$\sum_{i=1}^{u} \delta_i(2|1)^2 = \text{tr} D(2|1)^2. \tag{5.14}$$

The corrections to goodness-of-fit test X^2 are obtained as special cases of (5.9) and (5.12) by treating the model as nested within a saturated model (*i.e.*, a model where the unknown parameter θ is of length IJ).

Example

The previous methods were applied to data from the Canada Health Survey (1978-79). A brief description of the survey is provided in Section 2.

The data set examined consisted of the estimated counts of females aged 20-64 cross-classified by frequency of breast self-examination (with the 3 categories: monthly, quarterly, less often or never), education (with the 3 categories: secondary or less, some post-secondary, post-secondary) and age (with the 3 categories: 20-24, 25-44, 45-64).

The frequency of breast self-examination was considered to be the response variable, while education and age were taken as explanatory variables, so that the number of responses, $J + 1$, equalled 3 and the number of domains, I, was 9. Both response and explanatory variables are ordered.

Table 3

Survey Estimates of Cumulated Probabilities

	Age	Education	$C_{1(ik)}$	$C_{2(ik)}$
$i = 1, k = 1$	20-24	≤ Secondary	.25	.49
$k = 2$		< Post-Secondary	.25	.41
$k = 3$		≥ Post-Secondary	.23	.47
$i = 2, k = 1$	25-44	≤ Secondary	.25	.50
$k = 2$		< Post-Secondary	.27	.44
$k = 3$		≥ Post-Secondary	.26	.44
$i = 3, k = 1$	45-64	≤ Secondary	.28	.51
$k = 2$		< Post-Secondary	.24	.62
$k = 3$		≥ Post-Secondary	.29	.56

Table 4

Statistics for Testing Goodness of Fit and Nested Hypotheses

	Goodness of Fit (Age & Education)	Nested Hypothesis (Age only)
X^2	37.7	7.1
X_c^2	21.6	3.8
X_S^2	18.5*	3.7*
$\hat{\delta}.$	1.75	1.9
$\hat{a}^2$	0.83	0.1

* The Satterthwaite statistic has been adjusted to refer to the same χ^2 value as X_c^2.

The following model for the cumulated probabilities of the type described in equation (5.4), was considered:

$$\log\{C_j(ik)/(1 - C_j(ik))\} = \nu_j + \beta a_i + e_k \quad (j = 1,2; i = 1,2,3; k = 1,2,3) \qquad (5.15)$$

where $C_j(ik)$ is the j^{th} cumulated probability for the i^{th} age group and k^{th} education group. As well, $a_i = A_i - \bar{A}$, where A_i is the midpoint of the i^{th} age interval, and e_k is the effect of the k^{th} education group ($\sum e_k = 0$), ignoring the order of the education categories. Table 3 contains the survey estimates of the cumulated proportions. Table 4 contains the test statistics X^2, X_c^2 and X_S^2 for testing the goodness of fit of (5.15) and also for testing the nested hypothesis of no education effect, $e_k = 0$ for $k = 1,2$.

First, considering the goodness of fit of (5.15), if the survey design is ignored and the value of X^2 is referred to $\chi^2_{0.05}(13) = 22.4$, the upper 5% point of χ^2 with $IJ - 5 = 13$ d.f., we would reject the model. On the other hand, the value of X_c^2 or the value of X_S^2 when adjusted to refer to $\chi^2_{0.05}(13)$, is not significant at the 5% level, indicating that the model provides a good fit to the data.

For testing of the nested hypothesis, the value of X_c^2, or the value of X_S^2 when adjusted to refer to $\chi^2_{0.05}(2) = 5.99$ is not significant at the 5% level, indicating that the nested hypothesis of no education effect is tenable.

6. SOFTWARE

Implementation of the methodology of the previous sections requires two stages of computation — calculation of a vector of proportions, along with its estimated covariance matrix, and then calculation of model estimates, test statistics and their adjustments.

Surveys like the Canada Health Survey and the Labour Force Survey, from which examples have been presented, have complex designs and large data bases. Because of these two factors, calculation of covariance matrices was done on a mainframe computer. Custom SAS and Fortran programs were used for this purpose.

Computations required for the fitting and testing of goodness-of-fit models and subhypotheses were done either on the mainframe computer using SAS (and the MATRIX procedure in particular), or on a microcomputer using the GAUSS programming package. These programs are available to other analysts at Statistics Canada.

REFERENCES

AMEMIYA, T. (1985). *Advanced Econometrics.*Cambridge, Massachusetts: Harvard University Press.

BEDRICK, E.J. (1983). Adjusted goodness-of-fit tests for survey data. *Biometrika*, 70, 591-595.

BINDER, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys, *International Statistical Review*, 51, 279-292.

BOX, G.E.P., and COX, D.R. (1964). An analysis of transformations (with discussion). *Journal of the Royal Statistical Society*, Series B, 26, 211-252.

BOX, G.E.P., and COX, D.R. (1982). An analysis of transformations revisited, rebutted. *Journal of the American Statistical Association*, 77, 209-210.

FAY, R.E. (1985). A jack-knifed chi-squared test for complex samples. *Journal of the American Statistical Association*, 80, 148-157.

FULLER, W.A. (1975). Regression analysis for sample survey. *Sankhyā*, Series C, 37, 117-132.

FULLER, W.A. (1986). Estimators of the factor model for survey data. In *Advances in the Statistical Sciences*, Vol. I (Eds. MacNeill, I.B. and Umphrey, G.J.). Derdrecht, Holland: Reidell Publishing Co., 265-284.

GROSS, W.F. (1984). A note on chi-squared tests with survey data. *Journal of the Royal Statistical Society*, Series B, 46, 270-272.

GUERRERO, V.M., and JOHNSON, R.A. (1982). Use of the Box-Cox transformation with binary response models. *Biometrika*, 69, 309-314.

HABERMAN, S.J. (1982). Analysis of dispersion of multinomial responses. *Journal of the American Statistical Association*, 77, 568-580.

HIDIROGLOU, M.A., and RAO, J.N.K. (1987). Chi-squared tests with categorical data from complex surveys, Parts I and II. *Journal of Official Statistics*, 3, 117-132 and 133-140.

HINKLEY, D.V., and RUNGER, G. (1984), The analysis of transformed data. *Journal of the American Statistical Association*, 79, 302-309.

KISH, L., and FRANKEL, M.R. (1974). Inference from complex samples (with discussion). *Journal of the Royal Statistical Society*, Series B, 36, 1-37.

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.

KUMAR, S., and RAO, J.N.K. (1985). Fitting Box-Cox transformation models to labour force survey data. Unpublished Report, Social Surveys Methods Division, Statistics Canada, Ottawa.

McCULLAGH, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society*, Series B, 42, 109-142.

NATHAN, G., and HOLT, D. (1980). The effect of survey design on regression analysis. *Journal of the American Statistical Association*, 76, 681-689.

RAO, J.N.K., and SCOTT, A.J. (1984). On chi-squared tests for multi-way tables with cell proportions estimated from survey data. *Annals of Statistics*, 12, 46-60.

RAO, J.N.K., and SCOTT, A.J. (1987). On simple adjustments to chi-square tests with sample survey data. *Annals of Statistics*, 15, 385-397.

ROBERTS, G. (1985). *Contributions to Chi-Squared Tests with Survey Data*. Unpublished Ph.D. Thesis, Carleton University, Department of Mathematics and Statistics, Ottawa.

ROBERTS, G., RAO, J.N.K., and KUMAR, S. (1987). Logistic regression analysis of sample survey data. *Biometrika*, 74, 1-12.

SCOTT, A.J. (1986). Logistic regression analysis with survey data. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 25-30.

SCOTT, A.J., and HOLT, D. (1982). The effect of two-stage sampling on ordinary least squares methods. *Journal of the American Statistical Association*, 77, 848-854.

SCOTT, A.J., RAO, J.N.K., and THOMAS, D.R. (1989). Weighted least squares and quasi maximum likelihood estimation for categorical data under generalized linear models. *Linear Algebra and its Applications*, second special issue on Linear Algebra and Statistics, in press.

SINGH, A.C. (1985). On optimal asymptotic tests for analysis of categorical data from sample surveys. Statistics Canada Working Paper No. SSMD 86-002.

SINGH, A.C., and KUMAR, S. (1986). Categorical data analysis for complex surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 252-257.

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.

STATISTICS CANADA (1977). *Methodology of the Canadian Labour Force Survey*, 1976. Catalogue 71-526 occasional. Ottawa: Statistics Canada.

THOMAS, D.R., and RAO, J.N.K. (1987). Small-sample comparisons of level and power for simple goodness-of-fit statistics under cluster sampling. *Journal of the American Statistical Association*, 82, 630-636.

COMMENT

ROBERT E. FAY[1]

The authors have made an excellent contribution to the literature on the analysis of data from complex samples. By examining in turn four different models for categorical data: i) a log-linear model for a cross-classification; ii) a modification of the approach of Box and Cox to the transformation of binary data; iii) a problem of inference about parameters of a logistic regression model; and iv) a polytomous response model, the authors present solutions to important individual problems and illustrate the ways in which these flexible approaches to inference can be extended to other models for categorical data from complex samples. The applications are connected by an underlying theory, much of it previously appearing in Rao and Scott (1984), but this paper usefully presents in greater detail the implications of the general theory for specific models.

An omission from the paper is understandable but worth noting: for each model illustrated in the paper, replication provides an alternative strategy that, at times, may also be more convenient. In particular, the replication theory is complete for each of the applications, i), ii), and iv), to cross-classified data. In each case, tests of overall fit and comparisons of nested models can be assessed with the jackknifed chi-square test (Fay 1985) and standard errors for the parameters obtained through replication.

Replication also can provide standard errors and covariances for parameters of logistic regression models, as in iii), enabling in some cases a Wald-type test for equality of sets of regression parameters. It also appears likely that the jackknifing approach extends to the likelihood-ratio chi-square test in such situations involving continuous variables, although a firm proof of this conjecture is clearly required before application can be recommended. My point in calling attention to replication as a competing strategy for the problems presented in the paper is not to imply that it represents a methodologically superior approach to the methods of Rao and Scott (1984); instead, the availability of this methodology provides an additional choice to solve these and similar problems of inference. For example, the focus on replication for the estimation of variances from the current demographic surveys at the U.S. Census Bureau provides the potential to carry out analyses such as those presented in the paper.

I also want to point out that the methods presented and the analogues from replication theory have a potential importance beyond the realm of design-based inference from complex sample surveys, which is the focus of the paper. One of these involves the use of multiple imputation or related approaches intended to represent the uncertainty due to missing data. The implied interpretation of variance within the domain of design-based inference can be extended to include uncertainty from missing data without requiring changes to the methodology presented in the paper. The general methodology may also be applicable to some problems of inference from complex designed experiments, in which the design poses problems of clustering or stratification similar to complex sample surveys.

Of the four models discussed, however, I suggest that the Box and Cox transformation not be applied without consideration of alternative strategies, such as transformation of the x-variables instead. My own inclination would be to favor an analysis on a logistic scale, with possibly transformed predictors, unless the adaptation of the Box and Cox transformation obtains some distinct advantage, such as offering an additive model on the transformed scale in an instance where the logistic model does not provide as successful a fit without interaction terms.

I am delighted to have the opportunity to commend the authors on a useful and instructive paper.

[1] Robert E. Fay, U.S. Bureau of the Census, Washington, D.C. 20233.

COMMENT

C.J. SKINNER[1]

This paper provides an excellent discussion of a variety of applications of weighted least squares (WLS) and pseudo maximum likelihood (PML) procedures to categorical data. Its clear presentation and use of real survey examples will, I hope, help to encourage survey analysts to take account of complex designs in their analyses. As the authors indicate, analytical statistical procedures which take account of complex designs have been developed extensively in recent years (see *e.g.* Skinner, Holt and Smith 1989) and are even beginning to be referred to in standard computer software (*e.g.* SAS 1985, pp 61-67).

Commenting first on some specific aspects of the paper, I found Section 5 on polytomous variables to be especially valuable, given the wide occurrence of such data in surveys. A property of ordinal variables is that they may often be expected to possess monotonic relationships and so, for example, lack of monotonicity between the fitted values of $C_{1(ik)}$ (or $C_{2(ik)}$) and the education variable k in Table 3 makes the result of the corrected tests, that there is no evidence of an education effect, more plausible than the result of the uncorrected test.

The discussion of testing equality of two logistic regression models in Section 4 also seemed to me to be practically useful, although it would still seem to be possible theoretically to formulate this test as one of a nested hypothesis within the framework of Roberts, Rao and Kumar (1987).

Section 3 provides a useful illustration of how PML may be applied to general parametric models for categorical data. It is, however, gratifying that the more complex transformation model provides no significant improvement in fit over the logistic regression model, since the interpretation of the parameters of the transformation model is more difficult. For example, for the logistic model the coefficient for education may be interpreted as implying that the odds of being employed are increased by 16% for each additional year of education for males of a given age ($\exp(.1509) = 1.16$), whereas this interpretation is not generally available for the transformation model when $\lambda \neq 0$.

On a more general note I would be interested in the authors' views on the relative merits of WLS and PML. In the paper, these methods are presented quite separately, although both procedures would seem to be potentially applicable to a very wide class of models for categorical data under complex designs. Indeed both procedures are also applicable to models with continuous variables (Skinner, Holt and Smith 1989, Chapter 3); WLS requires just a statistic consistent for a known function of the parameters together with a consistent estimate of the covariance matrix of the statistic (Fuller 1984, Corollary 2), whereas PML is applicable very widely as described in Binder (1983). As a basis for discussion I list below a number of criteria on which WLS and PML might be compared; M1-M3 are relevant even under multinominal sampling, C1-C3 are specific to complex designs.

M1 **Flexibility** WLS may be more adaptable than PML for complex problems *e.g.* involving structural zeros.

M2 **Computation** WLS computation tends to have a more standard form.

M3 **Small cell counts** WLS is more sensitive to small counts, especially zeros.

C1 **Adaptability of multinomial methods to complex designs** WLS seems more easily adaptable.

[1] C.J. Skinner, University of Southampton, United Kingdom.

C2 **Efficiency** Under multinominal sampling WLS is usually asymptotically equivalent to PML (which is then just standard ML). It might be conjectured that WLS will always be at least as efficient as PML under complex designs, although this presupposes a 1-1 correspondence between WLS and PML estimation problems. If WLS is more efficient, is the gain usually negligible (*cf.* Scott and Holt 1982)? Are there general results here?

C3 **Degrees of freedom** WLS estimators and associated Wald tests may be unstable if the degrees of freedom used to estimate V_p are low.

ADDITIONAL REFERENCES

FULLER, W.A. (1984). Least squares and related analyses for complex survey designs. *Survey Methodology* 10, 97-118.

SAS Institute Inc. (1985). *SAS/IML User's Guide, Version 5 Edition*. Cary NC: SAS Institute Inc.

SKINNER, C.J., HOLT, D., and SMITH, T.M.F., Eds. (1989). *Analysis of Complex Surveys*. Chichester: Wiley.

COMMENT

E.A. MOLINA[1]

I would like to congratulate the authors on bringing together some recent methods developed for analyzing categorical data arising from sample surveys. The paper should be extremely useful for survey analysts who wish to take into account the impact of survey designs on the practical aspects of the analysis of survey data. In particular, it is important to emphasize that the methods discussed cover two different situations arising in practice: so called *primary analyses*, in which the researcher has all the relevant information at hand, and *secondary analyses*, in which the data provided do not include enough information about the population units to enable the calculation of full covariance matrices of the sample estimators.

The methods covered require the existence of a structural model for the data. There are situations, however, in which it is difficult to specify a single structural model that adequately describes categorical data. In large scale surveys there is often need to screen out many cross classifications at minimal cost. In such cases the use of measures of association is a common alternative. These non parametric methods were extended to sample survey data by Molina and Smith (1986, 1988).

For the primary analysis of survey data the paper concentrates on weighted least squares and Wald tests. The results in Scott, Rao and Thomas (1989) are summarized and the relationship with quasi-likelihood is mentioned. I think that an important conclusion from that paper should be included in this section, namely the need to take into account the survey constraints $K'p(X\beta) = \pi$ when using quasi-likelihood methods. The reader may not be aware of the importance of the careful choice of the g-inverse in equation (2.9). Quasi-likelihood methods are now widely used and the relationship with weighted least squares methods is a relevant one. In fact, quasi-likelihood functions represent an interesting alternative for the analysis of survey data. However, there are practical problems since the method requires that we specify the covariance matrix as a function of p, the variance function. Quasi-likelihoods are largely determined by these variance functions (see, *e.g.*, Morris 1982, and Jørgensen 1987). If a matrix of estimates is given instead of a function, the method would be equivalent to the use of a normal distribution.

Most of the paper is devoted to methods involving *pseudo likelihoods*. Since secondary analyses constitute the most common situation in practice, the methods presented are likely to be extensively used by survey analysts. I would like, however, to discuss some alternatives.

The study of the impact of survey design on Guerrero and Johnson's (1982) transformation models is an important addition to the literature. However, Nelder and Pregibon (1987) have proposed a family of functions, the *extended quasi-likelihoods*, that avoid some important disadvantages of transformation models and can be fitted with GLIM. If design effects are available, their methods can be adapted to survey data by incorporating them either in the variance functions or in the form of weights. Alternatively, design variables may be used to adjust the dispersion parameter in the models. In both cases, one advantage is that we can use the goodness of fit statistics and standard errors produced by GLIM under these models to examine the data without the introduction of further corrections.

These comments apply in general to the use of pseudo-likelihoods. The effect of ignoring the survey design may be treated as an increase or decrease in the expected variability that may be modelled as overdispersion or underdispersion by means of quasi-likelihoods or extended quasi-likelihoods. See, *e.g.*, Pocock *et al.* (1981), Breslow (1984), Williams (1982), among

[1] E.A. Molina, Universidad Simon Bolivar, Caracas and University of Southampton, United Kingdom.

others. As an example, I reanalyzed the data in Table 1. The analysis given in the paper is the correct one, since it incorporates the true covariance matrix. Suppose, however, that this matrix is not available and that only the cell design effects are at hand. Using GLIM I fitted model (2.12) with a Poisson error ignoring the sampling scheme. This gives $X^2 = 5.68$, $G^2 = 5.67$. The Rao and Scott (1987) approximation for the chi square statistic gives $X^2(\delta) = 5.68/2.25 = 2.52$. For the independence model the uncorrected values are $X^2 = 18.22$, $G^2 = 18.22$, and the correction gives $X^2(\delta) = 18.22/1.65 = 11.04$. What can be done if the deffs are not available?. A simple quasi-likelihood approach to overdispersion is to estimate the mean deviance for the larger model, $D = 5.68/3 = 1.89$, and to use the inverse of this value as a weight (or as a new scale parameter). This give $X^2 = 3.01$ for model (2.12) and $X^2 = 9.65$ for the independence model. The correct approach here is to use the excess in deviance (the difference between the log-likelihood ratio statistics) to test $\gamma = 0$, since G^2 will equate the degrees of freedom for the larger model. The value is 6.65, which is just significant at the 1% level. Both analyses are in agreement with the correct analysis given in the paper, but in other situations it may not be so. The quasi-likelihood model presented here is equivalent to assuming that the actual covariance matrix is a multiple of the one obtained under multinomial sampling, a model that may perform badly in several situations. The advantage is that it can be used when the only information available is that given by the variability inherent in the data, and the analysis performed in a standard statistical package like GLIM. If the deffs are available, other models involving them may be proposed, and a paper is in preparation.

There is, however, no completely satisfactory substitute for an analysis involving the actual covariance matrix. The objective of this contribution is to highlight other possibilities when the full covariance matrix is not known. Quasi-likelihoods offer a fertile ground for further exploration, particularly in relation to survey data. The paper under discussion presents several alternatives and is an important contribution to the field.

ADDITIONAL REFERENCES

JØRGENSEN, B. (1987). Exponential dispersion models. *Journal of the Royal Statistical Society* B 127-162.

MOLINA, E.A., and SMITH, T.M.F. (1986). The effect of sample design on the comparison of associations. *Biometrika* 73, 23-33.

MOLINA, E.A., and SMITH, T.M.F. (1988). The effect of sampling on operative measures of association. *International Statistical Review* 56, 235-242.

MORRIS, C.N. (1982). Natural exponential families with quadratic variance functions. *Annals of Statistics* 10, 65-80.

NELDER, J.A., and PREGIBON, D. (1987). An extended quasi-likelihood function. *Biometrika* 74, 221-232.

POCOCK, S.J., COOK, D.G., and BERESFORD, S.A.A. (1981). Regression of area mortality rates on explanatory variables: What weighting is appropriate? *Applied Statistics* 31, 286-295.

WILLIAMS, D.A. (1982). Extra binomial variation in logistic-linear models. *Applied Statistics* 31, 144-148.

RESPONSE FROM THE AUTHORS

We thank the three discussants, Fay, Molina and Skinner, for their useful comments and for suggesting additional methods useful in the analysis of cross-classified data from complex sample surveys.

(i) Response to comments of R.E. Fay

We agree with Fay that replication methodology and associated jackknife chi-squared tests provide viable alternatives to the methods presented here, provided the survey design permits the use of a replication method such as the jackknife or the balanced half-sample replication. His CPLX program indeed offers a comprehensive analysis option whenever estimates are available at the individual replicate level. Also, as noted in the Introduction, Fay's jackknife tests and Rao-Scott corrections have performed well under quite general conditions in simulation studies, unlike the Wald tests based on weighted least squares. Rao-Scott corrections are, however, also applicable to survey designs not permitting the use of a replication method.

The software systems for the Canada Health Survey and the Canadian Labour Force Survey were set up to readily provide the estimated covariance matrix of cell estimates but not the replicate level estimates. As a result, the implementation of jackknife tests would have required some changes in the software systems.

We are also thankful to Fay for pointing out that the methods presented here, and the analogues from replication theory, can also handle some problems of inference from complex designed experiments involving clustering and stratification. Indeed, one of us (J.N.K. Rao) recently used Rao-Scott type methods to fit dose-response models and to test hypotheses in teratological studies involving animal litters as experimental units (Rao and Colin 1989). These methods do not assume specific models for the intra-litter correlations, unlike other methods proposed in this area.

We considered Box-Cox transformation models since Guerrero and Johnson (1982) obtained significantly better fits on some Mexican data compared to the logit model. We agree with Fay, however, that the Box-Cox models should not be applied without consideration of alternative strategies, such as transforming the predictors. As noted by Fay, the Box-Cox approach would be useful in these cases where it would lead to additive models on the transformed scale while the logit model would require interaction terms.

(ii) Response to comments of E.A. Molina

Molina is correct in saying that measures of association can be used to screen out many cross classifications at minimal cost. His joint work with T.M.F. Smith on extending the classical theory for measures of association to sample survey data involving clustering and stratification is an important contribution.

As noted in the Introduction, we assumed throughout the paper that the user has access to a full estimated covariance matrix of cell estimates. However, such detailed information is often not available for secondary analyses, and in fact even cell deffs may not be available, as pointed out by Molina. In the latter case, Rao and Scott (1987) showed that an F statistic used in GLIM for testing a nested hypothesis, such as $\gamma = 0$ given the model (2.12), is asymptotically valid whenever the covariance matrix of cell estimates, $\hat{V}$, is proportional to the multinominal covariance matrix, $\hat{P}$. The F-test, however, is less powerful than the Rao-Scott tests, unless the denominator degrees of freedom are high. In the latter case, the F test might work well even if the condition $\hat{V} \propto \hat{P}$ is not satisfied (see Rao and Scott 1987, p. 392).

For the data in Table 1, $F = 6.63$ for testing $\gamma = 0$ given the model (2.12), which is not significant at the 5% level compared to $F_{1,3}(0.05) = 10.01$, the upper 1% of the F distribution with 1 and 3 degrees of freedom (d.f.). On the other hand, the Wald test W_1 and the Rao-Scott test, both requiring detailed information on the estimated covariance matrix, are significant at the 1% level compared to $\chi_1^2(0.01) = 6.63$. The F-test, therefore, appears to be less powerful here since the denominator d.f. is only 3. Molina's proposed test is, in fact, equal to F, but he was treating F as a χ^2 variable with 1 d.f. which may not be valid due to small denominator d.f.

The GLIM method does not provide a statistic for testing the goodness-of-fit of a model. Some information on the design effects is necessary for getting a valid test of goodness-of-fit.

(iii) Response to comments of C.J. Skinner

Skinner noted that the test of equality of two logistic regression models in Section 4 might be formulated as a test of a nested hypothesis within the framework of Roberts, Rao and Kumar (1987), using dummy x-variables. The framework of Roberts, Rao and Kumar, however, assumes one fixed sample size n whereas in Section 4 we have two fixed sample sizes n_1 and n_2 for the two time periods. As a result, their results would need careful modification in order to be applicable to the present case of test of equality of two logistic regression models. Moreover, the dummy variable approach would involve the determination of estimates of $2s$ parameters iteratively, whereas the approach in Section 4 requires two iterative solutions, each involving only s parameters. Thus, the dummy variable approach could lead to convergence problems if s is not small.

We treated WLS with singular covariance matrices separately in Section 2 since the logit-type models in the remaining sections do not involve singular covariance matrices. WLS can also be applied to logit-type models but the resulting estimators and associated Wald tests may be unstable if the degrees of freedom associated with the estimated covariance matrix, $\hat{V}_P$, are low (criterion C3 of Skinner). The six criteria proposed by Skinner for comparing WLS and PML are very useful. We prefer PML mainly on the basis of criterion C3. Regarding the relative efficiency of WLS and PML estimators under complex designs, no general results are available, but WLS estimators are not likely to be significantly more efficient (and in fact, may be less efficient) if the degrees of freedom associated with the estimated covariance matrix are low. Clearly, further research on the relative efficiency of WLS and PML estimators would be useful.

ADDITIONAL REFERENCES

RAO, J.N.K., and COLIN, D. (1988). Fitting dose-response models and hypothesis testing in teratological studies. Technical Report No. 116, Laboratory for Research in Statistics and Probability, Carleton University and University of Ottawa, Ottawa, Ontario.

Survey Methodology, December 1989
Vol. 15, No. 2, pp. 187–201
Statistics Canada

Simultaneous Confidence Intervals for Proportions Under Cluster Sampling

D. ROLAND THOMAS[1]

ABSTRACT

The paper describes a Monte Carlo study of simultaneous confidence interval procedures for $k > 2$ proportions, under a model of two-stage cluster sampling. The procedures investigated include: (i) standard multinomial intervals; (ii) Scheffé intervals based on sample estimates of the variances of cell proportions; (iii) Quesenberry-Hurst intervals adapted for clustered data using Rao and Scott's first and second order adjustments to X^2; (iv) simple Bonferroni intervals; (v) Bonferroni intervals based on transformations of the estimated proportions; (vi) Bonferroni intervals computed using the critical points of Student's t. In several realistic situations, actual coverage rates of the multinomial procedures were found to be seriously depressed compared to the nominal rate. The best performing intervals, from the point of view of coverage rates and coverage symmetry (an extension of an idea due to Jennings), were the t-based Bonferroni intervals derived using log and logit transformations. Of the Scheffé-like procedures, the best performance was provided by Quesenberry-Hurst intervals in combination with first-order Rao-Scott adjustments.

KEY WORDS: Simultaneous inference; Complex surveys; Monte Carlo.

1. INTRODUCTION

Survey results are often presented as estimated proportions (or percentages) of population units belonging to two or more distinct categories. Examples include many sociological studies (see for example Black and Myles 1986), marketing studies and opinion polls. As noted by Fitzpatrick and Scott (1987), inference on category proportions is often based on single binomial confidence intervals, even when more than two category proportions are being examined. This paper describes a study of several procedures for constructing simultaneous confidence intervals for the proportions π_i, $i = 1, \ldots, k$, of population units belonging to each of k distinct categories, using data from a two-stage cluster sample. Standard simultaneous confidence interval procedures for categorical data problems, reviewed by Hochberg and Tamane (1987), are based on the assumption of multinomially distributed sample counts, and are thus appropriate for data from simple random samples. When the data have been collected using sample survey designs that involve clustering, standard procedures are likely to perform poorly, as is the case when standard multinomial based tests are applied to data from complex sample surveys. In the latter case, it has been shown by many workers that clustering can lead to unacceptably high Type I error rates (see, for example, Fellegi 1980; Rao and Scott 1979, 1981; Holt, Scott and Ewing 1980). For simultaneous confidence intervals, therefore, it is natural to expect that clustering will lead to coverage probabilities that are lower than multinomial theory indicates.

Estimation of simultaneous confidence intervals (SCI's) is an important adjunct to hypothesis testing. The present study thus represents a natural follow-up to Thomas and Rao's (1987) investigation of test statistics for the simple goodness of fit problem, under

[1] D. Roland Thomas, School of Business, Carleton University, Ottawa, Ontario, K1S 5B6.

simulated cluster sampling. In this paper, adaptations of the standard SCI procedures are proposed, and their performance in small samples is evaluated using Monte Carlo techniques.

The cluster sampling model that is used in the Monte Carlo study is described in Section 2, and the SCI procedures to be examined are presented in Section 3. In Section 4, the design of the Monte Carlo experiment is described, together with procedures for evaluating confidence interval performance. The main results of the study are presented in Sections 5 through 7, followed in Section 8 by some final conclusions and recommendations.

2. THE CLUSTER SAMPLING MODEL

This investigation will focus on two-stage sampling in which a k-category sample of m units is drawn independently from each of r sampled clusters.

For a sample of size $n = mr$, let $\boldsymbol{m} = (m_1, \ldots, m_{k-1})'$ represent the category counts for the whole sample, where $m_k = n - \sum_{i=1}^{k-1} m_i$. In terms of proportions, let $\hat{\boldsymbol{\pi}} = (\hat{\pi}_1, \ldots, \hat{\pi}_{k-1})' = \boldsymbol{m}/n$ be the vector of category proportions for the full sample. Further, define $\boldsymbol{\pi} = \mathrm{E}(\hat{\boldsymbol{\pi}})$, where E denotes expectation under a suitable model of cluster sampling, and let $\boldsymbol{V}/n$ represent the $(k - 1) \times (k - 1)$ covariance matrix of $\hat{\boldsymbol{\pi}}$. Following Rao and Scott (1981), the ordinary design effect for the linear combination $\boldsymbol{c}'\hat{\boldsymbol{\pi}}$ of category proportions is $\boldsymbol{c}'\boldsymbol{V}\boldsymbol{c}/\boldsymbol{c}'\boldsymbol{P}\boldsymbol{c}$, where $\boldsymbol{P}$ is n times the covariance matrix of $\hat{\boldsymbol{\pi}}$ under multinomial sampling, *i.e.*, $\boldsymbol{P} = \mathrm{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}'$, and $\boldsymbol{c}$ is a vector of dimension $k - 1$. The largest design effect taken over all possible linear combinations is given by the largest eigenvalue of the design effect matrix $\boldsymbol{D} = \boldsymbol{P}^{-1}\boldsymbol{V}$. The eigenvalues of $\boldsymbol{D}$, denoted in decreasing order by $\lambda_1, \lambda_2, \ldots, \lambda_{k-1}$, were termed generalized design effects by Rao and Scott (1981), and provide a quantitative summary of the variance inflation associated with a particular design, relative to simple random sampling. Under the multinomial distribution, corresponding to simple random sampling from large populations, $\lambda_j = 1 \ \forall j$. Designs involving clustering usually yield generalized design effects greater than one on the average, *i.e.*, $\bar{\lambda} = \sum_{j=1}^{k-1} \lambda_j/(k - 1) > 1$. Furthermore, studies of real survey data (Hidiroglou and Rao 1987; Rao and Thomas 1988) reveal significant variation among the λ_j's. This is conveniently represented by their coefficient of variation, given by

$$a = \left(\sum_{j=1}^{k-1} \lambda_j^2/[(k - 1)\bar{\lambda}^2] - 1 \right)^{1/2}. \tag{1}$$

A suitable model of cluster sampling must therefore be capable of generating generalized design effects such that $\bar{\lambda} > 1$ and $a > 0$.

Brier (1981) proposed a model of two-stage cluster sampling in which individual clusters are represented by vectors of category probabilities, $\boldsymbol{p}_\ell = (p_{\ell 1}, p_{\ell 2}, \ldots, p_{\ell, k-1})'$, $\ell = 1, \ldots, r$, where for each cluster, $p_{\ell k} = 1 - \sum_{i=1}^{k-1} p_{\ell i}$. Each $\boldsymbol{p}_\ell$ was independently drawn from a Dirichlet distribution with mean $\boldsymbol{\pi}$, *i.e.* $\mathrm{E}(\boldsymbol{p}_\ell) = \boldsymbol{\pi}$, and second stage sampling of the m units per cluster was multinomial, conditional on the realized value of $\boldsymbol{p}_\ell$ for that cluster. Let the vector of counts for each cluster be $\boldsymbol{m}_\ell = (m_{\ell 1}, \ldots m_{\ell, k-1})$, *with* $m_{\ell k} = m - \sum_{i=1}^{k-1} m_{\ell i}$. Thus for the full sample, $\boldsymbol{m} = \sum_{\ell=1}^{r} \boldsymbol{m}_\ell$, and in terms of proportions, $\hat{\boldsymbol{\pi}} = \sum_{\ell=1}^{r} \hat{\boldsymbol{\pi}}_\ell$, where $\hat{\boldsymbol{\pi}}_\ell = \boldsymbol{m}_\ell/m$. Brier (1981) showed that under this model, $\mathrm{E}(\hat{\boldsymbol{\pi}}) = \boldsymbol{\pi}$ and $\boldsymbol{V}(\hat{\boldsymbol{\pi}}) = d\boldsymbol{P}/n$, *i.e.*, the covariance matrix of $\hat{\boldsymbol{\pi}}$ is proportional to the multinomial covariance matrix, with the constant of proportionality $d > 1$. Under this model, the design effect matrix is given by $\boldsymbol{D} = d\boldsymbol{I}_{k-1}$, where $\boldsymbol{I}_{k-1}$ is the identity matrix of order $k - 1$. Thus $\lambda_i = d \ \forall i$, so that $\bar{\lambda} = d$ and $a = 0$. Brier's model can therefore represent variance inflation $(\bar{\lambda} > 1)$, but cannot represent the unequal generalized design effects encountered in practice. Thomas and Rao (1987) used an extension of Brier's model in which the first stage p_ℓ's are sampled independently from a mixture of two Dirichlet distributions, representing a population composed of two distinct classes of clusters. This model, which is a special case of that proposed by Rao and Scott (1979), generates one distinct and $k - 2$ equal eigenvalues, with $\bar{\lambda}$ and a being explicit functions of the Dirichlet parameters. This greatly facilitates the design of the Monte Carlo study by allowing for convenient control of the values of the clustering measures $\bar{\lambda}$ and a. Since it satisfies the basic requirements outlined above ($\bar{\lambda} > 1, a > 0$), Thomas and Rao's (1987) model will be used in this study.

3. SIMULTANEOUS CONFIDENCE INTERVAL PROCEDURES

3.1 Scheffé Intervals

A standard Scheffé argument, based on the asymptotically exact probability statement

$$P\left(n(\hat{\pi} - \pi)' \hat{V}^{-1} (\hat{\pi} - \pi) \leq \chi^2_{k-1}(\alpha)\right) = 1 - \alpha \qquad (2)$$

leads to simultaneous confidence intervals for linear combinations, $\ell'\pi$, of the category probabilities, where ℓ is a vector of dimension $(k - 1)$. Appropriate choices of ℓ then yield SCI's on the individual cell probabilities given by

$$\pi_i \in \left\{\hat{\pi}_i \pm (\hat{v}_{ii})^{1/2} (A/n)^{1/2}\right\}, i = 1, \ldots, k, \qquad (3)$$

where $A = \chi^2_{k-1}(\alpha)$ is the upper α percent point of a chi-squared distribution on $k - 1$ degrees of freedom, and $\hat{v}_{ii}$ is the i^{th} diagonal element of a consistent estimator of V (as $r \rightarrow \infty$) given by

$$\hat{V} = \frac{n}{r(r - 1)} \sum_{l=1}^{r} (\hat{\pi}_\ell - \hat{\pi}) (\hat{\pi}_\ell - \hat{\pi})'. \qquad (4)$$

Note that when the endpoint of an interval lies outside [0, 1], definition (3) must be modified by truncating the endpoint to 0 or 1 as appropriate. For multinomial sampling, $\hat{v}_{ii}$ can be replaced by $\hat{\pi}_i(1 - \hat{\pi}_i)$, in which case the Scheffé intervals reduce to those proposed by Gold (1963). The latter will be referred to as Scheffé-Gold intervals. The Scheffé intervals of equation (3) will be conservative, *i.e.*, will have coverage exceeding $(1 - \alpha)$ asymptotically since they make use of only a finite number of the available ℓ directions (see Miller 1981, page 63). In fact, they will become very conservative as k increases, as can be shown using the following argument due to Goodman (1965). The coverage of the Scheffé intervals is equal to one minus the probability of occurrence of at least one of the events $\{(\hat{\pi}_i - \pi_i)^2/(\hat{v}_{ii}/n) > \chi^2_{(k-1)}(\alpha)\}$, $i = 1, \ldots, k$; since the random variables $(\hat{\pi}_i - \pi_i)^2/(\hat{v}_{ii}/n)$ each have chi-squared distributions on one degree of freedom asymptotically, the probability of each individual event can be evaluated. Using the Bonferroni inequality, lower bounds for the coverage can then be obtained; for a nominal coverage of 95% with $k = 3, 5, 8$ and 12, these bounds are .9571, .9896, .9986 and .9999 respectively.

3.2 Modified Quesenberry-Hurst Intervals

Under the assumption of multinominal sampling, Quesenberry and Hurst (1964) solved the large sample probability statement

$$P\left\{X^2 = n \sum_{i=1}^{k} \frac{(\hat{\pi}_i - \pi_i)^2}{\pi_i} \leq A\right\} = 1 - a \tag{5}$$

for the cell probabilities π_i, to get the SCI's

$$\pi_i \in \left\{\frac{\hat{\pi} + A/2n \pm (A/n)^{1/2} [\hat{\pi}_i (1 - \hat{\pi}_i) + A/4n]^{1/2}}{1 + A/2n}\right\}. \tag{6}$$

Under multinomial sampling, these intervals are asymptotically equivalent to Scheffé and Scheffé-Gold intervals, and will therefore exhibit similar asymptotic conservativeness.

Quesenberry-Hurst (Q-H) intervals can be adapted for use with clustered survey data using the first and second order corrections to the distribution of X^2 proposed by Rao and Scott (1981). Corresponding first and second order SCI's are obtained by replacing A in equation (3) by

$$A^{(1)} = \hat{\bar{\lambda}} A \text{ and } A^{(2)} = \hat{\bar{\lambda}}(1 + \hat{a}^2)\, \chi_\nu^2\,(\alpha) \tag{7}$$

respectively, where $\nu = (k - 1)/(1 + \hat{a}^2)$ and $\hat{\bar{\lambda}}$, an estimate of the mean of the generalized design effects, is given by (Rao and Scott, 1981)

$$\hat{\bar{\lambda}} = (k - 1)^{-1} \sum_{i=1}^{k} (1 - \hat{\pi}_i)\, \hat{d}_i, \tag{8}$$

where $\hat{d}_i$, $i = \ldots, k$ is an estimated cell design effect given by $\hat{d}_i = \hat{v}_{ii}/\hat{\pi}_i (1 - \hat{\pi}_i)$. The coefficient of variation, a, is estimated by replacing $\bar{\lambda}$ in equation (1) by $\hat{\bar{\lambda}}$, and $\sum \lambda_i^2$ by the estimate $\sum \hat{\lambda}_i^2 = \sum\sum \hat{v}_{ij}^2/\hat{\pi}_i\, \hat{\pi}_j$. It turns out (see Thomas 1989) that the second order modified intervals are unnecessarily conservative, so that only the first-order modified Q-H intervals will be discussed in the remainder of the paper.

3.3 Simple Bonferroni Intervals

Since (loosely speaking) each $\hat{\pi}_i$ is asymptotically $N(\pi_i, v_{ii}/n)$, the intervals

$$\pi_i \in \left\{\hat{\pi}_i \pm (\hat{v}_{ii}/n)^{1/2} z_{\alpha'/2}\right\}, \tag{9}$$

will have large sample coverage at least $(1 - \alpha)$ by the Bonferroni inequality, where $\alpha' = \alpha/k$ and $z_{\alpha'/2}$ is the upper $\alpha'/2$ percent point of the standard normal distribution. Intervals (9) are equivalent to Scheffé intervals with A in equation (3) replaced by $A^{(3)} = \chi_1^2(\alpha')$. As noted by Goodman (1965), they will be shorter than Scheffé intervals for the usual values of α and k; *e.g.*, $\alpha = 1\%, 5\%$, or 10%. Goodman's (1965) multinomial Bonferroni intervals are given by equation (9) with $\hat{v}_{ii}$ replaced by $\hat{\pi}_i\ (1 - \hat{\pi}_i)$. All endpoints of simple Bonferroni intervals that lie outside $[0, 1]$ will be truncated to 0 or 1 as appropriate.

3.4 Transformed Bonferroni Intervals

For suitably smooth g, $g(\hat{\pi}_i)$ will be asymptotically $N(g(\pi_i), [\, g_i'(\pi_i)]^2\, v_{ii}/n)$, where $g_i'(\pi_i)$ denotes the partial derivative $\partial g(\pi_i)/\partial \pi_i$ evaluated at π_i. Bonferroni intervals can then be obtained by inverting corresponding intervals on the $g(\pi_i)$'s, giving

$$\pi_i \in \left\{ g^{-1}(g(\hat{\pi}_i)\ \pm\ g_i'\,(\hat{\pi}_i)\ \ (\hat{v}_{ii}/n)^{1/2}\, z_{\alpha'/2}) \right\}. \tag{10}$$

Three g functions will be investigated: the square root $g_1(\pi_i) = \pi_i^{1/2}$ (previously investigated by Bailey 1980, for the case of multinomial sampling); the natural logarithm $g_2(\pi_i) = ln(\pi_i)$; and the logit $g_3(\pi_i) = ln(\pi_i/(1 - \pi_i))$. Interval endpoints that lie outside $[0, 1]$ will again be truncated to 0 or 1 as necessary.

Transformed Bonferroni intervals based on a jackknifed estimator of the variance of $g(\hat{\pi})$ have also been examined (see Thomas 1989). It was found that there is little advantage to using jackknifed variance estimates; Taylor series variance estimates are therefore recommended for their simplicity. Intervals based on jackknife variance estimates will not be considered further in this paper.

3.5 Variants of the Above Intervals

Scheffé Intervals: Following Thomas and Rao (1987), Scheffé intervals can be modified by replacing the critical constant A in equation (3) by $A^{(4)} = (k - 1)\ (r - 1)\ (r - k + 1)^{-1}$ $F_{(k-1),\,(r-k+1)}\ (\alpha)$, where $F_{(k-1),\,(r-k+1)}\ (\alpha)$ is the upper α percent point of an F distribution on $(k - 1)$ and $(r - k + 1)$ degrees of freedom.

Quesenberry-Hurst Intervals: Variants of the modified Quesenberry-Hurst (Q-H) intervals can also be defined, corresponding to the F forms of the first and second order corrected test statistic proposed by Thomas and Rao (1987). These again turn out to be conservative, and will not be considered further.

Bonferroni Intervals: Heuristic arguments (see the appendix to Thomas and Rao 1987) suggest that the simple Bonferroni intervals can be improved by replacing $z_{\alpha'/2}$ in (9) by $t_{r-1}(\alpha'/2)$, the upper $\alpha'/2$ percentage point of Student's t distribution on $r - 1$ degrees of freedom. This strategy will also be applied to the transformed Bonferroni intervals.

4. THE DESIGN OF THE MONTE CARLO STUDY

4.1 Parameters and Random Numbers

The parameters to be controlled are: (i) the nominal coverage level $(1 - \alpha)$ of the SCI; (ii) π, the model probability vector; (iii) k, the number of categories; (iv) r, the number of sample clusters; (v) m, the number of units drawn from each sampled cluster; (vi) $\bar{\lambda}$, the mean of the generalized design effects (eigenvalues); (vii) a, the coefficient of variation of the generalized design effects. The nature and degree of clustering is represented by the pair $(\bar{\lambda}, a)$ as follows: (a) multinomial sampling $(\bar{\lambda} = 1, a = 0)$; (b) constant design effect clustering $(\bar{\lambda} > 1, a = 0)$; (c) non-constant design effect clustering $(\bar{\lambda} > 1, a > 0)$.

Individual Monte Carlo experiments were run for particular combinations of k, $\bar{\lambda}$, a and r_{max}, the latter being the maximum number of clusters generated in one computer run. Most experiments were run at two values of $\bar{\lambda}$, namely 1.5 and 2.0, two values of a, namely $a = 0$ (constant design effects) and $a > 0$ (one level of non-constant design effects), for equiprobable categories ($\pi_i = 1/k, i = 1, \ldots, k$). Three values of k ($k = 3, 5, 8$) were initially selected to cover the range of numbers of categories commonly encountered in goodness-of-fit tests. An additional run was subsequently done for the case $k = 12$, $\bar{\lambda} = 2$ and $a > 0$ to check on the range of applicability of the results. The number of units per cluster was set at $m = 10$ for $k = 3, 5$ and 8, and at $m = 20$ for $k = 12$. Preliminary investigations showed coverage rates to be insensitive to the value of this parameter. For comparability of results over k, the non-zero settings of a were selected to make a/a_{max} the same for each selected value of k, where $a_{max} = (k - 2)^{1/2}$ is the maximum possible value of a. For $k = 5$, the non-zero value of a was set at 0.5, which is typical of the values encountered in practice, *e.g.*, $\hat{a} = 0.43$ for $k = 5$, as reported by Rao and Thomas (1988).

The initial focus on equiprobable categories allowed for a cost effective assessment of the influence of k, $\bar{\lambda}$ and a on coverage rates, and eliminated many of the possible SCI variants from further consideration. Additional experiments reported in Section 7 show that the procedures that passed this initial screening can in fact be applied when the cell probabilities are markedly unequal. Vectors of unequal probabilities were confined to the class $\pi(k, \mathrm{q}, \phi)$, defined by the elements $\pi_i = \phi, i = 1, \ldots, q$ and $\pi_i = (1 - q\phi)/(k - q)$, $i = q + 1, \ldots, k$.

For details of the generation of the random clusters from the mixture Dirichlet multinomial distribution, the reader is referred to Thomas and Rao (1987). Each Monte Carlo experiment consisted of 1000 sets of up to 100 independent clusters, grouped into nested subsets. All SCI procedures were applied in turn to each subset, using two nominal coverage levels (95% and 90%), thus improving the precision of comparisons between procedures at the same parameter settings, and between the same SCI procedures for different numbers of clusters. Most of the results presented will be for 95% nominal coverage; trends for 90% coverage were found to be qualitatively similar.

4.2 Evaluation Procedures

The percentage of Monte Carlo trials for which at least one of the k confidence intervals fails to cover the true parameter value is reported, and used for a preliminary screening of the main SCI procedures. This is a measure of the family error rate, which is equivalent to the actual significance level of the SCI when the latter is viewed as a test of goodness-of-fit. The family error rate, which will be referred to in this paper as the total error rate ER_T, is used in place of the more commonly reported actual coverage rate (equal to one hundred percent minus the total error rate) because it can be conveniently split into two one-sided rates which will provide information on the symmetry or 'unbiasedness' of each SCI procedure. Jennings (1987) argued that coverage rates alone can provide a misleading assessment of single parameter confidence interval procedures, and recommended that the number of times that an interval falls above and below the true parameter value should be separately reported. In this paper, Jennings' suggestion has been adapted to simultaneous confidence intervals on π_i, $i \in I$, where I is the index set $\{1, \ldots, k\}$, by counting the number of Monte Carlo trials for which:

(a) more intervals fall above their corresponding π_i, $i \in I$, than fall below;

(b) more intervals fall below their corresponding π_i, $i \in I$, than fall above;

(c) the same number (> 0) of intervals fall above their corresponding π_i, $i \in I$, as fall below.

Upper and lower error rates are then defined as $ER_U = [n_a + (n_c/2)]/N_t$ and $ER_L = [n_b + (n_c/2)]/N_t$, respectively, where N_t represents the number of Monte Carlo trials, and n_a, n_b and n_c denote the counts (a) through (c), respectively. The sum of ER_U and ER_L is clearly equal to the total error rate, ER_T. These one-sided error rates will be used to compare SCI procedures whose overall error rates are acceptably close to the nominal rate α, over a range of parameter settings and cluster strengths. Average interval lengths and corresponding standard errors have also been computed, and will be used as final discriminators in the selection of the recommended procedures.

5. A SUMMARY OF RESULTS FOR TOTAL ERROR RATES

All results in this section are given in terms of the total error rate ER_T, defined in Section 4. For lack of space, tables are presented only for the case of unequal design effects, $(a > 0)$, with $\bar{\lambda} = 2$. More detailed results are given in Thomas (1989). In interpreting the tabulated results, it should be noted that for 1000 Monte Carlo trials, binomial standard errors of point estimates of true ER_T's having magnitudes 5%, 10% and 20% are 0.7%, 0.9% and 1.3% respectively. As a general rule deviations from nominal rates, and differences between the error rates of different SCI procedures will be noted only when they are large enough to have practical significance, and exceed their Monte Carlo standard errors by a factor of at least two.

5.1 Multinomial Procedures

Results for multinomial intervals will only be summarized here; for details see Thomas (1989). Under cluster sampling, error rates for Goodman's Bonferroni intervals (see equation (9) with $\hat{v}_{ii}$ replaced by $\hat{\pi}_i(1 - \hat{\pi}))$ are unacceptably high except for values of $\bar{\lambda}$ close to 1, *i.e.*, unless the effect of clustering is small. The Scheffé-Gold and multinomial Quesenberry-Hurst intervals, on the other hand, can yield error rates that are close to the nominal value in certain cases, whenever their inherent conservativeness balances the error inflating effects of clustering (see also Andrews and Birdsall 1988). Unfortunately, this is not always the case; both procedures can display inflated error rates $(ER_T \geq 2\alpha)$ for realistic combinations of category numbers and clustering strengths.

Multinominal procedures should therefore not be used with complex survey data. Procedures are clearly required that directly account for the clustering, and provide good coverage for the required number of categories, over a wide range of clustering conditions.

5.2 The Scheffé Procedures

Total error rates for the χ^2-based Scheffé procedure of equation (3) and its F-based variant are summarized in Table 1 as functions of r, for the case $\alpha = 5\%$, $\bar{\lambda} = 2$ and $a > 0$. More detailed graphs are given in Thomas (1989).

For the values of k studied, ER_T for the χ^2-based Scheffé procedure of equation (3) increases rapidly as the number of clusters decreases, so that it should never be used for small numbers of clusters. The F-based variant, on the other hand, keeps ER_T reasonably close to or below $\alpha = 5\%$ for all r. As r increases, ER_T for F-based Scheffé remains fairly constant for the case $k = 3$, but becomes increasingly conservative for $k \geq 5$, as does the χ^2 version. These empirical trends with varying r can be explained in terms of two competing effects. As r increases, error rates for both procedures approach their asymptotic levels which are bounded above by 4.29%, 1.04% and 0.14%, for $k = 3$, 5 and 8 respectively (see Section 3.1).

Table 1

Total Error Rates for Scheffé and Modified Q-H Intervals;
$\alpha = 5\%$, $\bar{\lambda} = 2$, $m = 10$

			Total Error Rate (ER_T)		
k	a	r	Scheffé (χ^2 based)	Scheffé (F based)	Modified Q-H (first order)
3	.29	15	9.2	5.9	5.0
3	.29	30	5.7	4.7	5.1
3	.29	50	5.4	5.0	5.4
5	.5	15	8.8	5.2	4.3
5	.5	30	4.0	3.0	2.7
5	.5	50	2.5	2.0	2.0
8	.71	15	12.7	7.4	2.4
8	.71	30	4.2	3.0	2.7
8	.71	50	2.7	1.6	2.5
8	.71	100	0.8	0.7	2.3

As r decreases, however, the conservativeness of the Scheffé procedures (for $k \geq 5$) will be increasingly swamped by the effects of increasing non-normality of the estimated proportions, $\hat{\pi}$. For the F-based version, the inflation in error rate due to non-normality is less than for the chi-squared version of equation (3), with the result that ER_T for the F-based version never seriously exceeds the nominal 5% rate. For moderate levels of clustering ($\bar{\lambda} = 1.5$), the behaviour of the F-based procedure is qualitatively similar to that described above for the case $\bar{\lambda} = 2$. From the point of view of total error rate, therefore, the F-based Scheffé procedure is useable over a wide range of clustering situations, though its possible conservativeness is a disadvantage.

5.3 Modified Quesenberry-Hurst Intervals

Total error rates for the first order modified Quesenberry-Hurst (Q-H) procedure of Section 3.2 are also shown in Table 1 for $\alpha = 5\%$, $\bar{\lambda} = 2$ and $a > 0$.

Total error rates are close to or below the nominal 5% for all combinations of r and k shown. For moderate to large numbers of clusters ($r \geq 30$), error rates for $k = 5$, and 8 are very similar, being approximately one half of the nominal rate (true also when $k = 12$). For the case of constant design effects (see Thomas 1989), error rates for first order modified Q-H intervals are conservative for $k \geq 5$, particularly for large r. The absence of this Scheffé-like conservativeness for the more realistic case of unequal design effects shown in Table 1 can again be explained using the argument of Section 3.1. From equation (6), it is easily seen that the asymptotic coverage of the first-order modified Q-H intervals is given by one minus the probability that at least one of the random variables $(\hat{\pi}_i - \pi_i)^2/(\bar{\lambda}\pi_i(1 - \pi_i)/n)$, $i = 1, \ldots, k$, will exceed the critical value $\chi^2_{k-1}(\alpha)$ asymptotically. When $a > 0$, these individual random variables will not all be asymptotically distributed as chi-squared on one degree of freedom, so that the bound of Section 3.1 does not apply. The true bound on the error rate will be inflated since at least one of the random variables $(\hat{\pi}_i - \pi_i)^2/(\bar{\lambda}\pi_i(1 - \pi_i)/n)$ will be stochastically larger than $(\hat{\pi}_i - \pi_i)^2/(\nu_{ii}/n)$, whenever $a > 0$.

Trends for the case $\bar{\lambda} = 1.5$ are similar (Thomas 1989). Overall, the results show that from the point of view of total error rates, first-order modified Q-H intervals provide a safe but somewhat conservative SCI procedure under realistic clustering conditions.

5.4 Simple Bonferroni Intervals

Total error rates for the simple Bonferroni intervals given by equation (9) are summarized in Table 2 for the case $\alpha = 5\%$, $\bar{\lambda} = 2$, $a > 0$, and $k = 3, 5$ and 8. Also shown are corresponding error rates for the t-based variants described in Section 3.5.

From Table 2, it is evident that the error performance of both sets of SCI's is poor, both showing a strong tendency to high error rates for small to medium numbers of clusters when k, the number of categories, is five or more. Using critical values of Student's t distribution to compensate for the variability in the estimated variances of the category proportions clearly has the effect of generally lowering error rates. As can be seen from Table 2, however, this strategy is unable to prevent significant error rate inflation in the t-based intervals as the number of clusters decreases, except when $k = 3$. The trend to inflated error rates for small numbers of clusters (for both z and t-based intervals), is due to the increasing non-normality of the $\hat{\pi}_i$'s with decreasing r. This trend gets progressively more severe as k increases, which is to be expected since non-normality will become more pronounced, for a given value of r, as the values of the π_i's get smaller. This is precisely what happens with increasing k in the case under study, for which $\hat{\pi}_i = 1/k \; \forall \; i$.

When $k = 3$, error rates for the t-based procedure are essentially constant, and close to the nominal level. For $k = 8$, on the other hand, ER_T varies from close to 20% at $r = 15$ to approximately 8% at $r = 100$. From Table 2, and other results not shown, it appears that for $k \geq 8$, simple t-based intervals approach their Bonferroni limits very slowly as $r \rightarrow \infty$. Also, for $k \leq 5$, error rates are close to the nominal level for moderate to large numbers of clusters ($r \geq 40$). Results for constant design effects, and for the case $\bar{\lambda} = 1.5$ are consistent with the above. From the point of view of total error rates (or equivalently of coverage rates), it is clear that simple t-based Bonferroni intervals are useable in practice over a range of realistic clustering situations only if $k \leq 5$ and $r \geq 40$.

Table 2

Total Error Rates for z and t-Based Simple Bonferroni Intervals;
$\alpha = 5\%$, $\bar{\lambda} = 2$, $m = 10$

			Total Error Rate (ER_T)	
k	a	r	z-based	t-based
3	.29	15	10.0	5.6
3	.29	30	6.3	4.9
3	.29	50	6.5	5.5
5	.50	15	15.0	9.7
5	.50	30	8.8	7.2
5	.50	50	7.2	5.5
8	.71	15	29.6	19.1
8	.71	30	15.0	11.0
8	.71	50	11.5	9.8
8	.71	100	8.1	7.8

5.5 Transformed Bonferroni Intervals

The more detailed results given in Thomas (1989) demonstrate that the problem of error rate inflation exhibited by simple z-based Bonferroni intervals is not solved by the use of transformations alone. All three transformed z-based intervals again display severely inflated error rates for small to medium numbers of clusters. Fortunately, the effect of transformations on the t-based Bonferroni intervals is very different, as can be seen from the results summarized in Table 3.

For $k = 3$, 5 and 8, error rates for the log and logit intervals are close to the nominal 5% for all r values shown, with the logit intervals yielding slightly lower rates than the log intervals (see the footnote to Table 3). The t-based square root intervals, on the other hand, exhibit the undesirable characteristic of error rate inflation for small r, when $k \geq 8$; they will not be considered further. For large numbers of categories ($k = 12$), both log and logit intervals do exhibit some error rate inflation for intermediate numbers of clusters ($r = 30$). This is not a serious drawback, however, as this number of categories is rarely encountered in practice. Results for constant design effects, and for the case $\bar{\lambda} = 1.5$ are generally similar to those described above.

It thus appears that for the ranges of k, r, $\bar{\lambda}$ and a that are likely to be encountered in practice, log and logit transformations (which reduce the non-normality in $\hat{\pi}$) used in combination with t-based critical values (which compensate for the variability in the estimated variances) do yield intervals that provide the desired degree of control. These intervals will be explored further in Section 6 in terms of the symmetry of their error rates.

Table 3

Total Error Rates[1] for t-based Transformed Bonferroni Intervals; $\alpha = 5\%$, $\bar{\lambda} = 2$, $m = 10$ for $k \leq 8$, $m = 20$ for $k = 12$

			Total Error Rate (ER_T)		
			t-based Transformed Bonferroni		
k	a	r	Square Root	Log	Logit
3	.29	15	4.5	4.6	3.3
3	.29	30	3.6	4.0	3.5
3	.29	50	4.6	5.6	4.1
5	.5	15	6.4	4.7	4.6
5	.5	30	4.6	4.2	3.5
5	.5	50	4.3	4.5	4.0
8	.71	15	12.0	5.9	5.2
8	.71	30	6.2	6.6	5.2
8	.71	50	5.9	5.4	5.2
8	.71	100	4.9	3.9	4.2
12	.91	15	17.0	6.7	6.5
12	.91	30	12.9	10.1	10.2
12	.91	50	8.2	6.5	6.3

[1] For $k = 8$ and $r = 50$, the correlation between ER_T estimates for log and logit intervals is 0.92. Assuming this is typical for all r and k, the Monte Carlo standard error of the difference in log and logit error rates is approximately 0.3%.

Table 4

Percentage Asymmetry (PER_U)[1] in the Total Error Rate for the Viable Procedures; $a > 0$[2], $r = 50$, $m = 10$ for $k \leq 8$, $m = 20$ for $k = 12$

			$PER_U = (ER_U/ER_T) \times 100\%$			
			Scheffé	Modified Q-H	t-based Bonferroni	
α	k	$\bar{\lambda}$	(F-based)	(first order)	(log)	(logit)
5%	5	1.5	19.2	58.7	61.0	48.9
5%	5	2.0	0.0	45.0	61.1	48.8
5%	8	1.5	0.0	63.2	67.5	56.8
5%	8	2.0	0.0	65.2	64.9	49.0
5%	12	2.0	0.0	46.9	53.8	51.6
10%	5	1.5	16.3	49.4	59.2	48.4
10%	5	2.0	6.1	50.0	61.8	48.6
10%	8	1.5	0.0	60.7	67.3	55.8
10%	8	2.0	0.0	65.6	60.7	50.0
10%	12	2.0	0.0	47.5	56.0	51.4

[1] For $k = 8$, $\bar{\lambda} = 2$ and $\alpha = 5\%$, the correlation between PER_Uestimates for log and logit intervals is 0.82. Assuming this is typical, Monte Carlo standard errors for differences in log and logit PER_U's are approximately 4% and 3% for $\alpha = 5\%$ and 10%, respectively.

[2] For values of a for specific k, see Table 3.

6. ERROR RATE SYMMETRIES FOR THE VIABLE PROCEDURES

This section presents results on error rate symmetry based on the decomposition of the total error rate ER_Tinto its two additive components $ER_U$and ER_L, as described in Section 4. The measure used in the tables is $(ER_U/ER_T) \times 100\%$, *i.e.*, the upper error rate expressed as a percentage of the total error rate. It will be denoted PER_U. A symmetric SCI will have an empirical PER_U that is close to 50%; a PER_U that is greater (less) than 50% will indicate an increased probability of non-coverage due to intervals lying above (below) their respective π_i's. For values of percentage symmetry between 50% and 80%, 95% confidence intervals on the true PER_U are approximately $(PER_U \pm 14)\%$ and $(PER_U \pm 10)\%$ for total error rates of 5% and 10% respectively.

6.1 Modified Scheffé and Quesenberry-Hurst Intervals

Percentage symmetry results for the F-based Scheffé and the first order Quesenberry-Hurst (Q-H) intervals are given in Table 4 for a selection of parameter values. It can be seen that the Scheffé procedure displays extreme asymmetry, making it an unattractive SCI. The first order modified Q-H procedure displays only moderate asymmetry, and is therefore the better of the two in practice.

The source of the asymmetry in the Scheffé intervals is again the non-normality of the untransformed $\hat{\pi}_i$'s. In particular, the fact that "small" $\hat{\pi}_i$'s generate "small" estimates of the variances v_{ii} and hence shorter intervals (*cf.* the multinomial case where $\hat{v}_{ii} = \hat{\pi}_i(1 - \hat{\pi}_i)/n$, $i = 1, \ldots, k$) increases the probability that non-covering intervals with lie below their respective π_i's. This tendency to asymmetry will increase as the total error rate decreases, making the F-based Sheffé procedure particularly vulnerable to this effect. Since Scheffé intervals differ from simple Bonferroni intervals only through the critical constant used, asymmetry is also to be expected in the latter though it should not be as severe given that error rates for simple Bonferroni intervals are liberal. This is confirmed by study results, *e.g.*, $PER_u = 4.9\%$ for simple t-based Bonferroni intervals when $r = 50$, $k = 8$ and $a = 0.71$.

6.2 t-Based Transformed Bonferroni Intervals

Table 4 also gives percentage symmetry results for t-based Bonferroni intervals based on the log and logit transformations. The results of the table suggest that logit intervals do provide more symmetric coverage than the log intervals, when k is in the range 5 to 8. Thus logit intervals might be considered preferable in practice to log intervals from the point of view of error rate symmetry.

7. UNEQUAL CELL PROBABILITIES

Table 5 presents results on total error rates and error rate symmetry under unequal cell probabilities for the t-based log and logit transformed Bonferroni procedures, together with results for the first order modified Q-H procedure. Results are tabulated for six sets of unequal probabilities, three for the case $k = 5$, $\bar{\lambda} = 2$, $a = 0.5$, namely $\pi(5, 3, .3)$, $\pi(5, 2, .425)$ and $\pi(5, 1, .8)$, (see Section 4.1), and three for the case $k = 8$, $\bar{\lambda} = 2$, $a = 0.71$, namely $\pi(8, 3, .25)$, $\pi(8, 2, .35)$ and $\pi(8, 1,. 65)$. For each π vector, the remaining $k - q$ elements all equal 0.05. Results for equiprobable cells are also displayed in Table 5 for comparison.

It can be seen that deviations from equiprobability do affect total error rates for the modified Q-H procedure, particularly when $k = 8$. With the first element $\pi_1 = 0.65$ the total error rate of modified Q-H is close to its error rate under equiprobability. For the other two cases studied ($\pi_1 = \pi_2 = .35$, and $\pi_1 = \pi_2 = \pi_3 = 0.25$), total error rates are considerably lower, closer in fact to the modified Q-H results obtained for the constant design effect case (see Thomas 1989). This difference in total error rates occurs because the pattern of cell design effects is different for each set of unequal probabilities, though the pattern of generalized design effects (the λ's) remains the same ($\lambda_1 = 2 + 2\sqrt{3}$, $\lambda_j = 2 - \sqrt{3}/3$, $j = 2, \ldots, 7$ for $\bar{\lambda} = 2$, $a = \sqrt{2}/2 = .707$). When $\pi_1 = 0.65$, the cell design effects are $d_1 = 5.7$, $d_i = 1.82$, $i = 2, \ldots, 8$.

Table 5

The Effect of Unequal Cell Probabilities on the Total Error Rates (ER_T) and Percentage Asymmetries (PER_U) of the Modified Q-H and Transformed Bonferroni Procedures;
$r = 50$, $\bar{\lambda} = 2$, $a = 5\%$, $m = 10$

		Procedures					
		Modified Q-H (first order)		t-based Bonferroni			
				(log)		(logit)	
k	$\pi(k,q,\phi)$	ER_T	PER_U	ER_T	PER_U	ER_T	PER_U
5	$\pi(5,1,0.8)$	3.2	7.3	5.6	75.9	4.4	62.5
5	$\pi(5,2,0.425)$	1.4	82.1	4.8	57.2	4.6	47.8
5	$\pi(5,3,0.3)$	1.5	76.7	4.2	51.2	3.9	38.5
5	equi-prob.	2.0	45.0	4.5	61.1	4.0	48.8
8	$\pi(8,1,0.65)$	2.7	63.0	6.3	68.3	5.4	55.6
8	$\pi(8,2,0.35)$	0.6	83.3	4.9	58.2	4.4	51.2
8	$\pi(8,3,0.25)$	0.7	100	5.2	68.2	4.6	63.1
8	equi-prob.	2.5	66.5	6.0	64.0	5.2	49.0

Use of a uniform adjustment factor ($\hat{\lambda}$) will thus seriously underestimate the variance of the first estimated cell probability, leading to inflation of the error rate of the modified Q-H procedure. That the nominal error rate $\alpha = 5\%$ is not exceeded is due to the inherent conservativeness of modified Q-H intervals in the constant design effect case (see Section 5.3). When $\pi_1 = \pi_2 = 0.35$, corresponding design effects are $d_1 = d_2 = 2.36, d_i = 1.97, i = 3, \ldots, 8$. These are much closer to constant design effects ($d_i = 2.0, i = 1, \ldots, 8$) hence the conservative behaviour of the intervals in this case. It can also be seen from Table 5 that conservative ER_T's are associated with highly asymmetric error rates.

Despite the variation in cell design effects implied by the different probability vectors of Table 5, it can be seen that the transformed Bonferroni procedures exhibit very stable performance. Total error rates (for 50 clusters) are close to the nominal rate ($\alpha = 5\%$) for both log and logit intervals, and neither exhibits serious asymmetry. Total error rates corresponding to unequal probabilities do decrease with decreasing r over the range $r = 50$ to $r = 15$ when $k = 8$ (results not shown). Variations in ER_T are not severe, however; when $r = 15$ clusters the minimum rate for the cases examined is approximately 2%.

8. SUMMARY, CONCLUSIONS AND RECOMMENDATIONS

In the search for procedures that take direct account of the survey design and that provide adequate control of error rates and error rate symmetry over a wide range of problem and clustering situations, Scheffé intervals based on estimated cell variances must be rejected: the chi-squared version of equation (3) on the grounds of poor error control, and the F-based version on the grounds of extreme asymmetry. Modifications to Quesenberry-Hurst intervals are somewhat conservative, though the version based on the first order Rao-Scott correction does provide a viable procedure. For Bonferroni intervals, the benefits of using critical points of the t-distribution instead of the standard normal are substantial. Even so, intervals based on $\hat{\pi}$ and its square root provide inadequate control of total error rates, particularly for small numbers of clusters when the distribution of $\hat{\pi}$ becomes increasingly non-normal. On the other hand, t-based Bonferroni intervals using both the log and logit transformations provide good control of total error rates and error rate symmetry, and are clearly superior to all other competing intervals. Both log and logit transformed intervals (t-based) also appear to provide good control of error rates and error rate symmetry when the cell probabilities are unequal, differing in the cases studied by a ratio (maximum to minimum) of up to sixteen. From the point of view of total error rates there is little to choose between the log and logit intervals, though error rates for the latter are consistently a little lower. Logit intervals are superior from the point of view of symmetry, however. Estimates of confidence interval lengths (detailed results not shown) also favour the logit intervals, despite their slightly lower error rates. For example, for the equiprobable case with $\alpha = 5\%$, $k = 5$, $\bar{\lambda} = 2$, $a = 0.5$ and $r = 50$, the average length of the confidence interval on π_1 (expressed as a 95% confidence interval) was .1915 ± .0014 for the log-based interval, and .1850 ± .0014 for the logit-based interval. For the case of unequal probabilities, with $\alpha = 5\%$, $k = 8$, $\bar{\lambda} = 2$, $a = 0.71$, $r = 50$, $\pi_1 = 0.65$ and $\pi_2 = 0.05$ (see Table 5), 95% confidence intervals for the average lengths of log and logit intervals were: for π_1, .2865 ± .0012 and .2776 ± .0011, respectively; for π_2, .0806 ± .0010 and .0789 ± .0011, respectively.

Before final recommendations are made, it is necessary to consider possible limitations imposed by the design of the Monte Carlo study. A potentially limiting feature is the use of a single specific sampling design, namely two-stage cluster sampling with SRS at the second stage, given that practitioners will encounter data collected using a range of survey designs that might include stratification and multiple levels of unit selection. For large samples, the relevant distribution theory requires knowledge only of first and second moments, assuming that a suitable central limit theorem applies (see for example Rao and Scott 1981). This study will therefore yield valid recommendations for large numbers of clusters, or more generally for large numbers of degrees of freedom for variance estimation (Rao and Thomas 1988), as long as the covariance matrix V/n and hence the generalized design effects can be appropriately modelled. Since the Dirichlet mixture model used in this study yields generalized design effects having means and coefficients of variation that are typical of those found in practice, recommendations based on a large number of clusters or degrees of freedom (fifty or more) can be made with confidence. For small to moderate numbers of clusters, quantitative results may differ from design to design. Since the basic mechanisms underlying the results exhibited in this study, namely increasing non-normality of $\hat{\pi}$ for decreasing r plus the inherent conservativeness of Scheffé-like procedures, will apply in general, it is expected that the qualitative trends for the different statistics examined will be generalizable across a wide variety of designs, even when the number of clusters is not large. The basic aim of the study has been to identify procedures whose control of error rates is robust to variations in the study parameters, namely the number of categories, the number of clusters, the strength of clustering, and the skewness of the vector of category probabilities. The combination of parameters examined has covered much of the range likely to be encountered in practice, so it is reasonable to suggest that the robustness exhibited by the log and logit transformed Bonferroni intervals might extend to variations in survey design, for moderate numbers of clusters (or degrees of freedom). Further research on this question is clearly required.

Subject to these caveats, t-based Bonferroni simultaneous confidence intervals based on the logit transformation are recommended for assessing up to $k = 12$ proportions of varying magnitude, under realistic clustering conditions. If conservativeness is deemed to be an asset, the first-order modified Quesenberry-Hurst procedure can be safely used. Both procedures require only a knowledge of the variances (or design effects) of the estimated cell proportions.

ACKNOWLEDGMENTS

I wish to thank J.N.K. Rao for many valuable discussions and for his comments on a draft of this paper. Thanks are also due to Steve Brockwell both for useful suggestions and for excellent programming assistance, and to Paul Bertelman for programming assistance during the latter part of the project. Finally I wish to thank two reviewers and an associate editor for suggestions that significantly improved the presentation of the results. This research was supported by a grant from the Natural Science and Engineering Research Council of Canada.

REFERENCES

ANDREWS, R.W., and BIRDSALL, W.C. (1988). Simultaneous confidence intervals: a comparison under complex sampling. Paper presented at the 1988 American Statistical Association Annual Meeting, Chicago.

BAILEY, B.J.R. (1980). Large sample simultaneous confidence intervals for the multinominal probabilities based on transformation of the cell frequencies. *Technometrics*, 22, 583-589.

BLACK, D., and MYLES, J. (1986). Dependent industrialization and the Canadian class structure: a comparative analysis of Canada, the United States, and Sweden. *Canadian Review of Sociology and Anthropology*, 23, 157-181.

BRIER, S.S. (1980). Analysis of contingency tables under cluster sampling. *Biometrika*, 67, 591-596.

FELLEGI, I.P. (1980). Approximate tests of independence and goodness of fit based on stratified multistage samples. *Journal of the American Statistical Association*, 75, 261-268.

FITZPATRICK, S., and SCOTT, A.J. (1987). Quick simultaneous confidence intervals for multinomial proportions. *Journal of the American Statistical Association*, 82, 875-878.

GOLD, R.Z. (1963). Tests auxiliary to χ^2 tests in a Markov chain. *Annals of Mathematical Statistics*, 34, 56-74.

GOODMAN, L.A. (1965). On simultaneous confidence intervals for multinomial proportions. *Technometrics*, 7, 247-254.

HIDIROGLOU, M.A., and RAO, J.N.K. (1987). Chi-squared tests with categorical data from complex surveys: Part I - simple goodness-of-fit, homogeneity and independence in a two-way table with applications to the Canada Health Survey (1978-1979). *Journal of Official Statistics*, 3, 117-132.

HOCHBERG, Y., and TAMANE, A.C. (1987). *Multiple Comparison Procedures*. New York: Wiley.

HOLT, D., SCOTT, A.J., and EWING, P.D. (1980). Chi-squared tests with survey data. *Journal of the Royal Statistical Society*, Ser. A, 143, 303-320.

JENNINGS, D.E. (1987). How do we judge confidence-interval adequacy? *The American Statistician*, 41, 335-337.

MILLER, R.G. (1981). *Simultaneous Statistical Inference*, 2nd Edition. New York: Springer-Verlag.

QUESENBERRY, C.P., and HURST, D.C. (1964). Large sample simultaneous confidence intervals for multinomial proportions. *Technometrics*, 6, 191-195.

RAO, J.N.K., and SCOTT, A.J. (1979). Chi-squared tests for analysis of categorical data from complex surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 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, 261-230.

RAO, J.N.K., and THOMAS, D.R. (1988). The analysis of cross-classified categorical data from complex sample surveys. *Sociological Methodology*, 18, 213-269.

THOMAS, D.R. (1989). An investigation of simultaneous confidence interval procedures for proportions, under cluster sampling. Working Paper WPS 89-02, School of Business, Carleton University.

THOMAS, D.R., and RAO, J.N.K. (1987). Small-sample comparisons of level and power for simple goodness-of-fit statistics under cluster sampling. *Journal of the American Statistical Association*, 82, 630-636.

Survey Methodology, December 1989
Vol. 15, No. 2, pp. 203–223
Statistics Canada

Logistic Regression Under Complex Survey Designs

JORGE G. MOREL[1]

ABSTRACT

Estimation procedures for obtaining consistent estimators of the parameters of a generalized logistic function and of its asymptotic covariance matrix under complex survey designs are presented. A correction in the Taylor estimator of the covariance matrix is made to produce a positive definite covariance matrix. The correction also reduces the small sample bias. The estimation procedure is first presented for cluster sampling and then extended to more complex situations. A Monte Carlo study is conducted to examine the small sample properties of F-tests constructed from alternative covariance matrices. The maximum likelihood estimation method where the survey design is completely ignored is compared with the usual Taylor's series expansion method and with the modified Taylor procedure.

KEY WORDS: Pseudo-likelihood; CPLX procedure; Cluster sampling; Adjusted covariance matrix.

1. INTRODUCTION

In the last few years a lot of attention has been given to the problems that arise when chi-square tests based on the multinomial distribution are applied to data obtained from complex sample designs. It has been shown that the effects of stratification and clustering on the chi-square tests may lead to a distortion of nominal significance levels. Holt, Scott and Ewings (1980) proposed modified Pearson chi-square statistics tests of goodness-of-fit, homogeneity, and independence in two-way contingency tables. Rao and Scott (1981) presented similar tests for complex sample surveys. In all these cases, the correction factor requires only the knowledge of variance estimates (or design effects) for individual cells. Bedrick (1983) derived a correction factor for testing the fit of hierarchical log linear models with closed form parameter estimates. Rao and Scott (1984) presented more extensive methods of using design effects to obtain chi-square tests for complex surveys. They generalized their previous results to multiway tables. Fay (1985) presented the adjustments to the Pearson and likelihood test statistics through a jackknife approach.

The use of the conditional logistic model, Cox (1970), has become increasingly popular in the context of complex survey designs. Under suitable conditions, Binder (1983), proved the asymptotic normality of design-based sampling distribution for a family of parameter estimators that cannot be defined explicitly as a function of other statistics from the sample. His results are applied to binary logistic models. Further applications to the Canada Health Survey are also found in Binder *et al.* (1984).

Chambless and Boyle (1985) derived a general asymptotic distribution theory for stratified random samples with a fixed number of strata and increasing stratum sample sizes. Their theoretical results were illustrated with logistic regression and discrete proportional hazardsmodels. Albert and Lesaffre (1986) discussed the logistic discrimination method for classifying multivariate observations into one of several populations. They restrict their attention to discrimination between qualitatively distinct groups.

[1] Jorge G. Morel is Assistant Professor of the Department of Epidemiology and Biostatistics, University of South Florida, Tampa, Florida 33612.

Extensions to the case where the response consists of a polychotomous variable have been done by Bull and Pederson (1987) and Morel (1987). They show, by using Taylor's series expansion, that the large sample variance of the beta estimates has the form

$$H^{-1}\,GH^{-1}$$

where H^{-1} is the covariance matrix that wrongly results from assuming independence and multinomial distribution in the response vector, and G is a matrix whose estimation is based in the complex survey design.

More recently, Roberts, Rao and Kumar (1987) showed how to make adjustments that take into account the survey design in computing the standard chi-square and the likelihood ratio test statistics for logistic regression analysis involving a binary response variable. The adjustments are based on certain generalized design effects. Their results can be applied to cases where the whole population has been divided into I domains of study, a large sample is obtained for each domain, and in each domain a proportion π_i, $i = 1, 2, \ldots, I$, is to be estimated. It is assumed

$$\pi_i = \left[1 + \exp\left(x_i \underset{\sim}{\beta}^0\right)\right]^{-1} \exp\left(x_i \underset{\sim}{\beta}^0\right),\ i = 1, 2, \ldots, I,$$

where x_i is a k-vector of known constants derived from the i-th domain and $\underset{\sim}{\beta}^0$ is a k-vector of unknown parameters. This procedure may be most useful when only the summary table of counts and variance adjustment factors are available, instead of the complete data set.

In this paper an estimation procedure is presented for obtaining consistent estimators of the parameter vector of a generalized logistic model and its asymptotic covariance matrix when a complex sampling design is employed. The resulting estimated covariance matrix is always positive definite and asymptotically equivalent to the one obtained from Taylor's series expansion. A correction for reducing the small sample bias in the estimated covariance matrix is also introduced. It is shown, via a Monte Carlo study, that this correction levels off the inflated Type I error that arises from ignoring the complex survey, faster than the Taylor's series expansion. In this sense the correction proposed here produces, for small samples, results that are superior to the usual delta-method.

The new procedure will be termed, henceforth, the CPLX procedure, or simply CPLX. The maximum likelihood estimation method and the Taylor's series expansion method will be termed MLE and TAYLOR, respectively. The CPLX procedure has been incorporated into PC CARP, a personal computer program for variance estimation with large scale surveys, see Schnell *et al.* (1988).

2. LOGISTIC REGRESSION WITH CLUSTER SAMPLING

Consider first single-stage cluster sampling where n clusters or primary sampling units are taken with known probabilities with replacement from a finite population or without replacement from a very large population. Let m_j represent the size of the j-th cluster, $j = 1, 2, \ldots, n$, and let $y^*_{j\ell}$, $\ell = 1, 2, \ldots, m_j$ denote $(d + 1)$ dimensional classification vectors. The vector $y^*_{j\ell}$ consists entirely of zeros except for position r which will contain a one if the ℓ-th unit selected from the j-th cluster falls in the r-th category. Let $x_{j\ell}$ be a k-dimensional row vector of explanatory variables associated with the ℓ-th unit selected from the j-th cluster.

Then, for each $j = 1, 2, \ldots, n$, and each $\ell = 1, 2, \ldots, m_j$, the expectation of the r-th element of $y_{j\ell}^*$ is determined by a logistic relationship as

$$\begin{aligned}\pi_{j\ell r} = E\{y_{j\ell r}\} &= \left[1 + \sum_{s=1}^{d} \exp(x_{j\ell}\beta_s^0)\right]^{-1} \exp(x_{j\ell}\beta_r^0) \quad r = 1, 2, \ldots, d \\ &= 1 - \sum_{s=1}^{d} \pi_{j\ell s}, \qquad\qquad r = d + 1. \end{aligned} \tag{2.1}$$

Because the expected value function is nonlinear in the parameter vector $\underset{\sim}{\beta}^0 = (\underset{\sim}{\beta}_1^{0\prime}, \underset{\sim}{\beta}_2^{0\prime}, \ldots, \underset{\sim}{\beta}_d^{0\prime})'$, it is necessary to use nonlinear estimation methods. Define the pseudo log-likelihood $L_n(\underset{\sim}{\beta})$ as

$$L_n(\underset{\sim}{\beta}) = \sum_{j=1}^{n} \sum_{\ell=1}^{m_j} w_j(\log \underset{\sim}{\pi}_{j\ell}^*)' \, y_{j\ell}^*, \tag{2.2}$$

where $\underset{\sim}{\pi}_{j\ell}^* = (\pi_{j\ell 1}, \ldots, \pi_{j\ell, d+1})'$ and w_j is the sampling weight for the $j\ell$-th sampling unit. This function can be viewed as a weighted log likelihood function, where the weights are the sampling weights and the $y_{j\ell}^*$'s are distributed as multinomial random variables. If the sampling weights are all one, then (2.2) becomes the log-likelihood function under the assumption that the $y_{j\ell}^*$'s are independently multinomially distributed.

Let $\hat{\underset{\sim}{\beta}}_{\text{PSEUDO}}$ be the estimator of $\underset{\sim}{\beta}^0$ that maximizes (2.2). This estimator is a solution to the system of equations

$$\sum_{j=1}^{n} \sum_{\ell=1}^{m_j} w_j \, G(\underset{\sim}{\beta}, x_{j\ell}) \left[\text{Diag}(\underset{\sim}{\pi}_{j\ell}^*)\right]^{-1} (y_{j\ell}^* - \underset{\sim}{\pi}_{j\ell}^*) = \mathbf{0}, \tag{2.3}$$

where

$$G(\underset{\sim}{\beta}, x_{j\ell}) = \left[(I_{d \times d}, \mathbf{0}_{d \times 1}) \otimes x_{j\ell}'\right] \Delta(\underset{\sim}{\pi}_{j\ell}^*),$$

$$\Delta(\underset{\sim}{\pi}_{j\ell}^*) = \text{Diag}(\underset{\sim}{\pi}_{j\ell}^*) - \underset{\sim}{\pi}_{j\ell}^*(\underset{\sim}{\pi}_{j\ell}^*)',$$

and ⊗ denotes the Kronecker product.

The asymptotic normality of $\hat{\underset{\sim}{\beta}}_{\text{PSEUDO}}$ can be proved by defining the parameters of interest implicity as in (2.2) and then by extending the results given in Binder (1983). An alternative approach can be derived by making use of the pseudo-likelihood assumption and Proposition 1 in Dale (1986). Binder and Dale both provide the necessary regularity conditions.

As n increases,

$$\begin{aligned}\sqrt{n}(\hat{\underset{\sim}{\beta}}_{\text{PSEUDO}} - \underset{\sim}{\beta}^0) &= \sqrt{n}\left[H_n(\underset{\sim}{\beta}^0)\right]^{-1} U_n(\underset{\sim}{\beta}^0) \\ &\xrightarrow{L} N_{dk}(\mathbf{0}, \lim_{n \to \infty} \left[H_n(\underset{\sim}{\beta}^0)\right]^{-1} G_n\left[H_n(\underset{\sim}{\beta}^0)\right]^{-1}) \end{aligned} \tag{2.4}$$

where,

$$H_n(\underset{\sim}{\beta}^0) = \sum_{j=1}^{n} \sum_{\ell=1}^{m_j} w_j \, \Delta(\underset{\sim}{\pi}_{j\ell}) \otimes x'_{j\ell} x_{j\ell},$$

$$U_n(\underset{\sim}{\beta}^0) = \sum_{j=1}^{n} \sum_{\ell=1}^{m_j} w_j(y_{j\ell} - \underset{\sim}{\pi}_{j\ell}) \otimes x'_{j\ell},$$

$$G_n = \sum_{j=1}^{n} \sum_{\ell=1}^{m_j} w_j^2 \operatorname{Var}(y_{j\ell}) \otimes x'_{j\ell} x_{j\ell},$$

$y_{j\ell}$ and $\underset{\sim}{\pi}_{j\ell}$ are the vectors $y^*_{j\ell}$ and $\underset{\sim}{\pi}^*_{j\ell}$, without their last elements, respectively and N_{dk} denotes a dk-multivariate normal distribution.

Nelder and Wedderburn (1972) have shown that under binomial assumption, the pseudo log-likelihood function (2.2) can be solved by an iterative weighted least-squares procedure. Haberman (1974, p.48) shows that under regularity conditions a modified Newton-Raphson converges to the maximum likelihood estimator for the multinomial case. His proof does not depend on the existence of any consistent estimator of $\underset{\sim}{\beta}^0$ which allows the iterative algorithm to be initialized at $\hat{\underset{\sim}{\beta}} = \mathbf{0}$. Jennrich and Moore (1975) proved that when the multinomial assumption holds, the common Gauss-Newton algorithm for finding the maximum likelihood estimator of $\underset{\sim}{\beta}^0$ becomes the Newton-Raphson algorithm. Because of this equivalence of those algorithms and because a modified Newton-Raphson procedure always converge, we have adopted the modified Gauss-Newton algorithm described by Gallant (1987, p.318).

CPLX first finds $\hat{\underset{\sim}{\beta}}_{\text{PSEUDO}}$ using an iterative procedure in which the estimate of $\underset{\sim}{\beta}^0$ at the q-th step is

$$\hat{\underset{\sim}{\beta}}_{[q,i(q)]} = \hat{\underset{\sim}{\beta}}_{[q-1,i(q-1)]} + (0.5)^{i(q)} \left[H_n\left(\hat{\underset{\sim}{\beta}}_{[q-1,\, i(q-1)]}\right)\right]^{-1} U_n\left(\hat{\underset{\sim}{\beta}}_{[q-1,\, i(q-1)]}\right) \tag{2.5}$$

where $i(q)$ is a nonnegative integer such that

$$L_n\left(\hat{\underset{\sim}{\beta}}_{[q,i(q)]}\right) > L_n\left(\hat{\underset{\sim}{\beta}}_{[q-1,i(q-1)]}\right). \tag{2.6}$$

The modification of the iteration algorithm provided by $i(q)$ guarantees the convergence of the procedure. The iteration is initiated by setting $\hat{\underset{\sim}{\beta}}_{(0)} = \mathbf{0}$. The algorithm is declared to have converged when the condition

$$\frac{L_n\left(\hat{\underset{\sim}{\beta}}_{[q,i(q)]}\right) - L_n\left(\hat{\underset{\sim}{\beta}}_{[q-1,i(q-1)]}\right)}{\left|L_n\left(\hat{\underset{\sim}{\beta}}_{[q,i(q)]}\right)\right| + 10^{-5}} < \epsilon \tag{2.7}$$

is satisfied, where ϵ can be 10^{-8}.

Observe that a consistent estimator of $H_n(\underset{\sim}{\beta}^0)$ is $H_n(\hat{\underset{\sim}{\beta}}_{\text{PSEUDO}})$ and a distribution free estimator of G_n is

$$\overset{*}{G}_n = (n - 1)^{-1} n \sum_{j=1}^{n} (d_j - \bar{d})\,(d_j - \bar{d})', \tag{2.8}$$

where

$$d_j = \sum_{t=1}^{m_j} w_j(y_{jt} - \underset{\sim}{\pi}_{jt}) \otimes x_{jt}',$$

and $\bar{d} = n^{-1} \sum_{j=1}^{n} d_j$. If within each cluster, the y_{jt}^*'s are independent and identically distributed according to a multinomial random vector with parameters $(\underset{\sim}{\pi}_j^*, 1)$, then it can be easily shown that the expectation of $\overset{*}{G}_n$ is precisely $H_n(\underset{\sim}{\beta}^0)$. In practice the $\underset{\sim}{\pi}_{jt}$'s in (2.8) are replaced with $\hat{\underset{\sim}{\pi}}_{jt}$ where $\hat{\pi}_{jt_r}$ is defined as in (2.1) with $\hat{\beta}_{\text{PSEUDO}}$ substituted by $\underset{\sim}{\beta}^0$, and a small correction is applied to obtain the estimator

$$\hat{G}_n = (n^* - k)^{-1}\,(n^* - 1)\,(n - 1)^{-1} n \sum_{j=1}^{n} (\hat{d}_j - \hat{\bar{d}})\,(\hat{d}_j - \hat{\bar{d}})', \tag{2.9}$$

where

$$\hat{d}_j = \sum_{t=1}^{m_j} w_j(y_{jt} - \hat{\underset{\sim}{\pi}}_{jt}) \otimes x_{jt}',$$

$$\hat{\bar{d}} = n^{-1} \sum_{j=1}^{n} \hat{d}_j \quad \text{and} \quad n^* = \sum_{j=1}^{n} m_j.$$

The factor

$$(n^* - k)^{-1}\,(n^* - 1)\,(n - 1)^{-1}\,n$$

reduces to $(n - k)^{-1} n$ if each cluster contains exactly one element. The factor $(n - k)^{-1} n$ is the degrees of freedom correction applied to the residual mean square for ordinary least squares in which k parameters are estimated. The quantity in (2.9) is well defined for two or more clusters and the factor $(n^* - k)^{-1}\,(n^* - 1)$ should reduce the small sample bias associated with using the estimated function to calculate deviations. Therefore, a consistent estimator of the asymptotic covariance matrix of $\hat{\underset{\sim}{\beta}}_{\text{PSEUDO}}$ under the cluster sampling design is

$$\tilde{A}_n = \left[H_n(\hat{\underset{\sim}{\beta}}_{\text{PSEUDO}})\right]^{-1} \hat{G}_n \left[H_n(\hat{\underset{\sim}{\beta}}_{\text{PSEUDO}})\right]^{-1} \tag{2.10}$$

which can be used to test any hypothesis of the form H_O: $C\,\underset{\sim}{\beta}^0 = \underset{\sim}{\delta}^*$. Under the null hypothesis, by Moore (1977)

$$\left(C\hat{\underset{\sim}{\beta}}_{\text{PSEUDO}} - \underset{\sim}{\delta}^*\right)' \left[C\,\tilde{A}_n\,C'\right]^{-1} \left(C\hat{\underset{\sim}{\beta}}_{\text{PSEUDO}} - \underset{\sim}{\delta}^*\right) \tag{2.11}$$

converges in law to a chi-square distribution with $\upsilon = \text{rank}\,(C\tilde{A}_n C')$ degrees of freedom. Here, $[C\tilde{A}_n C']^{-1}$ is any generalized inverse of $C\tilde{A}_n C'$.

The sums of squares and products matrix used in the construction of $\hat{G}_n$ is based on n observations, where n is the number of clusters. By analogy to the Hotelling T^2 statistic, it is natural to adjust for degrees of freedom by multiplying (2.11) by the ratio

$$\frac{n-\upsilon}{\upsilon(n-1)} \tag{2.12}$$

to obtain an approximate F statistic with υ and $n - \upsilon$ degrees of freedom. In our case, this adjustment has the disadvantage that υ may exceed n in a sample with a small number of clusters but a large number of individual elements.

The covariance matrix constructed as if the elemental observations are a simple random sample is biased, but it can be used to make a small sample adjustment in the estimated covariance matrix. One might view the usual small sample degrees-of-freedom adjustment as the operation of adding to an initial estimator of the covariance matrix the quantity $(n-\upsilon)^{-1}\,\upsilon\,\hat{V}$, where $\hat{V}$ is also an estimator of the covariance matrix. In the usual case, $\hat{V}$ is also the initial estimator. In our case, we make the adjustment using the covariance matrix based on the elements as the second $\hat{V}$. In our case, the use of the elemental covariance matrix has the advantage that the resulting sum is always positive definite. The adjustment is a function of the number of parameter estimated, dk. The adjustment is

(1) if $n > 3dk - 2$

$$\hat{A}_n = \tilde{A}_n + (n-dk)^{-1}\,(dk-1)\,\gamma^*\big[H_n(\hat{\underset{\sim}{\beta}}_{\text{PSEUDO}})\big]^{-1}, \tag{2.13}$$

(2) if $n \leq 3dk - 2$

$$\hat{A}_n = \tilde{A}_n + 0.5\,\gamma^*\,\big[H_n(\hat{\underset{\sim}{\beta}}_{\text{PSEUDO}})\big]^{-1}, \tag{2.14}$$

where $\gamma^* = \max(1, \text{tr}\{[H_n(\hat{\underset{\sim}{\beta}}_{\text{PSEUDO}})]^{-1}\hat{G}_n\}/dk)$. The upper bound of 0.5 for correction in (2.14) is arbitrary. Then, an approximate F-test with υ and $n-\upsilon$ degrees of freedom is obtained by substituting $\hat{A}_n$ for $\tilde{A}_n$ in (2.11) and dividing the resulting quadratic form by υ. In practice, the approximate degrees of freedom can be taken to be υ and infinity.

3. A MONTE CARLO STUDY

In this section a Monte Carlo study is conducted to examine properties of F-Tests (2.11) involving model parameters. Data are generated under two different sampling schemes that correspond to single-stage cluster sampling where the primary units all have the same sampling weight and are taken from an infinite population. In the first sampling scheme all the elements within the cluster have the same explanatory vector x and therefore, the same conditional mean (2.1). This is the case where the logistic regression becomes weighted in the sense of several responses y's with the same covariate vector x. Different degrees of intra-class correlation are induced among the y's belonging to the same cluster.

The second sampling scheme, unlike the first, places different vectors of covariates for different subjects within the cluster. The conditional mean (2.1) is also satisfied and different degrees of intra-class correlation are controlled. The effect of the intra-class correlation is studied for both sampling schemes under three different estimation procedures: MLE where the clustering effect is completely ignored, TAYLOR where the large sample covariance matrix (2.10) is used, and CPLX where the adjusted covariance matrix (2.13-2.14) is employed. These last two procedures, for large samples, are asymptotically equivalent. For small samples CPLX performs better than TAYLOR.

3.1 Sampling Scheme I

Suppose that $x_1, x_2, \ldots, x_n$ are k-dimensional independent and identically distributed normal random vectors with vector mean $\underset{\sim}{\mu}$ and covariance matrix $\underset{\sim}{\Sigma}$. For each j, $j = 1, 2, \ldots, n$, suppose that given x_j, the random vectors $y_{j0}^0, y_{j1}^0, \ldots, y_{j,m_j}^0$ are independent and identically distributed multinomial random vectors, with parameters $(\underset{\sim}{\pi}_j^*, 1)$, where $\underset{\sim}{\pi}_j^*$ satisfies the logistic function (2.1) evaluated at the true parameter vector $\underset{\sim}{\beta}^0$ and at $x = x_j$. Let $U_{j1}, U_{j2}, \ldots, U_{j,m_j}$ be a set of independent and identically distributed uniform (0,1) random variables. For a known and fixed ζ, $0 \leq \zeta \leq 1$, define

$$y_{j\ell}^* \equiv y_{j0}^0 \quad \text{if} \quad U_{j\ell} \leq \zeta \tag{3.1.1}$$

and

$$y_{j\ell}^* \equiv y_{j\ell}^0 \quad \text{if} \quad U_{j\ell} > \zeta, \tag{3.1.2}$$

$\ell = 1, 2, \ldots m_j$.

It can be shown that within the j-th cluster,

$$\mathrm{E}(y_{j\ell}^*) = \underset{\sim}{\pi}_j^*, \tag{3.1.3}$$

$$\mathrm{Cov}\left(y_{j\ell}^*, y_{jt}^*\right) = \Delta\left(\underset{\sim}{\pi}_j^*\right) \quad \text{if} \quad \ell = t, \tag{3.1.4}$$

and

$$\mathrm{Cov}\left(y_{j\ell}^*, y_{jt}^*\right) = \zeta^2\,\Delta\left(\underset{\sim}{\pi}_j^*\right) \quad \text{if} \quad \ell \neq t. \tag{3.1.5}$$

Therefore, given x_j, the random vector $t_j = \sum_{\ell=1}^{m_j} y_{j\ell}^*$ does not have a multinomial distribution. Instead

$$\mathrm{E}\left(m_j^{-1}\,t_j\right) = \underset{\sim}{\pi}_j^* \tag{3.1.6}$$

and

$$\mathrm{Var}\left(m_j^{-1}\,t_j\right) = \left[1 + \zeta^2\,(m_j - 1)\right] m_j^{-1}\,\Delta\left(\underset{\sim}{\pi}_j^*\right), \tag{3.1.7}$$

where ζ^2 represents the intra-cluster correlation. Furthermore, if the m_j's are constant, *i.e.*, $m_j = m$, the factor $\phi = [1 + \zeta^2(m - 1)]$ corresponds to the design effect defined by Kish (1965, p.258). An estimate of the design effect ϕ is

$$\hat{\phi} = (dk)^{-1} \left[\sum_{\ell=1}^{dk} \hat{a}_{(i,i)} / \hat{h}^{(i,i)} \right] \bar{w}^{-1}, \tag{3.1.8}$$

where $\hat{a}_{(i,i)}$ and $\hat{h}^{(i,i)}$ represent the (i,i)-th elements of $\hat{A}_n$ in (2.13)-(2.14) and $[H_n(\hat{\beta}_{\text{PSEUDO}})]^{-1}$, respectively, and $\bar{w}$ is the average of the sampling weights for the entire sample.

Under this sampling scheme, data $(\underset{\sim}{x}_j, \underset{\sim}{y}_{j\ell}^{*})$, $j = 1, 2, \ldots, n$, $\ell = 1, 2, \ldots, m$, were generated with $k = 4$, $d = 3$, $m = 21$, and parameters

$$\underset{\sim}{\mu} = (1, -2, 1, 5)', \tag{3.1.9}$$

$$\underset{\sim}{\Sigma} = \text{Diag}(0, 25, 25, 25), \tag{3.1.10}$$

$$\underset{\sim}{\beta}_1^0 = (-0.3, -0.1, 0.1, 0.2), \tag{3.1.11}$$

$$\underset{\sim}{\beta}_2^0 = (0.2, -0.2, -0.2, 0.1), \tag{3.1.12}$$

and

$$\underset{\sim}{\beta}_3^0 = (-0.1, 0.3, -0.3, 0.1). \tag{3.1.13}$$

Based on (3.1.9)-(3.1.13), 1000 sets of samples with n clusters of size m, were generated according to (3.1.1)-(3.1.2) for different values of n, ζ^2, and ϕ. The estimated Type I errors obtained from comparing the F-tests of H_0: $\underset{\sim}{\beta} = \underset{\sim}{\beta}^0$ against $F(12, \infty; 0.05) = 1.753$ were computed under the three different estimation procedures: MLE, CPLX and TAYLOR. A measure of the distortion of the estimated Type I errors relative to the nominal 0.05 is the relative bias which is defined as

$$(0.05)^{-1} \mid \text{Estimated Type I error - } 0.05 \mid. \tag{3.1.14}$$

Relative biases of the estimated Type I errors are reported in Table 3.1. For data generated with no intra-class correlation, ($\zeta^2 = 0$) the MLE procedure, as it is expected, provides small relative bias of the estimated nominal 5% level. CPLX produces in this case relative biases slightly greater than MLE.This is the penalty of estimating extra parameters in (2.13-2.14).

The MLE procedure shows a strong distortion of the estimated Type I error when a positive intra-class correlation is present. This distortion increases as the intra-class correlation ζ^2 gets bigger. In the case where $\zeta^2 = 0.15$ ($\phi = 4$) the relative bias of the estimated Type I error is about 18 indicating an inflated Type I error of about 95%. For the CPLX procedure, the

Table 3.1

Relative Bias of the Estimated Type I Error for the *F*-test of H_0: $\underset{\sim}{\beta} = \underset{\sim}{\beta}^0$ with nominal 0.05 Level under Sampling Scheme I

			Procedure		
n	ζ^2	ϕ	MLE	CPLX	TAYLOR
20	0.00	1	0.24	0.60	16.42
20	0.05	2	9.66	3.68	17.06
20	0.10	3	15.24	3.98	17.44
20	0.15	4	17.74	4.00	17.70
30	0.00	1	0.08	0.06	12.82
30	0.05	2	9.84	1.20	13.74
30	0.10	3	15.52	1.76	14.22
30	0.15	4	17.74	1.86	14.68
40	0.00	1	0.04	0.32	9.66
40	0.05	2	9.98	0.82	9.62
40	0.10	3	16.20	1.02	11.66
40	0.15	4	17.74	1.80	11.66
50	0.00	1	0.06	0.50	7.40
50	0.05	2	9.76	1.44	8.38
50	0.10	3	16.00	1.96	9.32
50	0.15	4	17.80	2.20	9.70
100	0.00	1	0.06	0.90	2.68
100	0.05	2	10.02	1.66	3.90
100	0.10	3	16.26	2.06	4.70
100	0.15	4	17.78	2.24	5.10
200	0.00	1	0.02	0.74	1.28
200	0.05	2	10.46	1.00	1.64
200	0.10	3	16.30	0.88	1.88
200	0.15	4	18.00	1.52	2.12
400	0.00	1	0.02	0.44	0.70
400	0.05	2	10.14	0.66	0.90
400	0.10	3	16.56	0.64	1.00
400	0.15	4	17.86	0.56	0.84
800	0.00	1	0.08	0.32	0.40
800	0.05	2	10.36	0.22	0.36
800	0.10	3	16.04	0.68	0.80
800	0.15	4	18.12	0.50	0.54

relative bias decreases as the sample size increases from $n = 20$ to the cutting point of correction (2.14) which is 34 in this case. Then it slightly increases as the sample size approaches $n = 100$ and then decreases as the sample size keeps getting bigger. This pattern will be observed throughout the whole simulation. It represents the effect of the correction (2.13-2.14) in small samples.

The Taylor procedure has large relative biases when the sample sizes are small. It varies from 17 to 7 for sample sizes between $n = 20$ and $n = 50$. For large samples both methods CPLX and TAYLOR, provide as expected, similar results. In general, the CPLX shows relative biases smaller than the TAYLOR method.

If the F statistics used for testing H_0: $\underset{\sim}{\beta} = \underset{\sim}{\beta}^0$ are multiplied by the number of parameters being tested, the resulting statistic is distributed as a chi-square random variable with 12 degrees of freedom. The Monte Carlo means and variances for these chi-square statistics are presented in Table 3.2.

As expected, the MLE method produces means and variances around 12 and 24, respectively, when the design effect ϕ is one. CPLX has in this case means around 12 with greater variances that decrease when the sample size gets bigger. However, in the presence of any intraclass correlation, the means and variances under MLE are too large, while CPLX shows consistency with the asymptotic theory and the correction introduced in (2.13-2.14). The TAYLOR method has extremely high variances when the sample size is small. A possible explanation for this is that in some replications of the simulation the covariance matrix (2.10) was ill- conditioned producing very large quadratic forms for (2.11). This problem attenuates when the sample size is bigger. Both methods, CPLX and TAYLOR, become asymptotic equivalent for large samples.

Monte Carlo properties for the estimator (3.1.8) of the design effect are presented in Table 3.3 for both CPLX and TAYLOR methods. The CPLX procedure shows smaller biases and slightly large standard errors. Both methods perform fairly well.

For each category r, $r = 1, 2, 3$ and each covariate s, $s = 1, 2, 3, 4$, "t" statistics for the individual coefficient estimates were also computed as

$$\text{“t”} = \left[\text{Var}\left(\hat{\underset{\sim}{\beta}}_{rs}\right)\right]^{-0.5}\left(\hat{\underset{\sim}{\beta}}_{rs} - \underset{\sim}{\beta}^0_{rs}\right). \tag{3.1.15}$$

The twelve "t"statistics provided by the CPLX estimation procedure were grouped together and the simulated percentiles were computed. Similar computations were performed for the MLE "t" statistics. Consequently, for each run the percentiles are based on 12,000 "t" values. Once these percentiles were calculated, the relative biases were estimated as

$$(\text{Standard Normal Percentile})^{-1}|\ \text{Estimated Percentile} - \text{Standard Normal Percentile}\ |. \tag{3.1.16}$$

The results of the relative bias for the estimated 5th and 95th percentiles for the "t" statistics are presented in Table 3.4 for both MLE and CPLX procedures. Under the MLE it is expected that these relative biases be close to $\phi^{0.5} - 1$. This is true because the "t" statistics under MLE are inflated by the factor $\phi^{0.5}$. This is clearly seen in Table 3.4 under the two columns for the MLE percentiles. The CPLX procedure has satisfactory relative biases for small sample. These biases become negligible, as expected, when the sample sizes get bigger.

Table 3.2

Monte Carlo Properties of the Chi-square Statistic of H_0: $\underset{\sim}{\beta} = \underset{\sim}{\beta}^0$ under Sampling Scheme I

			Procedure					
			MLE		CPLX		TAYLOR	
n	ζ^2	ϕ	Mean	Variance	Mean	Variance	Mean	Variance
20	0.00	1	11.5	22.2	12.0	32.7	81.9	12x10^{3}
20	0.05	2	23.9	134.3	16.5	81.2	116.6	8x10^{4}
20	0.10	3	34.2	239.9	16.6	77.8	94.5	12x10^{3}
20	0.15	4	43.8	403.2	17.3	89.3	140.3	19x10^{4}
30	0.00	1	11.8	25.1	11.2	28.5	35.1	702.3
30	0.05	2	23.8	121.4	13.2	41.2	34.1	691.6
30	0.10	3	35.8	268.1	13.8	46.3	41.2	12x10^{2}
30	0.15	4	46.7	450.1	14.1	51.1	44.5	16x10^{2}
40	0.00	1	12.2	24.3	11.9	30.3	25.8	268.3
40	0.05	2	23.2	96.5	12.6	33.6	25.4	201.4
40	0.10	3	35.4	247.7	13.5	43.3	29.1	340.4
40	0.15	4	46.2	428.9	13.8	44.4	30.2	331.4
50	0.00	1	11.9	25.5	12.4	34.6	21.0	140.8
50	0.05	2	23.9	112.5	13.7	43.8	22.7	153.6
50	0.10	3	35.8	231.0	14.3	46.0	24.6	195.8
50	0.15	4	46.7	424.0	14.5	55.4	25.2	234.6
100	0.00	1	12.1	23.6	13.2	35.0	15.8	55.0
100	0.05	2	23.9	102.6	13.8	39.2	16.5	62.1
100	0.10	3	36.5	233.9	14.6	47.0	17.6	75.8
100	0.15	4	47.5	350.4	14.6	43.0	17.9	70.6
200	0.00	1	11.7	24.1	12.6	32.4	13.6	38.2
200	0.05	2	23.9	93.9	13.1	33.1	14.1	39.1
200	0.10	3	35.7	194.1	13.3	31.5	14.3	37.4
200	0.15	4	48.0	399.6	13.5	35.7	14.6	42.7
400	0.00	1	11.9	24.9	12.3	29.3	12.7	31.3
400	0.05	2	24.1	96.6	12.7	29.2	13.1	31.3
400	0.10	3	36.9	208.5	13.1	29.2	13.6	31.4
400	0.15	4	47.3	390.7	12.7	31.6	13.1	34.0
800	0.00	1	11.9	24.0	12.1	26.4	12.3	27.2
800	0.05	2	24.0	99.3	12.3	27.3	12.5	28.2
800	0.10	3	36.4	239.3	12.6	30.1	12.8	31.1
800	0.15	4	48.7	396.3	12.6	26.7	12.7	27.5

Table 3.3

Monte Carlo Properties of $\hat{\phi}$ under Sampling Scheme I

			Procedure			
			CPLX		TAYLOR	
n	ζ^2	ϕ	Rel. Bias	S.E.	Rel. Bias	S.E.
20	0.00	1	0.28	0.23	0.23	0.22
20	0.05	2	0.01	0.63	0.35	0.48
20	0.10	3	0.07	0.93	0.40	0.70
20	0.15	4	0.15	1.15	0.46	0.85
30	0.00	1	0.33	0.22	0.17	0.20
30	0.05	2	0.14	0.62	0.25	0.47
30	0.10	3	0.08	0.88	0.30	0.66
30	0.15	4	0.04	1.18	0.33	0.90
40	0.00	1	0.26	0.18	0.14	0.18
40	0.05	2	0.14	0.53	0.19	0.42
40	0.10	3	0.10	0.83	0.22	0.67
40	0.15	4	0.07	1.13	0.25	0.91
50	0.00	1	0.18	0.18	0.11	0.17
50	0.05	2	0.09	0.48	0.16	0.41
50	0.10	3	0.07	0.75	0.18	0.64
50	0.15	4	0.04	0.97	0.21	0.83
100	0.00	1	0.07	0.13	0.06	0.13
100	0.05	2	0.04	0.34	0.08	0.32
100	0.10	3	0.01	0.54	0.10	0.51
100	0.15	4	0.01	0.69	0.11	0.65
200	0.00	1	0.03	0.10	0.03	0.09
200	0.05	2	0.02	0.25	0.04	0.24
200	0.10	3	0.01	0.38	0.05	0.36
200	0.15	4	0.01	0.49	0.05	0.48
400	0.00	1	0.01	0.07	0.01	0.07
400	0.05	2	0.01	0.19	0.02	0.19
400	0.10	3	0.00	0.27	0.02	0.27
400	0.15	4	0.00	0.37	0.02	0.37
800	0.00	1	0.01	0.05	0.01	0.05
800	0.05	2	0.00	0.13	0.01	0.13
800	0.10	3	0.00	0.19	0.01	0.18
800	0.15	4	0.00	0.24	0.01	0.24

Table 3.4

Relative Bias of the Estimated 5th and 95th Percentiles for the "*t*" Statistics for the Coefficient Estimates under Sampling Scheme I

			Procedure			
			MLE Percentile		CPLX Percentile	
n	ζ^2	$\phi^{0.5} - 1$	5th	95th	5th	95th
20	0.00	0.00	0.02	0.00	0.10	0.09
20	0.05	0.41	0.40	0.38	0.04	0.02
20	0.10	0.73	0.68	0.65	0.07	0.04
20	0.15	1.00	0.84	0.79	0.07	0.04
30	0.00	0.00	0.00	0.02	0.10	0.09
30	0.05	0.41	0.43	0.38	0.01	0.02
30	0.10	0.73	0.73	0.70	0.02	0.01
30	0.15	1.00	0.97	0.91	0.01	0.01
40	0.00	0.00	0.01	0.01	0.07	0.08
40	0.05	0.41	0.38	0.41	0.03	0.02
40	0.10	0.73	0.70	0.72	0.03	0.01
40	0.15	1.00	0.96	0.93	0.01	0.03
50	0.00	0.00	0.01	0.01	0.05	0.07
50	0.05	0.41	0.43	0.40	0.00	0.01
50	0.10	0.73	0.71	0.70	0.01	0.00
50	0.15	1.00	0.97	0.96	0.02	0.01
100	0.00	0.00	0.00	0.02	0.01	0.00
100	0.05	0.41	0.42	0.42	0.02	0.01
100	0.10	0.73	0.71	0.74	0.01	0.03
100	0.15	1.00	1.03	0.99	0.04	0.04
200	0.00	0.00	0.01	0.01	0.00	0.00
200	0.05	0.41	0.42	0.43	0.01	0.01
200	0.10	0.73	0.71	0.72	0.01	0.01
200	0.15	1.00	1.00	1.00	0.02	0.02
400	0.00	0.00	0.01	0.01	0.01	0.01
400	0.05	0.41	0.39	0.40	0.01	0.00
400	0.10	0.73	0.76	0.77	0.03	0.04
400	0.15	1.00	1.02	0.89	0.02	0.00
800	0.00	0.00	0.00	0.01	0.00	0.01
800	0.05	0.41	0.43	0.44	0.01	0.02
800	0.10	0.73	0.76	0.70	0.02	0.01
800	0.15	1.00	1.07	1.04	0.04	0.02

3.2 Sampling Scheme II

Let $x_1, x_2, \ldots, x_n$ be a set of k-dimensional independent and identically distributed normal random vectors with vector mean $\underset{\sim}{\mu}$ and covariance matrix $\underset{\sim}{\Sigma}_B$. These vectors x represent cluster means for the explanatory variables in the logistic function (2.1). Suppose that for the j-th cluster, $j = 1, 2, \ldots, n$, $x^0_{j0}, x^0_{j1}, \ldots, x^0_{j,m_j}$ are independent and identically distributed normal random vectors with vector mean x_j and covariance matrix $\underset{\sim}{\Sigma}_W$. Given $x^0_{j\ell}$, $\ell = 0, 1, \ldots, m_j$, the $(d + 1)$-dimensional random vector $y^0_{j\ell}$ has a multinomial distribution with parameters $(\underset{\sim}{\pi}^0_{j\ell}, 1)$, where the elements of $\underset{\sim}{\pi}^0_{j\ell}$ satisfy the logistic function (2.1) evaluated at the true parameter vector $\underset{\sim}{\beta}^0$ and at $x = x^0_{j\ell}$. Furthermore, suppose that given the $x^0_{j\ell}$'s, the $y^0_{j\ell}$'s are independent.

Let $U_{j1}, U_{j2}, \ldots, U_{j,m_j}$ be m_j independent and identically distributed uniform (0,1) random variables that are also jointly independent from the $x^0_{j\ell}$'s and from the $y^0_{j\ell}$'s. Let ζ be a fixed and known number, $0 \le \zeta \le 1$. Then define $(x_{j\ell}, y^*_{j\ell})$, $\ell = 1, 2, \ldots, m_j$ in the following way:

$$\left(x_{j\ell}, y^*_{j\ell}\right) \equiv \left(x^0_{j0}, y^0_{j0}\right) \text{ if } U_{j\ell} \le \zeta \tag{3.2.1}$$

and

$$\left(x_{j\ell}, y^*_{j\ell}\right) \equiv \left(x^0_{j\ell}, y^0_{j\ell}\right) \text{ if } U_{j\ell} > \zeta. \tag{3.2.2}$$

Observe that within each cluster, the $x_{j\ell}$'s all have the same vector of conditional means x_j and that the covariance matrix between $x_{j\ell}$ and x_{jt} is $\underset{\sim}{\Sigma}_W$ if $\ell = t$ and $\zeta^2\, \underset{\sim}{\Sigma}_W$ otherwise. Also, note that the conditional mean of each $y^*_{j\ell}$ is the logistic function (2.1) evaluated at $\underset{\sim}{\beta}^0$ and $x = x_{j\ell}$, and that the vectors $(x_{j\ell}, y^*_{j\ell})$, $\ell = 1, 2, \ldots, m_j$, exhibit an intra-class correlation of ζ^2 and an approximate design effect of $\phi = [1 + \zeta^2\,(m - 1)]$ when all the m_j's are constant.

Data $(x_{j\ell}, y^*_{j\ell})$, $j = 1, 2, \ldots, n$, $\ell = 1, 2, \ldots, m_j$, were generated under this cluster sampling scheme with $k=4$, $d=3$, and parameters

$$\underset{\sim}{\mu} = \left(1, -6, 4, 8\right)', \tag{3.2.3}$$

$$\underset{\sim}{\Sigma}_B = \text{Diag}\left(0, 25, 25, 49\right), \tag{3.2.4}$$

$$\underset{\sim}{\Sigma}_W = \text{Diag}\left(0, 25, 36, 36\right), \tag{3.2.5}$$

$$\underset{\sim}{\beta}^0_1 = \left(0.30, -0.05, -0.06, 0.08\right), \tag{3.2.6}$$

$$\underset{\sim}{\beta}^0_2 = \left(0.06, -0.08, -0.10, 0.07\right), \tag{3.2.7}$$

and

$$\underset{\sim}{\beta}^0_3 = \left(0.70, -0.08, -0.10, 0.11\right), \tag{3.2.8}$$

Based on (3.2.3)–(3.2.8), 1000 sets of samples with n clusters of size $m_j = m = 6$, were generated according to (3.2.1)–(3.2.2) for different values of n, ζ^2 and ϕ. The relative biases defined in (3.1.14) of the estimated Type I errors from comparing the F-tests of H_0: $\underset{\sim}{\beta} = \underset{\sim}{\beta}^0$ against $F(12, \infty; 0.05) = 1.753$ are presented in Table 3.5 under three different estimation techniques: MLE, CPLX and TAYLOR.

In the presence of intra-class correlation, there is a strong distortion of the Type I error for MLE even in the case where ζ^2 is relatively small ($\zeta^2 = 0.2$) for cluster size $m = 6$. This distortion is reflected in the relative bias which ranges from approximately 7 to 18. These values indicate inflated Type I errors between 40% and 95%. The CPLX procedure provides satisfactory relative biases even for the case of small samples. The TAYLOR procedure has too high values for small samples. It becomes equivalent to CPLX for large samples. One more time CPLX seems to be superior to TAYLOR when the sample size is small.

Table 3.5

Relative Bias of the Estimated Type I Error for the F-test of H_0: $\underset{\sim}{\beta} = \underset{\sim}{\beta}^0$ with Nominal 0.05 Level under Sampling Scheme II

			Procedure		
n	ζ^2	ϕ	MLE	CPLX	TAYLOR
20	0.0	1	0.54	0.46	13.52
20	0.2	2	7.30	0.46	12.96
20	0.4	3	13.70	0.68	13.96
20	0.6	4	17.08	0.60	14.72
30	0.0	1	0.28	0.78	7.78
30	0.2	2	8.72	0.72	8.16
30	0.4	3	14.84	0.72	9.32
30	0.6	4	17.50	0.82	9.23
40	0.0	1	0.36	0.56	5.16
40	0.2	2	9.28	0.56	5.76
40	0.4	3	15.38	0.64	5.84
40	0.6	4	17.76	0.70	5.80
50	0.0	1	0.44	0.56	3.44
50	0.2	2	9.34	0.08	4.86
50	0.4	3	15.48	0.38	4.36
50	0.6	4	17.56	0.46	4.16
100	0.0	1	0.16	0.04	1.26
100	0.2	2	9.46	0.26	1.46
100	0.4	3	15.94	0.44	2.00
100	0.6	4	18.16	0.14	1.46
200	0.0	1	0.10	0.26	0.76
200	0.2	2	10.20	0.34	0.82
200	0.4	3	16.22	0.02	0.48
200	0.6	4	18.06	0.06	0.52

Table 3.6

Monte Carlo Properties of the Chi-square Statistic of H_0: $\underset{\sim}{\beta} = \underset{\sim}{\beta}^0$ under Sampling Scheme II

			Procedure					
			MLE		CPLX		TAYLOR	
n	ζ^2	f	Mean	Variance	Mean	Variance	Mean	Variance
20	0.0	1	11.3	18.9	10.2	19.7	40.5	15x10^{2}
20	0.2	2	20.3	62.8	10.5	21.4	39.2	11x10^{2}
20	0.4	3	28.3	106.4	10.5	18.4	111.3	42x10^{5}
20	0.6	4	35.2	152.6	10.3	18.2	11x10^{3}	50x10^{9}
30	0.0	1	11.6	21.6	9.4	16.3	22.0	147.3
30	0.2	2	21.8	75.2	9.9	17.5	22.7	161.2
30	0.4	3	30.4	117.6	9.8	16.5	24.3	224.6
30	0.6	4	39.3	191.0	9.5	14.5	24x10^{2}	60x10^{8}
40	0.0	1	11.6	21.3	9.9	19.4	18.1	86.7
40	0.2	2	22.4	76.5	10.4	18.3	18.9	80.8
40	0.4	3	31.8	153.2	10.2	17.8	19.2	90.4
40	0.6	4	41.4	223.1	10.1	16.9	19.3	104.4
50	0.0	1	11.5	19.9	10.6	20.0	16.1	56.9
50	0.2	2	22.7	80.6	11.4	23.9	17.5	70.9
50	0.4	3	32.3	160.1	11.1	22.9	17.4	73.7
50	0.6	4	41.7	262.3	10.7	19.7	17.0	63.8
100	0.0	1	11.8	21.5	11.8	25.2	13.9	36.2
100	0.2	2	22.9	87.3	11.9	27.0	14.0	38.5
100	0.4	3	34.7	191.8	12.3	27.9	14.4	40.7
100	0.6	4	45.1	297.7	12.0	25.0	14.1	37.2
200	0.0	1	12.0	23.8	12.1	26.3	13.0	30.3
200	0.2	2	24.0	88.6	12.4	25.9	13.3	30.0
200	0.4	3	34.5	175.2	12.0	23.3	12.8	27.0
200	0.6	4	46.8	320.0	12.2	24.0	13.0	27.9

Monte Carlo properties of the chi-square statistics of H_0: $\underset{\sim}{\beta} = \underset{\sim}{\beta}^0$ (chi-square = 12 × F) are presented in Table 3.6 for the three estimation procedures under study. CPLX shows means and variances slightly below 12 and 24, respectively, when the sample sizes are small. This underestimation vanishes when the sample size increases. The TAYLOR procedure has too large means and variances when the sample size is small. For instance, for $\zeta^2 = 0.6$, the variance is in the order of billions when n is 30 or less. For large samples, both CPLX and TAYLOR, seem to provide similar results. The MLE method has acceptable results only when $\zeta^2 = 0.00$. Otherwise the estimated mean and variances are too large.

Table 3.7

Monte Carlo Properties of $\hat{\phi}$ under Sampling Scheme II

			Procedure			
			CPLX		TAYLOR	
n	ζ^2	ϕ	Rel. Bias	S.E.	Rel. Bias	S.E.
20	0.0	1	0.48	0.22	0.04	0.20
20	0.2	2	0.16	0.53	0.26	0.42
20	0.4	3	0.05	0.87	0.34	0.72
20	0.6	4	0.01	1.24	0.39	1.03
30	0.0	1	0.49	0.18	0.02	0.16
30	0.2	2	0.25	0.48	0.19	0.40
30	0.4	3	0.19	0.84	0.24	0.69
30	0.6	4	0.16	1.12	0.27	0.94
40	0.0	1	0.38	0.16	0.02	0.14
40	0.2	2	0.22	0.45	0.14	0.38
40	0.4	3	0.16	0.70	0.20	0.60
40	0.6	4	0.16	0.98	0.19	0.86
50	0.0	1	0.27	0.14	0.02	0.13
50	0.2	2	0.15	0.42	0.12	0.37
50	0.4	3	0.12	0.67	0.15	0.60
50	0.6	4	0.11	0.89	0.16	0.81
100	0.0	1	0.12	0.10	0.01	0.10
100	0.2	2	0.06	0.32	0.07	0.31
100	0.4	3	0.05	0.50	0.07	0.48
100	0.6	4	0.06	0.59	0.07	0.57
200	0.0	1	0.05	0.07	0.01	0.07
200	0.2	2	0.03	0.24	0.03	0.23
200	0.4	3	0.02	0.34	0.04	0.33
200	0.6	4	0.02	0.40	0.03	0.40

Monte Carlo properties for the estimator of the design effect proposed in (3.1.8) are presented in Table 3.7 under the CPLX and TAYLOR procedures. The TAYLOR procedure seems to perform slightly better than CPLX for small samples. Both procedures, in general, provide reasonable values. They seem to be equivalent for large samples.

Table 3.8

Relative Bias of the Estimated 5th and 95th Percentiles for the "*t*" Statistics for the Coefficient Estimates under Sampling Scheme II

			Procedure			
			MLE Percentile		CPLX Percentile	
n	ζ^2	$\phi^{0.5} - 1$	5th	95th	5th	95th
20	0.0	0.00	0.01	0.00	0.15	0.18
20	0.2	0.41	0.37	0.32	0.06	0.09
20	0.4	0.73	0.63	0.57	0.02	0.05
20	0.6	1.00	0.79	0.74	0.05	0.05
30	0.0	0.00	0.02	0.00	0.15	0.16
30	0.2	0.41	0.39	0.38	0.10	0.10
30	0.4	0.73	0.68	0.63	0.07	0.08
30	0.6	1.00	0.91	0.86	0.05	0.07
40	0.0	0.00	0.01	0.00	0.12	0.15
40	0.2	0.41	0.39	0.40	0.10	0.06
40	0.4	0.73	0.65	0.60	0.07	0.09
40	0.6	1.00	0.99	0.89	0.04	0.05
50	0.0	0.00	0.01	0.01	0.10	0.10
50	0.2	0.41	0.39	0.40	0.05	0.04
50	0.4	0.73	0.73	0.72	0.02	0.01
50	0.6	1.00	1.00	0.95	0.00	0.01
100	0.0	0.00	0.01	0.01	0.04	0.05
100	0.2	0.41	0.40	0.37	0.02	0.02
100	0.4	0.73	0.72	0.73	0.00	0.00
100	0.6	1.00	1.00	1.02	0.01	0.02
200	0.0	0.00	0.02	0.01	0.00	0.01
200	0.2	0.41	0.40	0.45	0.01	0.02
200	0.4	0.73	0.71	0.68	0.01	0.01
200	0.6	1.00	1.03	0.95	0.02	0.02

The relative biases (3.1.16) of the 5th and 95th percentiles of the "*t*" statistics (3.1.15) are presented in Table 3.8 under the MLE and CPLX procedures. MLE has a relative bias, as expected, close to zero in the absence of intra-class correlation. This bias increases when the ζ^2 gets bigger. On the other hand, CPLX has small relative bias in general and for large sample this bias becomes negligible.

4. EXTENSION TO STRATIFIED SAMPLING AND MORE COMPLEX DESIGNS

A generalization of CPLX procedure to stratified sampling can be done as follows. Suppose that the population has been divided into $i = 1, 2, \ldots, L$ strata. Let m_{ij} represent the size of the j-th cluster in the i-th stratum, n_i the number of clusters selected in the i-th stratum, and $y^*_{ij\ell}$ the multinomial response of the ℓ-th element in the j-th cluster in the i-th stratum, $\ell = 1, 2, \ldots, m_{ij}, j = 1, 2, \ldots, n_i, i = 1, 2, \ldots, L$. It is assumed that $\underset{\sim}{\pi}^*_{ij\ell}$, the expected value of $y^*_{ij\ell}$, satisfies the logistic relationship (2.1) for a given explanatory vector $x_{ij\ell}$.

A consistent estimator of $\underset{\sim}{\beta}^0$, say $\underset{\sim}{\hat{\beta}}_{\text{PSEUDO}}$, can be found by maximizing the function

$$L_n(\underset{\sim}{\beta}) = \sum_{i=1}^{L} \sum_{j=1}^{n_i} \sum_{\ell=1}^{m_{ij}} w_{ij}(\log \underset{\sim}{\pi}^*_{ij\ell})' \, y^*_{ij\ell}. \tag{4.1}$$

Algorithm (2.5) is performed with three indexes i, j, ℓ. The adjustment given by (2.13) and (2.14) is applied with

$$n = \sum_{i=1}^{L} n_i, \tag{4.2}$$

$$H_n(\underset{\sim}{\hat{\beta}}_{\text{PSEUDO}}) = \sum_{i=1}^{L} \sum_{j=1}^{n_i} \sum_{\ell=1}^{m_{ij}} w_{ij} \, \Delta(\hat{\underset{\sim}{\pi}}^*_{ij\ell}) \otimes x'_{ij\ell} x_{ij\ell}, \tag{4.3}$$

$$\hat{G} = \left[(n^* - k)^{-1} (n^* - 1)\right] \sum_{i=1}^{L} (n_i - 1)^{-1} n_i(1 - f_i) \sum_{j=1}^{n_i} (\hat{d}_{ij} - \hat{\bar{d}}_i)(\hat{d}_{ij} - \hat{\bar{d}}_i)', \tag{4.4}$$

$$\hat{d}_{ij} = \sum_{\ell=1}^{m_{ij}} w_{ij}(y_{ij\ell} - \hat{\underset{\sim}{\pi}}_{ij\ell}) \otimes x'_{ij\ell}, \tag{4.5}$$

$$\hat{\bar{d}}_i = n_i^{-1} \sum_{j=1}^{n_i} \hat{d}_{ij}, \tag{4.6}$$

$$f_i = \text{sampling rate of } i\text{-th stratum, and} \tag{4.7}$$

$$n^* = \sum_{i=1}^{L} \sum_{j=1}^{n_i} m_{ij}. \tag{4.8}$$

The estimation procedure can be extended in a stepwise manner to multi-stage sampling designs by maximizing (4.1) up to elemental units. The summation of (4.3) should be extended in order to include all the final sampling units. The key part is (4.4). The construction of $\hat{G}$ must be based on the complex survey. This could be a difficult task for multi-stage sampling. Results for stratified two-stage sampling are presented in Fuller, *et al.* (1986, p. 82).

5. SUMMARY

In this paper, we have outlined a methodology for obtaining asymptotic normal estimators of the parameters of a generalized logistic function involving a multinomial response variable under complex survey designs. A consistent estimator of the asymptotic covariance matrix under the complex sampling design is (2.10), which results from the usual Taylor's series expansion. This covariance matrix produces for large samples correct Type I errors for the F-tests involving model parameters. More important, it is shown that correction (2.13-2.14) provides a covariance matrix that reduces the small sample bias. This adjusted covariance matrix has some important characteristics:

1. It levels off the inflated Type I error, originated from ignoring the complex survey, faster than the usual delta-method.
2. It is positive definite when $H_n(\hat{\underset{\sim}{\beta}}_{\mathrm{PSEUDO}})$ is positive definite regardless if (2.9) is singular or not.
3. It is asymptotic equivalent to (2.10).

The results of a Monte Carlo study were reported in Section 3. Data satisfying the logistic conditional mean (2.1) were generated under two different single-stage cluster sampling schemes. It was studied, among other things, the effect of the intra-class correlation and the design effect on the relative biases of the estimated Type I errors for the F-tests of H_0: $\underset{\sim}{\beta} = \underset{\sim}{\beta}^0$. The simulation showed, as expected, a strong relative bias when the naive maximum likelihood method is employed. For small samples, the Monte Carlo results favor the use of the adjusted covariance matrix over the one that arises from the usual delta-method.

ACKNOWLEDGEMENTS

This work was begun while the author was a student at Iowa State University. I thank Professor Wayne A. Fuller for introducing me to the topic and for suggesting a number of the small sample modifications that were incorporated into the estimation procedure. The author wants also to thank the referees for useful comments.

REFERENCES

ALBERT, A., and LESAFFRE, E. (1986). Multiple group logistic discrimination. *Computers and Mathematics with Applications*, 12A, 209-224.

BEDRICK, E.J. (1983). Adjusted chi-square tests for cross-classified tables of survey data. *Biometrika*, 70, 591-595.

BINDER, D.A. (1983). On the variance of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51, 279-292.

BINDER, D.A., GRATTON, M.A., HIDIROGLOU, M.A., KUMAR, S., and RAO, J.N.K. (1984). Analysis of categorical data from surveys with complex designs: some Canadian experiences. *Survey Methodology*, 10, 141-156.

BULL, S.B., and PEDERSON, L.L. (1987). Variance for polychotomous logistic regression using complex survey data. *Proceedings of the Section on Survey Research Methods, American Statistical Association*.

CHAMBLESS, L.E., and BOYLE, K.E. (1985). Maximum likelihood methods for complex sample data: Logistic regression and discrete proportional hazards models. *Communications in Statistics, Theory and Methods*, 14, 1377-1392.

COX, D.R. (1970). *The Analysis of Binary Data*. London: Methuen.

DALE, J.R. (1986). Asymptotic normality of goodness-of-fit statistics for sparse product multinomials. *Journal of the Royal Statistical Society*, Ser. B, 48, 48-59.

FAY, R.E. (1985). A jackknife chi-squared test for complex samples. *Journal of the American Statistical Association*, 80, 148-157.

FULLER, W.A., KENNEDY, W., SCHNELL, D., SULLIVAN, G., and PARK, H.J. (1986). *PC CARP*. Statistical Laboratory, Iowa State University, Ames, Iowa.

GALLANT, A.R. (1987). *Nonlinear Statistical Methods*. New York: John Wiley & Sons.

HABERMAN, S.J. (1974). *The Analysis of Frequency Data*. Chicago: The University of Chicago Press.

HOLT, D., SCOTT, A.J., and EWINGS, P.D. (1980). Chi-squared tests with survey data. *Journal of the Royal Statistical Society*, Ser. A, 143, 303-320.

JENNRICH, R.I., and MOORE, R.H. (1975). Maximum likelihood estimation by means of nonlinear least squares. *Proceedings of the Section on Statistical Computing, American Statistical Association*.

KISH, L. (1965). *Survey Sampling*. New York: John Wiley & Sons.

MOORE, D.S. (1977). Generalized inverses, Wald's method, and the construction of chi-squared tests of fit. *Journal of the American Statistical Association*, 72, 131-137.

MOREL, J. (1987). Multivariate nonlinear models for vectors of proportions: A generalized least squares approach. Unpublished Ph.D. dissertation. Iowa State University, Ames, Iowa.

NELDER, J.A., and WEDDERBURN, R.W.M. (1972). Generalized linear models. *Journal of the Royal Statistical Society*, Ser. A, 135, 370-384.

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. *The Annals of Statistics*, 12, 46-60.

ROBERTS, G., RAO, J.N.K., and KUMAR, S. (1987). Logistic regression analysis of sample survey data. *Biometrika*, 74, 1-12.

SCHNELL, D., KENNEDY, W.J., SULLIVAN, G., PARK, H.J., and FULLER, W.A. (1988). Personal computer variance software for complex surveys. *Survey Methodology*, 14, 59-69.

Survey Methodology, December 1989
Vol. 15, No. 2, pp. 225–235
Statistics Canada

Randomized Response Sampling from Dichotomous Populations with Continuous Randomization

LeROY A. FRANKLIN[1]

ABSTRACT

A randomized response model for sampling from dichotomous populations is developed in this paper. The model permits the use of continuous randomization and multiple trials per respondent. The special case of randomization with normal distributions is considered, and a computer simulation of such a sampling procedure is presented as an initial exploration into the effects such a scheme has on the amount of information in the sample. A portable electronic device is discussed which would implement the presented model. The results of a study taken, using the electronic randomizing device, is presented. The results show that randomized response sampling is a superior technique to direct questioning for at least some sensitive questions.

KEY WORDS: Randomized response; Randomization with continuous distributions; Computer simulation.

1. INTRODUCTION

Surveys often seek to estimate the proportion of individuals satisfying a particular condition. If the condition involves a highly personal or controversial subject (*e.g.*, seeking new employment, sexual behavior) or of an illegal nature (*e.g.* drug usage, criminal activities), survey respondents may be reluctant to answer honestly or may refuse to answer a direct question as to whether they satisfy the condition of interest. In such cases, it is difficult to make inferences about proportions on the basis of a survey in which sensitive questions are asked directly.

Randomized response sampling plans utilize a stochastic or randomizing device to enable respondents to provide answers to sensitive questions without fully revealing information regarding the sensitive issue. The actual outcome of the device for a particular respondent is observed by the respondent but not by the interviewer. However, the properties of the device are known to the experimenter, and this enables the experimenter to make inferences about the proportion of interest without knowing specifically about any single individual. The stochastic device introduces noise into the information-gathering process, but the resulting loss of information may be preferable to the uncontrollable noise introduced by nonresponse or lying when direct questions are used.

The original randomized response model was proposed by Warner (1965) and involved a dichotomous randomization for a dichotomous population. His model was studied from a Bayesian viewpoint in Winkler and Franklin (1979). The randomized response model with two or more trials per respondent was introduced by Gould, Shah and Abernathy (1969) and further developed by Liu and Chow (1976). Both papers demonstrated the superiority of the multiple trials per respondent in improving the efficiency of the estimate over the single trial model of Warner's. However, both also note that multiple trials might produce simultaneously

[1] Dr. LeRoy A. Franklin, Department of System and Decision Sciences, Indiana State University, School of Business, Terre Haute, Indiana 47809.

growing suspicion and lowered "truth telling" over the single trial model. The survey paper prepared by Horvitz, Greenberg, and Abernathy (1976) discusses several other plans with discrete randomization devices. In addition a thorough theoretical development and review of results is contained in the recent volume by Chaudhuri and Mukerjee (1988) entitled "Randomized Response: Theory and Techniques." A more general model, using either discrete or continuous randomization, is presented in Warner (1971) and these more general models were discussed from a Bayesian viewpoint by Pitz (1980), Smouse (1984), and O'Hagen (1987). A few surveys have actually been undertaken, some showing the randomized response methods are superior to direct survey methods (*e.g.* Gould *et al.* 1969 and Liu and Chow 1976) and a few others of uncertain results (*e.g.* Brewer 1981). However, only Poole (1974) developed a specific continuous randomization distribution (uniform) to estimate a continuous distribution and this was implemented by having respondents report their answer multiplied by a number chosen randomly from a random number table.

In this paper, we consider a randomized response model for sampling from a dichotomous population, but using a continuous randomization distribution. With Warner's original randomized response technique, the randomizing device determines which question the respondent answers. But with the method developed in this paper, the question for a respondent is fixed by whether or not he belongs to the sensitive group. The randomization here chooses values from two distributions (one for "yes" and the other for "no") and the respondent provides the value appropriate to his group membership. Multiple trials are incorporated into the model by having the respondent provide a single multi-digit response. This provides a potential benefit over usual multiple trial techniques in that the respondent perceives he/she has provided just one answer when in fact the multi-digit response incorporates several trials of the respondent.

The general model, for which the randomization can be handled via any type of distribution, is presented in Section 2. The special case in which the randomization involves normal distributions is discussed in Section 3, along with an approximating procedure for assessing the effect of randomization and multiple trials per respondent. Section 4 presents a computer simulation investigating the role that specific choices of means and standard deviations play in the efficiency of surveying by using normal distribution randomization with multiple trials. Section 5 presents a way of implementing normal distributions as the randomizing distribution through the use of a computerized, electronic device that generates and displays random normal values. Such a device was felt to be potentially superior to "drawing cards" or "flipping a spinner" since these methods may not be properly implemented by the respondent or the interviewer. The results of a survey taken using that electronic device to investigate five sensitive questions are examined in Section 6. Finally, a summary and a brief discussion of design issues are considered in Section 7.

2. THE MODEL

Suppose that we are interested in θ, the proportion of individuals belonging to Group A among the members of a particular population. A simple random sample of n individuals is chosen from the population with $n \geq 1$, where we assume that the population is large enough relative to n so that the sampling process can be viewed effectively as sampling with replacement. A total of k trials are conducted with each respondent, where $k \geq 1$. On trial j for respondent i, random values are drawn from the distribution functions G_{ij} and H_{ij}. The respondent sees both values and is asked to report the value from G_{ij} if he or she belongs to Group A and the value from H_{ij} otherwise. The researcher knows the exact form of G_{ij} and H_{ij} but sees only the value reported by the respondent, denoted by z_{ij}, and, thus, does not know from which distribution it came.

Inferences must be made about θ based on the kn sample observations z_{ij}, with $i = 1, \ldots, n$ and $j = 1, \ldots, k$. For convenience, we assume in the remainder of this paper that G_{ij} and H_{ij} are absolutely continuous with corresponding densities g_{ij} and h_{ij}; the development for the discrete case is analogous. The conditional density function of z_{ij} given θ is $\theta \, g_{ij} \, (z_{ij}) + (1 - \theta) \, h_{ij} \, (z_{ij})$, and the likelihood function for the entire experiment is:

$$L(\underset{\sim}{z} \mid \theta) = \prod_{i=1}^{n} \left[\theta \prod_{j=1}^{k} g_{ij} \, (z_{ij}) + (1 - \theta) \prod_{j=1}^{k} h_{ij} \, (z_{ij}) \right] \text{ for } 0 \leq \theta \leq 1, \qquad (2.1)$$

where $\underset{\sim}{z} = (\underset{\sim}{z}_1, \ldots, \underset{\sim}{z}_n)$ and $\underset{\sim}{z}_i = (z_{i1}, \ldots, z_{ik})$.

Expanding the likelihood function using the binomial theorem allows the likelihood function to be written in the form

$$L(\underset{\sim}{z} \mid \theta) = \sum_{t=0}^{n} \alpha_t \, \theta^t \, (1 - \theta)^{n-t} \text{ where } 0 \leq \theta \leq 1 \text{ and} \qquad (2.2)$$

$$\alpha_t = \sum_{s=1}^{c} \left[\prod_{i \in C_{ts}} \prod_{j=1}^{k} g_{ij} \, (z_{ij}) \right] \left[\prod_{i \notin C_{ts}} \prod_{j=1}^{k} h_{ij} \, (z_{ij}) \right], \text{ with} \qquad (2.3)$$

$C_{t1}, \ldots, C_{tc}$ representing the $c = \binom{n}{t}$ combinations of t items out of n. Here $\theta^t (1 - \theta)^{n-t}$ is the Bernoulli likelihood conditional upon exactly t respondents being in Group A, and α_t is the likelihood of $\underset{\sim}{z}$ given t. The mixture form in 2.2 arises because we are unable to observe a specific t in our sample.

A special case of (2.1) arises when we assume that the same randomizing distributions are used for all n respondents. Thus, $g_{ij} = g_j$ and $h_{ij} = h_j$ for $i = 1 \ldots n$ and thus (2.1) reduces to

$$L(\underset{\sim}{z} \mid \theta) = \prod_{i=1}^{n} \left[\theta \prod_{j=1}^{k} g_i \, (z_{ij}) + (1 - \theta) \prod_{j=1}^{k} h_j \, (z_{ij}) \right] \text{ for } 0 \leq \theta \leq 1. \qquad (2.4)$$

Whichever the form, in order to find the maximum likelihood estimates, a direct computer grid search must be made. This is feasible since θ is only a one-dimensional quantity and is restricted to the interval from 0 to 1. This can be easily accomplished by using well-known search techniques applied to the log of the likelihood function. (See, for example, Kennedy and Gentle 1980).

3. RANDOMIZATION WITH NORMAL DISTRIBUTIONS

Although any continuous distribution (*e.g.* Weibull, uniform, *etc.*) can be used as the randomizing distribution in the model discussed in Section 2, in this section only the normal distribution will be examined. Furthermore, suppose that the same randomization distributions are used for all respondents, so that form (2.4) is the appropriate likelihood. Thus, g_j and h_j are normal densities with means μ_{gj} and μ_{hj} and standard deviations σ_{gj} and σ_{hj}, respectively. Then the likelihood function in Section 2 can be related to these normal densities.

The amount of information that can be obtained about θ obviously depends on the means and standard deviations that are chosen. At one extreme, if $\mu_{gj} = \mu_{hj}$ and $\sigma_{gj} = \sigma_{hj}$ for $j = 1, \ldots, k$, then θ drops out of the likelihood function and $\underset{\sim}{z}$ (the sample) will provide no information about θ. At the other extreme, if $\mid \mu_{gj} - \mu_{hj} \mid \rightarrow \infty$ for any j with σ_{gj} and σ_{hj} fixed or if $\sigma_{gj} \rightarrow 0$ and $\sigma_{hj} \rightarrow 0$ for any j with a fixed $\mid \mu_{gj} - \mu_{hj} \mid \neq 0$, then we are effectively able to determine which group each respondent belongs to and the sampling process thus approaches Bernoulli sampling in θ.

An approximation to $L(\underset{\sim}{z} \mid \theta)$ as developed by Winkler and Franklin (1979) makes it easier to assess the effect of randomization and multiple trials with the choice of specific means and standard deviations. That is, for each sample, we can approximate the actual likelihood function given by (2.4) with an approximate likelihood function of the form

$$L^*(r^*, n^* \mid \theta) = \theta^{r^*} (1 - \theta)^{n^*-r^*}. \tag{3.1}$$

Taking the first and second derivations of the log of the approximating likelihood (3.1) and solving to find the maximum $(\hat{\theta})$ and the curvature at that maximum yields:

$$\hat{\theta} = \frac{r^*}{n^*} \tag{3.2}$$

and

$$\left[\frac{\partial^2 \log L^*(r^*, n^* \mid \theta)}{\partial \theta^2}\right]_{\theta = \hat{\theta}} = -\frac{n^*}{\hat{\theta}(1 - \hat{\theta})}. \tag{3.3}$$

Next taking the first derivative of the log of the exact likelihood (2.4) and setting it to equal zero gives the equation that will yield the exact maximum likelihood estimate for θ:

$$\sum_{i=1}^{n} \frac{\gamma_i - \eta_i}{\theta\gamma_i + (1-\theta)\eta_i} = 0 \text{ where } \gamma_i = \prod_{j=1}^{k} g_j(z_{ij}),\ \eta_i = \prod_{j=1}^{k} h_j(z_{ij}). \tag{3.4}$$

A grid search produces for (3.4) its solution $(\hat{\theta}_r)$. Taking the second derivative of the log of the exact likelihood (2.4) yields:

$$\left[\frac{\partial^2 \log L(\underset{\sim}{z} \mid \theta)}{\partial \theta^2}\right] = -\sum_{i=1}^{n} \frac{[\gamma_i - \eta_i]^2}{[\theta\gamma_i + (1 - \theta)\eta_i]^2}. \tag{3.5}$$

Substituting $\hat{\theta}_r$ into (3.5) gives the curvature of the actual log likelihood at $\hat{\theta}_r$ (the maximum). Equations (3.2) and (3.3) are two equations in two unknowns, r^* and n^*. Setting (3.2) $= \hat{\theta}_r$ and (3.3) $=$ (3.5) allows us to solve for r^* and n^* so that the approximating log likelihood has the same maximum $\hat{\theta} = \hat{\theta}_r$, and curvature at that maximum as does the actual log likelihood. Thus, the randomized response sample outcome of $\underset{\sim}{z}$ can be thought of as approximately equivalent to a non-randomized response sample (*i.e.* regular Bernoulli sampling) with r^* members out of n^* in the sensitive group. In this sense, n^* can be thought of as a rough measure of the amount of information in the randomized response sample which is of size n.

4. A COMPUTER SIMULATED INVESTIGATION OF THE CHOICE OF MEANS AND STANDARD DEVIATIONS

To investigate the impact of a given set of means and standard deviations for the normal randomizing distributions as well as the impact the size of θ and k (the number of trials) has upon r^* and n^* the randomized response sampling process was simulated by generating, via computer, repeated samples from a Bernoulli process with parameter θ and k sets of two-digit responses for each sample. In our simulation, we let $\mu_{gj} = 50$, $\mu_{hj} = 40$, and $\sigma_{gj} = \sigma_{hj} = \sigma$ for $j = 1, \ldots, k$. We considered two values of θ (.10 and .25), two values of σ (6 and 9), three values of n (50, 200, and 500), and three values of k (1, 2, and 3). Such values were chosen since they will register two-digit deviates that would overlap in distribution considerably and provided then a bench mark for later choices in the actual survey environment. For each of the 36 combinations of parameters, we replicated the sampling procedure 25 times. The solutions of r^* and n^* were found numerically for each sample, and the average values of n^* for the 25 replications with each set of parameter values are given in Table 1.

The average values of n^* vary considerably. At the worst extreme, when $\sigma = 9$, $\theta = .10$, and only one trial per respondent is used, n^* tends to be only 10-15 percent of n. On the other hand, when $\sigma = 6$, $\theta = .25$, and three trials are used per respondent, n^* is about 75 percent of n. As expected, the average value of n^* (the effective sample size) increases as n (the number of respondents) increases or as k (the number of trials per respondent) increases. In addition, decreasing σ or increasing θ also leads to a higher n^*.

For each combination of parameters, the mean and variance of $\hat{\theta}$ over the 25 trials were determined. The average values of $\hat{\theta}$ are very close (within 5%) to the corresponding values of θ, and the variance of $\hat{\theta}$ tends to increase as the average n^* decreases and, hence, tends to validate the simulation.

Table 1

Average Values of the Effective Sample Size (n^*) for Various Sample Sizes (n) and the Number of Trials per Respondent (k)

		$\theta = .10$		$\theta = .25$	
n	k	$\sigma = 6$	$\sigma = 9$	$\sigma = 6$	$\sigma = 9$
50	1	16.2	7.0	17.3	9.2
	2	27.3	13.1	30.6	17.8
	3	32.6	18.1	38.2	23.6
200	1	58.3	24.8	79.0	41.2
	2	103.1	49.6	124.4	72.9
	3	136.6	77.7	151.0	97.7
500	1	148.4	59.6	196.9	103.6
	2	261.1	129.3	309.5	181.2
	3	345.8	193.1	375.6	242.7

5. A PORTABLE, COMPUTERIZED RANDOMIZING DEVICE

Randomized-response sampling, using randomization with normal distributions and multiple trials, provides flexibility to the experimenter, who can select means and variances as well as the number of respondents and the number of trials per respondent. However, this flexibility is not of any value, unless the sampling scheme actually can be implemented in practice. The sampling scheme utilizing Bernoulli randomization can be implemented in a number of ways (*e.g.*, with cards or colored beads). However, the scheme developed in this paper requires generation of random normal values by some portable device.

A computerized, electronic device was built around the Intel 8080 microprocessor to generate and display random normal values. Each value is obtained by summing 16 uniformly distributed random numbers and transforming that sum to achieve a normal deviate with the desired mean and standard deviation. From the Central Limit Theorem, the resulting values should be approximately normally distributed, and extensive tests indicate that the values produced by the device do indeed behave like random normal values. This technique was chosen over other possible methods of generating normal deviates due to the simplicity of programming such a method in machine instructions for this specific microprocessor. For more details concerning the generation of the random normal values and the testing of the device, see Franklin (1977), Kennedy and Gentle (1980), as well as Knuth (1969).

The final, resulting device was approximately the size of a cigar box and is easily held in the hand. Power can be supplied either by a battery pack or by an extension cord.

For display purposes, the random normal values are truncated to two digits, and the device is designed to display six such two-digit numbers simultaneously in "windows" of six digits each. One window displays values chosen from g_1, g_2, and g_3 which appears as a single six-digit number in the "Yes" window. The other window displays values chosen from h_1, h_2, and h_3 which also appears as a single six-digit number for "No". The six means and standard deviations are stored permanently in the device, but they can be changed easily by using a small, detachable keyboard.

The actual surveying process is accomplished in the following manner. First, the interviewer asks the respondent a sensitive question about Group A. The respondent then pushes a button to activate the device, and two six-digit numbers appear in the windows within about one quarter of a second. If the respondent is a member of Group A, the number in the first window (the "Yes" window) is reported; otherwise, the number in the second window (the "No" window) is reported. To convince the respondent of the "randomness" of the values, he or she is encouraged to press the button several times and to observe the resulting numbers before the sensitive question is actually asked. Note that although $k = 3$, the respondent perceives a response as a single six-digit number, and we are thus actually obtaining three trials with a single six digit response. Hence, the advantage of multiple trials per respondent is exploited without the usual accompanying disadvantages coming into play.

6. SURVEY RESULTS AND CONCLUSIONS

Two simultaneous, but independent, surveys were conducted on the campus of a large urban university of students enrolled in that university. The first asked five sensitive questions of a respondent by the direct question method. The second asked the same five sensitive questions of a different respondent but using Randomized Response Sampling with continuous randomization implemented by the electronic device presented in the previous section. For the study $k = 3$ and $\mu_{g_1} = \mu_{g_2} = \mu_{g_3} = 40$ and $\mu_{h_1} = \mu_{h_2} = \mu_{h_3} = 50$ with $\sigma_{g_j} = \sigma_{h_j} = 5$ for $j = 1, 2, 3$. These values were chosen in accordance to the finding of the computer simulation discussed in Section 4. A different group of students was systematically selected (one in five) for each of the two surveys from students on the campus and individually interviewed. Each student surveyed was given a brief introduction as to the purpose of the survey and asked if they wished to participate. Less than 10% of all individuals stopped by both survey teams declined to participate. If the individual was willing to participate, he/she was then asked to provide his/her social security number to verify that he/she was, indeed, enrolled in the university. All respondents of both surveys had their social security number checked against an administrative master list of students and those not recorded as enrolled students were eliminated from the study (less then 5 percent of those surveyed).

Requiring their social security number also deliberately injected the element of associating the individual's identity with his responses. For many surveys (*i.e.* telephone, mail-in questionnaires, house-to-house surveys, *etc.*), this is the case and plays a significant role in the willingness of a respondent to answer truthfully. It was felt that it was precisely in such "revealing" circumstances that randomized response sampling can benefit the researcher most. The resulting sample sizes for the direct and randomized response methods were $n_1 = 473$ and $n_2 = 477$. The five sensitive questions were:

Q1 — "Have you ever cheated on an exam here at this university?"
Q2 — "Would you ever cheat on your income tax?"
Q3 — "Would you ever steal from an employer?"
Q4 — "Have you smoked any marijuana in the last 30 days?"
Q5 — "Have you ever participated in a homosexual act?"

All five questions were felt to be sufficiently sensitive so that any gains by randomized response sampling over direct sampling could be easily apparent. In addition, as a final question, the respondents in the randomized response group were asked "Do you think your friends would be more willing to tell the truth if they were asked sensitive questions by this technique?" This was asked in an effort to measure the acceptance and confidence of the person being interviewed that this particular randomized response technique did provide personal protection and anonymity.

The estimates of the proportion of respondents who are in the sensitive group are presented in Table 2 for both direct ($\hat{\theta}_{id}$) and randomized response ($\hat{\theta}_{ir}$) for question i along with the estimate of n_i^* (the effective sample size) for the randomized response method using the method discussed in Section 3. Also is presented the z value of a one-sided test of hypothesis H_o: $\theta_{id} - \theta_{ir} = 0$ vs H_a: $\theta_{id} - \theta_{ir} < 0$, along with the observed p-values. The tests were conducted using n_1 and n_i^* as sample sizes and hence give a much more conservative result than if n_1 and n_2 were utilized.

It is noteworthy that the randomized response method gave a higher estimate of θ for each of the five sensitive questions than the direct survey method. Furthermore, for Questions 1, 2, and 5, the randomization response method gave statistically significantly higher estimates of θ (p-values $< .001$ for all three) than the direct survey method. Hence, there seems to be conclusive evidence that, at least for some sensitive issues, the randomized response method with continuous randomization does provide better estimates of population proportions. It should also be noted that by our choices of μ_{g_j}, μ_{h_j}, σ_{g_j} and σ_{h_j} and $k = 3$ that n_i^* typically was 75 to 85 percent of the original sample size n_2 and thus most of the information was "recovered" by our randomized response method.

Table 2

Estimates of θ and Results of Testing Equality of θ's for Direct and Randomized Response Sampling with Respective Sample Sizes of $n_1 = 473$ and $n_2 = 477$

Question			Effective sample size		
i	$\hat{\theta}_{id}$	$\hat{\theta}_{ir}$	n_i^*	z-value	p-value
1	.0634	.2013	394.5	6.098	<.0001
2	.1797	.2941	408.1	3.997	<.0001
3	.1078	.1207	384.8	.583	.2810
4	.1882	.1942	409.5	.234	.4091
5	.0042	.0355	339.0	3.341	.0004

Furthermore, it is instructive to consider the nonsignificant results for Questions 3 and 4. This information (if the three significant results are ignored) could lead an observer to conclude that randomized response techniques are not particularly advantageous over direct questioning. However, in the light of the three significant differences revealed, this lack of significance perhaps could be interpreted as the question really was not "sensitive enough" to lead to dramatic differences in θ's or even that the question was "so sensitive" that the respondent chose to lie even with the randomized response technique. In addition, Question 1 "Have you ever cheated on an exam?" seemed to the experimenter to be relatively "unsensitive" but in retrospect the answer to this question when tied to the social security number of the respondent (given before the questioning process started) presented a much more threatening circumstance than was initially realized. Thus, perhaps some of the confusion about the efficacy of the randomized response technique is related to the "true sensitivity" of the question for the interviewee as opposed to the "perceived sensitivity" by the interviewer or experimenter. These aspects need further examination.

Finally, 88.9% (424 of the 477) felt "their friends would be more likely to answer truthfully sensitive questions by this randomized response technique." While some reservations may be expressed by the respondents' "desire to please the interviewer," nevertheless, this overwhelming percentage coupled with the significant differences already discussed seem strong evidence that this technique was accepted and felt to be protective of the interviewee.

7. DISCUSSION

The model developed in this paper permits the use of continuous, as well as discrete, randomizing distributions in utilizing randomized response sampling from a dichotomous population. In order to implement the model using randomization with normal distributions, a computerized, electronic device was also developed and discussed. The device is portable, has programmable means and standard deviations for the six normal distributions and provides from a single six digit response, three separate two digit trials. Such a system has both potential advantages and disadvantages over other randomized response techniques.

First, as alluded to in the introduction, a computerized randomizing device could be superior to the standard randomized response methods of "drawing cards" or "flipping a spinner" since these methods may not be properly implemented by either the respondent or the interviewer which would induce uncontrolled error. (See Abernathy, Greenberg and Horvitz (1970) for a discussion of the problems of "insufficient card shuffling" and "card loss" as well as insufficient interviewer training). Since the production of the randomizing values is computerized, the distributional problems that can and have accompanied the use of cards, beads, and spinners are eliminated because the problem of "random selection of values" is taken out of the hands of the interviewer *and* respondent and placed in the "hands" of the computer. If the computerized device fails, it is usually a complete, catastrophic crash of the whole chip which is readily apparent and very, very rare.

The second (and perhaps greatest) advantage is in the ability of the device to present a choice of two numbers each six digits in length from which the respondent chooses to answer "yes" or "no". But what seems to the respondent as a single six digit answer is in fact three separate two digit answers and in effect provides three trials per respondent. Thus, the benefits of multiple trials per respondent are gained but, since the respondent is unaware of the multiple trials format, without the usual accompanying disadvantages (noted by Liu and Chow 1976) coming into play.

In addition, the freedom to choose the six means and six standard deviations provides the experimenter with additional flexibility over standard randomized response techniques. For instance, if it is felt that the differences in the first two digits are most noticeable to respondents, the experimenter can make μ_{h_1} and σ_{h_1} close to (or even equal to) μ_{g_1}, and σ_{g_1}, respectively. Similarly, if the middle two digits might receive the least attention, the experimenter could attempt to gain the most information from these values by separating μ_{h_2} and μ_{g_2} the furthest. It is also possible to wire the displays in other than the obvious manner. For instance, the two digits of the first random normal value could appear as the fifth and second digits of the six digit number instead of the first and second digits. This flexibility in wiring, together with the the choices of parameters should provide a sampling scheme that is quite informative to the researcher without seemingly to threaten the respondent.

It should also be noted that while for this particular microprocessor it was convenient to utilize randomization with normal distributions, several other continuous distributions (*e.g.* uniform, Weibull) or even multi-valued discrete distributions (*e.g.* multinomial or poisson) could have been used. Further investigation into newer microprocessors as well as different randomizing distributions is recommended.

There are, however, some potential disadvantages associated with this particular randomized response technique. The cost of such a randomizing device since it involves a microprocessor is the order of fifteen hundred to two thousand dollars to produce. However, its versatility in wiring and programming would hopefully allow a device to be used in many investigations over several years and thus help to defray its rather high cost.

More difficult to quantify is the respondent's perception of the computerized device and the degree of confidence or suspicion he/she might have about the device. Do respondents fear that the computerized device is somehow "storing" their answer that somehow later can be deciphered to expose them? From the survey results, it seems that greater truth telling was secured by using the computerized randomizing devices over the direct survey method. Nevertheless, further study is recommended to compare this randomized response technique which uses the computerized device with other more standard randomized response techniques.

In practice, several matters are relevant in the consideration of design issues (*i.e.*, the selection of means and standard deviations for the device). In order to gain more information for a given sample size, we should increase $| \mu_{g_j} - \mu_{h_j} |$ and decrease σ_{g_j} and σ_{h_j} for $j = 1, 2, 3$. However, as this is done, it will become clearer to the respondent that, despite the randomization, the response is very revealing concerning the respondent's group membership. As a result, the respondent may not answer honestly or may refuse to answer. Additional study is needed to determine optimal values for choice of means and standard deviations. The results in Table 1 give some indication of the effects of varying a common standard deviation. But from a practical viewpoint, the field survey seemed to indicate that the choice of means separated by two standard deviations was able to both gain the confidence of the respondent and (with the multiple trials) to gain back from 75 to 85 percent of the original sample size without the usual "loss of confidence" that accompanies multiple trial techniques.

In particular, the field trial compared the direct survey techniques with the randomized response using the electronic device discussed with $\mu_{h_j} = 40$ and $\mu_{g_j} = 50$ and $\sigma_{h_j} = \sigma_{g_j} = 5$ for $j = 1, 2, 3$ for the normal, randomizing distributions. Of the five sensitive questions which were asked of the two (independent) groups, the randomized response method provided significantly greater estimates ($p < .001$) than the direct method for three of the questions. In addition, 88.9% of the subjects interviewed by the randomized response technique felt "their friends would be more likely to tell the truth if they were asked sensitive questions by this technique". Thus, it seems that (for at least certain questions), this randomized response sampling technique achieved greater honesty in response than the direct sampling method.

The question of protection of the respondent's privacy needs to be discussed. It is not ethical to tell the respondent that his or her group membership is disguised by the randomization, if, in fact, the disguise is transparent to the researcher (*e.g.* for example, by recording only even numbers for "YES" and only odd numbers for "NO"). With the electronic device that has been discussed, it seems indeed possible to provide true privacy without losing much information. If the means and standard deviations are programmed into the device and are not provided to an interviewer, the interviewer will find it very difficult to discriminate between group members and non-group members in the interviewing process, particularly if the wiring is "scrambled". Thus, the flexibility that enables us to gain information without threatening the respondent also helps to disguise the actual group membership from the interviewer.

ACKNOWLEDGEMENTS

The author wishes to thank the referees and an Associate Editor for their constructive comments.

REFERENCES

ABERNATHY, J.R., GREENBERG, B.G., and HORVITZ, D.G. (1970). Estimates of induced abortion in urban North Carolina. *Demography*, 7, 19-29.

BARNARD, G.A. (1976). Discussion on the invited and contributed papers. *International Statistical Review*, 44, 226.

BREWER, K.R.W. (1981). Estimating marijuana usage using randomized response some parodoxical findings. *Australian Journal of Statistics*, 23, 139-148.

CAMPBELL, C., and JOINER, B.L. (1973). How to get the answer without being sure you've asked the question. *The American Statistician*, 27, 229-231.

CHAUDHURI, A., and MUKERJEE, R. (1988). *Randomized Response: Theory and Techniques*. New York: Marcel Dekker, Inc..

CHOW, L.P., LIU, P.T., and MOSELY, W.H. (1973). A new randomized response technique for study of contemporary social problems. Presented at the 101st Annual Meeting of the American Public Health Association, Statistics Section.

FRANKLIN, L.A. (1977). A Bayesian approach to randomized response sampling. Unpublished doctoral dissertation, Indiana University, Bloomington, IN.

GOULD, A.L., SHAH, B.U., and ABERNATHY, J.R. (1969). Unrelated question randomized response techniques with two trials per respondent. *Proceedings of the Section on Social Statistics American Statistical Association*, 351-359.

HORVITZ, D.G., GREENBERG, B.G., and ABERNATHY, J.R. (1976). Randomized response: A data-gathering device for sensitive questions. *International Statistics Review*, 44, 181-196.

KENNEDY, W.J., and GENTLE, J.E. (1980). *Statistical Computing*. New York: Marcel Dekker, Inc..

KNUTH, D.E. (1969). *Semi Numerical Algorithms*, (Volume 2). New York: Addison Wesley.

LIU, P.T., and CHOW, L.P. (1976). The efficiency of the multiple trial randomized response technique. *Biometrics*, 32, 607-618.

O'HAGAN, A. (1987). Bayes linear estimates for randomized response models. *Journal of the American Statistical Association*, 82, 580-585.

PITZ, G.F. (1980). Bayesian analysis of randomized response models. *Psychological Bulletin*, 87, 209-212.

POOLE, W.K. (1974). Estimation of the distribution function of a continuous type random variable through randomized response. *Journal of the American Statistical Association*, 69, 1002-1005.

SMOUSE, E.P. (1984). A note on Bayesian lease squares inference for finite population models. *Journal of the American Statistical Association*, 79, 390-392.

WARNER, S.L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60, 63-69.

WARNER, S.L. (1971). The linear randomized response model. *Journal of the American Statistical Association*, 66, 884-888.

WINKLER, R.L., and FRANKLIN, L.A. (1979). Warner's randomized response model: A Bayesian approach. *Journal of the American Statistical Association*, 74, 207-214.

Survey Methodology, December 1989
Vol. 15, No. 2, pp. 237–252
Statistics Canada

Small Area Estimates of Proportions Via Empirical Bayes Techniques

BRENDA MacGIBBON[1] and THOMAS J. TOMBERLIN[2]

ABSTRACT

Empirical Bayes techniques are applied to the problem of "small area" estimation of proportions. Such methods have been previously used to advantage in a variety of situations, as described, for example, by Morris (1983). The basic idea here consists of incorporating random effects and nested random effects into models which reflect the complex structure of a multi-stage sample design, as was originally proposed by Dempster and Tomberlin (1980). Estimates of proportions can be obtained, together with associated estimates of uncertainty. These techniques are applied to simulated data in a Monte Carlo study which compares several available techniques for small area estimation.

KEY WORDS: Logistic regression; Random effects models; Bayes estimation; EM algorithm.

1. INTRODUCTION

1.1 The Problem

Complex multi-stage surveys are used to obtain estimates of proportions in many research disciplines (*e.g.*, epidemiology, economics, criminology *etc.*). Not only are estimates for local areas and other special subgroups required, but there is also a need for reliable measures of the accuracy of these estimates. This suggests to us the need for improved methodologies for this estimation problem and related statistical inference.

In addition, the techniques based on the standard normal theory used by Fay and Herriot (1979) to estimate income, a continuous random variable, in small areas are no longer directly applicable to the problem of estimating proportions for discrete outcome variables. Here, it is the logit transform of the proportion, not the proportion itself, that will be modelled in a linear way. This creates the same problems of estimation as in classical statistical logistic regression theory. (See Haberman 1978.) Unfortunately, fewer attempts have been made to solve this obviously more complex problem in small area estimation.

In order to address the problem of inference from a relatively thinly spread complex, multistage survey to small areas or domains not necessarily included in the survey, we have chosen an explicitly model-based approach. This was proposed originally by Dempster and Tomberlin (1980) for the estimation of census undercount from a post-enumeration survey. The methodology uses both a random effects, multiple logistic regression model and empirical Bayes techniques. This directly yields estimates of uncertainty associated with the estimated proportions for small areas via a Bayesian paradigm. This explicitly model-based method differs substantially from the implicitly model-based approach of the synthetic estimation techniques of Gonzalez and Hoza (1976, 1978), Gonzalez and Waksberg (1975), and others.

[1] Brenda MacGibbon, Department of Decision Sciences and MIS, Concordia University, 1455 Blvd de Maisonneuve W., Montréal, Québec H3G 1M8 and Département de mathématiques et d'informatique, Université du Québec à Montréal, C.P. 8888, Suc. "A", Montréal, Québec H3C 3P8.

[2] Thomas J. Tomberlin, Department of Decision Sciences and MIS, Concordia University, 1455 Blvd de Maisonneuve W., Montréal, Québec H3G 1M8.

As a typical complex survey will often be a nested structure of primary sampling units (PSU's), secondary sampling units (SSU's) within PSU's, tertiary sampling units (TSU's) within SSU's and, finally, households within TSU's; the explicitly model-based approach will allow us to take into account the complexitly of the sample design. The purpose of introducing a random effects model is to allow the data to determine, by empirical Bayes techniques, an appropriate compromise between the classical unbiased estimates which depend only on data in the specific local area, and the fixed effects estimates which pool information across areas.

In Section 1.2, a literature review is given and a solution to the problem of estimating proportions for small areas is proposed. The model and its associated estimates are made explicit in Sections 2 and 3 respectively. The results are applied to simulated data in a Monte Carlo study presented in Section 4.

1.2 The Review and a Proposed Solution to the Problem

Because of the growing need for small area statistics in recent years, and because reliable estimates for small areas or subdomains are not usually directly available by classical sample survey methods, several researchers have focused on the problem of small area estimation. This has necessitated the use of explicitly or implicitly model-based methods which allow for "borrowing strength" across small areas in order to increase the effective sample size for estimation, and hence the accuracy of the resulting estimates. Although much of the research in this area has applied linear model techniques and concentrated on the estimation of means or totals, rather than proportions, a discussion of the literature on these estimators and the criteria used to evaluate them can add valuable insight into our problem.

Classical theory dictates that estimators should be design-consistent and, if possible essentially design-unbiased. However these estimators are not always particularly useful when the sample sizes are small.

Gonzalez (1973) described the method of synthetic estimation as follows: "An unbiased estimate is obtained from a sample survey for a large area; when this estimate is used to derive estimates for sub-areas on the assumption that the small areas have the same characteristics as the larger area, we identify these estimates as synthetic estimates." It seems its first reported use was by the U.S. National Center for Health Statistics (1968) for the calculation of state estimates of long and short term disability rates. Various authors subsequently tried to formalize this concept of synthetic estimation, in particular, for means of continuous outcome variables, using both *ad hoc* and model-based approaches. Gonzalez (1973), Gonzalez and Waksberg (1975), Gonzalez and Hoza (1976) and Levy and French (1978) used previous census data to form post-strata which are subsequently used to combine information across small areas under the assumption that the mean response is similar across a section of these areas. Levy (1971), Ericksen (1973, 1974) and O'Hare (1976) employed regression methods in order to incorporate auxiliary information in small area estimation.The accuracy of this method has been evaluated in terms of its average sampling mean squared error over all small areas in a region.

Ericksen (1974) warned that there is no systematic methodology for the assessment of the bias or accuracy of synthetic estimators. Despite these shortcomings, synthetic estimation still remains a potentially powerful and attractive tool. There have been many reported empirical evaluations both on actual and simulated data sets of synthetic estimation in recent years, including Levy (1971), Gonzalez (1973), Gonzalez and Hoza (1978), and Schaible (1979). Several of these types of studies are described in a volume edited by Platek and Singh (1986).

Royall (1970, 1973), using a model-based approach, also considered the problem of estimating totals in finite populations, when auxiliary information is available. He established a probability model of the relationship between the variable of interest and the auxiliary variable and then derived optimal subdomain predictors.

Holt, Smith and Tomberlin (1979) and Laake (1979) applied the predictive approach of Royall to the problem of small area estimation. Laake (1979) found that in contrast to the synthetic approach, where biased estimators are usually obtained without an explicit method of estimating the bias, the prediction approach yielded estimates of mean squared error (MSE) as a tool for the comparison of estimators. In the problem of estimating small area totals, Holt, Smith and Tomberlin (1979) specified various possibilities of population structure in order to model the assumed relationship across subareas. With a specified model, it becomes possible to determine whether or not it is supported by the data and also to study the effect of model misspecification on the bias of the observed estimators. Under different models, the variance of the estimator, the estimate of the variance and MSE change. They built model-based confidence intervals, which have interpretations in terms of repeated realizations under the superpopulation model.

Purcell and Kish (1979, 1980) reviewed the different existing techniques of small area estimation, subdividing them into the following broad categories, regression-based procedures, the use of empirical Bayes and of Bayesian methods, superpopulation prediction theory, clustering techniques, and categorical data analysis methods. They underlined the fact that small area domain estimation should not be considered as a homogeneous problem, but that there exist many other interacting factors such as domain size which should be taken into account when choosing the type of estimator. Särndal (1984) later confirmed this.

The most serious shortcoming of model-dependent estimators is that useful estimates of mean squared errors are not available using fixed effects models because associated variance estimates do not reflect the bias inherent in estimates based on models having a reduced set of parameters. Two different approaches were then taken to the problem of small area estimation.

Fay and Herriot (1979) used the James-Stein theory of estimation (James and Stein 1961) on sample data to determine estimates of income for small places from the 1970 US Census of Population and Housing. In fact, they used an empirical Bayes approach which originated with Robbins (1955) and has been described by Efron and Morris (1975), thus formalizing the meritorious suggestion of Madow and Hansen (1975) of forming a weighted average of the sample and regression estimates. A similar approach by Schaible, *et. al.* (1977) gives a method for arriving at a composite estimator which is the weighted average of the unbiased and synthetic estimators. For other examples of empirical Bayes methods for small area estimation based on standard normal theory see Stroud (1987) and Cressie (1988).

Battese, Harter and Fuller (1988), using a prediction approach, proposed a nested error regression model in order to estimate means. A more general model, a random coefficients regression model, had been previously proposed for a similar problem by Dempster, Rubin and Tsutakawa (1981). They used Bayesian techniques to estimate fixed and random effects in covariance component models when the covariances and variances are tentatively assumed to be known and the EM algorithm to subsequently estimate these unknown parameters. The introduction of random effects models not only allows for standard maximum likelihood estimation, but also provides measures of the reliability of the final estimates of the parameters in the form of posterior variances.

Ericksen (1980) suggested using the mean squared error (MSE) to evaluate effectiveness of regression in small area estimation. He attempted to answer such questions as: When should more predictor variables be added to the regression equation? Should James-Stein weighting procedures be used when the synthetic and the regression estimate are far apart? He also warned of the effects of outliers on both the resulting estimate and its estimated error. Perhaps the effect on small area estimators of the failure of the linear model assumptions should be more seriously studied.

Although applied to the estimation of counts such as unemployment and mortality statistics, most of these techniques described were designed primarily for continuous outcome variables. Purcell and Kish (1980) introduced a categorical data analysis method for obtaining estimates of counts for small domains. Essentially, their methodology involves fitting log-linear models to the data, omitting some of the higher order interaction terms and obtaining estimates by the iterative proportional fitting algorithm described by Deming and Stephan (1940). We propose to extend these models to the problem of estimation of proportions in small domains as originally conceived by Dempster and Tomberlin (1980) by applying empirical Bayes techniques to logistic regression models with random effects. This would have the added advantage that a measure of uncertainty of the small area estimates would be available through the approximate posterior variances. The estimator proposed here is similar in nature to the composite one used by Schaible *et. al.* (1977) for unemployment rates, the principal difference being in the method for choosing the weights. We feel, however, that the empirical Bayesian paradigm gives a more natural and intuitive method for determining the weights. Empirical Bayes estimation based on simple logistic random effects has already proven useful in studying regional variation in mortality rates by Miao (1977). Somewhat more complex random effects models have been used for proportions on data from the World Fertility Survey (Wong and Mason, 1985) and for Poisson parameters on automobile insurance data (Weisberg, Tomberlin, and Chatterjee 1984 and Tomberlin 1988).

Roberts, Rao and Kumar (1987) fitted logistic regression models to binary outcome data obtained using complex sampling schemes, constructed "pseudo-maximum likelihood" estimators, and compared their estimates to unbiased ones. They also proposed a goodness-of-fit test for their model, which takes the sampling design into account. A fundamental difference between our approach and that of Roberts, *et. al.*, is that by incorporating the characteristics of the sample design into the model, we can estimate parameters, and obtain readily interpretable measures of their reliability by means of standard maximum likelihood techniques.

2. THE MODEL

Following the framework of Dempster and Tomberlin (1980), in its most general form, we specify a model which describes the probabilities associated with individuals in the population as a function of categorical variables, continuous covariates and sampling characteristics. The models we consider in this paper are specific examples of the following,

$$\text{logit}\ (\ \pi_{\mu\nu}) \ = \ \theta_\mu \ + \ X_{\mu\nu}\ \underset{\sim}{\beta} \ + \ \phi_\nu \tag{2.1}$$

where $\pi_{\mu\nu}$ represents the probability of a "response" for the ν-th unit in the μ-th cell, the subscript μ refers to a set of categorical variable covariates, and the subscript ν refers to a set of nested sampling characteristics, indicating PSU, SSU within PSU, and so on. The parameter θ_μ represents a sum of fixed classification effects, the parameter ϕ_ν represents a sum of random effects associated with sampling characteristics, the vector $X_{\mu\nu}$ represents a vector of quantitative covariates, and the parameter $\underset{\sim}{\beta}$ is a vector of fixed logistic linear regression parameters. The random effects parameters are assumed to have some parametric distribution, usually a multivariate normal distribution. The probabilities $\pi_{\mu\nu}$ are obtained by inverting the logit transformation as follows,

$$\pi_{\mu\nu} = [1 + \exp\{-(\theta_\mu + X_{\mu\nu} + \phi_\nu)\}]^{-1}. \tag{2.2}$$

For purposes of illustration, consider the following simple example. Let the proportion of interest be the labour force participation rate. Suppose we have one classification variable indicating sex and one continuous covariate indicating the age of the individual. Suppose further that the sample design is a simple, two stage cluster sample. In the first stage, a sample of counties is drawn and simple random samples of individuals within selected counties are drawn at the second stage.

For estimation purposes, consider the following model,

$$\text{logit }(\pi_{\mu\nu}) = \theta_\mu + X_{\mu\nu}\beta + \phi_i \tag{2.3}$$

$$\phi_i \sim \text{i.i.d. Normal }(0, \sigma^2). \tag{2.4}$$

Here, the classification subscript, μ, indicates the sex of the individual; the sampling characteristics subscript, $\nu = ij$, indicates the j-th individual within the i-th PSU; $X_{\mu\nu}$ indicates the age of the individual and ϕ_i is a random effect associated with the i-th PSU.

The consequence of assuming that the PSU effects are independent, identically distributed is that PSU departures away from the fixed part of the model are treated as exchangeable; that is, apart from effects of age and sex, no systematic information exists regarding differential employment rates among the counties in the population. Obviously in a realistic situation, such information would exist, for example, dominant industry, distance from principal markets, retail sales, etc. In such cases, this auxiliary information should be incorporated into the model. However, for purposes of illustration, we will continue with this simple model. The choice of a normal distribution of the error terms is a mathematical convenience, and the consequences of this choice must also be evaluated after actual data analysis. Extensions from the simple model described in (2.3-4) to include additional covariates, both categorical and quantitative is straight forward.

In theory, extensions to the model allowing for more complex sample designs is also simple. For example, data drawn using a three stage sample could be modelled using nested random effects as follows.

$$\text{logit }(\pi_{\mu\nu}) = \theta_\mu + X_{\mu\nu}\beta + \phi_i + \phi_{j(i)} \tag{2.5}$$

$$\phi_i \sim \text{Normal }(0, \sigma_1^2)$$

$$\phi_{j(i)} \sim \text{Normal }(0,\sigma_2^2).$$

Here, the sampling characteristics subscript, $\nu = ijk$ refers to the k-th individual within the j-th SSU within the i-th PSU. The parameter ϕ_i is the random effect associated with the i-th PSU, and $\phi_{j(i)}$ is the nested random effect associated with the j-th SSU within the i-th PSU. Stratification variables could also be incorporated within the fixed effects part of the model. While it is simple to write down the models corresponding to sample designs with several stages, without further research, it is not yet clear how difficult it will be to produce estimates based on these more complex models.

In an actual application, it would be necessary to use the data to identify predictor variables. This would require the development of some sort of model selection techniques. While not the primary focus of this paper, one might conceive of such a technique being based on an initial analysis using conventional variable selection techniques for logistic regression models as described by Haberman (1978), for example. Such an analysis could be conducted, ignoring the random effects parameters. Having chosen a set of predictors, the random effects would then be incorporated in the manner dictated by the sample design.

3. ESTIMATES

In this section, we develop empirical Bayes estimates for the simple model described in equations (2.3-4). First, it is assumed that the variance component, σ^2, is known, and Bayesian estimates of the probabilities $\pi_{\mu ij}$ are obtained. Then, the EM algorithm, as described by Dempster, Laird and Rubin (1977), is used to obtain the maximum likelihood estimate of σ^2 allowing for empirical Bayes estimates. Finally, posterior variances of these estimates are obtained. The development of these estimates is similar to that described by Laird (1978) and by Tomberlin (1988).

3.1 Bayes Estimates

As noted by Laird (1978) in her analysis of contingency tables, by Dempster, Rubin and Tsutakawa (1981) in their analysis of variance components for linear models, and by Tomberlin (1988) in his analysis of Poisson data, a Bayesian analysis of a mixed model such as described in (2.3-4) can be obtained by placing a flat prior on the fixed parameters, θ_μ and β and the proper prior given in (2.4) on the random parameters, ϕ_i.

Let the vector of 0-1 outcome variables indicating membership in the labour force be represented by y and let $\underline{\pi}$ represent a vector of the individual probabilities $\pi_{\mu ij}$. The data are then distributed as a product binomial given by,

$$p(y \mid \underline{\pi}) \propto \prod_{\mu ij} \pi_{\mu ij}^{y_{\mu ij}} (1 - \pi_{\mu ij})^{(1 - y_{\mu ij})}. \tag{3.1}$$

The prior distribution of the parameters is given by,

$$p(\underline{\theta}, \underline{\phi}, \beta \mid \sigma^2) \propto \exp\left[- \sum_i \frac{\phi_i^2}{2\sigma^2} \right]. \tag{3.2}$$

Thus, the joint distribution of the data, y, and the parameters is given by,

$$\begin{aligned} p(y, \underline{\theta}, \underline{\phi}, \beta \mid \sigma^2, \mathrm{X}) &= p(y \mid \underline{\theta}, \underline{\phi}, \beta, \sigma^2, \mathrm{X})\, p(\underline{\theta}, \underline{\phi}, \beta \mid \sigma^2, \mathrm{X}) \qquad (3.3) \\ &\propto \left[\prod_{\mu ij} \pi_{\mu ij}^{y_{\mu ij}} (1 - \pi_{\mu ij})^{(1 - y_{\mu ij})} \right] \exp\left[- \sum_i \frac{\phi_i^2}{2\sigma^2} \right]. \end{aligned}$$

From this, the posterior distribution of the parameters is given by,

$$p(\underline{\theta}, \underline{\phi}, \beta \mid y, \sigma^2, \mathrm{X}) \frac{p(y, \underline{\theta}, \underline{\phi}, \beta \mid \sigma^2, \mathrm{X})}{p(y \mid \sigma^2, \mathrm{X})}. \tag{3.4}$$

It is not feasible to obtain a closed form expression for the posterior given in (3.4) due to the intractable integration required to obtain the marginal distribution of y. Here we adopt the approximation employed by Laird (1978) and by Tomberlin (1988). The posterior is expressed as a multivariate normal distribution having its mean at the mode of (3.4) and covariance matrix equal to the inverse of the information matrix evaluated at the mode.

Obtaining the mode requires solving the following set of equations. This can be accomplished by using a multivariate Newton-Raphson algorithm.

$$\sum_{\mu ij} y_{\mu ij} \mathrm{X}_{\mu ij} = \sum_{\mu ij} \hat{\pi}_{\mu ij} \mathrm{X}_{\mu ij} \tag{3.5}$$

$$\sum_{ij} y_{\mu ij} = \sum_{ij} \hat{\pi}_{\mu ii} \tag{3.6}$$

$$\sum_{\mu j} (y_{\mu ij} - \hat{\pi}_{\mu ij}) - \frac{\hat{\phi}_i}{\sigma^2} = 0. \tag{3.7}$$

The posterior covariance matrix of the parameters is found by inverting the negative of the second derivative matrix of the log of (3.4) taken with respect to the parameters, and evaluated at the mode. Note that neither the equations for the mode, nor the covariance matrix involve the intractable denominator of (3.4).

Elements of the inverse of the posterior covariance matrix are given by,

$$\frac{-\partial^2}{\partial \beta^2} = \sum_{\mu ij} \hat{\pi}_{\mu ij} (1 - \hat{\pi}_{\mu ij}) \mathrm{X}^2_{\mu ij} \tag{3.8}$$

$$\frac{-\partial^2}{\partial \theta^2_\mu} = \sum_{ij} \hat{\pi}_{\mu ij} (1 - \hat{\pi}_{\mu ij}) \tag{3.9}$$

$$\frac{-\partial^2}{\partial \phi^2_i} = \sum_{\mu j} \hat{\pi}_{\mu ij} (1 - \hat{\pi}_{\mu ij}) - \frac{1}{\sigma^2} \tag{3.10}$$

$$\frac{-\partial^2}{\partial \beta \, \partial \theta_\mu} = \sum_{ij} \hat{\pi}_{\mu ij} (1 - \hat{\pi}_{\mu ij}) \mathrm{X}_{\mu ij} \tag{3.11}$$

$$\frac{-\partial^2}{\partial \beta \, \partial \phi_i} = \sum_{\mu j} \hat{\pi}_{\mu ij} (1 - \hat{\pi}_{\mu ij}) \mathrm{X}_{\mu ij} \tag{3.12}$$

$$\frac{-\partial^2}{\partial \theta_\mu \, \partial \phi_i} = \sum_{j} \hat{\pi}_{\mu ij} (1 - \hat{\pi}_{\mu i j}). \tag{3.13}$$

3.2 Empirical Bayes Estimates

To obtain empirical Bayes estimates, the prior variance, σ^2, must be estimated from the data. A reliable estimate requires a reasonable number of PSU's in the sample; otherwise, if the number of PSU's is too small, a purely Bayesian approach is recommended. We propose to estimate the prior variance using an EM algorithm as described by Dempster, Laird and Rubin (1977). The general framework for the estimates is similar to that employed by Laird (1978) for contingency table analysis, and Tomberlin (1988) for Poisson data in a two way classification. The estimates for the simple two-stage sample are obtained in exactly the same way as used by Leonard (1988).

The algorithm is initiated by choosing a starting value, $\sigma^2_{(0)}$, for the variance component. The posterior distribution of the random effects, ϕ_i, is then obtained by carrying out a Bayesian analysis as described in Section 2. This posterior distribution is then used to implement the E-step. The expected value of the sufficient statistic is calculated conditional on the data. The M-step is then completed by merely calculating the maximum likelihood function of the sufficient statistics. For a more complete description of the EM algorithm for regular exponential densities, see Dempster, Laird and Rubin (1977). The process is then repeated with a Bayesian analysis based on the updated estimate of the variance component, $\sigma^2_{(1)}$. The algorithm is continued until it converges.

3.3 Estimates of Small Area Proportions

Estimates together with associated posterior variances and covariances for parameters of the model given in (2.3-4) are presented in Sections 3.1 and 3.2. These estimated parameters are then employed to obtain estimates for small area proportions using a predictive approach. Assuming that the sample sizes within each area are small compared to those of the corresponding populations, this can be accomplished by averaging the individual estimated probabilities:

$$\hat{p}_i = \frac{\sum_{\mu j} \hat{\pi}_{\mu i j}}{N_i} \tag{3.14}$$

where N_i is the number of individuals in the i-th small area, and where the estimated probability associated with the μij-th individual, $\hat{\pi}_{\mu ij}$ is obtained by inverting the logistic function as follows,

$$\hat{\pi}_{\mu ij} = [1 + \exp\{-(\hat{\theta}_\mu + X_{\mu ij}\hat{\beta} + \hat{\phi}_i)\}]^{-1}. \tag{3.15}$$

To develop posterior variances for the estimates of small area proportions, it is convenient to adopt a more conventional notation for the linear part of the model, using dummy variables to indicate classifications. Let $Z_{\mu ij}$ represent a vector of predictor variables, both quantitative and qualitative, associated with the μij-th individual and let $\underline{\Gamma}$ represent a vector of the parameters of the model. Then,

$$Z^T_{\mu ij}\, \underline{\Gamma} = \theta_\mu + X_{\mu ij}\, \beta + \phi_i, \tag{3.16}$$

$$\hat{\pi}_{\mu ij} = [1 + \exp(-Z_{\mu ij}^T \hat{\underline{\Gamma}})]^{-1}. \tag{3.17}$$

Then, using a standard Taylor Series method, the posterior variance of the estimated small area proportion can be approximated as,

$$\mathrm{Var}(\hat{p}_i) = \left[\sum_{\mu j} Z_{\mu ij}^T \hat{\pi}_{\mu ij} (1 - \hat{\pi}_{\mu ij}) \right] \frac{\hat{\underline{\Sigma}}_{\Gamma}}{N_i^2} \left[\sum_{\mu j} Z_{\mu ij} \hat{\pi}_{\mu ij} (1 - \hat{\pi}_{\mu ij}) \right]. \tag{3.18}$$

Here, $\hat{\underline{\Sigma}}_{\Gamma}$ is the posterior covariance matrix of the estimated logistic regression parameters $\hat{\underline{\Gamma}}$.

Should the samples within small areas be substantial parts of the associated populations within those areas, then some additional gains in precision could be made by predicting only for the non-sampled units, in the spirit of the finite population sampling prediction methods originally described by Royall (1970).

4. THE SIMULATION STUDY

A simulation study was carried out to illustrate the characteristics of three different methodologies for producing local area estimates of proportions. The three methods evaluated were, the classical unbiased estimates, model-based estimates similar to the straightforward "synthetic estimates" of Gonzalez and Hoza (1978), and a modification of the proposed empirical Bayes estimates described in section 3, above. Data were simulated for a two-stage sample design. The 15 primary sampling units (PSU's) were also treated as the local areas for which individual estimates of labour force participation rates were required. Within each of the 15 PSU's, simple random samples of 25 individuals were drawn, for a total sample size of 375. The local area populations were assumed to be infinite so that complications associated with finite population sampling could be avoided.

As evaluations for local area estimates were required, it was decided to simulate resampling at the second stage only. That is, the same 15 PSU's were drawn for each of the simulation studies. Each replicate consisted of a different sample drawn within these PSU's. The study was based on 205 replications.

The data were generated using the model described in equation (2.3). The parameters were defined as follows,

$$\begin{aligned} \theta_1 &= -0.5 \\ \theta_2 &= -1.0 \\ \beta &= 0.1. \end{aligned} \tag{4.1}$$

The random parameters ϕ_i were generated from a normal distribution having mean zero and standard deviation 0.25. The $\pi_{\mu\nu}$ were obtained by inverting the logistic transformation as given in equation (3.15).

Here, θ_1 and θ_2 are the fixed effects associated with men and women respectively. That is, the odds ratio for labour force participation of men to that of women is $\exp[0.5] = 1.65$. The parameter β is the slope parameter associated with age, and the ϕ_i are the logistic random effects associated with the 15 PSU's, or local areas.

Table 1

Population Labour Force Participation Rates by Local Area

Local Area	1	2	3	4	5	6	7	8
Participation Rate	0.79	0.79	0.96	0.88	0.90	0.95	0.86	0.96
Local Area	9	10	11	12	13	14	15	
Participation Rate	0.61	0.87	0.81	0.91	0.94	0.92	0.83	

The predictor variables, were generated with identical distributions for each of the 15 local areas. Age was distributed uniformly on the interval 20 to 40 years, the sex of each individual was drawn from a Bernoulli distribution with proportion 0.5, and the two predictor variables were assumed to be independently distributed. The population labour force participation rates for the 15 local areas are displayed in Table 1. As each local area was assumed to have the same distribution on the predictor variables, the only source of variation from area to area was the random local area effects, the ϕ_i. The random nature of these effects can produce a substantial variation in local area participation rates as is particularly evidenced by local area 9.

The observed local area sample proportions were used as unbiased estimates. The synthetic estimator was based on the following fixed effects, logit model,

$$\text{logit}\ (\pi_{\mu\nu}) = \theta_\mu \tag{4.2}$$

where, $\pi_{\mu\nu}$ and θ_μ are defined as for the random effects model in (2.3). Notice, only data from a particular local area are used to form the unbiased estimator while data are pooled from all local areas to obtain the synthetic estimator. However, the synthetic estimators will be biased to a degree which depends on the extent that model (4.2) fails to capture differences between local areas.

The third estimator studied here is a modification of the proposed empirical Bayes estimator described in Section 3. Due to the amount of computer time required to estimate the variance component associated with the local area effects, in fact, the Bayes estimator described in Section 3.1 was employed. The prior variance used for these estimates was the known value of the variance given in (4.1) used to simulate the data. As a result of this compromise, the results for the "empirical Bayes" estimator given below would be expected to be somewhat better than those which would be obtained using a true empirical Bayes estimator. However, sensitivity analyses aimed at determining the effect of changes in the prior variance indicate that the results which would be obtained using the empirical Bayes estimator would not be expected to substantially differ from those reported here for the modified estimator.

To look at bias, (in the classical sense of design-based inference) the estimates were averaged over all 205 replicates. Averages for each of the 15 local areas, for each estimation method are presented in Figure 1. The population rates are plotted as the "True Proportions". These rates are almost exactly the same as the average unbiased estimates, and for the most part, are not visible on the graph. This confirms the unbiased nature of the classical estimates.

The synthetic estimates do not vary much from local area to local area. As each local area rate is based on the same pooled, fixed parameter estimates, the only source of variability from

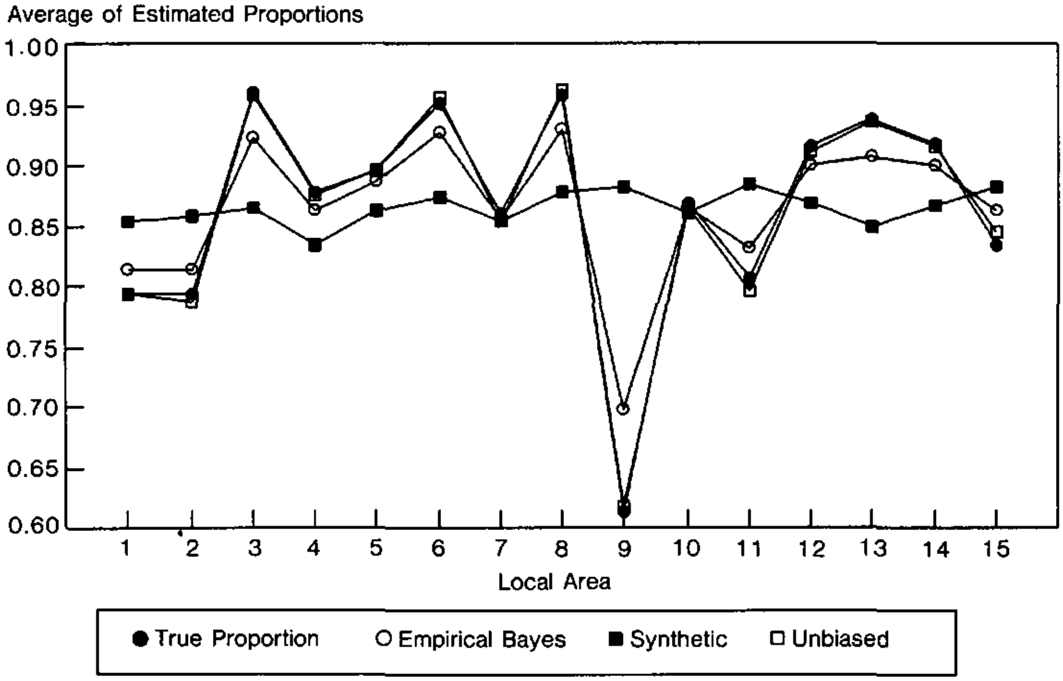

Figure 1. Averages of the estimated labour force participation rates for each of the three estimation methods plotted by local area

local area to local area is the small variability in the realized distributions of the predictor variables. The bias of this estimator can be large, as for example is the case for local area 9, where the synthetic method has a large positive bias. On the other hand, it should be noted that the synthetic method could not be expected to perform very well where there is little variability between the local area distributions of predictor variables.

The averages of the proposed estimates are in between the two extremes of the unbiased and synthetic estimates. They are biased, again in the classical sense, but their biases are smaller than those of the fixed effects model synthetic estimators.

Empirical Root Mean Square Errors (RMSE) were also calculated for each of the three estimators. These are presented in Figure 2. This plot demonstrates graphically where the synthetic estimator performs well and where it performs poorly. For local areas 7 and 10, where the local area effect is close to zero, the expected value of the synthetic estimator is very close to the population proportion. In these areas, the synthetic estimator has by far the smallest RMSE. By pooling data from the whole sample, it obtains a small sampling variance. On the other hand, in local area 9 where the local area effect is quite large, the associated RMSE for the synthetic estimator is also very large, due to its large bias. The modified empirical Bayes estimator obtains most of the reduction in RMSE that results from pooling the data across local areas, without suffering from the large bias associated with the synthetic estimator in those areas with large local area effects. In all but two cases, the modified empirical Bayes estimator achieves a smaller RMSE than the unbiased estimator. For local area 3, the RMSE's for the two estimators are about the same, and for local area 9, with a large local area effect, that of the modified empirical Bayes estimator is somewhat larger than that of the unbiased estimator. In short, the modified empirical Bayes estimator is sometimes the best of the three and never the worst.

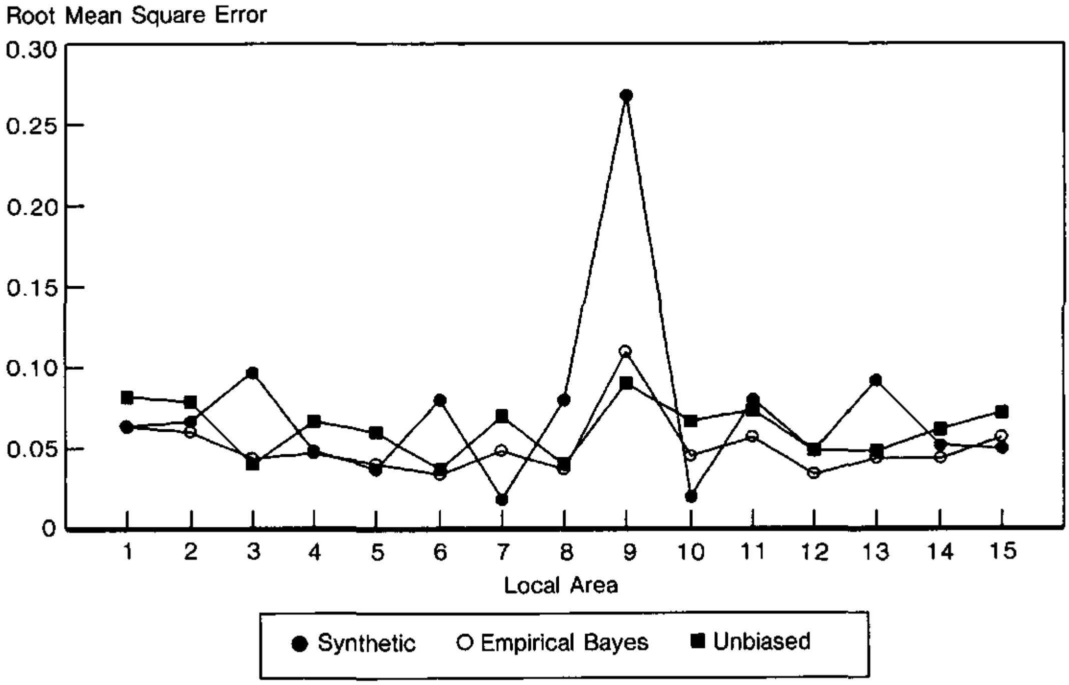

Figure 2. Empirical Root Mean Square Errors associated with each of the three estimation techniques plotted by local area

One of the principal shortcomings of the usual, fixed effects synthetic estimators is the difficulty in obtaining useful measures of associated accuracy. One can only obtain measures of sampling variances. Measures of bias which reflect model inadequacies are not available. For unbiased estimates, on the other hand, the usual estimates of sampling variability are also mean square error estimates as there is no bias. For empirical Bayes estimates, measures of uncertainty are available from the posterior covariance matrix of the parameters. These posterior variances reflect sampling variability as well as the "bias" which comes from simple fixed effects model inadequacies. This latter source of uncertainty is captured via the variability in the local area effects parameters.

The usefulness of these measures of uncertainty are compared graphically in Figure 3. The vertical axis corresponds to the empirical root mean square error (RMSE) which is obtained by comparing the individual replicate estimates with the known population proportions for each local area. The horizontal axis corresponds to the "reported RMSE". For the classical unbiased estimates, these are merely the sampling standard deviations for simple random sampling. For the synthetic estimates, they are also sampling standard deviations, corrected for the cluster sampling. The "reported RMSE" for empirical Bayes estimates are the square roots of the posterior variances of the estimated proportions which were obtained using the methods described in Section 3.2 above.

Note that the points corresponding to the unbiased estimates lie along a line indicating that the reported RMSE's are very close to the empirical RMSE's. This is as expected since there is no bias in these, so the reported RMSE's and the empirical RMSE's are merely sampling standard deviations. As opposed to this, the points corresponding to the synthetic estimates are in a cluster above 0.015 to 0.020 on the horizontal axes. For these estimates,

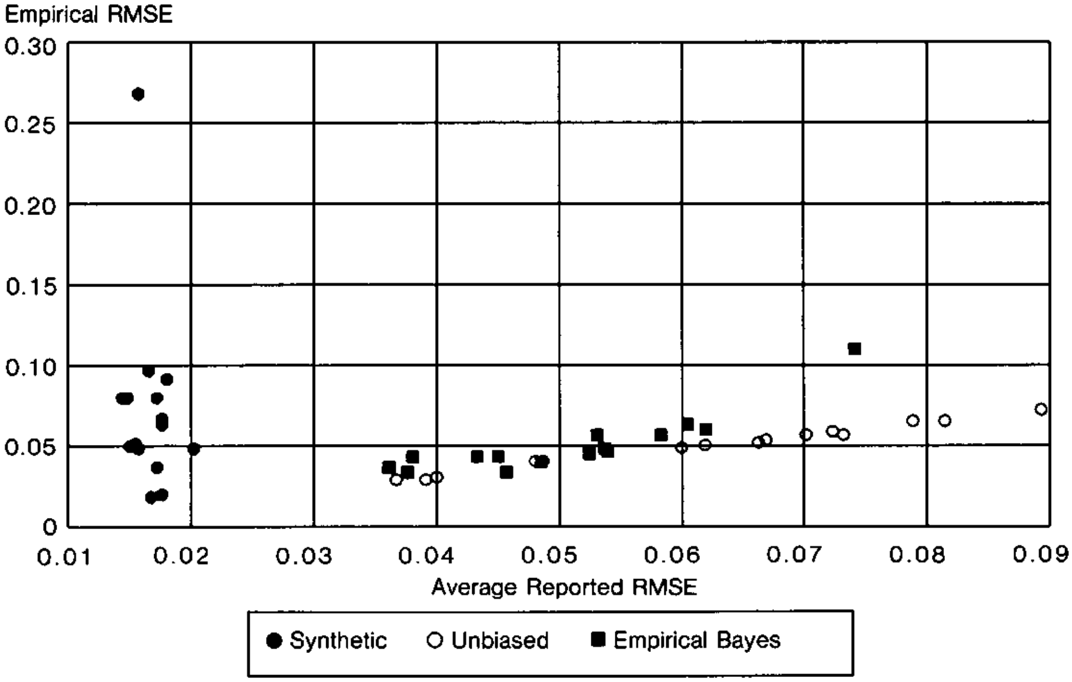

Figure 3. Empirical Mean Square Error vs "Reported Mean Square Errors" for each of the three estimation techniques

the "Reported RMSE's" are estimates of sampling standard deviations, which for these pooled estimates are all quite small. However, the empirical RMSE's for these estimates are quite a different story. They range from 0.015 to 0.100, with one outlier in excess of 0.250 (local area 9). Sampling variances alone are not sufficient to describe the uncertainty associated with the estimates.

The case for the modified empirical Bayes estimators is again in between these two extremes. However, with respect to the relationship between reported RMSE and empirical RMSE it is much closer to the corresponding relationship for the unbiased estimators. With the exception of the point associated with local area 9, the average reported RMSE's are very close to the corresponding empirical RMSE's.

5. CONCLUSIONS

In the simple simulation of a two-stage sample where PSU's correspond to local areas, the modified empirical Bayes estimators have been shown to be superior, overall to two standard alternatives. These have been evaluated in three ways, design-bias, root mean square error, and validity of estimable measures of uncertainty. The classical estimator is shown to be superior in terms of design-bias, as expected since it is design unbiased. In addition, valid estimates of RMSE's are available using standard techniques. However, these estimators suffer from large RMSE's due to the fact that they are formed from limited amounts of data. Indeed, unlike the other two alternatives, no estimates can be formed at all for local areas not in the sample.

At the other extreme, the synthetic estimator is far more stable than either of its competitors. Since all estimates are based on data from the whole sample, associated sampling variances are much smaller than those of the other two estimators. On the other hand, this estimator is unable to adjust for local areas which are quite different from the rest. This is the case, even when data are available in the sample that would indicate such a difference. As important, estimates of uncertainty in the form of sampling standard deviations for this estimator are particularly misleading since they are unable to account for departures from the fixed effects model.

As a compromise between these two estimators, the modified empirical Bayes estimator performs well on all three assessments. By using the data from the specific local areas to the extent it is reliable, this estimator avoids the large biases associated with the synthetic estimator. On the other hand, by pooling information from the whole sample, it has smaller sampling variances than the unbiased estimator, and generally smaller RMSE's. Finally, posterior variances are available as useful measures of uncertainty.

Several tasks remain in the investigation of the proposed estimators. First, the effect of using true empirical Bayes estimators instead of modified ones must be assessed. Some guidelines for minimum number of sampling units for valid empirical Bayes inference are required. True empirical Bayes estimates employ estimated prior variances and methods which account for this additional uncertainty are required. For example, the bootstrap techniques investigated by Laird and Louis (1987) could be used. Second, the estimation techniques need to be generalized to handle three and more stages of sampling. While the theoretical extension is trivial, the computational implications are not. Finally, these techniques must be applied to real data before recommending their adoption as a standard alternative for local area estimation.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the helpful suggestions of an associate editor and a referee, and the financial support of NSERC of Canada, and Concordia University.

REFERENCES

BATTESE, G. E., HARTER, R. M., and FULLER, W. A. (1988). An error-components model for prediction of county crop areas using survey and satellite data, *Journal of the American Statistical Association*, 83, 28-36.

CRESSIE, N. (1988). When are census counts improved by adjustment? *Survey Methodology*, 14, 191-208.

DEMING, W. E., and STEPHAN, F. F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics*, 11, 427-444.

DEMPSTER, A. P., LAIRD, N. M., and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of The Royal Statistical Society*, Ser. B, 39, 1-38.

DEMPSTER, A. P., RUBIN, D. B., and TSUTAKAWA, R.K. (1981). Estimation in covariance components models. *Journal of the American Statistical Association*, 76, 341-353.

DEMPSTER, A. P., and TOMBERLIN, T. J. (1980). The analysis of census undercount from a postenumeration survey. *Proceedings of the Conference on Census Undercount*, 88-94.

ERICKSEN, E. P. (1973). A method for combining sample survey data and symptomatic indicators to obtain population estimates for local areas. *Demography* 10, 137-159.

ERICKSEN, E. P. (1974). A regression method for estimating population changes of local areas. *Journal of the American Statistical Association*, 69, 867-875.

ERICKSEN, E. P. (1980). Can regression be used to estimate local undercount adjustments? *Proceedings of the Conference on Census Undercount*, 55-61.

EFRON, B., and MORRIS, C. (1975). Data analysis using Stein's estimator and its generalizations. *Journal of the American Statistical Association*, 70, 311-319.

FAY, R. E., and HERRIOT, R. A. (1979). Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 269-277.

GONZALEZ, M. E. (1973). Use and evaluation of synthetic estimates. *Proceedings of the Section on Social Statistics, American Statistical Association*, 33-36.

GONZALEZ, M. E., and HOZA, C. (1976). Small area estimation of unemployment. *Proceedings of the Section on Social Statistics, American Statistical Association*, 437-443.

GONZALEZ, M. E., and HOZA, C. (1978). Small area estimation with application to unemployment and housing estimates. *Journal of the American Statistical Association*, 73, 7-15.

GONZALEZ, M. E., and WAKSBERG, J. L. (1975). Estimation of the error of synthetic estimates. Unpublished paper presented at the first meeting of the International Association of Survey Statisticians, Vienna.

HABERMAN, S. J. (1978). *Analysis of Qualitative Data Volume 1: Introductory Topics.* New York: Academic Press.

HOLT, D., SMITH, T. M. F., and TOMBERLIN, T. J. (1979). A model-based approach to estimation for small subgroups of a population. *Journal of the American Statistical Association*, 74, 405-410.

JAMES, W., and STEIN, C. (1961). Estimation with quadratic loss. *Proceedings of the Fourth Berkeley Symposium of Mathematical Statistics and Probability*, Vol. 1, Berkeley: University of California Press, 361-379.

LAAKE, P. (1979). A predictive approach to subdomain estimation in finite populations. *Journal of the American Statistical Association*, 74, 355-358.

LAIRD, N. (1978). Empirical Bayes methods for two-way contingency tables. *Biometrika*, 65, 581-590.

LAIRD, N., and LOUIS, T. (1987). Empirical Bayes confidence intervals based on bootstrap samples (with discussion). *Journal of the American Statistical Association*, 82, 739-757.

LEONARD, K. J. (1988). Credit scoring via linear logistic models with random parameters. Ph. D. Dissertation, Department of Decision Sciences and Management Information Systems, Concordia University, Montréal.

LEVY, P. S., (1971). The use of mortality data in evaluating synthetic estimates. *Proceedings of the Section on Social Statistics, American Statistical Association*, 323-331.

LEVY, P. S., and FRENCH, D. K. (1978). Estimation of health characteristics. *Vital and Health Statistics*, Ser. 2, No. 75, NCHS, Washington, DC.

MADOW, W. G., and HANSEN, M. H. (1975) On statistical models and estimation in sample surveys. Contributed Papers, 40th Session of the International Statistical Institute, Warsaw, Poland, 554-557.

MIAO, L. L. (1977). An empirical Bayes approach to analysis of inter-area variation, Ph. D. Dissertation, Department of Statistics, Harvard University.

MORRIS, C. (1983). Parametric empirical Bayes inference: theory and applications. *Journal of the American Statistical Association*, 78, 47-54.

O'HARE, W. (1976). Report on a multiple regression method for making population estimates. *Demography*, 13, 369-379.

PLATEK, R., and SINGH, M. P. (1986). *Small Area Statistics--An International Symposium '85* (Contributed Papers), Technical Report Series of the Laboratory for Research in Statistics and Probability, Carleton University-University of Ottawa, Canada.

PURCELL, N. J., and KISH, L. (1979). Estimation for small domains. *Biometrics*, 35, 365-384.

PURCELL, N. J., and KISH, L. (1980). Postcensal estimates for local areas (or domains). *Bulletin of the International Statistical Institute*, 48, 3-18.

ROBERTS, G., RAO, J. N. K., and KUMAR, S. (1987). Logistic regression analysis of sample survey data. *Biometrika*, 74, 1-12.

ROBBINS, H. I. (1955). An empirical Bayes approach to statistics. *Proceedings of the Third Berkeley Symposium*. Berkeley: University of California Press, 157-164.

ROYALL, R. M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika*, 57, 377-387.

ROYALL, R. M. (1973). Discussion of papers by Gonzalez and Ericksen. *Proceedings of the Section on Social Statistics, American Statistical Association*, 42-43.

SÄRNDAL, C. E. (1984). Design consistent versus model dependent estimation for small domains. *Journal of the American Statistical Association*, 79, 624-631.

SCHAIBLE, W. L. (1979). A composite estimator for small area statistics. In *Synthetic Estimates for Small Areas* (NIDA Research Monograph 24), edited by J. Steinberg. Rockville, MD: National Institute on Drug Abuse, 36-53.

STROUD, T. W. F. (1987). Bayes and empirical Bayes approaches to small area estimation. In *Small Area Statistics*, (Eds. R. Platek, J. N. K. Rao, C. E. Särndal and M. P. Singh.). New York: Wiley, 124-137.

TOMBERLIN, T. J. (1988). Predicting accident frequencies for drivers classified by two factors. *Journal of the American Statistical Association*, 83, 309-321.

U.S. National Center for Public Health Statistics (1968). *Synthetic State Estimates of Disability*, PHS Publication No. 1759.

WEISBERG, H. I., TOMBERLIN, T. J., and CHATTERJEE, S. (1984). Predicting insurance losses under a cross-classification: a comparison of alternative approaches. *Journal of Business and Economic Statistics*, 2, 170-178.

WONG, G. Y., and MASON, W. M. (1985). The hierarchical logistic regression model for multilevel analysis. *Journal of the American Statistical Association*, 80, 513-524.

Survey Methodology, December 1989
Vol. 15, No. 2, pp. 253–260
Statistics Canada

Updating Size Measures in a PPSWOR Design

ALAN SUNTER[1]

ABSTRACT

It is sometimes required that a PPSWOR sample of first stage units (psu's) in a multistage population survey design be updated to take account of new size measures that have become available for the whole population of such units. However, because of a considerable investment in within-psu mapping, segmentation, listing, enumerator recruitment, *etc.*, we would like to retain the same sample psu's if possible, consistent with the requirement that selection probabilities may now be regarded as being proportional to the new size measures. The method described in this article differs from methods already described in the literature in that it is valid for any sample size and does not require enumeration of all possible samples. Further, it does not require that the old and the new sampling methods be the same and hence it provides a convenient way not only of updating size measures but also of switching to a new sampling method.

KEY WORDS: PPSWOR; Sample updating; PPS sequential sampling.

1. INTRODUCTION

It is sometimes required that a PPSWOR sample of first stage units (psu's) in a multistage population survey design be updated to take account of new size measures that have become available for the whole population of such units. This occurs, for example, when the psu's are census enumeration areas (or collections of census enumeration areas) and a new census has made new population/housing counts available or when, because of observed uneven growth in EA populations in an intercensal period, it is decided to do an interim update of size measures in a sampling stratum. However, because of a considerable investment in within-psu mapping, segmentation, listing, enumerator recruitment, *etc.*, we would like to retain the same sample psu's if possible, consistent with the requirement that selection probabilities, originally proportional to the old size measures, may now be regarded as being proportional to the new ones. A comprehensive treatment of the problem for $n = 1$ is given by Kish and Scott (1971) and is itself a generalization of a method given earlier by Keyfitz (1951). They point out that their method may be extended without difficulty to with replacement sampling (PPSWR) for $n > 1$. Their method may also be used (Drew, Choudhry, and Gray 1978; Platek and Singh 1978) for $n > 1$ when the PPSWOR procedure used is that due to Rao, Hartley and Cochran (1962), since this method involves the formation of n random groups and subsequent selection of a single psu from each group. It breaks down however if we wish, as indeed we probably would, to form new random groups according to the new size measures. Fellegi (1966) provides two methods applicable to a PPSWOR sample of $n = 2$ drawn by the Fellegi (1963) procedure.

The method given in this paper is similar to the second Fellegi method and, when applied to the examples in the Fellegi paper, gives very similar results. Unlike that method, however, it does not require the enumeration of all possible samples and hence is a feasible procedure for any value of n and N. Although it is formally applicable to any PPSWOR method for which it is feasible to calculate the selection probability of any sample selected it has its highest utility for PPSWOR methods in which all, or nearly all, n-tuple subsets are possible samples with

[1] Alan Sunter, President, A.B. Sunter Research Design & Analysis Inc., 63 Fifth Av., Ottawa, Canada, K1S 2M3.

probabilities approximately proportional to the product of their unit probabilities. The method of this type, used for purposes of illustration, is the author's pps sequential method (Sunter 1986, 1989).

2. REPLACEMENT PROCEDURE THEORY

We wish to reselect a PPSWOR sample, originally selected with probabilities $\{\pi_{11}, \pi_{12}, \ldots, \pi_{1n}\}$ proportional to original size measures $\{z_{11}, z_{12}, \ldots, z_{1n}\}$ under a new set of probabilities $\{\pi_{21}, \pi_{22}, \ldots, \pi_{2n}\}$ proportional to new size measures $\{z_{21}, z_{22}, \ldots, z_{2n}\}$. However, we want to do this in such a way that we have a high probability of retaining the original sample.

We assume that for any particular n-tuple S, including of course S', the original sample actually selected, it is possible to calculate both $P_1(S)$, its selection probability under the original scheme, and $P_2(S)$, its selection probability under a new scheme. For many samples in many schemes (*e.g.* pps systematic) one or both of these probabilities may be zero although, obviously, $P_1(S')$ cannot be zero.

The procedure is as follows:

Step 1: (a) Calculate $P_1(S')$, $P_2(S')$.

(b) If $P_2(S') \geq P_1(S')$ then retain the sample.

(c) If $P_2(S') < P_1(S')$ retain the sample with probability $P_2(S')/P_1(S')$. If rejected proceed to Step 2.

Step 2: (a) If the original sample was not retained then draw a new sample, S_1 say, with probability $P_2(S_1)$. If $P_2(S_1) < P_1(S_1)$ then reject the sample, otherwise retain with probability $1 - P_1(S_1)/P_2(S_1)$. If rejected proceed to Step 2(b).

(b) If the Step 2(a) sample was not retained then draw a new sample, S_2 say, and proceed as for Step 2(a).

(c), (d), ... Repeat the Step 2(a), 2(b), ... procedure until a sample is retained.

The sample eventually retained by this process has the required probability structure for both unit probabilities and unit pair joint probabilities. In other words, it may be regarded as having been drawn under the new scheme. In particular, since it has the same joint probability stucture, it has the same sampling variance.

Let P^* denote the probability that the process does not terminate at Step 1, P^{**} the conditional probability that it does not terminate at Step 2(a) given that it did not terminate at Step 1. Obviously P^{**} is then also the conditional probability that the process does not terminate at any subsequent step given that it did not terminate at any step preceding that step. We now have

$$P^* = \sum_{i:P_2(S_i) < P_1(S_i)} (1 - P_2(S_i)/P_1(S_i))P_1(S_i)$$

$$= \sum_{i:P_2(S_i) < P_1(S_i)} (P_1(S_i) - P_2(S_i)) \qquad (1)$$

where i now indexes the n-tuple subsets of the N population units, and

$$\begin{aligned} P^{**} &= 1 - \sum_{i:P_1(S_i) < P_2(S_i)} (1 - P_1(S_i)/P_2(S_i))P_2(S_i) \\ &= 1 - \sum_{i:P_1(S_i) < P_2(S_i)} (P_2(S_i) - P_1(S_i)) \end{aligned} \tag{2}$$

while, since $\sum_i P_1(S_i) = \sum_i P_2(S_i) = 1$, it is easy to see that the summation terms on the right of (1) and (2) respectively must be equal and we have $P^* = 1 - P^{**}$.

Denoting ultimate selection probability by P' we now have, by design:

For $i{:}P_2(S_i) < P_1(S_i)$

$$\begin{aligned} P'(S_i) &= P_1(S_i)\ (P_2(S_i)/P_1(S_i)) \\ &= P_2(S_i), \text{ as required.} \end{aligned}$$

For $i{:}P_2(S_i) \geq P_1(S_i)$

$$\begin{aligned} P'(S_i) &= P_1(S_i) + P^*(1 - P_1(S_i)/P_2(S_i))P_2(S_i) \\ &\quad + P^*P^{**}(1 - P_1(S_i)/P_2(S_i))P_2(S_i) \\ &\quad + P^*(P^{**})^2(1 - P_1(S_i)/P_2(S_i))P_2(S_i) \\ &\quad + P^*(P^{**})^3(1 - P_1(S_i)/P_2(S_i))P_2(S_i) \\ &\quad + \ldots \\ &= P_1(S_i) + P^*(P_2(S_i) - P_1(S_i))/(1 - P^{**}) \\ &= P_2(S_i) \end{aligned}$$

as required.

Finally, we observe that the expected number of Step 2 "trials", given that the original sample was not retained at Step 1, is given by the binomial waiting time distribution as $1/(1 - P^{**}) = 1/P^*$.

3. APPLICATION AND EXAMPLES

The new scheme need not be the same (even apart from the change in unit probabilities) as the old one. We could switch, for example, from a sample originally drawn under pps systematic sampling to one drawn under the author's (Sunter 1986, 1989) pps sequential scheme or even from PPSWR (pps with replacement) to a PPSWOR scheme. In the latter case, of course, an original sample with multiple inclusions of a single psu has zero probability of selection in the new PPSWOR scheme. The procedure may still be used, it may be noted, even if we have included new psu's in the stratum but are retaining the same sample size.

The procedure probably has its highest practical utility, as measured by its probability of retaining the same sample, when both the old and the new schemes are such that all, or nearly all, samples are possible and their probabilities are approximately proportional to the product of their unit selection probabilities. Under these circumstances, and provided that the changes in size measures are not extreme, $P_1(S_i)$ and $P_2(S_i)$ tend to have about the same values so that the probability of retaining the same sample will be relatively high. A practical PPSWOR method with the required properties is the author's, referred to above. Since we will use this method in the examples of the next section, we now describe it. There are two variants, in both of which we have to find a suitable ordering of the population and accumulate the size measures (which we assume to be scaled to sum to 1), in reverse order (so to speak), to give:

$$Z_i = \sum_{i}^{N} z_j; \; i = 1, 2, \ldots, N.$$

Variant 1: Order the population in any way such that

(a) $nz_i \leq Z_i$; $i = 1, 2, ..N - n$

(b) $(n - i)z_i < Z_i$; $i = n, n + 1, \ldots, N - 1$.

Then select units until exactly n have been selected according to:

$$P(U_i \mid n_i) = \begin{cases} 1 \text{ if } n_i = N - i + 1 \\ n_i z_i / Z_i \text{ otherwise} \end{cases}$$

where n_i is the number of sample units still required to be selected when we arrive at the i-th population unit.

It is always possible to satisfy the ordering requirements (a) and (b). For example ordering by increasing size obviously satisfies both as does ordering by decreasing size down to the point (if any) at which (b) fails and then by increasing size. The latter ordering has some advantage in that it tends to minimize the slight (and, for practical purposes, negligible) deviation from strict pps for the last n units (see Sunter 1986). Variant 2 avoids these deviations altogether by taking advantage of the fact that if it occurs that there are $n_i + 1$ units remaining in the population for any i, then it is usually possible to simply discard one of these units with appropriate probability and retain the others.

Variant 2: Order the population in any way such that

(a) $nz_i \leq Z_i$; $i = 1, 2, ..N - n - 1$

(b) $(n - i)z_j < Z_i$; $j \geq i \geq N - n$.

Then

(i) select according to $P(U_i \mid n_i) = nz_i/Z_i$ until either $n_i = 0$ or $n_i = N - i$, then

(ii) if $n_i > 0$ discard one of the remaining units, say that indexed j, with probability $1 - n_i z_j / Z_i$ and select the others.

An algorithm for finding an ordering satisfying the requirements for Variant 2 is given in Sunter(1986) and is incorporated in the program used for the simulations of the next section. In both variants π_{ij} maybe calculated according to

$$\pi_{ij} = n(n - 1)z_i z_j \tau_{ij}$$

where $i < j$ (in the indexing of the ordering actually used) and

$$\tau_1 = 1/Z_2$$

$$\tau_i = (1/Z_i + 1)(1 - z_1/Z_2) \ \ldots \ (1 - z_{i-1}/Z_i).$$

These expressions are exact for $i < j \leq N - n + 1$, and provide a very close approximation otherwise. They are easily calculated and give the method the advantage, unique among practical procedures for PPSWOR with $n > 2$, of the availability of variance estimation with negligible bias.

Pascal-like pseudocode for a routine that selects a sample according to Variant 1, at the same time calculating its probability and the value of τ_i for each selected unit, is given in an Appendix. It is easily extended to Variant 2 or modified to the calculation of P(S) for an already selected sample.

3.1 Example 1

To illustrate these procedures we take first an example with $n = 2$ and $N = 4$, small enough for sample enumeration and manual calculation, where it will be seen that, in order to obtain the "new" size measures, we simply inverted the order of the original assignment. The Variant 2 ordering algorithm mentioned above gives (4,1,2,3) for the first set of size measures and (1,4,3,2) for the second. There are six possible samples, listed in column (1) of Table 2, whose probabilities under the Variant 2 algorithm are easily calculated, with results shown in columns (2) and (3). Column (4) gives the probability of retaining this sample at Step 1, given that it was the original selection. Column (5) gives the conditional probability of retention at any subsequent Step 2, given that no sample was retained at a preceding step.

It may be verified that the overall probability of retention of the same sample, given by the sum of the products of the values in columns (2) and (4), is 0.5465. This value may be compared with the overall probability of retention of the same sample when the new sample is selected independently, given by $\sum_i P_1(S_i)P_2(S_i) = 0.1168$. Thus even in this rather extreme example, we have considerably increased the likelihood of retaining the same sample.

Table 1

Selection Probabilities

PSU	z_{1i}	z_{2i}
1	0.15	0.35
2	0.20	0.30
3	0.30	0.20
4	0.35	0.15

Table 2

(1) Sample	(2) $P_1(S)$	(3) $P_2(S)$	(4) $P_{2\|1}(S)$	(5) $P_{2\|2}(S)$
1,2	0.0231	0.3231	1.0	0.9286
1,3	0.1154	0.2154	1.0	0.4643
1,4	0.1615	0.1615	1.0	0
2,3	0.1615	0.1615	1.0	0
2,4	0.2154	0.1154	0.5357	0
3,4	0.3231	0.0231	0.0715	0

3.2 Example 2

In a more realistic set of examples we now take $n = 4$ for a population of 100 psu's with "original" size measures independently assigned from the uniform or rectangular distribution $R(1,3)$. "New" size measures are assigned in a number of ways, described below. For these examples it is no longer feasible to enumerate all possible samples or to perform the sample selection and sample probability calculations manually. However, writing a computer program to do the latter and to apply the reselection procedure was a straightforward task. The program was used to perform 200 iterations, for each example, of selection of a sample using Sunter's Variant 2 with probabilities proportional to the first set of size measures with subsequent application of the procedures described above for reselection of a sample with probabilities proportional to the second set of size meaures. The program, running on an XT-compatible operating at 7.16 MHz, generated and sorted the populations of size measures and performed 200 iterations of the sample selection and reselection in about three minutes.

Case 1, in which we have assigned new size measures from the same distribution independently of their original values, may be seen as a "worst practical case" scenario. Case 2, in which 10% of the psu's have doubled in size with the rest remaining unchanged, is an approximation of a "scattered development" scenario. Case 3 illustrates the random perturbation of size measures by an amount rectangularly distributed over an interval equal to the original size measure. From Table 3 it may be seen that with probabiliities ranging from 0.67 in the "worst case" scenario to 0.81 in the "scattered development" scenario, we retain the original sample. For those cases in which the original sample is rejected the average number of Step 2 trials required to select a new sample agreed closely with the predicted value of $1/P^*$.

Table 3

200 Iterations of a Size Measure Update Procedure, $n = 4$, $M = 100$;
Original Size Measures from $R(1,3)$

Case	Source of π_{2i}	Step 1 Retentions	Average Step 2 Trials	Estimated P^*
1	$z_{2i} \approx R(1,3)$	134	2.98	0.33
2	$z_{2i} = 2^* z_{1i}$ for 10% of psu's	153	5.53	0.19
3	$z_{21} = R(z_{1i}/2, 3z_{1i}/2)$	154	4.17	0.25

ACKNOWLEDGEMENTS

The author would like to thank the Editor and two referees for helpful comments and the correction of errors in the original presentation.

APPENDIX

Pseudocode for Variant 1 of PPS Sequential Sampling

It is assumed here that the population of size measures has already been given a suitable ordering, say by the algorithm given in Sunter (1986) and that its index, i, in this ordering identifies the unit. Size measures, scaled to sum to 1, are stored in an array z[1 .. PopSize] with their cumulative values (accumulated from PopSize down to 1) stored in an array Z[1 .. PopSize]. The meaning of the variables will be clear from the names that they are given. The results are to be stored in an array Sample [1 .. SamSize,1 .. 3] in which the elements are population index i, unit probability π_i, and τ_i respectively. "Random" is a function that returns a random number uniformly distributed on the the interval (0,1). The indentations in the code written below are intended to facilitate the visual pairing of the begin/end's that delineate a compound statement.

```
{Variables initialization}
  $i$ = 1; SamProb = 1; NumRem = SamSize; Gamma = $1/Z[2]$;
  {Sampling routine}
  while NumRem > 0 do
begin
  if $i > 1$ and $i <$ PopSize then
    Gamma = Gamma*$(1 - z[i - 1}/Z[i])$*$Z[i]/Z[i + 1]$;
  if i = PopSize − NumRem + 1 or Random < = Numrem*$z[i]/Z[i]$
  then
  begin
    if $i$ < > PopSize − NumRem + 1 then
      SamProb = SamProb*NumRem*$z[i]/Z[i]$;
    NumRem = NumRem − 1;
    Sample[SamSize − NumRem,1] = $i$;
    Sample[SamSize − NumRem,2] = SamSize*$z[i]$;
    Sample[SamSize − NumRem,3] = Gamma;
  end else SamProb = SamProb*(1 − NumRem*$z[i]/Z[i]$);
  $i = i + 1$;
end.
```

REFERENCES

DREW, J.D., CHOUDHRY, G.H., and GRAY, G.B. (1978). Some methods for updating sample survey frames and their effects on estimation. *Survey Methodology*, 4, 225-263.

FELLEGI, I.P. (1963). Sampling with varying probabilities without replacement: rotating and non-rotating samples. *Journal of the American Statistical Association*, 58, 183-201.

FELLEGI, I.P. (1966). Changing the probabilities of selection when two units are selected with PPS without replacement. *Proceedings of the Social Statistics Section, American Statistical Association*, 434-442.

KEYFITZ, N. (1951). Sampling with probabilities proportional to size. *Journal of the American Statistical Association*, 58, 183-201.

KISH, L., and SCOTT, A., (1971). Retaining units after changing strata and probabilities. *Journal of the American Statistical Association*, 66, 461-470.

RAO, J.N.K., HARTLEY, H.O., and COCHRAN, W.G. (1962). On a simple procedure of unequal probability sampling without replacement. *Journal of the Royal Statistical Society*, Series B, 24, 482-490.

PLATEK, R., and SINGH, M.P. (1978). A strategy for updating continuous surveys. *Metrika*, 25, 1-7.

SUNTER, A.B. (1986). Solutions to the problem of unequal probability sampling without replacement. *International Statistical Review*, 54, 33-50.

SUNTER, A.B. (1989). PPS Sampling in multistage designs: does it matter which method? Manuscript submitted to *Journal of Official Statistics*.

Survey Methodology, December 1989
Vol. 15, No. 2, pp. 261–270
Statistics Canada

The Use of Administrative Records for Estimating Population in Canada[1]

RAVI B.P. VERMA and RONALD RABY[2]

ABSTRACT

This paper examines the adequacy of estimates of emigrants from Canada and interprovincial migration data from the Family Allowance files and Revenue Canada tax files. The application of these data files in estimating total population for Canada, provinces and territories, was evaluated with reference to the 1986 Census counts. It was found that these two administrative files provided consistent and reasonably accurate series of data on emigration and interprovincial migration from 1981 to 1986. Consequently, the population estimates were fairly accurate. The estimate of emigrants derived from the Family Allowance file could be improved by using the ratio of adult to child emigrant rates computed from Employment and Immigration Canada's immigration file.

KEY WORDS: Interprovincial migration; Emigration; Population estimates; Census counts; Accuracy.

1. INTRODUCTION

The national Census, conducted every five years since 1951, provides a wide range of demographic data on the Canadian population. However, unlike some other industrialized countries, Canada does not have a continuous population registration to derive basic demographic data and track the movement of people over different geographic areas for non-census years. To fill this gap, since the 1940s Statistics Canada has developed a program of population and family estimates. For example, population estimates for Canada, provinces and territories, census divisions, and census metropolitan areas are produced using the latest census counts and several administrative data sources, including: Revenue Canada tax files and Family Allowance files for migration; Vital Statistics registration for births and deaths; and Immigrant Visa and Record of Landing Registration for immigration.

The strengths and weaknesses of these administrative files for estimating population and migration compared with 1981 Census data have been discussed elsewhere. (Statistics Canada 1987; Verma and Parent 1985; Norris, Britton and Verma 1982). In this paper, the accuracy of estimates of the components of population change for provinces and territories using the Family Allowance and Revenue Canada data sources will be evaluated by comparison with the 1986 Census counts. This evaluation will compare 1971, 1976 and 1981 data.

The paper is presented in the following sections: data sources and the methods of estimation; results of the evaluation; and conclusions and discussion.

2. DATA SOURCES AND THE METHODS OF ESTIMATION

This section describes the procedures for estimating total population, interprovincial migration, and emigration.

[1] Revised version of a paper presented at Statistics Canada Symposium on Statistical Uses of Administrative Data, November 1987.

[2] Ravi B.P. Verma and Ronald Raby, Demography Division, Statistics Canada, 4-A Jean Talon Building, Ottawa, Ontario, K1A 0T6.

2.1 Total Population

Quarterly and annual estimates of the total population of Canada and the provinces and territories, and annual totals for census divisions and census metropolitan areas, are produced by the component method. At the national level, the number of births and immigrants are added to, and the number of deaths and emigrants subtracted from, the base population (taken from the latest Census of Canada). By province and for smaller areas, estimates of internal migration are also taken into account.

The component method is expressed as follows:

$$\hat{P}(t + i) = P(t) + [B(t,t + i) - D(t,t + i) + I(t,t + i) - E(t,t + i)] + N(t,t + i). \tag{1}$$

Where, for any given province:

$\hat{P}(t + i)$ = estimate of population at time $t + i$

$P(t)$ = Census population counts at time t

B = number of births between time t and $t + i$

D = number of deaths between time t and $t + i$

I = number of immigrants between time t and $t + i$

E = number of emigrants between time t and $t + i$

N = number of net interprovincial immigrants between time t and $t + i$

$(t,t + i)$ = interval between the last census date and the reference date of the estimate.

2.2 Interprovincial Migration

Two administrative files are used to produce annual and quarterly estimates of interprovincial migration. Preliminary estimates are derived from Family Allowance files, while final figures are estimated from Revenue Canada income tax files.

2.2.1 Preliminary Estimates

The number of adult migrants is estimated using child migration figures derived from Family Allowance files, and ratios of adult out-migration rates to child out-migration rates ($f_{j,k}$) based on the most recent Revenue Canada tax file (calculated for 1 or 2 years before the reference date). Recipients of Family Allowance cheques must notify the Department of Health and Welfare of changes in address. These changes are compiled monthly for both province of origin and destination, by size of family (the number of children per family receiving the allowance). Coverage of the population by Family Allowance is comparable to that of the census (Statistics Canada 1987, p. 46). Estimates of the number of interprovincial out-migrants for all age groups are calculated as follows:

$$\hat{M}_{(j,k),18+} = \frac{M_{(j,k),0\text{-}17}}{P_{j,0\text{-}17}} \cdot f_{(j,k)} \cdot P_{j,18+} \tag{2}$$

$$f_{(j,k)} = \frac{M'_{(j,k),18+}}{\hat{P}_{j,18+}} \div \frac{M'_{(j,k),0\text{-}17}}{\hat{P}_{j,0\text{-}17}} \tag{3}$$

$$\hat{M}_{(j,k),0+} = \hat{M}_{(j,k),18+} + M_{(j,k),0\text{-}17} \tag{4}$$

where:

$\hat{M}_{(j,k),0+}$ = estimated total number of persons out-migrating from province j to province k

$\hat{M}_{(j,k),18+}$ = estimated number of adult out-migrants (aged 18+) from province j to province k

$M'_{(j,k),18+}$ = number of adult out-migrants from province j to province k derived from Revenue Canada tax files

$M'_{(j,k),0\text{-}17}$ = number of child out-migrants (aged 0-17) from province j to province k derived from Revenue Canada tax files

$M_{(j,k),0\text{-}17}$ = number of child out-migrants from province j to province k, based on Family Allowance files

$P_{j,18+}$ = estimated number of adults in province j, the difference between the total population estimates and estimates of the child population based on Family Allowance files

$P_{j,0\text{-}17}$ = total number of children receiving Family Allowance payments in province j

$f_{(j,k)}$ = estimation factor for adult migrants from province of origin j to province of destination k, based on estimates of migration from Revenue Canada tax files

$\hat{P}_{j,18+}$ = number of adults in province j, Demography Division population estimates

$\hat{P}_{j,0\text{-}17}$ = number of children in province j, Demography Division population estimates.

2.2.2 Final Estimates

Revenue Canada tax files are used to produce final estimates of interprovincial migrants. All individuals receiving an annual income above a specified minimum are required to file an income tax return by the end of April of each year. Migrant tax filers are identified by comparing area of residence from two consecutive tax returns. Information on the number and ages of dependents is imputed from the total amount of personal exemptions claimed by filers. An adjustment is made for segments of the population not covered by the Revenue Canada system; this includes people who neither file an income tax return nor appear as dependents in another filer's return (Norris and Standish 1983; Statistics Canada 1987).

2.3 Emigration

In Canada no system exists for recording emigrants; hence, their numbers must be estimated. Revenue Canada income tax files with an "out-of-Canada" address one year and an "in-Canada" address for the previous year are used to identify emigrants. The emigrant status of children under 17 years of age is determined from change of address notifications from Family Allowance recipients. By combining information from these two administrative files, both preliminary and final estimates of emigrants are generated. The estimation procedures are similar to those used to estimate preliminary interprovincial migration:

$$\hat{E}_j = \left[\frac{E_{j,0\text{-}17}}{P_{j,0\text{-}17}} \cdot f_c \cdot P_{j,18+}\right] + E_{j,0\text{-}17} \tag{5}$$

$$f_c = \frac{E'_{c,18+}}{\hat{P}_{c,18+}} \div \frac{E'_{c,0\text{-}17}}{\hat{P}_{c,0\text{-}17}} \tag{6}$$

$$\hat{E}_c = \sum_{j=1}^{12} \left[\hat{E}_j\right] \tag{7}$$

where:

$\hat{E}_j$ = estimated annual number of emigrants from province j

$\hat{E}_c$ = estimated annual number of emigrants from Canada

$E_{j,0\text{-}17}$ = number of emigrants from province j aged 0 to 17 who were eligible for Family Allowance

$P_{j,0\text{-}17}$ = number of children in province j who are eligible for Family Allowance

$P_{j,18+}$ = adult population of province j obtained by subtracting the number of children eligible for Family Allowance from the total estimated population

f_c = annual adjustment factor for estimating adult emigration from Canada, based on Revenue Canada tax files.

$E'_{c,18+}$ and $E'_{c,0\text{-}17}$ = estimated numbers of adult and child emigrants from Canada, based on Revenue Canada tax files.

$\hat{P}_{c,18+}$ and $\hat{P}_{c,0\text{-}17}$ = estimated June 1st population of adults and children for Canada, based on the component method.

The method of estimating the number of emigrants was modified in March 1989, affecting estimates after 1986. The new method combines counts by age of emigrants from Canada to the United States (from the U.S. Department of Justice, Immigration and Naturalization Service), and estimates of the numbers of emigrants from Canada to countries other than the U.S. based on Family Allowance files and an f_c factor calculated from immigration files (see Raby, Martel and Cartier 1989).

3. EVALUATION OF ESTIMATES OF THE COMPONENTS OF POPULATION CHANGE

Each component of population change (births, deaths, immigrants, emigrants and interprovincial migrants) may contain a degree of bias and error. However, the data on births, deaths and immigration can be regarded as more accurate than the estimates of emigrants and interprovincial migrants. In 1982, the methods of estimating emigrants and internal migration were thoroughly updated (see Statistics Canada 1987). These revised methods are evaluated below.

Table 1

Estimates of Emigrants by Different Methods, Canada, 1976-1981 and 1981-1986

Method	1976-81	1981-86
Residual*		
(a) Unadjusted	277,558	476,373
(b) Adjusted for Undercoverage	196,955[1]	134,857[1]
(c) Adjusted for Net Undercoverage	194,155[2]	218,148[2]
Revenue Canada Tax File	207,420	165,272
Family Allowance Method	278,624	235,481
Reverse Record Check	296,724	288,376

*Residual Method:
Emigrants = ([Births − Deaths] + [Immigrants]) − Intercensal growth of population between time t and $t + 5$.

[1] The undercoverage rates were 2.04% for the 1976 Census, 2.01% for the 1981 Census, and 3.21% for the 1986 Census.
[2] The 1976, 1981 and 1986 Census net undercoverage rates were 1.53%, 1.51% and 2.40% respectively. They are estimated using the U.S. experience of overcoverage which is 25% of the undercoverage rate.
Source: Demography Division, Statistics Canada.

3.1 Emigration Data

Table 1 presents estimates of emigrants from Canada by using different methods and data sources for 1976-1981 and 1981-1986. For 1981-1986, the estimate using the residual method is considerably higher than the estimate based on the Family Allowance file. The residual method subtracts the population growth between 1981 and 1986, unadjusted for census undercoverage, from natural increase and immigration. Since births, deaths and immigration data are assumed to be accurate, the higher estimate by the residual method can be attributed to the difference in undercoverage rates for 1981 and 1986. After adjusting the 1981 and 1986 Census counts for undercoverage (2.01% and 3.21% respectively), the estimate by the residual method was found to be 134,857. This figure is lower than estimates obtained using both the Family Allowance file (235,481) and the Revenue Canada tax file (165,272).

This low estimate may result from different rates of overcoverage in the 1981 and 1986 Censuses. No estimate of overcoverage is calculated in the Reverse Record Check study, but the rate can be assumed to be similar to the U.S. Census rate which is 25% of the undercoverage rate. After adjusting the 1981 and 1986 Census counts for net coverage rates of 1.51% and 2.40% respectively, the residual estimate (218,148) was close to the Family Allowance-based estimate (235,481).

For 1976-1981, the estimating methods do not produce similar results. The number of emigrants estimated by the residual method adjusted for net undercoverage was 194,155, which is close to the estimate based on Revenue Canada tax files (207,420), but considerably lower than the Family Allowance method estimate (278,624) or the Reverse Record Check estimate (296,724).

One possible source of error in the current method is the f_c factors, which are adult-child emigrant ratios, estimating the number of emigrants aged 18+ from 1981-1986. These ratios were obtained from the emigration data provided by the Revenue Canada tax files.

Table 2 shows f_c values derived from different data sources. The f_c factors from the Revenue Canada tax files are less than unity and higher than unity from the three other data sources: interprovincial migration data from income tax files, immigration files, and data on Canadian emigrants to the United States. The estimates of emigrants from these sources are also higher than the Revenue Canada-based estimate.

Table 2

Estimates of Emigrants by Family Allowance Method Using Different Values of f_c (Adult-Child Emigrant Ratios), 1981-1986

Data Source of f_c	Value of f_c Factor					Number of Emigrants
	1981-82	1982-83	1983-84	1984-85	1985-86	
1. Revenue Canada Tax Files	0.8698	0.8768	0.9052	0.8592	0.8592	235,481
2. Interprovincial Migration Data from Income Tax Files	1.0760	1.1000	1.0664	1.0290	1.0029	265,816
3. EIC Immigration Data	1.0801	1.0926	1.1723	1.1254	1.0694	275,762
4. Canadian Emigrants to the U.S.A.	1.2300	1.2774	1.3196	1.3745	1.4232	316,268

Source: Demography Division, Statistics Canada.

Each f_c factor source shows annual variations. The f_c factors for Canadians emigrating to the United States are particularly high, indicating that 23% to 42% more adults emigrated to the U.S. than did children. This is not surprising, as the southern American states have always been attractive to retirees. Hence the f_c factor based on U.S. data may not be suitable for estimating Canadian emigrants to countries other than the U.S.

Similarly, the f_c factors for interprovincial migration, based on the income tax file, suggest that adult migrants have exceeded child migrants by up to 10% from 1981 to 1986. However, the adult migrant group likely contains a high proportion of younger adults, who tend to move more often between provinces than other age groups. Hence this data source is also very specific and thus not suitable for computing the overall f_c factor.

According to some authors (Beaujot and Rappak 1988), emigrant and immigrant flow data are associated, making it possible to compute an f_c factor from the Emloyment and Immigration Canada (EIC) immigration file. f_c factors from the EIC immigration file are intermediate between those derived from interprovincial immigrant data and U.S. emigrant data. The figure based on the f_c factor from the immigration file (275,762) is higher than the official estimate of emigrants (235,481), but is close to that derived from the 1986 Reverse Record Check study (288,376). If the official estimate of the number of emigrants were increased to 275,762, the 1986 error of closure between the population estimate and census counts would be reduced from 0.95% to 0.79%.

In sum, for the 1981-86 period the estimates of emigrants seemed to be improved by taking f_c factors from the Canada Employment and Immigration (EIC) immigrant file rather than the Revenue Canada tax file.

Yet in March 1989, it was discovered that emigrant estimates based on Family Allowance files and an f_c factor derived from EIC immigration data were still too low after 1986. This seems to be a result of the high proportion (33%) of Canadian emigrants destined for the U.S. from 1981 to 1986, according to U.S. data.

An analysis was also made of a method combining U.S. Department of Justice, Immigration and Naturalization Service data on the numbers emigrating to the U.S. from Canada; child emigrant counts (ages 0-17) from Family Allowance files and an f_c factor obtained from the EIC immigration file for all countries other than the U.S. For 1981 to 1986, the estimated number of emigrants by this method was 285,413. This revised estimate is much closer to the Reverse Record Check study figure (288,376).

Table 3

Estimates of Net Interprovincial Migration from 1986 Census Data on Mobility, Family Allowance Files, Income Tax Files, and Residual Method, Canada, Provinces and Territories, 1981-1986

Geographic Area	1986 Census[1]	Family Allowance Files	Income Tax Files	Residual Method[2]
CANADA	0	0	0	−238,178
Nfld.	−16,550	−14,837	−15,051	−26,111
P.E.I.	1,540	293	751	−509
N.S.	6,275	5,204	6,895	−4,095
N.B.	−1,370	−2,239	−65	−11,212
Que.	−63,295	−76,040	−81,254	−167,286
Ont.	99,355	115,497	121,767	57,147
Man.	−1,555	−3,700	−2,634	−8,180
Sask.	−2,820	−668	−2,974	−13,564
Alta.	−27,665	−34,073	−31,676	−50,811
B.C.	9,500	13,289	7,382	−12,418
Yukon	−2,665	−2,381	−2,775	−1,643
N.W.T.	−755	−345	−366	504

[1] Population 5 years of age and over.

[2] The residual method for estimating net interprovincial migration is:

$$\text{Net Migration} = \text{Growth of Census Population between time } t \text{ and } t + 5 - [(\text{Births} - \text{Deaths}) + (\text{Immigration} - \text{Emigration})].$$

Source: Demography Division, Statistics Canada.

3.2 Interprovincial Migration Data

To test the accuracy of estimates of interprovincial migration obtained from the Revenue Canada tax file, two evaluations were conducted: (i) a comparison of sets of interprovincial migration data derived from the Revenue Canada tax files and Family Allowance files; and (ii) a comparison of the errors of closure of population estimates for two sets of internal migration data.

Table 3 presents net interprovincial migration estimates derived from four sources: 1986 Census data on mobility; the Revenue Canada tax file; the Family Allowance file; and the residual-based net migration estimate. For all provinces, estimates of internal migration derived from the 1986 Census mobility data, the Revenue Canada tax file and Family Allowance files were consistent on the direction of net migration. All sources except the residual-based method show positive net migration for Prince Edward Island, Nova Scotia, Ontario and British Columbia. In other provinces, net migration was negative.

The estimates of net interprovincial migration from Family Allowance files and Revenue Canada tax files are not strictly comparable to the residual method. By definition, the sum of net interprovincial migration in Canada, should be zero. However, the sum produced using the residual method is about 238,000. In addition, the differences between the residual-based and the Revenue Canada/Family Allowance-based net interprovincial migration estimates are very high in Newfoundland, New Brunswick, Quebec, Ontario and Alberta.

The coefficient of variation (the ratio of the standard deviation of the average absolute error of closure for the provinces to the average absolute error of closure) was used to measure the relative accuracy of the internal migration estimates. The other estimates of the components of population change were assumed to be accurate. Statistically, a coefficient of variation of 20% to 30% is normally acceptable.

Table 4

Error of Closure Between Alternative Population Estimates and Census Counts by Province and Territory 1971, 1976, 1981 and 1986

Geographic Area	Error of Closure[1] (%)							
	1971		1976		1981		1986	
	Income Tax	FA	Income Tax	FA	Income Tax	FA	Income Tax	FA
Newfoundland	−2.08	−1.64	0.49	1.34	1.63	2.30	1.97	2.01
Prince Edward Island	−2.09	−2.01	0.17	2.11	−0.05	1.02	0.99	0.63
Nova Scotia	−1.68	−2.39	−0.20	1.18	0.30	0.40	1.24	1.04
New Brunswick	−1.93	−2.65	−1.29	1.81	0.13	0.54	1.58	1.04
Quebec	−0.33	−0.97	−0.05	−0.18	−0.30	−0.07	1.32	1.40
Ontario	0.11	0.99	0.15	0.16	0.64	0.37	0.72	0.65
Manitoba	0.29	0.38	−0.27	0.39	1.07	0.87	0.51	0.41
Saskatchewan	0.44	−0.33	0.45	0.37	−0.31	0.28	1.08	1.31
Alberta	−0.14	0.52	−1.07	−1.11	−2.39	−2.64	0.73	0.63
British Columbia	0.01	−1.34	0.28	−1.10	0.03	−0.07	0.59	0.79
Yukon	−5.36	−5.99	−0.87	3.79	−1.98	2.06	−4.78	−3.10
Northwest Territories	−2.12	2.64	−12.98	−3.39	−7.08	0.43	−1.44	−1.40
Average Absolute Error								
10 provinces	0.91	1.33	0.44	0.97	0.69	0.86	1.07	1.01
Provinces and Territories	1.38	1.82	1.52	1.41	1.33	0.92	1.41	1.22

Note: From 1976 to 1980, Revenue Canada data for children were available for age group 0-15 only. Therefore the $f_{(j,k)}$ factors were calculated using migrants aged 0-15 and 16+ instead of 0-17 and 18+.

[1] Error of closure is calculated using the following equation:

$$\text{Error of closure} = \left(\frac{\text{Estimate} - \text{Census count}}{\text{Census count}}\right) \times 100$$

Income Tax: Revenue Canada Income Tax File. FA: Family Allowance File.

Source: Estimates of interprovincial migration based on Family Allowance data, Demography Division, Statistics Canada.
Estimates of interprovincial migration based on tax data, Small Area and Administrative Development Division, Statistics Canada.

Table 5

Coefficients of Variation of the Average Absolute Error of Closure between the Population Estimates and Census Counts among Provinces ($n = 10$), by Source of Interprovincial Migration Estimates, 1966-1971, 1971-1976, 1976-1981 and 1981-1986

Period ($t,t + 5$)	Source	AAE ($t + 5$)	Standard Deviation	Coefficient of Variation (%)
		(1)	(2)	(3) = (2 ÷ 1) × 100
1966-1971	Income Tax	0.91	0.2863	31
	FA	1.33	0.2642	20
1971-1976	Income Tax	0.44	0.1317	30
	FA	0.97	0.2135	22
1976-1981	Income Tax	0.69	0.2463	36
	FA	0.86	0.2855	33
1981-1986	Income Tax	1.07	0.1496	14
	FA	1.01	0.1570	16

Note: AAE: Average absolute error of closure.
Income Tax: Revenue Canada Income Tax File.
FA: Family Allowance File.

Source: Demography Division, Statistics Canada.

However, one could argue that the coefficient of variation is not a good indicator of the quality of internal migration data. For example, a set of estimates with an absolute error of closure of 10% for every province would give a coefficient of variation of zeros and consequently would be preferable to a set of estimates with closure errors ranging between – 1.0% and 1.0%. For cases like this, a quality measure that takes into account the size of the absolute error of closure as well as the standard deviation of absolute closure errors is clearly required. However, the likelihood of the provinces having the same absolute error of closure is extremely low (see Table 5), hence, the application of the coefficient of variation in this paper seemed to be valid.

Table 5 shows the coefficient of variation (computed from figures in Table 4) for population estimates based on two sets of internal migration estimates and the census counts for 1971, 1976, 1981 and 1986. Before 1976, the coefficients of variation for migration data from tax files were 50% higher for data from the Family Allowance file. This was expected, since the method for estimating migration from tax files was in the developmental stage. Futhermore, in estimating the number of interprovincial migrants, the f_j factor (adult to child migration rates) was based on Census mobility data, an approach found to be less satisfactory than the current method. However, for 1976-1981 and 1981-1986, the gap in the coefficient of variation between the tax and Family Allowance migration data narrowed considerably.

The tax-based migration data coefficient of variation was 9% higher in 1981 and 12% lower in 1986 than the coefficient of variation based on the Family Allowance file. Hence, the two sets of data are comparable, producing similar provincial estimates and errors of closure with the same level of variation among provinces. Since the coefficient of variation for each set is under 20%, they provide acceptable data on internal migration.

In conclusion, estimates of interprovincial migration from the Revenue Canada tax files for 1981-1986 are consistent with estimates from the Family Allowance file. By province, they yield small variations in the errors of closure.

4. CONCLUSION AND DISCUSSION

The Family Allowance files and Revenue Canada tax files play important roles in providing consistent emigration and internal migration estimates for Canada, and for the provinces and territories. For 1981 to 1986, estimates of emigrants and interprovincial migrants obtained from these files are acceptable for estimating total population.

Nationally the error of closure (the difference between the population estimates and census counts) for 1986 was higher than for the census years 1971, 1976 and 1981. In addition, the errors of closure by province in 1986 were positively biased, indicating that in all provinces the estimates were higher than census counts.

These discrepancies are largely a result of differences in coverage of the 1981 Census population, which was used as the bench-mark, and coverage of the 1986 Census population. The Reverse Record Check estimate of the 1981 undercoverage rate for Canada was 2.01%. The estimate for the 1986 Census was considerably higher, 3.21%.

Errors in the estimates of the other components of change may also partly account for the discrepancies.

ACKNOWLEDGEMENT

We are grateful to the reviewers of this article for their comments and suggestions.

REFERENCES

BEAUJOT, R., and RAPPAK, J.P. (1988). *Emigration from Canada: its Importance and Interpretation.* Ottawa: Employment and Immigration Canada.

NORRIS, D., BRITTON, M., and VERMA, R.B.P. (1982). The use of administrative records for estimating migration and population. *Statistics of Income and Related Administrative Record Research: 1982,* Washington, D.C.: Department of the Treasury, Internal Revenue Service.

NORRIS, D., and STANDISH, L.D. (1983). A technical report on the development of migration data from taxation records. Technical Report, Small Area and Administrative Data Development Division, Statistics Canada.

RABY, R., MARTEL, J., and CARTIER, G. (1989). Issues in the current postcensal population estimates. Paper presented at the Federal-Provincial Committee on Demography, Ottawa.

STATISTICS CANADA (1987). *Population Estimation Methods, Canada.* Catalogue 91-528E, Statistics Canada.

VERMA, R.B.P., and PARENT, P. (1985). An overview of the strengths and weaknesses of the selected administrative data files. *Survey Methodology*, 11, 171-179.

Survey Methodology, December 1989
Vol. 15, No. 2, pp. 271–280
Statistics Canada

Confidence Intervals for Postcensal Population Estimates: A Case Study for Local Areas

DAVID A. SWANSON[1]

ABSTRACT

This paper presents a technique for developing appropriate confidence intervals around postcensal population estimates using a modification of the ratio-correlation method termed the rank-order procedure. It is shown that the Wilcoxon test can be used to decide if a given ratio-correlation model is stable over time. If stability is indicated, then the confidence intervals associated with the data used in model construction are appropriate for postcensal estimates. If stability is not indicated, the confidence intervals associated with the data used in model construction are not appropriate, and, moreover, likely to overstate the precision of postcensal estimates. Given instability, it is shown that confidence intervals appropriate for postcensal estimates can be derived using the rank-order procedure. An empirical example is provided using county population estimates for Washington state.

KEY WORDS: Population estimation; Confidence intervals; Ratio-correlation regression.

1. INTRODUCTION

A method of generating confidence intervals for postcensal estimates was not available until Espenshade and Tayman (1982) introduced a time-series regression estimation technique utilizing age-specific postcensal death rates. The Espenshade-Tayman technique represents an important breakthrough in estimation technology; however, like most breakthroughs it has limitations, of which two are notable:

1. The technique is likely to be unsatisfactory at the subprovincial or substate level (Espenshade and Tayman 1982); and
2. It is a major departure from the standard regression technique used in Canada and the United States for estimating county-equivalent populations, namely, ratio-correlation. This departure is a particularly salient issue in terms of data requirements and the experience of people responsible for making county-equivalent and other subprovincial level population estimates. (Statistics Canada 1987). The term "county equivalent" is defined as a Census Division in Canada (Statistics Canada 1987) and as a county in nearly all U.S. states; notable exceptions in the U.S. include Alaska, in which county-equivalents are Census Areas, Louisiana, where Parishes functions as counties, and Virginia, in which "independent cities" are included as county-equivalents.

This paper presents a means of developing confidence intervals for postcensal county-equivalent populations using the rank-order procedure, a modification of the ratio-correlation method introduced by Swanson (1980) that exploits causal modeling concepts to take into account postcensal structural changes in a given ratio-correlation model.

There are three issues relevant to the development of confidence intervals made using the ratio-correlation method. The first has to do with model stability over time. If the structure of associations among model variables is invariant over time, then the confidence intervals

[1] David A. Swanson, Department of Sociology, Pacific Lutheran University, Tacoma, Washington 98447, U.S.A.

constructed in regard to the model data set will apply to the population estimates generated by the model from the estimation data set. Although it has been consistently documented that it is not prudent to assume model invariance (D'Allesandro and Tayman 1980; Ericksen 1973, 1974; Mandell and Tayman 1982; Namboodiri 1972; O'Hare 1976, 1980; Smith and Mandell 1984; Spar and Martin 1979; Swanson 1980; Swanson and Prevost 1986; Swanson and Tedrow 1984; Tayman and Schafer 1982; Verma *et al.* 1983), it would be useful to have a testing procedure for stability. This leads to the second issue, namely, the use of a statistical test. If the test indicates that stability can not be assumed, and yet confidence intervals associated with, say, a model constructed using 1960-70 data, are applied to estimates generated for, say, 1979, they are likely to overstate the level of precision in the 1979 estimates. Thus, the third issue is the need for a procedure that will generate appropriate confidence intervals.

In the report that follows, a description of ratio-correlation is provided along with the modification that forms the basis for developing appropriate confidence intervals. Next, the logic for developing these confidence intervals is formally described, followed by an empirical example showing both the test for instability and the generation of both "inappropriate" and "appropriate" confidence intervals.

2. METHODOLOGY FOR POPULATION ESTIMATION

Ratio-correlation is a regression method designed to measure the temporal change in county-equivalent population proportions using observed temporal change in proportions of symptomatic indicators such as registered voters, covered employment and public school enrollment. The temporal change is measured by simply taking a ratio of proportions at two points in time.

Since enumerated population numbers for all county-equivalents are available only from the federal census, a ratio-correlation regression model is always constructed using two points in time separated by a regular number of years. It is formally described as

$$\mathrm{R}_{it} = a_o + \sum_{j=1}^{k} (b_j)\ (X_i)_{jt} + \epsilon$$

where

a_o = the intercept term to be estimated

b_j = the regression coefficient to be estimated

ϵ = the error term

j = symptomatic indicator, $(1 \le j \le k)$

i = county-equivalent $(1 \le i \le n)$

t = the year of the most recent census

and

$$\mathrm{R}_{it} = \left[\frac{P_{i,t}}{\sum P_{i,t}}\right] \div \left[\frac{P_{i,t-z}}{\sum P_{i,t-z}}\right] \qquad (1.\mathrm{A})$$

$$(X_i)_{t,j} = \left[\frac{S_{i,t}}{\sum S_{i,t}}\right] \div \left[\frac{S_{i,t-z}}{\sum S_{i,t-z}}\right]_j \qquad (1.B)$$

where

Z = the number of years between each census

P = Population

S = Symptomatic Indicator

Once a model is constructed, it is used to develop a postcensal estimate for time $t + x$ by substituting $(S_{i,\,t+x}/\sum S_{i,t+x})_j$ into the numerator of the right-hand side of equation [1.B] while $(S_{i,t}/\sum S_{i,t})_j$ is substituted into the denominator of the right-hand side of equation [1.B]. This means that once $\hat{R}_{i,t+x}$ is obtained, an actual population for area i at time = $t + x$ is developed by introducing an independently estimated total population, P_{t+x}, into equation [1.A] and algebraically solving equation [1.A] for $P_{i,t+x}$. Since $\sum \hat{P}_{i,t+x}$ does not usually equal the independently derived total, P_{t+x}, an adjustment is made to force the summed population figures to the independently estimated total.

One limitation of ratio-correlation is that its structure is invariant over time, which is why the rank order procedure was introduced by Swanson (1980). The rank-order procedure is based on the fact that information contained in the zero-order correlations found in an estimation data set can be exploited due to work by Land (1969, Chapter IV); work that is based on the fundamental theorem underlying path analysis as developed by Wright (1921). It involves a theoretical reversal of the dependent variable in the regression model, the population variable, as an unmeasured, causally prior variable and a just-identified structure - a minimum of three predictor variables (in the regression model), the covariance of which is assumed to be due to the fact that they are all causally related to the population variable.

3. METHODOLOGY FOR CONFIDENCE INTERVALS ESTIMATION

If the relationships found among the variables in the model data set remain stable over time (as shown through the rank-order procedure) then the same relationships should be found among the variables in the estimation data set. This stability would indicate that the S.E.E. associated with the model data set is appropriate for generating confidence intervals for the estimation data set. However, if stability does not exist, then the S.E.E. associated with the model data set is not appropriate, and may, in fact, generate confidence intervals that overstate the precision of postcensal estimates. These considerations lead to the question of determining stability through statistical inference.

In answering the question just posed, consider that we are examining related pairs of variables. This implies that the Wilcoxon matched-pairs signed rank test could be used (Mosteller and Rourke 1973). In using this test, the null hypothesis is that there are no differences between the population estimates (scores) produced by the unmodified and modified regresion models.

The key to developing confidence intervals for postcensal county equivalent population estimates is found in the fact that the rank-order procedure generates a set of regression coefficients for the estimation data set. From these coefficients, estimates of R^2 and the S.E.E. for the estimation data set can be developed, and the estimated S.E.E. leads directly to the development of confidence intervals. First, recall that the coefficient of multiple determination, R^2, is simply the sum of the products of each zero-order correlation between an independent variable and the dependent variable, and the standardized regression coefficient for each independent variable (Hayes 1973), so that S.E.E. is (Hayes 1973)

$$\text{S.E.E.} = \left[\frac{(n)\ (S_y^2)\ (1 - R^2)}{n - 2}\right]^{1/2}$$

where

n = number of cases (county-equivalents)

S_y^2 = variance of the dependent variable

R^2 = coefficient of multiple determination

The formula for generating a confidence interval around a given estimated value for a point on a (population) regression line is provided by Kmenta (1971)

$$Y_i \pm (t_{n-2,\alpha/2})\ (\text{S.E.E.})$$

An important point to realize is that the confidence interval is not directly generated for a population estimate, rather it is for the estimated ratio of proportions, or $R_{it\ +\ x}$. However, as shown by Espenshade and Tayman (1982), a confidence interval around one variable can be translated for another variable algebraically substituted for the first. Thus, by finding the lower and upper confidence boundaries of $R_{it\ +\ x'}$ these lower and upper confidence boundaries can be translated into the population values:

$$(R_{it+x}) \pm (t_{n-2,\alpha/2})\ (\text{S.E.E.})$$

$$= \left[\frac{P_{it+x}}{\sum P_{it+x}}\right] \div \left[\frac{P_{it}}{\sum P_{it}}\right] \pm (t_{n-2,\alpha/2})\ (\text{S.E.E.})$$

which leads to

$$\text{L.L.}\ (\hat{P}_{it+x}) =$$

$$\left[\frac{P_{it}}{\sum P_{it}}\right] (\textstyle\sum P_{it+x}) \left[(\hat{R}_{it+x}) - (t_{n-2,\alpha/2})\ (\text{S.\hat{E}.E.})\right]$$

and

$$\text{U.L.}\ (\hat{P}_{it+x}) =$$

$$\left[\frac{P_{it}}{\sum P_{it}}\right]. (\textstyle\sum P_{it+x}) \left[(\hat{R}_{it+x}) + (t_{n-2,\alpha/2})\ (\text{S.\hat{E}.E.})\right]$$

4. EMPIRICAL STUDY

Table 1.A in Swanson (1980) gives the zero-order correlations relating to a ratio-correlation model for estimating county civilian populations under sixty-five years from employment, voters, and grades 1-8 enrollment for the state of Washington, for the period 1950-1960. Characteristics of the model constructed from these data are given in Table 1.B. while Tables 2.A and 2.B provide similar results for the 1960-1970 period as found in Swanson (1980). This latter set forms the estimation data over which the procedure will be described.

Although full knowledge of the estimation data set is available, the procedure is used as if this were not the case. Of course, what is known in any estimation problem is the zero-order correlation matrix for the independent variables, which is used in conjunction with the fundamental theorem of path analysis to estimate the coefficients for the modified model. Using the complete rank-order procedure, the modified model (Swanson 1980) is:

$$Y = 0.046618 + 0.066786X_1 + 0.50727X_2 + 0.38736X_3.$$

Estimates for 1970 of the county civilian population under sixty-five years of age (adjusted to the independently estimated state total) resulting from the preceding modified model are presented in Table 1 along with the actual enumerated populations.

The Wilcoxon test was conducted for the Washington data using the procedure in the SPSSx NPAR Tests command (SPSS 1986). To save space, the unmodified and modified population estimates are not presented. They can be found in Table 3 of Swanson (1980). Under the null hypothesis, the probability of obtaining $Z = -3.2096$ is 0.0013. Thus, the null hypothesis is rejected and it is assumed that instability exists for Washington counties in going from the model constructed using 1960/1950 data to the true unknown model associated with 1970/1960 data.

As a note of interest, the Chow test (Chow 1960) validated the results of the Wilcoxon test by showing that the difference between the "true" 1970-1960 ratio-correlation model and the 1960/1950 ratio-correlation model was statistically significant.

Had the results of the Wilcoxon test led us not to reject the null hypothesis, we would have used the unmodified coefficients from the 1960/1950 model data set to generate 1970 population estimates for Washington counties. Further, the S.E.E. for this same model (0.05022) would have been used to generate confidence intervals for the 1970 estimates. However, the results of the Wilcoxon test led us to reject the null hypothesis in this case. This indicates the modified coefficients developed using the rank-order procedure should be used in lieu of the unmodified model. Further, it indicates the need for a revised S.E.E., one that is not likely to overstate the precision of the 1970 estimates.

Using the estimated values found in the 1970 example data for Washington state (Swanson 1980) we find

$$\hat{R}^2 = (0.07533)\,(0.75290) + (0.47085)\,(0.92146) + (0.49481)\,(0.88082) = 0.926$$

and

$$(\text{S.}\hat{\text{E}}\text{.E.}) = \left[\frac{(39)\,(0.2145)^2\,(1 - 0.926)}{39\text{-}2}\right]^{1/2}$$
$$= 0.0599$$

Table 1
90% Confidence Interval for the Estimated Civilian Population Under Sixty-Five Years by County, State of Washington 1970

County	Enumerated Population	Lower Limit	Estimated Population	Upper Limit	90% Confidence Interval (in percent)
Adams	11102	10335	11458	12581	± 9.80
Asotin	11862	10469	11814	13154	± 11.38
Benton	63144	60405	67511	74616	± 10.53
Chelan	35862	31733	36177	40620	± 12.28
Clallam	30023	28063	31294	34525	± 10.32
Clark	116663	101183	111437	121690	± 9.20
Columbia	3771	3683	4161	4639	± 11.49
Cowlitz	62586	55170	61581	67992	± 10.41
Douglas	15287	14569	16252	17935	± 10.36
Ferry	3336	2963	3397	3831	± 12.78
Franklin	23983	21960	24631	27302	± 10.84
Garfield	2546	2447	2761	3075	± 11.37
Grant	38921	37561	42606	47651	± 11.84
Grays Harbor	52583	46294	52114	57935	± 11.17
Island	20589	20512	22148	24040	± 7.39
Jefferson	9235	8440	9473	10506	± 10.90
King	1054271	935664	1037937	1140203	± 9.85
Kitsap	86529	77022	85821	94619	± 10.25
Kittitas	22764	17649	19863	22077	± 11.15
Klickitat	10729	10440	11923	13406	± 12.44
Lewis	39265	35747	40122	44497	± 10.90
Lincoln	8168	7939	9107	10275	± 12.83
Mason	18411	16057	17827	19596	± 9.93
Okanogan	22952	21002	23795	25688	± 10.97
Pacific	13310	11270	12795	14320	± 11.92
Pend Oreille	5185	5147	5893	6639	± 12.86
Pierce	339048	314272	346728	379184	± 9.36
San Juan	3089	2636	2918	3201	± 9.66
Skagit	45703	43255	48758	54261	± 11.29
Skamania	5330	4787	5358	5929	± 10.66
Snohomish	245193	213164	231996	250827	± 8.12
Spokane	251057	227372	256723	286072	± 11.43
Stevens	15178	13869	15780	17692	± 12.11
Thurston	68719	63644	69540	75436	± 8.48
Wahkiakum	3137	3033	3397	3761	± 10.72
Walla Walla	36608	33727	38271	42812	± 11.87
Whatcom	72111	63218	70670	78122	± 10.54
Whitman	34843	28960	32409	35858	± 10.64
Yakima	128960	120347	136203	152219	± 11.69

Note, that from Table 2 in Swanson (1980), the actual R^2 and S.E.E. values are 0.878 and 0.05077, respectively. In comparison with the actual S.E.E. of 0.05077, the estimated S.E.E. is higher. This is appropriate given that we are more uncertain about the precision of estimates generated by the rank-order procedure than we would be about the precision associated with the "true" model, if in fact, the true model was obtainable. With the rank-order procedure, we can now generate a confidence band from the following formula:

$$Y_i \pm (t_{37,\alpha/2})\ (0.0599)$$

In Table 1 an empirical example using a 90% confidence interval is given for the 1970 estimated county population figures presented also in Table 1. Here, the 90% confidence interval is given by:

$$\left[\frac{P_{i1960}}{2522141}\right] (3032053) \left[(\hat{R}_{i1970}) \pm (1.69)\ (0.0599)\right]$$

In examining the confidence intervals given in Table 1 in combination with the enumerated populations provided, it is found that in only one county (Kittitas) is the enumerated population outside of the 90% confidence interval. In this instance, the enumerated population exceeds the upper limit by 687 people. At a 90% level of confidence, the intervals are fairly wide, with a mean of 10.81, a minimum of ± 7.39 percent for Island county and a maximum of ± 12.83 percent in Lincoln County. Compare these with the mean of the absolute percent errors associated with the 1970 estimates, which is 4.89 (Swanson 1980). This comparison suggests that the 90% level generates intervals that are too broad for practical use. Given this, it is of interest to consider which level of confidence would be more appropriate. It is also of interest to consider the effect of using the unmodified S.E.E. (0.05022) from the 1960/1950 model. We would expect that the confidence intervals generated by the unmodified model would be too optimistic. That is, at a given level of confidence, there would be fewer than expected counties for which the interval encompassed the actual population. To explore these issues, Table 2 was constructed.

In Table 2, two distinct sets of information are provided. For both sets, however, a comparison is made between the unmodified and modified estimates and their associated confidence intervals. In regard to the issue of expecting optimistic confidence intervals for the 1970 estimates generated by the unmodified model, Table 2 indicates that at varying levels of confidence ranging from 90% down to 50%, the intervals are, indeed, optimistic in that for only two of the six levels examined are the expected number of county estimates within the specified level of precision. At the 80% level, for example, only 28 (72 percent) of the counties have enumerated 1970 populations within the confidence interval specified around the estimates; at the 60% level, only 22 (56%) of the counties have enumerated 1970 populations within the confidence interval specified around the estimates.

The second aspect of Table 2 is the mean interval associated with a given level of confidence. At the 90% level, the mean of the intervals associated with the unmodified model is 9.10 percent; for the modified model it is 10.81 percent. At the 50% level, the means are 3.66% and 4.35%, respectively. Thus, it is clear that the 60% and 50% levels of confidence generate a mean interval that is more in line with the mean absolute percent error, which is 4.88 for the modified model.

Table 2

Number (%) of Counties in Which Actual 1970 Population was Inside the Confidence Interval

Level of	Unmodified	Modified
Confidence	S.E.E. (0.05022)	S.E.E. (0.0599)
90%	35 (89.7%)	38 (97.4%)
80%	28 (71.8%)	33 (84.6%)
70%	24 (61.5%)	29 (80.6%)
66.66%	24 (61.5%)	26 (66.66%)
60%	22 (56.4%)	23 (59.0%)
50%	20 (51.3%)	22 (56.4%)
	Mean Interval (in percent)	
	Unmodified	Modified
	S.E.E. (0.05022)	S.E.E. (0.0599)
90%	9.10	10.81
80%	7.02	8.38
70%	5.66	6.75
66.66%	5.59	6.40
60%	4.59	5.47
50%	3.66	4.35

In examining the issue of confidence intervals, it appears that a procedure is needed for generating confidence intervals that are not misleading in terms of the precision of postcensal county-equivalent population estimates. However, guidance is also needed on selecting a given level of confidence that is appropriate for the estimates. Of interest in this regard is the work of Stoto (1983) on empirical confidence intervals for population projections. One of Stoto's (1983:18) findings is the high and low population projections produced for the United States by the Bureau of the Census (1977) correspond to a 66.66% confidence interval. It may be the case that for county-equivalent postcensal populations, that the 66.66% confidence level is also appropriate, although in this test this level of confidence generates a mean interval of 6.4 percent for the modified estimates, which is somewhat above their mean percent error (4.9). Another consideration is the length of time between the year for which a postcensal estimate is desired and the preceding census. In the example, the maximum period of postcensal time in the United States was used, 10 years. For each county, we have, in essence, a situation in which maximum uncertainty exists in regard to estimates. From this perspective, the relatively wide interval generated for each county at a 90 percent level of confidence is appropriate. We would expect that structural model changes occur relative to time. Hence, a narrower band would likely be generated in the first year following the end-census year of model construction than in the second year; and so on through the intercensal period.

5. CONCLUSION

At this point it should be clear that the rank-order procedure is not being presented as a fully-validated technique for constructing confidence intervals around postcensal county-equivalent population estimates. However, it appears to offer a reasonable starting point. Even with its limitations, the use of the Wilcoxon test and the confidence intervals developed using the rank-order procedure appears capable of providing benefits to those responsible for making such postcensal population estimates. In the first place, as noted by Espenshade and Tayman (1983), it is important to provide the users of postcensal population estimates some notion of their accuracy as do both the Wilcoxon test and the confidence intervals. Second, with the selection of appropriate confidence intervals, a formal means is available for resolving disputes over the population of a given county-equivalent by using hypothesis testing procedures. Third, S.E.E. can be used as a basis for selecting one model over another. This means that a set of different ratio-correlation models could be considered for any given postcensal estimation year and, further, that a formal criterion is available for selecting one model over another. This feature could be useful in the event that the ratio-correlation estimates generated by a federal, provincial or state demographic center, are challenged in a given postcensal year, an event that has become more frequent, especially in the U.S. (D'Allesandro 1987).

ACKNOWLEDGMENTS

The Author is grateful to Jeff Tayman, anonymous referees, and the editorial staff for comments and suggestions. Peggy Jobe typed a draft of this paper and Carey Taylor typed the final version.

REFERENCES

CHOW, G. (1960). Tests of equality between sets of coefficients in two linear regressions. *Econometrics*, 28, 591-605.

D'ALLESANDRO, F. (1987). Should applied demographers take out liability insurance? Paper presented at the Annual Meeting of The Population Association of America.

D'ALLESANDRO, F., and TAYMAN, J. (1980). Ridge regression for population estimation: Some insights and clarification. *Staff Document No.* 56. Office of Financial Management, State of Washington: Olympia, Washington.

DRAPER, N.R., and SMITH, H. (1981). *Applied Regression Analysis, 2nd Edition*. New York: Wiley.

ERICKSEN, E.P. (1973). A method for combining sample survey data and symptomatic indicators to obtain population estimates for local areas. *Demography*, 10, 137-160.

ERICKSEN, E.P. (1974). A regression method for estimating population changes of local areas. *Journal of the American Statistical Association*, 69, 867-875.

ESPENSHADE, T.J., and TAYMAN, J. (1982). Confidence intervals for postcensal state population estimates. *Demography*, 19, 191-210.

HAYS, W.L. (1973). *Statistics for the Social Sciences*. New York: Holt, Rinehart and Winston.

KMENTA, J. (1971). *Elements of Econometrics*. New York: Macmillan.

LAND, K.C. (1969). Explorations in mathematical sociology. Unpublished Ph.D. dissertation. University of Texas, Austin.

MANDELL, M., and TAYMAN, J. (1982). Measuring temporal stability in regression models of population estimation. *Demography*, 19, 1351-46.

MOSTELLER, F., and ROURKE, R. (1973). *Sturdy Statistics*. Reading, Massachusetts: Addison-Wesley.

NAMBOODIRI, N.K. (1972). On the ratio-correlation and related methods of subnational population estimation. *Demography*, 9, 443-453.

O'HARE, W. (1976). Report on a multiple regression method for making population estimates. *Demography*, 13, 369-379.

O'HARE, W. (1980). A note on the use of regression methods in population estimates. *Demography*, 17, 341-343.

SMITH, S., and MANDELL, M. (1984). A comparison of local population estimates: The housing unit method versus component II, regression, and administrative records. *Journal of the American Statistical Association*, 99, 292-289.

SPAR, M., and MARTIN, J. (1979). Refinements to regression-based estimates of postcensal population characteristics. *Review of Public Data Use*, 7, 16-22.

SPSS, Inc. (1986). *SPSSx User's Guide*. Chicago: SPSS, Inc.

STATISTICS CANADA (1987). *Population Estimation Methods, Canada*. Catalogue No. 91-528E, Statistics Canada.

STOTO, M.A. (1983). The accuracy of population projections. *Journal of the American Statistical Association*, 78, 13-20.

SWANSON, D. (1980). Improving accuracy in multiple regression estimates of population using principles from causal modeling. *Demography*, 17, 413-427.

SWANSON, D., and PREVOST, R. (1986). Identifying extreme errors in ratio-correlation estimates of population. Presented at the Annual Meeting of the Population Association of America.

SWANSON, D., and TEDROW, L. (1984). Improving the measurement of temporal change in regression models used for county population estimates. *Demography*, 21, 373-381.

TAYMAN, J., and SCHAFER, E. (1985). The impact of coefficient drift and measurement error on the accuracy of ratio-correlation population estimates. *The Review of Regional Studies*, 15, 3-10.

U.S. BUREAU OF THE CENSUS (1977). Projections of the Population of the United States, 1977 to 2050. *Current Population Reports. Series P-25 No. 704*. Washington, D.C.: U.S. Government Printing Office.

VERMA, R.V.P., BASAVARAJAPPA, K.G., and BENDER, R.K. (1983). The regression estimates of population for subprovincial areas in Canada. *Survey Methodology*, 9, 219-240.

WRIGHT, S. (1921). Correlation and causation. *Journal of Agricultural Research*, 20, 557-585.

ACKNOWLEDGEMENTS

The Survey Methodology journal wishes to thank the following persons who have served as referees, sometimes more than once, during 1989.

M.G. Arellano, *University of California at San Francisco*

T.R. Balakrishnan, *University of Western Ontario*

M. Bankier, *Statistics Canada*

K. G. Basavarajappa, *Statistics Canada*

D.R. Bellhouse, *University of Western Ontario*

L. Biggeri, *University of Florence*

D. Binder, *Statistics Canada*

L. Blais, *Statistics Canada*

M. Brick, *Westat*

P. Cholette, *Statistics Canada*

G.H. Choudhry, *Statistics Canada*

T.C. Chua, *National University of Singapore*

C.D. Cowan, *National Center for Education Statistics*

E.B. Dagum, *Statistics Canada*

W.A. Fuller, *Iowa State University*

J. Gambino, *Statistics Canada*

J. Gentleman, *Statistics Canada*

M. Gonzalez, *U. S. Office of Management and Budget*

G.B. Gray, *Statistics Canada*

M. Hidiroglou, *Statistics Canada*

S. Hillmer, *University of Kansas*

D. Holt, *University of Southampton*

P. Hoyt, *Statistics Canada*

T.B. Jabine, *Statistical Consultant*

G. Kalton, *University of Michigan*

P.S. Kott, *U.S. Department of Agriculture*

P. Krishnan, *University of Alberta*

S. Kumar, *Statistics Canada*

N.M. Lalu, *University of Alberta*

E. Langlet, *Statistics Canada*

P. Lavallée, *Statistics Canada*

B. Lefrançois, *Statistics Canada*

R.J.A. Little, *University of California at Los Angeles*

L. Mach, *Statistics Canada*

E. Martin, *U.S. Bureau of the Census*

K. Namboodiri, *Ohio State University*

M. Otto, *U.S. Bureau of the Census*

J.N.K. Rao, *Carleton University*

D.B. Rubin, *Harvard University*

I. Sande, *Bell Communications Research*

C.E. Särndal, *University of Montreal*

J. Schafer, *Harvard University*

N. Schenker, *University of California at Los Angeles*

F. Scheuren, *U. S. Internal Revenue Service*

G.A. Shababb, *NPD/Nielsen, Inc.*

A.C. Singh, *Statistics Canada*

C.J. Skinner, *University of Southampton*

J. Smith, *Statistics Canada*

B.D. Spencer, *Northwestern University*

K.P. Srinath, *Statistics Canada*

S. Sudman, *University of Illinois*

A. Sunter, *A.B. Sunter Research Design & Analysis, Inc.*

J.-L. Tambay, *Statistics Canada*

A. Théberge, *Statistics Canada*

V. Tremblay, *Statplus*

A.R. Tupek, *U.S. Bureau of Labor Statistics*

R. Verma, *Statistics Canada*

G. Werking, *U.S. Bureau of Labor Statistics*

W.E. Winkler, *U.S. Bureau of the Census*

K. Wolter, *A. C. Nielsen*

Acknowledgements are also due to those who assisted during the production of the 1989 issues: C. VanBastelaar (Photocomposition), G. Gaulin (Author Services) and M. Haight (Translation Services). Finally we wish to acknowledge J. Clarke, J. Dufresne, M. Kent, C. Lacroix, C. Larabie and D. Lemire for their support with coordination, typing and copy editing.

SPECIAL OFFER

Copies of the proceedings of recent Statistics Canada symposia are still available, and may be purchased at a nominal cost. These are:

Symposium 87: ***Statistical Uses of Administrative Data.*** Two volumes, English and French. Regular price, each: $35.

These are now available at $10 each or $12 for both languages.

Symposium 88: ***The Impact of High Technology on Survey Taking.*** Bilingual, English and French. Regular price, each: $20.

These are now available at $10.

Cheques or money orders should be made payable to:

"The Receiver General for Canada".

Requests for these volumes, along with cheque or money order should be sent to:

Production Manager
Survey Methodology
Statistics Canada
4-C2 Jean Talon Building
Tunney's Pasture
Ottawa, Ontario
Canada K1A 0T6

GUIDELINES FOR MANUSCRIPTS

Before having a manuscript typed for submission, please examine a recent issue (Vol. 10, No. 2 and onward) of Survey Methodology as a guide and note particularly the following points:

l. Layout

1.1 Manuscripts should be typed on white bond paper of standard size (8½ × 11 inch), one side only, entirely double spaced with margins of at least 1½ inches on all sides.

1.2 The manuscripts should be divided into numbered sections with suitable verbal titles.

1.3 The name and address of each author should be given as a footnote on the first page of the manuscript.

1.4 Acknowledgements should appear at the end of the text.

1.5 Any appendix should be placed after the acknowledgements but before the list of references.

2. Abstract

The manuscript should begin with an abstract consisting of one paragraph followed by three to six key words. Avoid mathematical expressions in the abstract.

3. Style

3.1 Avoid footnotes, abbreviations, and acronyms.

3.2 Mathematical symbols will be italicized unless specified otherwise except for functional symbols such as "exp(·)" and "log(·)", etc.

3.3 Short formulae should be left in the text but everything in the text should fit in single spacing. Long and important equations should be separated from the text and numbered consecutively with arabic numerals on the right if they are to be referred to later.

3.4 Write fractions in the text using a solidus.

3.5 Distinguish between ambiguous characters, (e.g., w, ω; o, O, 0; l, 1).

3.6 Italics are used for emphasis. Indicate italics by underlining on the manuscript.

4. Figures and Tables

4.1 All figures and tables should be numbered consecutively with arabic numerals, with titles which are as nearly self explanatory as possible, at the bottom for figures and at the top for tables.

4.2 They should be put on separate pages with an indication of their appropriate placement in the text. (Normally they should appear near where they are first referred to).

5. References

5.1 References in the text should be cited with authors' names and the date of publication. If part of a reference is cited, indicate after the reference, e.g., Cochran (1977, p. 164).

5.2 The list of references at the end of the manuscript should be arranged alphabetically and for the same author chronologically. Distinguish publications of the same author in the same year by attaching a, b, c to the year of publication. Journal titles should not be abbreviated. Follow the same format used in recent issues.