(3

SURVEY METHODOLOGY

Catalogue 12-001-XPB

A JOURNAL PUBLISHED BY STATISTICS CANADA

DECEMBER 1996

•

VOLUME 22

•

NUMBER 2

Statistics Canada Statistique Canada

Canada

SURVEY METHODOLOGY

A JOURNAL PUBLISHED BY STATISTICS CANADA

DECEMBER 1996 • VOLUME 22 • NUMBER 2

Published by authority of the Minister
responsible for Statistics Canada

December 1996

Price: Canada: $45.00
United States: US$50.00
Other countries: US$55.00

Catalogue no. 12-001-XPB
Frequency: Semi-annual

ISSN 0714-0045

Ottawa

Statistics Canada Statistique Canada

Canada

SURVEY METHODOLOGY

A Journal Published by Statistics Canada

Survey Methodology 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

Survey Methodology 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

Survey Methodology is published twice a year. Authors are invited to submit their manuscripts in either English or French to the Editor, Dr. M.P. Singh, Household Survey Methods Division, Statistics Canada, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6. Four nonreturnable copies of each manuscript prepared following the guidelines given in the Journal are requested.

Subscription Rates

The price of Survey Methodology (Catalogue no. 12-001-XPB) is $45 per year in Canada, US $50 in the United States, and US $55 per year for other countries. Subscription order should be sent to Statistics Canada, Operations and Integration Division, Circulation Management, 120 Parkdale Avenue, Ottawa, Ontario, Canada K1A 0T6. A reduced price is available to members of the American Statistical Association, the International Association of Survey Statisticians, the American Association for Public Opinion Research, and the Statistical Society of Canada.

SURVEY METHODOLOGY

A Journal Published by Statistics Canada

Volume 22, Number 2, December 1996

CONTENTS

Survey Methodology, December 1996
Vol. 22, No. 2, pp. 105-106
Statistics Canada

In This Issue

This issue of *Survey Methodology* begins with a special section entitled **Weighting and Estimation** which contains four papers.

The first paper in this special section, by Singh and Mohl, gives an overview of calibration methods from a different perspective, with the objective of gaining a better heuristic understanding of these methods. Deville and Särndal presented calibration methods as minimizing the overall distance of the final weights from the survey weights, subject to the restriction that estimates of totals of certain covariates match known population totals. Singh and Mohl present different calibration methods as being derived from different models for the weight adjustment factors. Computational algorithms for different methods are provided in an appendix, and a numerical example is given to illustrate how the resulting weight adjustment factors might vary among the different methods.

Stukel, Hidiroglou and Särndal also investigate calibration estimators, the class of design-based point estimators developed by Deville and Särndal. These estimators are derived from distance functions and allow for restricting of the final weights such that they are positive or upwardly bounded, thus avoiding the usual problem of negative weights that arises from using the regression estimator. Through simulation, the properties of a number of these estimators based on different distance functions are studied; particular emphasis is given to the properties of the corresponding variance estimators, specifically the Jackknife and the Taylor. The surprising conclusion is that the bias of both the point estimators and the corresponding variance estimators is minimal, even under severe restricting of the final weights.

Jayasuriya and Valliant compare three methods of deriving household weights for the Consumer Expenditure Survey of the U.S. Bureau of Labor Statistics. Survey weights are usually calibrated to population totals of individual level characteristics, resulting in different final weights for individuals in the same household. The principal person method defines the final weight for the household to be the same as that for a particular person in the household. The regression approach replaces the vector of auxiliary variables for each individual in a household by the household average, resulting in identical calibrated weights for persons in the same household. Another option is obtained by restricting the weight adjustment factors to avoid extreme or negative weights. Variations on these methods are compared with respect to the final weights and the estimated CVs for a variety of household expenditure categories.

In the final paper in the section on **Weighting and Estimation**, Chen and Chen consider the problem of confidence interval estimation for a finite population average when auxiliary information is available. Noting the earlier results of Royall and Cumberland that show that naive use of existing design-based methods results in confidence intervals with very poor conditional coverage probabilities, they suggest transformations of the data which improve the adherence to the underlying normality assumption and thus improve the coverage rates. Auxiliary information is incorporated in two ways: either directly into the inference when auxiliary information is known for each unit or through calibration with empirical likelihood when auxiliary information is known only at the population level. Through simulation applied to six real populations, they show that their methods perform well.

In their paper, Thompson and Fisher modify the one and two sample McNemar tests for use with complex survey data. They then apply the modified two sample test to data from the U.S. Bureau of the Census Current Population Survey's Split Panel Study to test whether or not the shift to computer assisted telephone interviewing using a redesigned questionnaire would affect the estimates of unemployment. Results of this test are discussed and compared to other research on the effect of CATI on unemployment estimates.

Eltinge and Jang suggest ways for evaluating the stability of estimates of variance components (specifically within-PSU variance estimators) and other related quantities, under a complex three-stage design. As measures, they consider a simple design-based variance estimator of the within-PSU variance estimator, as well as an estimated "degrees of freedom" approach. A simulation based method permits the assessment as to whether an observed stability measure is consistent with standard assumptions regarding variance estimator stability. They apply the proposed methods to NHANES III data and show that true stability properties may vary substantially across variables, and that within-PSU variance estimators can be substantially less stable than one would anticipate from using a simple count of secondary units within each stratum.

Berger discusses Chao's plan for sequentially selecting an unequal-probability sample of fixed size without replacement. In this context, he suggests an approximation of the second-order probabilities of inclusion in order to obtain an approximate estimator of the variance for the Horvitz and Thompson estimator. This variance is then compared to apprroximations given for other procedures or selection plans. Equivalence conditions for these approximations are presented.

Cowling, Chambers, Lindsay and Parameswaran present two techniques for producing spatially smoothed data and consider their implications in both small and large area estimation. For the small area application, the sample weights are spatially smoothed using a modified linear regression approach, which results in a decrease in the variance but an increase in the bias of the estimates. For the large area application, a nonparametric regression method is used to spatially smooth the data and then the smoothed data is mapped using a Geographic Information System package. The results of a simulation study are presented, in which the most appropriate method and level of smoothing for use in the maps is investigated.

Brick, Waksberg and Keeter suggest using information on interruptions of telephone service so as to adjust the survey estimates to compensate for undercoverage bias. The data collected on telephone service interruptions serve to reduce the bias, but at the same time the variance is likely to increase owing to the greater variability of the sampling weights. The results obtained from a national survey show a significant potential for reducing the mean square error of the estimates under certain conditions.

Finally, Pandher uses a model based approach to find an optimal partition of a survey population into take-all and take-some strata. The approach assumes that there is a single variable of interest and that probability proportional to size sampling is used in the take-some stratum. An algorithm is presented for determining the optimal cut point between the take-all and take-some groups. A key requirement for the algorithm is that the model expectation of the variance is a convex function of the number of units in the take-all stratum, which depends on the model assumptions and the form of the inclusion probabilities. The method is applied to Statistics Canada's Local Government Finance Survey.

The Editor

Survey Methodology, December 1996
Vol. 22, No. 2, pp. 107-115
Statistics Canada

Understanding Calibration Estimators in Survey Sampling

A.C. SINGH and C.A. MOHL[1]

ABSTRACT

There exist well known methods due to Deville and Särndal (1992) which adjust sampling weights to meet benchmark constraints and range restrictions. The resulting estimators are known as calibration estimators. There also exists an earlier, but perhaps not as well known, method due to Huang and Fuller (1978). In addition, alternative methods were developed by Singh (1993), who showed that similar to the result of Deville-Särndal, all these methods are asymptotically equivalent to the regression method. The purpose of this paper is threefold: (i) to attempt to provide a simple heuristic justification of all calibration estimators (including both well known and not so well known) by taking a non-traditional approach; to do this, a model (instead of the distance function) for the weight adjustment factor is first chosen and then a suitable method of model fitting is shown to correspond to the distance minimization solution, (ii) to provide to practitioners computational algorithms as a quick reference, and (iii) to illustrate how various methods might compare in terms of distribution of weight adjustment factors, point estimates, estimated precision, and computational burden by giving numerical examples based on a real data set. Some interesting observations can be made by means of a descriptive analysis of numerical results which indicate that while all the calibration methods seem to behave similarly to the regression method for loose bounds, they however seem to behave differently for tight bounds.

KEY WORDS: Benchmark constraints; Distance minimization; Non-negative weights; Range restrictions.

1. INTRODUCTION

In providing estimates from sample surveys, sampling weights are commonly adjusted to obtain calibrated weights in order to match totals or benchmark constraints (BCs) for auxiliary variables. The methods of regression and raking are often used for this purpose. Although these methods have good asymptotic properties (see Deville and Särndal 1992), they may lead to calibrated weights with undesirable (finite sample) properties. The regression method can give negative weights while the raking procedure can produce very high weights. For this reason, range restrictions (RRs) may be imposed on the calibrated weights. It would be desirable to have a calibration method which (i) produces calibrated weights close to the original sampling weights; this can be achieved via minimization of a suitable distance function between the two sets of weights, (ii) meets BCs, and (iii) satisfies RRs. There exist several methods in the literature for weight adjustment under BCs and RRs, see *e.g.*, Deville and Särndal (1992, henceforth referred to as DS) for recent developments, and Huang and Fuller (1978) for earlier developments. For a review, as well as some further work, see Singh (1993, henceforth referred to as Singh). These methods are iterative in nature and can be classified into two families. Family I consists of methods which satisfy BCs after each iteration and continue to iterate until RRs are met. Family II, on the other hand, consists of methods which satisfy RRs after each iteration and continue to iterate until BCs are met. Methods of DS belong to family II while that of Huang-Fuller belongs to family I. Two additional methods, one for each family, were proposed by Singh. Using arguments similar to DS, Singh extended the remarkable result of DS by showing that all of the methods in families I and II are asymptotically equivalent to the regression method.

In Section 2, a non-traditional approach is followed in introducing each method which is expected to help in understanding of calibration estimators. The functional form of the weight adjustment factor is first heuristically motivated and later on a connection between a suitable method of model fitting and minimization of the distance function is made. Alongside, computational algorithms are given as a quick reference for practitioners. A computer program in GAUSS software is available from the second author; see also Singh and Mohl (1997). In Section 3, numerical examples are presented to illustrate various methods using data from Statistics Canada's Family Expenditure (FAMEX) survey. It is of practical interest to see how different calibration methods might compare for a real data set. In particular, we examine by means of a descriptive analysis the impact of RRs on the computational burden, distribution of weight adjustment factors, point estimates and their variance. Related comparative studies on calibration methods based on real data sets are due to Deville, Särndal and Sautory (1993) and Stukel and Boyer (1993). These studies, however, are restricted to family II methods and are primarily concerned with the distribution of weight adjustment factors. Finally, Section 4 contains a discussion.

[1] A.C. Singh, Methodology Research Advisory Group, and C.A. Mohl, Health Statistics Methods Section, Household Survey Methods Division, Statistics Canada, Ottawa, K1A 0T6.

2. HEURISTIC JUSTIFICATION OF CALIBRATION ESTIMATORS

We will use the following notation. Let n, N denote respectively the sample size and the population size. Let h_k denote the initial or h-weight (used in the expansion or Horvitz-Thompson estimator $\sum_{k=1}^{n} y_k h_k$) for the k-th element where y_k is the value of the study variable. It is assumed that the h-weights include adjustments for any non-response. The parameter of interest is the population total for y, denoted by τ_y. For each k, there are p-auxiliary variables, x_{kj}, $j = 1, ..., p$ for which the population total or benchmark constraint, $\tau_{xj} = \sum_{k=1}^{N} x_{kj}$ for each j is assumed to be known. The transposed p-vector $\boldsymbol{x}_k$ denotes $(x_{k1}, ..., x_{kp})$, the k-th row of the $n \times p$ matrix $\boldsymbol{X}$. Let $c_k^{(v)}$ denote the calibrated or c-weight for the k-th element at the v-th iteration. At $v = 0$, $c_k^{(v)} = h_k$. The expansion estimators of population totals for variables y and x_j using c-weights at the v-th iteration are denoted by $\hat{\tau}_y^{(v)}$ and $\hat{\tau}_{xj}^{(v)}$ respectively.

The RRs are specified by the condition $L \le g_k \le U$ where $g_k = c_k/h_k$ and $L < 1 < U$, where L and U denote suitable lower and upper bounds. The adjustment factors (*i.e.*, g_k's) are also called g-weights. First we consider the unrestricted case (*i.e.*, calibration without RRs) and then the restricted case. All methods in the restricted case require iterations for finding a solution. It is assumed that the iterative process converges in a finite number of iterations.

The criterion for convergence is defined as follows. For the iterative process to meet RRs, a tolerance level ϵ (*e.g.*, .005 or .01) for family I is defined so that the process terminates if the maximum absolute relative error (ARE) for RRs is $\le \epsilon$. Similarly, a tolerance level ($\delta > 0$) for family II is defined for meeting BCs by iterations. The reason for this is that our primary goal is not minimization of the distance function, but to find a solution which satisfies BCs and RRs. In addition to ϵ and δ, a parameter v_{max} is defined which limits the number of iterations.

There are seven methods considered in this paper, two for the unrestricted case, two for restricted case in family I and the remaining three also for the restricted case but in family II. We have given alternative names to existing methods to facilitate understanding of the relationship between different methods. The naming convention is based on the well known distance measures used in the analysis of count data.

Note that since all the methods are asymptotically equivalent to the regression method, the asymptotic variance of $\hat{\tau}_y$ can be estimated for each method by $\sum_k \sum_l (\pi_{kl} - \pi_k \pi_l)\ \pi_{kl}^{-1}\ (e_k g_k)(e_l g_l)$, as in DS (equation 3.4) where π_k, π_{kl} are respectively the first and second order inclusion probabilities, e_k are the sample residuals $y_k - \hat{\boldsymbol{B}}' \boldsymbol{x}_k$ with $\hat{\boldsymbol{B}}' = (\boldsymbol{y}' \Gamma_0 \boldsymbol{X})(\boldsymbol{X}' \Gamma_0 \boldsymbol{X})^{-1}$, and Γ_0 is the $n \times n$ matrix diag($\boldsymbol{h}$).

2.1 METHOD 1 (Linear Regression or Unrestricted Modified Chi Square, MCS-u)

This method is the simplest and gives rise to the popular generalized regression estimator of Särndal (1980). Here, the model for the adjustment factor is taken to be linear in $\boldsymbol{x}$, *i.e.*, $g_k = 1 + \boldsymbol{x}_k' \lambda$, for some p-vector of model parameters λ which satisfies BCs. That is, $\sum_{k=1}^{n} h_k (1 + \boldsymbol{x}_k' \lambda) x_{kj} = \tau_{xj}$, for all j. This gives rise to $\lambda^{\text{MCS-u}}$ as $(\boldsymbol{X}' \Gamma_0 \boldsymbol{X})^{-1} (\tau_x - \hat{\tau}_x^{(0)})$. The c-weights remain close to the h-weights in the sense that the above choice of g-weights minimizes the distance function, $\Delta^{\text{MCS-u}}(\boldsymbol{c}, \boldsymbol{h}) = \sum_{k=1}^{n} (c_k - h_k)^2 / h_k$ subject to BCs. Note that the g-weights could be negative for some k. This is rather undesirable in practice although the simplicity of the method is quite attractive. The computational algorithm for MCS-u is given in Appendix A1.

2.2 METHOD 2 (Raking or Unrestricted Modified Discrimination Information, MDI-u)

This method is also commonly used. Here, the model for the adjustment factor g_k is taken as $\exp(\boldsymbol{x}_k' \lambda)$, thus making it necessarily non-negative. Unlike the case of method 1, the model parameter vector $\lambda^{\text{MDI-u}}$ is obtained iteratively to meet BCs. The iterations can be started with $\lambda^{\text{MCS-u}}$ from the GR-estimator, *i.e.*, for iteration 1, set $\lambda^{(1)} = \lambda^{\text{MCS-u}}$, which implies $c_k^{(1)} = h_k \exp(\boldsymbol{x}_k' \lambda^{(1)})$. These c-weights, in general, do not satisfy BCs. For iteration 2 of this method, the $\lambda^{(1)}$ is adjusted (by a term of smaller order) to define $\lambda^{(2)}$ as $\lambda^{(1)} + (\boldsymbol{X}' \Gamma_1 \boldsymbol{X})^{-1} (\tau_x - \hat{\tau}_x^{(1)})$, where $\Gamma_1 = \text{diag}\ (\boldsymbol{c}^{(1)})$. The λ term is defined similarly for further iterations until convergence, *i.e.*, until BCs are met. The c-weights remain close to h-weights because iterations used in the above method constitute the Newton-Raphson steps for minimizing the distance function, $\Delta^{\text{MDI-u}}(\boldsymbol{c}, \boldsymbol{h}) = \sum_{k=1}^{n} [c_k \log(c_k/h_k) - c_k + h_k]$ subject to BCs. Note that although the g-weights are non-negative, they could be very high which is clearly not desirable in practice. The computational algorithm for MDI-u is given in Appendix A2.

2.3 METHOD 3 (Modified Huang-Fuller or Scaled Modified Chi Square, SMCS)

This method belongs to family I of the restricted case and is a slight modification of the method due to Huang and Fuller as given in Singh; see also Fuller, Loughin, and Baker (1994). As in regression, the model for the adjustment factor is taken to be linear in $\boldsymbol{x}$. To facilitate the satisfaction of RRs by these adjustments, a scaling factor q_k, $(0 < q_k \le 1)$, is used for each k so that the change in h-weights for those units whose g_k's tend to go outside the bounds $[L, U]$ is reduced. Thus, the g-weight is given by $g_k = 1 + q_k \boldsymbol{x}_k' \lambda$ where the model parameters $\boldsymbol{q}$ and λ are chosen iteratively in the sense that λ is found for a given $\boldsymbol{q}$ and then $\boldsymbol{q}$ is found for a given λ. We start with $q_k^{(0)} = 1$ for all k and set $\lambda^{(1)} = \lambda^{\text{MCS-u}}$ for iteration 1. Now, clearly $\boldsymbol{c}^{(1)}$ satisfies BCs but RRs need not be satisfied. Depending on the location of g_k's in relation to $[L, U]$, a working rule can be used to define q_k's so that the q_k's discount more for those units which are farther outside of the boundaries than those which are nearer. The scaling factors $q_k^{(1)}$ so determined, define in turn $\lambda^{(2)}$ for iteration 2 as $(\boldsymbol{X}' \Gamma_1 \boldsymbol{X})^{-1} (\tau_x - \hat{\tau}_x^{(1)})$ where $\Gamma_1 = \text{diag}(q_k^{[1]} h_k)$, $q_k^{[1]} = q_k^{(0)} q_k^{(1)}$,

$\lambda^{(2)}$ satisfying BCs after the iteration. Note that under usual regularity conditions, $\lambda^{(2)}$ differs from $\lambda^{(1)}$ only by a term of smaller order, since the maximum absolute difference $|q_k^{(1)} - 1|$ is small. Next, if $c^{(2)}$ after iteration 2 does not satisfy RRs, the scaling factors $q_k^{(2)}$ are defined appropriately and compounded with $q_k^{[1]}$ to get $q_k^{[2]}$ for use in iteration 3. The $\lambda^{(3)}$ for iteration 3 is then obtained as before so that BCs are satisfied after the iteration. Iterations continue until convergence, *i.e.*, until RRs are met. The weight vector c^{SMCS} is close to h because at each iteration ν, $c^{(\nu)}$ minimizes the distance function $\Delta_\nu^{SMCS}(c,h) = \sum_{k=1}^n (c_k - h_k)^2/h_k q_k^{[\nu-1]}$ subject to BCs, where $q_k^{[\nu-1]} = q_k^{(0)} q_k^{(1)} \ldots q_k^{(\nu-1)}$ for $\nu \geq 1$. Note that unlike the previous methods, the distance function varies from iteration to iteration.

The computational algorithm for SMCS is given in Appendix A3. Note that in the algorithm, $[L, U]$ is shrunk to $[L', U']$ by means of a parameter α where $L' = \alpha L + 1 - \alpha$, $U' = \alpha U + 1 - \alpha$, and $0 < \alpha \leq 1$. This implies that some units that are inside $[L, U]$ but close to the boundary are also discounted. This helps to speed up the convergence. Another parameter β, $0 \leq \beta \leq 1$ is also introduced to allow differential discounting of different units.

2.4 METHOD 4 (Shrinkage-Minimization, SM)

This method also belongs to family I and is due to Singh. As in regression, the model for the adjustment factor is taken to be linear in x, but a new parameter termed the shrinkage factor ψ_k $(0 < \psi_k \leq 1)$ is used for each k so that g_k's meet RRs, *i.e.*, g_k is set at $(1 + \psi_k x_k' \lambda(k))$. Notice that λ is allowed to depend on k through ψ_k and x_k. Unlike SMCS, here the g-weights, after discounting, satisfy RRs exactly, *i.e.*, those g-weights which are outside $[L, U]$ are shrunk to lie on or inside the boundary. Therefore, ψ_k's can be defined quite easily in practice. The model parameters ψ and λ are chosen iteratively in a manner analogous to that for SMCS. We start with $\psi_k^{(0)} = 1$ and set $\lambda^{(1)} = \lambda^{MCS\text{-}u}$ for iteration 1 to obtain $g_k^{(1)}$ as $(1 + \psi_k^{(0)} x_k' \lambda^{(1)})$. Clearly BCs are satisfied after the iteration but RRs need not be. Before iteration 2, $g_k^{(1)}$ is shrunk by $\psi_k^{(1)}$ to obtain $g_k^{(1)*}$ as $(1 + \psi_k^{[1]} x_k' \lambda^{(1)})$ where $\psi_k^{[1]} = \psi_k^{(0)} \psi_k^{(1)}$, which meets RRs. Given $\psi^{[1]}$, $\lambda^{(2)}(k)$ is obtained as $\lambda^{(1)} + (1/\psi_k^{[1]})(X' \Gamma_1 X)^{-1} (\tau_x - \hat{\tau}_x^{(1)*}) + x_k' (X' \Gamma_1 X)^{-1} (\tau_x - \hat{\tau}_x^{(1)*}) \lambda^{(1)}$ where $\Gamma_1 = \text{diag}(c^{(1)*})$, $c_k^{(1)*} = h_k g_k^{(1)*}$, and $\hat{\tau}_x^{(1)*}$ is the expansion estimator using $c^{(1)*}$-weights. Again BCs are satisfied after the iteration but RRs need not be. Note that $\lambda^{(2)}(k)$ differs from $\lambda^{(1)}$ by a term of smaller order uniformly over k. Iterations are continued until convergence, *i.e.*, until RRs are met. The weight vector c^{SM} is close to h because at each iteration $\nu \geq 1$, $c^{(\nu)}$ minimizes the distance function, $\Delta_\nu^{SM}(c, c^{(\nu-1)*}) = \sum_{k=1}^n (c_k - c_k^{(\nu-1)*})^2/c_k^{(\nu-1)*}$ subject to BCs. Note that in practice $c^{(\nu)*}$ can be obtained directly from $c^{(\nu)}$ without having to calculate $\psi^{(\nu)}$ separately. As with SMCS, the distance function depends on the iteration.

The computational algorithm is given in Appendix A4. Recall that in the above method, if a g-weight falls outside of the L and U boundaries, an adjustment is made to bring the g-weight back to the L or U boundary. A new parameter α $(0 < \alpha \leq 1)$ is introduced to allow the user to bring the g-weight farther inside the boundary to a point L' or U' $(L' = \alpha L + 1 - \alpha, \quad U' = \alpha U + 1 - \alpha)$. This is somewhat similar to the α parameter of SMCS. Another parameter $\eta(0 < \eta \leq \alpha \leq 1)$ is introduced to adjust the g-weights to the level L' or U' also for those units which are within $[L, U]$, but close to the boundary in that they are outside $[L'', U'']$ where $L'' = \eta L + 1 - \eta$, $U'' = \eta U + 1 - \eta$. All these parameters help speed up the convergence in general.

2.5 METHOD 5 (Linear Truncated or Restricted Modified Chi Square, MCS-r)

This well known method belongs to family II of the restricted case and is due to DS. As in SM, the model for the adjustment factor is taken to be linear in x with a new parameter termed the truncation factor ϕ_k $(0 < \phi_k \leq 1)$ which is used for each k so that g_k's meet RRs, *i.e.*, g_k is set at $(1 + \phi_k x_k' \lambda(k))$. The only difference between the truncation factor ϕ_k used here and the shrinkage factor used in SM is that here those g-weights which are outside $[L, U]$ are always adjusted to lie exactly on the boundary. The model parameters are chosen iteratively. Initially we set $\phi_k^{(0)} = 1$ and at iteration 1, $\lambda^{(1)} = \lambda^{MCS\text{-}u}$ to obtain $\tilde{g}_k^{(1)} = (1 + \phi_k^{(0)} x_k' \lambda^{(1)})$, which is further adjusted (or truncated) to obtain $g_k^{(1)}$ as $(1 + \phi_k^{[1]} x_k' \lambda^{(1)})$ where $\phi_k^{[1]} = \phi_k^{(0)} \phi_k^{(1)}$, so that RRs are met. However, $g^{(1)}$ may not satisfy BCs. Note that the difference between $g^{(1)}$ and $g^{MCS\text{-}u}$ is of smaller order. Now, for iteration 2, $\lambda^{(1)}$ is adjusted by a term of smaller order (uniformly over k) to define $\lambda^{(2)}(k)$ as $\lambda^{(1)} + (1/\phi_k^{[1]})(X' \Gamma_1 X)^{-1} (\tau_x - \tau_x^{(1)})$, where $\Gamma_1 = \text{diag}(h)$ except that the diagonal elements are truncated to zero for all those k for which $\phi_k^{(1)} < 1$, *i.e.*, those units which were truncated at the previous iteration. This discounting of diagonal elements is somewhat similar to using a zero scaling factor in SMCS. In the second iteration, we have $\tilde{g}_k^{(2)} = 1 + \phi_k^{[1]} x_k' \lambda^{(2)}(k)$ and the truncation factors $\phi_k^{(2)}$ are used to obtain $g_k^{(2)}$ which satisfy RRs. The successive iterations are defined in a similar manner. Clearly, unlike SM, here RRs are met at each iteration. Iterations are continued until BCs are met. The weight vector, $c^{MCS\text{-}r}$ is close to h because the iterations defined above constitute the Newton-Raphson steps for minimizing the distance function $\Delta^{MCS\text{-}r}(c,h) = \sum_k (c_k - h_k)^2/h_k$ if $Lh_k \leq c_k \leq Uh_k$; ∞ otherwise, subject to BCs. The computational algorithm is given in Appendix A5. Note that, in practice, it is more convenient to work with $g_k^{(\nu)}$ directly without having to compute $\phi_k^{(\nu)}$ separately.

2.6 METHOD 6 (Restricted Modified Discrimination Information or MDI-r)

This method also belongs to family II and was proposed by Singh following the lines of DS in developing MCS-r. It is related to MDI-u in the same way as MCS-r is to MCS-u. The basic idea is to adjust parameters ϕ and λ in the adjustment factor $g_k = \phi_k \exp(x_k' \lambda)$ so that RRs and BCs are satisfied. The truncation parameter ϕ is similar to that for MCS-r. This

is done iteratively. Similar to MCS-r, at iteration 1 we set $\tilde{g}_k^{(1)} = \phi_k^{(0)} \exp(x_k' \lambda^{(1)})$ where $\phi_k^{(0)} = 1, \lambda^{(1)} = \lambda^{\text{MCS-u}}$, which is further adjusted by a term of smaller order to obtain $g_k^{(1)}$ as $\phi_k^{(1)} \exp(x_k' \lambda^{(1)})$ so that RRs are met, *i.e.*, it lies in $[L, U]$. Next for iteration 2, $g_k^{(1)}$ is adjusted by a term of smaller order to obtain $g_k^{(2)}$ as $\phi_k^{(2)} \exp(x_k' \lambda^{(2)})$, where $\lambda^{(2)} = \lambda^{(1)} + (X' \Gamma_1 X)^{-1} (\tau_x - \hat{\tau}_x^{(1)})$, and $\Gamma_1 = \text{diag}(h_k g_k^{(1)})$ except that the diagonal elements are truncated to 0 for all those k for which $\phi_k^{(1)} < 1$. The truncation factors $\phi_k^{(2)}$ are used to ensure that RRs are met. Iterations are continued until convergence, *i.e.*, until BCs are met. The weight vector $c^{\text{MDI-r}}$ is close to h because the iterations defined above constitute the Newton-Raphson steps for minimizing the distance function $\Delta^{\text{MDI-r}}(c, h) = \sum_{k=1}^{n} [c_k \log(c_k/h_k) - c_k + h_k]$ if $Lh_k \le c_k \le Uh_k$; ∞ otherwise, subject to BCs. Note that in practice, the truncation factors are not needed separately to compute $g_k^{(v)}$. Appendix A6 gives the computational algorithm for MDI-r.

2.7 METHOD 7 (Logit or Generalized Modified Discrimination Information, GMDI)

This is the last method considered. This well known method of family II is due to DS. As in the raking method, we start with $\exp(x_k' \lambda)$ and an inverse logit-type transformation is used to ensure that the adjustment factor satisfies RRs. The model for the adjustment factor is given by $g_k = [(U - 1) + (1 - L) \exp(A x_k' \lambda)]^{-1} [L(U - 1) + U(1 - L) \exp(A x_k' \lambda)]$, where $A = (1 - L)^{-1} (U - 1)^{-1} (U - L)$. This adjustment factor, unlike other methods, lies necessarily inside the interval $[L, U]$, *i.e.*, does not take boundary values. As $L \to 0$ and $U \to \infty$, the factor reduces to the familiar inverse logit form, $\exp(x_k' \lambda)/[1 + \exp(x_k' \lambda)]$. The model parameter λ is obtained iteratively to meet BCs. Starting with $\lambda^{\text{MCS-u}}$ as $\lambda^{(1)}$ for iteration 1, we adjust by a smaller order term to obtain $\lambda^{(2)}$ as $\lambda^{(1)} + (X' \Gamma_1 X)^{-1} (\tau_x - \hat{\tau}_x^{(1)})$ where $\Gamma_1 = \text{diag}(h_k d_k^{(1)})$, $d_k^{(1)} = (U - 1)^{-1} (1 - L)^{-1} (U - g_k^{(1)}) (g_k^{(1)} - L)$. Further iterations are done in a similar manner until BCs are met. The weight-vector c^{GMDI} is close to h in the sense that subject to BCs, the above iterative process corresponds to the Newton-Raphson algorithm for minimizing the distance function $\Delta^{\text{GMDI}}(c, h)$ given by $A^{-1} \sum_{k=1}^{n} h_k [(g_k - L) \log\{(1 - L)^{-1} (g_k - L)\} + (U - g_k) \log\{(U - 1)^{-1} (U - g_k)\}]$. Appendix A7 gives the computational algorithm for GMDI.

3. NUMERICAL EXAMPLES

3.1 Data Description

We consider application of the seven adjustment methods described above to data from the 1990 Statistics Canada's Family Expenditure (FAMEX) Survey for the two cities (or domains) of Regina and Saskatoon in the province of Saskatchewan. Four study variables are considered: annual expenditures on owned dwelling for repair and renovation, furniture and equipment, ladies' clothing , and men's clothing. The FAMEX survey is a supplementary survey to the Canadian Labour Force Survey (LFS) and, therefore, is based on the LFS design – a multistage stratified cluster sample of households, see Singh *et al.* (1990). Samples are drawn independently from the two cities of Regina and Saskatoon. Respectively for the two cities, the numbers of strata are 30 and 34, and the numbers of primary sampling units (PSUs) selected in the sample are 111 and 94. The total numbers of sampled households are 321 and 278, while the corresponding numbers (n) of individuals are 797 and 712.

3.2 Benchmark Constraints, Range Restrictions and Common Weights per Household

The number (p) of BCs is four for each domain. They correspond to the demographic population counts for the four groups: age < 15, age ≥ 15, one person households, and households with two or more persons. The corresponding counts are 40696, 139047, 12746, and 48457 for Regina, and 42544, 139299, 20628, and 52059 for Saskatoon. Thus, the total numbers of households for the two domains are 61203 and 72687 respectively and the corresponding population sizes (N) are 179743 and 181843. The auxiliary x-variables here are indicators for the above four groups.

For Regina, (min, max) of g-weights are obtained as (-0.72, 2.74) and (0.19, 3.95) respectively for regression and raking methods. It is therefore of interest to make them nonnegative for regression and to reduce the high weights for raking. Two types of RRs are chosen: one has somewhat loose bounds with $L = 1/5$ and $U = 5$ and the other has somewhat tight bounds with $L = 2/5$ and $U = 5/2$. For Saskatoon, (min, max) of g-weights are obtained as (0.86, 1.08) and (0.87, 1.09) respectively for regression and raking methods. Note that both methods give g-weights close to 1, and therefore there is no real need for RRs. However, for the sake of illustration, we choose $L = 0.88$ and $U = 1.12$.

The initial sampling weights or h-weights of individuals in the same household are common and equal to the weight of that household. It is desirable that after calibration, all members of a household have the same c-weights. This can be achieved by modifying the X matrix so that x_j-values for each person in the same household are common and equal to the average value for the household, see, *e.g.*, Lemaître and Dufour (1987). We also perform an initial scaling on the h-weights so that they add up to N; this is similar to the Hájek modification of the Horvitz-Thompson estimator. This scaling essentially redefines $[L, U]$ to make them meaningful for calibration of h-weights.

3.3 Descriptive Measures for Comparison

For comparing various methods, we consider four types of descriptive measures:

(i) Summary statistics for the distribution of the g-weights,
(ii) Point estimates for several variables,
(iii) Estimated precision of the calibration estimates, and
(iv) Computational burden imposed by each method.

The first measure consists of a graphical summary using a box plot for g-weights, and the standard deviation of g-weights, SD(g), defined as $[N^{-1} \sum_{k=1}^{n} h_k (g_k - 1)^2]^{1/2}$. Note

that the mean of g-weights, *i.e.*, $N^{-1}\sum_{k=1}^{n} h_k g_k$, is 1 in view of the fact that $\sum h_k = \sum c_k = N$, and the SD(g) also equals $[N^{-1}\sum_{k=1}^{n}(c_k - h_k)^2/h_k]^{1/2}$, the square root of a normalized chi-square type distance for measuring closeness between h- and c-weights. For comparing point estimates and their precision for estimating parameter for each variable y of interest, we compute relative difference (RD) and relative precision (RP) with respect to the MCS-u weights, *i.e.*, relative to the regression estimator. Denoting an estimator based on c-weights as a c-estimator, we have RD as (c-estimator minus regression estimator) divided by the regression estimator, and RP as SE(regression estimator) divided by SE(c-estimator). Note that for the numerical examples under consideration, variances are computed using jackknifing by deleting PSUs. Finally, the computational burden is expressed in terms of the number of iterations. Testing has shown that for all the restricted methods, each iteration takes a similar amount of time and hence a good comparison of their computational burden is the number of iterations required for convergence.

3.4 Specification of Other Parameters

We also need to specify some other parameters, namely, α, β for SMCS, and α, η for SM. Empirically, values of $\alpha = 0.67$, $\eta = 0.9$ and $\beta = 0.8$ are found to perform well. The tolerance levels ϵ for family I and δ for family II are set at 0.01, and ν_{max} is set at 10.

3.5 Results: A Descriptive Analysis

3.5.1 Distribution of g-weights

We first consider the Regina data. Figure 1 gives a box plot of the distribution of g-weights with $L = 0.4$ and $U = 2.5$. Note that there are negative g-weights (and hence negative c-weights) for MCS-u and large g-weights (which produce large c-weights) for the MDI-u method. For MCS-u, the fraction of g-weights < 0 is 4.9%, the fraction < 0.4 is 5.9%, the fraction above 2.5 is 1.25% while above 3.5 is 0%. For MDI-u, the fraction below 0.4 is 4.9%, the fraction > 2.5 is 4.3% and above 3.5 is 1.25%. Thus, both methods yield c-weights which are out of bounds with respect to RRs with tight bounds. The range restricted methods all have median g-weights between 0.65 and 0.75; the SMCS g-weights show, however, the most clustering around the median. Table 1 shows that under loose bounds, the SD(g) for each restricted method is slightly higher (about 7%) than the regression method, but for tight bounds, the difference increases to about 15% for family I and about 10% for family II.

Now for the Saskatoon data, Figure 2 gives a box plot of g-weights with $L = 0.88$ and $U = 1.12$. For both regression and raking methods, about 5.6% are below L and 0% are above U. All methods have similar interquartile range for g-weights with medians slightly above 1. Also it is seen from Table 1 that SD(g) for all the methods (restricted and unrestricted) are about the same and quite small.

Table 1
Number of Iterations and SD(g)
($\alpha = .67, \beta = .8, \eta = .9, \epsilon = \delta = .01, \nu_{max} = 10$)

Method	Regina				Saskatoon	
	$L = 0.2, U = 5.0$ (Loose bounds)		$L = 0.4, U = 2.5$ (Tight bounds)		$L = 0.88, U = 1.12$	
	Number of iterations	SD(g)	Number of iterations	SD(g)	Number of iterations	SD(g)
Family I						
SMCS	2	0.647	3	0.702	2	0.071
SM	2	0.636	4	0.689	2	0.070
Family II						
MCS-r	2	0.628	3	0.654	1	0.069
MDI-r	3	0.642	3	0.660	1	0.069
GMDI	3	0.640	3	0.659	2	0.069

Note: For the unrestricted (or no bounds) case, the number of iterations and SD(g) are: for Regina MCS-u and MDI-u are (1,0.599) and (3,0.647) respectively; for Saskatoon MCS-u and MDI-u are (1,0.070) and (1,0.069) respectively.

3.5.2 Relative Difference of Point Estimates

Tables 2(a) and (b) show that for Regina, under loose bounds RD is small for all the methods for each of the variables. In fact, it is negligible except for the variable "owned dwelling" for which it is generally under 4%. However, under tight bounds, it increases somewhat but remains small with values ranging between 1% and 5%. For Saskatoon (Table 2c), under the given bounds RD is negligible for all the methods.

3.5.3 Estimated Relative Precision of Estimates

For Regina, under loose bounds, RP is generally within 5% (of the precision of the regression estimator) for all methods and all variables except for MDI-r with the variable "ladies' clothing" for which it is lower by 9%. However, under tight bounds, RP varies more and is now generally within 9% except for SMCS and SM with the variable "Men's clothing" (RP is lower by 20%) and MDI-r for the variable "Ladies' clothing" for which RP is lower by 11%. For Saskatoon (Table 2c), under the chosen bounds RP is close to 1 for all cases.

3.5.4 Computational Burden

For Regina (Table 1), under loose bounds each method takes two or three iterations. As the bounds are tightened, most of the methods require more iterations to converge. To see how tightly the bounds could be squeezed before encountering convergence problems, three more sets of bounds were used with [L, U] = [0.425, 2.35], [0.45, 2.22] and [0.475, 2.11]. These results are not shown in the table. With ν_{max} as 10, the SM method does not converge for [0.425, 2.35]. The SMCS and GMDI methods do not converge for [0.45, 2.22] and the MCS-r and MDI-r finally have

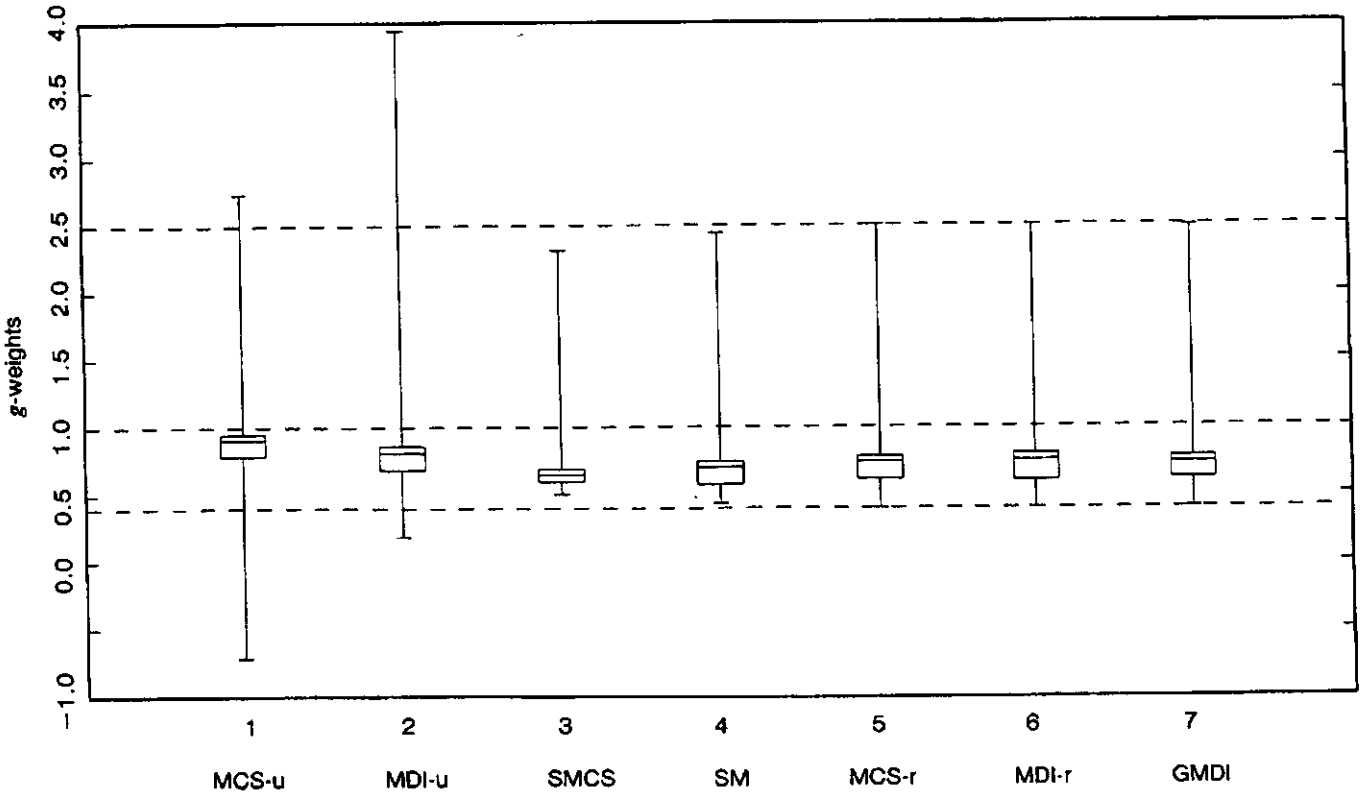


Figure 1. Box Plot: g-weights for Regina FAMEX data ($L = 0.4$, $U = 2.5$)

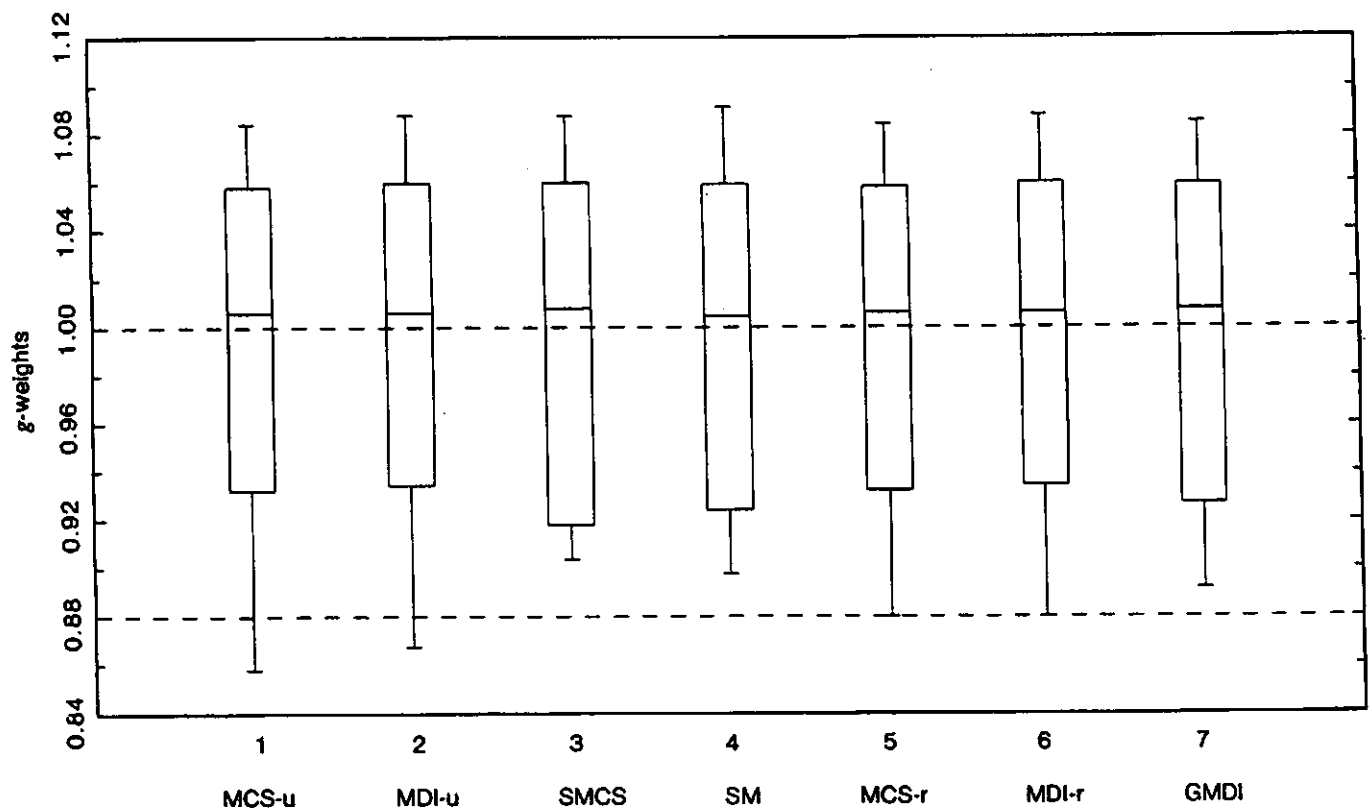


Figure 2. Box Plot: g-weights for Saskatoon FAMEX data ($L = 0.88$, $U = 1.12$)

Table 2a
Difference in Point Estimates and Precision Relative to Regression Estimator ($\alpha = .67, \beta = .8, \eta = .9, \epsilon = \delta = .01, \nu_{max} = 10$)
Regina: $L = 0.2$, $U = 5.0$ (Loose Bounds)

	Owned Dwelling		Furniture\Equipment	
	RD	RP	RD	RP
Family I				
SMCS	-0.043	1.047	0.001	1.032
SM	-0.036	1.032	-0.002	1.040
Family II				
MCS-r	-0.032	1.035	0.002	1.034
MDI-r	-0.033	0.991	-0.008	1.037
GMDI	-0.037	0.999	-0.004	1.041
	Ladies' Clothing		Men's Clothing	
Family I				
SMCS	0.015	0.931	0.009	0.952
SM	0.010	0.951	0.006	0.968
Family II				
MCS-r	0.011	0.950	0.008	0.964
MDI-r	0.007	0.911	-0.001	0.961
GMDI	0.009	0.940	0.002	0.968

Notes:
1. RD and RP denote respectively "relative difference" and "relative precision".
2. For the unrestricted (or no bounds) case, the corresponding measures for the raking (MDI-u) method relative to regression are (-0.034, 1.005), (-0.008, 1.049), (0.004, 0.968) and (0.002, 0.980) for the four study variables respectively.

Table 2b
Difference in Point Estimates and Precision Relative to Regression Estimator ($\alpha = .67, \beta = .8, \eta = .9, \epsilon = \delta = .01, \nu_{max} = 10$)
Regina: $L = 0.4$, $U = 2.5$ (Tight Bounds)

	Owned Dwelling		Furniture\Equipment	
	RD	RP	RD	RP
Family I				
SMCS	-0.056	1.100	0.012	1.000
SM	-0.055	0.992	0.017	0.919
Family II				
MCS-r	-0.048	1.073	0.008	0.952
MDI-r	-0.045	1.087	0.012	0.965
GMDI	-0.047	1.077	0.009	1.006
	Ladies' Clothing		Men's Clothing	
Family I				
SMCS	0.024	0.917	0.038	0.808
SM	0.025	0.917	0.024	0.801
Family II				
MCS-r	0.020	0.904	0.012	0.922
MDI-r	0.025	0.888	0.012	0.922
GMDI	0.021	0.938	0.018	0.917

Note: During the jackknifing procedure, the SM method failed to converge in ten iterations for four pseudo-replicates (out of a total of 111).

Table 2c
Difference in Point Estimates and Precision Relative to Regression Estimator ($\alpha = .67, \beta = .8, \eta = .9, \epsilon = \delta = .01, \nu_{max} = 10$)
Saskatoon: $L = 0.88$, $U = 1.12$

	Owned Dwelling		Furniture\Equipment	
	RD	RP	RD	RP
Family I				
SMCS	-0.001	1.001	-0.001	0.999
SM	-0.000	1.001	-0.000	0.999
Family II				
MCS-r	0.000	0.999	0.000	1.000
MDI-r	0.002	0.997	0.002	0.994
GMDI	-0.000	1.007	-0.000	0.990
	Ladies' Clothing		Men's Clothing	
Family I				
SMCS	0.000	1.013	-0.001	0.999
SM	-0.000	1.002	-0.000	0.998
Family II				
MCS-r	0.000	0.990	0.000	0.994
MDI-r	0.002	1.001	0.002	0.983
GMDI	0.000	0.977	-0.000	0.990

Notes:
1. For the unrestricted (or no bounds) case, the corresponding measures for the raking (MDI-u) method relative to regression are (0.002, 1.000), (0.002, 1.000), (0.002, 1.002) and (0.002, 0.995) for the four study variables respectively.
2. During the jackknifing procedure, the SM method failed to converge in ten iterations for two pseudo-replicates (out of a total of 94).

convergence problems for [0.475, 2.11]. For Saskatoon (Table 1), under the chosen bounds each method takes only one or two iterations. With ν_{max} as 10, as bounds are tightened to [0.92, 1.08], SM does not converge. At [0.93, 1.07], SMCS, MCS-r, and MDI-r have convergence problems, and finally at [0.96, 1.06], GMDI has problems.

4. DISCUSSION

Although numerical results for a few variables for two different domains considered in this paper are quite limited to draw general conclusions, the results based on a descriptive analysis are nevertheless interesting and may provide some indications which might be useful in practice. These can be summarized in the following observations. For loose bounds, all the restricted methods seem to perform almost at par with the regression method. However, for tight bounds, there seem to be a difference in point estimates and especially in estimated precision. This observation clearly needs further study in light of the fact that all methods are asymptotically equivalent to the regression method. A simulation study in this regard would be desirable. The recent study of Stukel, Hidiroglou, and Särndal (1996) sheds some light on this issue. Moreover, for tight bounds, there may not be convergence

under the specified number of iterations even if a solution exists. This problem may be more apparent in dealing with jackknife replicates. Therefore, caution should be exercised in choosing the maximum number of iterations for tight bounds. Finally, in practice, it is possible that even with minimal requirements on BCs and RRs, none of the calibration estimators converge within a reasonable number of iterations. In this situation, it would be of interest to investigate whether the (asymptotic) design consistency of calibration estimators could be preserved while allowing deviation from BCs. The idea of using ridge regression by Bardsley and Chambers (1984), although not in the design-based context, may be useful for this purpose. This problem is currently being investigated in collaboration with J.N.K. Rao.

APPENDIX

Here we provide computational algorithms for all seven methods of weight adjustment. These algorithms were used to write computer programs in GAUSS software for the numerical examples presented in this paper.

In all the methods, some form of the following expression denoted by the n-vector $f^{(\nu)}$, is used repeatedly for computing $c_k^{(\nu)}$ for $\nu = 1, 2, \ldots$.

$$f^{(\nu)} \equiv X (X' \Gamma_{\nu-1} X)^{-1} (\tau_x - \hat{\tau}_x^{(\nu-1)}) \qquad (1)$$

where $\Gamma_{(\nu-1)}$ is an $n \times n$ diagonal matrix defined below in the algorithm for each method. Initially $\Gamma_0 = \text{diag}(h)$ and $\hat{\tau}_x^{(0)} = \sum x_k h_k$.

A1. METHOD 1 (MCS-u)

The solution is non-iterative and is given in two steps as follows.

(i) Compute $f_k^{(1)}$, $k = 1$ to n from (1) by setting $\Gamma_{(\nu-1)} = \Gamma_0$.

(ii) Compute g_k as $1 + f_k^{(1)}$ and then $c_k^{\text{MCS-u}}$ as $h_k g_k$.

A2. METHOD 2 (MDI-u)

The solution is obtained iteratively by the following steps for $\nu = 1, 2, \ldots$.

(i) Set the tolerance level $\delta \geq 0$ for meeting BCs at some small value.

(ii) For the ν-th iteration, compute $f_k^{(\nu)}$, $k = 1$ to n, from (1) by setting $\Gamma_{\nu-1} = \text{diag}(c_k^{(\nu-1)})$.

(iii) For $\nu = 1, 2, \ldots$ compute $g_k^{(\nu)}$ as $g_k^{(\nu-1)} \exp(f_k^{(\nu)})$, $g_k^{(0)} = 1$ and then $c_k^{(\nu)}$ from $h_k g_k^{(\nu)}$.

(iv) Repeat steps (ii)-(iii) until the BCs are met up to the tolerance level δ or the number of iterations is at its maximum, ν_{max}. The last iteration gives $c_k^{\text{MDI-u}}$.

A3. METHOD 3 (SMCS)

The solution is obtained iteratively as follows.

(i) Set the RRs, *i.e.*, choose L and U, $L < 1 < U$.

(ii) Set the tolerance level $\epsilon \geq 0$ at a small value for meeting the RRs.

(iii) Choose a parameter α between 0 and 1 (*e.g.* 2/3) and set $L' = \alpha L + 1 - \alpha$, $U' = \alpha U + 1 - \alpha$. The default value of 1 for α is also allowed in which case $L' = L$, $U' = U$.

(iv) For the ν-th iteration with $g_k^{(0)} = 1$, define $\xi_k^{(\nu-1)} = (g_k^{(\nu-1)} - 1)/(L' - 1)$ if $g_k^{(\nu-1)} \leq 1$; $(g_k^{(\nu-1)} - 1)/(U' - 1)$ otherwise.

(v) Choose another parameter β between 0 and 1 (*e.g.*, 4/5). Set $q_k^{(\nu-1)} = 1$ if $\xi_k^{(\nu-1)} < 1/2$; $1 - \beta(\xi_k^{(\nu-1)} - 1/2)^2$ if $1/2 \leq \xi_k^{(\nu-1)} < 1$; $(1 - \beta/4)/\xi_k^{(\nu-1)}$ if $\xi_k^{(\nu-1)} \geq 1$ and then define for $\nu = 1, 2, \ldots$, $q_k^{[\nu-1]} = q_k^{(0)} \ldots q_k^{(\nu-1)}$ where $q_k^{(0)} = 1$. Note compounding of q-factors in defining $q_k^{[\nu-1]}$.

(vi) Compute $f_k^{(\nu)}$ from (1) by setting $\Gamma_{\nu-1} = \text{diag}(h_k q_k^{[\nu-1]})$, and $\hat{\tau}_x^{(\nu-1)} = \hat{\tau}_x^{(0)}$ for all ν.

(vii) Find $g_k^{(\nu)}$ as $1 + q_k^{[\nu-1]} f_k^{(\nu)}$ and then $c_k^{(\nu)}$ as $h_k g_k^{(\nu)}$.

(viii) Repeat steps (iv)-(vii) until the RRs are met up to the tolerance level ϵ or $\nu = \nu_{max}$. The last iteration gives c_k^{SMCS}. The value of β should remain the same at each iteration.

A4. METHOD 4 (SM)

This method consists of the following steps performed iteratively.

(i)-(ii) Same as in Method 3.

(iii) Choose parameters α, η, $0 < \alpha \leq \eta \leq 1$, (*e.g.*, $\alpha = 2/3$, $\eta = 9/10$) and define

$$L' = \alpha L + (1 - \alpha),\ U' = \alpha U + (1 - \alpha)$$
$$L'' = \eta L + (1 - \eta), U'' = \eta U + (1 - \eta).$$

The default option for α and η is 1 in which case $L' = L'' = L$, $U' = U'' = U$.

(iv) (Shrinkage). The $c_k^{(\nu)}$ from the ν-th iteration is shrunk to obtain $c_k^{(\nu)*}$ according to $c_k^{(\nu)*} = L' h_k$ if $c_k^{(\nu)} < L'' h_k$; $U' h_k$ if $c_k^{(\nu)} > U'' h_k$; $c_k^{(\nu)}$ otherwise. For $\nu = 0$, $c_k^{(0)} = c_k^{(0)*} = h_k$.

(v) (Minimization). Find $f_k^{(\nu)}$ from (1) by setting $\Gamma_{\nu-1} = \text{diag}(c_k^{(\nu-1)*})$ and $\hat{\tau}_x^{(\nu-1)} = \hat{\tau}_x^{(\nu-1)*}$.

(vi) Compute $g_k^{(\nu)}$ as $g_k^{(\nu-1)*}(1 + f_k^{(\nu)})$ where $g_k^{(\nu-1)*} = c_k^{(\nu-1)*}/h_k$ and then $c_k^{(\nu)}$ from $h_k g_k^{(\nu)}$.

(vii) Repeat steps (iv)-(vi) until the RRs are satisfied up to tolerance ϵ or $\nu = \nu_{max}$. The last iteration gives c_k^{SM}.

A5. METHOD 5 (MCS-r)

The iterative algorithm consists of the following steps.

(i) Set L and U.

(ii) Set the tolerance level $\delta \geq 0$ for meeting the BCs.

(iii) Compute $f_k^{(\nu)}$ from (1) by setting $\Gamma_{\nu-1} = \text{diag}(h_k a_k^{(\nu-1)})$ where $a_k^{(\nu-1)} = 1$ if $g_k^{(\nu-1)}$ was truncated to L or U, and 0 otherwise.

(iv) Set $g_k^{(0)} = 1$ and compute $g_k^{(\nu)}$ as $g_k^{(\nu-1)} + f_k^{(\nu)}$ if $L \leq g_k^{(\nu)} \leq U$; otherwise truncate $g_k^{(\nu)}$ to L or U as the case may be, and then $c_k^{(\nu)}$ as $h_k g_k^{(\nu)}$.

(v) Repeat steps (iii)-(iv) until BCs are met at the tolerance level δ or $\nu = \nu_{max}$. The last iteration gives $c_k^{\text{MCS-r}}$.

A6. METHOD 6 (MDI-r)

The iterative algorithm consists of the following steps.

(i)-(ii) Same as in Method 5.

(iii) Compute $f_k^{(v)}$ from (1) by setting $\Gamma_{v-1} = \text{diag}(c_k^{(v-1)} a_k^{(v-1)})$ where $a_k^{(v-1)}$ is defined as in Step (iii) of Method 5.

(iv) Set $g_k^{(0)} = 1$ and compute $g_k^{(v)} = g_k^{(v-1)} \exp(f_k^{(v)})$ if $L \le g_k^{(v)} \le U$; otherwise truncate $g_k^{(v)}$ to L or U as the case may be, and then $c_k^{(v)}$ as $h_k g_k^{(v)}$.

(v) Repeat steps (iii)-(iv) until BCs are satisfied at tolerance δ or $v = v_{max}$. The last iteration gives c_k^{MDI-r}.

A7. METHOD 7 (GMDI)

The iterative algorithm consists of the following steps.

(i)-(ii) Same as in Method 5.

(iii) Compute $f_k^{(v)}$ from (1) by setting $\Gamma_{v-1} = \text{diag}(h_k d_k^{(v-1)})$ where $d_k^{(v-1)}$ is analogous to $d_k^{(1)}$ of Section 2.7.

(iv) Using $\boldsymbol{x}_k' \lambda^{(v)} = \boldsymbol{x}_k' \lambda^{(v-1)} + f_k^{(v)}$, find $g_k^{(v)}$ from the formula for g_k given in Section 2.7, and then $c_k^{(v)}$ as $h_k g_k^{(v)}$.

(v) Repeat steps (iii)-(iv) until BCs are met at tolerance δ or $v = v_{max}$. The last iteration gives c_k^{GMDI}.

ACKNOWLEDGEMENTS

The authors are grateful to C.-E. Särndal for helpful discussions and to U. Nevraumont for providing the FAMEX data. The authors are also grateful to referees for very useful comments. The first author's research was supported in part by an NSERC grant held at Carleton University, Ottawa.

REFERENCES

BARDSLEY, P., and CHAMBERS, R.L. (1984). Multipurpose estimation from unbalanced samples. *Applied Statistics*, 33, 290-299.

DEVILLE, J.-C., and SÄRNDAL, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.

DEVILLE, J.-C., SÄRNDAL, C.-E., and SAUTORY, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, 88, 1013-1020.

HUANG, E.T., and FULLER, W.A. (1978). Nonnegative regression estimation for sample survey data. *Proceedings of the Social Statistics Section, American Statistical Association*, 300-305.

FULLER, W.A., LOUGHIN, M.M., and BAKER, H.D. (1994). Regression weighting in the presence of nonresponse with application to the 1987-88 Nationwide Food Consumption Survey. *Survey Methodology*, 20, 75-85.

LEMAÎTRE, G., and DUFOUR, J. (1987). An integrated method for weighting persons and families. *Survey Methodology*, 13, 199-207.

SÄRNDAL, C.-E. (1980). On π-inverse weighting versus best linear unbiased weighting in probability sampling. *Biometrika*, 67, 639-650.

SINGH, A.C. (1993). On weight adjustment in survey sampling. Discussion paper for the 18th meeting of the Advisory Committee on Statistical Methods, Statistics Canada, Ottawa, October 25-26.

SINGH, A.C., and MOHL, C.A. (1997). Calibration estimators with application to FAMEX survey and computer program documentation. Methodology Branch Working Paper, Statistics Canada.

SINGH, M.P., DREW, J.D., GAMBINO, J.G., and MAYDA, F. (1990). *Methodology of the Canadian Labour Force Survey: 1984-1990*. Catalogue No. 71-526, Statistics Canada.

STUKEL, D.M., and BOYER, R. (1992). Calibration estimation: an application to the Canadian Labour Force Survey. Methodology Branch Working Paper, SSMD 92-009E, Statistics Canada.

STUKEL, D.M., HIDIROGLOU, M.A., and SÄRNDAL, C.-E. (1996). Variance estimation for calibration estimators: A comparison of jackknifing versus Taylor linearization. *Survey Methodology*, 22, 117-125.

Survey Methodology, December 1996
Vol. 22, No. 2, pp. 117-125
Statistics Canada

Variance Estimation for Calibration Estimators: A Comparison of Jackknifing Versus Taylor Linearization

DIANA M. STUKEL, MICHAEL A. HIDIROGLOU and CARL-ERIK SÄRNDAL[1]

ABSTRACT

The use of auxiliary information in estimation procedures in complex surveys, such as Statistics Canada's Labour Force Survey, is becoming increasingly sophisticated. In the past, regression and raking ratio estimation were the commonly used procedures for incorporating auxiliary data into the estimation process. However, the weights associated with these estimators could be negative or highly positive. Recent theoretical developments by Deville and Särndal (1992) in the construction of "restricted" weights, which can be forced to be positive and upwardly bounded, has led us to study the properties of the resulting estimators. In this paper, we investigate the properties of a number of such weight generating procedures, as well as their corresponding estimated variances. In particular, two variance estimation procedures are investigated via a Monte Carlo simulation study based on Labour Force Survey data; they are Jackknifing and Taylor Linearization. The conclusion is that the bias of both the point estimators and the variance estimators is minimal, even under severe "restricting" of the final weights.

KEY WORDS: Auxiliary information; Raking ratio estimators; Regression estimators; Restricted weighting.

1. INTRODUCTION

Auxiliary information has many uses in survey sampling. One typical use is its incorporation at the estimation stage through the use of regression estimators or raking ratio estimators. For these estimators, a unit's sampling weight is multiplied by an adjustment factor to produce the final weight. A well-known shortcoming associated with the regression estimator is that some of the adjustment factors may be negative, resulting in negative final weights. On the other hand, for the raking ratio estimator, some adjustment factors may be very large and positive, resulting in unduly large final weights. These shortcomings can be overcome by considering a family of estimators, known as "calibration estimators". Developed by Deville and Särndal (1992), the estimators in this family incorporate auxiliary information, and in certain cases, non-negative weights can be ensured by prespecifying lower and upper bounds on the weights. These "calibration" weights are obtained by minimizing functions which measure the distances between original sampling weights and final calibrated weights, while respecting a set of benchmarking constraints. Huang and Fuller (1978) and Singh and Mohl (1996) have developed similar estimators which maintain the above properties. Ordinarily, there are very small differences between the point estimates corresponding to the various distance functions.

Historically, Statistics Canada's Labour Force Survey (LFS) has used, at different points in time, both the Taylor and Jackknife variance estimation techniques in tandem with regression and raking ratio estimators. Recently, the LFS has also allowed for the option of using other calibration estimators in addition to the previously available regression estimator, to eliminate the problem of potential negative weights. It is therefore of interest to investigate the behaviour of these point estimators and their corresponding Taylor and Jackknife variance estimators, particularly for those estimators that allow bounding on the weights. Therein lies the main focus of this paper. Now, both the Taylor and the Jackknife have their advantages. The Taylor method is computationally much less intensive than the Jackknife method, but requires working out new expressions for each different parameter that is considered; this is particularly a burden in multipurpose surveys where many different parameters may be of interest. On the other hand, for the Jackknife method, cumbersome variance expressions need not be derived for each new parameter; only the functional form of the point estimator itself is required.

The paper is structured as follows: section 2 provides the theoretical underpinnings of calibration estimation and introduces a family of related distance functions. In section 3, variances for calibration estimators are discussed. Section 4 provides the results of a Monte Carlo simulation study, in which the bias of both the point estimators and their corresponding Taylor and Jackknife variance estimators (relative to a "true" variance) is tracked, for a variety of distance functions from calibration theory. In section 5, some concluding remarks are made.

2. DISTANCE FUNCTIONS AND CALIBRATION ESTIMATORS

We begin by introducing the basic idea behind calibration estimation. Let $U = \{1, ..., k, ..., N\}$ denote the index set for

[1] Diana M. Stukel, Household Survey Methods Division, and Michael A. Hidiroglou, Business Survey Methods Division, Statistics Canada, Ottawa, Ontario, K1A 0T6; Carl-Erik Särndal, Département de Mathématiques et de Statistique, Université de Montréal, C.P. 6128, Succursale A, Montréal, P.Q., H3C 3J7.

the N units of a finite population of units. In survey sampling, one is often interested in estimating parameters of a finite population such as totals, means and ratios. For the sake of simplicity, we will focus on totals, although the ideas presented in this paper may easily be extended to include other parameters. Thus, suppose the objective is to estimate the population total $Y = \sum_{k\in U} y_k$, where y_k is the value of y, the variable of interest for the k-th population unit.

A probability sample s is drawn from U by a given sampling design which induces the inclusion probabilities $\pi_k = P(k\in s)$. These are assumed known and positive. Let $a_k = 1/\pi_k$ be the sampling weight associated with the k-th unit. Finally, let the auxiliary information be specified in the form of known population totals of one or more auxiliary variables.

An elementary estimator of Y is the Horvitz-Thompson (HT) estimator:

$$\hat{Y}_a = \sum_{k\in s} a_k y_k.$$

The HT estimator possibly but not necessarily (depending on the sampling design) incorporates auxiliary information at the design stage only; what is sought is an improved estimator which incorporates the auxiliary information at the estimation stage, as well. The incorporation of auxiliary information can be reflected in the creation of new weights, denoted by w_k; $k\in s$. The new estimator is then of the form:

$$\hat{Y}_w = \sum_{k\in s} w_k y_k. \tag{2.1}$$

The approach of Deville and Särndal (1992) and Deville, Särndal and Sautory (1993) involves determining these new weights $\{w_k; k\in s\}$ by making them as close as possible to the original sampling weights $\{a_k; k\in s\}$ according to a specified distance function. Constraints placed on the new weights are such that, when applied to each of the auxiliary variables, the known population total X is reproduced. That is,

$$\sum_{k\in s} w_k x_k = X \tag{2.2}$$

is required to hold, leading to a problem in constrained minimization. Here $x_k' = (x_{1k}, x_{2k}, ..., x_{pk})$ is a vector of length p containing the values of the auxiliary variables for the k-th individual, and the auxiliary information available from an external source is summarized by the known vector total $X = \sum_{k\in U} x_k$.

We denote the distance from w_k to a_k by $F^*(w_k, a_k)$. Deville and Särndal (1992) limit their discussions to distance functions of the form $F^*(w_k, a_k) = a_k c_k F(w_k/a_k)$ where $w_k/a_k = g_k$, the ratio of the final calibrated weight to original sampling weight, is called the "g-factor". Here c_k is a known positive weight unrelated to a_k; the uniform weighting $c_k = 1$ is often used in applications. Note that equation (2.1) can alternatively be written as:

$$\hat{Y}_w = \sum_{k\in s} a_k g_k y_k.$$

It is assumed that F is non-negative and convex, and that $F(1) = 0$, implying that when $w_k = a_k$ the distance between the weights is zero. Moreover, it is required that F' is continuous, one-to-one, and that $F'(1) = 0$ and $F''(1) > 0$ which makes $w_k = a_k$ a local minimum. (See Deville, Särndal and Sautory 1993.) The total distance, $\sum_{k\in s} a_k c_k F(w_k/a_k)$, is minimized subject to the constraint (2.2). That is,

$$\sum_{k\in s} a_k c_k F(w_k/a_k) - \lambda'\left(\sum_{k\in s} w_k x_k - X\right)$$

is minimized with respect to the w_k, where λ is a p-vector of Lagrange multipliers. Differentiating with respect to w_k, equating to zero, and solving for w_k leads to the calibrated weights $w_k = a_k g_k = a_k g(\lambda' x_k/c_k)$ where g is the inverse function of f and $f(z) = dF(z)/dz$. To compute w_k, one must first obtain λ as the solution of the calibration equation implied by (2.2), namely,

$$\sum_{k\in s} a_k g(\lambda' x_k/c_k) x_k = X. \tag{2.3}$$

The solution of this (possibly) nonlinear system of p equations in p unknowns may require the use of some iterative procedure, such as the Newton-Raphson method.

A number of distance functions are considered by Deville and Särndal (1992), Huang and Fuller (1978) and Singh and Mohl (1996). Two important distance functions which we first discuss are the Generalized Least Squares (GLS) distance function and the Raking Ratio (RR) distance function, both given in Deville and Särndal(1992).

The GLS distance function is defined by:

$$F^*(w_k, a_k) = F^*_{GLS}(w_k, a_k)$$
$$= c_k(w_k - a_k)^2/a_k = a_k c_k (w_k/a_k - 1)^2. \tag{2.4}$$

It generates the well-known generalized regression estimator (GREG), which encompasses as special cases the ratio estimator, the simple regression estimator, and the simple post-stratified estimator, among others. It follows from (2.3) that the calibrated weights corresponding to the GLS distance function are:

$$w_k = a_k g_k = a_k[1 + (X - \hat{X}_a)'\left(\sum_{j\in s} a_j x_j x_j'/c_j\right)^{-1} x_k/c_k]$$

where $\hat{X}_a = \sum_{k\in s} a_k x_k$ is the HT estimator of X. The corresponding estimator of Y can be written in the usual regression estimator form as

$$\hat{Y}_{w(GREG)} = \hat{Y}_a + (X - \hat{X}_a)'\hat{\beta} \tag{2.5}$$

where

$$\hat{\beta} = \left(\sum_{k\in s} a_k x_k x_k'/c_k\right)^{-1} \sum_{k\in s} a_k x_k y_k/c_k. \tag{2.6}$$

Thus, the regression estimator can be thought of as the HT estimator plus an adjustment term. A drawback of the GLS distance function is that it may give rise to negative weights, particularly if the system is overconstrained. In practice, negative weights are rare; however, it is desirable to eliminate them entirely since it may be difficult to give them any meaningful interpretation.

The Raking Ratio (RR) distance function is defined by:

$$F^*(w_k,a_k) = F^*_{RR}(w_k,a_k)$$
$$= c_k[w_k\log(w_k/a_k) - w_k + a_k] \qquad (2.7)$$
$$= a_k c_k[(w_k/a_k)\log(w_k/a_k) - (w_k/a_k) + 1].$$

Solving for g-factors using the RR distance function and the constraint defined by equation (2.3) can be shown to be equivalent to using the Iterative Proportional Fitting (IPF) algorithm of Deming and Stephan (1940) when calibrating on known marginals of frequency tables of dimension two or higher. Unlike the GLS distance function, which has a closed form solution, the calibration equations for the RR distance function can only be solved iteratively. Computer software exists for this purpose; for example, the CALMAR software (see Deville, Särndal and Sautory 1993) solves the calibration equations for the RR distance function using the Newton-Raphson method, rather than the IPF algorithm originally proposed by Deming and Stephan. The RR distance function always ensures positive weights; however, it also has the undesirable property that some of the resulting calibration weights can be excessively large.

Neither the possibility of negative weights produced by the GLS distance function nor the possibility of large positive weights produced by the RR distance function are desirable. One can define restricted distance functions whereby the range of the resulting weights w_k are limited. This is achieved by imposing restrictions on the distance function $F(w_k/a_k)$ in such a way that the g-factors $g_k = w_k/a_k$ are bounded within a prespecified interval. To this end, one can specify a lower bound L and an upper bound U, such that $L < 1 < U$. To guarantee positive weights, one would choose $L > 0$. Now, Deville and Särndal (1992) define restricted versions of the two distance functions given above; they are: the Restricted GLS (RGLS) distance function and the Restricted Raking Ratio (RRR) or Logit distance function. Two other methods of restricting final weights are proposed by Huang and Fuller (1978) and Singh and Mohl (1996). All four restricted distance functions are considered in this paper; they are also discussed in detail in Singh and Mohl (1996), but from a different perspective.

The Restricted GLS distance function is defined by:

$$F^*(w_k,a_k) = F^*_{RGLS}(w_k,a_k) = \begin{cases} c_k(w_k - a_k)^2/a_k & \text{if } L < w_k/a_k < U \\ \infty & \text{otherwise.} \end{cases} \qquad (2.8)$$

The Restricted RR (or Logit) distance function is defined by:

$$F^*(w_k,a_k) = F^*_{RRR}(w_k,a_k) = \begin{cases} A^{-1}c_k[(w_k/a_k - L)\log[(w_k/a_k - L)/(1 - L)] \\ \quad + (U - w_k/a_k)\log[(U - w_k/a_k)/(U - 1)]] \\ \qquad\qquad \text{if } L < w_k/a_k < U \\ \infty \quad \text{otherwise} \end{cases} \qquad (2.9)$$

where $A = (U - L)/\{(1 - L)(U - 1)\}$. The specification $L = 0$, $U = \infty$ gives the RR distance function. It is easy to show that the Restricted GLS and Restricted RR distance functions share the property that the corresponding weights w_k satisfy $L < w_k/a_k < U$.

Now, Huang and Fuller (1978) propose a method for adjusting regression weights such that the calibration constraints given by equation (2.2) are satisfied and such that the g-factors are restricted to lie close to one. Singh and Mohl (1996) show that their method can be written in terms of minimizing a distance function which changes from iteration to iteration. Singh and Mohl also modify the original method to allow for arbitrary restrictions on the g-factors, similar to the restricted distance functions above, and show that the estimator resulting from the modified distance function is asymptotically equivalent to the regression estimator. The Modified Huang-Fuller (MHF) distance function is given by:

$$F^*(w_k^{(v-1)}, a_k) = F^{(v)*}_{MHF}(w_k^{(v-1)}, a_k)$$
$$= (w_k^{(v-1)} - a_k)^2/a_k q_k^{(v-1)*}; \; v = 1,2,\ldots \qquad (2.10)$$

where $q_k^{(v-1)*} = q_k^{(v-1)}\ldots q_k^{(1)} q_k^{(0)}$ with $q_k^{(0)} = 1$ and where v is the iteration number. Here,

$$q_k^{(v-1)} = \begin{cases} 1 & \text{if } \xi_k^{(v-1)} < .5 \\ 1 - \delta(\xi_k^{(v-1)} - .5)^2 & \text{if } .5 \le \xi_k^{(v-1)} < 1 \\ (1 - \delta/4)/\xi_k^{(v-1)} & \text{if } \xi_k^{(v-1)} \ge 1 \end{cases}$$

for δ arbitrarily chosen such that $0 < \delta < 1$. Also

$$\xi_k^{(v-1)} = \begin{cases} (g_k^{(v-1)} - 1)/(L' - 1) & \text{if } g_k^{(v-1)} \le 1 \\ (g_k^{(v-1)} - 1)/(U' - 1) & \text{otherwise} \end{cases}$$

where $L' = \alpha L + 1 - \alpha$ and $U' = \alpha U + 1 - \alpha$ for α arbitrarily chosen such that $0 < \alpha < 1$ and L and U are as in earlier restricted distance functions. The parameters α and δ serve to speed up the convergence of the iterative algorithm used to provide a solution. Singh and Mohl (1996) empirically test a variety of values for these parameters using large data sets, and suggest that $\alpha = .67$ and $\delta = .8$ work well in practice. Finally, the g-factor at each iteration is

$$g_k^{(v-1)} = 1 + (X - \hat{X}_w^{(v-2)})' \left(\sum_{j \in s} a_j q_j^{(v-2)*} x_j x_j' \right)^{-1} x_k; \quad v = 2,3,...$$

where $\hat{X}_w^{(v-2)} = \sum_{k \in s} w_k^{(v-2)} x_k$; $v = 2, 3, ...$ and where $w_k^{(v-2)} = a_k g_k^{(v-2)}$; $v = 2, 3, ...$ Starting values are given by $g_k^{(0)} = 1$ and $w_k^{(0)} = a_k$.

Singh and Mohl (1996) also propose a new distance function which changes from iteration to iteration called the Shrinkage-Minimization (SM) distance function, and show that the estimator resulting from this distance function is also asymptotically equivalent to the regression estimator. It is given by:

$$F^*(w_k^{(v-1)}, a_k) = F_{SM}^{(v)*}(w_k^{(v-1)}, a_k) = (w_k^{(v-1)} - a_k^{(v-1)*})^2 / a_k^{(v-1)*}; \quad v = 1, 2, ... \quad (2.11)$$

where

$$a_k^{(v-1)*} = \begin{cases} L' a_k & \text{if } w_k^{(v-1)} < L'' a_k \\ U' a_k & \text{if } w_k^{(v-1)} > U'' a_k \\ w_k^{(v-1)} & \text{otherwise.} \end{cases} \quad v = 2, 3, ...$$

Terms in the above equations are defined as follows: $L' = \alpha L + (1 - \alpha)$, $U' = \alpha U + (1 - \alpha)$, $L'' = \eta L + (1 - \eta)$ and $U'' = \eta U + (1 - \eta)$ for α and η arbitrarily chosen such that $0 < \alpha < \eta \leq 1$. As before, the parameters α and η serve to speed up the convergence of the iterative algorithm used to provide a solution; Singh and Mohl (1996) suggest that $\alpha = .67$ and $\eta = .9$ work well in practice. Finally, $w_k^{(v-1)} = a_k g_k^{(v-1)}$; $v = 2, 3, ...$ where

$$g_k^{(v-1)} = \frac{a_k^{(v-2)*}}{a_k} \left[1 + (X - \hat{X}_w^{(v-2)})' \left(\sum_{j \in s} a_j^{(v-2)*} x_j x_j' \right)^{-1} x_k \right]; \quad v = 2,3,...$$

and where $\hat{X}_w^{(v-2)}$ is as before. Starting values are given by $a_k^{(0)*} = a_k$ and $w_k^{(0)} = a_k$.

A property of the Modified Huang-Fuller and Shrinkage-Minimization distance functions is that the calibration constraints (equation (2.2)) are met at every iteration whereas the range restrictions on the g-factors are met only upon convergence. For the Restricted GLS and Restricted Raking Ratio distance functions, the range restrictions on the g-factors are met at every iteration whereas the calibration constraints are only met upon convergence. Now, it is often useful to specify an upper bound on the number of iterations to convergence; this feature may be programmed into the iterative algorithm for operational expediency. If this upper bound is exceeded due to slow convergence, the iterative algorithm may be terminated prematurely. Regardless, for the Modified Huang-Fuller and Shrinkage-Minimization distance functions, the calibration constraints will be met. Likewise, for the Restricted GLS and Restricted Raking Ratio distance functions, the range restrictions will be met.

Now, the behaviour of the g-factors from some of the distance functions has been studied extensively; see, for example, Deville, Särndal and Sautory (1993). Stukel and Boyer (1992) empirically show that the GLS and RR distance functions, as well as their restricted counterparts having loose bounds imposed on them, give g-factors whose distributions over a given data set adhere to normality rather closely. However, as the bounds on the restricted distance functions are squeezed together more closely, the distributions exhibit a "pile-up" of g-factors at the lower and upper bounds. Regardless, even under extreme squeezing, the restricted distance functions seem to give point estimates that are close to their unrestricted counterparts, as the results of our empirical study will verify. However, the biases of both the point and variance estimators under extreme squeezing on the restricted distance functions have not been investigated. This investigation is of interest to surveys such as the LFS, where an augmentation to the current estimation system has been implemented, which now allows users the option of choosing from amongst the Restricted GLS distance function and the Shrinkage-Minimization distance function, in addition to the previously available GLS distance function.

3. VARIANCE ESTIMATION FOR CALIBRATION ESTIMATORS

The exact variance of the calibration estimator $\hat{Y}_w$ is intractable since the point estimator itself is nonlinear. In addition, there is no explicit unbiased method of variance estimation. Therefore, approximately unbiased methods, such as the Taylor and the Jackknife, are often used in practice.

Now, for stratified multistage designs, "with replacement" sampling is not often used in practice since the possibility of drawing the same unit more than once is unappealing. Therefore, the preponderance of surveys use "without replacement" sampling, at least at the first stage of sampling. Even so, if the first stage sampling fraction is small (say, less than 10 percent as a rule of thumb), it may be reasonable to use a simplified variance formula that assumes "with replacement" sampling at the first stage of sampling. For the generalized regression estimator (GLS distance function) under a stratified multi-stage design this simplification of the variance estimator yields:

$$\hat{V}_T^*(\hat{Y}_{w(\text{GREG})}) = \sum_{h=1}^{L} \frac{n_h}{n_h - 1} \sum_{i=1}^{n_h} \left[\sum_{k \in s_{hi}} a_{hik} e_{hik} - \frac{1}{n_h} \sum_{i=1}^{n_h} \sum_{k \in s_{hi}} a_{hik} e_{hik} \right]^2 \quad (3.1)$$

where s_{hi} is the sample of individuals in the i-th primary sampling unit (PSU) and the h-th stratum, a_{hik} is the original sampling weight under the stratified multi-stage design for

sampled individual k in PSU i and stratum h, and n_h is the number of sampled PSUs in stratum h. Also $e_{hik} = y_{hik} - x'_{hik}\hat{\beta}$ is the estimated residual associated with the regression estimator where $\hat{\beta} = (\sum_{hik\epsilon s} a_{hik} x_{hik} x'_{hik}/c_{hik})^{-1} \sum_{hik\epsilon s} a_{hik} x_{hik} y_{hik}/c_{hik}$. For many designs, the "with replacement" formula given by (3.1) overestimates the true variance (see Särndal, Swensson and Wretman 1992, section 4.6). Note that although, technically speaking, this simplified variance estimator is *not* the Taylor variance estimator, it is often referred to as such for historical reasons and so will it be in this paper.

An improvement to equation (3.1), which includes the g-factor in the variance formula (recall that $w_{hik} = a_{hik} g_{hik}$), is suggested by Hidiroglou, Fuller and Hickman (1980). It is given by:

$$\hat{V}_T(\hat{Y}_{w(\mathrm{GREG})}) = \sum_{h=1}^{L} \frac{n_h}{n_h - 1} \sum_{i=1}^{n_h} \left[\sum_{k\epsilon s_{hi}} w_{hik} e_{hik} - \frac{1}{n_h} \sum_{i=1}^{n_h} \sum_{k\epsilon s_{hi}} w_{hik} e_{hik} \right]^2. \quad (3.2)$$

An analogue of equation (3.2) is also suggested by Särndal (1982) in the context of two-stage sampling, but for Yates-Grundy type variance estimators. Now Deville and Särndal (1992) show that any distance function which obeys a set of general conditions will produce an estimator that is asymptotically equivalent to the one produced by the GLS distance function, that is, $\hat{Y}_{w(\mathrm{GREG})}$ given by (2.5). Singh and Mohl (1996) extend this result to include the Modified Huang-Fuller and Shrinkage-Minimization distance functions. As a result, the asymptotic variance of the calibration estimator $\hat{Y}_w$ can be considered to be roughly equal to that of $\hat{Y}_{w(\mathrm{GREG})}$. This observation leads to a method for estimating the Taylor variance which is common to all calibration estimators, namely, to estimate the variance of $\hat{Y}_w$ using a modification of the Taylor variance estimator employed for $\hat{Y}_{w(\mathrm{GREG})}$, rather then rederiving the Taylor formula for each of the distance functions separately. Thus, whenever a variance estimator associated with a distance function different from the GLS is required, equation (3.2) is used, replacing the final weights $\{w_{hik}\}$ from the GLS distance function with those from the distance function in question.

It is straightforward to apply the Jackknife procedure to obtain a variance estimator for $\hat{Y}_w$, regardless of the distance function used to obtain the final calibrated weights. An expression for the variance formula under a stratified multi-stage design using with replacement sampling at the first stage is given by:

$$\hat{V}_J(\hat{Y}_w) = \sum_{h=1}^{L} \frac{n_h - 1}{n_h} \sum_{i=1}^{n_h} (\hat{Y}_w(hi) - \hat{Y}_w)^2 \quad (3.3)$$

where $\hat{Y}_w(hi)$ is often referred to as the "replicate estimator"; "replicates" are formed by taking what remains of the sample after removing PSU i from stratum h. Thus, $\hat{Y}_w(hi)$ is calculated by recomputing $\hat{Y}_w$ after removing the i-th PSU from the h-th stratum, $h = 1, ..., L$; $i = 1, ..., n_h$, *i.e.*, with the original sampling weights altered to reflect the PSU removal and the g-factors recalculated based on the reduced sample or replicate. Finally, the Jackknife estimator is constructed by repeatedly removing PSUs one at a time, calculating the corresponding replicate estimator, and then assembling the final estimator using (3.3). The Jackknife variance estimator given by (3.3) is the most conservative among the four variations suggested in the extensive discussion on the subject by Wolter (1985).

It is interesting to note that, for the GREG estimator, Yung and Rao (1996) obtain (3.2) as an approximation to the Jackknife variance estimator given by (3.3); they call (3.2) the "Jackknife Linearization Variance Estimator". Their simulation study shows that biases (both conditional and unconditional) of the Taylor variance estimator (equation (3.1)), the Jackknife Linearization variance estimator (equation (3.2)) and the Jackknife variance estimator (equation (3.3)) behave similarly. While their simulation focuses on variance estimators for the unrestricted GREG estimator, our simulation study, which we discuss next, focuses on variance estimators for the GREG as well as for estimators based on other restricted and unrestricted distance functions.

4. MONTE CARLO SIMULATION STUDY

4.1 Design of the Study

In order to compare the performance of the calibration estimators and their corresponding Taylor and Jackknife variance estimators, we undertook a Monte Carlo simulation study, in which we investigated their finite sample design-based frequentist properties.

December 1990 Labour Force Survey (LFS) sample data for the province of Newfoundland was used to simulate a finite population, from which repeated samples were drawn. The LFS is the largest ongoing household sample survey conducted by Statistics Canada. Monthly data relating to the labour market is collected using a complex multi-stage sampling design with several levels of stratification. The details of the design of the survey prior to the 1991 redesign can be found in Singh, Drew, Gambino and Mayda (1990). In general, provinces are stratified into "economic regions", which are large areas of similar economic structure; Newfoundland has four such economic regions. The economic regions are further substratified into "self-representing units" (SRUs) and "non self-representing units" (NSRUs), which are, in turn, further substratified into lower level substrata. SRUs are cities whose population exceeds 15,000, such as St. John's and Cornerbrook, in the case of Newfoundland. Now, the lowest level of stratification in Newfoundland yielded 45 strata, each of which contained less than 6 primary sampling units (PSUs), which was an insufficient number from which to sample, for the purposes of the simulation. Thus, the 45 strata were collapsed down to 18, each containing between 6 and 18 PSUs. In collapsing the strata,

economic regions were kept intact, as were the Census Metropolitan Areas (CMAs) of St. John's and Cornerbrook.

For the Monte Carlo study, $R = 4{,}000$ samples, each of size approximately 1,000, were drawn from the Newfoundland "population" (which was of size 9,152), according to a two-stage design. For collapsed strata belonging to NSRUs, two PSUs were selected at the first stage using Probability Proportional to Size (PPS) with replacement (WR) sampling, where the size measure used was the number of dwellings in the PSU. At the second stage, one in five dwellings were selected from the sampled PSUs using Simple Random Sampling (SRS) without replacement (WOR). For collapsed strata belonging to SRUs, three PSUs were selected at the first stage using PPS WR sampling. At the second stage, all the dwellings in the sampled PSUs were selected, reducing this part of the design to one-stage take-all cluster sampling. This feature was necessary since there were not enough dwellings per PSU to subsample in SRUs. The selection of two PSUs in NSRU strata versus three in SRU strata was driven by the fact that, in general, NSRU strata had fewer population PSUs from which to sample than did SRU strata. In all, there were 47 sampled PSUs. In either case (NSRUs or SRUs), all dwelling members were included in the sample. Although this design is a hybrid between a one and two-stage design, we shall refer to it as a two-stage design, for convenience.

We took Y, the total number of unemployed, to be the parameter of interest. This was calculated from the finite population by: $Y = \sum_{k \in U} y_k = \sum_{k=1}^{9152} y_k$ where $y_k = 1$ if individual k was unemployed; 0 otherwise. For each of the $R = 4{,}000$ samples, we calculated $\hat{Y}_w$, the estimated total number of unemployed as $\hat{Y}_w = \sum_{k \in s} w_k y_k$. The $\{ w_k : k \in s \}$ were determined by the following six distance functions discussed earlier:

(1) the Generalized Least Squares (GLS) Distance Function (equation (2.4)),
(2) the Raking Ratio (RR) Distance Function (equation (2.7)),
(3) the Restricted GLS (RGLS) Distance Function (equation (2.8)),
(4) the Restricted RR (RRR) or Logit Distance Function (equation (2.9)),
(5) the Modified Huang-Fuller (MHF) Distance Function ($\alpha = .67$, $\delta = .8$) (equation (2.10)), and
(6) the Shrinkage-Minimization (SM) Distance Function ($\alpha = .67$, $\eta = .9$) (equation (2.11)).

For the latter four distance functions, the following four sets of bounds were imposed on each to restrict the minimization: (i) $L = 0$, $U = 4$, (ii) $L = .4$, $U = 2$, (iii) $L = .68$, $U = 1.6$ and (iv) $L = .8$, $U = 1.3$. This yielded a total of eighteen point estimators. For each of the eighteen point estimators, the calibration used auxiliary information based on Census projections at the province level for 10 mutually exclusive and exhaustive age/sex categories (age categories: < = 14, 15-24, 25-44, 45-64, > = 65 crossed with the two sexes) and the four economic regions of Newfoundland. Thus, the auxiliary information for each individual was a vector of length fourteen having exactly two ones and twelve zeros. However, for computational purposes, the dimensionality of the vector had to be reduced to thirteen when using the Newton-Raphson procedure to solve equation (2.3). For the first four distance functions, we set $c_k = 1$.

For each of the $R = 4{,}000$ samples and each of the eighteen point estimators, we calculated the Jackknife variance estimator given by equation (3.3). We also calculated the Taylor variance estimator given by equation (3.2), and the modification suggested in section 3 was used for distance functions other than the GLS. Note that since PPSWR, rather than PPSWOR, was used at the first stage of sampling, the use of the variance estimator given by equation (3.2) was entirely appropriate for our simulation. Finally, for the GLS distance function only, the formula (3.1) was calculated to observe the impact of omitting g-factors from the variance estimator.

For each of the six distance functions given above, a number of frequentist properties were investigated. These are given below.

(A) The Percent Relative Bias of the Estimated Number of Unemployed (with respect to the population value) is estimated by:

$$\frac{E_M(\hat{Y}_w) - Y}{Y} * 100 \qquad (4.1)$$

where

$$E_M(\hat{Y}_w) = \frac{1}{R} \sum_{r=1}^{R} \hat{Y}_{w_r}$$

is the Monte Carlo expectation of the point estimator $\hat{Y}_w$ taken over the R samples, and $\hat{Y}_{w_r}$ is the value of $\hat{Y}_w$ for sample r.

(B) The Percent Relative Bias of the Taylor/Jackknife Variance Estimator (with respect to the true variance) is estimated by:

$$\frac{(E_M(\hat{V}(\hat{Y}_w)) - V_{\text{true}})}{V_{\text{true}}} * 100 \qquad (4.2)$$

where

$$E_M(\hat{V}(\hat{Y}_w)) = \frac{1}{R} \sum_{r=1}^{R} \hat{V}_r(\hat{Y}_w)$$

and

$$V_{\text{true}} = \frac{1}{R} \sum_{r=1}^{R} (\hat{Y}_{w_r} - E_M(\hat{Y}_w))^2$$

and $\hat{V}_r(\hat{Y}_w)$ is the value of $\hat{V}(\hat{Y}_w)$ (Taylor or Jackknife) for sample r.

(C) The Percent Coefficient of Variation of the Taylor/Jackknife Variance Estimator (with respect to the true variance) is estimated by:

$$\frac{\sqrt{\frac{1}{R}\sum_{r=1}^{R}(\hat{V}_r(\hat{Y}_w) - V_{true})^2}}{V_{true}} * 100 \qquad (4.3)$$

i.e., the root mean squared error of the variance estimator divided by the true variance, expressed as a percentage. Although most studies focus on the *bias* of the variance estimators, it is also of secondary interest to look at the *coefficient of variation* of the variance estimators to see how variable the variance estimates themselves are.

Note that in equations (4.2) and (4.3), it may have been more appropriate to make comparisons relative to a "true mean squared error" rather than a "true variance". However, for our simulation, the relative biases were so small that the differences between the two types of comparisons are virtually negligible.

Finally, in order to assess the appropriateness of the choice of number of repeated samples, we calculated Monte Carlo errors, using as a measure the Percent Coefficient of Variation of $E_M(\hat{V}(\hat{Y}_w))$, given by:

$$\frac{\sqrt{\frac{1}{R^2}\sum_{r=1}^{R}[\hat{V}_r(\hat{Y}_w) - E_M(\hat{V}(\hat{Y}_w))]^2}}{E_M(\hat{V}(\hat{Y}_w))} * 100. \qquad (4.4)$$

The Monte Carlo errors were found to be consistently low (between .99% and 3.60%) for both the Jackknife and Taylor using $R = 4{,}000$, indicating stable results.

4.2 Results of the Study

Table 1 gives the Percent Relative Bias of the Point Estimators (equation (4.1)) as well as the Percent Relative Bias of the Taylor and Jackknife Variance Estimators (equation (4.2)) and the Percent CVs of the Taylor and Jackknife Variance Estimators (equation (4.3)). The percent relative bias for all the point estimates (column two) is negligible, ranging in value from 0.10% to 0.52%, but much less than 1% in all cases. The fact that all point estimates have a similar bias seems reasonable, given the asymptotic equivalence of all calibration estimators to the regression estimator.

The third column gives the percent relative bias of the Taylor variance estimator. Here, the true variance is always underestimated, but never by more than 6.2%. In the case of the regression estimator, it appears to make little difference whether or not the g-factor is included in the variance formula (equation (3.1) versus (3.2)); the bias improves only slightly for the case of the g-factor included (-5.82% versus -6.01%). The Jackknife variance estimator (column four), on the other hand, outperforms the Taylor variance estimator uniformly. The Jackknife almost always underestimates the true variance, but by less than 2% in all cases.

To produce a solution, all distance functions but the GLS required an iterative algorithm. This being the case, some of the 4,000 samples experienced convergence problems, particularly in the case of extreme bounding on the g-factors. Those samples for which the algorithm did not converge were discarded. Thus, they did not contribute to the various Monte Carlo measures. The number of such discarded samples is

Table 1
Percent Relative Bias of the Point Estimators, and Percent Relative Bias and Percent CV of the Taylor and Jackknife Variance Estimators (Sample Size About 1000)

Distance Function		Percent Relative Bias Point Estimator	Percent Relative Bias Taylor Variance	Percent Relative Bias Jackknife Variance	Percent CV Taylor Variance	Percent CV Jackknife Variance	Number of Discarded Samples (From 4000)
GLS (Regression)		.11	-6.01 (eq 3.1) -5.82 (eq 3.2)	-1.73	60.79 (eq 3.1) 59.60 (eq 3.2)	62.86	0
Restricted GLS	($L = 0, U = 4$)	.11	-5.82	-1.73	59.60	62.86	0
	($L = .4, U = 2$)	.10	-5.36	-1.27	59.93	63.21	32
Raking Ratio		.52	-6.20	0.84	59.45	63.35	0
Restricted RR	($L = 0, U = 4$)	.50	-6.09	-0.31	59.48	63.47	0
	($L = .4, U = 2$)	.46	-5.69	-0.39	59.81	64.21	32
Modified Huang-Fuller	($L = 0, U = 4$)	.11	-5.82	-1.73	59.60	62.86	0
	($L = .4, U = 2$)	.10	-5.36	-1.20	59.94	63.27	32
Shrinkage-Minimization	($L = 0, U = 4$)	.11	-5.82	-1.73	59.60	62.86	0
	($L = .4, U = 2$)	.10	-5.36	-1.27	59.94	63.25	32

indicated in the last column of Table 1. In the case of extreme bounds ($L = .68$, $U = 1.6$ and $L = .8$, $U = 1.3$), so many samples were discarded (between 231 and 234 for the cases $L = .68$, $U = 1.6$ and between 1,562 and 1,602 for the cases $L = .8$, $U = 1.3$) that the results were not considered reliable, and so are not reported here. However, these tighter bounds were of interest, so the simulation was rerun using approximately double the sample size (increase from roughly 1,000 to 2,000). Note that Deville and Särndal (1992) show that convergence is achieved for all distance functions with probability one as the sample size increases.

Columns five and six of Table 1 give the Percent CVs of the Taylor and Jackknife Variance Estimators. The coefficients of variation are similar for all distance functions, ranging in value from 59.45% to 64.21%. However, the CVs corresponding to the Jackknife are always slightly larger than that of Taylor. Coefficients of variation of this magnitude, although large, have been encountered in other simulation studies relating to variances. See, for example, Kovačević, Yung and Pandher (1995). However, we were interested in seeing if the key results relating to the bias of the variance estimators would still hold if the CVs were lowered. Therefore, at the suggestion of a referee, we reran the simulation, increasing the number of PSUs drawn from 47 to 83, since CVs of variance estimators are known to be approximately inversely related to the number of PSUs drawn. The PSUs were increased in such a way that the overall design was made self-weighting; this approach appeared to have the greatest effect on lowering the CVs. The second stage of sampling remained the same as before. Rerunning the simulation had the secondary benefit of roughly doubling the sample size, and thus, solving the convergence problems referred to in the last paragraph.

The results from the second run of the simulation are reported in Table 2. The last column in Table 2 shows the reduced number of discarded samples due to convergence problems. The fifth and sixth column of this table show that the CVs are significantly reduced to between 22.70% and 24.2% with the Jackknife consistently exhibiting slightly higher values. Now, as before, the percent relative bias in the point estimator is negligible, always being well under 1%. In the previous run, the percent relative biases for the Taylor estimator were always roughly -6%; here, they are always about -3%, again implying underestimation of the true variance. Once more, in the case of the GLS distance function, there is very little difference in the bias that results from using equation (3.1) versus (3.2). The percent relative bias in the Jackknife estimator (always roughly -1.5%) is consistently

Table 2
Percent Relative Bias of the Point Estimators, and Percent Relative Bias and Percent CV of the Taylor and Jackknife Variance Estimators (Sample Size About 2000)

Distance Function		Percent Relative Bias Point Estimator	Percent Relative Bias Taylor Variance	Percent Relative Bias Jackknife Variance	Percent CV Taylor Variance	Percent CV Jackknife Variance	Number of Discarded Samples (From 4000)
GLS (Regression)		.02	-2.71 (eq 3.1) -2.61 (eq 3.2)	-1.43	23.03 (eq 3.1) 22.84 (eq 3.2)	23.29	0
Restricted GLS	($L = 0$, $U = 4$)	.02	-2.61	-1.43	22.84	23.29	0
	($L = .4$, $U = 2$)	.02	-2.61	-1.43	22.84	23.29	0
	($L = .68$, $U = 1.6$)	.02	-2.61	-1.44	22.84	23.29	0
	($L = .8$, $U = 1.3$)	.02	-2.75	-1.56	22.70	23.15	118
Raking Ratio		.25	-2.75	-1.15	22.84	23.43	0
Restricted RR	($L = 0$, $U = 4$)	.17	-2.67	-1.36	22.84	23.30	0
	($L = .4$, $U = 2$)	.16	-2.70	-1.42	22.84	23.29	0
	($L = .68$, $U = 1.6$)	.31	-2.77	-0.49	22.83	24.20	0
	($L = .8$, $U = 1.3$)	.27	-2.91	*	22.70	*	118
Modified Huang-Fuller	($L = 0$, $U = 4$)	.02	-2.61	-1.43	22.84	23.29	0
	($L = .4$, $U = 2$)	.02	-2.61	-1.43	22.84	23.29	0
	($L = .68$, $U = 1.6$)	.02	-2.61	-1.44	22.84	23.29	0
	($L = .8$, U = 1.3)	.02	-2.58	-1.36	22.73	23.18	116
Shrinkage-Minimization	($L = 0$, $U = 4$)	.02	-2.61	-1.43	22.84	23.29	0
	($L = .4$, $U = 2$)	.02	-2.61	-1.43	22.84	23.29	0
	($L = .68$, $U = 1.6$)	.02	-2.61	-1.44	22.84	23.29	0
	($L = .8$, $U = 1.3$)	.02	-2.61	-1.24	22.73	23.63	118

smaller in absolute value than that of Taylor. For the Jackknife estimator, there is one case (Restricted RR (L = .8, U = 1.3)) where there were convergence problems; those results are omitted, indicated by a "*". Surprisingly, for both the Taylor and Jackknife, there is virtually no change in bias for the restricted distance functions as the bounds are made successively more tight. In fact, there seems to be very little difference in the percent relative bias across all of the distance functions, for both the Taylor and the Jackknife. Note that for the rerun of the simulation, the Monte Carlo errors ranged between .37% and 2.13%.

5. CONCLUSIONS

This paper focused on exploring the behaviour of point estimators and their corresponding Taylor and Jackknife variance estimators for a number of different distance functions available through calibration theory. Particular emphasis was given to those distance functions which allowed range restrictions to be imposed on the g-factors, eliminating the possibility of negative and high positive final weights. All of the point estimators which were investigated exhibited a negligible bias.

Both the Jackknife and Taylor variance estimators exhibited small underestimation of the true variance, although the Jackknife consistently had smaller biases (in absolute value) than the Taylor. The most striking result was that, for both Taylor and Jackknife, the biases remained roughly the same in the cases of extreme bounding on the g-factors as in the cases of less restrictive bounding. In general, however, caution should be exercised in the use of extreme bounds, due to the convergence problems that may be experienced, particularly when Jackknifing is used for variance estimation and the point estimators must be recalculated repeatedly. If the main objective of using the restricted distance functions is to eliminate the possibility of negative or high positive weights, then modest bounds on the g-factors should suffice.

As a final remark, it is interesting to note that roughly 97% of the computing time was spent Jackknifing while the remaining 3% was spent on Taylor linearization. This rather extreme difference in computation time may give the Taylor method an advantageous edge if measures of precision are required for a large number of domains. However, given recent developments in the computational efficiency of the Jackknife variance estimator (for example, the program WESVARPC (1995)), it may be possible to offset this imbalance. Even so, it should be noted that, at this time, WESVARPC has improved the computational efficiency for designs having only two PSUs per stratum, and poststratified estimators having only one dimension.

In conclusion, since our study does not conclusively show either variance estimator to be clearly superior and shows both to behave reasonably well for all distance functions, it is up to the user to decide which variance/ distance function combination best fits the system requirements.

ACKNOWLEDGEMENTS

The authors would like to thank Chris Mohl for providing us with some of the computer code which we used in the simulation study. We would also like to thank an Associate Editor and two referees for useful comments on the earlier version of this paper.

REFERENCES

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.

DEVILLE, J.-C., and SÄRNDAL, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.

DEVILLE, J.-C., SÄRNDAL, C.-E., and SAUTORY, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, 88, 1013-1020.

HIDIROGLOU, M.A., FULLER, W.A., and HICKMAN, R.D. (1980). SUPERCARP, Department of Statistics, Iowa State University, Ames, Iowa.

HUANG, E.T., and FULLER, W.A. (1978). Nonnegative regression estimation for sample survey data. *Proceedings of the Social Statistics Section, American Statistical Association*, 300-305.

KOVAČEVIĆ, M.S., YUNG, W., and PANDHER, G.S. (1995). Estimating the sampling variances of measures of income inequality and polarization – An empirical study. Methodology Branch Working Paper, HSMD-95-007E, Statistics Canada.

SÄRNDAL, C.-E. (1982). Implications of survey design for generalized regression estimation of linear functions. *Journal of Statistical Planning and Inference*, 7, 155-170.

SÄRNDAL, C.-E., SWENSSON, B., and WRETMAN, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

SINGH, A.C., and MOHL, C.A. (1996). Understanding calibration estimators in survey sampling. *Survey Methodology*, 22, 107-115.

SINGH, M.P., DREW, J.D., GAMBINO, J.G., and MAYDA, F. (1990). *Methodology of the Canadian Labour Force Survey: 1984-1990*. Catalogue No. 71-526, Statistics Canada.

STUKEL, D.M., and BOYER, R. (1992). Calibration estimation: An application to the Canadian Labour Force Survey. Methodology Branch Working Paper, SSMD-92-009E, Statistics Canada.

WESVARPC (1995). Westat Inc., Rockville, Maryland.

WOLTER, K.M. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag.

YUNG, W., and RAO, J.N.K. (1996). Jackknife linearization variance estimators under stratified multi-stage sampling. *Survey Methodology*, 22, 23-31.

Survey Methodology, December 1996
Vol. 22, No. 2, pp. 127-137
Statistics Canada

An Application of Restricted Regression Estimation in a Household Survey

BODHINI R. JAYASURIYA and RICHARD VALLIANT[1]

ABSTRACT

This paper empirically compares three estimation methods – regression, restricted regression, and principal person – used in a household survey of consumer expenditures. The three methods are applied to post-stratification which is important in many household surveys to adjust for under-coverage of the target population. Post-stratum population counts are typically available from an external census for numbers of persons but not for numbers of households. If household estimates are needed, a single weight must be assigned to each household while using the person counts for post-stratification. This is easily accomplished with regression estimators of totals or means by using person counts in each household's auxiliary data. Restricted regression estimation refines the weights by controlling extremes and can produce estimators with lower variance than Horvitz-Thompson estimators while still adhering to the population controls. The regression methods also allow controls to be used for both person-level and household-level counts and quantitative auxiliaries. With the principal person method, persons are classified into post-strata and person weights are ratio adjusted to achieve population control totals. This leads to each person in a household potentially having a different weight. The weight associated with the "principal person" is then selected as the household weight. We will compare estimated means from the three methods and their estimated standard errors for a number of expenditures from the Consumer Expenditure survey sponsored by the U.S. Bureau of Labor Statistics.

KEY WORDS: Calibration; Principal person method; Replication variance; Restricted regression.

1. INTRODUCTION

A signal problem in large household surveys is under-coverage of the target population often arising from differential response rates among population subgroups and frame deficiencies. Post-stratification is one method used at the estimation stage to reduce mean square errors based on information that affect the response variables. The estimator is constructed in such a way that the estimated total number of individuals falling into each post-stratum is equal to the true population count. Post-stratum population counts are typically available from an external census for numbers of persons but not always for numbers of households. If household estimates are needed, a single weight must be assigned to each household while using the person counts for post-stratification. Regression estimators of totals or means accomplish this by using person counts in each household's auxiliary data. Restricted regression estimation controls extreme weights and can produce estimators with lower variance than the Horvitz-Thompson estimator while still adhering to the population controls. An alternative used by some surveys is the Principal Person (PP) method (Alexander 1987) in which the household weight is based on the individual designated as the"principal person" in each household. Persons are classified into post-strata and person weights are ratio adjusted to achieve population control totals, leading to the possibility that each person in a household may have a different weight. The weight associated with the principal person is then assigned to the household. This *ad hoc* method is difficult to analyze theoretically. The regression estimators discussed in this paper, while easily adjusting for the population under-count, automatically provide a household weight that is not based on any particular one of its members. Lemaître and Dufour (1987) address Statistics Canada's use of the regression estimator in this regard.

There are a growing number of precedents for the use of regression estimators in surveys both in the theoretical literature and in actual survey practice. Statistics Canada has incorporated the general regression estimator into its generalized estimation system (GES) software that is now used in many of its surveys (Estevao, Hidiroglou and Särndal 1995). Fuller, Loughin and Baker (1993) discuss an application to the USDA Nationwide Food Consumption Survey. One of the attractions of regression estimation is that many of the standard techniques in surveys including the post-stratification estimator mentioned above are special cases of regression estimators. The regression estimator also more flexibly incorporates auxiliary data than other more common methods. In a household survey, for example, both person-level and household-level auxiliaries that can be qualitative or quantitative are easily accommodated. Other works related to regression estimation and post-stratification include Bethlehem and Keller (1987), Casady and Valliant (1993), Deville and Särndal (1992), Deville, Särndal and Sautory (1993) and Zieschang (1990).

In this study we compare the regression estimator with the PP estimator currently in use at the Bureau of Labor Statistics (BLS). Each estimator can be written in the form of a weighted sum of the sample values of the response variable. Then each weight is traditionally interpreted as the number of

[1] Bodhini R. Jayasuriya and Richard Valliant, U.S. Bureau of Labor Statistics, 2 Massachusetts Avenue, N.E., Room 4915, Washington, DC 20212, U.S.A.

individuals in the population who would have the corresponding value of the response variable. This interpretation requires that each weight be greater than or equal to one. The ordinary least-squares regression estimator has the disadvantage that it can produce non-positive weights. A number of ways are suggested in the literature on how to overcome this problem. Possibly the easiest is the method introduced by Deville and Särndal (1992) which can remove any negative weights as well as control extreme weights. The restricted regression estimators produced by these new weights are also compared to the original regression estimator and the PP estimator.

In Section 2, the three different estimators are presented. Section 3 is an application of these procedures to the Consumer Expenditure (CE) Survey at BLS – the same setting as in Zieschang (1990). We compare the coefficients of variation for a number of the survey target variables for the full population and for a number of domains. Section 4 provides a summary of our conclusions.

2. REGRESSION, CALIBRATION AND PRINCIPAL PERSON ESTIMATION

First, we give a brief introduction to the regression estimator. A sample s of size n is selected from a finite population U of size N. Let the probability of selection of the i-th unit be π_i. The sample could be two-stage and the unit could be either the primary sampling unit or the secondary sampling unit. There is no need here to complicate the notation with explicit subscripts for the different stages of sampling. Let the variable of interest be denoted by y and suppose that its value at the i-th unit, y_i, is observed for each $i \in s$. Assume the existence of K auxiliary variables $x_1, x_2, ..., x_K$ whose values at each $i \in s$ are available. Define $\boldsymbol{x}_i = (x_{i1}, x_{i2}, ..., x_{iK})'$, for each $i \in U$, where x_{ik} denotes the value of the variable x_k at unit i. Let $\boldsymbol{X} = (X_1, ..., X_K)'$ denote the K-dimensional vector of known population totals of the variables $x_1, x_2, ..., x_K$. The regression estimator is then motivated by the working model ξ:

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_K x_{iK} + \varepsilon_i \qquad (2.1)$$

for $i = 1, ..., N$. Here, $\beta_1, ..., \beta_K$ are unknown model parameters. The ε_i are random errors with $E_\xi(\varepsilon_i) = 0$ and $\mathrm{var}_\xi(\varepsilon_i) = \sigma_i^2$ for $i = 1, ..., N$. The term "working model" is used to emphasize the fact that the model is likely to be wrong to some degree. In the CE, the unit of analysis, indexed by i, is a consumer unit (CU), which is similar to a household and defined in more detail in Section 3. The value y_i might be the total food expenditures by the CU and the x_{ik}'s might be various CU characteristics like numbers of people of different ages, or CU income, that have an effect on the CU's expenditure on food. The variance of expenditures might be dependent on CU size so that having σ_i^2 proportional to the number of persons in the CU might be reasonable. We include an intercept in some of our models by setting the first auxiliary variable, x_1, equal to 1.

A linear regression estimator of the population total of y is defined to be

$$\hat{y}_R = \hat{y}_\pi + (\boldsymbol{X} - \hat{\boldsymbol{x}}_\pi)' \hat{\boldsymbol{\beta}} \qquad (2.2)$$

where $\hat{y}_\pi$ denotes the π-estimator (or Horvitz-Thompson estimator) of the population total of y, *i.e.*,

$$\hat{y}_\pi = \sum_{i \in s} a_i y_i \qquad (2.3)$$

with $a_i = 1/\pi_i$. Also, $\hat{\boldsymbol{x}}_\pi = (\hat{x}_{1\pi}, ..., \hat{x}_{K\pi})'$ is the vector of π-estimators of the population totals of the variables $x_1, x_2, ..., x_K$ and

$$\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, ..., \hat{\beta}_K)' = \left[\sum_{i \in s} \frac{a_i \boldsymbol{x}_i \boldsymbol{x}_i'}{\sigma_i^2} \right]^{-1} \sum_{i \in s} \frac{a_i \boldsymbol{x}_i y_i}{\sigma_i^2}. \qquad (2.4)$$

We assume that $\sum_{i \in s} a_i \boldsymbol{x}_i \boldsymbol{x}_i' / \sigma_i^2$ is nonsingular. Even if model (2.1) fails to some degree, $\hat{y}_R / N$ is a design consistent estimator of the population mean $\overline{Y}$ irrespective of whether the assumed model is true or false. This is clear from (2.2). If $\hat{y}_\pi / N$ and $\hat{\boldsymbol{x}}_\pi / N$ are design consistent estimators of $\overline{Y}$ and of $\overline{\boldsymbol{X}}$, the vector of population means of the auxiliaries, then the second term in $\hat{y}_R / N$ converges to zero while the first converges to $\overline{Y}$. For more details, see Särndal, Swensson and Wretman (1992).

The regression estimator $\hat{y}_R$ can also be expressed as a weighted sum of the sample y_i's, which is a desirable feature for survey operations. It is easily seen that (2.2) can be re-written as $\hat{y}_R = \sum_{i \in s} w_i y_i$ with

$$w_i = a_i \left[1 + (\boldsymbol{X} - \hat{\boldsymbol{x}}_\pi)' \boldsymbol{A}^{-1} \frac{\boldsymbol{x}_i}{\sigma_i^2} \right] \qquad (2.5)$$

where $\boldsymbol{A} = \sum_{i \in s} a_i \boldsymbol{x}_i \boldsymbol{x}_i' / \sigma_i^2$. The weights do depend on the sample through the x_i's that are in the sample, but this is also true of many survey estimators, including the post-stratification estimator. However, these weights do not depend on the particular y variable being studied, implying that one set of w_i weights can be used for all estimates.

A mean per unit is estimated in the obvious way: $\hat{\bar{y}}_R = \hat{y}_R / \hat{N}$ where $\hat{N} = \sum_{i \in s} w_i$. If we estimate the totals of the auxiliaries $\boldsymbol{x}_i$, then

$$\sum_{i \in s} w_i \boldsymbol{x}_i' = \sum_{i \in s} \left[a_i \boldsymbol{x}_i' + (\boldsymbol{X} - \hat{\boldsymbol{x}}_\pi)' \boldsymbol{A}^{-1} \frac{a_i \boldsymbol{x}_i \boldsymbol{x}_i'}{\sigma_i^2} \right] = \boldsymbol{X}', \qquad (2.6)$$

i.e., we reproduce the known population totals. This is also a characteristic of the post-stratification estimator.

The estimator of β in (2.4) does not account for any correlation among the errors in model (2.1). In clustered

populations, units that are geographically near each other, *e.g.*, CU's in the same neighborhood, may be correlated. Using a full covariance matrix V may be more nearly optimal (*e.g.*, see Casady and Valliant 1993 and Rao 1994). Though use of a full covariance matrix V may lower the variance of $\hat{\beta}$, the elements of V will depend on the particular y being studied, and estimation of V is generally a nuisance. Consequently, it is interesting and practical to consider the simple case of $V = \text{diag}(\sigma_i^2)$ that leads to (2.2). Note that when the design-variance $\text{var}_p(\hat{y}_R)$ is estimated, it will be necessary to use a method that properly reflects clustering and other design complexities.

The regression estimator has the disadvantage that the weights can be unreasonably large, small or, even negative. The restricted calibration estimators of Deville and Särndal (1992), introduced next, add constraints to control the size of the weights. Calibration estimators are formed by minimizing a given distance, F, between some initial weight and the final weight, subject to constraints. The constraints can involve the available auxiliary variables thus incorporating them into the estimator. The regression estimator presented above is a special case of the calibration estimator in which F is defined to be the generalized least squares (GLS) distance function,

$$F(w_i, a_i) = \frac{a_i c_i}{2}\left(\frac{w_i}{a_i} - 1\right)^2$$

for $i = 1, ..., n$, with c_i a known, positive weight (*e.g.*, $c_i = \sigma_i^2$ or $c_i = 1$) associated with unit i, and w_i, the final weight. The total sample distance $\sum_{i \in s} F(w_i, a_i)$ is minimized subject to the constraints,

$$\sum_{i \in s} w_i \boldsymbol{x}_i = \boldsymbol{X}. \tag{2.7}$$

In this form, the weights of the regression estimator of the population total of y given in (2.5) can be written as,

$$w_i = a_i g(c_i^{-1} \lambda' \boldsymbol{x}_i) \tag{2.8}$$

for $i = 1, ..., n$ where

$$g(u) = 1 + u, \tag{2.9}$$

for $u \in \Re$ and λ is a Lagrange multiplier evaluated in the minimization process. The particular form of w_i with $c_i = \sigma_i^2$ for the regression estimator was given in (2.5). To eliminate extremes, the weights can be refined by restricting g so that

$$g(u) = \begin{cases} L & \text{if } u < L - 1 \\ 1 + u & \text{if } L - 1 \le u \le U - 1 \\ U & \text{if } u > U - 1. \end{cases} \tag{2.10}$$

With this definition of g, the weights w_i satisfy

$$L < w_i / a_i < U \tag{2.11}$$

for $i = 1, ..., n$ so that L and U can be chosen in such a way as to reflect the desired deviation from the initial weights a_i. Choosing $L > 0$ ensures that the weights are positive, and U is picked to be appropriately small to prohibit large weights. The restricted regression weights must be solved for iteratively; one easily programmed algorithm is given in Stukel and Boyer (1992). Another method of restricting weights is ridge regression as used by Bardsley and Chambers (1984).

In most household surveys, post-stratification serves primarily as an adjustment for under-coverage of the target population by the frame and the sample. In the U.S., there are few reliable population counts of households to use in post-stratification. Consequently, population counts of persons are usually used for the post-strata control totals. This disagreement in the unit of analysis (the household) and the unit of post-stratification (the person) when a household characteristic is of interest led to the development of the PP method that is used in the CE and Current Population Surveys.

In the PP method described in Alexander (1987), a household begins the weighting process with a single base weight, a_i, that is then adjusted for non-response. The adjusted weight is assigned to each person in the household and the person weights are then further adjusted to force them to sum to known population controls of persons by age, race, and sex. This last adjustment can result in persons having different weights within the same household. The household is then assigned the weight of the person designated as the "principal person" in the household. This method has an element of arbitrariness and is difficult to analyze mathematically. The intent of this research was not to see if the PP method could be improved upon, but rather to use the current implementation of PP as a convenient baseline for measuring the performance of other estimators.

The regression and restricted regression estimators can be formulated in such a way that population person controls are satisfied, all persons in a household retain the same weight, and no arbitrary choice among person weights is needed to assign a household weight. This is accomplished by defining the auxiliary variables at the household level. For example, if there were three age post-strata and household i has 1, 0, and 2 persons in these post-strata, the auxiliary data vector would be $\boldsymbol{x}_i = (1,0,2)'$. Note that this formulation is different from Lemaître and Dufour (1987) who defined the auxiliary variables at the person level and assigned the average of the household data – (1/3, 0, 2/3) in the example – to each person. Those authors used this "average" method because they were interested in estimates both for persons, *e.g.*, number employed, and for households, *e.g.*, economic families. We, on the other hand, need only a household weight since our target variables (*i.e.*, y) like shelter or utility expenditures are collected at the household level.

3. AN APPLICATION

We compare the three estimators (*i.e.*, regression, restricted regression (with $L = .5$, $U = 4$), and principal person) by an

application to the estimated means and their estimated standard errors for a number of expenditures from the CE Survey sponsored by the Bureau of Labor Statistics.

The CE Survey gathers information on the spending patterns and living costs of the American consumers. There are two parts to the survey, a quarterly interview and a weekly diary survey. The Interview Survey collects detailed data on the types of expenditures which respondents can be expected to recall for a period of three months or longer (*e.g.*, property, automobiles, major appliances) – an estimated sixty to seventy percent of total household expenditures. The Diary Survey is completed at home by the respondent family for two consecutive 1-week periods and collects data on all the expenses of the family in that time period. The sample is selected in two stages with geographic primary sampling units at the first stage and households at the second.

We evaluated the estimators described above for a number of expenditures from the Interview Survey. Data collected during the second quarter of 1992 consisting of $n = 5156$ CU's were used. The CE Survey's primary unit of analysis is the consumer unit, an economic family within a household. A consumer unit (CU) consists of individuals in the household who share expenditures. Thus, there may be more than one CU in a household.

Five different sets of auxiliary variables ($\boldsymbol{x}_i$'s in the notation of Section 2) were studied. They were chosen by testing the adequacy of model (2.1) for the selected expenditures with different combinations of the available auxiliary variables. Combinations of auxiliaries were identified in which each estimated regression coefficient was significant in an ordinary least squares regression at the 5% level. A key step that substantially improved the fit of the models was simply including an intercept. Factored into the selection of auxiliaries was also the knowledge that the survey has more under-coverage of Blacks than non-Blacks and that this needed to be accounted for by post-stratification. We viewed this method of variable selection as exploratory and, consequently, a number of combinations were studied to determine which set produced the best estimators of mean expenditures. The 56 post-strata based on age/race/sex currently in use in the CE were included. (The 56 are routinely collapsed in actual CE operations because of small sample sizes in some cells.) Other variables that were statistically significant in various combinations were region (NE, MW, S, W), urbanicity (urban/rural) by region, age of reference person of the CU (< 25, 25-34, 35-44, 45-64, 65+), household tenure (owner/renter), income before taxes of the CU, and the 56 post-strata collapsed by sex and some of the age categories to form 10 age/race categories. Based on this information, weights (2.8) were computed using g given in (2.9) – regwts – and (2.10) – calwts. For both the regression and restricted regression weights, we set a_i equal to the adjusted base weight, *i.e.*, $1/\pi_i$ times a non-response adjustment. In order for the matrix $\boldsymbol{A}$ in Section 2 to be nonsingular, one of the categories in some auxiliaries, like region, was omitted from each $\boldsymbol{x}_i$. For this application, the population totals necessary to evaluate $\boldsymbol{X} = (X_1,...,X_K)'$ were obtained mostly from the Statistical Abstract of the United States (1993) whose sources are the 1990 Census figures and the Current Population Reports published by the U.S. Bureau of the Census. When an intercept is used, the appropriate control total for that variable is the number of CU's in the population for which we used the PP estimate as a surrogate. The combinations of auxiliaries used to form the different weights are given in Table 1. Regwts0, with 56 age/race/sex post-strata uses the largest number of post-strata. The 56 are the starting point for the PP method but are usually collapsed to 30-40 because of small cell sizes. When computing calwts0, those 56 post-strata were collapsed to 45 since the constraints imposed by the L and U bounds could cause singularity in the matrix based algorithm.

Table 1
Weights and Their Corresponding Auxiliary Variables

Weights	Auxiliary Variables	K
regwts0	Age/race/sex	56
regwts1	Intercept, age/race/sex, region, urban × region	18
regwts2	Intercept, age/race/sex, region, urban × region, age of reference person, housing tenure, family income before taxes	24
calwts0	Age/race/sex	45
calwts1	Intercept, age/race/sex, region, urban × region	18
calwts2	Intercept, age/race/sex, region, urban × region, age of reference person, housing tenure, family income before taxes	24
calwts3	Intercept, age/race/sex, region, urban × region, family income before taxes (truncated at $500,000)	19
calwts4	Intercept, age/race/sex, region, urban × region, age of reference person, housing tenure	23
PP	Age/race/sex	56[1]

[1] The initial set of 56 is usually collapsed to 30-40 because of small sample sizes in some cells.

3.1 Comparisons of Weights

A variety of comparisons of weights produced by the different methods were made, only a few of which can be mentioned here. Figure 1 shows plots of the PP weights, regwts0, calwts0, and calwts1 versus the adjusted base weights. For PP and regwts0, the adjustments to go from a_i to w_i are much more variable than for calwts0 and calwts1, which employ the $L = 0.5$ and $U = 4$ restrictions. High variability among the w_i can lead to expenditure estimates with high variance and to poor confidence interval coverage since large sample normality may not hold. Even though (2.11) implies that $a_i/2 < w_i < 4a_i$ for each i for the calwts, the lower right panel in Figure 1 shows that the calwts1 satisfy $a_i/2 < w_i \leq 2a_i$, for each i. Thus, setting $U = 2$ or 3 would have little effect on calwts1. Calwts0 would have been slightly affected by setting $U = 2$ since a few points were outside the upper reference line. The upper two panels indicate that the PP weights and regwts0 do not conform to the restriction $a_i/2 < w_i < 2a_i$.

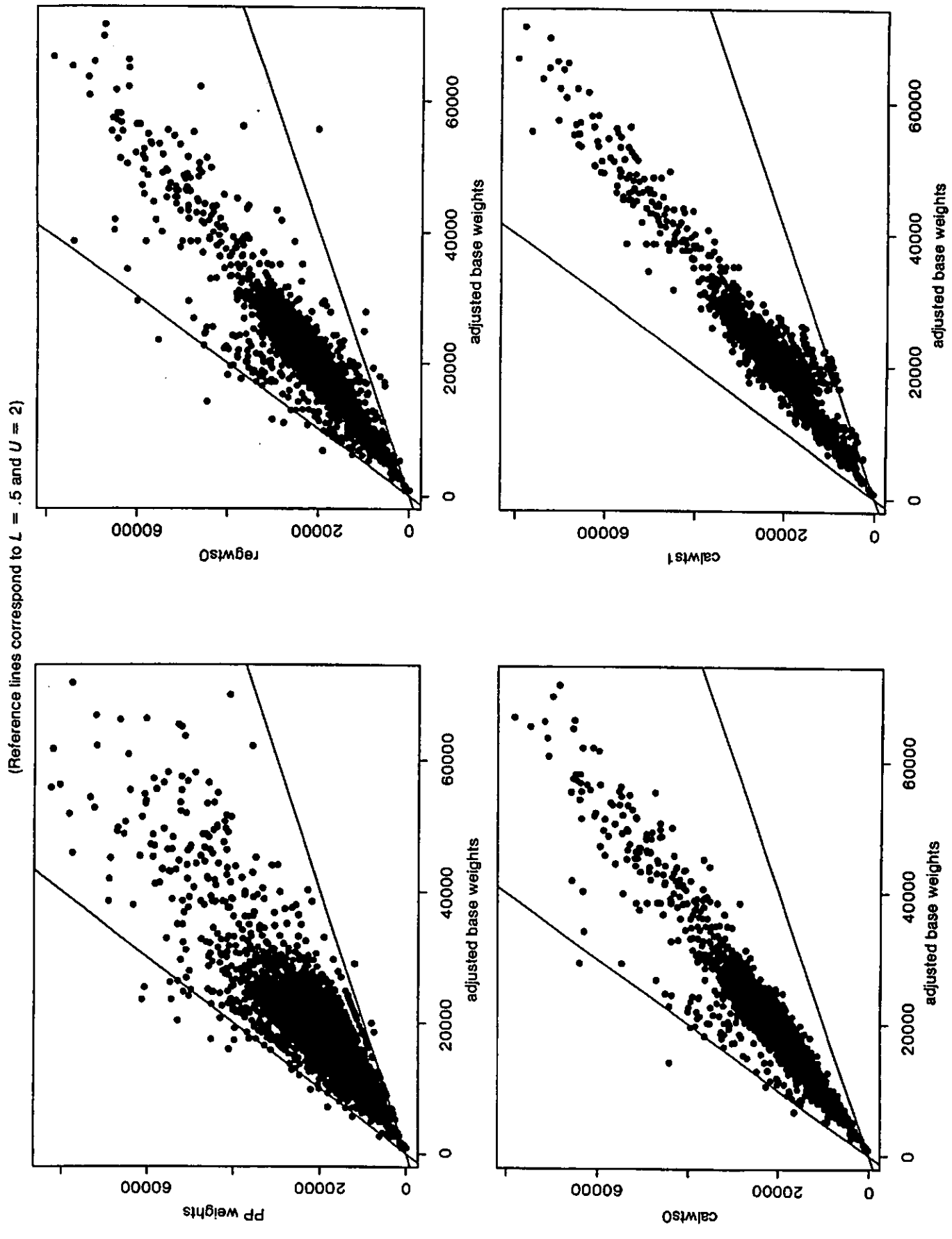


Figure 1. Four sets of weights plotted versus adjusted base weights

The concern about negative regression weights was minor in the application. In the full sample, only one CU had a negative weight for regwts1 and regwts2 while regwts0 had no negative weights. However, in the replicates used for variance estimation, described in Section 3.2, 2 or 3 CU's did have negative weights in many replicates so that using the L restriction was more important there.

3.2 Precision of Estimates from the Different Methods

Although comparison of weights is instructive, the methods must ultimately be judged based on the level of estimated CU means and their precision. The standard errors of these estimators were computed via the method of balanced half sampling (BHS) using 44 replicates as currently implemented in the CE for the PP estimator. The BHS estimator is constructed to reflect the stratification and the clustering that is used in the CE. A half sample is constructed in a prescribed way (McCarthy 1969) to contain one half of the first-stage sample units in a survey. Defining the mean per CU based on CU's in half-sample α to be $\hat{\bar{y}}_{R(\alpha)}$ and that for the full sample to be $\hat{\bar{y}}_R$, the BHS estimate of variance is $V_{BHS}(\hat{\bar{y}}_{R(\alpha)}) = \sum_{\alpha=1}^{44}(\hat{\bar{y}}_{R(\alpha)} - \hat{\bar{y}}_R)^2/44$. To compute each $\hat{\bar{y}}_{R(\alpha)}$, the same estimation steps used for the full sample are repeated for the CU's in the half-sample. As the expenditure estimates from the CE Survey are published for various inter domains of interest, we computed the means and the standard errors for a few chosen domains as well. For each of these, the coefficient of variation (cv) was computed and then its ratio to the cv of the PP weight estimate was calculated.

For each type of weight, if the ratio of each expenditure cv to that of the PP weights is less than one, an improvement over the PP estimate is indicated since, for all the weights, the expenditure mean estimates were very close to those of the PP estimates. We computed the ratios of cv's and the ratios of means for each of the sets of weights described in Table 1, for each of the chosen expenditures, and for each of the following domains:

(1) Age of Reference Person: < 25, 25-34, 35-44, 45-54, 55-64, 65+
(2) Region: NE, MW, S, W
(3) Size of CU: 1, 2, 3, 4, 5+
(4) Composition of Household: Husband and wife only, Husband and wife + children, Other Husband and wife, One parent + at least one child < 18, Single person and other CU's
(5) Household Tenure: Owner, Renter
(6) Race of Reference Person: Black, Non-Black.

We will discuss only domains (1) – (3) here. In addition, ratios for all CU's, *i.e.*, the total across the domains, were computed for each expenditure and are shown in Table 2. For All Expenditures, regwts2, calwts2, and calwts3, with ratios of .79, .78, and .75, provide substantial reduction in cv compared to PP. For less aggregated expenditures regwts1 or calwts1 provide reasonably consistent improvements over PP without the losses incurred by some of the other weights for expenditures like Furniture, Personal insurance and pensions, and its sub-category Pensions and social security.

Table 2
Ratios to PP cv of cv's for the Different Weighting Methods
The Minimum Ratio is Highlighted in Each Row

Expenditure	regwts			calwts				
	0	1	2	0	1	2	3	4
All expenditures	0.98	0.90	0.79	0.98	0.90	0.78	0.75	0.87
Shelter	0.93	0.85	0.75	0.93	0.85	0.74	0.72	0.84
Utilities	1.08	1.03	0.94	1.07	1.03	0.88	0.91	0.92
Furniture	1.08	1.21	3.52	1.06	1.21	2.58	2.57	1.17
Major appliances	1.08	1.06	1.04	1.06	1.08	1.09	1.00	1.03
All vehicles	0.90	0.89	0.98	0.91	0.89	0.98	0.97	0.90
New cars, trucks	0.95	0.91	1.01	0.96	0.91	1.02	1.02	0.91
Used cars, trucks	0.98	0.94	0.96	0.97	0.94	0.97	0.96	0.95
Gasoline, motor oil	1.17	1.11	1.03	1.12	1.10	0.99	0.94	1.10
Health care	1.05	0.97	0.86	1.07	0.97	0.85	0.87	0.94
Education	0.92	0.93	1.04	0.91	0.93	1.06	1.07	0.88
Cash contributions	1.01	1.02	1.28	1.01	1.02	1.30	1.29	1.03
Personal insurance, pensions	1.00	0.97	1.64	1.01	0.98	1.24	0.98	0.95
Life, other personal insurance	1.08	1.02	1.53	1.08	0.98	1.38	1.33	1.01
Pensions, social security	1.00	0.99	1.75	1.01	0.99	1.34	1.06	0.97

Trellis plots (Cleveland 1993) of the cv and mean ratios for calwts0 and calwts1 are given in Figures 2-4. Calwts0 is pictured because it is the nearest calibration equivalent to the current method of post-stratification. Calwts1 appears to be the best of the alternatives we have examined in the sense of improving the All Expenditures estimates while providing consistent performance for individual expenditure groups. In each panel of the plots a vertical reference line is drawn at 1, the point of equality between the calibration results and those for the PP method. The lower row in each plot presents ratios of means from calwts0 and calwts1 to the PP means and illustrates that with a few exceptions the levels of the means from the two restricted regression choices are about the same as from PP.

The two calibration choices, in the main, improve cv's compared to PP, *i.e.*, cv ratios tend to be less than 1, for most domains and expenditures, and calwts1 is somewhat better than calwts0. For the age-of-reference-person domains < 25 and 65+, for example, 12 of the 15 expenditures have calwts1 ratios of less than 1. For CU sizes 1-4 the numbers of cv ratios less than or equal to 1 are 12, 9, 9, and 11. There are exceptions, of course. For the South region only 6 of 15 expenditures have calwts1 cv ratios less than or equal to 1.

Calwts2 and calwts3, which used family income before taxes as one of the auxiliaries, had somewhat erratic performance for domains, sometimes making major improvements over PP but occasionally showing serious losses. This is connected to the nature of the family income variable itself. For the entire data set of 5156 CU's, income before taxes was positive for 4698 CU's, zero for 450 CU's and negative for 8 CU's. The zeroes are incomplete income reporters while the negatives are for families that had business losses added to other income. In either case, these CU's vitiate the usefulness

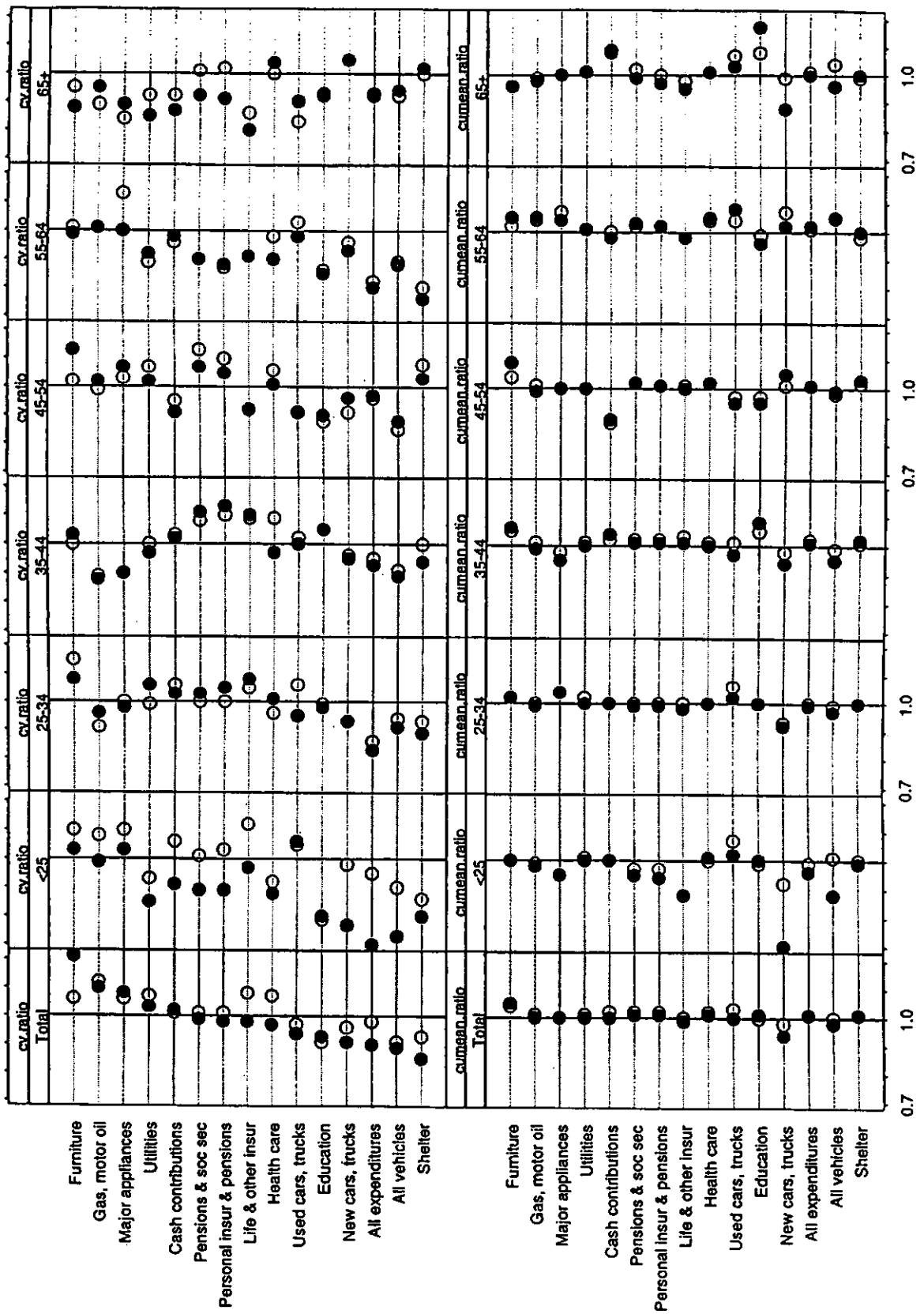


Figure 2. Ratios to PP of cv's and means for two weighting methods by age of reference person

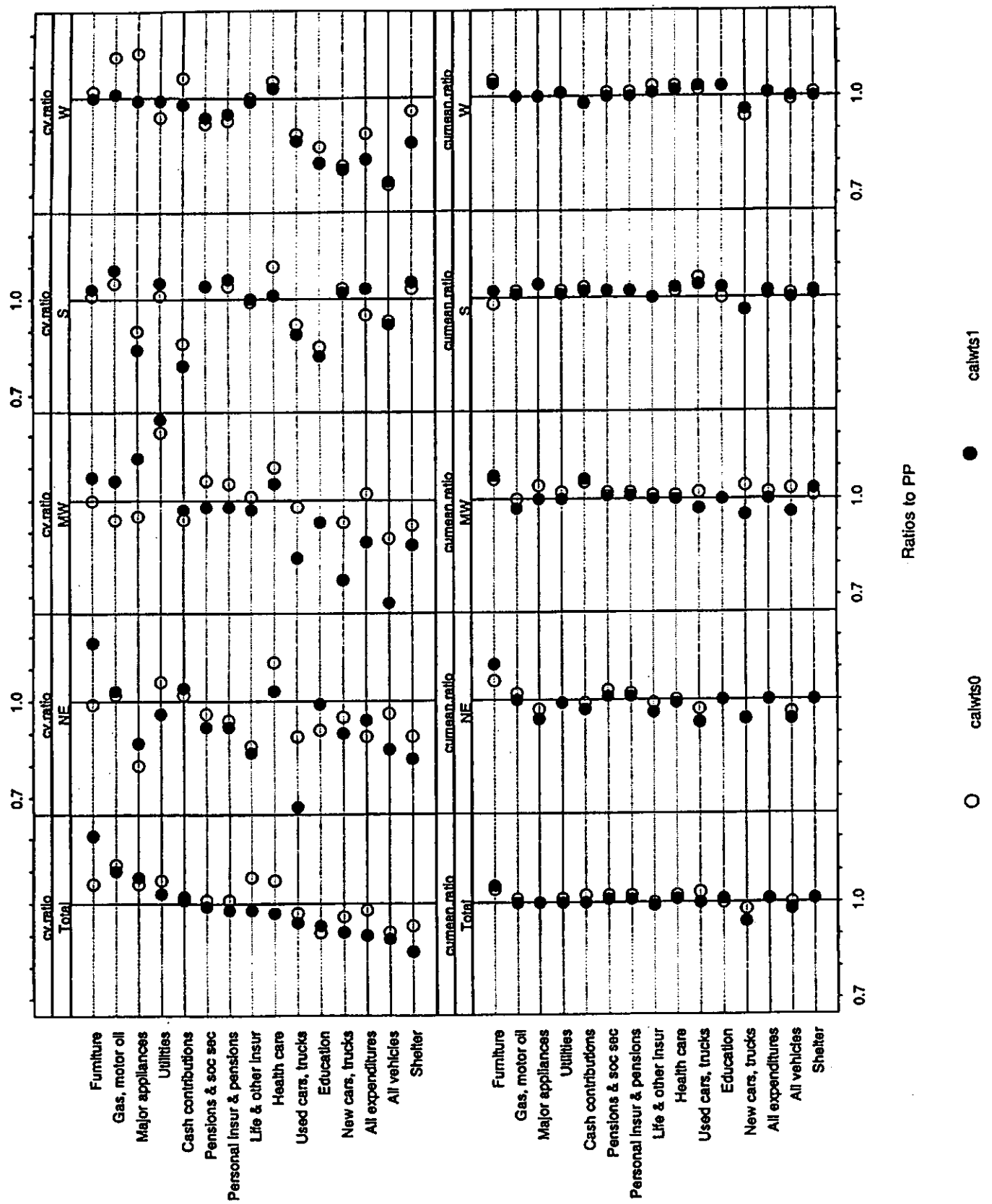


Figure 3. Ratios to PP of cv's and means for two weighting methods by region

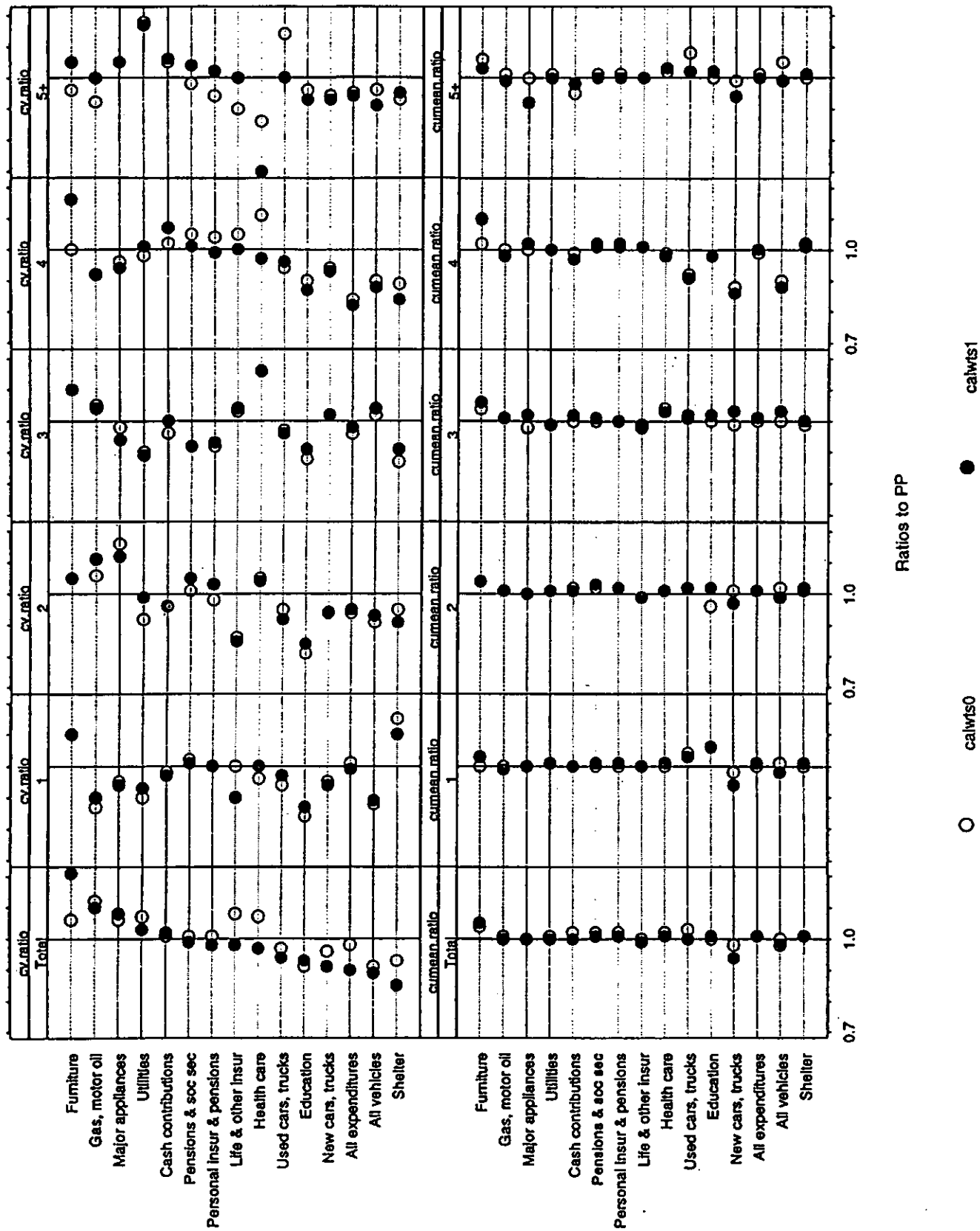


Figure 4. Ratios to PP of cv's and means for two weighting methods by size of CU

of this variable in predicting expenditures. Perhaps, use of another measure of income combined with item imputations for missing incomes would improve calwts2 and calwts3 for domain estimation.

Taking all of the above into consideration, regwts1, calwts1 and calwts4 are efficient choices in this application. Calwts1 has the advantage of non-negative weights over regwts1. Since calwts4 requires 23 auxiliary variables as opposed to calwts1's 18, calwts1 is the more parsimonious choice. Subsequent to the analysis discussed here, we performed a similar study using a full year's data for both the Interview and Diary Surveys for 1990. Results were similar to those reported here and a final set of 24 auxiliaries was adopted based on number of persons by age, race, sex, region, urban × region, and number of CU's by tenure, and an intercept. The conversion of CE estimation to restricted regression is now underway.

4. CONCLUSION

The objective of this study was to investigate methods for deriving household weights that did not depend on the weight of one single member of the household. Different types of weights based on the regression estimation procedure were presented and their relative merits evaluated. Regression estimation incorporates the current survey post-stratification methods in which the weighted sum of the persons in each post-stratum is forced to be equal to an independent census count of that number. This is accomplished via auxiliary variables that are incorporated into the regression model. It also automatically produces for each sample household a weight that does not depend on any single one of its members.

We studied eight types of weights that came from five different regression models. In order to eliminate the undesirable negative weights that can result from ordinary least-squares regression estimation, restricted regression estimators were adapted to the present problem. Restricted regression has the flexibility to restrict the possible deviation of each final weight from its base weight while adhering to the properties discussed above. This, in particular, allows the constraint of positive weights. The restricted regression weights are easily computed via matrix-oriented software like S-Plus™ or SAS/IML™.

Restricted regression, and more generally, restricted calibration have a number of attractive features for household surveys, like the one studied here, but also for surveys of other types of units like hospitals, schools, or business establishments where a variety of auxiliary data may be available. Given past data on target variables, standard model building procedures can be used for the selection of auxiliary variables. The properties of regression estimation can be used to choose the predictors optimally in order to reduce the redundancy of information that gets incorporated into the survey estimation procedure. This is one of the greatest advantages of using an estimator that has a vast and tested literature behind it. Good predictors may include qualitative variables, *e.g.*, age, race, type of hospital (general medical, psychiatric, *etc.*), type of business (manufacturing, retail trade, *etc.*) that might be often used in stratification or post-stratification. The predictors can also be quantitative variables like family income, annual sales, number of students at different levels, or the number of inpatient days to name but a few. In our application, including an intercept also led to noticeably smaller standard errors of survey estimates. The regression approach also allows data at different levels to be easily incorporated in estimation. In the household survey studied here, auxiliaries on both persons and households were included.

The immense flexibility of regression gives practitioners options they might not otherwise have. If new, pertinent predictor variables become available, software for regression estimation can accommodate them simply by changing the matrix of auxiliaries and vector of population controls. Software that is rigidly written to perform only post-stratification or ratio estimation with a single auxiliary, for example, might have to undergo a major overhaul to change the estimator. Of course, if the estimator is one of the less general post-stratification or the ratio types, regression software will often handle it as a special case. In the United States, an extremely large continuing household survey is being contemplated (Love, Alexander and Dalzell 1995) that will provide very precise estimates of many characteristics that may be used as control totals in smaller surveys. The restricted regression approach positions the CE Survey to smoothly incorporate such new data in estimation should it become available.

ACKNOWLEDGMENTS

Any opinions expressed are those of the authors and do not represent policy of the Bureau of Labor Statistics. We are grateful to the referees and editor for their useful comments.

REFERENCES

ALEXANDER, C.H. (1987). A class of methods for using person controls in household weighting. *Survey Methodology*, 13, 183-198.

BARDSLEY, P., and CHAMBERS, R.L. (1984). Multipurpose estimation from unbalanced samples. *Applied Statistics*, 33, 290-299.

BETHLEHEM, J.G., and KELLER, W.J. (1987). Linear weighting of sample survey data. *Journal of Official Statistics*, 3, 141-153.

CASADY, R.J., and VALLIANT, R. (1993). Conditional properties of post-stratified estimators under normal theory. *Survey Methodology*, 19, 183-192.

CLEVELAND, W.S. (1993). *Visualizing Data*. Summit, New Jersey: Hobart Press.

DEVILLE, J.-C., and SÄRNDAL, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.

DEVILLE, J.-C., SÄRNDAL, C.-E., and SAUTORY, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, 88, 1013-1020.

ESTEVAO, V., HIDIROGLOU, M.A., and SÄRNDAL, C.-E. (1995). Methodological principles for a generalized estimating system at Statistics Canada. *Journal of Official Statistics*, 11, 181-204.

FULLER, W.A., LOUGHIN, M.M., and BAKER, H.D. (1993). Regression weighting for the 1987-1988, Nationwide Food Consumption Survey. Unpublished report submitted to the United States Department of Agriculture.

LEMAÎTRE, G., and DUFOUR, J. (1987). An integrated method for weighting persons and families. *Survey Methodology*, 13, 199-207.

LOVE, S., ALEXANDER, C.H., and DALZELL, D. (1995). Constructing a major survey – operational plans and issues of continuous measurement. *Proceedings of the Section on Survey Research Methods, American Statistical Association*. To appear.

McCARTHY, P.J. (1969). Pseudo-replication: half-samples. *Review of the International Statistical Institute*, 37, 239-264.

RAO, J.N.K. (1994). Estimating totals and distribution functions using auxiliary information at the estimation stage. *Journal of Official Statistics*, 10, 153-165.

SÄRNDAL, C.-E., SWENSSON, B., and WRETMAN, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

STUKEL, D.M., and BOYER, R. (1992). Calibration estimation: an application to the Canadian Labor Force Survey. Working Paper, SSMD-92-009E, Ottawa: Statistics Canada.

ZIESCHANG, K.D. (1990). Sample weighting methods and estimation methods in the Consumer Expenditure Survey. *Journal of the American Statistical Association*, 85, 986-1001.

Survey Methodology, December 1996
Vol. 22, No. 2, pp. 139-146
Statistics Canada

A Transformation Method for Finite Population Sampling Calibrated With Empirical Likelihood

GEMAI CHEN and JIAHUA CHEN[1]

ABSTRACT

In this paper, we study a confidence interval estimation method for a finite population average when some auxiliary information is available. As demonstrated by Royall and Cumberland in a series of empirical studies, naive use of existing methods to construct confidence intervals for population averages may result in very poor conditional coverage probabilities, conditional on the sample mean of the covariate. When this happens, we propose to transform the data to improve the precision of the normal approximation. The transformed data are then used to make inference on the original population average, and the auxiliary information is incorporated into the inference directly, or by calibration with empirical likelihood. Our approach is design-based. We apply our approach to six real populations and find that when transformation is needed, our approach performs well compared to the usual regression method.

KEY WORDS: Finite population; Sampling; Confidence interval; Transformation; Empirical likelihood.

1. INTRODUCTION

Let (x_i, y_i), $i = 1, 2, ..., N$ be values associated with N units in a finite population. For unit i, y_i is the variable of interest and x_i is an auxiliary variable. One of the most extensively studied finite population problems is the estimation of the population average $\bar{y} = (y_1 + ... + y_N)/N$ (or total $N\bar{y}$) under various sampling schemes. We shall focus on the simple random sampling scheme in this paper, because the nature of the problems we want to study can be better seen from this scheme and the results obtained here can be easily generalized into other sampling schemes of which the simple random sampling scheme is the building block.

It is often true that some information about the auxiliary variable x is known and can be used to make inference about $\bar{y}$. For example, let $S = \{1, ..., i, ..., N\}$ and let $s \subset S$ be a simple random sample of size n. When $\bar{x} = (x_1 + ... + x_N)/N$ is known, and x and y are correlated, the population average $\bar{y}$ can be estimated by the ratio estimator $\hat{\bar{y}} = (\bar{y}_s/\bar{x}_s)\bar{x}$, or by the regression estimator $\hat{\bar{y}} = \bar{y}_s + b(\bar{x} - \bar{x}_s)$, where $\bar{x}_s$ and $\bar{y}_s$ are the sample averages of x and y, respectively, and $b = \sum(x_i - \bar{x}_s)(y_i - \bar{y}_s)/\sum(x_i - \bar{x}_s)^2$.

Under very general conditions, both the ratio estimator and the regression estimator are asymptotically normal; see Scott and Wu (1981), Bickel and Freedman (1984), and Theorem 2.1 of Section 2. Hence, if v is a carefully chosen estimator of the variance of $\hat{\bar{y}}$, the standardized variable $(\hat{\bar{y}} - \bar{y})/\sqrt{v}$ is customarily treated to have the standard normal distribution. Therefore, if z_α denotes the upper α-percentile of the standard normal distribution, then

$$(\hat{\bar{y}} - z_\alpha\sqrt{v},\ \hat{\bar{y}} + z_\alpha\sqrt{v}) \qquad (1.1)$$

will produce an approximate $100\ (1 - 2\alpha)\%$ confidence interval for $\bar{y}$.

Confidence interval (1.1) is widely used in practice. However, problems arise when it is applied to certain populations. Royall and Cumberland (1981a, 1981b, 1985) studied the ratio and regression estimators and applied them to six real populations where strong correlations between x and y seemed to exist. (See Section 3 for a summary of the six populations.) Various estimators of the variance of $\hat{\bar{y}}$ were used. It was found that the actual conditional coverage rate of the confidence interval (1.1), conditional on $\bar{x}_s$, depended heavily on the size of $\bar{x}_s$ and were usually much lower than the claimed coverage rate, even with the most adaptive variance estimator. For example, the 95% confidence interval for a population named Counties 70 had a conditional coverage rate 76% with the jackknife variance estimator when $\bar{x}_s$ was small, and the conditional coverage rate could go as low as 50% with other variance estimators.

The above mentioned studies point to the need to construct confidence intervals that "will live up to their name" (Royall and Cumberland 1985, p. 359). However, up to now there has been little progress made in this direction. In this paper, we present some results from studying an alternative procedure for constructing confidence intervals and from applying it to the six populations studied by Royall and Cumberland and many others. As will be shown in Section 3, the conditional coverage rate of our confidence intervals is more accurate.

Two important ideas, namely, transformation and empirical likelihood, are used simultaneously to attack the problems encountered by Royall and Cumberland in particular, and to develop a new procedure in general. As explained in Cochran (1977, p. 150), the preference in sample survey theory is to make, at most, limited assumptions about the frequency

[1] Gemai Chen, Department of Mathematics and Statistics, University of Regina, Regina, Saskatchewan, Canada, S4S 0A2; Jiahua Chen, Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario, Canada, N2L 3G1.

distribution followed by the data in the sample. However, ratio or regression estimator can help obtain increased precision by taking advantage of the correlation between y_i and x_i. This, of course, can be described by some assumption(s), such as an approximate linear relationship between y and x. Although almost no further assumptions are necessary to use the ratio or regression approach, the procedure (1.1) is clearly based on an normal approximation. But as it is well known, the normal approximation can be very poor when the population distribution is severely skewed and the sample size is small. In terms of procedure (1.1), the closer the estimator distribution is to the normal, the better one can construct confidence intervals. If the population distribution is severely skewed, a transformation may produce a population distribution that is at least more symmetric, so that the normal approximation for the estimator is more accurate.

When using the ratio and regression estimators, knowing $\bar{x}$ is crucial to gain improvement over the use of sample mean. In our proposed procedure, the complete information about the auxiliary variable x can be incorporated. But if $\bar{x}$ is the only auxiliary information available, it is difficult to use this information directly when a transformation is involved, because any non-linear transformation obscures the link between $\bar{x}$ and $\bar{y}$. In this second case, we find the method of empirical likelihood very helpful in solving our problem; see particularly Owen (1988, 1990) and Chen and Qin (1992) for references. The empirical likelihood method in this situation can also be regarded as a calibration method as discussed in Deville and Särndal (1992). This approach rescues us from losing information about x after transforming the data.

There have been many discussions on how to use transformations to make better inference on the transformed scale (Box and Cox 1964; Carroll and Ruppert 1988; Calvin and Sedransk 1991, and the references therein). There have also been some studies on how to make inference on the original scale, after a transformation is applied (Carroll and Ruppert 1984; Elliott 1977). What is new with our procedure is the attempt to link the above two steps by combining transformation with auxiliary information and/or by applying empirical likelihood method when necessary.

The details of our procedure are given in Section 2. Then our procedure is applied to the six populations studied by Royall and Cumberland in Section 3. The validity of our procedure in an arbitrary setting is demonstrated in Section 4 and some comments are made at the end of the paper.

2. THE NEW PROCEDURE

As mentioned in the last section, a problem with the confidence interval (1.1) is that it will fail if the distribution of $(\hat{\bar{y}} - \bar{y})/\sqrt{v}$ is severely asymmetric and far from the normal distribution. The problem can be inherited from the skewness of the population distribution. When the skewness is severe, a central confidence interval procedure like (1.1) is doomed to fail. The basic model employed by Royall and Cumberland (1981a, 1981b, 1985) is

$$y_i = \alpha + \beta x_i + \epsilon_i, \tag{2.1}$$

with $E(\epsilon_i) = 0$, $V(\epsilon_i) = \sigma^2$ and $\mathrm{Cov}(\epsilon_i, \epsilon_j) = 0$, for $i \neq j$. It is easy to find that for the six real populations studied by Royall and Cumberland, the corresponding error distributions are very skewed. These observations lead us to consider transforming the variables y and/or x, and consider the model

$$h(y_i) = \alpha + \beta g(x_i) + \sigma\epsilon_i, \tag{2.2}$$

where $h(\cdot)$ and $g(\cdot)$ are two monotone functions. There are many families of transformations suggested in the literature. One commonly used family is the Box-Cox power transformation family defined by

$$f(x,\lambda) = \begin{cases} (x^\lambda - 1)/\lambda & \text{when } \lambda \neq 0, \\ \log(x) & \text{when } \lambda = 0. \end{cases}$$

Model (2.1) is a special case of (2.2) when both h and g equal $f(x, 1)$.

The choice of transformations in model (2.2) might be suggested by an examination of the sample x's and y's based on a possible model relationship, or by our subject knowledge about the population under investigation. For example, for the six populations discussed in Royall and Cumberland, the population distributions are severely skewed towards the right which can be learned from the nature of the finite populations. Therefore, a log transformation may make them all less skewed. Other more objective methods of choosing transformations are discussed in Section 4.

We emphasize that models (2.1) and (2.2) are used here to motivate transformations, point estimators, or confidence interval procedures. Our study of conditional coverage rates will, however, be based on the probability measure generated by the design, as in Royall and Cumberland (1985). For this purpose, we embed our finite population in a sequence of populations indexed by k. This means that a sub-index k is needed to write $N = N_k$ and $n = n_k$, *etc.*, but for simplicity, we will suppress the index k if there is no possibility for confusion.

Let $v_i = h(y_i)$, $u_i = g(x_i)$, $\bar{v}_N = N^{-1}\sum_{i=1}^N v_i$ and $\bar{u}_N = N^{-1}\sum_{i=1}^N u_i$. Define

$$\beta_N = \frac{\sum_{i=1}^N (u_i - \bar{u}_N)v_i}{\sum_{i=1}^N (u_i - \bar{u}_N)^2},$$

$$\alpha_N = \bar{v}_N - \beta_N \bar{u}_N,$$

$$e_i = v_i - (\alpha_N + \beta_N u_i),$$

$$\sigma_N^2 = \frac{1}{N-1}\sum_{i=1}^N e_i^2.$$

Suppose $s \subset S$ is a simple random sample of size n. We similarly define

$$\hat{\beta} = \frac{\sum_{i\in s}(u_i - \bar{u}_s)v_i}{\sum_{i\in s}(u_i - \bar{u}_s)^2},$$

$$\hat{\alpha} = \bar{v}_s - \hat{\beta}\,\bar{u}_s,$$

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i\in s}(v_i - \hat{\alpha} - \hat{\beta}u_i)^2,$$

where $\bar{u}_s$ and $\bar{v}_s$ are the sample averages.

Denote the inverse function of $h(\cdot)$ by $h^{-1}(\cdot)$. Then the fitted value of y_i is

$$\hat{y}_i = h^{-1}(\hat{\alpha} + \hat{\beta}\,u_i). \tag{2.3}$$

We discuss confidence interval estimation of $\bar{y}$ in two cases. In the first case where all x_i $(i = 1, ..., N)$ are known, a natural estimator of $\bar{y}$ is $(\sum_{i\in s} y_i + \sum_{i\in s}\hat{y}_i)/N$. However, for the purpose of constructing confidence intervals for $\bar{y}$, we study the distribution of

$$\hat{\bar{y}}(\hat{\alpha},\hat{\beta}) = \frac{1}{N}\sum_{i=1}^{N}\hat{y}_i = \int_{-\infty}^{\infty} h^{-1}(\hat{\alpha} + \hat{\beta}u)\,dF_N(u) \tag{2.4}$$

instead, where $F_N(u)$ is the empirical distribution function of the u_i $(i = 1, ..., N)$. Clearly, the distribution of $\hat{\bar{y}}(\hat{\alpha},\hat{\beta})$ is determined by the distribution of $(\hat{\alpha},\hat{\beta})$ which is descibed in the following design-based theorem.

Theorem 2.1 Suppose that when $k \to \infty$, both $n = n_k$ and $N - n = N_k - n_k$ go to ∞ and

1. $\bar{u} = \lim_{k\to\infty} N^{-1}\sum_{i=1}^{N} u_i$ exists.
2. $N^{-1}\sum_{i=1}^{N} u_i^4 = O(1)$.
3. $\sigma_u^2 = \lim_{k\to\infty}\sigma_{u,N}^2 = \lim_{k\to\infty}(N-1)^{-1}\sum_{i=1}^{N}(u_i - \bar{u}_N)^2$ exists and is greater than zero.
4. $\sigma^2 = \lim_{k\to\infty}\sigma_N^2 = \lim_{k\to\infty}(N-1)^{-1}\sum_{i=1}^{N} e_i^2$ exists and is greater than zero.
5. $N^{-1}\sum_{i=1}^{N}|e_i|^3 = O(1)$, $N^{-1}\sum_{i=1}^{N}|(u_i - \bar{u}_N)e_i|^3 = O(1)$.
6. $r = \lim_{k\to\infty}(\sigma_{u,N}^2\,\sigma_N^2)^{-1}N^{-1}\sum_{i=1}^{N}(u_i - \bar{u}_N)^2 e_i^2$ exists and is greater than zero.
7. $f = \lim_{k\to\infty} n/N$ exists and is less than 1.

Then

(1) $\sqrt{n}(\hat{\alpha} - \alpha_N, \hat{\beta} - \beta_N)'$ converges in distribution to the bivariate normal distribution $N_2(0,\Sigma)$, where

$$\Sigma = \begin{pmatrix} 1 + \frac{\bar{u}^2}{\sigma_u^2}r & -\frac{\bar{u}}{\sigma_u^2}r \\ -\frac{\bar{u}}{\sigma_u^2}r & \frac{1}{\sigma_u^2}r \end{pmatrix}(1-f)\,\sigma^2.$$

(2) Let B_n be any joint $100(1-\gamma)\%$ confidence region for (α_N, β_N) and define G_n by

$$G_n = \{\hat{\bar{y}}(\alpha,\beta) : (\alpha,\beta) \in B_n\}, \tag{2.5}$$

then,

$$\text{Prob}\{\bar{y}(\alpha_N, \beta_N) \in G_n\} \geq 1 - \gamma,$$

where $\bar{y}(\alpha_N, \beta_N) = \sum_{i=1}^{N} h^{-1}(\alpha_N + \beta_N u_i)/N$.

The proof is deferred to the Appendix.

We note that without underlying normality on the errors, it is not easy to get an exact confidence region B_n for (α_N, β_N) for a specified confidence level $1 - \gamma$. The B_n used in the following discussion and the expressions built upon it are, therefore, approximate.

Theorem 2.1 allows us to construct confidence intervals for $\bar{y}(\alpha_N,\beta_N)$, but $\bar{y}(\alpha_N,\beta_N)$ is not equal to $\bar{y}$ in general. This is an intrinsic problem as long as a non-linear transformation is used. If only a point estimator is needed, we would use the regression estimator currently, and we suggest that the methodology developed in this paper be used for interval estimation. Bias corrections for $\hat{\bar{y}}(\hat{\alpha},\hat{\beta})$ are, however, possible, and a specific one is used in our simulation study. Work on general corrections is under study.

According to Theorem 2.1, G_n is a conservative confidence interval for $\bar{y}(\alpha_N,\beta_N)$, which can also be regarded as an approximate confidence interval for $\bar{y}$. To improve the coverage rate of G_n, observe that the contours of $\hat{\bar{y}}(\alpha,\beta)$ in a small neighborhood of $O = (\hat{\alpha},\hat{\beta})$ are approximately parallel straight lines on the $\alpha\beta$ plane; see Figure 1. Let (a, b) be the

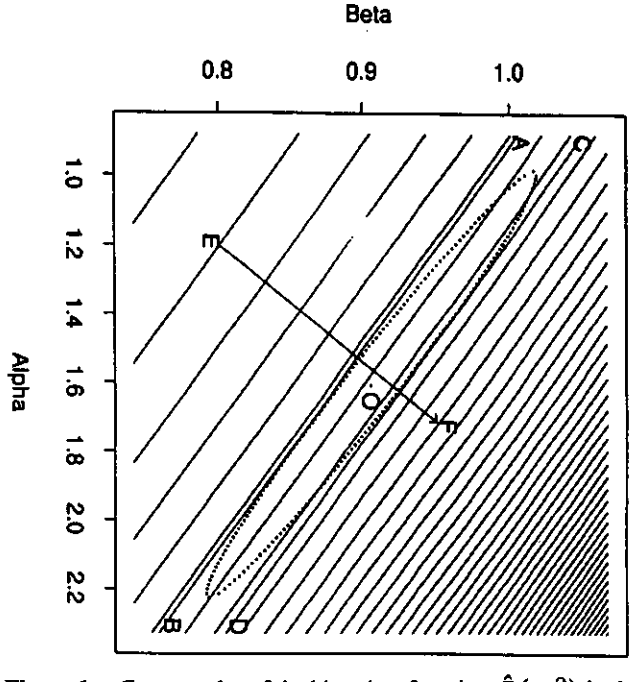


Figure 1. Contour plot of the bi-variate function $\hat{\bar{y}}(\alpha,\beta)$ in the neighbourhood of $O = (\hat{\alpha},\hat{\beta})$, based on a random sample of size 32 taken from population Cancer

directional cosines of the direction $\overrightarrow{EF}$ along which the contours increase. Then $\hat{\bar{y}}(\alpha,\beta)$ is (approximately) a monotone function of $T_n = a(\alpha - \hat{\alpha}) + b(\beta - \hat{\beta})$, where T_n is the corresponding change along the direction $\overrightarrow{EF}$ to the changes in α and β. A natural choice of B_n is

$$B_n = \{(\alpha,\beta) : |a(\alpha - \hat{\alpha}) + b(\beta - \hat{\beta})| \le c\hat{\sigma}t(\gamma/2; n-2)\},$$

where $c^2 = \mathrm{Var}(T_n)/\sigma^2$, $\mathrm{Var}(T_n)$ is the variance of T_n, and $t(\gamma/2; n-2)$ is the upper $\gamma/2$-percentile of the t distribution with $n-2$ degrees of freedom. This B_n is the region between two parallel straight lines AB and CD in Figure 1.

A drawback of the above B_n is that it is an *unbounded* region. If the contours of $\hat{\bar{y}}(\alpha,\beta)$ are not close to be parallel and/or straight, this B_n will lead to very conservative confidence intervals. To guard against this possibility, we construct a bounded elliptic region C_n defined by those (α, β) that satisfy

$$\left\{ n(\alpha - \hat{\alpha})^2 + 2n\bar{u}_s(\alpha - \hat{\alpha})(\beta - \hat{\beta}) + n\left(\bar{u}_s^2 + r_s^{-1}\frac{\sum_{i\in s}(u_i - \bar{u}_s)^2}{n-1} \right)(\beta - \hat{\beta})^2 \right\} \le \left(1 - \frac{n}{N}\right)\hat{\sigma}^2 t^2(\gamma/2; n-2),$$

where $(1 - n/N)$ is part of the variances of $\hat{\alpha}$ and $\hat{\beta}$, because we are doing sampling without replacement from a finite population, and

$$r_s = \frac{n^{-1}\sum_{i\in s}(u_i - \bar{u}_N)^2(v_i - \hat{\alpha} - \hat{\beta}u_i)^2}{\left\{n^{-1}\sum_{i\in s}(u_i - \bar{u}_N)^2\right\}\left\{(n-2)^{-1}\sum_{i\in s}(v_i - \hat{\alpha} - \hat{\beta}u_i)^2\right\}} \tag{2.6}$$

is a sample estimate of the quantity r in Theorem 2.1. The C_n thus defined is represented by the region inside the ellipse in Figure 1 and has the property that it touches both boundary lines of B_n regardless of the direction (a, b). Therefore, when $\hat{\bar{y}}(\alpha,\beta)$ is indeed a monotone function of T_n, C_n produces the same confidence interval for $\bar{y}$ as B_n does. However, C_n is less vulnerable than B_n if the contours of $\hat{\bar{y}}(\alpha,\beta)$ are not close to be parallel and/or straight, because C_n shrinks to one point as n increases. A confidence interval for $\bar{y}$ corresponding to C_n is defined as

$$I_n = \{\hat{\bar{y}}(\alpha,\beta) : (\alpha,\beta) \in C_n\}. \tag{2.7}$$

As the error distributions are more symmetric after the transformation, the new confidence interval based on C_n is therefore expected to be better than the confidence interval without transformation. Note that since all x_i are known, other approaches, such as optimal stratification and post-stratification, may be better. However, optimal stratification may not be possible in some cases as discussed in Cochran (1977, p. 134). Also research is needed on the use of post-stratification when the error distributions are severely skewed.

We now turn to the discussion of the second case where $\bar{x} = (x_1 + \ldots + x_N)/N$ is known, but x_i, $i = 1, \ldots, N$, are unknown. If we want to proceed as in the first case, one approach is to estimate $F_N(u)$ and somehow make use of the information in $\bar{x}$. The following empirical likelihood methodology is found to be an effective way of doing this. We outline the main ideas here; the interested reader should consult Owen (1988, 1990) and Chen and Qin (1992) for more details. The key idea is to maximize the (empirical) likelihood functions under various restrictions formed by the knowledge about some aspects of the parameters. For example, in our problem, the knowledge is $\bar{x}$. It is shown by Chen and Qin (1992) that the resulting estimators with the presence of restrictions are asymptotically more efficient than those without restrictions.

Specifically, we estimate $F_N(u)$ in (2.4) by

$$\hat{F}_N(u) = \sum_{i\in s} p_i I[u_i \le u], \tag{2.8}$$

where the p_i are chosen by maximizing

$$\prod_{i\in s} p_i \tag{2.9}$$

subject to

$$p_i \ge 0, \quad \sum_{i\in s} p_i = 1, \quad \sum_{i\in s} p_i x_i = \bar{x}. \tag{2.10}$$

If y_i, $i \in s$ are regarded as realizations of the random variables Y_i, $i \in s$, with distribution function F, the p_i in (2.9) can be defined by $p_i = F(Y_i) - F(Y_i-)$, and (2.9) is called the empirical likelihood function in Owen (1990).

Deville and Särndal (1992) look at the above approach from a calibration point of view. They suggest using unequal weights for different units sampled to reflect their different contributions, while keeping $\sum p_i x_i = \bar{x}$. It is believed that if these weights give a perfect estimate of $\bar{x}$, they should also be good for estimating $\bar{y}$.

The solution to (2.9) and (2.10) will not exist if either the minimum x value in a sample is greater than or equal to $\bar{x}$, or the maximum x value in a sample is less than or equal to $\bar{x}$. When this happens, one remedy is to replace (2.9) with

$$\sum_{i\in s}(np_i - 1)^2, \tag{2.11}$$

subject to a milder constraint

$$\sum_{i\in s} p_i = 1, \quad \sum_{i\in s} p_i x_i = \bar{x}. \tag{2.12}$$

Under (2.11) and (2.12), we have

$$p_i = \frac{1}{n} + (\bar{x} - \bar{x}_s)(x_i - \bar{x}_s)\Big/\sum_{i\in s}(x_i - \bar{x}_s)^2, \tag{2.13}$$

which always exists unless all the x_i in the sample are the same. The latter situation corresponds to the lack of a covariate, which implies $p_i = n^{-1}$ if $\bar{x} = x_i$, or the solution does not exist if $\bar{x} \neq x_i$. The function given in (2.11) is called the Euclidean likelihood, which is asymptotically equivalent to the empirical likelihood (2.9) (Owen 1990).

For our simulation study in Section 3, we suggest a bias correction to be used in our computation. If $h(w) = g(w) = \log(w)$, we suggest a corrected estimator of $\bar{y}$ as

$$\hat{\bar{y}}^*(\hat{\alpha},\hat{\beta}) = \int_{-\infty}^{\infty} \exp\left\{\hat{\alpha} + \hat{\beta} u_i + \frac{1}{2}\hat{\sigma}^2\right\} F_N(u), \qquad (2.14)$$

if all u_i, $i = 1, \ldots, N$ are known, and replace $F_N(u)$ by $\hat{F}_N(u)$ and $\bar{u}_N$ in (2.6) by $\bar{u}_s$ when only $\bar{x}$ is known. This correction is motivated by model-based considerations under a normality assumption. Correspondingly, I_n of (2.7) is corrected as

$$I_n^* = \{\hat{\bar{y}}^*(\alpha,\beta) : (\alpha,\beta) \in C_n\}. \qquad (2.15)$$

When other power transformations are used, similar corrections can be made using the results in Pankratz and Dudley (1987).

3. APPLICATION TO SIX REAL POPULATIONS

The six real populations studied by Royall and Cumberland (1981a, 1981b, 1985) are summarized in Table 1. Attention was given to the variety in the type of data (demographic, economic, *etc.*), and in the logical relationship between the x and y variables, when these populations were chosen. Note that we have added 1 to the y values in population Cancer in order to take the log transformation.

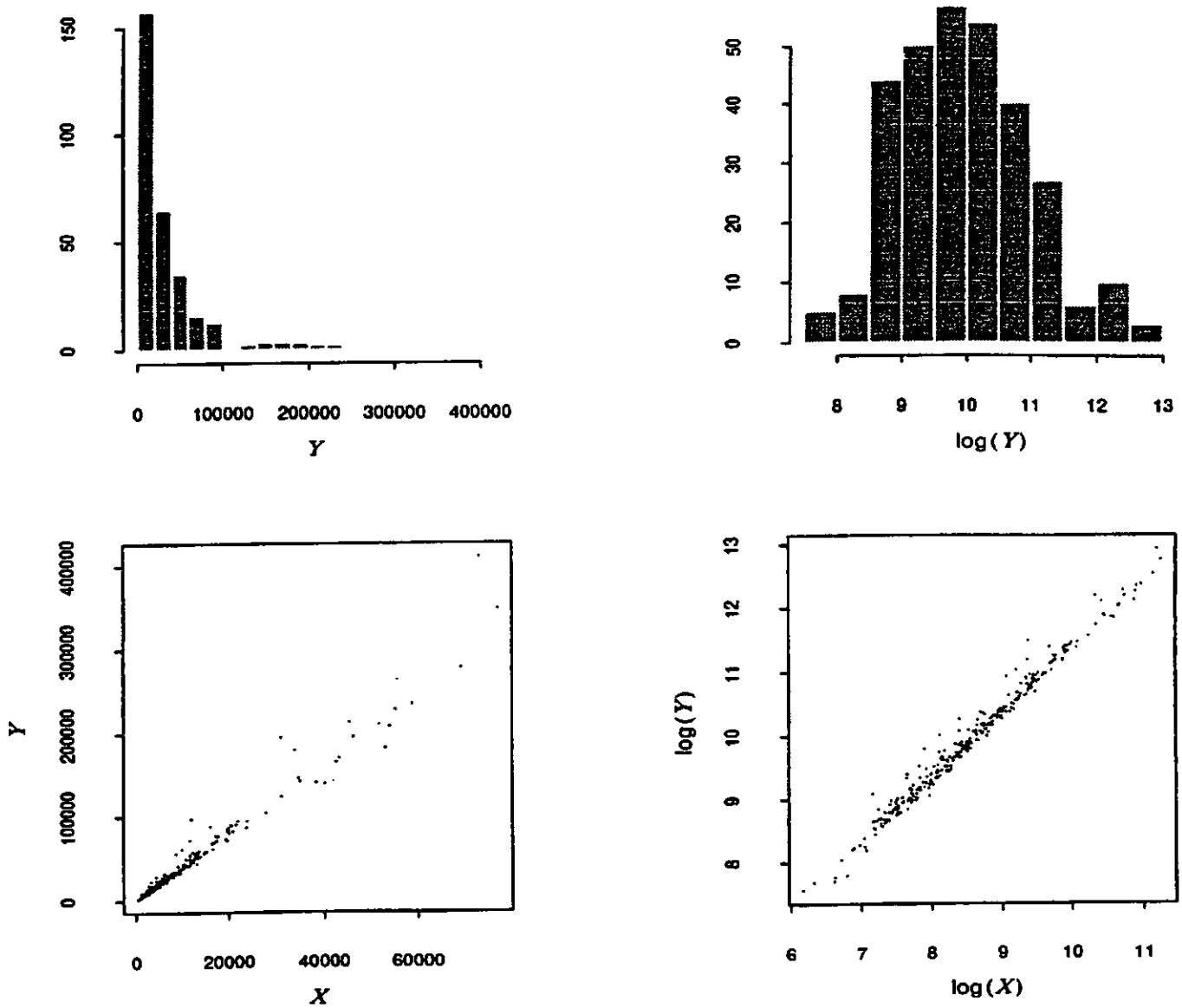


Figure 2. Histograms and scatter plots for the population Counties 70 before and after taking the log transformation

Table 1
Summaries of the Six Populations

Population	N	$\bar{x}$	$\bar{y}$	$\rho(x, y)$	$\rho(\log(x), \log(y))$
Cancer	301	1.1288×10^4	4.0847×10^1	0.967	0.948
Cities	125	2.6602×10^5	2.8553×10^5	0.947	0.953
Counties 60	304	8.9312×10^3	3.2916×10^4	0.998	0.998
Counties 70	304	8.9312×10^3	3.6984×10^4	0.982	0.991
Hospitals	393	2.7470×10^2	8.1465×10^2	0.911	0.943
Sales	331	2.3164×10^9	2.4078×10^9	0.997	0.985

The Counties 70 data are plotted in Figure 2. The histogram of y clearly indicates that the population distribution is severely skewed, while the same plot for $\log(y)$ shows a substantial improvement. Also, the scatter plot of $\log(y)$ *vs.* $\log(x)$ shows a better linear relationship than the scatter plot of y *vs.* x. The need and the benefit of taking transformation is therefore obvious. Similar comments can also be made for populations Cities, Counties 60 and Hospitals. For populations Cancer and Sales, the log transformation (or any other power transformations) seem to weaken the linear relationship that exists between x and y.

Now, we illustrate our new procedure by assuming $h = g = \log$ in (2.2). Equations (2.9) to (2.15) are used to perform the calculations. As in Royall and Cumberland (1981b, 1985), for each of the six populations, we take a simple random sample of size 32 and calculate $\bar{x}_s$, $\hat{\bar{y}}^*(\hat{\alpha}, \hat{\beta})$ and construct a 95% confidence interval I_{32}^*. We repeat this process 10,000 times for each population. The results are reported in Table 2 under the title "Transformation Method" when all x values are known, and under the title "Empirical Likelihood Method" when only $\bar{x}$ is known. The term ratio denotes the average length of the confidence intervals divided by the root mean square error for each population. The non-coverage rate (Ncr) is the proportion of intervals that fail to contain the population average $\bar{y}$. The quantities under the titles "Regression Method (regression variance)" and "Regression Method (jackknife variance)" are obtained using the same method of Royall and Cumberland (1981b) when the usual regression variance and the jackknife variance of $\hat{\bar{y}}$ are used, respectively, but for 10,000 random samples instead of the original 1,000 samples. The results under "Empirical Likelihood Method (created population)" are to be explained in the next Section.

Next, we follow Royall and Cumberland to make *design based* inference and to study the *conditional* coverage properties of several interval estimation procedures. Specifically, we divide the confidence intervals into 20 groups according to the size of $\bar{x}_s$, and plot the proportions of intervals in each group that fail to contain the population average $\bar{y}$. For each specific group, the proportion of those intervals that lay above (below) $\bar{y}$ is plotted above (below) the horizontal line. Figure 3 contains such plots for the Counties 70 data. The top two plots show the non-coverage rates of the regression method using the usual regression variance and the jackknife variance for $\hat{\bar{y}}$; the middle two plots show the non-coverage rates of our new procedure. The bottom left plot will be explained in Section 4. As can be seen clearly, our new procedure with a log transformation produces substantial improvement. For populations Cities, Counties 60 and Hospitals, our new procedure also produces some improvement (plots are not shown here). For populations Cancer and Sales, the new procedure produces very conservative results. This is likely due to the fact that the log transformation (or any power transformation) actually weakens the linear relationship between x and y.

Table 2
Simulation results based on 10,000 simple random samples of size 32

	Cancer	Cities	Counties 60	Counties 70	Hospitals	Sales
	Regression Method (regression variance)					
Ratio	3.26	3.65	3.05	2.90	3.62	2.94
Ncr	0.141	0.116	0.146	0.271	0.098	0.176
	Regression Method (jackknife variance)					
Ratio	4.03	3.88	4.03	3.57	3.93	3.95
Ncr	0.081	0.102	0.083	0.192	0.068	0.079
	Transformation Method (all x values are known)					
Ratio	5.08	4.00	3.75	3.76	4.04	5.41
Ncr	0.018	0.074	0.053	0.069	0.042	0.001
	Empirical Likelihood Method (only $\bar{x}$ is known)					
Ratio	5.12	3.74	3.37	3.69	4.15	4.90
Ncr	0.017	0.082	0.081	0.082	0.037	0.006
	Empirical Likelihood Method (created population)					
Ratio	3.92	3.92	3.97	3.96	3.90	3.99
Ncr	0.057	0.059	0.055	0.058	0.059	0.059

We have also performed simulations for sample sizes 16 and 64, and/or for target coverage rate 90%. The results are very similar to what we have presented.

4. DISCUSSION

We use the log transformation in some of our discussions because it is perhaps the most frequently used transformation in practice. Nevertheless, there exist more objective methods to select transformations. One such a method is the well known Box-Cox power transformation which we have mentioned; see Box and Cox (1964), Box and Tidwell (1962), Carroll and Ruppert (1988). Another recent method is based on a procedure called alternating conditional expectation (ACE) (Breiman and Friedman 1985, De Veaux and Steele 1989).

There are other possibilities to improve conditional coverage rate. One such a possibility is to employ asymmetrical error distributions such as the inverse Gaussian family (Whitmore 1983). Another possibility is to adopt quasi-likelihood (Nelder and Pregibon 1987) to finite population problems.

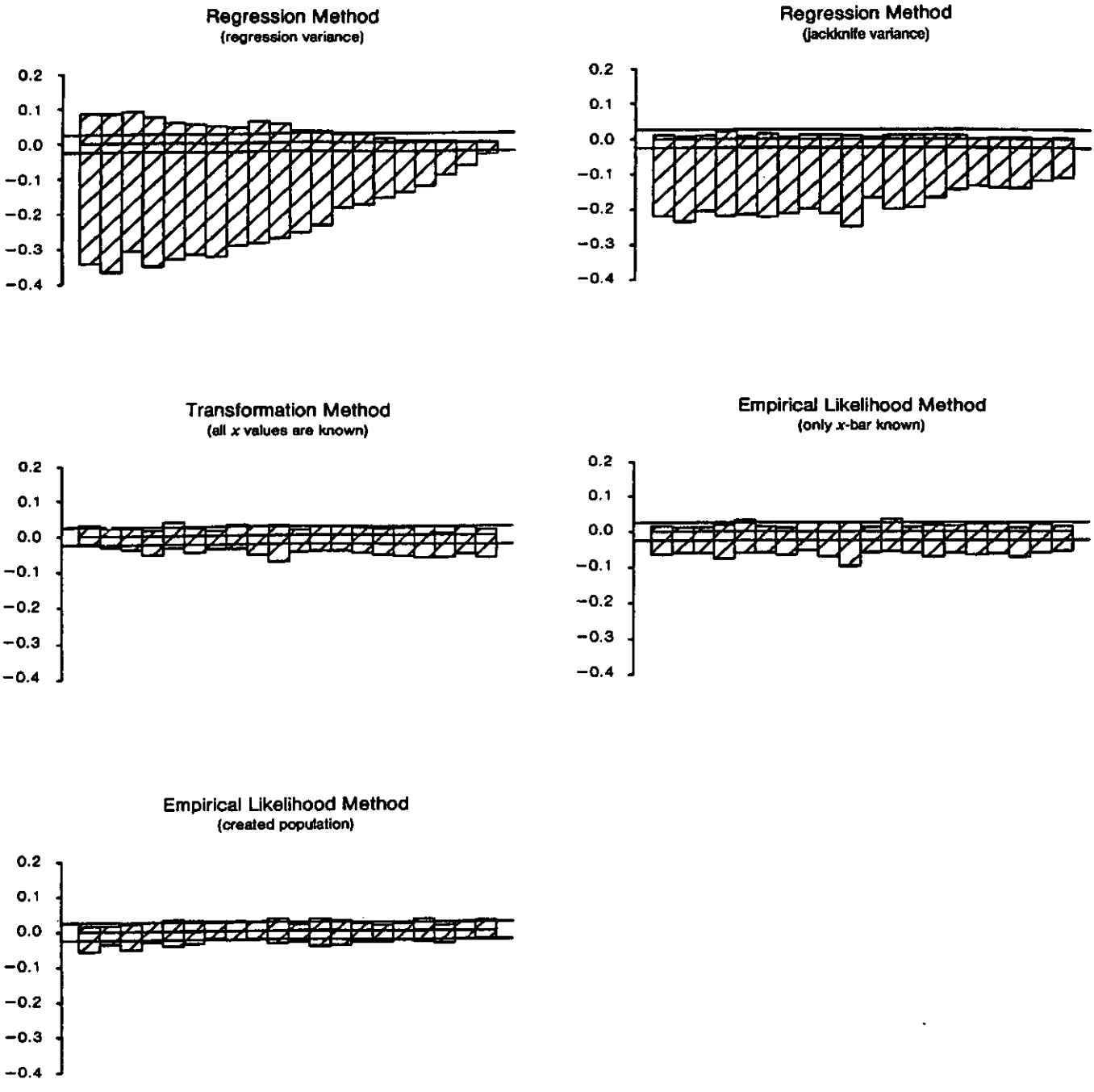


Figure 3. Plots of conditional non-coverage rates for the population Counties 70 based on 10,000 simple random samples of size 32. Reference lines are drawn at 2.5% and the expected non-coverage rate is 5%

The validity of our new procedure is also demonstrated in the following simulation study. For each of the six real populations, we create a new population by replacing the original y_i values with

$$y_i^* = \exp\{\hat{\alpha} + \hat{\beta} \log(x_i) \hat{\sigma} \epsilon_i\},$$

where $\hat{\alpha}, \hat{\beta}$ and $\hat{\sigma}$ are the parameter estimates from fitting model (2.2) with $h = g = \log$ to the old population, and ϵ_i are generated as i.i.d. standard normal variates. Using the six created populations which are fixed, we repeat the simulations as in Section 3 for the case where only $\bar{x}$ is known. Table 2 contains the summary of this simulation study, and the non-coverage plot for the Counties 70 data is shown at the bottom left corner of Figure 3. (Non-coverage plots for other populations look very similar to this plot.) It is clear from this study that when the finite population is generated from a super-population model like (2.2) with a normal error distribution, our new procedure gives the correct conditional coverage rates. Furthermore, we decrease the correlation between

x and y to as low as 0.5 for each of the six populations by increasing θ and repeat the above simulations. The results are as good as those shown in Table 2 and Figure 3.

Although only the simple random sampling scheme is considered in this paper, the proposed procedure is applicable as long as (i) there is a linear correlation between $h(y)$ and $g(x)$ for some monotone functions h and g, and (ii) either $F_N(u)$ or $\hat{F}_N(u)$ can be found. Since the six populations studied here are carefully chosen to be representative, our new procedure is expected to be useful to study other finite populations.

ACKNOWLEDGEMENTS

We would like to thank the referee and the Editor for their comments which greatly improved the presentation. Both authors are supported by grants from the Natural Sciences and Engineering Research Council of Canada.

APPENDIX

Proof of Theorem 2.1 (1). For any given real numbers t_1 and t_2, we have

$$t_1(\hat{\alpha} - \alpha_N) + t_2(\hat{\beta} - \beta_N) =$$

$$t_1 n^{-1} \sum_{i \in s} e_i + \frac{t_2 - t_1 \bar{u}_s}{\sum_{i \in s}(u_i - \bar{u}_s)^2} \sum_{i \in s} (u_i - \bar{u}_s) e_i.$$

From Conditions 1, 2 and 3, we have

$$\bar{u}_s \to \bar{u}, \quad n^{-1} \sum_{i \in s} (u_i - \bar{u}_s)^2 \to \sigma_u^2.$$

Therefore, we can write

$$t_1(\hat{\alpha} - \alpha_N) + t_2(\hat{\beta} - \beta_N) =$$

$$t_1 n^{-1} \sum_{i \in s} e_i + \frac{t_2 - t_1 \bar{u}}{\sigma_u^2} n^{-1} \sum_{i \in s} (u_i - \bar{u}) e_i + o_p(n^{-1/2}).$$

The Lindeberg-Hájek condition is satisfied for $t_1 e_i + t_2 - t_1 \bar{u}/\sigma_u^2 (u_i - \bar{u}) e_i$ under the moment condition 5, see Hájek (1960), Scott and Wu (1981) and Bickel and Freedman (1984). Together with Conditions 4, 6 and 7, the desired result follows by using the Cramér-Wold device.

Proof of Theorem 2.1 (2). Because there may be other values $(\alpha', \beta') \notin B_n$ for which $\hat{\bar{y}}(\alpha', \beta') = \hat{\bar{y}}(\alpha, \beta)$ for some $(\alpha, \beta) \in B_n, G_n$ is always conservative.

REFERENCES

BICKEL, P.J., and FREEDMAN, D.A. (1984). Asymptotic normality and the bootstrap in stratified sampling. *The Annals of Statistics*, 12, 470-482.

BOX, G.E.P., and TIDWELL, P.W. (1962). Transformations of the independent variables. *Technometrics*, 4, 531-550.

BOX, G.E.P., and COX, D.R. (1964). An analysis of transformations. *Journal of The Royal Statistical Society*, Series B, 26, 211-243, discussions 244-252.

BREIMAN, L., and FRIEDMAN, J. (1985). Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association*, 80, 580-597.

CARROLL, R.J., and RUPPERT, D. (1981). On prediction and power transformation family. *Biometrika*, 68, 609-615.

CARROLL, R.J., and RUPPERT, D. (1988). *Transformation and Weighting in Regression*. London: Chapman and Hall.

CALVIN, J.A., and SEDRANSK, J. (1991). Bayesian and frequentist predictive inference for the patterns of care studies. *Journal of the American Statistical Association*, 86, 36-54.

CHEN, J., and QIN, J. (1992). Empirical likelihood estimation for finite populations and the effective usage of auxiliary information. *Biometrika*, 80, 107-116.

COCHRAN, W.G. (1977). *Sampling Techniques*. (3rd ed.) New York: John Wiley.

DE VEAUX, R.D., and STEELE, J.M. (1989). ACE guided-transformation method for estimation of the coefficient of soil-water diffusivity. *Technometrics*, 31, 91-98.

DEVILLE, J., and SÄRNDAL, C. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.

ELLIOTT, J.M. (1977). Statistical analysis of samples of benthic invertebrates. *Freshwater Biological Scientific Publication*, No. 25, (2nd ed.).

HÁJEK, J. (1960). Limiting Distributions in Simple Random Sampling From a Finite Population. Publications in Mathematics of the Hungarian Academy of Science, 5, 361-374.

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

OWEN, A. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75, 237-249.

OWEN, A. (1990). Empirical likelihood confidence regions. *The Annals of Statistics*, 18, 90-120.

PANKRATZ, A., and DUDLEY, U. (1987). Forecasts of power-transformed series. *Journal of Forecasting*, 6, 239-248.

ROYALL, R.M., and CUMBERLAND, W.G. (1981a). An empirical study of the ratio estimator and estimators of its variance. *Journal of the American Statistical Association*, 76, 66-88.

ROYALL, R.M., and CUMBERLAND, W.G. (1981b). The finite-population linear regression estimator and estimators of its variance – An empirical study. *Journal of the American Statistical Association*, 76, 924-930.

ROYALL, R.M., and CUMBERLAND, W.G. (1985). Conditional coverage properties of finite population confidence intervals. *Journal of the American Statistical Association*, 80, 355-359.

SÄRNDAL, C., SWENSSON, B., and WRETMAN, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

SCOTT, A., and WU, C.F. (1981). On the asymptotic distribution of the ratio and regression estimators. *Journal of the American Statistical Association*, 76, 98-102.

WHITMORE, G.A. (1983). A regression method for censored inverse-gaussian data. *The Canadian Journal of Statistics*, 11, 305-315.

Survey Methodology, December 1996
Vol. 22, No. 2, pp. 147-155
Statistics Canada

The Application of McNemar Tests to the Current Population Survey's Split Panel Study

KATHERINE JENNY THOMPSON and ROBIN FISHER[1]

ABSTRACT

Results from the Current Population Survey split panel studies indicated a centralized computer-assisted telephone interviewing (CATI) effect on labor force estimates. One hypothesis is that the CATI interviewing increased the probability of respondent's changing their reported labor force status. The two sample McNemar test is appropriate for testing this type of hypothesis: the hypothesis of interest is that the marginal changes in each of two independent sample's tables are equal. We show two adaptations of this test to complex survey data, along with applications from the Current Population Survey's Parallel Survey split panel data and from the Current Population Survey's CATI Phase-in data.

KEY WORDS: Current Population Survey; Parallel survey; Nonparametric statistics.

1. INTRODUCTION

Results from the Current Population Survey's Parallel Survey split panel study and from the Current Population Survey's CATI Phase-in Project provided some indication of a centralized computer-assisted telephone interviewing (CATI) effect on the United States' monthly labor force estimates (Thompson 1994 and Shoemaker 1993). One hypothesis is that the CATI interviewing increased the probability of respondent's changing their reported labor force status from the first (personal) interview to the second (CATI) interview.

The two sample McNemar test is appropriate for testing this type of hypothesis. The McNemar test (1947) has been generalized to a two sample situation where the hypothesis of interest is that the marginal changes in each of two independent samples' 2×2 tables are equal (Feuer and Kessler 1989). The application presented was for a two sample cohort analysis and assumed simple random sampling.

Certain modifications of the test statistic for a McNemar test are necessary for a complex survey data application. First, because the data are not obtained through a simple random sample and are weighted, a separate estimate of the variance is required. Second, unless the survey has a longitudinal design, a separate link of individuals in two consecutive months' of data must be performed. In general, such a link will include some false matches and exclude some true matches. This adds another source of variance.

We show two adaptations of this test to complex survey data. In particular, we present these tests along with applications to the Current Population Survey's Parallel Survey split panel study and from the Current Population Survey's CATI Phase-in Project. In Section 2 we describe these test modifications including background on the one and two-sample McNemar tests (Section 2.1), modifications for complex survey data (Section 2.2), and some remarks on applications to several months' data (Section 2.3). Section 3 presents applications of these tests specifically to Current Population Survey Parallel Survey Data and to Current Population Survey CATI Phase-in data including background on the two studies (Section 3.1), details of the panel estimates and variance estimates (Section 3.2), diagnostics (Section 3.3), and results (Section 3.4). We make some concluding remarks in Section 4. Details of covariance estimation are included in the appendix.

2. TEST AND MODIFICATIONS

2.1 General

A sample is randomly split into two independent representative samples (split panels). After a baseline measurement is taken in both panels, a new technique is administered in one panel, the treatment panel. The other panel serves as a control.

The records are linked longitudinally after the second measured. A matched response can be +, −, or * (missing). Since this is matched data, the "**" cell will be empty.

This scenario is represented pictorially as

Treatment Panel

Month 2
Treatment

Month 1 / No Treatment	+	−	*	
+	x_{++}	x_{+-}	x_{+*}	$x_{+\cdot}$
−	x_{-+}	x_{--}	x_{-*}	$x_{-\cdot}$
*	x_{*+}	x_{*-}		$x_{*\cdot}$
	$x_{\cdot+}$	$x_{\cdot-}$	$x_{\cdot*}$	n

[1] Katherine Jenny Thompson, Economic Statistical Methods and Programming Division, and Robin Fisher, Housing and Household Economic Statistics Division, United States Bureau of the Census, Washington, DC 20233, U.S.A.

Control Panel
Month 2
No Treatment

Month 1 No Treatment	+	-	*	
+	x'_{++}	x'_{+-}	x'_{+*}	$x'_{+.}$
-	x'_{-+}	x'_{--}	x'_{-*}	$x'_{-.}$
*	x'_{*+}	x'_{*-}		$x'_{*.}$
	$x'_{.+}$	$x'_{.-}$	$x'_{.*}$	n'

where n is not necessarily equal to n'.

For each panel, define

$M_{(12)}$ as the set of cases which have month 1 and month 2 responses (matched cases). This set contains $n_{(12)} = (x_{++} + x_{+-} + x_{-+} + x_{--})$ elements;

$M_{(10)}$ as the set of cases which have month 1 responses, but no month 2 response. This set contains $n_{(10)} = (x_{+*} + x_{-*})$ elements;

$M_{(02)}$ as the set of cases which have month 2 responses, but no month 1 response. This set contains $n_{(02)} = (x_{*+} + x_{*-})$ elements.

Note that the n's are sample sizes and do not have weights.

First, consider the one-sample case. Traditionally, the one-sample McNemar test statistic is constructed from the $n_{(12)}$ and $n'_{(12)}$ matched responses, where a prime (') indicates the control panel. In the one-sample scenario, we test the hypothesis

H_0: $p_{+-} = p_{-+}$, where the p's refer to cell probabilities
H_1: Not H_0

i.e., the hypothesis that the movement from one state to the other (+ to -, or - to +) is zero. We also refer to this movement as the flux.

The one-sample test can be a useful diagnostic in the two-sample situation. We examine the Control panel estimates to see if there is zero movement. Any significant movement in the Treatment panel can be measured as a deviation from zero flux or as a change in the probability of a "+."

The two-sample hypothesis is

H_0: $(p_{-+} - p_{+-}) = (p'_{-+} - p'_{+-})$
H_1: Not H_0.

In other words, the difference in the probabilities of switching in the two directions is the same, regardless of the treatment, or equivalently, the difference in panel fluxes is zero.

The Feuer and Kessler generalization (1989) to a two-sample McNemar test (described in 2.2.1 below) is confined to the $M_{(12)}$ and $M'_{(12)}$ linked sets. We can add an additional assumption, however, to allow the unmatched responses to be included in computation of the test statistics. This assumption motivates the discussion in Section 2.2.2.

2.2 Complex Survey Modifications

2.2.1 Modification One: Longitudinally Linked Data

This method is a straightforward application of the two-sample McNemar test, using longitudinally linked data from a complex survey.

To construct the test statistic, we examine the cell probabilities and note that

$$\begin{aligned}[p_{-+} - p_{+-}] &= [(p_{++} + p_{-+}) - (p_{++} + p_{+-})] \\ &= [p^*_{.+} - p^*_{+.}] \\ &= p^*_2 - p^*_1\end{aligned}$$

where p^*_2 is the marginal probability of a + response month 2, given a matched response for both months; and p^*_1 is the marginal probability of a + response month 1, given a matched response for both months.

The one-sample test statistic constructed from this panel's data is

$$Z^*_1 = \frac{p^\circ_2 - p^\circ_1}{\sqrt{\mathrm{Var}(p^\circ_2 - p^\circ_1)}},$$

where

$$p^\circ_1 = \frac{x_{++} + x_{+-}}{n_{(12)}}, \quad p^\circ_2 = \frac{x_{++} + x_{-+}}{n_{(12)}}.$$

Given two independent panels, the two-sample test statistic is

$$Z^* = \frac{(p^\circ_2 - p^\circ_1) - (p'^\circ_2 - p'^\circ_1)}{\sqrt{\mathrm{Var}(p^\circ_2 - p^\circ_1) + \mathrm{Var}(p'^\circ_2 - p'^\circ_1)}},$$

where

$$p'^\circ_1 = \frac{x'_{++} + x'_{+-}}{n'_{(12)}}, \quad p'^\circ_2 = \frac{x'_{++} + x'_{-+}}{n'_{(12)}}.$$

These results hold regardless of sample design. To extend the results to a complex survey application, we use weighted estimates and complex survey variances and covariances in place of simple random sample variances.

If the survey is designed to collect longitudinal data, then this modification is a natural extension of the method described by Feuer and Kessler. For this type of survey design, an effective mechanism to link individuals from month to month is presumably in place. Often, however, this is not the case, and one data set must be physically linked to another. Consequently, the $n_{(12)}$ elements in the domain will contain some false matches, and some actual matches may be inadvertently excluded. Both the record weights and variance estimates will need to be adjusted to account for the matching. Jabine and Scheuren (1986) provide an excellent summary of the methodological issues arising from the use of linked data, both for model-based and ad-hoc ("hard") record linkage techniques.

2.2.2 Modification Two: Unlinked Data

This method omits the longitudinal linkage step altogether, noting that the construction of the traditional McNemar test statistic can be expressed in terms of estimates of marginal probabilities. Assume that under the null hypothesis, the expected value of $(p_{*.} - p_{.*})$ is zero. This is described for a simple random sampling application in Marascuilo *et al.* (1988).

The domain for the first month of data is given $M_{(12)} \cup M_{(10)}$ which contains $n_{(12)} + n_{(10)} = n_1$ elements. The domain for the second month of data is given by $M_{(12)} \cup M_{(02)}$ which contains $n_{(12)} + n_{(02)} = n_2$ elements.

The one-sample test statistic constructed from the unlinked data is given by

$$Z_1 = \frac{p_2 - p_1}{\sqrt{\text{Var}(p_2 - p_1)}},$$

where

$$p_1 = \frac{x_{+.}}{n_1}, \quad p_2 = \frac{x_{.+}}{n_2}.$$

Given two independent panels, the two-sample test statistic is

$$Z = \frac{(p_2 - p_1) - (p_2' - p_1')}{\sqrt{\text{Var}(p_2 - p_1) + \text{Var}(p_2' - p_1')}},$$

where

$$p_1' = \frac{x_{+.}'}{n_1'}, \quad p_2' = \frac{x_{.+}'}{n_2'}.$$

As with the application described in 2.2.1, all estimates are weighted estimates, and variances are complex survey variances.

2.3 Linear Combinations

We can use our estimated covariance matrix to test linear combinations of $\underline{\lambda}_T$, $\underline{\lambda}_C$, and $\underline{\delta}$ over time, where $\underline{\lambda}_T = \underline{p}_2 - \underline{p}_1$, $\underline{\lambda}_C = \underline{p}_2' - \underline{p}_1'$, and $\underline{\delta} = \underline{\lambda}_T - \underline{\lambda}_C$, and $\underline{p}_1, \underline{p}_2, \underline{p}_1'$ and $\underline{p}_2'$ are vectors containing the marginal probabilities for the time period under consideration.

General linear hypotheses of the form $\underline{K}'\underline{\mu}$ are now easily tested. One might wish to test for contrast by time period, for example testing the average difference from January through June against the remainder of the year's data. Perhaps the most interesting (to our applications) of these tests is of the hypothesis H_0: $\underline{1}'\underline{\mu} = 0$, where $\underline{\mu}$ is the expected value of one of the vectors described above.

Another test of particular interest is the "omnibus hypothesis," where we test H_0: $\underline{\mu} = \underline{0}$. The test statistics for this hypothesis are $\underline{\lambda}_T' \sum_{\lambda(T)}^{-1} \underline{\lambda}_T$, $\underline{\lambda}_C' \sum_{\lambda(C)}^{-1} \underline{\lambda}_C$, and $\underline{\lambda}_\delta' \sum_{\lambda(\delta)}^{-1} \underline{\lambda}_\delta$, each of which has an approximate chi-squared distribution with r degrees of freedom, where r is the dimension of the vector of interest.

3. APPLICATIONS

In this section, we apply the one and two-sample McNemar techniques for unlinked data outlined in 2.2.2 and 2.3 to two separate sets of data: the Current Population Survey's Parallel Survey split panel data and Current Population Survey CATI Phase-in data. Tables 1 and 2 (section 3.4.1) provide the results for Parallel Survey split panel data. Tables 3 and 4 (section 3.4.2) provide the results for the Current Population Survey CATI Phase-in data.

3.1 Background

The official monthly civilian labor force estimates from January 1994 onward are based on data from a comprehensively redesigned Current Population Survey. The redesign included implementation of a new, fully computerized questionnaire, and an increase in centralized computer-assisted telephone interviewing (CATI). To gauge the effect of the Current Population Survey redesign on published estimates, a Parallel Survey was conducted using the new questionnaire and data collection procedures from July 1992 through December 1993. Special studies were embedded in both the Parallel Survey and the Current Population Survey during the same time period to provide data for testing hypotheses about the effects of the new methodological differences on labor force estimates: the Parallel Survey split panel study and the Current Population Survey CATI Phase-in Project (a continuation of the study presented in Shoemaker 1993).

The effect of increased centralized computer-assisted telephone interviewing was of particular interest. Findings from the study described in Shoemaker (1993) had shown that including centralized telephone interviews tended to yield a larger unemployment rate. The two-sample McNemar test appeared to be a good vehicle for examining this phenomenon. In both the Current Population Survey and the Parallel Survey, households are interviewed for 4 consecutive months, not interviewed for the next 8 consecutive months, and then interviewed for another 4 consecutive months. The first and fifth interviews are conducted by a personal visit, and the subsequent interviews are conducted by telephone whenever possible. Thus the first and fifth interviews provide a baseline measurement of labor force status; the second and sixth interviews provide a "post-treatment" measurement of labor force status.

To create the panels for both studies, sample within selected sample areas was randomly divided into two representative panels using systematic sampling methods. The treatment panel was designated as CATI eligible. This meant that the sample households in the panel were eligible for interview at a centralized facility after the initial (first and fifth) interviews. To be interviewed by CATI, a respondent must have a telephone and speak English or Spanish, and must agree to be interviewed in subsequent months by telephone. Not all households in this panel were interviewed by CATI. The other panel served as a control.

The monthly unemployment rate is the primary statistic of interest published from Current Population Survey data. This rate is defined as the estimated number of unemployed persons divided by the estimated number of persons in the civilian labor force (the denominator does not include military personnel, persons under sixteen years old, or people who are no longer looking for work, or retired persons). Our primary goal was to understand how including CATI interviews influenced the probability of changing labor force status, in this case from unemployed to not unemployed (or vice versa). Our statistics for the one and two-sample McNemar tests used unemployment to population ratios, rather than unemployment rates. This allowed for a slightly more precise estimate of the proportion by decreasing the variability of the test statistic.

3.2 Estimates

Each month/panel estimate is an unbiased estimate. That is, the weights used to produce the estimates were strictly a function of the probability of selection: each weight is the product of the baseweight (the inverse probability of selection for a PSU), the weighting control factor (an adjustment for field subsampling), and a split panel factor (an adjustement for the probability of inclusion in a split panel). The split panel factor for the Parallel Survey study was constant by design: nine tenths of the sample was randomly assigned to the treatment panel. The split panel factors for the CPS CATI Phase-in were not constant: the sample in the treatment panel varied on a monthly level, as more sample was randomly assigned to CATI facilities.

Variances of levels were computed with generalized variance functions (GVFs). For more details, see Fisher *et al.* (1993). Robert Fay used his VPLX software (Fay 1990) to calculate replicate estimates of correlation between rotation groups for unemployed and for civilian labor force using September 1992 through December 1993 data from the Current Population Survey. We used these correlations for the test statistics based on unlinked data, assuming that they would not differ by survey (Current Population Survey versus Parallel Survey) or by geography (national versus subnational). We derived an expression for the within-panel correlation for civilian population by relating previously calculated autocorrelations (Fisher and McGuinness 1993) and variance estimates to the individual rotation group estimates. See the appendix for details of the estimation of the correlations.

We did not use the linked modification in our applications for several reasons. The primary reason was the difficulty of longitudinally matching the data. Moreover, we were unable to evaluate the success of our matching. Finally, we did not have any estimates of correlation for the linked data.

Implicit in our analysis of the unlinked data is the assumption that the probability of a nonresponse (or a nonmatch) is random. We assume that the probability of a nonresponse one month is independent of the respondent's labor force classification in the previous month. This assumption is not universally recognized. In fact, Stasny and Fienberg (1984) argue the reverse, and propose several alternative discrete-time models for the use of longitudinally linked CPS data. In our application, the estimates of marginal probabilities based on our (perhaps) poorly matched linked data were almost identical to the estimates based on unlinked data, and so we feel that our analysis did not suffer particularly from our assumption.

3.3 Diagnostics

Small expected sample sizes in individual cells will result in highly variable and consequently unreliable tests. We are not aware of a general method of calculating adequate sample sizes for this type of analysis using complex survey data. Instead, as a naive approach we used a slightly modified version of the traditional Pearson chi-squared test diagnostic to form a cut-off value as follows:

As defined in Section 2.2.2, let

$x_{+\cdot}$ = unweighted unemployed persons in month 1;
$x_{-\cdot}$ = unweighted not-unemployed persons in month 1;
$x_{\cdot+}$ = unweighted unemployed persons in month 2;
$x_{\cdot-}$ = unweighted not-unemployed persons in month 2.

Recall that in the case of the usual contingency table, $E[+-] = x_{+\cdot} x_{\cdot-} / n_{(12)}$, $E[-+] = x_{-\cdot} x_{\cdot+} / n_{(12)}$ under the assumption of independence (and ignoring missing values). In our estimates of expected cell size, we used unlinked marginal data. The sample sizes for the two margins corresponding to the two months are different; that is, the denominators of the expected cell totals are different depending on which margin we examine. Because we could not observe $n_{(12)}$, we estimated it by the geometric mean of n_1 and n_2, which seemed to most closely resemble the expression for the expected cell size. We have not evaluated the effectiveness of the geometric mean versus alternative estimators.

A commonly used rule in contingency table analysis is that expected cell sizes should be at least five. However, both the Current Population Survey and Parallel Survey designs are highly clustered, and we felt that the cut-off value should be adjusted upwards. Accordingly, we multiplied the cut-off value by a design effect. We further increased the cut-off value for expected cell sizes to compensate for the correlation between the rows and columns of our tables to arrive at our final cut-off expected cell size of ten.

3.4 Results

3.4.1 Parallel Survey Split Panel Study

This section presents the formal results from the one and two-sample McNemar tests using unlinked Parallel Survey split panel data. Although this data was collected monthly, small expected cell sizes in the control panel led us to omit several sets of adjacent months from this analysis. Table 1

Table 1
One-Sample McNemar Tests for Individual Parallel Survey Panels – Unlinked Data

Time Frame	Treatment Panel			
	$p_2 - p_1$	$se(p_2 - p_1)$	Z-Statistic	*P*-Value
10/92 – 11/92	-0.62	0.29	-2.18	0.03
11/92 – 12/92	-0.47	0.28	-1.68	0.09
04/93 – 05/93	-0.76	0.27	-2.84	0.00
06/93 – 07/93	-0.04	0.27	-0.16	0.88
08/93 – 09/93	-0.66	0.27	-2.42	0.02
	Control Panel			
	$p_2' - p_1'$	$se(p_2' - p_1')$	Z-Statistic	*P*-Value
10/92 – 11/92	2.44	0.81	3.02	0.00
11/92 – 12/92	0.11	0.83	0.14	0.89
04/93 – 05/93	0.20	0.72	0.27	0.78
06/93 – 07/93	0.97	0.71	1.38	0.17
08/93 – 09/93	-1.73	0.68	-2.54	0.01

provides summary statistics for the one-sample "monthly" tests for each panel which were based on unlinked data from the Parallel Survey's split panels. Table 2 provides summary statistics for the two-sample tests based on unlinked data.

The reported values of p_1, p_2, p_1', and p_2' are percentages of estimated unemployed to estimated total population for the panel. Recall that p_1 and p_1' are the panel ratio of estimated unemployed from the first and fifth interviews to the estimated panel population from the first and fifth interviews; p_2 and p_2' are the panel ratio of estimated unemployed from the second and sixth interviews to the estimated panel population from the second and sixth interviews. Data from the time frame of February 1993 – March 1993 are omitted: a CATI facility was closed during the March interview week because of a blizzard.

The one-sample McNemar tests in Table 1 test the probability that the proportion unemployed does not change between the initial and the subsequent interview within the same panel. We use the Control panel to examine the unemployment flux from one month to the next in the absence of CATI. Note that the two significant point estimates are in the opposite direction.

The entire vector of differences of proportions was found to be significantly different from the zero vector (p-value = 0.00), but the sum of the individual components was not found to be significant (p-value = 0.24). Consequently, we did not test any further linear combinations.

We expected a certain amount of month-in-sample bias to be present in these estimates. In Adams (Bureau of the Census 1991), the estimates of p_1 constructed from the first and fifth months in sample of the full Current Population Survey were roughly six percent larger than their respective second and sixth month-in-sample analogues (p_2). Consequently, estimates of $(p_2 - p_1)$ calculated from the full Current Population Survey data were generally negative. As seen in Table 1, this was not the case with the Parallel Survey Control panel's estimates: counter to our intuition, the estimated difference $(p_2' - p_1')$ is generally positive. This could be a function of the time difference, a geographic difference, or a design difference. Adams used 1987 data from the Current Population Survey to calculate national estimates of biases associated with rotation groups. Thus in each of these one-sample tests, the net movements are intertwined with an unmeasured effect from month-in-sample bias.

Note the negative unemployment flux in the Treatment panel. This observation is supported by the significant result from the formal test of the omnibus hypothesis (p-value = 0.00), and the significant result for the hypothesis $\underline{1}'\underline{\mu} = 0$ (p-value = 0.00).

The two-sample McNemar test results are presented below.

Table 2
Two-Sample McNemar Tests – Unlinked Parallel Survey Data

Time Frame	$(p_2 - p_1) - (p_2' - p_1')$	$se[(p_2 - p_1) - (p_2' - p_1')]$	Z-Statistic	*P*-Value
10/92 – 11/92	-3.06	0.86	-3.58	0.00
11/92 – 12/92	-0.58	0.88	-0.66	0.51
04/93 – 05/93	-0.95	0.77	-1.24	0.22
06/93 – 07/93	-1.02	0.76	-1.34	0.18
08/93 – 09/93	1.08	0.74	1.47	0.14

Individually, the monthly results do not demonstrate a clear difference in the unemployment flux between the two panels. On the other hand, the omnibus test statistic is significant (p-value = 0.00). The mean unemployment flux seems to be lower in the treatment panel as evidenced by the significant test results of the hypothesis $\underline{1}'\underline{\mu} = 0$, where $\underline{\mu}$ is the vector of $((p_2 - p_1) - (p_2' - p_1'))_i$'s, with each element corresponding to a month's estimate (p-value = 0.01).

In these tests, we make statements about contrasts in a table of probabilities, looking for indicators of the effect of a treatment on unemployment movement. As mentioned earlier, some month-in-sample bias is present in the one-sample tests. The tested hypotheses examine combinations of the net movement within a panel and month-in-sample bias. This problem is somewhat mitigated in the two-sample tests. Indeed, if month-in-sample bias is an additive term which affects both panels equally, it will cancel out of the test statistic. Moreover, this effect will be alleviated somewhat in the two-sample test even if it is not the same between the two panels or is multiplicative. Our preliminary sensitivity analysis bore this out: we found that the one-sample tests were sensitive to month-in-sample bias, but that the two-sample tests were not.

The two-sample t-tests presented in Thompson (1994) failed to detect a difference by panel in mean unemployment rate using the Parallel Survey split panel data. This contrasts with the Current Population Survey CATI Phase-in results: over two years, the CATI (Treatment) panel had consistently significantly higher unemployment rates than the non-CATI

(Control) panel. See Shoemaker (1993). In this analysis of Parallel Survey split panel data, we have evidence that the expected value of the proportion unemployed is lower in the presence of CATI. There are, however, some problems with the data. First, as previously mentioned, there is some confounding in the Treatment (CATI) panel, since not all respondents in this panel have their second interview conducted from a centralized telephone facility. Second, in each month the expected sample size in the Control panel cells was near ten, which could be small enough to make the distribution behave unpredictably. This latter problem is not an issue with the Current Population Survey CATI Phase-in study analysis presented in 3.4.2.

3.4.2 Current Population Survey CATI Phase-in Project Results

The Current Population Survey CATI Phase-in project was a continuation of the study presented in Shoemaker (1993). The primary purpose of this study was to measure the effect of including CATI interviewing on the unemployment rate. CATI interviewers in this study used an automated version of the old Current Population Survey paper questionnaire, which had a slightly modified version of the lead-in labor force question. More details are provided in Thompson (1994). The data considered in this paper are from the same time period as the Parallel Survey split panel data examined in 3.4.1: October 1992 through December 1993, again omitting the February 1993 – March 1993 time frame. Expected cell sizes in both the Treatment (CATI) and Control (non-CATI) panels were well over one hundred, and so all other contiguous months of data are included.

The one-sample McNemar test results for both panels are presented in Table 3. Test statistics are constructed with unlinked data. The reported values of p_1, p_2, p_1', and p_2' are percentages of estimated unemployed to estimated total population for the panel.

As with the Parallel Survey split panel data, the one-sample McNemar tests using the CATI Phase-in data test the probability that the proportion unemployed does not change between the initial and the subsequent interview within the same panel. Again, we use the Control panel to estimate the unemployment flux from one month to the next in the absence of CATI. The monthly tests for the Control panel do not appear to exhibit any particular movement. Furthermore, the omnibus hypothesis test was not significant (p-value = 0.29), so we did not test any further linear combinations.

Again basing our expectations on the effects of month-in-sample bias presented in Adams (1991), we believed that the Control panel estimate of p_1' (from the first and fifth months-in-sample) would be larger than its respective second and sixth month-in-sample analog, p_2'. On the average, this was the case: although quite variable, the estimates of p_1' are on the average about 4 percent larger than the estimates of p_2'. Because both panels are representative samples from the same parent sample, we assume that the month-in-sample bias behaves similarly in both panels. The Treatment (CATI) panel estimates of p_2 are *larger* on the average than the estimates of p_1. Given the Control panel's estimates behavior, this phenomenon provides some evidence of a CATI effect.

Table 3
One-Sample McNemar Tests for Individual Current Population Survey Panels – Unlinked Data

Time Frame	Treatment Panel			
	$p_2 - p_1$	se($p_2 - p_1$)	Z-Statistic	P-Value
10/92 – 11/92	1.13	0.16	7.63	0.00
11/92 – 12/92	0.07	0.17	0.44	0.66
12/92 – 01/93	0.43	0.13	3.46	0.00
01/93 – 02/93	0.00	0.14	0.03	0.97
03/93 – 04/93	-0.25	0.14	-1.81	0.07
04/93 – 05/93	0.63	0.13	4.99	0.00
05/93 – 06/93	0.88	0.13	6.56	0.00
06/93 – 07/93	0.84	0.13	6.49	0.00
07/93 – 08/93	-0.07	0.14	-0.51	0.61
08/93 – 09/93	0.42	0.13	3.17	0.00
09/93 – 10/93	0.06	0.12	0.52	0.60
10/93 – 11/93	1.05	0.12	8.45	0.00
11/93 – 12/93	0.18	0.14	1.27	0.20
	Control Panel			
	$p_2' - p_1'$	se($p_2' - p_1'$)	Z-Statistic	P-Value
10/92 – 11/92	0.05	0.47	0.11	0.92
11/92 – 12/92	-0.14	0.47	-0.30	0.76
12/92 – 01/93	0.72	0.43	1.68	0.09
01/93 – 02/93	-0.91	0.43	-2.11	0.03
03/93 – 04/93	-0.16	0.39	-0.40	0.69
04/93 – 05/93	-0.18	0.43	-0.42	0.67
05/93 – 06/93	0.47	0.38	1.22	0.22
06/93 – 07/93	-0.32	0.46	-0.68	0.49
07/93 – 08/93	-0.52	0.39	-1.32	0.19
08/93 – 09/93	-0.54	0.44	-1.21	0.23
09/93 – 10/93	-0.08	0.37	-0.22	0.83
10/93 – 11/93	-0.63	0.42	-1.50	0.13
11/93 – 12/93	-0.09	0.37	-0.23	0.82

Note the movement in the Treatment panel from *not* unemployed to unemployed. This observation is supported by the significant result from the formal test of the omnibus hypothesis (p-value = 0.00), and the significant result for the hypothesis $\underline{1}'\underline{\mu} = 0$ (p-value = 0.00). In contrast to the Parallel Survey results provided in 3.4.1, this data provides some evidence that unemployment rate is higher in the presence of CATI. This evidence is further supported by the two sample McNemar test results provided Table 4. The individual monthly results in Table 4 provide some evidence of difference in the unemployment flux between two panels. Furthermore, the omnibus test is significant (p-value = 0.00). The mean unemployment flux in the Treatment panel seems to be higher as evidenced by the significant test results of the hypothesis $\underline{1}'\underline{\mu} = 0$.

The two-sample t-tests presented in Thompson (1994) also detected a *positive* difference by panel in mean unemployment rate using the Current Population Survey split panel data

Table 4
Two-Sample McNemar Tests – Unlinked Current Population Survey Data

Time Frame	$(p_2 - p_1) - (p_2' - p_1')$	$se[(p_2 - p_1) - (p_2' - p_1')]$	Z-Statistic	P-Value
10/92 – 11/92	1.18	0.50	2.38	0.02
11/92 – 12/92	0.22	0.50	0.43	0.67
12/92 – 01/93	-0.29	0.45	-0.64	0.52
01/93 – 02/93	0.92	0.45	2.03	0.04
03/93 – 04/93	-0.10	0.42	-0.23	0.81
04/93 – 05/93	0.81	0.45	1.81	0.07
05/93 – 06/93	0.41	0.41	1.01	0.31
06/93 – 07/93	1.16	0.48	2.41	0.02
07/93 – 08/93	0.45	0.42	1.07	0.28
08/93 – 09/93	0.95	0.46	2.06	0.04
09/93 – 10/93	0.14	0.39	0.37	0.71
10/93 – 11/93	1.69	0.44	3.83	0.00
11/93 – 12/93	0.26	0.40	0.66	0.51

i.e., including CATI interviews resulted in a *higher* unemployment rate. These results were consistent with the Current Population Survey CATI Phase-in results presented in Shoemaker (1993). This analysis of Current Population Survey split panel data reinforces that conclusion. Again, it is impossible to attribute the positive net migration from not unemployed to unemployed entirely to the effect of CATI: the same confounding described in 3.4.1 is present in this Treatment (CATI) panel.

3.5 Discussion

Our results appear to yield opposite conclusions about the effect of CATI on unemployment flux. The CATI effect is not, however, the same in both tests.

Perhaps the key difference is the questionnaire. The Parallel Survey data was collected using the newly redesigned Current Population Survey questionnaire. The new questionnaire was designed as an automated instrument. In contrast, the old Current Population Survey questionnaire used for the Current Population Survey CATI Phase-in Project was designed as a paper instrument. Field interviewers were required to memorize complicated skip patterns. To minimize respondent burden, both versions of the Current Population Survey questionnaire are designed for an average interview length of twenty minutes. Using an automated questionnaire, an interviewer can collect more (and more detailed) information in the same amount of time, since she no longer has to determine the path of the interview. Besides the automation difference, the wording of the labor force questions differs between the two questionnaires.

Parallel Survey interviews were conducted using the same questionnaire both in the field interviews (using a laptop computer) and in the CATI facilities. In contrast, the Current Population Survey CATI Phase-in interviews used two different versions of the old questionnaire: a paper version for the field interviews; and an automated version, with a slightly modified lead-in labor force question for the CATI interviews.

Given these questionnaire differences, and the caveats about the Parallel Survey split panel data, we view our results as preliminary. Instead, we recommend pursuing this examination using one and two-sample McNemar techniques on the new Current Population Survey split panel data, which uses the old CATI Phase-in design and the redesigned, fully automated questionnaire.

4. CONCLUSION

We have presented two modifications of the one and two-sample McNemar tests using complex survey data, with applications from the unlinked data modification. If the survey does not have a longitudinal design, then the application using the linked data will have an unknown variance/covariance structure and will include a variance component due to matching error. In this case, using the unlinked data makes sense with respect to the model's interpretation, although the statistic based on the (unlinked) estimates of marginal probabilities may be inferior to a well-developed linked model. If the survey has a longitudinal design, then the first method may be preferred, as it is a straight-forward extension of the traditional test, and consequently, the interpretation is equivalent to the textbook interpretation.

The two-sample McNemar test is not the sole approach one might use in the situation described in section 2.2.2. Another approach to the unlinked form of this problem would be to use a log-linear model for a $2 \times 2 \times 2$ contingency table as in Rao and Scott (1984). In either case, there are trade-offs. The interpretation of the McNemar test is intuitive: it is a cause and effect model, or a repeated measures type of experimental design. The $2 \times 2 \times 2$ contingency table model's interpretation is perhaps less intuitive. Note, however, that the test statistic for the McNemar tests are "Wald-like" statistics, which are often considered to be less efficient than the chi-squared type, *e.g.*, Fay (1985). It is also worth noting that unlike the Rao-Scott formulation, the approach described in this paper makes explicit provisions for the use of linked data.

Areas for future research include investigations into the power of these tests in the context of complex sample data, variance/covariance estimation for linked data including matching error variance contributions, and the difference in efficiency in the two approaches. In data analytical applications, one and two-sample McNemar tests seem to have uses in comparing aspects of different survey methods or effects on responses within a method over time. The approach is nonparametric in its conception; when the approximation is good, it avoids pitfalls that may be associated with model-based tests.

ACKNOWLEDGEMENTS

We thank James Hartman, Alfredo Navarro, James Roebuck, Lynn Weidman, the referee, and the editors for their useful comments. We also thank Sue Chandler for her typing

of this paper. This paper reports the general results of research undertaken by Census Bureau staff. The views expressed are attributable to the authors and do not necessarily reflect those of the Census Bureau.

APPENDIX

For the unlinked data modification of the McNemar Test, $(p_2 - p_1)$ is estimated by $X_{+\cdot}/N_1 + X_{\cdot+}/N_2$ where $X_{+\cdot}$, $X_{\cdot+}$, N_1, and N_2 are weighted estimates, and

$$\widehat{\mathrm{Var}}(p_2 - p_1) = \left[\frac{X_{+\cdot}}{N_1}\right]^2\left[\frac{\mathrm{Var}(X_{+\cdot})}{X_{+\cdot}^2} - \frac{\mathrm{Var}(N_1)}{N_1^2}\right] + \left[\frac{X_{\cdot+}}{N_2}\right]^2\left[\frac{\mathrm{Var}(X_{\cdot+})}{X_{\cdot+}^2} - \frac{\mathrm{Var}(N_2)}{N_2^2}\right] - \frac{2X_{+\cdot}X_{\cdot+}}{N_1 N_2}\left[\frac{\mathrm{Cov}(X_{+\cdot}, X_{\cdot+})}{X_{+\cdot}X_{\cdot+}} - \frac{\mathrm{Var}(N_1)}{N_1^2} - \frac{\mathrm{Var}(N_2)}{N_2^2} + \frac{\mathrm{se}(N_1)\,\mathrm{se}(N_2)}{N_1 N_2}\right].$$

In this appendix we discuss the derivation of the covariance term in the variance estimate, considering only the unlinked data.

Consider the within-panel correlation

$$\mathrm{Cov}(X_{+\cdot}, X_{\cdot+}) = \mathrm{Cov}\left(\sum_{j=1,5} X_{1,j}, \sum_{j=2,6} X_{2,j}\right) \qquad \text{(A1)}$$

where $X_{i,j}$ is a weighted sample level for month i, month-in-sample (MIS) j. Note that $X_{1,j}$ and $X_{2,j+1}$ are from the same rotation group unless $j = 4$ since a rotation group is out of sample for eight months after being in for four.

We assumed that the correlations between $X_{i,j}$ and $X_{k,m}$ can be decomposed into three separate categories:

1) A within-rotation-group correlation,

$$\mathrm{Cov}(X_{i,j}, X_{i+1,j+1}) = r_1, \text{ when } j = 1,2,3,5,6,7.$$

2) A within-month-between-rotation group correlation,

$$\mathrm{Cov}(X_{i,j}, X_{i,k}) = \omega, \; k \neq j, \text{ and}$$

3) A between-rotation-group between-month correlation.

$$\mathrm{Cov}(X_{i,j}, X_{i+1,k}) = \gamma, \; k \neq j+1 \text{ or } j = 3.$$

Replicate estimates of these correlations were available.

The covariance in (A1) becomes

$$\begin{aligned}\mathrm{Cov}(X_{+\cdot}, X_{\cdot+}) &= \mathrm{Cov}(X_{1,1} + X_{1,5}, X_{2,2} + X_{2,6}) \\ &= \mathrm{Cov}(X_{1,1}, X_{2,2}) + \mathrm{Cov}(X_{1,1}, X_{2,6}) + \\ &\quad \mathrm{Cov}(X_{1,5}, X_{2,2}) + \mathrm{Cov}(X_{1,5}, X_{2,6}) \\ &= 2(r_1 + \gamma)\,\mathrm{Var}(X_{i,j}), \end{aligned} \qquad \text{(A2)}$$

using the simplifying assumption that $\mathrm{Var}(X_{i,j})$ is constant for all i and j. The variance for a full month's estimate, $\mathrm{Var}(\sum_{j=1}^{8} X_{i,j})$ is available in the form of a generalized variance function (GVF). We use this estimate to calculate $\mathrm{Var}(X_{i,j})$ by applying the following derivation:

$$\begin{aligned}\mathrm{Var}\left(\sum_{j=1}^{8} X_{i,j}\right) &= \sum_j \sum_k \mathrm{Cov}(X_{i,j}, X_{i,k}) \\ &= \sum_j \mathrm{Var}(X_{i,j}) + \sum_{j \neq k} \mathrm{Cov}(X_{i,j}, X_{i,k}) \\ &= (8 + 56\omega)\,\mathrm{Var}(X_{i,j})\end{aligned}$$

so

$$\mathrm{Var}(X_{i,j}) = (8 + 56\omega)^{-1}\,\mathrm{Var}\left(\sum_{j=1}^{8} X_{i,j}\right). \qquad \text{(A3)}$$

REFERENCES

ADAMS, D.E. (1991). Current population survey month-in-sample (MIS) bias index research. Internal memorandum, Demographic Statistical Methods Division, U.S. Bureau of the Census, Washington, DC.

DONNER, A., and LI, K.Y.R. (1990). The relationship between chi-square statistics from matched and unmatched analyses. *Journal of Clinical Epidemiology*, 43, 827-831.

FAY, R. (1990). VPLX: Variance estimates for complex samples. *Proceedings of the Section on Survey Research Methods, American Statistical Association.*

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

FEUER, E.J., and KESSLER, L.G. (1989). Test statistic and sample size for a two-sample McNemar test. *Biometrics*, 45, 629-636.

FISHER, R., ROBISON, E., THOMPSON, J., and WELCH, M. (1993). Variance estimation in the current population survey overlap test. *Proceedings of the Section on Survey Research Methods, American Statistical Association.*

FISHER, R., and McGUINNESS, R. (1993). Correlations and adjustment factors for current population survey. Internal memorandum, Demographic Statistical Methods Division, U.S. Bureau of the Census, Washington, DC.

HOGUE, C. (1984). History of the problems encountered estimating gross flows. *Proceedings of the Conference on Gross Flows in Labor Force Statistics*, 1-7.

JABINE, T.B., and SCHEUREN, F.J. (1986). Record linkages for statistical purposes: Methodological Issues. *Journal of Official Statistics*, 2, 3, 255-277.

McNEMAR, Q. (1947). Note on the sampling error of the differences between correlated proportions of percentages. *Psychometrika*, 12, 153-157.

MARASCUILO, L.A., OMELICH, C.L., and GOKHALE, D.V. (1988). Planned and post hoc methods for multiple-sample McNemar (1947) tests with missing data. *Psychological Bulletin*, 103, 238-245.

RAO, J.N.K., and WU, C.F.J. (1988). Resampling inference with complex survey data. *Journal of the American Statistical Association*, 83, 231-241.

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.

SHOEMAKER, H.H. (1993). Results from the current population survey CATI phase-in project. *Proceedings of the Section on Survey Research Methods, American Statistical Association.*

STASNY, E.A., and FIENBERG, S.E. (1984). Some stochastic models for estimating gross flows in the presence of nonrandom nonresponse. *Proceedings of the Conference on Gross Flows in Labor Force Statistics*, 25-39.

THOMPSON, J. (1994). Mode effects analysis of labor force estimates. Current Population Survey Overlap Analysis Team Technical Report 3, U.S. Bureau of the Census, Washington, DC.

THOMPSON, J., and FISHER, R. (1994). Two sample McNemar test for complex surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association.*

Survey Methodology, December 1996
Vol. 22, No. 2, pp. 157-165
Statistics Canada

Stability Measures for Variance Component Estimators Under a Stratified Multistage Design

J.L. ELTINGE and D.S. JANG[1]

ABSTRACT

In work with sample surveys, we often use estimators of the variance components associated with sampling within and between primary sample units. For these applications, it can be important to have some indication of whether the variance component estimators are stable, *i.e.*, have relatively low variance. This paper discusses several data-based measures of the stability of design-based variance component estimators and related quantities. The development emphasizes methods that can be applied to surveys with moderate or large numbers of strata and small numbers of primary sample units per stratum. We direct principal attention toward the design variance of a within-PSU variance estimator, and two related degrees-of-freedom terms. A simulation-based method allows one to assess whether an observed stability measure is consistent with standard assumptions regarding variance estimator stability. We also develop two sets of stability measures for design-based estimators of between-PSU variance components and the ratio of the overall variance to the within-PSU variance. The proposed methods are applied to interview and examination data from the U.S. Third National Health and Nutrition Examination Survey (NHANES III). These results indicate that the true stability properties may vary substantially across variables. In addition, for some variables, within-PSU variance estimators appear to be considerably less stable than one would anticipate from a simple count of secondary units within each stratum.

KEY WORDS: Between-PSU variance; Complex sample design; Degrees of freedom; Diagnostic; Design-based analysis; Satterthwaite approximation; Stratum collapse; U.S. Third National Health and Nutrition Examination Survey (NHANES III); Within-PSU variance.

1. INTRODUCTION

In work with sample surveys, it is often desirable to have good estimates of the variance components attributable to sampling within and between primary sample units (PSUs). For example, the magnitude of an estimated within-PSU variance, relative to a between-PSU variance, may influence decisions regarding sample allocation and related design issues (*e.g.*, Hansen *et al.* 1953, Chapter 7). Similar relative-magnitude properties affect the bias of certain variance estimators derived under simplifying assumptions regarding the sample design (*e.g.*, Korn and Graubard 1995, p. 278-279, 287; and Wolter 1985, p. 44-46). Also, some survey analysts have a general interest in identification of surveys and variables for which the between-PSU component of variance is substantially greater than zero. See, *e.g.*, Herzog and Scheuren (1976, p. 398) and Wolter (1985, p. 47) for related comments. In addition, Jang and Eltinge (1996) give an example for which there is some interest in the within-PSU variances by themselves.

In some application work, estimates of within-PSU variances and related quantities are reported with the apparent assumption that the estimates are stable, *i.e.*, have relatively low variances. This paper shows that it can be important to carry out data-based checks of this assumption of stability, and that some relatively simple checking methods follow from standard design-based ideas. We emphasize methods that can be applied to designs with a moderate or large number of strata and a small number of PSUs selected per stratum.

Subsection 2.1 reviews the relevant estimators of within-PSU variances and overall stratum-level variances. Subsection 2.2 identifies two distinct components of the variance of the within-PSU variance estimator. Subsection 2.3 presents simple design-based estimators of the variances of two within-PSU variance estimators. Section 3 develops two related degrees-of-freedom measures.

Section 4 examines the extent to which related design-based methods can be used to assess the stability of quantities that depend both on the within-PSU variance estimator and on the overall stratum-level variance estimator. Principal attention is directed toward an estimator of the between-PSU variance and an estimator of the ratio of the overall stratum-level variance divided by the within-PSU variance. Section 4.2 discusses one set of methods based on the stability measures from Section 2 and some moderately restrictive moment assumptions. Section 4.3 outlines a second set of methods based on stratum collapse.

Section 5 applies the main ideas of Sections 2 through 4 to variance estimates computed for the U.S. Third National Health and Nutrition Examination Survey. Section 5 also uses a simple simulation-based method to assess the consistency of the observed measures with standard assumptions regarding variance estimator stability. The Section 5 results suggest that the true stability of within-PSU variance estimators can be substantially less than anticipated from a simple count of the number of secondary units contributing to each PSU. In addition, the results indicate that the stability properties of

[1] J.L. Eltinge and D.S. Jang, Department of Statistics, Texas A&M University, College Station, TX 77843-3143, U.S.A.

within-PSU variance estimators and related quantities can vary substantially across different variables collected in the same survey. Section 6 gives additional comments on the methods and empirical results presented here.

2. WITHIN-PSU AND OVERALL STRATUM-LEVEL VARIANCE ESTIMATORS

2.1 General Notation

In principle, we could use either design-based or model-based methods to examine within-PSU and between-PSU variance components. The present work will take a design-based approach. This is consistent with some related previous literature, *e.g.*, Wolter (1985, p. 40-41, 47). The design-based approach will be especially useful in highlighting some strengths and limitations of the proposed stability-assessment methods. For example, in Section 2.3 this approach will give us some indication of specific design features that may affect variance estimator stability. Also, in Section 4 the design-based approach will help to clarify the extent to which certain moment restrictions are needed to justify one set of stability measures.

Following the notation and ideas in Wolter (1985, p. 43-47), consider a stratified multistage sample design with L strata and with N_h primary sampling units (PSUs) contained in stratum $h = 1, 2, ..., L$. We select n_h PSUs with replacement and with per-draw selection probabilities p_{hi}. Within selected PSU (h,i), we select n_{hi} secondary sample units (SSUs) with replacement and with per-draw selection probabilities p_{hij}. Further subsampling is carried out within a selected SSU to obtain n_{hij} individual elements for interview or examination. The stability-assessment methods developed here are intended primarily for designs with moderate or large L, relatively small n_h (*e.g.*, $n_h = 2$), and relatively large n_{hi}. Designs with these characteristics are often used in large household interview surveys, *e.g.*, the health survey discussed in Section 4.

We will focus on estimation of a population total $Y = \sum_{h=1}^{L} Y_h$, where $Y_h = \sum_{i=1}^{N_h} Y_{hi}$, $Y_{hi} = \sum_{j=1}^{N_{hi}} \sum_{k=1}^{N_{hij}} Y_{hijk}$, Y_{hijk} is a survey item for element k in SSU j in PSU i in stratum h, N_{hi} is the number of SSUs in PSU (h,i), and N_{hij} is the number of elements in SSU (h, i, j). Extensions to nonlinear functions of population totals are straightforward and will be considered in the applications in Section 5. A standard design-based estimator of Y is $\hat{Y} = \sum_{h=1}^{L} \hat{Y}_h$ where

$$\hat{Y}_h = \sum_{i=1}^{n_h} \sum_{j=1}^{n_{hi}} \sum_{k=1}^{n_{hij}} w_{hijk} y_{hijk}, \qquad (2.1)$$

w_{hijk} is the customary weight derived from selection probabilities and sample sizes to ensure unbiased estimation of each Y_h, and the lower-case terms y_{hijk} denote sample observations. In subsequent work, it will be useful to rewrite expression (2.1) as

$$\hat{Y}_h = n_h^{-1} \sum_{i=1}^{n_h} p_{hi}^{-1} \hat{Y}_{hi},$$

where $\hat{Y}_{hi} = n_{hi}^{-1} \sum_{j=1}^{n_{hi}} z_{hij}$ and $z_{hij} = n_h n_{hi} p_{hi} \sum_{k=1}^{n_{hij}} w_{hijk} y_{hijk}$.

2.2 Within- and Between-PSU Variances

Throughout this discussion, expectations and variances will be defined with respect to the sample design. Under the conditions stated above, the variance of $\hat{Y}$ is $V(\hat{Y}) = \sum_{h=1}^{L} V_h$, where $V_h = V_{Bh} + V_{Wh}$, $V_{Bh} = V(n_h^{-1} \sum_{i=1}^{n_h} p_{hi}^{-1} Y_{hi})$, $V_{Wh} = n_h^{-1} \sum_{i=1}^{N_h} p_{hi}^{-1} \sigma_{2hi}^2$, and $\sigma_{2hi}^2 = V(\hat{Y}_{hi} - Y_{hi} \mid h,i)$; see, *e.g.*, Wolter (1985, p. 42). Note especially that Y_{hi} is the true population total for selected PSU (h, i), and that σ_{2hi}^2 reflects the variability in $\hat{Y}_{hi} - Y_{hi}$ attributable to subsampling at the SSU and finer levels.

A customary unbiased estimator of the overall stratum-level variance V_h is

$$\hat{V}(\hat{Y}_h) = n_h^{-1} (n_h - 1)^{-1} \sum_{i=1}^{n_h} (p_{hi}^{-1} \hat{Y}_{hi} - \hat{Y}_h)^2,$$

and the corresponding estimator of $V(\hat{Y}) = \sum_{h=1}^{L} V(\hat{Y}_h)$ is $\hat{V}(\hat{Y}) = \sum_{h=1}^{L} \hat{V}(\hat{Y}_h)$.

Now consider estimation of the within-PSU variance V_{Wh}. Since $\hat{Y}_{hi}$ is a sample mean of the independent and identically distributed terms z_{hij}, standard arguments show that for a given PSU (h,i), an unbiased estimator of σ_{2hi}^2 is $\hat{\sigma}_{2hi}^2 = n_{hi}^{-1} (n_{hi} - 1)^{-1} \sum_{j=1}^{n_{hi}} (z_{hij} - \hat{Y}_{hi})^2$. Thus, an unbiased estimator of V_{Wh} is

$$\hat{V}_{Wh} = \sum_{i=1}^{n_h} n_h^{-2} p_{hi}^{-2} \hat{\sigma}_{2hi}^2 = \sum_{i=1}^{n_h} n_{hi}^{-1} (n_{hi} - 1)^{-1} \sum_{j=1}^{n_{hi}} (x_{hij} - \bar{x}_{hi})^2,$$

where $x_{hij} = n_{hi} \sum_{k=1}^{n_{hij}} w_{hijk} y_{hijk}$ and $\bar{x}_{hi} = n_{hi}^{-1} \sum_{j=1}^{n_{hi}} x_{hij}$. Note that the latter expression for $\hat{V}_{Wh}$ uses only sample sizes, the observations y_{hijk} and the customary weights w_{hijk}.

2.3 The Variance of $\hat{V}_{Wh}$

A direct modification of standard conditional-moment arguments shows that the variance of $\hat{V}_{Wh}$ is $\gamma_{Bh} + \gamma_{Wh}$, where

$$\gamma_{Bh} = V(n_h^{-2} \sum_{i=1}^{n_h} p_{hi}^{-2} \sigma_{2hi}^2)$$

and

$$\gamma_{Wh} = n_h^{-3} \sum_{i=1}^{N_h} p_{hi}^{-3} V(\hat{\sigma}_{2hi}^2 \mid h,i).$$

Thus, the variance of $\hat{V}_{Wh}$ itself depends on a sum of between- and within-PSU variances, and the relative magnitudes of γ_{Bh} and γ_{Wh} depend on trade-offs among σ_{2hi}^2, p_{hi} and n_{hi}. For example, under regularity conditions, the terms $V(\hat{\sigma}_{2hi}^2 \mid h,i)$ are approximately inversely proportional to n_{hi}. Thus, if the n_{hi} are uniformly large within stratum h, then γ_{Wh}

may be relatively small. Also, if the terms $p_{hi}^{-2}\sigma_{2hi}^2$ are approximately constant within a given stratum, then γ_{Bh} may be relatively small. Conversely, marked heterogeneity of $p_{hi}^{-2}\sigma_{2hi}^2$ may inflate γ_{Bh} and thus inflate $V(\hat{V}_{Wh})$ as well.

In addition, note that under the stated design conditions, $\hat{V}_{Wh}$ is the sample mean of the independent and identically distributed terms $n_h^{-1}p_{hi}^{-2}\delta_{2hi}^2$. Thus, an unbiased estimator of the variance of $\hat{V}_{Wh}$ is

$$\tilde{V}(\hat{V}_{Wh}) = n_h^{-1}(n_h - 1)^{-1}\sum_{i=1}^{n_h}(n_h^{-1}p_{hi}^{-2}\delta_{2hi}^2 - \hat{V}_{Wh})^2. \quad (2.2)$$

Some applications focus on the full-population level, rather than on individual strata, and so the "within-PSU" contribution of interest is the sum of the within-PSU variances, $V_W = \sum_{h=1}^{L} V_{Wh}$. Under the conditions given above, an unbiased estimator of V_W is $\hat{V}_W = \sum_{h=1}^{L}\hat{V}_{Wh}$. Also, since our sampling and subsampling are independent across strata, we have $V(\hat{V}_W) = \sum_{h=1}^{L}(\gamma_{Bh} + \gamma_{Wh})$, and an unbiased estimator of $V(\hat{V}_W)$ is

$$\tilde{V}(\hat{V}_W) = \sum_{h=1}^{L}\tilde{V}(\hat{V}_{Wh}).$$

Finally, note that the preceding development used the assumption of sampling with replacement at both the primary- and secondary-unit levels. Two applications of result (2.4.16) in Wolter (1985, p. 46) show that under mild conditions that hold for many, but not all, without-replacement designs, $\hat{V}_{Wh}$ will be unbiased or conservative for the true within-PSU variance; and $\tilde{V}(\hat{V}_{Wh})$ will be unbiased or conservative for the true variance of $\hat{V}_W$. A formal technical statement and proof of this result is available from the authors.

2.4 Balanced Interpretation of Stability Measures

The remainder of this paper uses $\tilde{V}(\hat{V}_{Wh})$ and related quantities to assess the stability of variance-component estimators. In working with these results, it is useful to remember that data-based measures of variance estimator stability are justifiably viewed with some caution, because they are functions of fourth sample moments, and thus are themselves subject to a considerable amount of sampling variability. See, *e.g.*, Fuller (1984, p. 111). This caution carries over to the proposed estimator $\tilde{V}(\hat{V}_{Wh})$ and to the related statistics discussed in Sections 3 and 4 below.

However, one should not overstate this caution to the point of making no attempt at data-based assessment of variance estimator stability. The estimator $\tilde{V}(\hat{V}_{Wh})$, and the related measures in Sections 3 and 4, are relatively simple to compute, and provide diagnostics that can help to identify variables for which:

(a) the instability of $\hat{V}_{Wh}$ is especially problematic; or

(b) the instability of $\hat{V}_{Wh}$ has a substantial effect on the precision of estimators of the relative magnitudes of between-PSU and within-PSU variances.

Consequently, interpretation of specific values of $\tilde{V}(\hat{V}_{Wh})$ and related stability measures should reflect a balance between the abovementioned general caution and a recognition of their potential diagnostic value.

3. TWO STABILITY MEASURES FOR WITHIN-PSU VARIANCE ESTIMATORS

3.1 Degrees-of-Freedom Diagnostics for Variance Estimator Stability

Some analysts prefer to express variance estimator stability through "degrees of freedom" measures related to the Satterthwaite (1941, 1946) approximation. To introduce this idea, consider a general variance estimator $\hat{V}$, and note that $\{E(\hat{V})\}^{-1}d\hat{V}$ has the same first and second moments as a chi-square random variable on d degrees of freedom, where d is the solution to the equation,

$$2\{E(\hat{V})\}^2 - V(\hat{V})d = 0.$$

If the distribution of $\{E(\hat{V})\}^{-1}d\hat{V}$ is indeed well approximated by a chi-square distribution, then d may be viewed fairly literally as a "degrees of freedom" term. Otherwise, d can be viewed as twice the inverse of the squared coefficient of variation of $\hat{V}$. In either case, d has a certain appeal because it is scale-free, and can be tied fairly directly to notions of "effective sample size" in the evaluation of variance estimator performance. Subsection 3.3 gives related comments for two special cases.

Given an unbiased estimator $\tilde{V}(\hat{V})$ of the variance of $\hat{V}$, one may compute a "degrees of freedom" estimator $\hat{d}$ as the solution to the unbiased estimating equation

$$2\{\hat{V}^2 - \tilde{V}(\hat{V})\} - \tilde{V}(\hat{V})d = 0, \quad (3.1)$$

i.e., $\hat{d} = \{\tilde{V}(\hat{V})\}^{-1}2\hat{V}^2 - 2$. Under mild regularity conditions, $d^{-1}\hat{d}$ converges in probability to one, provided $\{V(\hat{V})\}^{-1}\tilde{V}(\hat{V})$ and $\{E(\hat{V})\}^{-1}\hat{V}$ both converge in probability to one.

3.2 Degrees-of-Freedom Diagnostics for Pooled and Stratum-Level Estimators of Within-PSU Variances

We can apply these general degrees-of-freedom ideas to the within-PSU variance estimators $\hat{V}_{Wh}$ and $\hat{V}_W$ developed in Section 2. First consider the case in which there is intrinsic interest in the stability of individual stratum-level estimators $\hat{V}_{Wh}$. The associated "degrees of freedom" measure is $d_{Wh} = \{V(\hat{V}_{Wh})\}^{-1}2V_{Wh}^2$. For designs with large n_h, one may use (3.1) to compute estimators $\hat{d}_{Wh} = \{\tilde{V}(\hat{V}_{Wh})\}^{-1}2\hat{V}_{Wh}^2 - 2$ separately for each stratum. For designs with small n_h (*e.g.*, $n_h = 2$ for each stratum), the estimator $\hat{d}_{Wh}$ itself may be very unstable.

Consequently, it also is useful to consider the alternative combined estimator

$$\hat{d}_{W0} = \left\{\sum_{h=1}^{L} \tilde{V}(\hat{V}_{Wh})\right\}^{-1} 2\sum_{h=1}^{L} \hat{V}_{Wh}^2 - 2,$$

under the assumption that all d_{Wh} equal a common value d_{W0}.

Now consider the pooled within-PSU variance estimator $\hat{V}_W$ developed in Section 2.3. The resulting "degrees of freedom" measure is $d_{WF} = \{\sum_{h=1}^{L} V(\hat{V}_{Wh})\}^{-1} 2V_W^2$, and expression (3.1) suggests the estimator

$$\hat{d}_{WF} = \left\{\sum_{h=1}^{L} \tilde{V}(\hat{V}_{Wh})\right\}^{-1} 2\hat{V}_W^2 - 2.$$

3.3 Comparison of d_{W0} and d_{WF} to Direct SSU Counts

To interpret $\hat{d}_{W0}$ and $\hat{d}_{WF}$ as stability measures, consider the following idealized setting. Assume that for all h, the PSU counts n_h are equal to a common value n_1, say; and that for all h and i, the SSU counts n_{hi} are equal to a common value n_{11}. In addition, assume that the terms $p_{hi}^{-2}\sigma_{2hi}^2$ are constant within each stratum; and that, conditional on (h, i), each $\sigma_{2hi}^{-2}(n_{11} - 1)\,\hat{\sigma}_{2hi}^2$ is distributed as a chi-square random variable on $n_{11} - 1$ degrees of freedom. Then routine arguments show that $d_{W0} = n_1(n_{11} - 1)$. If the preceding assumptions are satisfied approximately, and if the product $n_1(n_{11} - 1)$ is large (greater than 40, say), then a data user may be inclined to view $\hat{V}_{Wh}$ as relatively stable, or equivalently, to view the errors $\hat{V}_{Wh} - V_{Wh}$ as negligible. This appears to be the reasoning used implicitly when estimates $\hat{V}_{Wh}$ are treated as known values in design or analysis work. However, the application in Section 5 will give some examples for which this reasoning is problematic, so that evaluation of the estimates $\hat{d}_{W0}$ is important.

Also, under the idealized conditions described above, and under the additional assumption that the V_{Wh} are all equal, we have $d_{WF} = Ln_1(n_{11} - 1)$.

4. COMPARISON OF WITHIN-PSU AND OVERALL STRATUM-LEVEL VARIANCES

4.1 Estimators of Between-PSU Variances and Related Variance Ratios

Section 1 cited some applications that hinge on the magnitude of V_{Wh} relative to V_h. The specifics of the relative-magnitude comparisons vary with the individual application, but interest generally focuses on differences or ratios. For example, recall that $V_{Bh} = V_h - V_{Wh}$ and define the overall between-PSU variance term $V_B = \sum_{h=1}^{L} V_{Bh}$. In addition, note that unbiased estimators of V_{Bh} and V_B are $\hat{V}_{Bh} = \hat{V}_h - \hat{V}_{Wh}$ and $\hat{V}_B = \sum_{h=1}^{L} \hat{V}_{Bh}$ respectively.

Similarly, define the ratio $R_{WV} = V_W^{-1}V(\hat{Y})$, the magnitude of the overall variance $V(\hat{Y})$ relative to the within-PSU contribution V_W. A direct estimator of R_{WV} is $\hat{R}_{WV} = \hat{V}_W^{-1}\hat{V}(\hat{Y})$. Note that if $V_{Wh}^{-1}V_h = R_{WV}$ for all h, then $\hat{R}_{WV}$ could also be viewed as a pooled estimator of this common *stratum*-level ratio.

For both $\hat{V}_B$ and $\hat{R}_{WV}$, stability assessment involves the variance of $\hat{V}_h$ and the covariance of $\hat{V}_{Wh}$ with $\hat{V}_h$. Estimation of the these moments can be somewhat problematic for surveys that select small numbers of PSUs from each stratum. We consider two approaches to resolving this problem. Section 4.2 uses moderate restrictions on the moment structure of $(\hat{V}_{Wh}, \hat{V}_h)$ to develop estimators $V(\hat{V}_h)$ and related quantities. Section 4.3 uses stratum collapse to develop alternative stability measures.

4.2 Stability Measures Based on $\tilde{V}(\hat{V}_{Wh})$ and Moment Conditions

4.2.1 Moment Conditions

Under moderate moment restrictions, we can estimate the variance of $\hat{V}_h$ directly from $\hat{V}_h$ itself. Specifically, assume that the variance of $\hat{V}_h$ equals $(n_h - 1)^{-1}2V_h^2$; this would hold, *e.g.*, under the standard assumption that $V_h^{-1}(n_h - 1)\hat{V}_h$ is distributed as a chi-square random variable on $n_h - 1$ degrees of freedom. As in Sections 2 and 3, we continue to assume that $\hat{V}_h$ is unbiased for V_h. Then routine moment arguments show that $(n_h + 1)^{-1}2\hat{V}_h^2$ is an unbiased estimator of the variance of $\hat{V}_h$.

In the remainder of Section 4.2, we will also assume that $\mathrm{Cov}(\hat{V}_{Wh}, \hat{V}_h) = 0$. Routine conditional-moment arguments show that this will hold if the terms $p_{hi}^{-2}\sigma_{2hi}^2$ are equal within a given stratum; and if, conditional on (h,i,j), the SSU-level estimates x_{hij} are normally distributed, so that $\hat{\sigma}_{2hi}^2$ is conditionally independent of $\hat{Y}_{hi}$.

4.2.2 Stability Measures

Under the conditions stated in Section 4.2.1, unbiased estimators of $V(\hat{V}_{Bh})$ and $V(\hat{V}_B)$ are

$$\tilde{V}(\hat{V}_{Bh}) = (n_h + 1)^{-1}2\hat{V}_h^2 + \tilde{V}(\hat{V}_{Wh}) \qquad (4.1)$$

and $\tilde{V}(\hat{V}_B) = \sum_{h=1}^{L} \tilde{V}(\hat{V}_{Bh})$, where $\tilde{V}(\hat{V}_{Wh})$ is defined in expression (2.2). Also, under the same conditions routine ratio-estimation arguments lead to the variance estimator

$$\tilde{V}(\hat{R}_{WV}) = \hat{V}_W^{-2}\sum_{h=1}^{L}\left\{(n_h + 1)^{-1}2\hat{V}_h^2 + \hat{R}_{WV}^2\,\tilde{V}(\hat{V}_{Wh})\right\}. \qquad (4.2)$$

4.3 Alternative Stability Measures Based on Stratum Collapse

The assumptions of Section 4.2.1 may be problematic in some applications. For example, for some survey designs and variables, the SSU-level estimators x_{hij} may have markedly nonnormal distributions, so the assumption $\mathrm{Cov}(\hat{V}_{Wh}, \hat{V}_h) = 0$ may not hold. For these cases, one may consider the use of stratum collapse to produce alternative estimators of $V(\hat{V}_B)$ and $V(\hat{R}_{WV})$.

Specifically, partition the set of L strata into G prespecified groups, with L_g strata contained in group S_g, $g = 1, ..., G$. With this new notation, note that

$$(\hat{V}(\hat{Y}),\hat{V}_W,\hat{V}_B) = \sum_{g=1}^{G} \sum_{h\in S_g} (\hat{V}_h,\hat{V}_{Wh},\hat{V}_{Bh}).$$

Standard stratum-collapse methods (*e.g.*, Wolter 1985, Section 2.5) then lead to the alternative variance estimator,

$$V_{cs}^*(\hat{V}_B) = \sum_{g=1}^{G} (L_g - 1)^{-1} L_g \sum_{h\in S_g} D_{gh}^2,$$

where $D_{gh} = \hat{V}_{Bh} - L_g^{-1}\sum_{j\in S_g}\hat{V}_{Bj}$. Similarly, a collapsed-stratum variance estimator for $\hat{R}_{WV}$ is,

$$V_{cs}^*(\hat{R}_{WV}) = (\hat{V}_W)^{-2}\sum_{g=1}^{G} (L_g - 1)^{-1} L_g \sum_{h\in S_g} C_{gh}^2,$$

where $C_{gh} = (\hat{V}_h - \hat{R}_{WV}\hat{V}_{Wh}) - L_g^{-1}\sum_{j\in S_g}(\hat{V}_j - \hat{R}_{WV}\hat{V}_{Wj})$.

In general, collapsed-stratum variance estimators require some care in interpretation; see, *e.g.*, Rust and Kalton (1985), Wolter (1985, Section 2.5) and references cited therein. For example, collapsed-stratum variance estimators generally will be conservative. In addition, for cases with moderate L, the variance estimators $V_{cs}^*(\hat{V}_B)$ and $V_{cs}^*(\hat{R}_{WV})$ may themselves have limited stability.

5. APPLICATION TO THE U.S. THIRD NATIONAL HEALTH AND NUTRITION EXAMINATION SURVEY

5.1 Sample Design and Estimation Methods

The methods proposed in Sections 2 through 4 were applied to data from Phase I of the Third National Health and Nutrition Examination Survey (NHANES III). National Center for Health Statistics (1996) gives a general description of this survey, including special characteristics associated with Phase I (data collected between 1988 and 1991). For the present discussion, three aspects are of special interest. First, variance estimators were constructed on the basis of a collapsed design involving $L = 22$ strata (large groups of counties), with two primary sample units (generally individual counties) selected per stratum. Second, each selected PSU had a relatively large number of selected SSUs (generally groups of city blocks, or similar rural areas). The number of selected SSUs within each stratum ranged from 30 to 63, with a mean of 45.8.

Third, additional subsampling within each SSU led to selection of the survey elements (individual noninstitutionalized U.S. civilians). Each selected person was asked to respond to a health questionnaire and to participate in a detailed medical examination. Twelve of the resulting variables are listed in Table 1.

Standard weighted ratio estimates $\hat{\theta}$ were computed for the population means of each of the twelve variables listed in Table 1. The first two columns of Table 2 present the corresponding variance estimates $\hat{V}(\hat{\theta})$ and $\hat{V}_W$. As part of a larger study of the within-PSU variances V_{Wh} discussed in Jang and Eltinge (1996), there was considerable interest in the stability of the individual estimates $\hat{V}_{Wh}$. Since we had $n_h = 2$ for each stratum, the reasoning in Section 3.2 indicated that it was not feasible to examine the individual terms $\hat{d}_{Wh}$. Consequently, Section 5.2 will examine the pooled measure $\hat{d}_{W0}$ of the stability of the $\hat{V}_{Wh}$ and will also present some related simulation-based tests and diagnostic plots.

Table 1
Twelve NHANES III Variables

Variable name	Description
HAE2	Told by health professional that you had hypertension (indicator variable)
HAE7	Told by health professional that your blood cholesterol was high (indicator variable)
HAD1	Told by health professional that you had diabetes (indicator variable)
HAR3	Do you smoke cigarettes now?
BMPHT	Height
BMPWT	Weight
HDRESULT	HDL cholesterol
TCRESULT	Serum total cholesterol
LEAD	Blood lead, in micrograms per deciliter
log(LEAD)	Natural logarithm of blood lead
BP1K1	Systolic blood pressure
BP1K5	Diastolic blood pressure

Table 2
Variance Estimates and Stability Measures for Twelve NHANES III Variables

Variable name	$\hat{V}_W$	$\hat{V}(\hat{Y})$	$\hat{d}_{W0}$	$\hat{d}_{WF}$
HAE2	0.0000385	0.0000511	23.7	425.8
HAE7	0.0000821	0.000135	13.6	225.6
HAD1	0.00000956	0.00000749	8.8	160.6
HAR3	0.000122	0.000205	6.4	125.8
BMPHT	0.0223	0.0416	15.3	275.1
BMPWT	0.104	0.122	8.6	139.2
HDRESULT	0.0743	0.163	11.5	196.2
TCRESULT	0.590	0.860	21.2	353.9
LEAD	0.00388	0.00657	2.8	48.8
log(LEAD)	0.000211	0.000678	10.5	174.9
BP1K1	1.073	2.896	1.0	26.5
BP1K5	0.252	0.217	17.2	52.9

In addition, there was interest in the extent to which the variances of the $\hat{V}_{Wh}$ contributed to the variances of the pooled quantities $\hat{V}_B$ and $\hat{R}_{WV}$. Section 5.3 explores this question.

5.2 Within-PSU Variance Estimates and Associated Stability Measures

5.2.1 Comparison Across Variables

The final two columns of Table 2 report the degrees-of-freedom estimates $\hat{d}_{W0}$ and $\hat{d}_{WF}$ for the twelve NHANES III variables. Note especially that the stratum-level stability measures $\hat{d}_{W0}$ are relatively low, compared to the mean of 45.8 SSUs per stratum. For example, all of the variables have $\hat{d}_{W0}$ less than 24, and five (HAD1, HAR3, BMPWT, LEAD and BP1K1) have $\hat{d}_{W0}$ less than 10. Due to the interest in the $\hat{d}_{W0}$ described above, this led to two general questions.

(1) Are the observed $\hat{d}_{W0}$ consistent with the nominal degrees-of-freedom value d_{W0} that one would anticipate from the direct SSU counts $n_{h1} + n_{h2} - 2$?

(2) Conversely, are the observed $\hat{d}_{W0}$ consistent with distributional conditions that produce considerably smaller values of d_{W0}?

Standard large-sample-theory-based tests for (1) and (2) would have depended on eighth sample moments, and thus were inadvisable in the present case, due to the relatively small values of $L = 22$ and $n_h = 2$. Instead, the following simulation-based test was carried out.

5.2.2 Simulation-Based Interpretation of Stability Measures

This simulation work covers six cases involving different values of two terms. The first term, denoted d_{hi}, represents the degrees of freedom associated with the variance estimator $\hat{\sigma}^2_{2hi}$ in PSU (h, i). The second term, denoted R_{12}, is the ratio of the expressions $p_{hi}^{-2}\sigma^2_{2hi}$ in the first and second sample PSUs in stratum h.

In each of the six cases discussed below, independent pseudorandom variables g_{hi} were generated from a chi-square distribution on d_{hi} degrees of freedom for $h = 1, 2, ..., 22$ and $i = 1, 2$. Re-scaled variables $\hat{V}_{Whi} = d_{hi}^{-1} V_{Whi} g_{hi}$ were then computed, where V_{Whi} is a random variable equal to one with probability one-half and equal to R_{12} with probability one-half. The random variables g_{hi} and V_{Whi} are mutually independent. Finally, the sums $\hat{V}_{Wh} = \hat{V}_{Wh1} + \hat{V}_{Wh2}$ and the associated measures $\bar{V}(\hat{V}_{Wh})$, $\bar{V}(\hat{V}_W)$ and $\hat{d}_{W0}$ were computed. This was repeated 10,000 times.

Table 3 lists the values of d_{hi} and R_{12} covered in the six cases, and Table 4 lists the resulting simulated means, standard deviations and quantiles for $\hat{d}_{W0}$. When interpreting the results for these cases, note that randomness of the g_{hi} corresponds to the estimation error in the $\hat{\sigma}^2_{2hi}$ due to subsampling at the SSU and lower levels; and randomness of the V_{Whi} reflects the variability of the $p_{hi}^{-2}\sigma^2_{2hi}$ induced by sampling of PSUs within a given stratum.

Table 3
Cases Covered for the Simulated Quantiles

Cases	d	R_{12}
1	22	1
2	Obs. Dist.	1
3	5	1
4	22	9
5	Obs. Dist.	9
6	5	9

Case 1 uses $d_{hi} = 22$ and $R_{12} = 1$. Arguments from Section 3.3 show that the resulting $\hat{V}_{Wh}$ are distributed as constant multiples of a chi-square random variable with $d_{W0} = 44$ degrees of freedom. Thus, for Case 1, the choice of $d_{hi} = 22$ has led to simulated quantiles of $\hat{d}_{W0}$ that are approximately those that one would anticipate from the mean SSU count of 45.8 observed for Phase I of NHANES III, under the setting described in Section 3.4. Note that even in this idealized Case 1, the relative variability of the $\hat{d}_{W0}$ is fairly high.

Now compare the $\hat{d}_{W0}$ reported in Table 2 to the simulated quantiles from Case 1. All twelve of the observed $\hat{d}_{W0}$ fall below the 0.025 simulated quantile of 24.8; and ten of the twelve fall below the 0.005 quantile of 21.1. Thus, the $\hat{d}_{W0}$ observed for the NHANES III variables are not consistent with a nominal $d_{W0} = 44$ produced in the idealized setting covered by Case 1.

5.2.3 Simulation Under Alternative Conditions with Smaller d_{W0}

In general, the distribution of $\hat{d}_{W0}$ may deviate from that observed under the idealized Case 1 due to: (a) variability in the true SSU counts n_{hi}; (b) limited stability of the PSU-level estimates $\hat{\sigma}^2_{2hi}$; and (c) heterogeneity of the true PSU-level terms σ^2_{2hi}. Cases 2 through 6 cover the combined effects of these three factors.

Table 4
Simulated Quantiles for $\hat{d}_{W0}$

Cases	Mean	S.D.	$q_{.005}$	$q_{.01}$	$q_{.025}$	$q_{.05}$	$q_{.10}$	$q_{.25}$	$q_{.50}$	$q_{.75}$	$q_{.90}$	$q_{.95}$	$q_{.975}$	$q_{.99}$	$q_{.995}$
1	48.9	17.7	21.1	22.5	24.8	27.4	30.7	36.7	45.5	57.4	71.2	81.5	92.6	108.5	122.1
2	48.3	17.5	20.7	21.9	24.2	26.8	29.9	36.3	45.2	56.6	70.2	80.3	92.0	106.2	118.0
3	11.3	4.7	4.1	4.5	5.1	5.6	6.4	8.0	10.3	13.5	17.3	20.0	23.0	26.8	30.1
4	5.5	2.7	1.4	1.6	2.0	2.3	2.7	3.7	5.0	6.8	8.9	10.5	12.1	14.8	16.7
5	5.5	2.7	1.4	1.6	1.9	2.3	2.7	3.7	5.0	6.7	8.9	10.6	12.1	14.1	16.1
6	3.5	2.1	0.7	0.8	1.0	1.2	1.5	2.1	3.0	4.4	6.0	7.4	8.8	11.2	12.6

The design for Case 2 was identical to that for Case 1, except that the d_{hi} were random variables, selected with equal probabilities and with replacement from the 44 values $n_{hi} - 1$ corresponding to the 44 SSU counts n_{hi} in the original dataset. The resulting simulated quantiles of $\hat{d}_{W0}$ are similar to those for Case 1.

Case 3 uses $d_{hi} = 5$ and $R_{12} = 1$; the resulting $\hat{V}_{Wh}$ are distributed as constant multiples of chi-square random variables with $d_{W0} = 10$ degrees of freedom. The simulated quantiles for Case 3 were somewhat more consistent with the $\hat{d}_{W0}$ observed for the NHANES III dataset. For example, ten of the twelve variables have $\hat{d}_{W0}$ at or above the simulated 0.10 quantile of 6.4. However, two of the variables (lead and systolic blood pressure) had their $\hat{d}_{W0}$ below the simulated 0.005 quantile for Case 3.

Cases 4 through 6 cover more extreme cases of instability, induced by use of the scale factor $R_{12} = 9$. A scale factor different from one introduces a component of variability associated with sampling of PSUs with unequal σ^2_{2hi}, and causes the $\hat{V}_{Wh}$ to have distributions outside of the rescaled chi-square family. Cases 4 through 6 use the same d_{hi} values used in Cases 1 through 3, respectively. The smallest observed NHANES III $\hat{d}_{W0}$ values are somewhat more consistent with the simulated quantiles for Cases 4 through 6, although the $\hat{d}_{W0} = 1.0$ for systolic blood pressure still falls below the simulated 0.005 quantile for Cases 4 and 5, and is approximately equal to the simulated 0.025 quantile for Case 6.

In addition, note that the three largest observed $\hat{d}_{W0}$ values (for the hypertension indicator, the total cholesterol measure, and diastolic blood pressure) fall above the simulated upper 0.995 quantiles for each of cases 4 through 6. This, in conjunction with the abovementioned results for Cases 1 through 3, indicates that the twelve observed $\hat{d}_{W0}$ are consistent with settings that produce substantially different true d_{W0} values for different variables.

Taken together, these simulation results suggest that for the twelve NHANES III variables examined, the stability of $\hat{V}_{Wh}$ may be substantially worse than one would anticipate from a simple count of SSUs within each stratum; and that the true stability measures d_{W0} may vary substantially from one variable to the next.

5.2.4 Diagnostic Plots

In a purely numerical sense, $\hat{d}_{W0}$ depends on the magnitudes of the $\tilde{V}(\hat{V}_{Wh})$ relative to the terms $2\hat{V}^2_{Wh}$. Consequently, diagnostic plots of $\tilde{V}(\hat{V}_{Wh})^{1/2}$ against $\hat{V}_{Wh}$ are useful in the identification of specific patterns and "problem strata" that lead to unusually high or low $\hat{d}_{W0}$.

Figures 1 through 3 give plots for the variables HAE2 (diagnosed hypertension), log(blood lead), and blood lead, respectively. Each plot was constructed with horizontal and vertical axes on the same scale. The plot for HAE2 has the bulk of its points well below a line with slope = 1 and intercept = 0. In addition, the values of $\tilde{V}(\hat{V}_{Wh})^{1/2}$ that are large in an absolute sense are still substantially less than the corresponding $\hat{V}_{Wh}$. This is consistent with the relatively large degrees-of-freedom value $\hat{d}_{W0} = 23.7$. The plot for log(blood lead) shows a somewhat greater concentration of points near the line with slope = 1 and intercept = 0, which is consistent with the somewhat smaller value $\hat{d}_{W0} = 10.5$.

The plot for blood lead shows one apparent outlier: the largest value of $\tilde{V}(\hat{V}_{Wh})^{1/2}$ is approximately equal to the corresponding $\hat{V}_{Wh}$. For this stratum, we examined the terms $\hat{V}_{Wh}$ and $p_{hi}^{-2}\hat{\sigma}^2_{2hi}$ for unusual patterns, *e.g.*, extreme individual

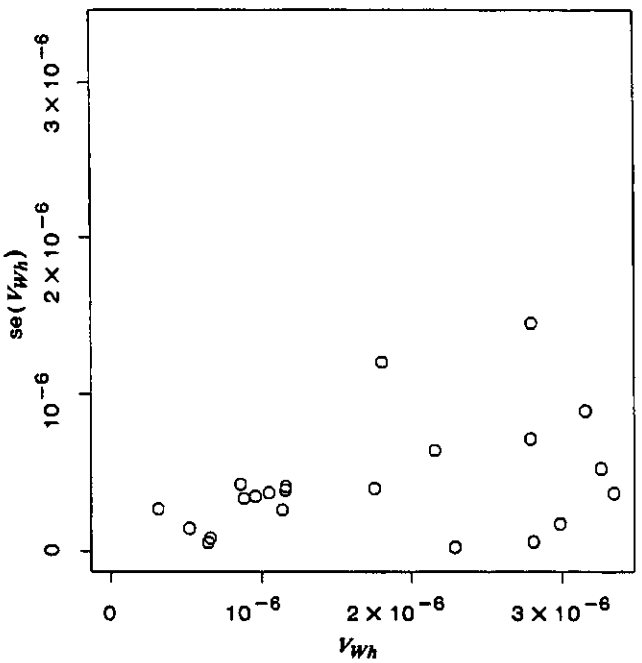


Figure 1. Plot of $\tilde{V}(\hat{V}_{Wh})^{1/2}$ against $\hat{V}_{Wh}$ for HAE2

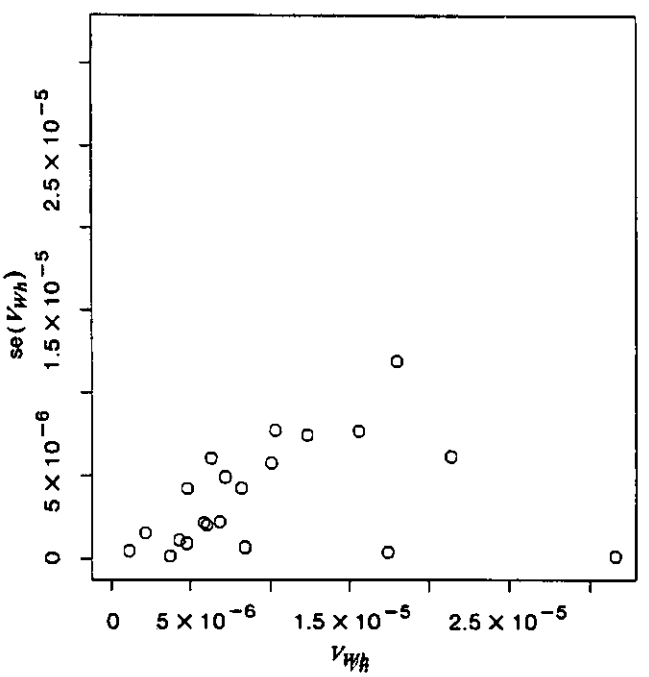


Figure 2. Plot of $\tilde{V}(\hat{V}_{Wh})^{1/2}$ against $\hat{V}_{Wh}$ for log (blood lead)

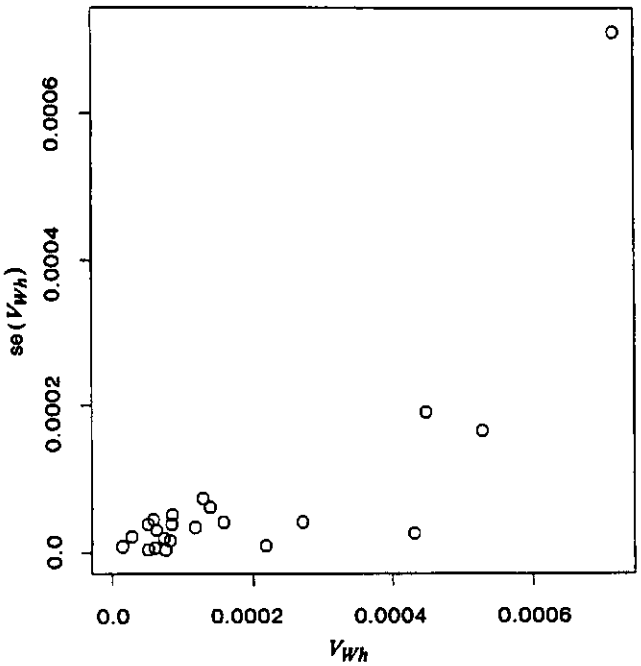


Figure 3. Plot of $\tilde{V}(\hat{V}_{Wh})^{1/2}$ against $\hat{V}_{Wh}$ for blood lead

values or extreme element-level weights. Here, one of the two associated $p_{hi}^{-2}\hat{\sigma}_{2hi}^2$ values was approximately equal to zero and the other was the largest of all the PSU-level terms $p_{hi}^{-2}\hat{\sigma}_{2hi}^2$. In addition, the stratum in question had the largest $\hat{V}_h$ value. However, this stratum did not display outlying values of $\tilde{V}(\hat{V}_{Wh})^{1/2}$ and $\hat{V}_h$ for other related variables, *e.g.*, log (blood lead). Thus, the unusual pattern observed for blood lead may be attributable to a few very high observed values for the blood lead variable, rather than to the sample design or weighting as such. Within this context, note that at the population level in the U.S., lead measurements tend to have a roughly lognormal distribution, and high lead measurements show some tendency to be clustered together due to environmental factors.

5.3 Between-PSU Variance Estimates and the Variance Ratio $\hat{R}_{WV}$

Table 5 presents the estimates $\hat{V}_B$ and $\hat{R}_{WV}$, and associated standard errors, for the twelve NHANES III variables. Of special interest are the columns labeled $\tilde{V}(\hat{V}_B)^{-1}\tilde{V}(\hat{V}_W)$, the proportion of the variance estimate $\tilde{V}(\hat{V}_B)$ that is attributable to the within-PSU variance term; and $\tilde{V}(\hat{R}_{WV})^{-1}\hat{V}_W^{-2}\hat{R}_{WV}^2\tilde{V}(\hat{V}_W)$, the corresponding proportion for $\hat{R}_{WV}$. Relatively large values for these proportions indicate that $\tilde{V}(\hat{V}_W)$ makes a substantial contribution to $\tilde{V}(\hat{V}_B)$ and $\tilde{V}(\hat{R}_{WV})$ for the variables in question.

Note that the proportion $\tilde{V}(\hat{R}_{WV})^{-1}\hat{V}_W^{-2}\hat{R}_{WV}^2\tilde{V}(\hat{V}_W)$ is greater than or equal to 0.3 for blood lead, BP1K1 (systolic blood pressure) and BP1K5 (diastolic blood pressure). For blood lead and BP1K1, the large proportions arise primarily because of the relatively large value of $\tilde{V}(\hat{V}_W)$. For BP1K5, $\tilde{V}(\hat{V}_W)$ is not as large on a relative scale, but the proportion $\tilde{V}(\hat{R}_{WV})^{-1}\hat{V}_W^{-2}\hat{R}_{WV}^2\tilde{V}(\hat{V}_W)$ is still large because $\hat{V}_W$ is not small relative to $\hat{V}(\hat{Y})$. For all three variables, the relatively large values of $\tilde{V}(\hat{R}_{WV})^{-1}\hat{V}_W^{-2}\hat{R}_{WV}^2\tilde{V}(\hat{V}_W)$ indicate that it is important to account for the variance $V(\hat{V}_W)$ when one considers the stability of $\hat{R}_{WV}$. For BP1K5, a similar comment applies to the effect of $V(\hat{V}_W)$ on the stability of $\hat{V}_B$.

Table 5
Estimates of $\hat{V}_B$ and $\hat{R}_{WV}$ for Twelve NHANES III Variables with Associated Standard Errors and Relative Within-PSU Contributions

Variable name	$\hat{V}_B$	se($\hat{V}_B$)	$\tilde{V}(\hat{V}_B)^{-1}\tilde{V}(\hat{V}_W)$
HAE2	0.0000126	0.0000188	0.020
HAE7	0.0000532	0.0000445	0.030
HAD1	-0.00000208	0.00000246	0.186
HAR3	0.0000825	0.0000703	0.047
BMPHT	0.0193	0.0114	0.027
BMPWT	0.0174	0.0400	0.096
HDRESULT	0.0887	0.0744	0.010
TCRESULT	0.270	0.253	0.031
LEAD	0.00269	0.00188	0.168
log(LEAD)	0.000468	0.000205	0.012
BP1K1	1.823	0.997	0.081
BP1K5	-0.0351	0.0793	0.367
	$\hat{R}_{WV}$	se($\hat{R}_{WV}$)	$\tilde{V}(\hat{R}_{WV})^{-1}\hat{V}_W^{-2}\hat{R}_{WV}^2\tilde{V}(\hat{V}_W)$
HAE2	1.327	0.491	0.034
HAE7	1.648	0.556	0.077
HAD1	0.783	0.247	0.123
HAR3	1.676	0.600	0.122
BMPHT	1.864	0.530	0.089
BMPWT	1.168	0.391	0.126
HDRESULT	2.193	1.020	0.047
TCRESULT	1.458	0.436	0.063
LEAD	1.694	0.555	0.367
log(LEAD)	3.221	1.025	0.112
BP1K1	2.699	1.142	0.391
BP1K5	0.861	0.300	0.300

6. DISCUSSION

This paper has presented three main ideas. First, due to the role that estimated within-PSU variances $\hat{V}_{Wh}$ play in survey design and analysis, it is important to account for the sampling error encountered in estimation of V_{Wh}. Second, standard design-based estimation methods lead to relatively simple estimators of the design variance of $\hat{V}_{Wh}$. In general, interpretation of these stability measures

requires some caution. However, they can provide useful diagnostics for the identification of variables for which the instability of $\hat{V}_{Wh}$ is especially problematic, or has an especially pronounced effect on the variance of related quantities like $\hat{V}_B$ and $\hat{R}_{WV}$. Third, the application to the U.S. Third National Health and Nutrition Examination Survey (NHANES III), and associated simulation work, indicated the following.

(i) For different sets of variables, the observed stability measures $\hat{d}_{W0}$ are consistent with substantially different sets of stability conditions.

(ii) For some variables, the estimators $\hat{V}_{Wh}$ are considerably less stable than one would anticipate from a direct count of secondary sample units.

(iii) For some variables, the estimated variance of $\hat{V}_{Wh}$ makes a substantial contribution to the estimated variances of the estimated between-PSU variance $\hat{V}_B$ and the variance ratio $\hat{R}_{WV}$.

ACKNOWLEDGEMENTS

We thank Van Parsons, Cliff Johnson and other NCHS statisticians for providing access to the NHANES III dataset; and for sharing a wealth of information regarding the NHANES III – Phase I project. We also thank Van Parsons and two anonymous referees for helpful comments on earlier versions of this paper. This research was supported in part by the National Center for Health Statistics. The views expressed in this paper are those of the authors and do not necessarily represent the policies of the National Center for Health Statistics.

REFERENCES

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

HANSEN, M.H., HURWITZ, W.N., and MADOW, W.G. (1953). *Sample Survey Methods and Theory, Volume I: Methods and Applications*. New York: John Wiley.

HERZOG, T.N., and SCHEUREN, F.J. (1976). Dallying with some CPS design effects for proportions. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 396-401.

JANG, D.S., and ELTINGE, J.L. (1996). Use of Within-PSU Variances and Errors-in-Variables Regression to Assess the Stability of a Standard Design-Based Variance Estimator. Unpublished manuscript, Department of Statistics, Texas A&M University.

KORN, E.I., and GRAUBARD, B.G. (1995). Analysis of large health surveys: accounting for the sampling design. *Journal of the Royal Statistical Society*, Series A, 158, 263-295.

NATIONAL CENTER FOR HEALTH STATISTICS (1996). National Health and Nutrition Examination Survey III Report (in press). National Center for Health Statistics, Hyattsville, MD.

RUST, K., and KALTON, G. (1987). Strategies for collapsing strata for variance estimation. *Journal of Official Statistics*, 3, 69-81.

SATTERTHWAITE, F.E. (1941). Synthesis of variance. *Psychometrika*, 6, 309-316.

SATTERTHWAITE, F.E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin*, 2, 110-114.

WOLTER, K.M. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag.

Survey Methodology, December 1996
Vol. 22, No. 2, pp. 167-173
Statistics Canada

Asymptotic Variance for Sequential Sampling Without Replacement With Unequal Probabilities

YVES G. BERGER[1]

ABSTRACT

We propose a second-order inclusion probability approximation for the Chao plan (1982) to obtain an approximate variance estimator for the Horvitz and Thompson estimator. We will then compare this variance with other approximations provided for the randomized systematic sampling plan (Hartley and Rao 1962), the rejective sampling plan (Hájek 1964) and the Rao-Sampford sampling plan (Rao 1965 and Sampford 1967). Our conclusion will be that these approximations are equivalent if the first-order inclusion probabilities are small and if the size of the sample is large.

KEY WORDS: Sampling with replacement; Randomized systematic sampling plan; Rejective sampling plan; Rao-Sampford sampling plan; Inclusion probabilities; Horvitz-Thompson; Yates-Grundy.

1. INTRODUCTION

Consider a finite population U_N containing N units and a subset U_k of U_N comprising the first units k of U_N. Let $\pi_{(k;i)}$ denote the first-order inclusion probabilities for a population U_k. We assume that they are proportional to an auxiliary variable. These probabilities have two arguments: the size k of the population and the serial number i of the unit within the population. We assume that $\pi_{(k;i)} < 1$ for all i and that all $k > n$. This hypothesis has more chance of breaking down when k is small, *i.e.*, close to n. We can solve this problem by assuming that the values of the auxiliary variable show little dispersion for those units occurring at the beginning of the population.

Let $\pi_{(k;i,j)}$ denote the second-order inclusion probability of units i and j for a population U_k. These probabilities are dependent on the sampling plan used.

We will use the Horvitz-Thompson estimator (1951) to estimate the total $\sum_{i=1}^{N} Y_i$ of a variable Y. This estimator is given by

$$t_{HT} = \sum_{i \in S_N} \frac{Y_i}{\pi_{(N;i)}}; \qquad (1)$$

where S_N is a sample of U_N. We assume that the size of S_N is constant and equal to n.

Given that the size of the sample is fixed, a variance estimator of (1) is given by the Yates-Grundy estimator (1953),

$$\hat{V} = \sum_{j \in S_N} \sum_{i \in S_N; i<j} \frac{-\Delta_{(N;i,j)}}{\pi_{(N;i,j)}} \left[\frac{Y_i}{\pi_{(N;i)}} - \frac{Y_j}{\pi_{(N;j)}} \right]^2, \qquad (2)$$

where

$$\Delta_{(N;i,j)} = \pi_{(N;i,j)} - \pi_{(N;i)}\pi_{(N;i)}. \qquad (3)$$

Let us consider the sample size sequence $\{n_1, n_2, ..., n_\nu, ...\}$ and the population size sequence $\{N_1, N_2, ..., N_\nu, ...\}$, where n_ν and N_ν increase whenever $\nu \to \infty$. To simplify the problem we eliminate the index ν.

The asymptotic approach used here is that of Hájek (1964):

$$d = \sum_{j=1}^{N} \pi_{(N;j)}[1 - \pi_{(N;j)}] \to \infty,$$

which means that $n \to \infty$ and $(N - n) \to \infty$, given that $d \le \sum_{j=1}^{N} [1 - \pi(N;j)] = N - n$ and that $d \le \sum_{j=1}^{N} \pi(N;j) = n$.

In section 2, we introduce the Chao sampling plan (1982) as well as three results linked to first and second-order inclusion probabilities. In section 3, we provide an approximation of $\pi_{(N;i,j)}$. In section 4, we propose an approximation of the Yates-Grundy variance. Section 5 compares this variance approximation with other approximations proposed for the randomized systematic plan, the rejective plan and the Rao-Sampford plan. Two numerical examples are provided in section 6.

2. CHAO SAMPLING PLAN

This is a sampling plan without replacement with unequal probabilities, of fixed size. This method is a generalization of the method used by McLeod and Bellhouse (1983) for a simple plan.

Let S_K denote a sample of size n of U_k with a set $\{\pi_{(k;i)}: i \in U_k\}$ of first-order inclusion probabilities. The Chao plan provides for a sample S_{k+1} of size n of U_{k+1} with a set $\{\pi_{(k+1;i)}: i \in U_{k+1}\}$ of first-order inclusion probabilities. The method entails selecting the $(k + 1)$-th unit with the probability $\pi_{(k+1;k+1)}$. If this unit is not selected, then we take $S_{k+1} = S_k$; otherwise we take $S_{k+1} = S_k \cup \{k + 1\} \backslash \{j\}$, where j is a unit selected at random within S_k. The procedure starts from an initial sample $S_n = U_n$ comprising the first units n of the population.

[1] Yves Berger, Université Libre de Bruxelles, Laboratoire de Méthodologie du Traitement des Données, C.P. 124, Avenue Jeanne, 44, Bruxelles, Belgique, E-Mail : yvberger@ulb.ac.be

The Chao plan provides the advantage of being sequential. In fact, it allows us to select a sample through a simple sequential run of the population. The systematic plan is another sequential plan that is often used. However, the latter is inconvenient in that it induces zero second-order inclusion probabilities. We can avoid this problem by randomizing the systematic plan. In such a case, the population is ordered at random before the sample is selected. This operation eliminates in part the problem of zero second-order inclusion probabilities. As will be seen at the end of this section, the Chao plan offers the advantage of not having any zero second-order inclusion probabilities. Randomization is therefore not needed for the latter.

The rejective plan and the Rao-Sampford plan are inconvenient in that they are not sequential. In fact, the units are selected at random with replacement within the population. If a unit is selected twice, we are forced to select a new sample. These two plans, although they are more easily understood, are more difficult to implement than the Chao plan.

The following theorem, which is a direct application of the theorem given by Chao (1982), provides a relation between the first-order inclusion probability $\pi_{(k;i)}$ of the i-th unit of U_k and the first-order inclusion probability $\pi_{(k+1;i)}$ of the i-th unit of U_{k+1}.

Theorem 1

$$\pi_{(k+1;i)} = \begin{cases} [1 - \pi_{(k+1;k+1)} R_{(k;i)}]\, \pi_{(k;i)}, & \text{for } i < k+1; \\ \pi_{(k+1;k+1)}, & \text{for } i = k+1; \end{cases} \quad (4)$$

where

$$R_{(k;i)} = \begin{cases} \dfrac{1 - \pi_{(n+1;i)}}{\pi_{(n+1;n+1)}}, & \text{for } k = n, \\ \dfrac{1}{n}, & \text{for } k \geq n+1. \end{cases} \quad (5)$$

The second-order inclusion probabilities can be calculated iteratively using the following theorem:

Theorem 2 (Chao, 1982)

$$\pi_{(k;i,j)} = \begin{cases} \{1 - \pi_{(k;k)}[R_{(k-1;i)} + R_{(k-1;j)}]\}\pi_{(k-1;i,j)}, & \text{for } i<j<k; \\ \pi_{(k;k)}[1 - R_{(k-1;i)}]\, \pi_{(k-1;i)}, & \text{for } i<j=k. \end{cases}$$

Bethlehem and Schuerhoff (1984) give a sufficient and necessary condition for the second-order inclusion probabilities to be strictly positive for a population U_k:

$$\#\{i : i \leq \ell \text{ and } \pi_{(\ell;i)} = 1\} \neq n - 1, \text{ for } \ell \text{ such that } n < \ell \leq k.$$

Since $\pi_{(\ell;i)} < 1$ for all i and ℓ such that $i \leq \ell \leq k$, this condition is always met. Therefore, within the framework of this article, we will never have zero second-order inclusion probabilities.

Moreover, the quantity $\Delta_{(N;i,j)}$ is always negative if we use the Chao plan (Chao 1982, p. 656). Then the Yates-Grundy variance offers the advantage of always being positive.

3. APPROXIMATION OF SECOND-ORDER INCLUSION PROBABILITIES

The following theorem provides us with an asymptotic expression for second-order inclusion probabilities for the Chao plan.

Theorem 3

$$\pi_{(N;i,j)} \approx \begin{cases} \pi_{(N;i)}\pi_{(N;j)} \dfrac{n-1}{n - p_{(j)}}, & \text{if } j > n+1; \\ \pi_{(N;i)}\pi_{(N;j)} \dfrac{\pi_{(n+1;i)} + \pi_{(n+1;j)} - 1}{\pi_{(n+1;i)}\pi_{(n+1;j)}}, & \text{if } j \leq n+1; \end{cases} \quad (6)$$

where $p_{(j)} = \pi_{(j;j)}$ and $i<j$.

The proof of this theorem can be found in Appendix I.

Note that this approximation has a different structure depending on whether $j > n+1$ or $j \leq n+1$. To avoid this problem, we will use a plausible condition for the auxiliary variable so that these two structures will be equivalent. Let us consider the hypothesis given in the introduction, that the values of the auxiliary variable show little dispersion for the first units $n+1$ of the population. More precisely, we assume that the auxiliary variable is constant for the first units $n+1$, *i.e.*:

$$\pi_{(n+1;i)} = \frac{n}{n+1} \quad \text{for } i \leq n+1.$$

In this case,

$$\frac{\pi_{(n+1;i)} + \pi_{(n+1;j)} - 1}{\pi_{(n+1;i)}\pi_{(n+1;j)}} = \frac{n-1}{n - \pi_{(n+1;j)}}.$$

By using (6), we have the following approximation for second-order inclusion probabilities

$$\pi_{(N;i,j)} \approx \pi_{(N;i)}\pi_{(N;j)} \frac{n-1}{n - p_{(j)}} \quad \text{if } i < j; \quad (7)$$

where

$$p_{(j)} = \begin{cases} \pi_{(j;j)}, & \text{if } j > n+1, \\ \pi_{(n+1;j)}, & \text{if } j \leq n+1. \end{cases} \quad (8)$$

4. VARIANCE ESTIMATOR

Relation (7) leads to the following approximation for $\Delta_{(N;i,j)}$:

$$\tilde{\Delta}_{(N;i,j)} = \pi_{(N;i)}\pi_{(N;j)}\frac{p_{(j)} - 1}{n - p_{(j)}}, \quad \text{if} \quad i < j. \tag{9}$$

(2), (7) and (9) provide an asymptotic expression for the Yates-Grundy estimator.

$$\tilde{V}_C = \frac{1}{[n-1]}\sum_{j\in S_N}[1 - p_{(j)}]\sum_{i\in S_N; i<j}\left[\frac{Y_i}{\pi_{(N;i)}} - \frac{Y_j}{\pi_{(N;j)}}\right]^2. \tag{10}$$

But this expression tends to underestimate the variance. In fact, to establish relation (6), we use approximation (19) from Appendix I. This approximation always implies that:

$$\pi_{(N;i,j)} < \pi_{(N;i)}\pi_{(N;j)}\frac{n-1}{n - p_{(j)}}. \tag{11}$$

This can easily be verified if we observe that (20) is obtained from (18) using approximation (19). Inequality (11) is therefore true for $j > n + 1$. For $j \leq n + 1$, it is sufficient to observe that (21) is also obtained from (19). Inequality (11) implies that:

$$\frac{-\Delta_{(N;i,j)}}{\pi_{(N;i,j)}} > \frac{1 - p_{(j)}}{n-1}, \tag{12}$$

given that $\Delta_{(N;i,j)} < 0$. From (2), (10) and (12), we have effectively

$$\hat{V} > \tilde{V}_C.$$

To overcome this problem of variance underestimation, we plan to make an adjustment on (9). It is well known that:

$$\sum_{i=1;i\neq j}^{N}\pi_{(N;i,j)} = (n-1)\pi_{(N;j)}. \tag{13}$$

Approximation (7) does not abide by constraint (13). The adjustment involves assuming that the $p_{(j)}$ are unknown and selecting them so as to satisfy (13) for the second-order probability approximation, *i.e.*:

$$\sum_{i=1}^{j-1}\pi_{(N;i)}\pi_{(N;j)}\frac{n-1}{n-p_{(j)}} + \sum_{i=j+1}^{N}\pi_{(N;i)}\pi_{(N;j)}\frac{n-1}{n-p_{(i)}} = (n-1)\pi_{(N;j)}.$$

This constraint can be written as follows:

$$\sum_{i=1}^{j-1}\pi_{(N;i)} + \sum_{i=j+1}^{N}\pi_{(N;i)}\frac{n - p_{(j)}}{n - p_{(i)}} = n - p_{(j)}. \tag{14}$$

Given that $\sum_{j=1}^{N}\pi_{(N;j)} = n$, constraint (14) is practically verified if

$$p_{(i)} = \pi_{(N;i)} \tag{15}$$

$$\sum_{i=j+1}^{N}\pi_{(N;i)}\frac{n - \pi_{(N;j)}}{n - \pi_{(N;i)}} \approx \sum_{i=j+1}^{N}\pi_{(N;i)}. \tag{16}$$

Relation (16) is plausible given that the difference between the left and right sides of (16) has as its lower bound

$$\frac{1}{n}\sum_{i=j+1}^{N}\pi_{(N;i)}[\pi_{(N;i)} - \pi_{(N;j)}],$$

and as its upper bound

$$\frac{1}{n-1}\sum_{i=j+1}^{N}\pi_{(N;i)}[\pi_{(N;i)} - \pi_{(N;j)}].$$

These two bounds are close to zero when the $\pi_{(N;i)}$ show little dispersion. This means that solution (15) is appropriate when the $\pi_{(N;i)}$ are small. Furthermore, the greater the value of j, the closer the two bounds are to zero. Therefore, solution (15) verifies (13) all the more as j is large. This implies that our approximation (9) is very good for the duplicate pairs (i, j) $(i < j)$ such that the unit j is located at the end of the population. In fact, we want approximation (9) to be the best for the duplicate pairs (i, j) whose presence in the sample is highly probable (*i.e.*, for the pairs (i, j) $(i < j)$ for which $\pi_{(N;i,j)}$ is the largest). It is therefore preferable to place the units having high first-order inclusion probabilities at the end of the population.

If we choose to have $p_{(i)} = \pi_{(N;i)}$, we have $p_{(i)}$ smaller than (8). This leads to a larger variance approximation. This solution is all the more acceptable as it corresponds to the result of the simple plan without replacement. In fact, if we replace within (7) $\pi_{(N;i)}$, $\pi_{(N;j)}$ and $p_{(j)}$ by n/N, we obtain

$$\pi_{(N;i,j)} \approx \frac{n(n-1)}{N(N-1)}, \quad \text{if} \quad i > n + 1.$$

This expression corresponds, quite clearly, to the result of the simple plan without replacement.

In conclusion, we approximate $\Delta_{(N;i,j)}$ through (9) with $p_{(i)} = \pi_{(N;i)}$. We assume that the population is ordered in such a way that the units having small $\pi_{(N;i)}$ are located at the beginning of the population and that the units having large $\pi_{(N;i)}$ are located at the end of the population. We also assume that the $\pi_{(N;i)}$ do not show too much dispersion for the first units $n + 1$ of the population.

5. COMPARISON WITH OTHER PLANS

Instead of comparing the second-order inclusion probabilities, we will compare the quantities $-\Delta_{(N;i,j)}/\pi_{(N;i,j)}$ which are of some use in calculating the Yates-Grundy variance. We will examine what these quantities provide for the Chao plan, the randomized systematic plan (Hartley and Rao 1962), the rejective plan (Hájek 1964) and the Rao-Sampford plan (Rao 1965, and Sampford 1967).

Theorem 4

$$\frac{-\Delta_{(N;i,j)}}{\pi_{(N;i,j)}} \approx \begin{cases} \dfrac{1-\pi_{(N;j)}}{n-1}, & \text{For the Chao plan;} \\ \dfrac{1-\pi_{(N;i)}-\pi_{(N;j)}}{n-1}, & \text{for the randomized systematic plan;} \\ \dfrac{n[1-\pi_{(N;i)}][1-\pi_{(N;j)}]}{d(n-1)}, & \text{for the rejective plan and the Rao-Sampford plan.} \end{cases}$$

The proof of this theorem can be found in Appendix II.

It is important to note that the proposed approximation for the randomized systematic plan comes from Deville's approximation (p. 21) and not from the famous Hartley-Rao approximation (1962). We were not able to use the Hartley-Rao formula because the latter is based on the asymptotic hypothesis, n fixed and $N \to \infty$, which is different from that adopted in this paper.

We observe that if the $\pi_{(N;i)}$ are small, $-\Delta_{(N;i,j)}/\pi_{(N;i,j)}$ is equivalent for the Chao plan and for the systematic plan. However, we observe that $-\Delta_{(N;i,j)}/\pi_{(N;i,j)}$ is always smaller in the systematic case than it is in the Chao case. This is certainly due to the fact that the approximation for the systematic plan underestimates $-\Delta_{(N;i,j)}/\pi_{(N;i,j)}$. This can be confirmed by replacing $\pi_{(N;i)}$ and $\pi_{(N;j)}$ by n/N. We then have

$$\frac{-\Delta_{(N;i,j)}}{\pi_{(N;i,j)}} \approx \frac{N-2n}{N(n-1)},$$

for the randomized systematic plan. This is equivalent to a simple plan, thus

$$\frac{-\Delta_{(N;i,j)}}{\pi_{(N;i,j)}} = \frac{N-n}{N(n-1)}.$$

We intend to adjust the approximation of $-\Delta_{(N;i,j)}/\pi_{(N;i,j)}$ for the systematic plan by multiplying it by

$$\frac{N-n}{N-2n} = \frac{1-f}{1-2f},$$

where $f = n/N$ is the sampling rate.

The approximation of $-\Delta_{(N;i,j)}/\pi_{(N;i,j)}$ for the Chao plan is also of the same magnitude as that of the rejective plan. In fact, if the $\pi_{(N;i)}$ are small, we have the approximation

$$\frac{n[1-\pi_{(N;i)}]}{d} \approx \frac{n[1-\pi_{(N;i)}]}{[1-\pi_{(N;i)}]\sum_{j=1}^{N}\pi_{(N;j)}} \approx 1.$$

Therefore, the Yates-Grundy estimator is approximately the same whether we use the Chao plan, the randomized systematic plan, the rejective plan or the Rao-Sampford plan, for large n and small $\pi_{(N;i)}$.

6. NUMERICAL EXAMPLES

The two following examples correspond to two extreme cases. In the first example, the $\pi_{(N;i)}$ show little dispersion; in the second, they show much more dispersion. Let us consider a small sample of size 20. The population size is 50 so that the $\pi_{(N;i)}$ are not too small. We have willingly opted for a bad situation in order to show that even with a sample of size 20 and a small population, the asymptotic results nevertheless represent a good approximation.

Example 1

Let us consider the first-order inclusion probabilities represented in Figure 1.

Figure 1. First-order inclusion probabilities in the case of Example 1

Figure 2 shows, on the Y axis, the true values of $-\Delta_{(N;i,j)}/\pi_{(N;i,j)}$ for the Chao plan and, on the X axis, the approximations. We have also represented the straight line where the approximations are equal to the true values. The approximations are all the better as the points are close to the straight line.

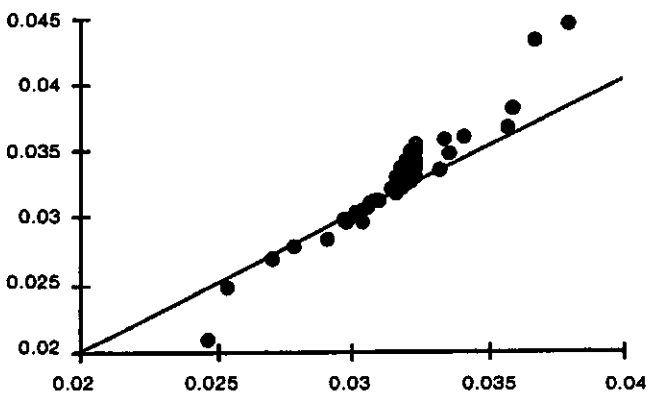


Figure 2. Approximations and true values of $-\Delta_{(N;i,j)}/\pi_{(N;i,j)}$, in the case of Example 1

We have a mean error of -0.000569 with a standard deviation of 0.0015996. This is very small in relation to the order of magnitude of the approximations. The centre of gravity of the scatter plot is located in (0.0313; 0.0318). It might seem surprising that there are less points at the left of the centre of gravity than at the right. This is simply due to the fact that most of the points at the left of the centre of gravity overlap.

We observe that the pairs (i, j) with $i < j$ such that $\pi_{(N;i)}$ is large correspond to points located on the left. They are the pairs showing the best approximation. Moreover, there is a high probability that these pairs are located within the sample given that $\pi_{(N;i)}$ is large. Therefore, our approximate variance (10) is definitely acceptable.

Example 2

The first-order inclusion probabilities are given in Figure 3. Here we notice that these probabilities are more dispersed than in Example 1. Figure 4 provides the true values as well as the approximations of $-\Delta_{(N;i,j)}/\pi_{(N;i,j)}$.

Figure 3. First-order inclusion probabilities in the case of Example 2

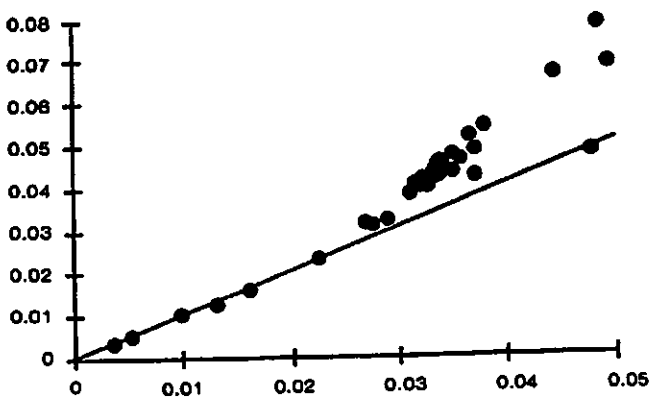


Figure 4. Approximations and true values of $-\Delta_{(N;i,j)}/\pi_{(N;i,j)}$, in the case of Example 2

We have a mean error of -0.006999 with a standard deviation of 0.006438. The centre of gravity of the scatter plot is located in (0.02957; 0.036606).

We reach the same conclusion as in Example 1. The second example leads to worse approximations. This is simply due to the high first-order inclusion probabilities.

7. CONCLUSION

The Chao plan provides a number of advantages: (i) it is sequential; (ii) the second-order inclusion probabilities are positive; and (iii) the Yates-Grundy variance is always positive. On the other hand, the second-order inclusion probabilities are difficult to calculate. That is why we propose to approximate them. We have observed that this approximation is better when the beginning of the population consists of units having small $\pi_{(N;i)}$ and the end of the population consists of units having large $\pi_{(N;i)}$. We have compared our approximation with other approximations provided for the randomized systematic plan, the rejective plan and the Rao-Sampford plan. We have concluded that these approximations are equivalent if the first-order inclusion probabilities are small and if the size of the sample is large. The two numerical examples which close this paper confirm the sound results of our approximation.

APPENDIX I

Proof of Theorem 3

Before proving this theorem, we will demonstrate the following two lemmas.

Lemma 1

$$\pi_{(k;i)} = p^*_{(i)} \prod_{\ell=a^*_i}^{k} \left[1 - \pi_{(\ell;i)} \frac{1}{n}\right];$$

where

$$p^*_{(i)} = \begin{cases} \pi_{(i;i)} & \text{if } i > n+1; \\ \pi_{(n+1;i)} & \text{if } i \le n+1; \end{cases}$$

$$a^*_i = \begin{cases} i+1 & \text{if } i > n+1; \\ n+2 & \text{if } i \le n+1. \end{cases} \qquad (17)$$

Lemma 2

$$\pi_{(k;i,j)} = q^*_{(j)} \prod_{\ell=a^*_j}^{k} \left[1 - \pi_{(\ell;\ell)} \frac{2}{n}\right];$$

where $i < j$,

$$q^*_{(j)} = \begin{cases} \pi_{(j-1;i)} \pi_{(j;j)} \left(1 - \dfrac{1}{n}\right) & \text{if } j > n+1; \\ \pi_{(n+1;i)} + \pi_{(n+1;j)} - 1 & \text{if } j \le n+1; \end{cases}$$

and a^*_i is defined by (17).

Now, with these two lemmas, we can demonstrate Theorem 3.

Proof of Theorem 3

Case 1: If $j > n+1$, using Lemma 2, we have

$$\pi_{(N;i,j)} = \pi_{(j-1;i)} \pi_{(j;j)} \left(1 - \frac{1}{n}\right) \prod_{\ell=j+1}^{N} \left[1 - \pi_{(\ell;\ell)} \frac{2}{n}\right].$$

On the basis of Lemma 1, this last expression becomes

$$\pi(N;i,j) = p^{*}_{(i)}\pi_{(j;j)}\left(1-\frac{1}{n}\right)\prod_{\ell=a_i^{\cdot}}^{j-1}\left[1-\pi_{(\ell;\ell)}\frac{1}{n}\right]\prod_{q=j+1}^{N}\left[1-\pi_{(q;q)}\frac{2}{n}\right].$$

By multiplying this last expression by

$$\left[\frac{1-\pi_{(j;j)}\frac{1}{n}}{1-\pi_{(j;j)}\frac{1}{n}}\right]\prod_{\ell=j+1}^{N}\left[\frac{1-\pi_{(\ell;\ell)}\frac{1}{n}}{1-\pi_{(\ell;\ell)}\frac{1}{n}}\right]\left[\frac{1-\pi_{(\ell;\ell)}\frac{2}{n}}{1-\pi_{(\ell;\ell)}\frac{2}{n}}\right]=1,$$

and by regrouping certain terms, we obtain

$$\pi_{(N;i,j)} = \pi_{(j;j)}p^{*}_{(i)}\left[\frac{n-1}{n-\pi_{(j;j)}}\right]\prod_{\ell=a_i^{\cdot}}^{N}\left[1-\pi_{(\ell;\ell)}\frac{1}{n}\right]\prod_{q=j+1}^{N}\left[\frac{1-\pi_{(q;q)}\frac{2}{n}}{1-\pi_{(q;q)}\frac{1}{n}}\right].$$

On the basis of Lemma 1, this last expression becomes

$$\pi_{(N;i,j)} = \left[\frac{n-1}{n-\pi_{(j;j)}}\right]\pi_{(N;i)}\pi_{(j;j)}\prod_{\ell=j+1}^{N}\left[\frac{1-\pi_{(\ell;\ell)}\frac{2}{n}}{1-\pi_{(\ell;\ell)}\frac{1}{n}}\right]. \quad (18)$$

If n is sufficiently large

$$\frac{1-\pi_{(\ell;\ell)}\frac{2}{n}}{1-\pi_{(\ell;\ell)}\frac{1}{n}} \approx \left[1-\pi_{(\ell;\ell)}\frac{2}{n}\right]\left[1+\pi_{(\ell;\ell)}\frac{1}{n}\right];$$

$$\approx 1+\frac{\pi_{(\ell;\ell)}}{n}-\frac{2\pi_{(\ell;\ell)}}{n}-\frac{2\pi^2_{(\ell;\ell)}}{n^2};$$

$$\approx 1-\frac{\pi_{(\ell;\ell)}}{n}. \quad (19)$$

Then (18) becomes,

$$\pi_{(N;i,j)} \approx \left[\frac{n-1}{n-\pi_{(j;j)}}\right]\pi_{(N;i)}\pi_{(j;j)}\prod_{\ell=j+1}^{N}\left[1-\pi_{(\ell;\ell)}\frac{1}{n}\right]. \quad (20)$$

Finally, on the basis of Lemma 1, this last expression can be written:

$$\pi_{(N;i,j)} \approx \pi_{(N;i)}\pi_{(N;j)}\frac{n-1}{n-\pi_{(j;j)}}.$$

Case 2: If $j \le n+1$, Lemma 2 provides

$$\pi_{(N;i,j)} = [\pi_{(n+1;i)}+\pi_{(n+1;j)}-1]\prod_{\ell=n+2}^{N}\left[1-\pi_{(\ell;\ell)}\frac{2}{n}\right];$$

in other words

$$\pi_{(N;i,j)} = \prod_{\ell=n+2}^{N}\left[1-\pi_{(\ell;\ell)}\frac{1}{n}\right]\prod_{q=n+2}^{N}\left[\frac{1-\pi_{(q;q)}\frac{2}{n}}{1-\pi_{(q;q)}\frac{1}{n}}\right][\pi_{(n+1;i)}+\pi_{(n+1;j)}-1].$$

By using approximation (19), we obtain

$$\pi_{(N;i,j)} \approx \left\{\prod_{\ell=n+2}^{N}\left[1-\pi_{(\ell;\ell)}\frac{1}{n}\right]\right\}^2 \pi_{(n+1;i)}\pi_{(n+1;j)}\frac{\pi_{(n+1;i)}+\pi_{(n+1;j)}-1}{\pi_{(n+1;i)}\pi_{(n+1;j)}}.$$

On the basis of Lemma 1, we obtain finally

$$\pi_{(N;i,j)} \approx \pi_{(N;i)}\pi_{(N;j)}\frac{\pi_{(n+1;i)}+\pi_{(n+1;j)}-1}{\pi_{(n+1;i)}\pi_{(n+1;j)}}. \quad (21)$$

Q.E.D.

APPENDIX II

Proof of Theorem 4

- For the Chao plan, it is sufficient to use (6), (9) and (15).
- For the randomized systematic plan, it is sufficient to use the approximation of the $\pi_{(N;i,j)}$ given by Deville (p. 21)

$$\pi_{(N;i,j)} \approx \pi_{(N;i)}\pi_{(N;j)}\frac{n-1}{n-\pi_{(N;i)}-\pi_{(N;j)}}. \quad (22)$$

This expression is obtained from the hypothesis

$$\mathrm{Max}_{1\le i\le N}\left\{\frac{\pi_{(N;i)}}{n}\right\} \to 0.$$

This last hypothesis is verified since $n \to \infty$.

- For the rejective plan, using Hájek's result (1964, p. 1508), we have

$$\frac{-\Delta_{(N;i,j)}}{\pi_{(N;i,j)}} \approx \frac{[1 - \pi_{(N;i)}][1 - \pi_{(N;j)}]}{d - [1 - \pi_{(N;i)}][1 - \pi_{(N;j)}]}, \qquad (23)$$

for $d \to \infty$. We note that (23) remains valid for the Rao-Sampford plan (see Hájek 1981, Theorem 8.2, p. 82). Using the approximation (Hájek 1964, p. 1521),

$$\{d - [1 - \pi_{(N;i)}][1 - \pi_{(N;j)}]\}^{-1} \approx \frac{n}{d(n-1)},$$

we obtain the result of the theorem.

Q.E.D.

ACKNOWLEDGEMENTS

The author wishes to thank the referees who submitted a number of constructive comments that led to considerable improvements.

REFERENCES

BETHLEHEM, J.G., and SCHUERHOFF, H. (1984). Second-order inclusion probabilities in sequential sampling without replacement with unequal probabilities. *Biometrika*, 71, 642-644.

CHAO, M.T. (1982). A general purpose unequal probability sampling plan. *Biometrika*, 69, 653-656.

DEVILLE, J.-C. (No date). Cours de sondage, Chapitre III: les outils de bases. Lecture notes, ENSAE, Paris.

HÁJEK, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. *Annals of Mathematical Statistics*, 35, 1491-1523.

HÁJEK, J. (1981). *Sampling from a Finite Population*. New York and Bassel: Marcel Dekker, Inc.

HARTLEY, H.O., and RAO, J.N.K. (1962). Sampling with unequal probabilities without replacement. *Annals of Mathematical Statistics*, 33, 350-374.

HORVITZ, D.G., and THOMPSON, D.J. (1951). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47,663-685.

McLEOD, A.I., and BELLHOUSE, D.R. (1983). A convenient algorithm for drawing a simple random sample. *Applied Statistics*, 32, 2.

RAO, J.N.K. (1965). On two simple schemes of unequal probability sampling without replacement. *Journal of the Indian Statistical Association*, 3, 173-180.

SAMPFORD, M.R. (1967). On sampling without replacement with unequal probabilities of selection. *Biometrika*, 54, 494-513.

YATES, F., and GRUNDY, P.M. (1953). Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society*, series B, 1, 253-261.

Survey Methodology, December 1996
Vol. 22, No. 2, pp. 175-183
Statistics Canada

Applications of Spatial Smoothing to Survey Data

ANN COWLING, RAY CHAMBERS, RAY LINDSAY and BHAMATHY PARAMESWARAN[1]

ABSTRACT

In this paper we present two applications of spatial smoothing using data collected in a large scale economic survey of Australian farms: one a small area and the other a large area application. In the small area application, we describe how the sample weights can be spatially smoothed in order to improve small area estimates. In the large area application, we give a method for spatially smoothing and then mapping the survey data. The standard method of weighting in the survey is a variant of linear regression weighting. For the small area application, this method is modified by introducing a constraint on the spatial variability of the weights. Results from a small scale empirical study indicate that this decreases the variance of the small area estimators as expected, but at the cost of an increase in their bias. In the large area application, we describe the nonparametric regression method used to spatially smooth the survey data as well as techniques for mapping this smoothed data using a Geographic Information System (GIS) package. We also present the results of a simulation study conducted to determine the most appropriate method and level of smoothing for use in the maps.

KEY WORDS: Kernel estimation; Mapping survey data; Small area estimation; Survey weighting.

1. INTRODUCTION

The Australian Bureau of Agricultural and Resource Economics (ABARE) is the applied economic research organisation attached to the Department of Primary Industries and Energy. Amongst its information gathering activities, ABARE conducts annual surveys of selected Australian agricultural industries which provide a broad range of information on the economic and physical characteristics of farm business units.

The largest survey is the Australian Agricultural and Grazing Industries Survey (AAGIS), which covers farm establishments with an estimated value of agricultural operations (EVAO) of $A22,500 or more in the last agricultural census that are classified to one of the broadacre industries – that is, cereal crop production, beef cattle production, and sheep and wool production. For the last two years, around 1650 farms have been included in the AAGIS sample, which is stratified by geographic area, industry, and EVAO. The sample farms are located throughout Australia with a non-uniform density. The latitude and longitude of the sample farms (defined in terms of the location of the farm "gate") is recorded as a regular part of the collection. This knowledge of the location of the surveyed farms enables the spatial smoothing techniques described in this paper to be used.

Traditionally, AAGIS estimates have been presented only as tables of numbers showing averages for all Australia, each state, and industries within states. However, the concern of rural industry and government about the combined impact of drought in some areas of Australia and the decline in certain commodity prices has highlighted the need for timely and detailed information on regional trends in farm performance. In particular, there has been a perceived need for information which portrays the spatial distribution of farm performance, reflecting actual variability in climate and production across Australia.

A highly effective way of presenting information on a spatial basis is to map the regional variation in economic performance of the surveyed farms. We use a nonparametric regression method to spatially smooth the farm level survey data, which is then presented in the form of a map. Recent improvement in computing power and the availability of high quality and affordable GIS packages have made this form of presentation a practical alternative to the traditional tabular method of presenting survey results.

Maps have been found to be a successful form of exposition for a number of reasons. First, estimates presented in a map are easily interpreted; when presented with too many tables it is very easy for a client to overlook local variations or be "swamped" by numbers. Next, maps make it easy for a client to relate the geographic variation in one variable with that of another. Finally, a colour map has great visual impact.

This demand for information on a spatial basis has resulted in an increased emphasis on small area estimates. One method of small area estimation (which originated naturally from smoothing survey data for presentation in maps) is to spatially smooth the sample weights. This reduces the variability of the small area estimates.

In Section 2, we examine a method of integrating geographical location into ABARE's survey weighting methods in order to make our small area estimates less variable. It is applied to sub-regional estimation within two Agricultural Regions in Section 3. In Section 4, we describe how kernel regression techniques can be used to produce

[1] Ann Cowling, CSIRO Division of Fisheries, GPO Box 1538, Hobart TAS 7001, Australia and Australian Bureau of Agricultural and Resource Economics; Ray Chambers, Department of Social Statistics, University of Southampton, Highfield, Southampton S017 1BJ, United Kingdom; Ray Lindsay and Bhamathy Parameswaran, Australian Bureau of Agricultural and Resource Economics, GPO Box 1563, Canberra ACT 2601, Australia.

maps which give a good indication of the local geographic variation of a surveyed variable. Two methods of mapping the smoothed data are discussed, both of which use ARC/INFO, a GIS software package. The results of a simulation study comparing various kernel regression methodologies for use in ABARE's maps are summarised in the Appendix.

2. SMALL AREA ESTIMATION BY SPATIALLY SMOOTHING SAMPLE WEIGHTS

The standard method used to compute sample weights at ABARE is described in Bardsley and Chambers (1984). It rests on the assumption that at some appropriate level of aggregation (say, Agricultural Region) the variable Y follows a linear model of the form

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \boldsymbol{V} \tag{2.1}$$

where $\boldsymbol{Y}$ is the N-vector of values of Y at this level of aggregation, $\boldsymbol{X}$ is a $N \times p$ matrix of values of a set of p benchmark variables, β is an unknown p-vector of regression coefficients and $\boldsymbol{V}$ is a N-vector of errors satisfying $E(\boldsymbol{V}) = \boldsymbol{0}$ and $\text{var}(\boldsymbol{V}) = \sigma^2\Omega$, where σ is an unknown scale parameter and Ω is a known $N \times N$ diagonal matrix having as its elements the measure of size of each farm, EVAO, introduced in the previous section.

Since this model is a multipurpose model, with the same set of benchmark variables used for each survey variable, the column dimension, p, of $\boldsymbol{X}$ is usually large. Typically, $\boldsymbol{X}$ consists of between 3 and 7 variables related to the main agricultural commodities produced by farms in the region together with dummy variables indicating industry strata within the region. Best linear unbiased estimation of the population total of a survey variable on the basis of such an overspecified model typically results in weights that are highly variable and often negative.

As discussed in Bardsley and Chambers (1984), negative weights are highly undesirable in a multi-purpose survey like AAGIS. In particular, such weights can lead to negative estimates of intrinsically positive quantities. This problem has been pointed out in the literature a number of times (see for example, Deville and Särndal 1992; Bankier, Rathwell and Majkowski 1992; and Fuller, Loughin and Baker 1994). The method used at ABARE to control for strictly positive sample weights is based on the ridge-type modification to the best linear unbiased weights suggested by Bardsley and Chambers (1984).

Given a sample of size n from a particular region, the ridge weighting approach determines the sample weight vector $\boldsymbol{w}$ by minimising the mean squared error criterion

$$Q = \lambda^{-1}\boldsymbol{B}^T\boldsymbol{C}\boldsymbol{B} + (\boldsymbol{w} - \boldsymbol{1})^T\omega(\boldsymbol{w} - \boldsymbol{1}). \tag{2.2}$$

Here $\boldsymbol{B} = \boldsymbol{T} - \boldsymbol{x}^T\boldsymbol{w}$ is a p-vector of benchmark biases, corresponding to the differences between the (known) population totals $\boldsymbol{T}$ of the p benchmark variables making up $\boldsymbol{X}$ and the corresponding survey estimates $\boldsymbol{x}^T\boldsymbol{w}$ of these totals, $\boldsymbol{C}$ is a $p \times p$ diagonal matrix of non-negative relative "costs" associated with these biases, ω is the sample component of Ω, $\boldsymbol{x}$ is the sample component of $\boldsymbol{X}$, $\boldsymbol{1}$ is a n-vector of ones and λ is a scaling constant which is chosen by the survey analyst. The value of $\boldsymbol{w}$ minimising Q is

$$\boldsymbol{w} = \boldsymbol{1} + \omega^{-1}\boldsymbol{x}(\lambda\boldsymbol{C}^{-1} + \boldsymbol{x}^T\omega^{-1}\boldsymbol{x})^{-1}(\boldsymbol{T} - \boldsymbol{x}^T\boldsymbol{1}). \tag{2.3}$$

The scale constant λ is called the ridge parameter associated with these weights. As λ increases from zero, the sample weights in $\boldsymbol{w}$ move away from their best linear unbiased values under the model (2.1) (namely, their values at $\lambda = 0$) and become less and less variable. That is, as λ increases, the variances of the survey estimates based on these weights decrease. On the other hand, as λ increases, these estimates become more biased under (2.1), so the components of $\boldsymbol{B}$ move away from their zero values at $\lambda = 0$ (where the sample weights define unbiased estimates under (2.1)). These components become larger and larger (in absolute terms) as λ increases.

The survey analyst makes a tradeoff between these two competing sources of "error" by choosing the smallest value of λ such that the sample weights in $\boldsymbol{w}$ stabilise at strictly positive values as close as possible to their best linear unbiased values under (2.1). This ensures that the components of $\boldsymbol{B}$ are as small as possible subject to this stability requirement. At ABARE, the value of λ is chosen so that the sample weights are at least unity.

Recent small area estimation research in ABARE has focussed on a method of modifying this ridge weighting procedure to create sample weights that are less spatially variable. We achieve this by modifying the mean squared error criterion Q in (2.2) to include a constraint on spatial variability, while continuing to regard the elements of the variable $\boldsymbol{Y}$ as being independent.

Let $\boldsymbol{K}$ be an $n \times n$ matrix reflecting Euclidean distance between sample farms, such that $\boldsymbol{K}$ is symmetric and non-negative, $\boldsymbol{K}_{ii} = 1$ for all i and $\boldsymbol{K}_{ij} \downarrow 0$ as the distance between farm i and farm j increases. Put $\boldsymbol{u} = \boldsymbol{w} - \boldsymbol{1}$. The aim is then to choose $\boldsymbol{u}$ so that when $\boldsymbol{K}_{ij}$ is large, the difference between $\boldsymbol{u}_i$ and $\boldsymbol{u}_j$ is small. That is, we seek to minimise a quantity of the form

$$\sum_{i\epsilon s}\sum_{j\epsilon s}\boldsymbol{K}_{ij}(\boldsymbol{u}_i - \boldsymbol{u}_j)^2 = 2(\boldsymbol{u}^{(2)})^T\boldsymbol{K}\boldsymbol{1} - 2\boldsymbol{u}^T\boldsymbol{K}\boldsymbol{u} \tag{2.4}$$

where $(\boldsymbol{u}^{(2)})_i = (\boldsymbol{u}_i)^2$. An appropriate modification to the mean squared error criterion (2.2) leads to minimisation of

$$Q^* = \lambda^{-1}\boldsymbol{B}^T\boldsymbol{C}\boldsymbol{B} + \boldsymbol{u}^T\omega\boldsymbol{u} + (\boldsymbol{u}^{(2)})^T\boldsymbol{K}\boldsymbol{1} - \boldsymbol{u}^T\boldsymbol{K}\boldsymbol{u}.$$

Minimising with respect to $\boldsymbol{u}$ leads to

$$\boldsymbol{u} = \eta^{-1}\boldsymbol{x}(\lambda\boldsymbol{C}^{-1} + \boldsymbol{x}^T\eta^{-1}\boldsymbol{x})^{-1}(\boldsymbol{T} - \boldsymbol{x}^T\boldsymbol{1})$$

provided η^{-1} exists, where

$$\eta = \text{diag}(\boldsymbol{K}\mathbf{1}) - \boldsymbol{K} + \omega. \tag{2.5}$$

Clearly, then,

$$\boldsymbol{w} = \mathbf{1} + \eta^{-1}\boldsymbol{x}(\lambda \boldsymbol{C}^{-1} + \boldsymbol{x}^T\eta^{-1}\boldsymbol{x})^{-1}(\boldsymbol{T} - \boldsymbol{x}^T\mathbf{1}). \tag{2.6}$$

It can be seen that the modified mean squared error criterion Q^* equally weights the spatial smoothness criterion given in (2.4), and the term corresponding to the variance of the prediction error of the sample estimates, $\boldsymbol{u}^T\omega\boldsymbol{u}$. As the scale of $\boldsymbol{K}$ was arbitrarily specified, the comparative weighting of the two criteria must be modified by "scaling up" the spatial matrix $\{\text{diag}(\boldsymbol{K}\mathbf{1}) - \boldsymbol{K}\}$ by a factor ϕ in order to make it comparable in size with the heteroscedasticity matrix ω, and by adding a parameter α, $0 \le \alpha \le 1$, to the expression for η in equation (2.5), so that

$$\eta = (1 - \alpha)\phi\{\text{diag}(\boldsymbol{K}\mathbf{1}) - \boldsymbol{K}\} + \alpha\omega.$$

These spatially smoothed sample weights can be derived in a second way, providing deeper insight into how they should be interpreted. This follows from noting that

$$\eta = \begin{bmatrix} \sigma_1^2 + \sum_{m\neq 1} \boldsymbol{K}_{1m} & -\boldsymbol{K}_{12} & \cdots & -\boldsymbol{K}_{1n} \\ -\boldsymbol{K}_{21} & \sigma_2^2 + \sum_{m\neq 2} \boldsymbol{K}_{2m} & \cdots & -\boldsymbol{K}_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ -\boldsymbol{K}_{n1} & -\boldsymbol{K}_{n2} & \cdots & \sigma_n^2 + \sum_{m\neq n} \boldsymbol{K}_{nm} \end{bmatrix}$$

can be expressed as $\eta = \boldsymbol{S}\boldsymbol{R}\boldsymbol{S}$, where $\boldsymbol{S}$ is a diagonal matrix with $\boldsymbol{S}_{ii} = (\sigma_i^2 + \sum_{m\neq i}\boldsymbol{K}_{im})^{1/2}$, and $\boldsymbol{R}$ is a correlation matrix with

$$\boldsymbol{R}_{ij} = \begin{cases} 1 & \text{if } i = j \\ -\boldsymbol{K}_{ij}\left\{\left(\sigma_i^2 + \sum_{m\neq i}\boldsymbol{K}_{im}\right)\left(\sigma_j^2 + \sum_{m\neq j}\boldsymbol{K}_{jm}\right)\right\}^{-1/2} & \text{if } i \neq j. \end{cases}$$

Thus the spatially smoothed sample weights can alternatively be derived as ridge-type regression weights based on the assumption that the variable $\boldsymbol{Y}$ follows a linear model of the form (2.1), with $\boldsymbol{V}$ redefined as satisfying $E(\boldsymbol{V}) = \mathbf{0}$, $\text{var}(\boldsymbol{Y}_i) = \sigma_i^2 + \sum_{m\neq i}\boldsymbol{K}_{im}$, and $\text{cov}(\boldsymbol{Y}_i, \boldsymbol{Y}_j) = -\boldsymbol{K}_{ij}$ for $i \neq j$. The usual ridge weighting procedure then leads directly to (2.6) with η defined by (2.5). Note that under this implied model neighbouring farms are negatively correlated.

This second method of derivation shows clearly that the introduction of spatial smoothness for the survey weights is at odds with standard concepts of statistical efficiency as far as estimation at the aggregate level is concerned. Since the spatial correlation between neighbouring farms will typically be positive, efficient survey estimation at the aggregate level will involve weighting based on (2.3) with ω replaced by a non-diagonal variance/covariance matrix reflecting this positive spatial correlation. These are not the weights that result when one imposes as spatial similarity constraint. Consequently, one could expect that such "large area efficient" weights would tend to be more dissimilar for neighbouring farms than they would be for farms that are far apart. That is, there is a price to pay in weighting – if less variable aggregate level estimates are required, then this tends to lead to more variable small area estimates. Conversely, if (2.6) is adopted as the method of weighting because of its desirable small area properties, then it can be expected that aggregate level estimates obtained by summing these small area estimates will be less efficient.

The spatially smooth sample weights (2.6) have been implemented using

$$\boldsymbol{K}_{ij} = \exp(-d\|z_i - z_j\|), \tag{2.7}$$

where $\|z_i - z_j\|$ is the distance between farm i and farm j and d is a constant controlling the radius of circle around the i-th farm within which spatial smoothing is applied. The smaller the value of d, the larger the radius of spatial smoothing. At present, the "scaling up" constant ϕ is computed as the ratio of the determinants of the $\boldsymbol{K}$ and ω matrices, raised to the power n^{-2}. An empirical evaluation of this method is described in the following Section.

3. AN APPLICATION OF SPATIALLY SMOOTHED SAMPLE WEIGHTING

Initial results from an evaluation of the first method of spatially smoothed ridge weighting described in the previous section are set out in Tables 1 to 3. These results are for two Agricultural Regions. The first, Region A, is in New South Wales. In spatial terms, this region is relatively homogeneous, being located in the southwestern corner of the state. The principal agricultural activities are wheat and rice production and wool and lamb production. The second, Region B, is in Western Australia. This region is more spatially heterogeneous, ranging from established cropping and wool production farms in the central west of the state to much larger livestock and cropping farms on marginal farming land in the south east of the state. The principal agricultural activities are wheat and legumes production and wool production.

Six variations of the spatially smoothed ridge weights (2.6) with $\boldsymbol{K}$ given by (2.7) were used in the evaluation, defined by values of $d = 0.05$ (weak spatial effects) and $d = 0.005$ (strong spatial effects), and values of $\alpha = 0.9$ (most emphasis on the standard ridge weights), $\alpha = 0.5$ (equal emphasis on standard ridge weights and spatially smooth weights) and $\alpha = 0.1$ (most emphasis on spatially smooth weights).

Table 1
Values (in relative percentage terms) of the biases associated with estimation of the benchmark variables corresponding to the principal agricultural commodities produced in Region A (sample size $n = 101$ farms) and Region B (sample size $n = 85$ farms) using the standard ridge weights (2.3) and the spatially smooth ridge weights (2.6)

		Wheat	Sheep	Rice
Region A				
Standard ridge weights		-0.50	5.0	13.0
Spatially smoothed ridge weights				
$d = 0.05$	$\alpha = 0.9$	-0.50	4.6	11.9
	$\alpha = 0.5$	-0.46	4.7	12.4
	$\alpha = 0.1$	0.07	6.2	17.4
$d = 0.005$	$\alpha = 0.9$	-0.40	4.9	12.7
	$\alpha = 0.5$	0.80	8.9	28.0
	$\alpha = 0.1$	9.20	25.0	60.0
		Wheat	**Sheep**	**Legumes**
Region B				
Standard ridge weights		0.43	-1.25	1.49
Spatially smoothed ridge weights				
$d = 0.05$	$\alpha = 0.9$	0.42	-1.16	1.37
	$\alpha = 0.5$	0.44	-1.14	1.40
	$\alpha = 0.1$	0.69	-1.25	2.53
$d = 0.005$	$\alpha = 0.9$	0.50	-1.20	1.68
	$\alpha = 0.5$	1.51	1.14	9.73
	$\alpha = 0.1$	26.57	19.61	45.46

Table 1 shows the relative biases associated with estimation of the population totals of the main commodity related benchmarks for each region under these different weighting systems, as well as the corresponding biases associated with the standard ridge weights. The increase in these biases as the amount of spatial smoothing in the weights is increased is evident. Since these production benchmarks are positively correlated with most of the economic variables measured in the survey, these benchmark biases can be expected to be translated into a corresponding upward bias in survey estimates based on these weights.

Figures 1 to 4 show the difference between the smoothed weights and the standard ridge weights for the two "extreme" combinations of α and d in both regions changes as the size (measured in terms of the logarithm of the estimated value of agricultural operations, or log(EVAO)) of the sample farms changes.

Observe that for relatively strong spatial smoothing (Figures 1 and 3), the effect of smoothing is to increase the weights of most of the larger sample farms, while dramatically decreasing the weights of a small number of smaller sample farms. Weak spatial smoothing (Figures 2 and 4) changes the weights much less, and there is little relationship between the size of the farm and the direction of weight change. Consequently, an upward shift in survey estimates for these regions could be expected with the introduction of

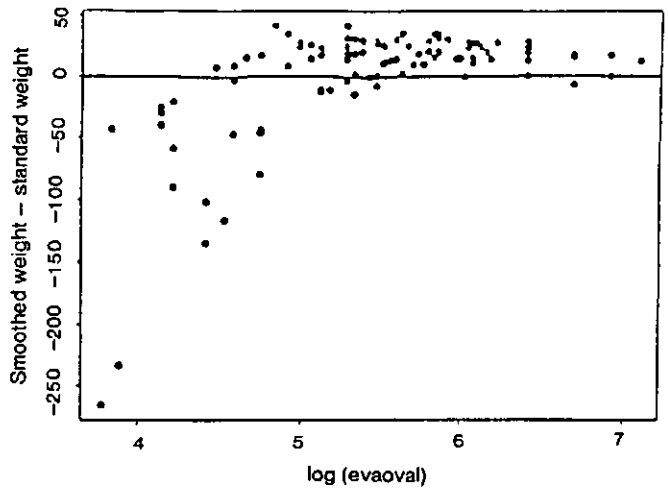


Figure 1. Difference between smoothed weight with $\alpha = 0.1$ and $d = 0.005$ and standard ridge weight, Region A

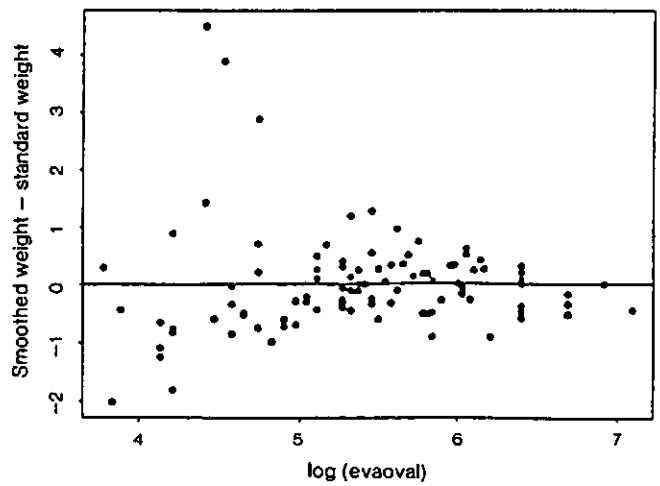


Figure 2. Difference between smoothed weight with $\alpha = 0.9$ and $d = 0.05$ and standard ridge weight, Region A

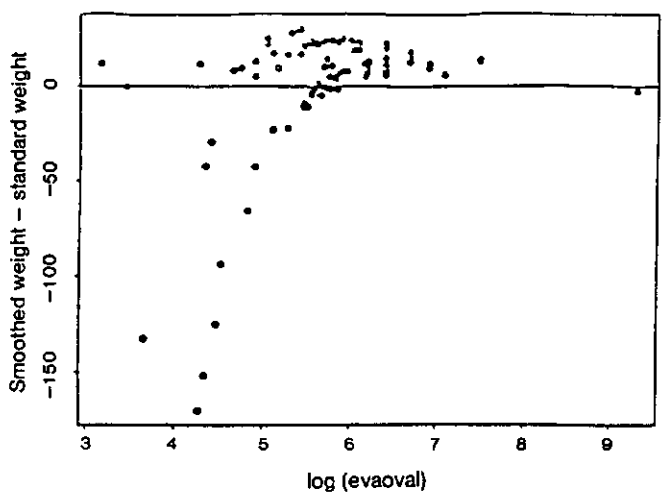


Figure 3. Difference between smoothed weight with $\alpha = 0.1$ and $d = 0.005$ and standard ridge weight, Region B

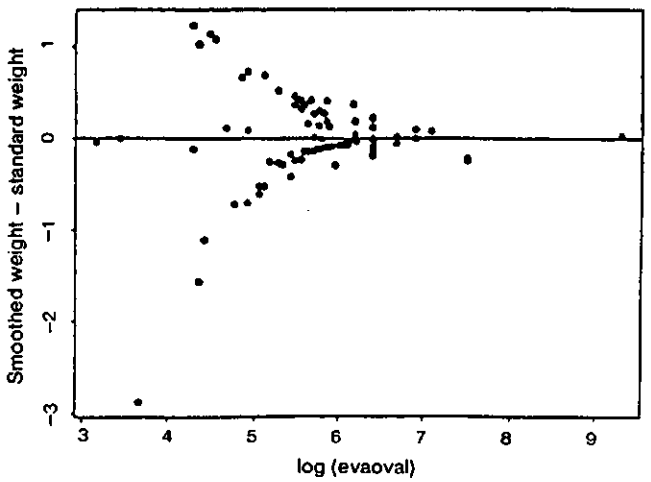


Figure 4. Difference between smoothed weight with $\alpha = 0.9$ and $d = 0.05$ and standard ridge weight, Region B

strongly spatially smoothed sample weights. Given the increased positive biases indicated in Table 1, this upward shift would be expected to be essentially due to the introduction of a positive bias in these estimates.

Is this increased bias compensated for by a lower standard error? To evaluate this question, survey estimates and estimated standard errors were computed for a key financial variable, total cash costs. These estimates are set out in Table 2 (Region A) and Table 3 (Region B). Estimates are provided both for each region and for small areas within each region, denoted SR-i in the table, with the index i ranging between 1 and 6 for Region A and between 1 and 7 for Region B.

Table 2
Estimates (with corresponding estimated standard errors in parentheses) of the average value of Y = total cash costs in subregions SR–1 to SR–6, making up Region A (sample size $n = 101$ farms), using the standard ridge weights (2.3) and the spatially smooth ridge weights (2.6)

	Standard weights	Spatially smoothed ridge weights					
		$d = 0.05$			$d = 0.005$		
		$\alpha = 0.9$	$\alpha = 0.5$	$\alpha = 0.1$	$\alpha = 0.9$	$\alpha = 0.5$	$\alpha = 0.1$
SR–1	100,618	100,453	101,297	107,263	102,059	112,635	135,419
	(24,551)	(24,511)	(23,906)	(20,487)	(23,474)	(18,923)	(18,011)
SR–2	115,320	115,417	116,002	120,362	116,917	126,165	153,707
	(26,754)	(26,661)	(26,448)	(25,637)	(26,423)	(25,990)	(27,975)
SR–3	167,524	167,453	167,486	168,257	167,709	170,781	187,683
	(28,479)	(28,467)	(28,473)	(28,426)	(28,175)	(26,471)	(24,211)
SR–4	182,940	180,317	177,838	163,556	176,257	174,077	192,296
	(106,471)	(105,485)	(101,012)	(74,418)	(97,823)	(69,109)	(43,651)
SR–5	132,050	132,083	132,389	134,786	132,490	136,369	151,046
	(25,089)	(25,096)	(25,154)	(25,475)	(25,173)	(24,410)	(23,110)
SR–6	132,493	132,184	133,204	141,623	133,763	147,652	192,781
	(44,385)	(44,546)	(44,757)	(46,736)	(45,078)	(46,953)	(53,105)
Region A	134,114	133,807	134,141	137,080	134,506	142,040	166,432
	(15,691)	(15,655)	(15,426)	(13,845)	(15,199)	(13,494)	(12,815)

Table 3
Estimates (with corresponding estimated standard errors in parentheses) of the average value of Y = total cash costs in subregions SR–1 to SR–7, making up Region B (sample size $n = 85$ farms), using the standard ridge weights (2.3) and the spatially smooth ridge weights (2.6)

	Standard weights	Spatially smoothed weights					
		$d = 0.05$			$d = 0.005$		
		$\alpha = 0.9$	$\alpha = 0.5$	$\alpha = 0.1$	$\alpha = 0.9$	$\alpha = 0.5$	$\alpha = 0.1$
SR–1	183,194	183,262	183,528	186,151	184,287	195,138	257,652
	(64,851)	(64,325)	(64,051)	(64,967)	(64,132)	(69,859)	(59,518)
SR–2	261,952	261,487	261,119	261,182	261,938	276,912	331,805
	(70,989)	(70,601)	(70,502)	(73,131)	(70,723)	(79,751)	(67,356)
SR–3	113,499	113,441	113,742	116,847	114,631	125,525	157,007
	(30,304)	(30,289)	(30,255)	(30,731)	(30,377)	(31,507)	(32,500)
SR–4	242,220	242,182	242,208	242,221	242,163	242,439	250,871
	(26,160)	(25,671)	(26,159)	(26,160)	(26,154)	(24,244)	(24,836)
SR–5	134,524	134,970	135,700	139,122	134,734	131,448	148,629
	(32,420)	(32,528)	(32,432)	(30,607)	(32,202)	(27,867)	(27,942)
SR–6	176,540	176,977	175,708	163,241	172,076	148,434	171,856
	(60,377)	(60,703)	(59,214)	(46,361)	(55,925)	(36,218)	(39,527)
SR–7	205,287	205,644	205,433	202,039	204,519	194,998	219,959
	(44,137)	(44,008)	(43,963)	(44,044)	(43,972)	(45,434)	(51,690)
Region B	176,283	176,342	176,397	176,822	176,294	179,998	216,445
	(19,039)	(18,869)	(18,874)	(18,213)	(18,511)	(18,540)	(17,099)

It is seen that, in general, the answer to the question posed above is yes. The estimated standard errors of the survey estimates decrease as the degree of spatial smoothness of the weights increases (from left to right across the tables). However, as expected, the estimates themselves also increase in size, becoming more and more positively biased. Overall, the gain due to reduced standard error seems to cancel out the increase in bias, except for the heaviest spatial smoothing ($\alpha = 0.1$, $d = 0.005$). In this latter case the increase in bias outweighs the reduction in standard error. The choice $\alpha = 0.1$ and $d = 0.05$ seems a good compromise, leading to reasonable (but not spectacular) bias-variance tradeoffs in Region A, and little change in the estimates in Region B.

4. ESTIMATION AND MAPPING OF LOCAL AVERAGES

A survey data map is a two-dimensional surface which estimates the spatial mean function of the survey variable in the population. In practice, such a map is obtained by applying a nonparametric regression technique to the weighted unit record data obtained in the survey.

At ABARE, we use kernel regression (a nonparametric technique) to produce maps which show the spatial variation of the estimated spatial mean function surfaces of key survey variables. These surfaces are obtained by replacing the observed sample values of these variables by locally weighted averages. In addition, for each local average map, a

corresponding map is produced which shows an estimate of the local variability of the variable of interest. We give below a brief outline of the technique: for clarity of exposition we deal only with the univariate case. See Ruppert and Wand (1994), Wand and Jones (1995, p140), and the references therein, for discussion of the multivariate case.

We assume that the finite population is generated as an iid sample $\{(Z_i, Y_i), i = 1, ..., N\}$ from a super population where Y_i is the value of a response variable Y observed at location Z_i. We suppose that the observations follow the model

$$Y_i = m(Z_i) + \epsilon_i, \qquad i = 1, ..., N$$

where $m(z) = E(Y \mid Z = z)$ is the conditional mean of Y given Z, and the ϵ_i are independent random variables with zero mean and variance $\sigma^2(z)$. Suppose that the error terms ϵ_i are independent of the process by which the sample is selected, so that the sample values $\{(Z_i, Y_i), i = 1, ..., n\}$ follow the same model, and write f for the density of $Z_1, ..., Z_n$.

A natural choice for the local average at any point z is then the mean of the values of the response variable for those observations with locations close to z, since observations from points far away will tend to have very different mean values. The local average is defined as a weighted mean

$$\hat{m}(z) = n^{-1} \sum_{i=1}^{n} W_i(z) Y_i$$

where the weights $\{W_i(z)\}$ depend on the locations $\{Z_i\}$ of the sample observations, and $\hat{m}(z)$ estimates $m(z)$.

The weights are constructed using a function K known as the kernel, which is continuous, bounded, symmetric and integrates to one. Various weight sequences have been proposed: the traditional Nadaraya-Watson weights (Nadaraya 1964 and Watson 1964) are

$$W_i(z) = h^{-1} K\{(z - Z_i)/h\} \Big/ \left[(nh)^{-1} \sum_{j=1}^{n} K\{(z - Z_j)/h\} \right],$$

where h is a scale factor known as the bandwidth. The kernel function K gives an observation close to z relatively more influence on the local average at this location than it gives to an observation further from z.

Where observations are sparse, a fixed-bandwidth window may contain few points and the corresponding estimator may therefore have a very high variance. This may be avoided by using the k-nearest-neighbour method in which a different bandwidth is used at each estimation point z. The bandwidth at z is the distance to the k-th nearest neighbour of z, so that there are always exactly k points in the bandwidth window. Let h_k be the distance between z and its k-th nearest neighbour. The k-nearest-neighbour Nadaraya-Watson weights are

$$W_{ih_k}(z) = h_k^{-1} K\{(z - Z_i)/h_k\} \Big/ \left[(nh_k)^{-1} \sum_{j=1}^{n} K\{(z - Z_j)/h_k\} \right].$$

We show in Table 4 the asymptotic mean squared error (MSE) properties of the usual (fixed-bandwidth) and k-nearest-neighbour estimators as given in Härdle (1990, p. 46).

Table 4
Asymptotic bias and variance of Nadaraya-Watson estimators; $c_K = \int K^2(u)du$, $d_K = \int u^2 K(u)du$

	Fixed-bandwidth	k-nearest-neighbour
Bias	$h^2 \frac{(m''f + 2m'f')(x)}{2f(x)} d_K$	$\left(\frac{k}{n}\right)^2 \frac{(m''f + 2m'f')(x)}{8f^3(x)} d_K$
Variance	$\frac{\sigma^2(x)}{nhf(x)} c_K$	$\frac{2\sigma^2(x)}{k} c_K$

Clearly, the bias of the estimated regression function can be reduced by using a smaller bandwidth h (number of nearest-neighbours k), but this leads to a noisy estimate $\hat{m}$ with local detail masking global features of the curve ($\hat{m}$ has high variance). If $h(k)$ is large, $\hat{m}$ is smoother but the global features are dampened ($\hat{m}$ has high bias and low variance). The bias, then, can only be reduced at the expense of variance and vice versa, with the bandwidth h determining the ratio of (squared) bias to variance.

In reality, the survey design and the spatial distribution of a survey variable Y will not be independent, so simple local averages for Y derived from the sample data will be misleading as estimates of the local population means of this variable. To overcome this problem the kernel weights are multiplied by the survey weights to get the final smoothing weights used for calculating the local average. This is equivalent to estimating the local population mean $m(z)$ of Y under the assumption that it is locally linear in the same benchmark variables as those used to model the overall population mean of Y.

A wide array of alternative kernel smoothing procedures have been discussed in the literature. As well as various sequences of smoothing weights $\{W_i\}$, there are different types of bandwidths, and several automatic bandwidth selection methods. A simulation study was therefore conducted to determine the most appropriate kernel methodology for use in ABARE's maps. This is described in the Appendix.

Uncertainty about the estimate of the spatial mean derived via kernel-based spatial smoothing can be represented by mapping the local variability of the variable of interest. Areas of high local variability correspond to areas where the map of the mean function is less precise and vice versa for areas of low local variability.

The usual method of determining confidence regions for a kernel curve estimate is the bootstrap; see Härdle (1990), Hall (1992), and references therein. However, for computational efficiency, we use the expectiles (Newey and Powell 1987) of the spatial distribution of Y to describe this

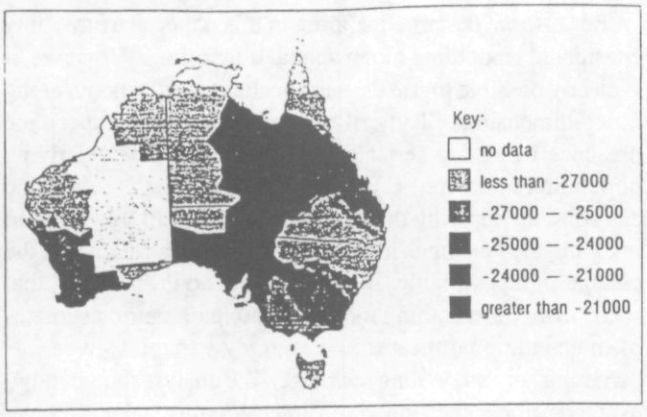


Figure 5. Polygon map of farm business profit in 1991-1992, all broadacre farm ($)

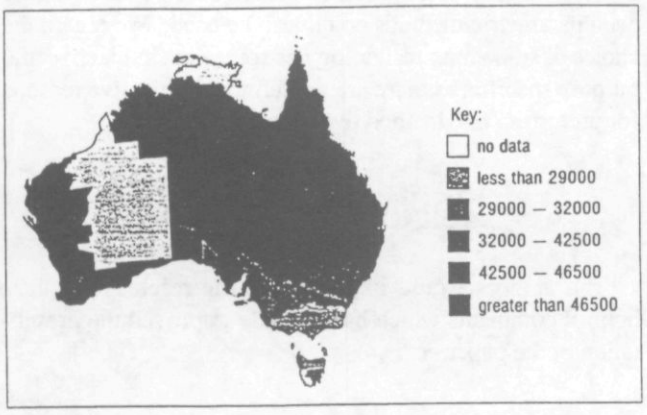


Figure 6. Polygon map of interexpectile range of farm business profit in 1991-1992, all broadacre farms ($)

local variability. An expectile bears the same relationship to the mean as the corresponding quantile does to the median. In particular, the difference between the 75th and 25th expectiles of a distribution is a measure of the spread of the distribution in the same way as the interquartile range is a measure of this spread. The smoothing program contains a module for non-parametric M-quantile regression (Breckling and Chambers 1988) which is used to fit a smooth surface to the expectiles of the Y-distribution at any location. The difference between the smoothed 75th and 25th expectile surfaces (the smooth expectile analogue of the interquartile range) is then mapped to show areas of high and low variability in the data.

Not surprisingly, this smooth interexpectile range tends to be highest in areas where the farms are sparsely located and the farm-to-farm variability in Y is therefore highest. The interexpectile range map corresponding to Figure 5 is shown in Figure 6. Note that these smoothed interexpectile range maps provide similar information to confidence bands at any particular point on the map. However, they do not have the same repeated sampling interpretation as confidence intervals, and hence should be treated as guides to, rather than measures of, the uncertainty associated with a particular map.

For confidentiality reasons, care must be taken when mapping the smoothed data for publication to ensure that the locations of the surveyed farms are not revealed. Another requirement is output quality compatible with desktop publication packages. Two procedures for generating the final maps that satisfy these requirements have been developed using ARC/INFO.

In the first method, a Thiessen polygon is constructed around each farm. The polygon defines the area closer to that farm than to any other farm. The farm location is not in the centre of its polygon, and the polygon shape does not resemble the shape of the farm, so the polygons conceal the locations of the survey farms, as shown in Figure 7. The whole of each polygon is coloured according to the smoothed value of Y at the farm location in that polygon. Usually ten colours are used in each map and the estimated population deciles of the smoothed data are used as boundaries for the colour area. The maps shown in this paper are black-and-white analogues of these colour maps.

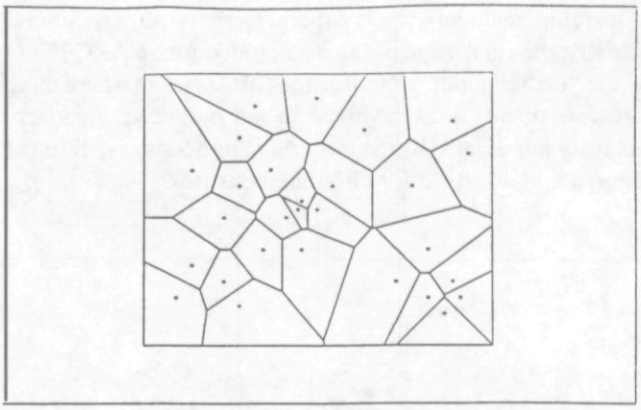


Figure 7. Thiessen polygons constructed around selected ABARE survey farms. Farm location is shown as a small square within each polygon

In the second method, smoothed values on a dense rectangular grid are used in place of smoothed values at the farm locations, and a further minor interpolation of the data is carried out in ARC/INFO. A continuous 3-dimensional surface which passes through the smoothed values at the grid points is built in two steps. As a first approximation, a faceted surface of triangles obtained by Delauney triangulation is constructed, and then a bivariate fifth degree polynomial is fitted within each triangle using Akima's algorithm (Akima 1978). The resulting continuous surface is then contoured using the estimated population deciles. Figure 8 is an example.

In this second method of presentation, the locations of the survey farms are not used in any way, thereby completely concealing the location of each survey farm. It also gives smooth contours, and the result is not as patchy as the polygon based map. Moreover, it is preferred by ABARE's graphics staff because it reduces the number of areas to be

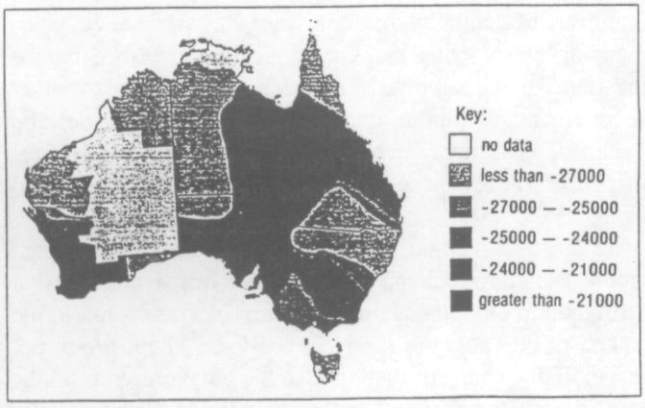


Figure 8. Contour map of farm business profit in 1991-1992, all broadacre farms ($)

separately coloured and has lower storage requirements, enabling the maps to be more readily manipulated in desktop publishing packages. Its disadvantage is that it uses more computing time in the ARC/INFO stage.

Since the above procedures interpolate across all of Australia, including areas where there is no agricultural activity, the final stage of the map production in ARC/INFO is the "blanking out" of those areas of Australia where there are few or no farms involved in the particular broadacre industry represented by the map. As Figure 9 shows, different areas are blanked out for different industries.

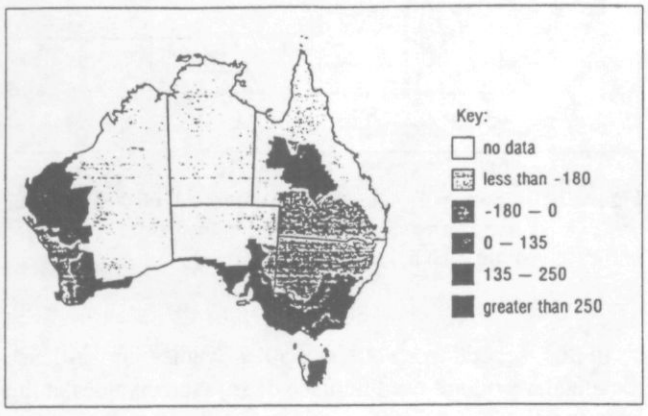


Figure 9. Polygon map showing expected change in wool production, 1991-92 to 1992-93, farms with 100 or more sheep in 1991-92 (kg)

5. DISCUSSION

In this paper we have demonstrated that when survey data has a spatial dimension, as in the case of the AAGIS, spatial smoothness concepts may be useful to the analyst. The concept can be used to modify survey weights to ensure less variable small area survey estimates. It may also be used to smooth the data along spatial dimensions before mapping the spatial mean function.

Because we describe mapping in this paper, we have only considered smoothing along spatial dimensions. However, it is clearly possible to use the same techniques to smooth along other dimensions. Thus, if there is reason to expect the presence of strong serial correlation when the underlying population is ordered according to some variable, then one can consider applying the methods described in this paper to mapping the "change" in the survey variables relative to the change in this variable. In doing so, it should be noted that such "maps" are nothing more than nonparametric estimates of the conditional means of the survey variables given this "ordering" or "smoothing" variable. The analyst should, however, remember the "curse of dimensionality": the effective sample size drops sharply with each additional smoothing variable used in these nonparametric techniques.

Finally, in mapping the survey data, we have used kernel-based estimation techniques. However, spline smoothing, or even parametric methods could also be used. We regard the choice of smoothing technology as somewhat subjective and purpose specific, as there are no definitive objective reasons for preferring one method over another.

ACKNOWLEDGEMENTS

The authors would like to thank the referees for their helpful comments which have greatly improved the presentation of the paper.

APPENDIX

In the last few years a number of optimality properties have been established for the locally-linear kernel weights (see for example Wand and Jones (1995) and references therein). We therefore compared Nadaraya-Watson (NW) and locally-linear (LL) weight sequences using fixed (FBW) and k-nearest-neighbour (NN) bandwidths with each weight sequence. For each of these combinations, we selected the bandwidth using least-squares cross-validation (CV), and an *ad hoc* method (detailed in the last paragraph of this section) aimed at reducing the speckledness of a map (SF).

Two criteria were used to evaluate the performance of each methodology. The first, MSE, is the obvious statistical criterion for assessing a biased estimator. The second criterion is more ABARE specific. As estimates are produced both in tables (by State) and in maps, the impression of the state average given by the map should be close to the tabulated value. We therefore used a weighted sum of the squared differences between the state averages of the raw and smoothed survey data (SB^2). This measure was also calculated at regional rather than state level (RB^2; there are between one and nine regions in each state).

Data were generated at the survey farm locations using three smooth functions with varying degrees of smoothness (measured by $\int m''$) and normal mixture errors. For example,

$$m_1(z) = 6.25 \times 10^4 \times \cos\left\{\frac{z_1 - 132.5}{2.25}\right\} \cos\left\{\frac{z_2 + 27.5}{1.75}\right\},$$

where z_1 and z_2 are the longitude and latitude of the point z. The functions $m_i(z)$ were scaled to have the same range as the smoothed values of a key survey variable, and the errors were scaled to have the same range as the residuals of the same variable after smoothing. Large variances were generated at locations with high residuals, and small variances at locations with low residuals. The simulation results based on the smooth function are given in Table 5.

Using MSE as the criterion for assessing methodology, the results were not consistent for the three functions $m_i(z)$. However, when either RB^2 or SB^2 was used as the performance measure, the LL estimator with k-nearest-neighbour bandwidth selected using SF outperformed the other methods by at least ten percent for each function $m_i(z)$, and is therefore the currently preferred methodology for producing ABARE's maps.

Table 5
Comparison of locally-linear (LL) and Nadaraya-Watson (NW) weight sequences, using fixed (FBW) and k-nearest-neighbour (NN) bandwidths selected by least-squares cross-validation (CV) and the criterion detailed below (SF). The results were obtained from 400 independent samples with mean function and normal mixture errors. The MSE values were calculated using the average over the finite population of $(y - \hat{m}(z))^2$

		$MSE \times 10^{-7}$		$RB^2 \times 10^{-7}$		$SB^2 \times 10^{-7}$	
		CV	SF	CV	SF	CV	SF
LL	FBW	39.64	93.93	4.44	1.67	1.33	0.39
	NN	20.50	22.83	2.22	1.35	0.37	0.14
NW	FBW	41.91	52.78	3.29	1.77	0.34	0.17
	NN	21.77	22.22	3.03	2.33	0.62	0.41

The bandwidth selection method aimed at reducing the speckledness of a map (SF) is a measure of the smoothness of the map: it measures how similar the smoothed value is at any farm to that of its neighbours. Let $p(i)$ be the survey estimate of the percentile of the smoothed variable at the i-th farm. Let S_i be the set of indices of the six farms closest to the i-th farm. In this method, the value of

$$SF(h) = (6n)^{-1} \sum_i \sum_{k \in S_i} |p(i) - p(k)|$$

is calculated. It is scale-free, and decreases monotonically as the bandwidth decreases. The chosen bandwidth is the smallest bandwidth with a sufficiently small ($< \epsilon$) rate of decrease of SF. The value of ϵ was chosen subjectively following detailed examination of maps of five key variables for five values of ϵ.

REFERENCES

BANKIER, M.D., RATHWELL, S., and MAJKOWSKI, M. (1992). Two step generalised least squares estimation in the 1991 Canadian Census. *Proceedings of the Workshop on Uses of Auxiliary Information in Surveys*. Statistics Sweden, Örebro, October 5-7.

BARDSLEY, P., and CHAMBERS, R.L. (1984). Multipurpose estimation from unbalanced samples. *Applied Statistics*, 33, 290-299.

BRECKLING, J., and CHAMBERS, R.L. (1988). *M*-quantiles. *Biometrika*, 75, 761-771.

DEVILLE, J.-C., and SÄRNDAL, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.

FULLER, W.A., LOUGHIN, M.M., and BAKER, H.D. (1994). Regression weighting in the presence of nonresponse with application to the 1987-1988 Nationwide Food Consumption Survey. *Survey Methodology*, 20, 75-85.

HALL, P. (1992). *The Bootstrap and Edgeworth Expansion*. New York: Springer-Verlag.

HÄRDLE, W. (1990). *Applied Nonparametric Regression*. Cambridge: Cambridge University Press.

NADARAYA, E.A. (1964). On estimating regression. *Theory of Probability and its Applications*, 10, 186-190.

NEWEY, W.K., and POWELL, J.L. (1987). Asymmetric least squares estimation and testing. *Econometrica*, 55, 819-847.

RUPPERT, D., and WAND, M.P. (1994). Multivariate locally weighted least squares regression. *Annals of Statistics*, 22, to appear.

WAND, M.P., and JONES, M.C. (1995). *Kernel Smoothing*. London: Chapman and Hall.

WATSON, G.S. (1964). Smooth regression analysis. *Sankhyā*, Series A, 26, 101-116.

Survey Methodology, December 1996
Vol. 22, No. 2, pp. 185-197
Statistics Canada

Using Data on Interruptions in Telephone Service as Coverage Adjustments

J. MICHAEL BRICK, JOSEPH WAKSBERG and SCOTT KEETER[1]

ABSTRACT

Telephone surveys in the U.S. are subject to coverage bias because about 6 percent of all households do not have a telephone at any particular point in time. The bias resulting from this undercoverage can be important since those who do not have a telephone are generally poorer and have other characteristics that differ from the telephone population. Poststratification and the other usual methods of adjustment often do not fully compensate for this bias. This research examines a procedure for adjusting the survey estimates based on the observation that some households have a telephone for only part of the year, often due to economic circumstances. By collecting data on interruptions in telephone service in the past year, statistical adjustments of the estimates can be made which may reduce the bias in the estimates but which at the same time increase variances because of greater variability in weights. This paper considers a method of adjustment using data collected from a national telephone survey. Estimates of the reductions in bias and the effect on the mean square error of the estimates are computed for a variety of statistics. The results show that when the estimates from the survey are highly related to economic conditions the telephone interruption adjustment procedure can improve the mean square error of the estimates.

KEY WORDS: Coverage; Bias; Weighting adjustment; Telephone sampling; RDD surveys.

1. INTRODUCTION

Telephone surveys provide a relatively economical method of data collection compared with face-to-face interviewing. However, telephone surveys in the U.S. are subject to an important source of bias that does not affect household surveys conducted with face-to-face interviewing: at present only 94 percent of households nationally have telephone service at any given time. Moreover, for some populations such as households with young children, coverage rates are even lower.

Weighting that includes poststratification based on demographic variables known to be associated with telephone coverage is effective in mitigating some of the consequences of coverage bias in telephone surveys. Postsurvey weighting is also generally used to compensate for nonresponse and other biases. But even when effective, weighting to known demographic totals only partially solves the problem of coverage bias, undercompensating for some variables (Massey and Botman 1988) and overcompensating for others (Brick, Burke, and West 1992).

This article describes a study of an alternative method for adjusting telephone survey data to compensate for coverage bias. The method, suggested by Keeter (1995), is based on the observation that telephone subscription is a dynamic condition not just across households in the population, but also within many households over time. A sizable number of U.S. households lose and gain telephone status during a given year. Because of this phenomenon, the telephone population at a given time includes households that have recently been in the nontelephone population. Despite considerable information on the size and characteristics of the nontelephone population, little is known about its dynamics over shorter time periods. Evidence from social workers, telephone companies, and others who deal with indigent households suggests that for many families, telephone subscription is episodic. Households may have a telephone when they can afford it, but the telephone may be turned off when times are harder, or when the bills get too large to manage, (Federal Communications Commission 1988). It is not known how many households change their telephone status and how long they stay in a particular status.

Keeter (1995) examined two household panel surveys to obtain estimates of the dynamics of telephone service subscription. Those households that changed telephone status (presence of a telephone in the household) are called 'transient' households. For data from one panel survey that collected data 12 months apart, half of the 6 percent of all households without a telephone at either time were transient. For the other panel survey in which data were collected only two months apart, one-fourth of the 6 percent of households without telephones at either point in time were transient. Since these estimates were based on observations at two points in time rather than continuous measurement, they underestimate the percent of households that are transient. Nevertheless, these results show that a substantial proportion of households without a telephone at a specific point in time is transient.

Another important condition that must be satisfied if the transient telephone households are to be useful in reducing

[1] J. Michael Brick and Joseph Waksberg, Westat, Inc., 1650 Research Blvd., Rockville, MD 20850, U.S.A.; Scott Keeter, Virginia Commonwealth University, Survey Research Laboratory, Richmond, VA 23284, U.S.A.

coverage bias involves the characteristics of transient households and nontelephone households. If the two groups are not similar, then the adjustments will not be effective. Using the panel data and data from several Virginia surveys, Keeter (1995) showed that the characteristics of the transient households are much more consistent with nontelephone households than telephone households.

These findings suggest the possibility that weighting adjustments that use the data from households that have telephones only sometimes during the year might be an improvement over the current practice. To evaluate this approach to adjusting the weights, questions were added to two national surveys conducted in 1993 by Westat. Both of these surveys were random digit dial (RDD) and computer assisted telephone surveys, and the data were collected in the telephone research centers of Westat.

One of the surveys is the National Household Education Survey of 1993 (NHES:93). The NHES:93 was conducted for the National Center for Education Statistics of the Department of Education in the spring of 1993 to study issues related to school readiness of young children and school safety and discipline of children in school. The other survey was the National Survey of Veterans (NSV) which was conducted in the second half of 1993 for the U.S. Department of Veterans Affairs. In this survey, adults were screened to determine if they were veterans, and the veterans were then asked about a variety of topics including their health, education, and financial status.

Below, we present estimates of the percentage of persons that experienced some interruption of telephone service, describe procedures for adjusting the survey weights using these data, and discuss the statistical implications of using the adjusted weights. The final section summarizes the findings and gives some considerations for using this technique in RDD telephone surveys.

2. ESTIMATES OF INTERRUPTIONS OF TELEPHONE SERVICE

Estimates of the percentage of persons with interruptions of telephone service from national surveys were needed to further examine the potential of reducing coverage biases using these data. Questions were added to the NSV and the NHES:93 for this purpose. In the NSV, about 23,000 households were screened and interviews were completed with over 5,500 eligible veterans. In the screening interview, all household members 14 years and over were enumerated and questions were asked about their characteristics and their veteran status. If a sampled adult was a veteran, then a more detailed interview was attempted. The results reported here are those asked about the adults enumerated in the screening interview which included only a few characteristics of the adults and the household.

In the NHES:93, 64,000 households were screened and nearly 30,000 interviews were conducted within those screened households. Two survey components were included: School Readiness (SR) and School Safety and Discipline (SS&D). Approximately 11,000 parents of 3- to 7-year-olds completed interviews on SR topics and about 12,700 parents of children in grades 3 through 12 were interviewed for the SS&D component. Data on interruptions in telephone service were collected from households in which at least one SR or SS&D interview was completed.

Since the responses to the questions in the NHES:93 were only obtained for those households that completed either an SR or SS&D interview, many characteristics of the children can be analyzed, but the data do not apply to as broad a population as the NSV. The NSV applies to all adults, but only limited data were collected on most of the adults. For all households that had completed an interview (a screening interview in the NSV and a more detailed interview in the NHES:93), a member of the household was asked if the household had experienced an interruption in telephone service in the last 12 months and how long it lasted.

Estimated Service Interruptions in the NSV and NHES:93

The estimated percentage of persons in households that had a telephone interruption of one day or more during the last 12 months varies substantially from survey to survey. Only 2.3 percent of adults had an interruption of one day or more based on the data from the NSV, while the percentage from the NHES:93 for younger children (the SR population of 3- to 7-year-olds) was 12.0 percent, and for the SS&D population of older children (grade 3 through 12) it was 9.2 percent.

Figure 1 shows estimates and 95 percent confidence intervals of the percentage of persons that had interruptions of one day or more along with estimates for those with interruptions of telephone service that lasted for at least one week and at least 4 weeks. While the percentages vary from sample to sample, the patterns of increase by length of interruption are relatively stable. The percentage with interruptions of one week or longer is less than half the percentage with any interruption, and the percentage with interruptions of 4 weeks or more is about one-fourth the percentage with any interruption.

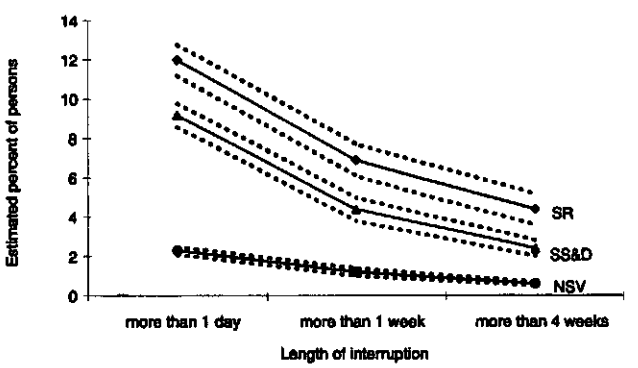


Figure 1. Estimated percentage of persons with interrupted telephone service from the three populations

The large difference in the estimates from the NSV and the NHES:93 comes from at least two important sources. The first source is that the populations were different. We would expect young children to live in households that experience more interruptions than older children and adults. Thornberry and Massey (1988) estimated that the telephone coverage rate for young children was lower than for any other age group. Thus, the difference of about 3 percent in the estimates of the percentage with an interruption between the younger (SR) and older (SS&D) children from the NHES:93 is reasonable.

The difference in the populations does not completely account for the large difference between the NSV and the NHES:93 estimates. An important reason for this difference is related to the way the questions were asked in the two surveys. The NHES:93 interview began by asking, "During the past 12 months, has your household ever been without telephone service for more than 24 hours?". In the NSV interview, respondents were asked if, "At any time during the past 12 months, has your household *not* had telephone service?". This was followed by a question that asked if the interruption was for at least 24 hours. Thus, the NSV version was a screening item followed by a more detailed question. This type of construction often depresses reports of subsequent activities, which is consistent with the lower NSV estimates.

A more important reason for the difference is probably due to the wording of the questions. With the NSV question, a 'no' response may have confused respondents because the question asks if they did *not* have telephone service. Converse and Presser (1986) discuss the problems that arise with this type of question construction. The wording for the NHES:93 is less confusing. The combination of the wording and the use of a screening item in the NSV is likely to be the main reason for the smaller estimate using the NSV questionnaire.

The difference in the estimates associated with the different ways of asking the interruption questions is evident from the estimates from two surveys conducted in Virginia by Virginia Community University. In a November 1993 survey, the items about telephone interruptions were asked using the NSV wording; in April 1994 the items were changed to the NHES:93 wording. The results from the surveys parallel the differences in the estimates between the NSV and the NHES:93. The November 1993 Virginia study estimated that 3 percent had an interruption in service in the last 12 months, while in April the estimated percentage was 9 percent. Thus, it is clear that the different ways of asking the questions heavily influenced the size of the estimates, and it suggests that the estimates from the NSV are biased downward. Some adults who did experience an interruption in telephone service during the previous 12 months probably responded incorrectly in the NSV.

Characteristics of Persons With Service Interruptions

Estimates of the percentage of persons who had a telephone interruption are examined below by the characteristics of the person to evaluate the potential of using these data to adjust for coverage bias. We estimated the percentage of persons in households with any interruption in service by characteristics collected in both the NSV and the NHES:93. These estimates are shown in the first part of table 1. Some differences in the distributions may be due to the different ways of asking the questions. For example, the education classification is different in the two surveys: in the NSV education is recorded for the oldest person in the household, while in the NHES:93 education is the highest for either of the parents of the child.

All subsequent analysis is restricted to NHES:93 data for two reasons. First, more data on the characteristics are available from the NHES:93 detailed SR and SS&D interviews than the NSV screening interview. Second, the telephone interruption estimate from the NSV is biased due to the wording of the item, as discussed earlier. Of course, the NHES:93 estimates apply to households with children which have higher nontelephone rates than the general population, and in that sense they do not reflect the situation for the total population.

Using the NHES:93 data, we find that the percents of persons with some interruption are relatively consistent for the SR and the SS&D populations (see table 1). The characteristics generally associated with lower economic status have the highest percentage with interruptions. For example, the percentage of children with interruptions in both the SR and SS&D populations is larger for those from households with lower household income than for those from households with higher income. Similarly, children participating in public assistance programs (WIC or free meals) have much higher rates of service interruptions than nonparticipants. However, the percentages of children in households with telephone interruptions are less variable for characteristics related to school readiness and school safety and discipline than for the socioeconomic items. Additional characteristics for both populations were examined and presented in Brick, Keeter, Waksberg and Bell (1996), but are not shown here. For most of the other substantive items, the differences in the percentage of persons with some interruption in telephone service were either not statistically significant or not large enough to be of great practical importance.

3. WEIGHT ADJUSTMENTS

In almost all sample surveys, the data collected from respondents are adjusted to account for nonresponse and noncoverage and to reduce the variability in the estimates by using auxiliary data from other data sources. One of the most important benefits of this type of adjustment in telephone samples is that it often reduces the bias associated with the undercoverage of persons living in households without telephones.

Kalton and Kasprzyk (1986) discuss adjustments to the base weights, classifying the adjustments into four categories: population weighting adjustments, sample weighting adjustments, raking ratio adjustments, and response probability

Table 1
Estimated Percentage of Persons With Any Interruptions in Telephone Service in Last 12 Months for Three Populations

	NSV		NHES:93 SR		NHES:93 SS&D	
	Estimate	Standard error	Estimate	Standard error	Estimate	Standard error
Total	2.3	0.1	12.0	0.4	9.2	0.3
Region						
Midwest	2.3	0.2	11.0	1.0	7.3	0.7
Northeast	2.0	0.2	9.5	1.2	9.0	0.8
South	2.6	0.2	13.6	0.7	10.8	0.6
West	2.4	0.2	12.5	0.9	9.2	0.8
Race/ethnicity[1]						
White	2.0	0.1	9.3	0.5	7.2	0.3
Black	3.5	0.4	19.8	1.5	14.7	1.1
Hispanic	3.9	0.5	17.2	1.5	14.1	1.1
Other	2.6	0.6	11.7	2.6	9.3	1.5
Education[2]						
Less than high school diploma	3.2	0.2	18.4	1.8	17.4	1.6
High school graduate	2.0	0.2	15.4	0.8	11.0	0.8
Some college	2.3	0.2	11.8	0.7	8.6	0.5
Bachelor's degree	1.6	0.2	5.5	0.8	5.3	0.8
Graduate school	2.2	0.3	5.2	0.7	4.5	0.6
Household income						
$10,000 or less			22.8	1.3	19.0	1.3
$10,001 to $20,000			19.9	1.4	15.7	1.1
$20,001 to $30,000			9.3	0.8	7.9	0.6
More than $30,000			5.5	0.5	5.0	0.3
Women, infant and children program participant[3]						
Yes			18.2	1.3		
No			8.0	0.6		
Free meal at school or center[4]						
Yes			21.1	1.2		
No			7.6	0.5		
Birth weight						
5.5 pounds or less			12.0	1.6		
Greater than 5.5 pounds			12.0	0.4		
School control						
Public					9.4	0.4
Private					7.5	1.1
Ease of obtaining marijuana at school[5]						
Very or fairly easy					9.7	0.6
Hard					8.0	0.8
Nearly impossible					9.0	0.7

[1] Race/ethnicity is reported for the oldest member in the NSV and for the child in the NHES:93.
[2] Education is for the oldest household member in the NSV and the most educated parent of the child in the NHES:93.
[3] Estimate restricted to preschoolers.
[4] Estimate applies to children except preschoolers.
[5] Estimate applies only to children in grades 6 through 12.
Source: U.S. Department of Veterans Affairs, National Survey of Veterans, summer/fall 1993, and U.S. Department of Education, National Household Education Survey, spring 1993.

adjustments. In the NHES:93, sample weighting adjustments and raking ratio adjustments were used. Sample weighting adjustments were used to account for differential nonresponse from sampled persons. Raking ratio adjustments were then used to make the specified marginal distributions of the sample correspond to totals from the October 1992 Current Population Survey (CPS). One of the most important benefits of the type of raking ratio adjustment used in the NHES:93 is that it reduces the bias associated with the undercoverage of persons living in households without telephones because the CPS covers persons in both telephone and nontelephone households.

The data on telephone service interruptions can be used to make a response probability adjustment. Response probability adjustments are constructed by assuming that each sampled unit has a probability of responding to the survey, estimating that probability, and then using the inverse of the estimated response probability as a weighting adjustment. The Politz and Simmons (1949) method is probably the best known application of the response probability adjustment procedure, and Kalton and Kasprzyk (1986) discuss others.

To apply this type of adjustment using the telephone service interruption data, assume that living in a telephone household is a dynamic phenomenon and that a probability distribution can be associated with this status. Conceptually, a survey is conducted by sampling from this distribution and observing only those members that live in telephone households at the time of the survey. The probability of living in a telephone household (the equivalent of the response probability) must then be estimated for each respondent. The inverse of the estimated probability is the coverage adjustment. This model assumes that each person can be assigned a probability of being in a household with a telephone and that the probability is between zero and one (but not equal to zero).

The data on whether or not a household had an interruption in telephone service and the length of that interruption are the basis for this type of adjustment. Persons are divided into two categories: those in households with interruptions in service and those in households without interruptions in service. The probability is assumed to be equal to one for persons in households without interruptions and their weights are not adjusted. The weights of persons in households with at least some interruptions in the last 12 months are adjusted to account for other households that have a probability of being covered of less than one. The adjustments may vary depending on the length of time they lived in nontelephone households and on other characteristics of the household. The purpose of having different adjustments is to account for the fact that some persons are more likely to live in nontelephone households than others.

Although the weighting adjustments may reduce the undercoverage bias, introducing adjustments also typically increases the variances of the estimates. Kish (1992) discusses the reasons for unequal weights as well as the consequences from using them in a variety of situations. He advocates a common statistical approach of balancing the bias reductions against the variance increases. If the weights reduce the bias of the estimates significantly, then it may be worthwhile accepting the variance increases. On the other hand, small reductions in bias associated with large variance increases are not recommended.

In the remainder of this section, the specific weighting adjustment procedures are described. The statistical properties of the weights developed under four alternative adjustment schemes are presented. The alternative weights are applied to the NHES:93 data and the decrease in the bias of the estimates is compared with the increase in the variance of the estimates due to the unequal weighting.

Adjustment Schemes

The first step was to decide how to classify the length of interruption in telephone service. Various lengths of interruptions were examined to determine cut-offs that discriminated between temporary interruptions, not due to economic causes and others. It was decided to use two categories for forming adjustment cells: one week or more, and one month or more.

Within each of the length-of-service interruption categories, the children were classified into adjustment cells based on either parental education or tenure (home ownership). Race/ethnicity was used to form cells within the parental education and tenure categories. These cells were chosen because the percentage of persons with interruptions varied by these characteristics and the corresponding data were also available from the CPS. Four adjustment schemes were defined using these items:

Scheme A1 – children in households that had a telephone service interruption of one week or more within categories defined by parental education (less than high school, high school diploma, college diploma or above) and race/ethnicity (Hispanic, black/non-Hispanic, white and other/non-Hispanic);

Scheme A2 – children in households that had a telephone service interruption of one month or more within categories defined by parental education and race/ethnicity;

Scheme B1 – children in households that had a telephone service interruption of one week or more within categories defined by tenure (own/other, rent) and race/ethnicity; and

Scheme B2 – children in households that had a telephone service interruption of one month or more within categories defined by tenure and race/ethnicity.

The adjustment factors for these schemes could not be obtained directly from the NHES:93 data because no data were collected from households without telephones. Instead, the adjustments were developed using both CPS and NHES:93 data and then applied to the NHES:93 weights.

To motivate the adjustment of the weights, consider partitioning the universe of persons into four components: t_1 is the number of persons in *telephone* households with *no telephone interruptions* in the past year; t_2 is the number of persons in *telephone* households with *some telephone interruptions* in the past year; t_3 is the number of persons in *nontelephone* households with *no telephone interruptions* in

the past year (*i.e.*, persons who lived in nontelephone households throughout the entire year); and t_4 is the number of persons in *nontelephone* households with *some telephone interruptions* in the past year. As noted above, the response probability model assumes $t_3 = 0$.

Using the CPS it is possible to estimate $t_1 + t_2$ and t_4 (assuming $t_3 = 0$); designate these estimates as $\hat{t}_1 + \hat{t}_2$ and $\hat{t}_4$, respectively. From the NHES:93, t_1 and t_2 can be estimated separately; call these estimates t_1^* and t_2^*, respectively. The bias in the NHES:93 estimates arises because they are from a telephone survey and do not include persons in nontelephone households (t_4).

A weight adjustment of $A = 1 + t_4/t_2$ would result in unbiased estimates of totals; however, this adjustment involves unknown, population quantities that must be estimated. Since t_2 can only be estimated from the NHES:93 and t_4 can only be estimated from the CPS (assuming $t_3 = 0$), the adjustment is expressed in ratios to reduce the bias due to estimating the totals from different surveys. The revised weight is

$$w_i' = w_i \left(1 + \delta_i \frac{\dfrac{\hat{t}_4}{\hat{t}_1 + \hat{t}_2}}{\dfrac{t_2^*}{t_1^* + t_2^*}} \right), \qquad (1)$$

where w_i is the NHES:93 weight adjusted for nonresponse of sampled persons but not yet raked to October 1992 CPS totals, $\delta_i = 1$ if the person lives in a household that had an interruption of telephone service in the last year and is zero otherwise. The quantity in parenthesis in (1) is an estimate of A, the weight adjustment.

Revised weights were computed separately for the SR and SS&D components. Rather than the overall adjustment as given in (1), the weight adjustments were computed within the cells defined for each of the four weighting schemes (A1, A2, B1, and B2). Table 2 shows the resulting adjustment factors for the SR and SS&D components. The adjustments in the first column are those for schemes A1 and B1. The second column contains the adjustment factors for schemes A2 and B2. The adjustment factors for the schemes based on the one month or more interruptions are greater than those based on the one week or more because the denominator of the ratio is, by definition, smaller for this classification (see Figure 1 for estimates of the percentage of persons with interruptions for each scheme).

The last weighting step rakes the four alternative weights to the same October 1992 CPS totals used in raking the standard NHES:93 person-level weights. The result of this process is the standard NHES:93 weight and four alternative weights based on different adjustment schemes. All five of the weights conform to the same marginal totals. The only difference in the weights is the adjustment for the telephone

Table 2
Weighting Cell Adjustments Factors, Based on Length of Interruption of Telephone Service

Factor	SR		SS&D	
	Length of service interruption			
	One week or more	One month or more	One week or more	One month or more
Cells defined by parental education and race/ethnicity (Schemes A1 and A2)				
Less than high school; Hispanic	5.75	16.35	4.89	8.52
Less than high school; black, non-Hispanic	5.10	6.72	4.26	5.95
Less than high school; white and other, non-Hispanic	4.98	5.37	3.81	4.86
High school diploma; Hispanic	2.31	2.76	2.67	4.51
High school diploma; black, non-Hispanic	2.65	3.73	3.06	4.71
High school diploma; white and other, non-Hispanic	2.16	2.79	2.18	3.09
College degree or more; Hispanic	1.34	2.33	1.96	8.22
College degree or more; black, non-Hispanic	1.77	2.64	1.35	8.83
College degree or more; white and other, non-Hispanic	1.58	2.09	1.91	3.48
Cells defined by tenure and race/ethnicity (Schemes B1 and B2)				
Renter; Hispanic	3.74	5.15	3.58	6.08
Renter; black, non-Hispanic	3.23	4.54	3.38	4.95
Renter; white and other, non-Hispanic	2.43	2.96	2.99	4.00
Owner/other; Hispanic	2.00	3.06	2.81	5.66
Owner/other; black, non-Hispanic	2.53	3.46	2.90	6.11
Owner/other; white and other, non-Hispanic	2.26	3.45	2.03	3.10

service interruption prior to raking. The standard weights are not further adjusted while the alternative weights have different adjustments depending on the scheme.

4. FINDINGS

As noted above, adjustment of the weights to reduce the bias increases the variability of the weights, thus increasing the variance of the estimates. Kish (1992) gives an approximate expression for this increase in variance arising from unequal weights. We call this expression for the increase in variance due to differential weights the variance inflation factor (*VIF*). The *VIF* can be written as

$$VIF = 1 + \mathrm{CV}^2 \text{ (weights)} \qquad (2)$$

where CV is the coefficient of variation of the weights.

Table 3 shows the *VIF* for the standard NHES:93 weights for each component. The SS&D component is broken down by the grade of the student, because youth were selected at different rates for these grade levels. The *VIF* for each of the components is about 1.4, indicating the variance is inflated by about 40 percent due to the variability in the standard weights. The *VIF* for the combined SS&D file is somewhat larger (1.5) because it includes youth who were sampled at different rates.

The other factors given in table 3 are the ratios of the *VIF* for the four alternative weights to the *VIF* for the standard weight. These ratios show how much greater the variances of estimates produced using the alternative weights are expected to be as compared to the variances of the standard NHES:93 weights.

Overall, the increase in variance due to the telephone interruption coverage adjustment are from 9 to 13 percent for schemes A1 and B1 in the SS&D component but up to 20 percent for the SR component. The ratios are larger for the schemes A2 and B2, ranging from 24 to 35 percent, with the largest ratio for Scheme A2 for the SR component. The larger ratios (hence *VIF*s) for the schemes based on interruptions of one month or more are a consequence of the larger and more variable factors shown in the second column of table 2. The ratios for the SR population are higher than the SS&D ratios.

4.1 Coverage Bias Reduction

If estimates of the same characteristics as those produced from the NHES:93 were available from an independent source and these benchmark estimates were free of telephone coverage bias, then it would be possible to compare the five estimates to the benchmark. However, benchmarks comparable to the estimates from the two components of the NHES:93 do not exist and other methods are needed to assess the bias-reducing potential of the coverage adjustments.

Due to of the lack of a benchmark, some model assumptions are required to assess the effectiveness of the adjustments. For this evaluation we assume that the adjustment procedures reduce the coverage bias. As a result of this assumption, the difference between the standard estimate and the adjusted estimate is considered an unbiased estimate of the decrease in the coverage bias resulting from using the procedures. Clearly, the coverage bias is not completely eliminated by any of the adjustment procedures. Even if the model were correct, the bias reductions from the data would still be subject to sampling error. Despite the problems with this assumption, this type of assumption is necessary to obtain some idea of the effectiveness of the adjustment. If the adjustment eliminates the bias, the mean square errors of the adjusted estimates are equal to the variances of the estimates, with no contribution from coverage bias. Therefore, the model assumption is favorable to the adjusted estimates, positing the adjusted estimates to be unbiased. The impact of this assumption is discussed critically after evidence of the effectiveness of the method is presented.

The estimate from each scheme can be compared to the standard NHES:93 estimate, and the difference between the standard estimate and the adjusted estimate is an estimate of the reduction in the coverage bias. With four adjusted estimates, four different estimates of bias reduction are possible. The estimated reduction in bias is

$$b_a = \hat{p}_s - \hat{p}_a, \qquad (3)$$

where b_a is the estimated bias reduction using adjustment scheme a (a = A1, A2, B1, or B2), $\hat{p}_s$ is the estimate of the proportion using the standard estimate, and $\hat{p}_a$ is the estimated proportion using adjustment scheme a.

Table 3
Ratios of Variance Inflation Factor Due to Coverage Adjustment

Component	Sample size	*VIF** standard weight	Ratio of scheme's *VIF* to standard weight's *VIF*			
			Scheme A1	Scheme A2	Scheme B1	Scheme B2
School Readiness	10,888	1.36	1.20	1.35	1.16	1.26
School Safety and Discipline						
3rd through 5th graders	2,563	1.37	1.12	1.25	1.13	1.26
6th through 12th graders	10,117	1.39	1.13	1.27	1.09	1.24
3rd through 12th graders	12,680	1.49	1.12	1.26	1.11	1.25

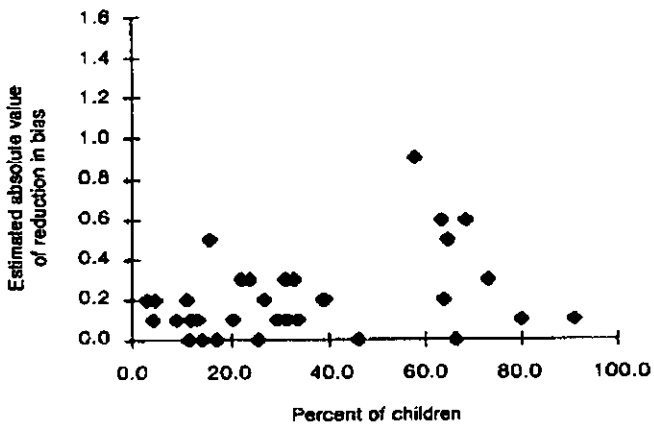
* *VIF* is the standard inflation factor. It is the coefficient of variation of the weights squared plus one.
Source: U.S. Department of Education, National Center for Education Statistics, National Household Education Survey, spring 1993.

The estimated reductions in bias under each adjustment weighting scheme are given in table 4. Estimates for additional characteristics are given in Brick *et al.* (1996). The bias reductions in the standard estimate assume each adjustment scheme eliminates the coverage bias.

The bias reduction estimates for most of the items in Table 4 are less than one percent and consistent in direction across the schemes. Before summarizing the estimates, we must account for the fact that the total number of children is constant for all the estimates due to the raking of the estimates to the CPS totals. The fixed total number of children across response categories has two consequences: it creates a negative correlation in the estimated reduction in bias across response categories; and it gives a false impression of the number of independent pieces of information in the tabled values.

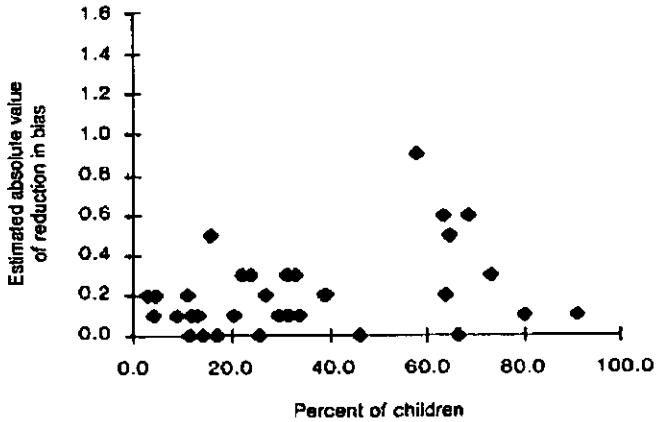
The approach taken to address to this problem in summarizing the bias estimates is to delete the estimate for one of the response categories for each item. The "no" response category for all items with "yes" and "no" response categories was deleted. For other types of variables, the response category with the smallest estimate was deleted.

Figure 2 presents the absolute value of the reduction in bias estimated using scheme A1 for the SR characteristics, and figure 3 is the same representation for the SS&D. These figures use all the estimates presented in Brick *et al.* (1996), rather than just those shown in table 4. For both components, the bias reductions are small. The largest absolute bias is 1.3 percent for SR and 0.9 percent for SS&D. The mean and median of the bias reductions and the absolute values of the bias reductions were also computed for each scheme and each component. For the SR component, the mean and median of the absolute value of the estimated bias reductions for the four schemes are between 0.2 and 0.4 percent. For the SS&D, the mean and median of the absolute values are between 0.1 and 0.3.



Source: U.S. Department of Education, National Center for Education Statistics, National Household Education Survey, spring 1993

Figure 2. Estimated reduction in absolute bias for School Readiness characteristics (scheme A1)



Source: U.S. Department of Education, National Center for Education Statistics, National Household Education Survey, spring 1993

Figure 3. Estimated reduction in absolute value of bias for School Safety and Discipline characteristics (scheme A1)

Bias Ratio

The size of the absolute reduction in bias is not a very useful statistical measure of the impact of the bias because it does not take the magnitude of the sampling error of the estimate into account. Cochran (1977) discusses the impact on confidence intervals as the ratio of the bias to the sampling error varies. For each scheme the bias ratio is given by

$$r_a = \frac{b_a}{\mathrm{se}(\hat{p}_s)}, \tag{4}$$

with the standard error of the standard estimate as the denominator. As the bias ratio increases, the chance of covering the population value departs significantly from the nominal confidence interval.

The bias ratios for selected characteristics are shown in Table 4. Many of the bias ratios for the SR items are large, even though the average and median ratios are near zero. Nearly half of the ratios for all the items examined are larger than 0.4 in absolute value. A ratio of 0.4 is large enough to reduce a nominal confidence interval from 95 percent to about 93 percent. For the SS&D items, the bias ratios are smaller, with only 15 percent of all the items having bias ratios greater than 0.4.

4.2 Mean Square Error

Since the variance is not an adequate measure of error for biased estimates, the mean square error of the estimates is used instead. The mean square error (MSE) is the sum of the variance and the square of the bias of the estimate.

The MSE can be estimated for the NHES:93 estimates by using the standard variance estimates and the bias reduction estimates presented above. The estimated MSE can be approximated as

$$\mathrm{MSE}_a = \mathrm{var}(\hat{p}_s) + b_a^2 \tag{5}$$

Table 4
Estimated Reduction in Bias and Bias Ratio for Selected Characteristics of the NHES:93

Characteristic	Standard estimate		Estimated reduction in bias				Bias ratio			
	Estimate	Standard error	Scheme A1	Scheme A2	Scheme B1	Scheme B2	Scheme A1	Scheme A2	Scheme B1	Scheme B2
School Readiness (SR) population										
Parental educational level										
Less than high school graduate	8.6	0.3	-1.7	-1.9	0.1	0.1	-5.7	-6.3	0.3	0.3
High school graduate or equivalent	33.9	0.8	0.4	0.3	-0.7	-1.0	0.5	0.4	-0.9	-1.3
Some college	57.5	0.7	1.3	1.6	0.6	0.9	1.9	2.3	0.9	1.3
Mother's employment status										
No mother in household	2.4	0.2	-0.1	-0.1	-0.1	-0.1	-0.5	-0.5	-0.5	-0.5
Employed 35 hours/week or more	34.3	0.5	0.5	0.8	0.2	0.5	1.0	1.6	0.4	1.0
Employed less than 35 hours/week	20.9	0.5	-0.1	-0.2	0.0	-0.2	-0.2	-0.4	0.0	-0.4
Seeking employment	6.6	0.4	0.0	-0.1	-0.1	-0.1	0.0	-0.3	-0.3	-0.3
Not in labor force	35.8	0.6	-0.4	-0.3	0.0	0.0	-0.7	-0.5	0.0	0.0
Father's employment status										
No father in household	26.3	0.5	-0.4	-0.6	0.0	-0.1	-0.8	-1.2	0.0	-0.2
Employed 35 hours/week or more	63.4	0.6	0.3	0.5	0.1	0.2	0.5	0.8	0.2	0.3
Employed less than 35 hours/week	3.8	0.3	0.0	-0.1	0.0	0.1	0.0	-0.3	0.0	0.3
Seeking employment	3.2	0.3	0.0	0.0	-0.1	-0.2	0.0	0.0	-0.3	-0.7
Not in labor force	3.3	0.2	0.1	0.2	0.0	0.1	0.5	1.0	0.0	0.5
Time since doctor visit for routine care										
Less than 1 year	84.1	0.4	0.4	0.4	0.2	0.1	1.0	1.0	0.5	0.2
Over 1 year	15.9	0.4	-0.4	-0.5	-0.2	-0.1	-1.0	-1.3	-0.5	-0.2
Birth weight										
5.5 pounds or less	93.3	0.3	-0.1	0.0	0.0	0.1	-0.3	0.0	0.0	0.3
Greater than 5.5 pounds	6.7	0.3	0.1	0.0	0.0	-0.1	0.3	0.0	0.0	-0.3
Child attending center-based program[1]										
Yes	52.6	0.8	0.9	0.3	0.8	0.6	1.1	0.4	1.0	0.8
No	47.4	0.8	-0.9	-0.3	-0.8	-0.6	-1.1	-0.4	-1.0	-0.8
Child ever attended center-based program[1]										
Yes	62.9	0.8	0.5	0.3	0.4	0.3	0.6	0.4	0.5	0.4
No	37.1	0.8	-0.5	-0.3	-0.4	-0.3	-0.6	-0.4	-0.5	-0.4
Attended center-based program prior to school[2]										
Yes	73.5	0.5	0.6	0.7	0.5	0.6	1.2	1.4	1.0	1.2
No	26.5	0.5	-0.6	-0.7	-0.5	-0.6	-1.2	-1.4	-1.0	-1.2
Women, Infant, and Children program participant[1]										
Yes	33.8	1.0	-0.6	-0.1	-0.8	-0.7	-0.6	-0.1	-0.8	-0.7
No	66.2	1.0	0.6	0.1	0.8	0.7	0.6	0.1	0.8	0.7
Free meal at school or center[2]										
Yes	35.8	0.6	-0.9	-1.1	-0.5	-0.5	-1.5	-1.8	-0.8	-0.8
No	64.2	0.6	0.9	1.1	0.5	0.5	1.5	1.8	0.8	0.8
Repeated kindergarten[3]										
Yes	5.7	0.4	-0.3	-0.5	-0.2	-0.2	-0.8	-1.3	-0.5	-0.5
No	94.3	0.4	0.3	0.5	0.2	0.2	0.7	1.3	0.5	0.5
School Safety and Discipline (SS&D) population										
Parental educational level										
Less than high school graduate	9.4	0.5	-1.2	-1.3	-0.3	-0.6	-2.4	-2.6	-0.6	-1.2
High school graduate or equivalent	32.7	0.6	0.3	0.0	-0.2	-0.6	0.5	0.0	-0.3	-1.0
Some college	57.9	0.5	0.9	1.3	0.5	1.1	1.8	2.6	1.0	2.2
Mother's employment status										
No mother in household	3.5	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Employed 35 hours/week or more	46.2	0.5	0.0	0.1	-0.1	0.1	0.0	0.2	-0.2	0.2
Employed less than 35 hours/week	20.3	0.5	0.1	0.0	0.0	-0.1	0.2	0.0	0.0	-0.2
Seeking employment	4.5	0.3	-0.2	-0.2	-0.2	-0.2	-0.7	-0.7	-0.7	-0.7
Not in labor force	25.5	0.5	0.0	0.1	0.2	0.2	0.0	0.2	0.4	0.4

Table 4
Estimated Reduction in Bias and Bias Ratio for Selected Characteristics of the NHES:93 – Concluded

Characteristic	Standard estimate		Estimated reduction in bias				Bias ratio			
	Estimate	Standard error	Scheme A1	Scheme A2	Scheme B1	Scheme B2	Scheme A1	Scheme A2	Scheme B1	Scheme B2
Father's employment status										
No father in household	26.8	0.6	-0.2	-0.2	-0.1	-0.2	-0.3	-0.3	-0.2	-0.3
Employed 35 hours/week or more	63.2	0.5	0.6	0.9	0.6	0.8	1.2	1.8	1.2	1.6
Employed less than 35 hours/week	3.1	0.2	-0.2	-0.2	-0.2	-0.2	-1.0	-1.0	-1.0	-1.0
Seeking employment	2.6	0.2	-0.2	-0.3	-0.2	-0.3	-1.0	-1.5	-1.0	-1.5
Not in labor force	4.3	0.3	-0.1	-0.1	-0.1	-0.1	-0.3	-0.3	-0.3	-0.3
School control										
Public	91.2	0.3	-0.1	-0.1	-0.1	-0.1	-0.3	-0.3	-0.3	-0.3
Private	8.8	0.3	0.1	0.1	0.1	0.1	0.3	0.3	0.3	0.3
Visitors required to sign in at school										
Yes	79.9	0.5	0.1	0.4	0.0	0.2	0.2	0.8	0.0	0.4
No	20.1	0.5	-0.1	-0.4	0.0	-0.2	-0.2	-0.8	0.0	-0.4
Had drug or alcohol ed program this year										
Yes	68.5	0.7	0.6	0.8	0.7	0.9	0.9	1.1	1.0	1.3
No	31.5	0.7	-0.6	-0.8	-0.7	-0.9	-0.9	-1.1	-1.0	-1.3
Students in fighting gangs at school[4]										
Yes	22.3	0.5	-0.3	-0.4	-0.3	-0.5	-0.6	-0.8	-0.6	-1.0
No	77.7	0.5	0.3	0.4	0.3	0.5	0.6	0.8	0.6	1.0
Ease of obtaining marijuana at school[4]										
Very or fairly easy	39.2	0.6	-0.2	-0.3	-0.2	-0.3	-0.3	-0.5	-0.3	-0.5
Hard	29.7	0.5	0.1	0.1	0.2	0.2	0.2	0.2	0.4	0.4
Nearly impossible	31.1	0.6	0.1	0.1	0.0	0.1	0.2	0.2	0.0	0.2
Fear of incident of crime at school										
None	66.1	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Fear of theft or robbery[5]	11.9	0.5	-0.1	-0.2	0.0	-0.2	-0.2	-0.4	0.0	-0.4
Fear of bullying or assault[5]	8.6	0.3	-0.1	-0.1	-0.1	-0.1	-0.3	-0.3	-0.3	-0.3
Fear of two or more types of incidents[5]	13.3	0.5	0.1	0.3	0.1	0.2	0.2	0.6	0.2	0.4
Knowledge of crime at school										
None	38.7	0.6	0.2	0.1	0.2	0.1	0.3	0.2	0.3	0.2
Fear of theft or robbery[5]	14.1	0.5	0.2	0.3	0.2	0.3	0.4	0.6	0.4	0.6
Fear of bullying or assault[5]	15.6	0.4	-0.5	-0.4	-0.4	-0.4	-1.3	-1.0	-1.0	-1.0
Fear of two or more types of incidents[5]	31.6	0.6	0.1	0.0	0.0	0.0	0.2	0.0	0.0	0.0
Victimization by crime										
Not victimized	73.0	0.5	0.3	0.2	0.3	0.2	0.6	0.4	0.6	0.4
Victim of theft or robbery[5]	10.9	0.3	-0.2	-0.1	-0.1	0.0	-0.7	-0.3	-0.3	0.0
Victim of bulling or assault[5]	8.9	0.3	-0.1	0.0	-0.2	-0.1	-0.3	0.0	-0.7	-0.3
Victim of two or more types of incidents[5]	7.2	0.3	0.0	0.0	0.0	-0.1	0.0	0.0	0.0	-0.3
Witnessed crime at school										
None	63.8	0.8	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2
Witnessed robbery[6]	0.6	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Witnessed bulling or assault[6]	24.1	0.8	-0.3	-0.3	-0.3	-0.3	-0.4	-0.4	-0.4	-0.4
Witnessed two or more types of incidents	11.4	0.4	0.0	0.1	0.0	0.0	0.0	0.2	0.0	0.0

[1] Applies to preschoolers only.
[2] Applies to all children except preschoolers.
[3] Applies to children in primary school only.
[4] Applies to students in grades 6 through 12 only.
[5] For the fear of incident, knowledge of crime, and victimized by crime variables, the second response category is used if either theft or robbery was reported but not both, the third response category is used if either bullying or assault was reported but not both.
[6] This response category is used if either bullying or assault was reported, but not both, was reported.
Note: Percents may not add to 100 because of rounding.
Source: U.S. Department of Education, National Center for Education Statistics, National Household Education Survey, spring 1993.

where $\hat{p}_s$ is the estimated proportion under the standard approach and b_a is the reduction in bias under scheme a. Because of the high correlation in the estimates of the bias from the four adjustment schemes, only the mean square errors for scheme A1 were computed. In Brick *et al.* (1996), the estimates using other schemes are shown to have negligible effects.

The mean square errors of the adjusted estimates are now contrasted with the variability in the standard NHES:93 estimates. The variance increase from adjusting the weights using the telephone service interruption data was expressed as a *VIF* in table 3. Multiplying the variance estimates of the standard estimates by the appropriate adjustment factor yields an approximate variance for the adjusted (presumably unbiased) estimates which are then compared to the mean square error of the standard estimates.

To aid in comparing the weighting procedures, ratios of the variance of the adjusted estimate to the mean square error for the standard estimate were tabulated (see Brick *et al.* 1996). The ratio is called the mean square ratio and can be written as

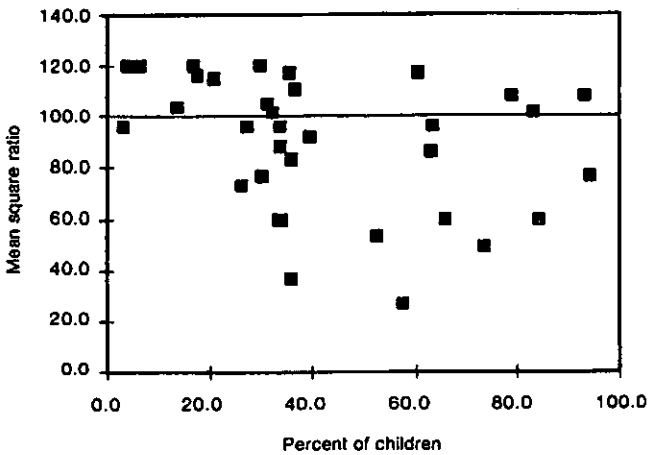
$$\text{msr}_a(\hat{p}) = \frac{100 \times \text{relative}VIF_a \times \text{var}(\hat{p}_s)}{\text{mse}_{A1}(\hat{p})}. \qquad (6)$$

Note that the mean square error is derived using the bias estimated from scheme A1 only, but it is used to compute the mean square ratios for all four schemes. As noted above, this simplification does not have much effect on the mean square ratios because the bias estimates are approximately the same across schemes.

The mean square ratios include contributions from the bias (in the mean square error estimates) and the variance (in the *VIF*). When the mean square ratio is 100, the variance of the adjusted estimate is exactly equal to the mean square error of the biased, standard estimate. A ratio less than 100 indicates that the bias reduction of the adjustment is greater than the variance increase that comes with it. A mean square ratio over 100 means that the variance increase associated with the adjustment is greater than the bias reduction.

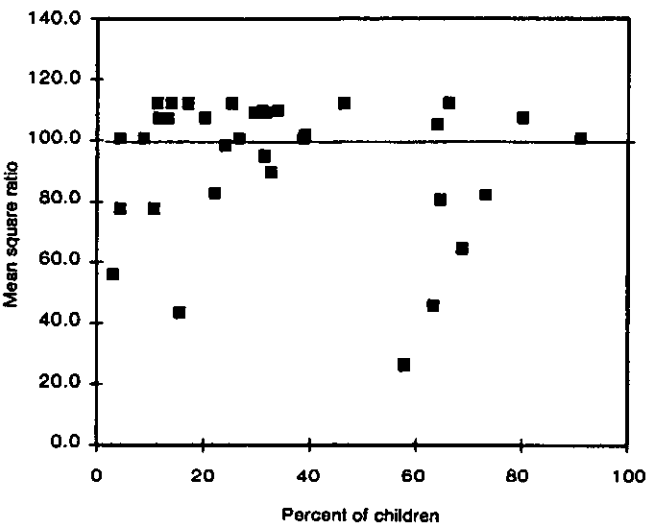
Figures 4 and 5 graphically present the msr for the two component surveys using scheme A1. In addition, Table 5 shows summary statistics for the msr for all four adjustment schemes. The distributions of mean square ratios for both components are very similar with the mean square ratios slightly lower for the SR component. The medians for schemes A1 and B1 (those based on interruptions of one week or more) are near the break-even point of 100. The means for these schemes are close to 90 and the figures confirm that the difference between the mean and medians is due to the skewed distributions of the mean square ratios.

A striking feature of the distributions of the mean square ratios for schemes A1 and B1 is the size of the ratios at the extremes of the distribution. The maximum mean square ratios for both components is 120, while some ratios are as small as 26. This means the maximum increase in the mean square error of the estimates is 20 percent, while the reductions in



Source: U.S. Department of Education, National Center for Education Statistics, National Household Education Survey, spring 1993

Figure 4. Estimated mean square ratios for selected School Readiness items (scheme A1)



Source: U.S. Department of Education, National Center for Education Statistics, National Household Education Survey, spring 1993

Figure 5. Estimated mean square ratios for selected School Safety and Discipline items (scheme A1)

mean square error for a number of other estimates are quite large. Thus, the penalty associated with adjusting even when the estimate is not biased is modest, but the benefits of adjusting when it is needed are impressive.

The distributions for the mean square ratios for schemes A1 and B1 are very similar, and the choice of which of these schemes should be used may be determined by nonstatistical issues, such as availability of data and the other types of adjustments required in the survey. The mean square ratios show that the adjusted weights reduce the mean square error for about half the estimates considered below those derived from the standard weights. The distributions of the mean square ratios for schemes A2 and B2 (those based on interruptions of 1 month or more) have medians and means

Table 5
Summaries of Distribution of Mean Square Ratios for Selected Characteristics of School Readiness and School Safety and Discipline Components

	Adjustment scheme			
	A1	A2	B1	B2
School Readiness				
Mean	89.8	101.0	86.8	94.2
Median	96.0	108.0	92.8	100.8
Minimum	27.0	30.3	26.1	28.3
Maximum	120.0	135.0	116.0	126.0
School Safety and Discipline				
Mean	93.3	104.9	92.2	103.9
Median	100.8	113.4	99.9	112.5
Minimum	26.4	29.7	26.2	29.5
Maximum	112.0	126.0	111.0	125.0

Source: U.S. Department of Education, National Center for Education Statistics, National Household Education Survey, spring 1993.

that are greater than 100. Essentially, these mean square ratios are shifted upward when compared with those of schemes A1 and B1, and are not recommended.

5. CONCLUSIONS

If the percentage of the target population living in non-telephone households is relatively large and the characteristics of those persons are different from those who live in telephone households, then the estimates may be susceptible to significant coverage bias. One method of addressing this problem without resorting to other modes of data collection is to adjust the weights to reduce the coverage bias. In this study, the weights for persons in households reporting an interruption in telephone service were increased to account for those without telephones.

The bias reduction estimates computed under the assumed model showed that the coverage adjustments for the SR component improved some of the estimates substantially, and did not do much harm to any statistics. The bias reduction estimates from the SS&D component, on the other hand, were not as substantively important. The adjustments reduced bias for both components, but they also increased the variability of the estimates. The distributions of the mean square ratios show that about half the estimates could be improved using the telephone service interruption adjustments. Furthermore, even for those estimates that were less accurate due to the variance increases associated with the differential weights, the magnitude of the increases were not large. In other words, the penalty for adjusting when it did not reduce the coverage bias was not very great. These findings suggest that the adjustments should be seriously considered.

The alternative weighting schemes performed differently with respect to the mean square ratios. The schemes based on interruptions of telephone service of one week or more were better than the schemes based on interruptions of one month or more. The bias adjustments resulting from using educational attainment by race/ethnicity categories were roughly equivalent to those using tenure by race/ethnicity.

The size of the sample is a relevant factor that should be taken into account when evaluating the use of the telephone service interruption adjustment. Bias ratios increase with the sample size because the bias is not affected while the sampling error of the estimate (the denominator of the bias ratio) decreases. Thus, the adjustments should be more beneficial in surveys with large sample sizes where the bias ratios might be expected to be large.

While the results of this study suggest that the adjustments could be useful for many estimates from telephone surveys, confirmation is needed before the adjustments are recommended. As discussed earlier, the estimates of the mean square errors in this study were based on the assumption that the adjusted estimates eliminated the bias of the estimates. This model assumption could not be verified because of the lack of benchmark data for comparison. The assumed model is very beneficial to the adjusted estimates in the sense that it results in lower bounds on the mean square errors for the adjusted estimates. Thus, the findings of this study should be taken as an indication that adjustment using data on interruptions in telephone service is a feasible method, but requires further study and evaluation.

The questions about interruptions in telephone service were recently added to the National Health Interview Survey, a survey conducted by the Census Bureau for the National Center for Health Statistics. The findings from this survey should be very useful in evaluating this method because the survey covers households without telephones by in-person interviews, eliminating the need for the critical model assumption used in this study.

ACKNOWLEDGEMENTS

The authors would like to thank the referees and editor for comments that substantially improved the methods and presentation of the material.

REFERENCES

BRICK, J., BURKE, J., and WEST, J. (1992). Telephone Undercoverage Bias of 14- to 21-year-olds and 3- to 5-year-olds. National Household Education Survey Technical Report No. 2, Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement, NCES 92-101.

BRICK, J., KEETER, S., WAKSBERG, J., and BELL, B. (1996). Adjusting for Coverage Bias Using Telephone Service Interruption Data. National Household Education Survey Technical Report, Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement, NCES 96-336.

COCHRAN, W. (1977). *Sampling Techniques*. New York: John Wiley and Sons, 12-15.

CONVERSE, J., and PRESSER, S. (1986). *Survey Questions, Handcrafting the Standardized Questionnaire*. Sage University Paper series on Quantitative Applications in the Social Sciences, 07-063. Beverly Hills: Sage Publishers.

FEDERAL COMMUNICATIONS COMMISSION (1988). Monitoring Report: CC Docket No. 87-339. Prepared by the Staff of the Federal-State Joint Board in CC Docket No. 80-286, Washington DC.

KALTON, G., and KASPRZYK, D. (1986). The treatment of missing survey data. *Survey Methodology*, 12, 1-16.

KEETER, S. (1995). Estimating noncoverage bias from a phone survey. *Public Opinion Quarterly*, 59, 196-217.

KISH, L. (1992). Weighting for unequal P_i. *Journal of Official Statistics*, 8, 183-200.

MASSEY, J., and BOTMAN, S. (1988). Weighting adjustments for random digit dialed surveys, Chapter 9 in *Telephone Survey Methodology*. Eds. Groves, Biemer, Lyberg, Massey, Nicholls, and Waksberg. New York: John Wiley and Sons, 143-160.

POLITZ, A., and SIMMONS, W. (1949). An attempt to get the 'not at homes' into the sample without callbacks. *Journal of the American Statistical Association*, 44, 9-31.

THORNBERRY, O., and MASSEY, J. (1988). Trends in United States telephone coverage across time and subgroups, Chapter 3 in *Telephone Survey Methodology*. Eds. Groves, Biemer, Lyberg, Massey, Nicholls, and Waksberg. New York: John Wiley and Sons, 25-50.

Survey Methodology, December 1996
Vol. 22, No. 2, pp. 199-204
Statistics Canada

Optimal Sample Redesign Under GREG in Skewed Populations With Application

GURUPDESH S. PANDHER [1]

ABSTRACT

Within a survey re-engineering context, the combined methodology developed in the paper addresses the problem of finding the minimal sample size for the generalized regression estimator in skewed survey populations (*e.g.*, business, institutional, agriculture populations). Three components necessary in identifying an efficient sample redesign strategy involve i) constructing an efficient partitioning between the "take-all" and "sampled" groups, ii) identifying an efficient sample selection scheme, and iii) finding the minimal sample size required to meet the desired precision constraint(s). A scheme named the "Transfer Algorithm" is devised to address the first issue (Pandher 1995) and is integrated with the other two components to arrive at a combined iterative procedure that converges to a globally minimal sample size and population partitioning under the imposed precision constraint. An equivalence result is obtained allowing the solution to the proposed algorithm to be alternatively determined in terms of simple quantities computable directly from the population auxiliary data. Results from the application of the proposed sample redesign methodology to the Local Government Survey in Ontario are reported. A 52% reduction in the total sample size is achieved for the regression estimator of the total at a minimum coefficient of variation of 2%.

KEY WORDS: Minimal sample size; Optimal sample selection; Precision constraint; Sampled group; Take-all group.

1. INTRODUCTION

In many survey situations additional information is available on all population units before the survey is undertaken. This auxiliary information is frequently useful in devising a more efficient sample design and estimation strategy. In a survey redesign context, the most optimal strategy holds the promise of offering the largest reduction in survey costs by requiring the lowest sample size necessary to meet the desired precision constraint on the estimates. In repeat surveys of skewed populations, an efficient sample design and estimation strategy may be realized by exploiting a) the correlation structure between the size-based auxiliary information x (*e.g.*, population of municipality, employees in a firm, farm acreage) and the survey variables y (*e.g.*, municipality expenditures, value of shipments, farm yield) and b) the variance relationship between the survey variable and the auxiliary size information.

In this paper, a comprehensive sample redesign methodology is developed for skewed populations with the ultimate objective of bringing about maximal reductions in the current sample size while ensuring a desired level of precision for the generalized regression estimator of the total. This work was motivated by the redesign of the Local Government Finance Survey (LGFS) conducted by Statistics Canada's Public Institutions Division. Financial information (*e.g.*, revenues, expenditures, debt, *etc.*) obtained from local government units is used in the estimation and publication of financial statistics on a provincial and national basis. Although the work presented in this paper is motivated by a concrete application, the sample design methodology devised applies generally to all surveys based on skewed populations (*e.g.*, agricultural, business, and institutional surveys).

In identifying an efficient new sample design, the overall methodology addresses and integrates the solution to three problems:

1) Creation of the "Take-all" and "Sampled Groups"

Since the variability of the survey response y_k tends to increase with the size of the unit x_k, it is common in skewed populations to sample the largest x-valued units with certainty in order to improve the efficiency of the population estimators. The demarcation of the population into the non-overlapping "take-all" $U_a = \{1, ..., N_a\}$ and "sampled" groups $U_b = \{1, ..., N_b\}$ is obtained through a new scheme named the "Transfer Algorithm".

2) Choosing an Efficient Sample Selection Scheme

Let $p(s; \lambda) = (p_a(s_a), p_b(s_b; \lambda))$ represent the complete sample design where the sample design parameter λ determines the type of sample selection implemented in the sampled group U_b. The sample inclusion probabilities due to $p_b(s_b; \lambda)$ may be expressed as $\pi_k(\lambda) = n_b(x_k^{\lambda/2} / \sum_{U_b} x_j^{\lambda/2})$, $k \in U_b$. Note that the parameter λ defines a broad class of sample designs with SRS ($\lambda = 0$) and pps ($\lambda = 2$) as particular cases. Design optimality results (Godambe and Joshi 1965) allow the identification of the most optimal value for the sample design parameter λ.

[1] Gurupdesh S. Pandher, Survey Analysis and Methods Development Section, Household Survey Methods Division, Methodology Branch, Statistics Canada, 16th Floor, R.H. Coats Building, Ottawa, Ontario, Canada, K1A 0T6.

3) Minimal Sample Size Determination

The third component of the overall methodology is aimed at finding the minimal sample size required to meet the imposed precision constraints for the estimator.

The combined procedure devised integrates these components to allow a new globally minimal sample size and optimal population partitioning to be determined under a flexible range of sample selection strategies (*e.g.*, SRS, pps, generalized pps). Firstly, the "Transfer Algorithm" is proposed which finds an optimal population allocation between the take-all and sampled population groups in the sense of minimizing the variance of the generalized regression estimator (GREG) of the total. Desirable mathematical properties of this algorithm such as existence and optimality of solution along with an equivalence result were established in Pandher (1995). The equivalence result allows the solution to be determined in terms of simple quantities computable directly from the population auxiliary data.

The Transfer Algorithm in then synthesized iteratively with the sample size determination step to find the minimal sample size needed to satisfy the imposed precision constraints through an iterative procedure. The combined methodology produces a sequence of sample sizes and population partitionings which converge to a globally optimal solution where further reductions in the sample size are not possible given the imposed precision constraint. An application of the procedure is given for Ontario using provincial data from the Local Government Finance Survey.

Lavallée and Hidiroglou (1988), Hidiroglou and Srinath (1993) (subsequently denoted as L&H and H&S, respectively), and Glasser (1962) have proposed alternative methodologies for constructing the take-all and sampled groups within the context of stratified SRS design. The proposed approach differs from other methods in three respects. Firstly, the population demarcation is obtained under a flexible range of sample selection strategies (*e.g.*, SRS, pps, generalized pps). Secondly, the criterion for constructing the population demarcation is based on minimizing the variance of the GREG estimator of the total under the desired sample selection strategy (Glasser and L&H base their allocation on minimizing the within-stratum sum-of-squares x; H&S use the total regression sum-of-squares under a regression model with a compulsory intercept assuming SRS). Thirdly, the proposed methodology explicitly captures the size-induced heteroscedasticity present in skewed survey populations which has been ignored in other frameworks.

Lastly, it is useful to qualify the sense in which the term "optimal" is used. Since, the redesign uses auxiliary information from a previous cycle of the survey to estimate the design parameters, there is a level of sub-optimality introduced in the redesign methodology by this lag. But as a practical matter, using the data from the most recent survey is the best that can be done. Once the design parameters have been estimated or are known however, the cut-offs and sample sizes required to achieve the desired precision yield the lowest anticipated design variance given that the estimates are true (or close to it). It is therefore, in this sense that the word "optimal" is used.

2. SURVEY FRAMEWORK

The model assisted survey framework is adopted for the skewed population whose auxiliary and survey characteristics are denoted by $C_U = \{(x_1, y_1), ..., (x_N, y_N)\}$. In this framework, underlying the class of generalized regression estimators for the population total are regression models (Särndal 1992, p. 255) exploiting the correlation between the survey variables y and the auxiliary covariates x. Different model assumptions on the deterministic and stochastic components of the underlying model lead to different regression estimators for the population total. For example, a ratio-form heteroscedastic model

$$y_k = \beta x_k + \epsilon_k, \tag{2.1}$$

with the error $\epsilon_k \sim (0, \sigma_k^2)$ and the variance structure given by $\sigma_k^2 = c\, x_k^{\gamma}$ (γ is the heteroscedasticity parameter) leads to the following GREG estimator:

$$\hat{t}_{Rb} = \sum_{U_b} x_k \hat{B} + \sum_{s_b} \frac{(y_k - x_k \hat{B})}{\pi_k} \tag{2.2}$$

where $\hat{B} = (\sum_s y_k/\pi_k)/(\sum_s x_k/\pi_k)$ is the sample-based probability weighted estimate of the population regression parameter B.

Given this estimation framework, the total across both groups $t = t_a + t_b$ is estimated by $\hat{t} = t_a + \hat{t}_{Rb}$ where $\hat{t}_a = t_a = \sum_{U_a} y_k$ since all units are sampled in the take-all group and $\hat{t}_{Rb}$ is the GREG estimator under the relevant model. The anticipated variance of $\hat{t}_{Rb}$ (defined as the variance with respect to both the design and the model, denoted p and ξ, respectively) is expressible as

$$V(\hat{t}_{Rb}) \equiv \mathrm{E}_\xi V_p(\hat{t}_{Rb}) \doteq \sum_{k \in U_b} \left(\frac{1}{\pi_k} - 1 \right) \sigma_k^2. \tag{2.3}$$

Furthermore, if σ_k^2 depends on the auxiliary measure x_k according to the formulation $\sigma_k^2 = c x_k^{\gamma}$ (2.4), then design optimality (Godambe and Joshi 1965) implies that the optimal sample inclusion probabilities are $\pi_k^*(\gamma) \propto x_k^{\gamma/2}$, $k \in U_b$. Therefore, the sample design $p_b^*(s_b; \lambda = \gamma)$ in the sampled sub-population, defining the first order inclusion probabilities $\pi_k^*(\gamma) = n(x_k^{\gamma/2}/\sum_U x_j^{\gamma/2})$, $k \in U_b$, minimizes the anticipated variance $V(\hat{t}_{Rb})$.

In the model assisted framework used in this paper, the auxiliary measure x_k is assumed to be a scalar. As noted by a referee, the more general case where x_k is a vector could be handled by fitting the appropriate parametric relationship $\sigma_k^2 = f(x_{k1}, ..., x_{kq})$ and using the estimated $\hat{\theta}_k$ in lieu of x_k in defining the inclusion probabilities. The approach for the multivariate x_k seems intuitively sound and is mentioned here for completeness but requires further study and investigation.

Three methods for estimating the heteroscedasticity parameter γ from past survey data called the "Least Squares Method", the "Maximum Likelihood Method", and the "Graphical Method" are described in Appendix A of Pandher (1995).

3. TRANSFER ALGORITHM

In this section, an iterative scheme named the "Transfer Algorithm" is proposed to determine the optimal demarcation between the take-all and sampled sub-populations under the sample design $p(s;\lambda)$. The criterion for this construction is based on finding a population partitioning minimizing the estimated anticipated variance of $\hat{t}_{Rb}$. An equivalence result from Pandher (1995) is used to find an alternative and simpler method of solution based entirely on quantities defined on the auxiliary population data.

The proposed scheme for constructing the take-all and sampled sub-populations, U_a and U_b, respectively, is based on the following idea. Initially, place all population units in the sampled group, labelling it $U_b^{(0)}$ (the superscript l represents the iteration cycle). Hence, the take-all group is an empty set $U_a^{(0)}=\{\varnothing\}$. The resulting population and sample size allocation at $l = 0$ is given by $N_a^{(0)} = 0$, $n_a^{(0)} = 0$, $N_b^{(0)} = N$, and $n_b^{(0)} = n_0$ where n_0 is the current sample size.

In a repeat survey setting, the variances σ_k^2 in (2.3) can be empirically modelled using the relation $\sigma_k^2 = c\, x_k^{\gamma}$ (2.4) where γ and c are estimated from previous sample data as mentioned before. Using the estimated version of (2.4) in (2.3) yields the following estimator for $V^{(l)}(\hat{t}_{Rb};\cdot)$:

$$\hat{V}^{(l)}(\hat{t}_R;\lambda,N_b^{(l)},n_b^{(l)}) = \sum_{k\in U_b^{(l)}} \left(\frac{1}{\pi_k(\lambda)} - 1\right) \hat{c}\, x_k^{\hat{\gamma}} \qquad (3.1)$$

where the largest l x-valued units have been removed from $U_b^{(0)}$. Note that λ is used here to parameterize the sample design to allow greater generality when $\lambda \neq \gamma$.

In the iterative algorithm, we start initially with all population units placed in $U_b^{(0)}$. Then for each iteration $l, 0 \le l < n$, the largest $l+1$ x-valued unit $x_{(N-l-1)}$ is transferred from $U_b^{(l)}$ to $U_a^{(l)}$ and the difference

$$\Delta(l) = \hat{V}^{(l+1)}(\hat{t}_{Rb};\lambda, N-l-1, n-l-1) - \hat{V}^{(l)}(\hat{t}_{Rb};\lambda, N-l, n-l) \qquad (3.2)$$

is computed. Negative values of $\Delta(l)$ mean that the transfer of the unit corresponding to the ordered value $x_{(N-l-1)}$ lead to a decrease in the variance. Moreover, such transfers continue to result in a reduction in the variance of $\hat{t}_{Rb}$ as long as $\Delta(l) < 0$. In general, for any iteration l, the relationship between the population and sample size allocations is described by the following relations: $N_b^{(l)} = N - l$, $n_b^{(l)} = n - l$, and $N_a^{(l)} = n_a^{(l)} = l$. These relations hold because the overall population and sample sizes must remain constant ($N = N_a^{(l)} + N_b^{(l)}$ and $n = n_a^{(l)} + n_b^{(l)}$) for all iterations.

The solution is also constrained by the condition $\pi_k(\lambda) < 1, k \in U_b^*(l^*)$. Let $l^*(\lambda)$, $0 \le l^* < n$, represent the solution to the Transfer Algorithm. Given the discussion above, the solution to the Transfer Algorithm under the sample design $p(s;\lambda)$ may be formulated as

$$l^*(\lambda) = \min_l \{l: [\pi_{(N-l)}(\lambda) < 1] \quad \text{and} \quad \hat{\Delta}(l) = [\hat{V}^{(l+1)}(\hat{t}_{Rb};\lambda) - \hat{V}^{(l)}(\hat{t}_{Rb};\lambda)] \ge 0, 0 \le l < n\}. \qquad (3.3)$$

The optimal population allocation to the take-all group $U_a^*(l^*)$ is then given by the population units coinciding with the l^* ordered units transferred to the take-all auxiliary vector $X_a^* = (x_{(N-l^*)}, x_{(N-l^*+1)}, \ldots, x_{(N)})$; correspondingly the sampled group $U_b^*(l^*)$ consists of the units corresponding to $X_b^* = (x_{(1)}, x_{(2)}, \ldots, x_{(N-l^*-1)})$.

Transferring a unit from $U_b^{(l)}$ to $U_a^{(l)}$ causes two opposite effects on the variance $V^{(l)}(\hat{t}_{Rb};\cdot)$. The reduction in the population size ($N_b^{(l+1)} = N_b^{(l)} - 1$) has the impact of decreasing the variance, while the equivalent reduction in the sample size ($n_b^{(l+1)} = n_b^{(l)} - 1$) has the reverse effect of increasing $V^{(l)}(\hat{t}_{Rb};\cdot)$. Somewhere in this process, a critical value l^*, $0 \le l^* < n$, exists which gives the optimal breakdown $\{U_a^*(l^*), U_b^*(l^*)\}$. Moreover, in Theorem 3 of Pandher (1995), it is shown that as long as the conditions $(x_{(N-l)}^{\lambda/2} - x_{(N-l-1)}^{\lambda/2}) \ge 0$ and $(x_{(N-l)}^{\gamma-\lambda/2} - x_{(N-l-1)}^{\gamma-\lambda/2}) \ge 0, 0 \le l < n$, hold, a solution to the Transfer Algorithm exists and that the system remains stable (optimal) upon reaching l^*. Stability further implies that the solution is optimal since the conditions leading to the solution do not change in the range $l^* \le l < n$. These two properties may be more precisely defined as follows:

Existence: $\exists\, l^*, 0 \le l^* < n$, such that $V^{(l^*+1)} - V^{(l^*)} \ge 0$ and $\pi_{(N-l^*)}^{(l^*)} < 1$.

Stability: If $V^{(l^*+1)} - V^{(l^*)} \ge 0$, then $V^{(l+1)} - V^{(l)} \ge 0$ and $\pi_{(N-l)}^{(l)} < 1$ for $0 \le l^* < l < n$.

An example of the application of the Transfer Algorithm to the LGF survey population of local municipalities in Ontario (with $N = 793$, $n = 108$, $\gamma = 2$, and $\lambda = 1$) is given in Figure 1. The curves are plotted for $l > 8$ because in the interval $0 < l \le 8$, the first condition of (3.3), namely $[\pi_{(N-l)}(\lambda) < 1]$, is not satisfied. The minimum value of $\hat{V}^{(l)}(\hat{t}_{Rb})$ is achieved at $l^* = 57$ where $\Delta(l^*) = \hat{V}^{(l^*+1)} - \hat{V}^{(l^*)} \ge 0$.

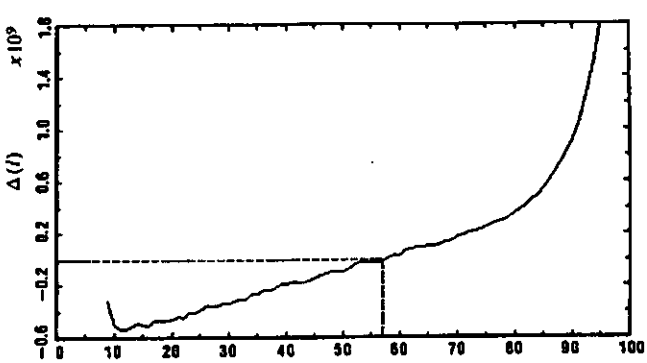


Figure 1. Changes in variance of regression estimator ($\lambda = 1$): $\Delta(l) = V^{(l+1)}(t; 1, N-l-1, n-l-1) - V^{(l)}(t; 1, N-l, n-l)$

Theorem 2 from the complete paper is an important result which allows the solution to the Transfer Algorithm to be equivalently expressed in terms of simpler quantities based on the auxiliary data. A brief sketch of the development of this theorem is given in the Appendix.

Theorem 2. Equivalent Solution to the Transfer Algorithm

The solution $l^*(\lambda)$ to the Transfer Algorithm stated in (3.3) in terms of $V^{(l)} - V^{(l-1)}$ and $\pi^{(l)}_{(N-l)}(\lambda)$ may also be equivalently expressed as

$$l^*(\lambda) = \begin{cases} \min_l \{l : n - l \le R(l;\gamma - \lambda/2), 0 \le l < n\}, & 0 \le \lambda < \gamma \\ \min_l \{l : n - l \le R(l;\gamma/2), 0 \le l < n\}, & \lambda = \gamma \\ \min_l \{l : n - l \le R(l;\lambda/2), 0 \le l < n\}, & \gamma < \lambda \le 2\gamma \end{cases}$$

where $R(l;\gamma - \lambda/2) = \sum_{k=1}^{N-l} x_{(k)}^{\gamma-\lambda/2} / x_{(N-l)}^{\gamma-\lambda/2}$ and $R(l;\lambda/2) = \sum_{k=1}^{N-l} x_{(k)}^{\lambda/2} / x_{(N-l)}^{\lambda/2}$ define the critical values.

This use of this theorem to find the optimal population allocation is illustrated graphically in Figure 2 (Ontario data). In this case, $0 \le \lambda < \gamma$, and the solution is determined by the behaviour of the functions $R(l;\gamma - \lambda/2)$ (the lower curve in the graph) and $n - l$. The same solution $l^* = 57$ is obtained as before.

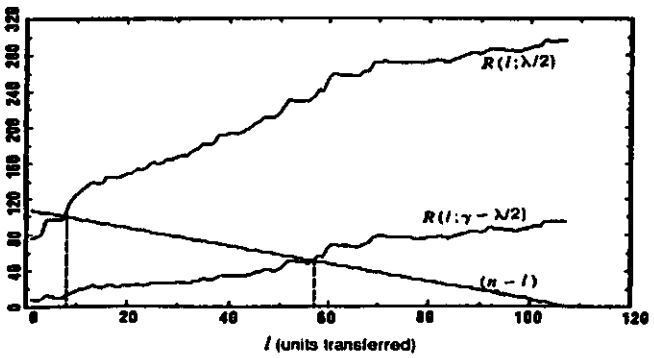


Figure 2. Use of $R(l;\gamma - \lambda/2)$, $R(l;\lambda/2)$, and $(n - l)$ to construct optimal take-all/sampled groups (Ontario)

4. SAMPLE SIZE DETERMINATION AND COMBINED ITERATIVE PROCEDURE

Given a sample design $p(s,\lambda)$, $0 \le \lambda \le 2\gamma$, with sample size n, the Transfer Algorithm yields an optimal construction of the take-all and sampled sub-populations, $U_a^*(l^*)$ and $U_b^*(l^*)$, respectively. Next, an expression for finding the minimal sample size is obtained which meets the imposed precision constraint – expressed in terms of the coefficient of variation CV_{min}. The sample determination step is then integrated with the Transfer Algorithm to develop a combined procedure which allows the survey designer to find the globally minimal sample size and optimal population partitioning.

4.1 Expression for New Sample Size

Let q represent the iteration cycle for the combined procedure and $n_q^* = n_{aq}^* + n_{bq}^*$ denote the total minimal sample size required to satisfy the precision constraint. Given the sample design $p_q(s,\lambda,l_q^*(\lambda,n_q))$, current sample size n_q, and the population partitioning $\{U_{aq}^*(l_q^*), U_{bq}^*(l_q^*)\}$, the precision constraint for $\hat{t}_R = t_a + \hat{t}_{Rb}$ may be stated formally as

$$CV_{min} \ge \frac{\hat{V}_q^{1/2}(\hat{t}_{Rb};\lambda, N - l_q^*, n_q - l_q^*)}{\hat{t}_R}. \qquad (4.1)$$

Solving this inequality for n_{bq}^* gives the following expression for the minimal sample size needed in the sampled group $U_{bq}^*(l_q^*)$ to meet the precision constraint:

$$n_{bq}^* = n_q^* - l_q^*(n_q) = \frac{X(l_q^*,\lambda/2)\,X(l_q^*,\hat{\gamma} - \lambda/2)\hat{c}}{\hat{t}_R^{\,2} CV_{min} + X(l_q^*,\hat{\gamma})\hat{c}} \qquad (4.2)$$

where $X(l_q^*,\lambda/2) = \sum_{k=1}^{N-l_q^*} x_{(k)}^{\lambda/2}$, $X(l_q^*,\hat{\gamma} - \lambda/2) = \sum_{k=1}^{N-l_q^*} x_{(k)}^{\hat{\gamma}-\lambda/2}$, and $\hat{t}_R$ may be estimated from past survey data corresponding to the period of the auxiliary information. The total new minimal sample size required to meet the precision constraint is then given by

$$n_q^* = n_{aq}^* + n_{bq}^* = l_q^*(n_q) + n_{bq}^*. \qquad (4.3)$$

4.2 Combined Sample Redesign Methodology

Next, note that the solution to the Transfer Algorithm l^* depends on the current total sample size: $l_q^*(\lambda) \equiv l_q^*(\lambda, n_q)$. Once the new minimal sample size n_q^* is determined, the existing partitioning $\{U_{aq}^*(l_q^*), U_{bq}^*(l_q^*)\}$ which was optimal at n_q is no longer optimal at the new minimal sample size n_q^* because $l^*(\lambda,n_q^*) \ne l^*(\lambda,n_q)$ if $n_q^* \ne n_q$. Therefore, letting $n_{q+1} = n_q^*$, a new population partitioning from the Transfer Algorithm based on $l_{q+1}^*(\lambda,n_{q+1})$, given by $\{U_{a,q+1}^*(l_{q+1}^*), U_{b,q+1}^*(l_{q+1}^*)\}$, is required to optimize the construction of the take-all and sampled sub-populations. Next, applying (4.2) over $U_{b,q+1}^*(l_{q+1}^*)$ gives a new minimal sample size $n_{q+1}^* = l_{q+1}^*(n_{q+1}) + n_{b,q+1}^*$ required to achieve the desired precision CV_{min}. Proceeding in this fashion, the combined scheme produces a sequence of population partitionings, sample sizes, and sample allocations

$$(l^*(\lambda,n_q), (n_{aq} = l_q^*, n_{bq} = n_q - l_q^*),$$

$$(N_{aq}^* = l_q^*, N_{bq}^* = N - l_q^*), (n_{aq}^* = l_q^*, n_{bq}^*)), \; q = 0,1,\ldots \qquad (4.4)$$

with $n_{q+1} = n_q^* = n_{aq}^* + n_{bq}^*$ and the initial value n_0 (current survey sample size). The combined procedure is repeated until further reductions in the minimal sample size can no longer be achieved. This leads to the stopping rule

$$q^* = \min_q \{q : n_{q+1}^* - n_q^* \ge 0\}. \qquad (4.5)$$

The optimality of the combined procedure can be established using Theorem 2 and is omitted here due to space (see Pandher 1995). The main result is that the combined procedure converges to a globally optimal solution along the path defined by (4.4) to a point where further reductions in the sample size are not possible (by reconstructing U_a^* and U_b^*) given the imposed precision constraint.

5. APPLICATION

The combined sample design procedure described above is now applied to the redesign of the Local Government Finance Survey in the province of Ontario. The survey response y in this application is the actual expenditures reported for sampled local government units for Ontario in 1989. The actual estimates are prepared 30 months after the end of the survey year from financial statements submitted by the local government units to the Department of Municipal Affairs (provincial). Population counts for the local government units from the nearest census (1991) are used as the auxiliary variable x. The population of local-level municipalities for Ontario consists of a total of 793 units of which a sample of 108 units is currently taken.

The results of applying the combined methodology to Ontario LGFS data are reported in Table 1. The level of desired precision CV_{min} was set at 2% for the total regression estimator $\hat{t}_R = t_a + \hat{t}_{Rb}$. Using the methods of Pandher (1995), the best value for the heteroscedasticity parameter γ in Ontario was determined to be $\hat{\gamma} = 2$; the corresponding proportionality constant was estimated to be $\hat{c} = .0825$. The near optimal sample design defined by $\lambda = \hat{\gamma}$ $(p(s; \hat{\gamma}))$ was used.

Table 1
Application of Combined Methodology to LGF Survey Data (Ontario, 1989)

Iteration (q)	n_q	$l_q(\lambda, n_q)$	n_{aq}^*	n_{bq}^*	n_q^*
0	108	39	39	18	57
1	57	16	16	34	50
2	50	12	12	38	50

For Ontario the combined scheme stopped at iteration $q^* = 2$. The globally optimal population partitioning between the take-all and sampled groups is $N_a^* = 16$ and $N_b^* = 777$. The new minimal total sample size is $n^* = 50$ with allocations $n_a^* = 16$ and $n_b^* = 34$. A total sample size reduction of $n_0 - n_2^* = 108 - 50 = 58$ is achieved at the desired CV of 2% for the regression estimator $\hat{t}_R = t_a + \hat{t}_{Rb}$.

6. CONCLUDING REMARKS

This paper provides a comprehensive methodology for identifying and implementing an efficient sample design for recurrent surveys of skewed populations. The combined procedure integrates the solution to the following three problems: i) identifying an efficient sample selection scheme, ii) constructing an efficient demarcation between the take-all and sampled population groups at a given sample size, and iii) determining the minimal sample size required to meet the precision constraint(s).

The equivalence result to the Transfer Algorithm (Pandher 1995) was used to create the take-all and sampled groups. The first two components were then combined with a sample size determination step through an iterative procedure. Under the stoping rule, the combined iterative procedure converges to a globally minimal sample size and optimal population partitioning. Results from the application of the proposed sample redesign methodology to the Local Government Survey in Ontario were reported. A 52% reduction in the total sample size was achieved for the regression estimator of the total $(\hat{t}_R = t_a + \hat{t}_{Rb})$ at the desired precision of CV = 2%.

ACKNOWLEDGEMENTS

The author would like to acknowledge the support of Public Institutions Division for sponsoring this work and thank M.P. Singh, H. Mantel, M.S. Kovacevic, S. Wu and the referees for their valuable comments on earlier drafts of the paper.

APPENDIX

A brief sketch of the development behind Theorem 2 (Equivalence Result) is given here; for technical details see Pandher (1995). The same paper also establishes the desirable mathematical properties of the Transfer Algorithm such as existence and optimality of solution as well as the optimality of the combined procedure.

Using the expression for the variance of $V^{(l)}(\hat{t}_{Rb}; \cdot)$ given in (3.1), the difference $V^{(l+1)} - V^{(l)}$ may be expressed as

$$\hat{V}^{(l+1)} - \hat{V}^{(l)} = c \frac{A(l)\ B(l)}{(n-l)\ (n-l-1)} \qquad \text{(A.1)}$$

where

$$A(l) = \sum_{j=1}^{N-l} x_{(j)}^{\lambda/2} - (n-l)\, x_{(N-l)}^{\lambda/2}$$

and

$$B(l) = \sum_{k=1}^{N-l} x_{(k)}^{\gamma-\lambda/2} - (n-l)\ x_{(N-l)}^{\gamma-\lambda/2}.$$

The condition $B(l) < 0$ may also be expressed as $n - l > R(l; \gamma - \lambda/2)$ where $R(l; \alpha) = \sum_{k=1}^{N-l} x_{(k)}^{\alpha} / x_{(N-l)}^{\alpha}$. Similarly, the condition $A(l) > 0$ corresponds to $n - l < R(l; \lambda/2)$. All possible states of the system defined by the Transfer Algorithm are summarized in Table A.1.

Table A.1
Outcomes for $V^{(l+1)} - V^{(l)} < 0$ and $V^{(l+1)} - V^{(l)} \geq 0$ in Terms of $n^{(l)} = n - l$

Behaviour of A and B	$V^{(l+1)} - V^{(l)} < 0$ Condition on $n^{(l)} = n - l$
$A(l) > 0$ $B(l) < 0$	$R(l; \gamma - \lambda/2) < n - l < R(l; \lambda/2)$ (T.1)
$A(l) < 0$ $B(l) > 0$	$R(l; \lambda/2) < n - l < R(l; \gamma - \lambda/2)$ (T.3)
	$V^{(l+1)} - V^{(l)} \geq 0$ Condition on $n^{(l)} = n - l$
$A(l) > 0$ $B(l) \geq 0$	$n - l \leq \min\{R(l; \lambda/2),\ R(l; \gamma - \lambda/2)\}$ (T.2)
$A(l) \leq 0$ $B(l) \leq 0$	$n - l \geq \max\{R(l; \lambda/2),\ R(l; \gamma - \lambda/2)\}$ (T.4)

The first column describes the behaviour of $A(l)$ and $B(l)$ leading to the outcome $V^{(l+1)} - V^{(l)} < 0$ and $V^{(l+1)} - V^{(l)} \geq 0$, respectively. The second column describes the equivalent condition in terms of $n^{(l)} = n - l$, $R(l; \gamma - \lambda/2)$, and $R(l; \lambda/2)$ corresponding to $V^{(l)} - V^{(l-1)} < 0$ and $V^{(l)} - V^{(l-1)} \geq 0$, respectively. An important condition required for the solution to the Transfer Algorithm $l^*(\lambda)$ is that $\pi_{(N-l)}(\lambda) < 1$ hold. It is easy to verify that $\pi_{(N-l)}(\lambda) < 1 \Rightarrow A(l) > 0$. In terms of the description for the Transfer Algorithm given in Table A.1, this condition means that the solution can occur only when both $A(l) > 0$ and $B(l) \geq 0$ or, equivalently, when $n - l$ satisfies condition (T.2).

Table A.1 completely enumerates all possible states of the system defined by the Transfer Algorithm. The correspondence between the internal cell quantities (computable directly from the auxiliary data and estimated parameters) and the margins ($A(l)$, $B(l)$, $V^{(l+1)} - V^{(l)}$) represents a tautology which leads directly to Theorem 2 (Equivalence Result). The behaviour of the system described in the table also depends on the sample design $p(s; \lambda)$ employed. The three relevant cases are:

a) $0 \leq \lambda < \gamma \Rightarrow [R(l; \gamma - \lambda/2) < R(l; \lambda/2)]$,

b) $\lambda = \gamma \Rightarrow [R(l; \gamma - \lambda/2) = R(l; \lambda/2)]$, and

c) $\gamma < \lambda \Rightarrow [R(l; \gamma - \lambda/2) > R(l; \lambda/2)]$.

In case a) the system starts ($l = 0$) in state (T.4), moves to (T.1) and then finally rests in state (T.2); state (T.3) is infeasible here. The solution to the Transfer Algorithm $l^*(\lambda)$ is given by the smallest l leading the system to move into state (T.2). In case b), the system starts in state (T.4) and moves to (T.2); (T.1) and (T.3) do not apply. Finally, in case c), the transition path is from (T.4) to (T.3) to (T.2); here (T.1) is invalid.

REFERENCES

GLASSER, G.J. (1962). On the complete coverage of large units in a statistical study. *Review of the International Statistical Institute,* 30, 28-32.

GODAMBE, V.P., and JOSHI, V.M. (1965). Admissibility and Bayes estimation in sampling finite populations. *Annals of Mathematical Statistics,* 36, 1702-1722.

HIDIROGLOU, M.A., and SRINATH, K.P. (1993). Problems associated with designing subannual business surveys. *Journal of Business and Economic Statistics,* 11, 397-405.

LAVALLÉE, P., and HIDIROGLOU, M.A. (1988). On the stratification of skewed populations. *Survey Methodology,* 14, 33-43.

PANDHER, G.S. (1995). Surveys of skewed populations: optimal sample redesign under the generalized regression estimator with applications to the Local Government Finance Survey. Methodology Branch Working Paper: HSMD-95-006, Statistics Canada.

SÄRNDAL, C.-E., SWENSSON, B., and WRETMAN, J. (1992). *Model Assisted Survey Sampling.* New York: Springer-Verlag.

ACKNOWLEDGEMENTS

Survey Methodology wishes to thank the following persons who have served as referees during 1996. An asterisk indicates that the person served more than once.

Y. Amamiya, *Iowa State University*
M. Bankier, *Statistics Canada*
* D.R. Bellhouse, *University of Western Ontario*
T.R. Belin, *University of California - Los Angeles*
W. Bell, *U.S. Bureau of the Census*
* D.A. Binder, *Statistics Canada*
F.J. Breidt, *Iowa State University*
K. Campbell, *Los Alamos National Laboratory*
C.M. Cassel, *VIKON, Statistics Sweden*
G. Chen, *PNN Laboratory*
G. Chen, *University of Regina*
J. Chen, *University of Waterloo*
C. Cordy, *Oregon State University*
C. D. Cowan, *Federal Deposit Insurance Corporation*
B.G. Cox, *Mathematica Policy Research, Inc.*
C. S. Davis, *University of Iowa*
* J. Denis, *Statistics Canada*
J.-C. Deville, *INSEE*
A.H. Dorfman, *U.S. Bureau of Labor Statistics*
J.D. Drew, *Statistics Canada*
J.-J. Droesbeke, *Université Libre de Bruxelles*
P. Duchesne, *Université de Montréal*
J.L. Eltinge, *Texas A&M University*
L. Ernst, *U.S. Bureau of Labor Statistics*
J.T. Fagan, *U.S. Bureau of the Census*
W.A. Fuller, *Iowa State University*
* J. Gambino, *Statistics Canada*
Y. Gervais, *Statistics Canada*
R.M. Groves, *University of Maryland*
* M.A. Hidiroglou, *Statistics Canada*
D. Holt, *Central Statistical Office, U.K.*
E. Hoy, *U.S. Bureau of the Census*
E. Johnson, *Educational Testing Service*
C. Julien, *Statistics Canada*
G. Kalton, *Westat, Inc.*
* P.S. Kott, *National Agricultural Statistical Service*
M. Kovacevic, *Statistics Canada*
R. Lachapelle, *Statistics Canada*
M. Latouche, *Statistics Canada*
P. Lavallée, *Statistics Canada*
S. Linacre, *Australian Bureau of Statistics*
* D. Malec, *National Center for Health Statistics*
* H. Mantel, *Statistics Canada*
H. Mariotte, *INSEE*
A. Mason, *East-West Center*
N. Mathiowetz, *University of Maryland*
S.M. Miller, *U.S. Bureau of Labor Statistics*
P.L. do Nascimento Silva, *University of Southampton*
L. Norberg, *Statistics Sweden*
D. Pfeffermann, *Hebrew University*
J. Qian, *National Opinion Research Center*
R. Raby, *Statistics Canada*
T.E. Raghunathan, *University of Michigan*
* J.N.K. Rao, *Carleton University*
* L.-P. Rivest, *Université Laval*
G. Roberts, *Statistics Canada*
K. Rust, *Westat, Inc.*
I. Sande, *Bell Communications Research, U.S.A.*
* C.-E. Särndal, *Université de Montréal*
* W.L. Schaible, *U.S. Bureau of Labor Statistics*
N. Schenker, *University of California - Los Angeles*
F.J. Scheuren, *George Washington University*
J. Sedransk, *Case Western Reserve University*
J.P. Shaffer, *Educational Testing Service*
G.M. Shapiro, *Westat Inc.*
* A.C. Singh, *Statistics Canada*
* M.P. Singh, *Statistics Canada*
* R. Sitter, *Simon Fraser University*
C.J. Skinner, *University of Southampton*
E.A. Stasny, *Ohio State University*
* D. Stukel, *Statistics Canada*
* R. Thomas, *Carleton University*
M. Thompson, *University of Waterloo*
Y. Tillé, *École nationale de statistique et de l'analyse de l'information*
D. Tiller, *U.S. Bureau of Labor Statistics*
R. Tourangeau, *National Opinion Research Center*
R. Valliant, *U.S. Bureau of Labor Statistics*
R.B.P. Verma, *Statistics Canada*
V.K. Verma, *University of Essex*
P.J. Waite, *U.S. Bureau of the Census*
J. Waksberg, *Westat, Inc.*
W.E. Winkler, *U.S. Bureau of the Census*
K.M. Wolter, *National Opinion Research Center*
S. Wu, *Statistics Canada*
E. Zanutto, *Harvard University*
* A. Zaslavsky, *Harvard University*

Acknowledgements are also due to those who assisted during the production of the 1996 issues: S. Beauchamp and L. Durocher (Composition Unit) and L. Perreault (Official Languages and Translation Division). Finally we wish to acknowledge S. DiLoreto, S.F. Bertrand, C. Larabie and D. Lemire of Household Survey Methods Division, for their support with coordination, typing and copy editing.

The Statistician

JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES D

CONTENTS VOLUME 45, No. 2, 1996

THE ROYAL STATISTICAL SOCIETY, 12 ERROL ST, LONDON, EC1Y 8LX, UK

CONTENTS TABLE DES MATIÈRES

Volume 24, No. 3, September/septembre 1996

Volume 24, No. 4, December/décembre 1996

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents

Volume 12, Number 1, 1996

Volume 12, Number 2, 1996

All inquires about submissions and subscriptions should be directed to the Chief Editor:
Lars Lyberg, R&D Department, Statistics Sweden, S - 115 81 Stockholm, Sweden.

GUIDELINES FOR MANUSCRIPTS

Before having a manuscript typed for submission, please examine a recent issue (Vol. 19, No. 1 and onward) of *Survey Methodology* as a guide and note particularly the following points:

1. 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.