SURVEY METHODOLOGY

Catalogue No. 12-001-XPB

A JOURNAL PUBLISHED BY STATISTICS CANADA

JUNE 1997

•

VOLUME 23

•

NUMBER 1

Statistics Canada Statistique Canada

Canada

SURVEY METHODOLOGY

A JOURNAL PUBLISHED BY STATISTICS CANADA

JUNE 1997 • VOLUME 23 • NUMBER 1

Published by authority of the Minister
responsible for Statistics Canada

July 1997

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 $47 per year in Canada and US $47 per year Outside Canada. Subscription order should be sent to Statistics Canada, Operations and Integration Division, Circulation Management, 120 Parkdale Avenue, Ottawa, Ontario, Canada K1A 0T6 or by dialling (613) 951-7277 or 1 800 700-1033, by fax (613) 951-1584 or 1 800 889-9734 or by Internet: order@statcan.ca. 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 23, Number 1, June 1997

CONTENTS

Survey Methodology, June 1997
Vol. 23, No. 1, pp. 1-2
Statistics Canada

In This Issue

This issue of *Survey Methodology* contains articles on a variety of topics. Stafford and Bellhouse, in the first paper, present the basic building blocks to develop a comprehensive computer algebra for survey sampling theory. They show that three basic techniques in sampling theory depend on the repeated application of rules that give rise to partitions. The methodology is illustrated through applications to moment calculation of the sample mean, the ratio estimator and the regression estimator under the special case of simple random sampling without replacement. The machine application to the methodology described was done in the programming language *Mathematica.*

Hinkins, Oh and Scheuren introduce a new strategy for analysis of data from complex surveys. They draw a sub-sample in such a way that the sub-sample may be considered to be a simple random sample from the original population and then apply standard procedures for IID data. They suggest repeating the procedure many times to recover information lost in sub-sampling the original sample. They show how to implement their approach for stratified element sampling, for one and two stage cluster sampling, and for two PSU per stratum designs.

Nascimento Silva and Skinner consider the problem of variable selection for regression estimation. They develop an approach based on minimizing the mean squared error of the resultant estimator. They empirically compare their approach to others using data from a 1988 test of Brazilian census procedures; the proposed procedures have good bias and mean squared error properties.

Eltinge and Yansaneh study the problem of formation of nonresponse adjustment cells. Within the general paradigms of estimated-probability and estimated-item based cells, they consider a variety of diagnostics for evaluating a set of adjustment cells. The diagnostic procedures include: comparison of estimates and standard errors for different numbers of adjustment cells; assessment of within-cell bias; assessment of cell widths relative to precision of estimated response probabilities; and comparisons of cell-based estimates to the unadjusted estimate.

Kovačević and Yung conduct an empirical study to compare variance estimation methods for measures of income inequality estimated from complex survey data. Variance estimation methods included in the study are: jackknife; bootstrap; grouped balanced half-sample method; repeatedly grouped balanced half-sample method; and a Taylor method based on estimating equations. After comparing relative bias, relative stability, and coverage properties of associated confidence intervals for a number of income inequality measures, they conclude that the Taylor method works best with the bootstrap method coming second.

Humphreys and Skinner investigate the use of the instrumental variable estimation method for estimation of gross flows among discrete states. This approach may be useful when external estimates of misclassification rates are not available. They numerically illustrate their method using data from the U.S. Panel Study of Income Dynamics and the two states "employed" and "not employed". They show that when measurement error is present, the unadjusted estimates can have considerable bias; this problem may be overcome by using suitable instrumental variables.

Waksberg, Judkins and Massey discuss issues involved in oversampling geographical areas to produce estimates for small domains of the population in demographic surveys, in conjunction with household screening. An empirical evaluation of the variance reduction is presented, along with an assessment of the sampling robustness over time. Simultaneous geographic oversampling for estimation of several small domains is discussed.

Losinger, in his paper, proposes a modified random groups standard error estimator for data from the U.S. Decenial Census sample. The usual random groups estimator has two undesirable properties for binomial variables: estimates of standard error for the "yes" and "no" responses are not equal; if all respondents answer "yes" the estimated standard error is not equal to zero. The essential idea of the proposed modification is to apply a ratio adjustment to each subgroup estimate so that subgroup estimates of population agree with the total.

Finally, Zeelenberg gives a simple technique, which exploits the use of differentials, to linearize design-based, nonlinear estimators. Ultimately, the linearized expressions allow one to obtain simple Taylor-based expressions for the variances of the nonlinear estimators. He illustrates the technique using two examples: the regression coefficient estimator and the regression estimator.

The Editor

Survey Methodology, June 1997
Vol. 23, No. 1, pp. 3-10
Statistics Canada

A Computer Algebra for Sample Survey Theory

J.E. STAFFORD and D.R. BELLHOUSE[1]

ABSTRACT

A system of procedures that can be used to automate complicated algebraic calculations frequently encountered in sample survey theory is introduced. It is shown that three basic techniques in sampling theory depend on the repeated application of rules that give rise to partitions: the computation of expected values under any unistage sampling design, the determination of unbiased or consistent estimators under these designs and the calculation of Taylor series expansions. The methodology is illustrated here through applications to moment calculations of the sample mean, the ratio estimator and the regression estimator under the special case of simple random sampling without replacement. The innovation presented here is that calculations can now be performed instantaneously on a computer without error and without reliance on existing formulae which may be long and involved. One other immediate benefit of this is that calculations can be performed where no formulae presently exist. The computer code developed to implement this methodology is available via anonymous ftp at *fisher.stats.uwo.ca*.

KEY WORDS: *k*-statistics; Partitions; Product moments; Ratio and regression estimators; Symbolic computation; Variance estimation.

1. INTRODUCTION

In classical sampling theory two general problems concern us. These are the determination of an unbiased estimator of a parameter θ and the calculation of moments of $\hat{\theta}$, the estimator of θ.

The basic method to handle expectations and unbiased estimation is to operate on sample and population nested sums respectively through the inclusion probabilities, either single or joint probabilities as appropriate. A nested sum is a sum over the range of one or more indices such that each term in the sum depends on indices of different value. An unbiased estimator of any population nested sum is the associated sample nested sum with the quantity under the summation divided by the appropriate inclusion probability. Similarly the expectation of any sample nested sum is the associated population nested sum with the quantity under the summation multiplied by the appropriate inclusion probability.

In sampling theory, as well as several other areas of statistics, many algebraic calculations depend on a partition of some kind. With particular reference to sampling, Wishart (1952) showed that basic moment calculations under simple random sampling without replacement relied heavily on partitions. Here we will use partitions to express the sum of products of means or totals as linear combinations of nested sums and vice versa.

In the results presented here we consider the situation in which θ and $\hat{\theta}$ can be expressed as smooth functions of means or totals, population or sample as appropriate. There are two possibilities: the smooth function under consideration can be expressed as the sum of products of means or totals, or the smooth function cannot be so expressed. When the second possibility is operative the function $\hat{\theta}$ is first linearized through a Taylor expansion and θ is expressed as the root of an estimating equation. We use integer partitions to obtain terms in the Taylor linearization of a function or for the root of a function. The end result is that θ and $\hat{\theta}$ can be expressed, either exactly or approximately, as the sum of products of means or totals. These in turn can be expressed in terms of linear combinations of nested sums and vice versa. Estimation of θ or calculation of the moments of $\hat{\theta}$ is then a three step procedure: (a) Express an estimating equation for θ or the estimator $\hat{\theta}$ as the sum of products of means or totals, using Taylor linearization when necessary. (b) Transform the expression obtained in the first step to a linear combination of nested sums. Then operate on these nested sums to obtain unbiased estimates or expectations as appropriate. (c) Transform the resulting nested sums in the second step back into a sum a products of means or totals.

The key to automation of sampling theory results is the use of partitions. In general, whether these partitions are simple partitions, like that of an integer, or more complicated, like a full partition, each results from the repeated application of a fundamental rule. When the rule is identified, the possibility of automating a calculation arises. Seemingly unrelated formulae can result from the same fundamental rule and one computer algebra tool can be constructive in implementing many different calculations.

The notation used in the paper is outlined in §2. A discussion of expectation operators is given in §3. The concept of partitioning is reviewed in §4 and a rule is provided which leads to a simple recursive method for the enumeration of partitions. Integer partitions and Taylor linearization is discussed in §5. It is shown in §6 how the enumeration of partitions leads to the automatic calculation of expected values of products of sample means and *k*-statistics

[1] J.E. Stafford and D.R. Bellhouse, Department of Statistical and Actuarial Sciences, University of Western Ontario, London, Ontario, N6A 5B7.

and to the derivation of unbiased estimators of products of finite populations means and k-statistics. Also in this section we apply the methodology to ratio and regression estimation.

Automation of these calculations and derivations will provide procedures which can be performed instantaneously and without error on a computer. Also, the reliance on formulae which may be long and involved is eliminated. A great deal of hand written algebra can be avoided. All computer code for the implementation of the methodology described here was written in the symbolic package *Mathematica* 2.0 which was installed on an IBM Risc 6000 with 64 megabytes of RAM. It is available via anonymous ftp at *fisher.stats.uwo.ca*. Although we use *Mathematica*, implementation in other environments such as *Maple*, *Macsyma* or *Reduce* is no doubt possible. For example, Kendall (1993) describes a system, implemented in *Reduce*, for the identification of invariant expressions. For a complete review of computer algebra in probability and statistics prior to 1991, see Kendall (1993).

2. SOME NOTATION

Consider a finite population of size N. A measurement of interest y_j is made on each unit $j, j \in U = \{1, ..., N\}$. In addition a single auxiliary variable x_j or possibly a $P \times 1$ vector of auxiliary variables $\boldsymbol{x}_j$ may be taken on the units. The p-th entry of this vector $\boldsymbol{x}_j$ is x_{pj}, where $p = 1, ..., P$. Several kinds of finite population parameters may be defined on the measurements y_j, x_j, or $\boldsymbol{x}_j$ for $j = 1, ..., N$. We denote a finite population parameter of interest by θ. Often θ can be expressed as a smooth function of finite population means, central moments and k-statistics. For convenience here we will deal only with means and k-statistics. Note that finite population variances and covariances are also second order k-statistics.

Not all N population elements are observed. Suppose that a sample s of size n is chosen from the population U by some sampling scheme. An estimator of θ, given by $\hat{\theta}$, is a smooth function of sample means and sample k-statistics.

In order to avoid much cumbersome summation notation we adapt the index notation of McCullagh (1987) to our purposes. For any j the vector $\boldsymbol{x}_j$ contain P entries so that each of these x-variables may be associated with one of the P indices. Suppose $\{i_1, ..., i_m\}$ is a subset of m of these P indices. In our adaptation of McCullagh's notation, $x_{i_j j}$ is now what we called the vector $\boldsymbol{x}_j$. Products of these indexed quantities become multidimensional arrays. For example the product $x_{i_1 j} x_{i_2 j} x_{i_3 j}$ is a three-dimensional array of dimension $P \times P \times P$.

Let M denote a finite population mean. The argument of M shows the structure of the summand in the mean. For example, $M(y) = \sum_{j \in U} y_j / N$ and $M(yy)$ or equivalently $M(y^2) = \sum_{j \in U} y_j^2 / N$. In index notation, for example,

$$M(x_{i_1} x_{i_2} x_{i_3}) = \sum_{j \in U} x_{i_1 j} x_{i_2 j} x_{i_3 j} / N \tag{1}$$

is a three-dimensional array. An element of this array is the mean of products in one of the permutations of the P elements taken three at a time in $\boldsymbol{x}_j$ where up to three of the elements may be alike. The (p,q,r)-th element of this array is $\sum_{j \in U} x_{pj} x_{qj} x_{rj}$ where $p, q, r = 1, ..., P$. The sample mean is denoted by m so that, for example,

$$m(x_{i_1} x_{i_2} x_{i_3}) = \sum_{j \in s} x_{i_1 j} x_{i_2 j} x_{i_3 j} / n. \tag{2}$$

For the purpose of making asymptotic expansions, since the variance of a given estimator $\hat{\theta}$ will be $O(n^{-1})$, we define a standardized variable for $\hat{\theta}$: it is the original variable $\hat{\theta}$ centered about its expectation and scaled by $1/\sqrt{n}$. That is,

$$z(\hat{\theta}) = [\hat{\theta} - E(\hat{\theta})] \sqrt{n}. \tag{3}$$

When necessary we use the summation convention of McCullagh (1987), where subscripts repeated as superscripts indicate implicit sums over that index. As a particular example, on assuming that the $\boldsymbol{x}_j$ are independent and identically distributed vectors from some infinite superpopulation, multivariate superpopulation moments can be obtained through the moment generating function which is expressed in this convention as

$$\mathrm{MGF}(\boldsymbol{t}) = 1 + \sum_{h=1}^{\infty} \mu_{i_1 \cdots i_h} \prod_{j=1}^{h} t^{i_j} / h!, \tag{4}$$

where

$$\mu_{i_1 \cdots i_h} = \frac{\partial^h}{\partial t_{i_1} \cdots \partial t_{i_h}} \mathrm{MGF}(\boldsymbol{t})\big|_{\boldsymbol{t}=0}. \tag{5}$$

By definition, the relationship between the moment generating function and the cumulant generating function is determined by the rule $\mathrm{MGF}(\boldsymbol{t}) = \exp\{\mathrm{K}(\boldsymbol{t})\}$, where

$$\mathrm{K}(\boldsymbol{t}) = \sum_{h=1}^{\infty} \kappa_{i_1 \cdots i_h} \prod_{j=1}^{h} t^{i_j} / h! \tag{6}$$

is the cumulant generating function, where

$$\kappa_{i_1 \cdots i_h} = \frac{\partial^h}{\partial t_{i_1} \cdots \partial t_{i_h}} \mathrm{K}(\boldsymbol{t})\big|_{\boldsymbol{t}=0}.$$

The finite population k-statistics, denoted by $K(\cdot)$, are defined as the unbiased (under the i.i.d. superpopulation model) estimators of the associated model cumulants. The number of arguments in K separated by commas denotes the order of the k-statistic. For example, the third order k-statistic $K(x_{i_1}, x_{i_2}, x_{i_3})$ is the model-unbiased estimate of (6), where

$$K(x_{i_1}, x_{i_2}, x_{i_3}) = \frac{N}{(N-1)(N-2)} \times \sum_{j \in U} [x_{i_1 j} - M(x_{i_1})][x_{i_2 j} - M(x_{i_2})][x_{i_3 j} - M(x_{i_3})]. \tag{7}$$

In the univariate case finite population k-statistics are described in Wishart (1952). In particular $K(y,y)$ and $K(y,y,y)$ in the current notation are K_2 and K_3 in Wishart's (1952) notation. The sample k-statistics, denoted by $k(\cdot)$ with the appropriate arguments, are defined as the unbiased

estimators under simple random sampling without replacement of the associated finite population k-statistics. As in Wishart (1952) the sample k-statistic can be obtained from the population k-statistic upon replacing N by n and upon taking the sum over $j \in s$ rather than all units in the finite population. For example,

$$k(x_{i_1}, x_{i_2}, x_{i_3}) = \frac{n}{(n-1)(n-2)} \times \sum_{j \in s} [x_{i_1 j} - m(x_{i_1})][x_{i_2 j} - m(x_{i_2})][x_{i_3 j} - m(x_{i_3})].$$

Note that if a comma is not present in the population or sample k-statistic, then the product of elements which appear together is required. For example, $K(xy)$ is the first order finite population k-statistic of a new variable which is the product of the measurements x_j and y_j for $j = 1, \ldots, N$; $K(x,y)$ is a second order k-statistic, in particular the finite population covariance between x and y.

3. OPERATORS

The expectation operator E can be applied directly to any sample nested sum to obtain a finite population nested sum. Likewise an unbiased estimator of any finite population nested sum is a sample nested sum. In terms of triple nested sums, for example,

$$E\left[\sum_{J_3 \in s} x_{i_1 j} x_{i_2 k} x_{i_3 l}\right] = \sum_{J_{3=1}}^{N} \pi_{jkl} x_{i_1 j} x_{i_2 k} x_{i_3 l} \qquad (8)$$

and

$$\sum_{J_{3=1}}^{N} x_{i_1 j} x_{i_2 k} x_{i_3 l} \sim \sum_{J_3 \in s} x_{i_1 j} x_{i_2 k} x_{i_3 l} / \pi_{jkl}, \qquad (9)$$

where J_3 is the index set $\{j, k, l\}$ such that $j \neq k \neq l$ and where π_{jkl} is a joint inclusion probability. Parallel expressions may be established for with replacement sampling schemes.

Note that m will be unbiased for the associated M under simple random sampling without replacement. In general for any sampling design of fixed size n,

$$E[m(x_{i_1} x_{i_2} x_{i_3})] = \frac{N}{n} M(x_{i_1} x_{i_2} x_{i_3} \pi)$$

and

$$M(x_{i_1} x_{i_2} x_{i_3}) \sim \frac{n}{N} m(x_{i_1} x_{i_2} x_{i_3} / \pi)$$

where $M(x_{i_1} x_{i_2} x_{i_3})$ and $m(x_{i_1} x_{i_2} x_{i_3})$ are defined in (1) and (2) respectively.

The whole operation of finding expectation of an estimator $\hat{\theta}$ or of finding an unbiased estimator for the parameter of θ may be represented schematically as

$$\sum\prod \rightarrow \sum\sum \rightarrow \sum\prod, \qquad (10)$$

where $\sum\prod$ denotes the sum of products and $\sum\sum$ denotes a sum of nested sums. If θ or $\hat{\theta}$ can be expressed as a $\sum\prod$ quantity, *i.e.*, a sum of products of means, then finding an unbiased estimator of θ or moments of $\hat{\theta}$ reduces to following the schema in (10) and applying the appropriate operator, such as those given in (8) or (9), to $\sum\sum$, the middle step in the schema. If θ or $\hat{\theta}$ are smooth functions of means but cannot be expressed directly as $\sum\prod$ quantities, then an initial step is required before applying the schema in (10). For $\hat{\theta}$ the initial step is to obtain a Taylor expansion of $\hat{\theta}$. For θ the initial step is to obtain an estimating equation and then to solve it for the parameter.

We illustrate the schema in (10) by considering the simple case of finding $E[\{m(x_{i_1})\}^2]$ under simple random sampling without replacement. The first operation is to express $\{m(x_{i_1})\}^2$ in terms of nested sums. In particular,

$$\{m(x_{i_1})\}^2 = \frac{1}{n^2} \sum_{j \in s} x_{i_1 j}^2 + \frac{1}{n^2} \sum_{j \neq k \in s} x_{i_1 j} x_{i_1 k}. \qquad (11)$$

This is the $\sum\prod \rightarrow \sum\sum$ step. Now the expectation operator can be applied to $\sum\sum$. On applying inclusion probabilities $\pi_j = n/N$ and $\pi_{jk} = n(n-1)/[N(N-1)]$, the expectation operation on (11) yields

$$\frac{1}{n^2} \frac{n}{N} \sum_{j=1}^{N} x_{i_1 j}^2 + \frac{1}{n^2} \frac{n(n-1)}{N(N-1)} \sum_{j \neq k=1}^{N} x_{i_1 j} x_{i_1 k}. \qquad (12)$$

Now the $\sum\sum \rightarrow \sum\prod$ step is applied. On expressing the nested sum in (12) as the sum of products, in particular $\sum_{j \neq k=1}^{N} x_{i_1 j} x_{i_1 k} = \sum_{j=1}^{N} x_{i_1 j} \sum_{j=1}^{N} x_{i_1 j} - \sum_{j=1}^{N} x_{i_1 j} x_{i_1 j}$, the third operation yields

$$E[\{(m(x_{i_1})\}^2] = \frac{N(n-1)}{(N-1)n} \{M(x_{i_1})\}^2 + \frac{N-n}{n(N-1)} M(x_{i_1}^2). \qquad (13)$$

In (13), $M(x_{i_1}) = K(x_{i_1})$ and $M(x_{i_1}^2) = [N/(N-1)] K(x_{i_1}, x_{i_1}) + K(x_{i_1}) K(x_{i_1})$ so that (13) can be reexpressed as

$$E(m(x_{i_1})^2) = \{K(x_{i_1})\}^2 + (N-n) K(x_{i_1}, x_{i_1})/(Nn). \qquad (14)$$

Likewise, following the schema in (10), the operations for finding an unbiased estimator of, for example, $\{M(x_{i_1})\}^2$ is similar to (11), (12) and (13). The estimand $\{M(x_{i_1})\}^2$ is expressed in nested sums similar to (11). These sums will be nested finite population sums. Similar to (12) the inclusion probabilities are applied. In this case the finite population sums are replaced by sample sums and summand is divided by the appropriate inclusion probability. Finally, similar to (13) the resulting nested sample sums are expressed as products of sums.

Each of the elementary operations to obtain an expected value through equations (11), (13) and (14), or to obtain an unbiased estimator, can be carried out using partitions. These operations are: expressing sums of products as nested sums and vice versa, and expressing means in terms of k-statistics and vice versa.

4. PARTITIONS AND FUNDAMENTAL PROCEDURES

Central to the automation of all algebraic calculations considered here is the notion of a partition. Partitioning as a focal point gives the appearance that the automated methods presented here are nothing more than an integer partition or a partition of an index set. While we assume that a partition of an integer is understood, a full partition requires a more formal definition.

Consider a set of m indices $I_m = \{i_1, ..., i_m\}$. A single partition P_m of I_m divides the m indices into $k \le m$ mutually exclusive and exhaustive subsets or blocks of I_m. We write $P_m = (b_1 \mid b_2 \mid ... \mid b_k)$, where the $b_1, ..., b_k$ are the blocks of I_m. P_m is unique up to permutations of indices within the blocks b_i. The block b_i is comprised of a subset of the indices of I_m. Elements within a block may be constrained to an alphabetical ordering and the blocks themselves may be ordered such that leading elements of each block are ordered alphabetically. This ensures the uniqueness of the partition P_m. In this case P_m would be called a standard ordered partition. Ordering the partitions in this manner does not offer any computational advantage and hence is not a requirement in what follows. The full partition of I_m is the set $\wp_m$ of all single partitions P_m of I_m.

Now we may identify the full partition of I_m in an algorithmic way via an inclusion-exclusion rule.

i. Let $\wp_1 = \{i_1\}$.

ii. An inclusion-exclusion rule determines the contribution to $\wp_t$ by a partition $P_{t-1} \in \wp_{t-1}$. In the inclusion part of the rule, the new index i_t is added as an element in turn to each of the blocks $b_1, ..., b_k$ which comprise P_{t-1}. If P_{t-1} has k blocks, k partitions for $\wp_t$ are created. In the exclusion part of the rule a new block containing the single index i_t is added to P_{t-1}.

For example, the full partition of $I_3 = \{i_1, i_2, i_3\}$ is given by the steps

i. $\wp_1 = \{(i_1)\}$

ii. $\wp_2 = \{(i_1 i_2), (i_1 \mid i_2)\}$ (15)

iii. $\wp_3 = \{(i_1 i_2 i_3), (i_1 i_2 \mid i_3), (i_1 i_3 \mid i_2), (i_1 \mid i_2 i_3), (i_1 \mid i_2 \mid i_3)\}$.

From step (i) to step (ii) the inclusion rule results in the partition $(i_1 i_2)$ and the exclusion rule results in $(i_1 \mid i_2)$. From step (ii) to step (iii) the inclusion rule results in the creation of the partitions $(i_1 i_2 i_3)$, $(i_1 i_3 \mid i_2)$, and $(i_1 \mid i_2 i_3)$. The exclusion rule yields the partitions $(i_1 i_2 \mid i_3)$ and $(i_1 \mid i_2 \mid i_3)$. This type of construction is easy to automate since it depends on a simple rule. Details of automating the partition of indices into full partitions and complementary set partitions are given in Stafford (1996).

Consider, for example, the classical problem of writing the model moments of the random vector x_{i_1} in terms of its cumulants. As in (5) we can identify the h-th moment array by differentiating MGF(t) in (4) h times and setting t equal to the zero vector. The result is the h-th coefficient in the expansion of MGF(t). Equivalently we can apply the same operation to $\exp\{K(t)\}$. In this case the result is a sum that depends on the coefficients of $K(t)$ in (6). For example, we may write the first three moments in terms of cumulants as follows:

$$\begin{aligned}
\mu_{i_1} &= \kappa_{i_1} \\
\mu_{i_1 i_2} &= \kappa_{i_1 i_2} + \kappa_{i_1}\kappa_{i_2} \\
\mu_{i_1 i_2 i_3} &= \kappa_{i_1 i_2 i_3} + \kappa_{i_1 i_2}\kappa_{i_3} + \kappa_{i_1 i_3}\kappa_{i_2} + \kappa_{i_1}\kappa_{i_2 i_3} + \kappa_{i_1}\kappa_{i_2}\kappa_{i_3}.
\end{aligned}$$

Now in each case the result is a sum over the full partitions given in (15). These partitions arise since the multiplication rule for differentiation mimics the inclusion-exclusion rule for the enumeration of the full partition.

The above result is applied to sampling theory where we consider the problem of finding the expected value of a product of sample sums. The calculation requires expanding the product of the sums to identify terms where the finite population expectation operator will behave differently due to differences in the values of inclusion probabilities and joint inclusion probabilities.

For example, the product of sums $\sum_{j\in s} x_{i_1 j} \sum_{j\in s} x_{i_2 j} \sum_{j\in s} x_{i_3 j}$ can be expressed as

$$\sum_{j\in s} x_{i_1 j} x_{i_2 j} x_{i_3 j} + \sum_{j*k\in s} x_{i_1 j} x_{i_2 j} x_{i_3 k} + \sum_{j*k\in s} x_{i_1 j} x_{i_2 k} x_{i_3 j} + \sum_{j*k\in s} x_{i_1 k} x_{i_2 j} x_{i_3 j} + \sum_{j*k*l\in s} x_{i_1 j} x_{i_2 k} x_{i_3 l}. \quad (16)$$

The result corresponds to the full partition of the indices $I_3 = \{i_1, i_2, i_3\}$ given by $\wp_3$ in (15). The order of the partitions in $\wp_3$ is the same as the order given for the terms in (16). For each partition in $\wp_3$, the variables in the same block have the same second index in the appropriate term in (16). For example, the partition $(i_1 i_3 \mid i_2)$ corresponds to the term $\sum_{j*k\in S} x_{i_1 j} x_{i_2 k} x_{i_3 j}$ in (16). Each term in the result can be identified by a partition of I_3 and each partition determines the manner in which the expected value operator will behave.

In general, we want to expand products of the form $\prod_{r=1}^{m} \sum_{j\in s} x_{i_r j}$, where the product is taken over the elements i_r of the index set $I_m = \{i_1, ..., i_m\}$. As in (16), the product can be expressed in terms of the full partition of I_m. This is because the iterative rule for expanding a product of sums mimics the inclusion-exclusion rule.

The expansion of the products of sums through partitions is demonstrated inductively as follows. Assume the product of the first $t - 1$ sums can be expressed as a sum over the full partition of the index set $I_{t-1} = \{i_1, ..., i_{t-1}\}$, in particular

$$\prod_{r=1}^{t-1} \left(\sum_{j\in s} x_{i_r j}\right) = \sum_{P_{t-1}\in \wp_{t-1}} X_{P_{t-1}}. \quad (17)$$

In (17) the term $X_{P_{t-1}}$ is the sum identified by the partition $P_{t-1} = (b_1 \mid ... \mid b_k)$, $k = 1, ..., t-1$. The blocks b_j indicate groups of variables with the same second index and so P_{t-1} induces an index set $J_k = \{j_1, ..., j_k\}$ of second indices. We can express $X_{P_{t-1}}$ as

$$X_{P_{t-1}} = \sum_{j_1 * \ldots * j_k \in s} \left(\prod_{j \in J_k} X_{b_j} \right), \qquad (18)$$

where X_{b_j} is a product of x's defined by the block b_j that all have the same second index. To illustrate (18), consider, for example, the third term of (16). Here $P_{t-1} = (i_1 i_3 | i_2)$ and $J_2 = \{j, k\}$ so that in (18) the sum is taken over $j \neq k \in s$ and the multiplicands of the product are $X_{b_j} = x_{i_1 j} x_{i_3 j}$ and $X_{b_k} = x_{i_2 k}$. Returning to the general discussion, when either side of (17) is multiplied by $\sum_{j \in s} x_{i_t j}$ the product of the first t sums is obtained. Now the product $X_{P_{t-1}} \sum_{j \in s} x_{i_t j}$ can be expressed as

$$\sum_{j_1 * \ldots * j_k \in s} \left(\sum_{l=1}^{k} x_{i_t j_l} \prod_{j \in J_k} X_{b_j} \right) + \sum_{j_1 * \ldots * j_k * j_{k+1} \in s} \left(\prod_{j \in J_k} X_{b_j} x_{i_t j_{k+1}} \right). \qquad (19)$$

The first term in (19) corresponds to the inclusion part of the rule and the second term in (19) corresponds to the exclusion part of the rule. When (19) is summed over all $P_{t-1} \in \wp_{t-1}$, the result will be a sum over the full partition of the first t indices given by I_t, *i.e.*, the sum over all $P_t \in \wp_t$.

Once the product of sums, $\prod_{r=1}^{m} \sum_{j \in s} x_{i_r j}$, is expanded into a sum of nested sums, the finite population expected value operator can be applied to each term so that the expected value of this product can be obtained. The expected value under simple random sampling without replacement of the product of sums results in a weighted sum of nested sums, with each sum taken over the finite population. We then wish to evaluate these nested sums.

In general we wish to evaluate the nested sum $\sum_{J_t} Y_{J_t}$ where J_t is the index set $\{j_1, \ldots, j_t\}$. The sum is taken over all $j_1 * \ldots * j_t$ with each $j_r = 1, \ldots, N$. The summand Y_{J_t} is the product $x_{i_1 j_1} x_{i_2 j_2} \ldots x_{i_t j_t}$. In the special case when $t = 3$ or $J_3 = \{j, k, l\}$ the nested sum can be written in terms of full sums as

$$\sum_{J_3} Y_{jkl} = \sum_{j * k * l = 1}^{N} Y_{jkl} = \sum_{j * k * l = 1}^{N} x_{i_1 j} x_{i_2 k} x_{i_3 l} =$$

$$2 \sum_{j=1}^{N} x_{i_1 j} x_{i_2 j} x_{i_3 j} - \sum_{j=1}^{N} x_{i_1 j} x_{i_2 j} \sum_{j=1}^{N} x_{i_3 j} - \sum_{j=1}^{N} x_{i_1 j} x_{i_3 j} \sum_{j=1}^{N} x_{i_2 j} -$$

$$\sum_{j=1}^{N} x_{i_1 j} \sum_{j=1}^{N} x_{i_2 j} x_{i_3 j} + \sum_{j=1}^{N} x_{i_1 j} \sum_{j=1}^{N} x_{i_2 j} \sum_{j=1}^{N} x_{i_3 j}. \qquad (20)$$

Note that the full sums in the rightmost expression in (20) result from the full partition $\wp_3$ in (15). The order of the partitions in $\wp_3$ is the same as the order of the terms on the right of (20). The subscripts on the right of (20) denote the block membership in $\wp_3$. For example, the partition $(i_1 i_3 | i_2)$ corresponds to the term $\sum_{j=1}^{N} x_{i_1 j} x_{i_3 j} \sum_{j=1}^{N} x_{i_2 j}$ in (20). Note also from (20) that the determination of a nested sum is complicated by the additional determination of the appropriate coefficients of the full sums.

In general the evaluation of finite population nested sums results from the repeated application of the rule

$$\sum_{j_1 * \ldots * j_t = 1}^{N} \left(\prod_{r=1}^{t} x_{i_r j_r} \right) = \sum_{j_1 * \ldots * j_{t-1} = 1}^{N} \left[\prod_{r=1}^{t-1} x_{i_r j_r} \left(\sum_{j_t = 1}^{N} x_{i_t j_t} \right) \right]$$

$$- \sum_{j_1 * \ldots * j_{t-1} = 1}^{N} \left[\sum_{l=1}^{t-1} x_{i_t j_l} \left(\prod_{r=1}^{t-1} x_{i_r j_r} \right) \right]. \qquad (21)$$

This expression mimics the inclusion-exclusion rule where the first set of sums on the right follows the exclusion part of the rule and the second set follows the inclusion part of the rule. Repeated application of (21) yields

$$\sum_{j_1 * \ldots * j_t = 1}^{N} \left(\prod_{r=1}^{t} x_{i_r j_r} \right) = \sum_{P_t \in \wp_t} (-1)^{|J_t| - |P_t|}$$

$$\times \left\{ \prod_{b_k \in P_t} \left[(|b_k| - 1)! \sum_{j=1}^{N} \left(\prod_{i_r \in b_k} x_{i_r j} \right) \right] \right\}$$

where $|J_t|, |P_t|$ and $|b_k|$ are the number of indices in J_t, the number of blocks in the single partition P_t and the number of elements in the block b_k respectively.

5. INTEGER PARTITIONS AND TAYLOR LINEARIZATION

Suppose that under some sampling design an estimator $\hat{\theta}$ of a parameter θ is of interest. The methodology described in §§2 to 4 may be used in moment calculations for $\hat{\theta}$ or to find unbiased estimators of these moments. Only in the simplest cases can this methodology be applied directly. Typically $\hat{\theta}$ must be linearized so that it becomes a polynomial function of sample means or sums which are $O_p(1)$ random variables with respect to the sampling design. Once $\hat{\theta}$ is linearized in this way the methodology of §§2 to 4 is applicable.

The objective of the linearization is to write $\hat{\theta}$ as an asymptotic expansion where terms descend in order by $1/\sqrt{n}$, specifically

$$\hat{\theta} = \hat{\theta}_0 + \hat{\theta}_1 / \sqrt{n} + \hat{\theta}_2 / n + \ldots, \qquad (22)$$

where $\hat{\theta}_i$ is the coefficient of the $n^{-i/2}$ term. Typically $\hat{\theta}$ is a product of quantities that can also be expanded in this way. For example, if the measurement of interest is y and one auxiliary variable x is present then θ might be $M(y)$ and the auxiliary information available is $M(x)$ as well as x_j for $j \in s$. Then $\hat{\theta} = M(x)m(y)/m(x)$, the simple ratio estimator, is a product of three quantities $M(x), m(y)$ and $1/m(x)$ all having asymptotic expansions of their own. The expansion of $M(x)$ is itself. From (3) the expansion for $m(y)$ yields $M(y) + z(m(y))/\sqrt{n}$. The expansion for $1/m(x)$ results from (3) and then applying a Taylor expansion to $[M(x) + z(m(x))/\sqrt{n}]^{-1}$.

In general any expansion of a function with sufficient regularity can be found if operators are defined to expand a function, say $g(\dot{e})$ where $\dot{e}$ is itself an expansion. We are interested in expanding functions of the form

$$g(\dot{e}) = \prod_{j=1}^{p} \dot{e}_j. \tag{23}$$

where $\dot{e}_j$ itself has the expansion $\sum_{i=0}^{\infty} e_{ij} n^{-i/2}$. In linearizing $\hat{\theta}$ the basic requirement is to define an operator that returns $\hat{\theta}_i$ in (22). The efficiency of this operator derives solely from a rule for expanding functions of the form given in (23). The calculations required are functions of integer partitions. For example the $1/n$ term in the expansion of $\prod_{j=1}^{3} \dot{e}_j$ is

$$e_{21}e_{02}e_{03} + e_{01}e_{22}e_{03} + e_{01}e_{02}e_{23} + e_{11}e_{12}e_{13} + e_{11}e_{02}e_{13} + e_{01}e_{12}e_{13}. \tag{24}$$

Collecting first indices for each term in the sum results in a list in which each element sums to 2: {(2,0,0), (0,2,0), (0,0,2), (1,1,0), (1,0,1), (0,1,1)}. On noting that the order $n^{-i/2}$ term in any expansion $\dot{e}_j$ is actually the $(i+1)$-th term in the sum $\sum_{i=0}^{\infty} e_{ij} n^{-i/2}$, we may modify the list derived from (24) so that entries identify the position of terms in a sum. The modification is to add 1 to each index value in the list. In the list derived from (25) this results in all partitions of the integer 5 into 3 blocks: {(3,1,1), (1,3,1), (1,1,3), (2,2,1), (2,1,2), (1,2,2)}. In general, the i-th term in the expansion of $\prod_{j=1}^{p} \dot{e}_j$ or $\dot{e}_j^p$, where p is a positive integer, is a sum over all partitions of the integer $i + p$ into p blocks. Consequently, using this methodology any term in the expansion of, for example, the ratio estimator can be found.

We illustrate this technique with ratio and regression estimation. The ratio estimator is given by

$$M(x)\,m(y)/m(x) \tag{25}$$

and the regression estimator by

$$k(y) + b[K(x) - k(x)] = k(y) + \frac{k(x,y)}{k(x,x)}[K(x) - k(x)] \tag{26}$$

in the notation of k-statistics.

On using (3) the ratio estimator (25) may be expressed as

$$M(x)\left[M(y) + \frac{z(y)}{\sqrt{n}}\right]\left[M(x) + \frac{z(x)}{\sqrt{n}}\right]^{-1}. \tag{27}$$

The expression in (27) may be expressed in terms of (24) with $p = 3$. The first term in (27) is the expansion $\sum_{i=0}^{\infty} e_{i1} n^{-i/2}$ with $e_{01} = M(x)$ and $e_{11} = e_{21} = \cdots = 0$. The first term in square brackets in (28) is the expansion $\sum_{i=0}^{\infty} e_{i2} n^{-i/2}$ where $e_{02} = M(y)$, $e_{12} = z(m(y))$ and $e_{22} = e_{32} = \cdots = 0$. The second term in square brackets is the expansion $\sum_{i=0}^{\infty} e_{i3} n^{-i/2}$ where $e_{i3} = (-1)^i \{z(m(y))\}^i / \{M(x)\}^{i+1}$. To get the $1/\sqrt{n}$ term in the expansion of (27), in which case $i = 1$ and $p = 3$, we need to find the integer partitions of 4 in blocks of 3. This yields the partitions (2,1,1), (1,2,1) and (1,1,2). On subtracting 1 from each index value in the list we obtain the list (1,0,0), (0,1,0), (0,0,1). Therefore the required term in the expansion is $(e_{11}e_{02}e_{03} + e_{01}e_{12}e_{03} + e_{01}e_{02}e_{13})/\sqrt{n}$ or equivalently $[z(m(y)) - M(y)z(m(x))/M(x)]/\sqrt{n}$. The $1/n$ term is obtained from (24) which reduces to

$$[M(y)\{z(x)\}^2/\{M(x)\}^2 - z(x)z(y)/M(x)]/n.$$

The regression estimator in (26) may be expressed as

$$K(y) + \frac{z(k(y))}{\sqrt{n}} + \left[K(x,y) + \frac{z(k(x,y))}{\sqrt{n}}\right] \times \left[K(x,x) + \frac{z(k(x,x))}{\sqrt{n}}\right]^{-1}\left[\frac{z(k(x))}{\sqrt{n}}\right] \tag{28}$$

using (3). The terms in the square brackets in (28) can be expanded in a similar fashion to the ratio estimator. In this case the terms in the expansions become: $e_{01} = K(x,y)$, $e_{11} = z(k(x,y))$ and $e_{21} = e_{31} = \cdots = 0$; $e_{i2} = (-1)^i \{z(k(x,x))\}^i / \{K(x,x)\}^{i+1}$ for $i = 0, 1, 2, \ldots$; and $e_{03} = 0$, $e_{13} = z(k(x))$ and $e_{23} = e_{33} = \cdots = 0$. Consequently, the $1/\sqrt{n}$ term in the expansion of the terms in the square brackets in (28) is

$$-\frac{K(x,y)z(k(x))}{K(x,x)\sqrt{n}}$$

and the $1/n$ term is

$$-\frac{1}{n}\left[\frac{z(k(x,y))}{K(x,x)} - \frac{K(x,y)z(k(x,x))}{K(x,x)^2}\right]z(k(x)).$$

These were obtained by the same argument that was used in the ratio estimator.

6. MACHINE APPLICATIONS TO THE CALCULATION OF EXPECTED VALUES OF SAMPLE STATISTICS AND THE DERIVATION OF UNBIASED ESTIMATORS

Since the machine application to the methodology described in §§3 to 5 was done in the programming language *Mathematica* we give a brief description of the operation of *Mathematica*. Then we describe the operators that were developed in *Mathematica* to provide a computer algebra for survey sampling theory.

Programming in *Mathematica* is carried out using expressions of the form $h[e_1, e_2, \ldots]$ where the object h is called the head of the expression and the e's are the elements of the expression. We have developed a number of machine expressions in *Mathematica* in the form of $h[e_1, e_2, \ldots]$ for operators which we apply to developing a computer algebra for sampling. All of these operators have been devised to

handle vectors as their arguments as well as scalars. There are four basic operators: $EV[\cdot]$ for expected value, $Cum[\cdot]$ for calculation of cumulants, $UE[\cdot]$ for unbiased estimator, and $Aexp[\cdot]$ for asymptotic expansion. There is also an operator to switch from notation using k-statistics to notation using means and vice versa.

The expected value operator $EV[\cdot]$ on sample statistics combines and carries out in *Mathematica* the three basic operations shown in the schema in (10). $EV[\cdot]$ contains two arguments, the first is the expression for which the expected value is to be obtained and the second is the sampling design which defines the inclusion probabilities. The application in *Mathematica* of $EV[\cdot]$ to $m(x_{i_1})m(x_{i_2})m(x_{i_3})$ under simple random sampling without replacement yields

$$K(x_{i_1})K(x_{i_2})K(x_{i_3}) + \frac{(N-n)(K(x_{i_1},x_{i_2})K(x_{i_3})}{Nn}$$

$$+ \frac{K(x_{i_1},x_{i_3})K(x_{i_2}) + K(x_{i_1})K(X_{i_2},x_{i_3}))}{Nn}$$

$$+ \frac{(N^2-3Nn+2n^2)K(x_{i_1},x_{i_2},x_{i_3})}{N^2n^2}$$

in the simplest expression of the output. Note that the result is a function of the full partition of $\{i_1,i_2,i_3\}$. If the operand is changed to $\{m(x_{i_1}) - M(x_{i_1})\} \times \{m(x_{i_2}) - M(x_{i_2})\} \times \{m(x_{i_3}) - M(x_{i_3})\}$, application of $EV[\cdot]$ yields

$$\frac{(N^2-3Nn+2n^2)K(x_{i_1},x_{i_2},x_{i_3})}{N^2n^2},$$

which was obtained by Nath (1968) for particular values of the indices i_1, i_2 and i_3. In fact, the results in Nath (1968, 1969) for the products of three and four means and the exact results in Raghunandanan and Srinivasan (1973) for up to a product of eight means can all be reproduced automatically with the software that has been developed.

To this point the sampling design used in each of the examples has been simple random sampling without replacement. Results under general sampling designs can be obtained. We illustrate these results for the operator $Cum[\cdot]$ which is used to obtain the cumulants of an estimator. Note that the second cumulant for an estimator is also the variance. The operator $Cum[\cdot]$ has three arguments. The first is an expression for the estimator, the second is the order of the cumulant and the third is the sampling design. Under general sampling designs, estimators can be expressed in terms of $\sum\prod$ in the schema given by (10) and the $\sum\prod$ can be expanded to obtain $\sum\sum$, the middle term in (10). There is, however, no general simplification to obtain the final term in (10). This is illustrated with the Horvitz-Thompson estimator of $M(y)$ given by $(n/N)m(y/\pi)$ in the notation developed here. Application of the operator $Cum[\cdot]$ under a general sampling design to obtain the third cumulant of the Horvitz-Thompson estimator yields

$$2\frac{\left\{\sum_{i=1}^{N} y_i\right\}^3}{N^3} - 3\frac{\left\{\sum_{i=1}^{N} y_i\right\}\left\{\sum_{i=1}^{N} \frac{y_i^2}{\pi_i}\right\}}{N^3} - 3\frac{\sum_{i=1}^{N} \frac{y_i^3}{\pi_i^2}}{N^3}$$

$$- 3\frac{\left\{\sum_{i=1}^{N} y_i\right\}\left\{\sum_{i=1}^{N}\sum_{j=1}^{N} \frac{\pi_{ij}y_iy_j}{(\pi_i\pi_j)}\right\}}{N^3} + 3\frac{\sum_{i=1}^{N}\sum_{j=1}^{N} \frac{\pi_{ij}y_iy_j^2}{(\pi_i\pi_j^2)}}{N^3}$$

$$+ \sum_{i=1}^{N}\sum_{h=1}^{N}\sum_{k=1}^{N} \frac{\frac{\pi_{ijk}y_iy_jy_k}{(\pi_i\pi_j\pi_k)}}{N^3}$$

where, for example, the term π_{ii} is the single inclusion probability π_i.

The operator $Aexp[\cdot]$ has two arguments, the function for which the expansion is required and the order of the expansion. This operator is used in combination with the $EV[\cdot]$ or $Cum[\cdot]$ operators to obtain approximate expectations or cumulants. This is illustrated in the case of the multiple linear regression estimator under simple random sampling without replacement. When there are q covariates the resulting regression estimator is given by

$$k(y) + b_{i_1}[K(x^{i_1}) - k(x^{i_1})] \qquad (29)$$

using index and k-statistics notation. In (29) the coefficient b_{i_1} is the vector resulting from the product $k(x_{i_1},y)\,ik(x^{i_1},x_{i_2})$ in index notation, where the $q \times q$ array $ik(x_{i_1},x_{i_2})$ is the inverse of the $q \times q$ array given by $k(x_{i_1},x_{i_2})$. Similarly we will use $IK(x_{i_1},x_{i_2})$ to denote the inverse of the finite population array $K(x_{i_1},x_{i_2})$. Derivation of the mean square error of (29) requires Taylor expansions of the elements of b_{i_1} followed by the appropriate moment calculations and collection of terms. The *Mathematica* command to obtain the approximate variance of (29) is obtained by first applying $Aexp[\cdot]$ to (29) with 2 as the order in the expansion. Then the operator $Cum[\cdot]$ is applied to the result with the following arguments: the result from the asymptotic expansion as the estimator, simple random sampling as the design and 2 for the order of the cumulant. This yields

$$\frac{(N-n)K(y,y)}{Nn} + \frac{(-N+n)K(x_{i_1},y)K(x_{i_2},y)IK(x^{i_1},x^{i_2})}{Nn}$$

in index notation as output.

Estimation is achieved through the operator $UE[\cdot]$ which has two arguments, the estimand and the sampling design. For example, application of $UE[\cdot]$ to $\{M(x)\}^2$ under simple random sampling yields

$$\frac{(Nn)\{k(x)\}^2 + (N-n)k(x,x)}{Nn}.$$

If the estimand cannot be expressed as a sum of nested sums, but instead can be expressed as the root of an estimating function, then $UE[\cdot]$ obtains a consistent estimator.

7. DISCUSSION OF FUTURE WORK

The basic building blocks to develop a comprehensive computer algebra for survey sampling theory have been given. The foundation of this algebra is based on the enumeration of partitions. Fundamental operations under partition enumeration include the evaluation of nested sums and Taylor series expansions. Once these operations have been completed then expectations of sample statistics can be calculated or unbiased estimators of population quantities can be determined.

The next phase in this work is to extend the unistage results to multistage and multiphase sampling. In both multistage and multiphase sampling the problem reduces to the computer evaluation of multiple sums under an expectation operator or the determination of an unbiased estimator of multiple finite population sums. The problem of multistage sampling is currently under investigation. Another current area of inquiry is to extend the algebra to superpopulation models.

Once the basic algebra is in place then research problems involving algebraically complex sampling formulae can be easily investigated.

ACKNOWLEDGEMENTS

The authors are grateful to David Andrews for some useful discussions on this topic. This work was supported by grants from the Natural Sciences and Engineering Research Councils of Canada and by a research contract from Statistics Canada.

REFERENCES

ANDREWS, D.F., and STAFFORD, J.E. (1993). Tools for the symbolic computation of asymptotic expansions. *Journal of the Royal Statistical Society (B)*, 55, 613-628.

KENDALL, W.S. (1993). Computer algebra in probability and statistics. *Statistica Neerlandica*, 47, 9-25.

McCULLAGH, P. (1987). *Tensor Methods in Statistics*. New York: Chapman and Hall.

NATH, S.N. (1968). On product moments from a finite universe. *Journal of the American Statistical Association*, 63, 535-541.

NATH, S.N. (1969). More results on product moments from a finite universe. *Journal of the American Statistical Association*, 64, 864-869.

RAGHUNANDANAN, K., and SRINIVASAN, R. (1973). Some product moments useful in sampling theory. *Journal of the American Statistical Association*, 68, 409-413.

STAFFORD, J.E. (1996). A note on symbolic Newton-Raphson, submitted for publication.

STAFFORD, J.E., and ANDREWS, D.F. (1993). A symbolic algorithm for studying adjustments to the profile likelihood. *Biometrika*, 80, 715-730.

WISHART, J. (1952). Moment coefficients of the k-statistics in samples from a finite population. *Biometrika*, 39, 1-13.

Survey Methodology, June 1997
Vol. 23, No. 1, pp. 11-21
Statistics Canada

Inverse Sampling Design Algorithms

SUSAN HINKINS, H. LOCK OH and FRITZ SCHEUREN[1]

ABSTRACT

In the main body of statistics, sampling is often disposed of by assuming a sampling process that selects random variables such that they are independent and identically distributed (IID). Important techniques, like regression and contingency table analysis, were developed largely in the IID world; hence, adjustments are needed to use them in complex survey settings. Rather than adjust the analysis, however, what is new in the present formulation is to draw a second sample from the original sample. In this second sample, the first set of selections are inverted, so as to yield at the end a simple random sample. Of course, to employ this two-step process to draw a single simple random sample from the usually much larger complex survey would be inefficient, so multiple simple random samples are drawn and a way to base inferences on them developed. Not all original samples can be inverted; but many practical special cases are discussed which cover a wide range of practices.

KEY WORDS: Finite population sampling; Inference in complex surveys; Resampling.

1. INTRODUCTION

The development of modern survey sampling is an extraordinary achievement (Bellhouse 1988; Hansen 1987; Kish 1995). The very richness in that development may have had the effect, though, of isolating survey sampling from the rest of statistics – where it is the richness of models that is given emphasis. In fact, it is a well-known commonplace that, in the main body of statistics, sampling is often disposed of by assuming a sampling process that selects random variables such that they are independent and identically distributed (IID).

Important techniques, like regression and contingency table analysis, were developed largely in this IID world; hence, adjustments are needed to use them in complex survey settings. Indeed, whole books have been written on this problem (Skinner, Holt and Smith 1989); and much time and effort have been devoted to it in software (like SUDAAN or WESVAR PC) specially written for surveys (See also Wolter 1985). With all that has been done already, can something more of value be added? We think we may have a contribution to offer on how to deal better with the "seam" which currently exists between IID and survey statistics.

Organizationally, the paper is divided into four sections. This introduction is Section 1. In Section 2 and 3 a general problem statement is provided and several "resolutions" are offered in a few of the better known designs. Our approach is to resample the complex sample to obtain an easier to analyze data structure. Specifically, we cover stratified element sampling, one and two-stage cluster samples, plus the important two PSU per stratum design (Section 2). Because any given resample is unlikely to contain all the information in the original survey, we look at what happens when the original complex sample is repeatedly resampled. A concrete illustration of our ideas is also given in Section 3; this has been taken from our practice and is based on a highly stratified Statistics of Income (SOI) sample of corporate tax returns (*e.g.*, Hughes, Mulrow, Hinkins, Collins and Uberall 1994). In a concluding section (Section 4), we discuss a few applications and some next steps needed for our still embryonic ideas to grow more useful.

2. PROBLEM STATEMENT AND POSSIBLE "RESOLUTIONS"

2.1 Motivation and Basic Approach

Suppose we wanted to apply an IID procedure to a complex survey sample. Suppose, too, that we wanted to take a fresh look at "solving" the seam problem that occurs because the survey design is not IID. How might one proceed? Well, there is a familiar expression that may fit our approach

If you only have a hammer, every problem turns into a nail.

Now, as samplers, we have a hammer and it is sampling itself. Can we turn the seam problem in surveys into a nail that can be dealt with by using another sampling design?

It is our contention that some of the time the answer to this question is "Yes." We call this second sample design an "Inverse Sampling Design Algorithm" – hence, the name of this paper.

A schematic might help visualize the algorithm (see figure 1). In the diagram two sampling approaches are compared – both yielding simple random samples from a population:

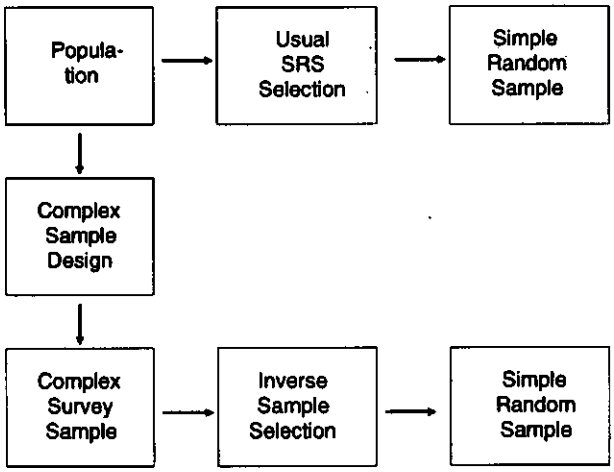
(1) The first design (top row) does this by employing a conventional direct simple random (SRS) selection process (*e.g.*, Cochrane 1977), such that all possible

[1] Susan Hinkins, Internal Revenue Service, Bozman, MT, U.S.A.; H. Lock Oh, Internal Revenue Service, Washington, DC, U.S.A.; Fritz Scheuren, Ernest and Young, 1402 Ruffner Rd., Alexandria, VA 22302 U.S.A.

samples of a given size have the same probability of selection. (Such designs are often impracticable or inefficient or both; hence, they are almost never used by survey samplers, despite their ubiquity in textbooks.)

(2) The second design envisions a two-step process. The first step is to sample the population in a complex way that focuses carefully on the nature of the population and the client's needs – using the client's resources frugally (this is the survey sampler's province, par excellence).

(3) What is new in our formulation is to draw a second (perhaps complex?) sample that inverts the first set of selections, so as to yield at the end a simple random sample. Of course, to employ this two-step process to draw a single simple random sample from the usually much larger complex survey would be inefficient, so we propose to create multiple simple random samples and base our inferences on them.



While elaborations are possible, the basic nature of the algorithms we are talking about should, by this point, be obvious. They can consist of just four basic steps:

(1) Invert, if you can, the existing complex design, so that simple random subsamples can be generated (to some useful degree of approximation).

(2) Potentially, apply your conventional statistical package directly to the subsample, since that is now appropriate.

(3) Repeat the subsampling and conventional analysis, in steps (1) and (2), over and over again.

(4) Retain, if you can, the flavour of the original randomization paradigm by using the distribution of subsample results as a basis of inference (rather than the original complex sample).

Notice some things that this approach is – and is not: First, it is extremely computer intensive – presupposing cheap, even very cheap computing. Second, it presupposes that practical inverse algorithms exist (which may not always be the case). Third, it also assumes that the original power of the full sample can be captured if enough subsamples are taken, so that no appreciable efficiency is lost. Fourth, as much as it may resemble the bootstrap (Efron 1979), we are not doing bootstrapping. There is no intent to mimic the original selections, as would be required to use the bootstrap properly (*e.g.*, McCarthy and Snowmen 1985; Rao and Wu 1988) – just the opposite; our goal here is to create a totally different and more analytically tractable set of subsamples from the original design.

2.2 Defining An Inverse Sampling Algorithm

Suppose that we wish to draw a simple random sample, without replacement, from a finite population of size N. Suppose further that the population is no longer available for sampling, but we have a sample selected from this population using a sample design D; let S_D denote this sample. Let S_m denote a second sample of size m that could be drawn from the population. An inverse sampling algorithm must describe how to select a sample from S_D so that for any given sample S_m

$$\Pr(\text{select } S_m \mid S_D) * \Pr(S_m \subset S_D) = \frac{1}{\binom{N}{m}}. \qquad (1)$$

The first step is to calculate the probability that an arbitrary but fixed sample S_m is contained in the sample S_D. Obviously, there are constraints on the size of the simple random sample (SRS) that can be drawn in this manner; the probability that S_D contains S_m cannot be zero. Certainly, therefore, the SRS cannot be larger than the size of the original sample S_D, and in fact the size of the SRS is generally required to be much smaller than the original complex sample.

The problem, then, is to find a general algorithm to select an SRS from a given sample S_D with the correct conditional probability. It is also necessary to check that valid probability functions are used. The following subsections show the inverse sampling algorithms for a few of the more common sample designs: stratified, cluster, multistage, and stratified multistage designs. We also give an example where an inverse algorithm at first does not appear feasible.

2.3 Inverting A Stratified Sample

In this subsection the inverse algorithm is given for a stratified sample with four strata. The algorithm generalizes for any number of strata. We have a stratified sample with fixed sample sizes n_h in each stratum h, and known stratum population sizes, $N_1 + N_2 + N_3 + N_4 = N$. Because a given sample of arbitrary size m from the population might be contained entirely within one stratum, the largest simple random sample that can be selected from a stratified sample is of size $m = \min\{n_h\}$.

For a given sample S_m, let (x_1, x_2, x_3, x_4) denote the number of units in each stratum. Each x_i will be between 0 and m, and $x_1 + x_2 + x_3 + x_4 = m$. The probability that S_m is contained in the stratified sample is equal to the number of stratified samples containing these m units divided by the total number of possible stratified samples, *i.e.*

$$\Pr(S_m \subset S_D) = \frac{\binom{N_1-x_1}{n_1-x_1}\binom{N_2-x_2}{n_2-x_2}\binom{N_3-x_3}{n_3-x_3}\binom{N_4-x_4}{n_4-x_4}}{\binom{N_1}{n_1}\binom{N_2}{n_2}\binom{N_3}{n_3}\binom{N_4}{n_4}}. \quad (2)$$

The algorithm for selecting a SRS from the stratified sample consists of the following three steps:

(1) Determine the size of the SRS to be selected: $m \le \min\{n_h\}$.

(2) Generate a realization $\{m_1, ..., m_4\}$ from a hypergeometric distribution, with probabilities

$$\Pr(m_1 = i_1, m_2 = i_2, ..., m_4 = i_4) = \frac{\binom{N_1}{i_1}\binom{N_2}{i_2}\binom{N_3}{i_3}\binom{N_4}{i_4}}{\binom{N}{m}} \quad (3)$$

where $i_1 + i_2 + i_3 + i_4 = m$ and $0 \le i_1 \le m, 0 \le i_2 \le m, 0 \le i_3 \le m, 0 \le i_4 \le m$.

(3) In each stratum h, select a simple random sample of size m_h, without replacement, from the n_h sample units.

The conditional probability of selecting the sample S_m given that it is contained in the stratified sample, is then

$$\frac{\binom{N_1}{x_1}\cdots\binom{N_4}{x_4}}{\binom{N}{m}} \cdot \frac{1}{\binom{n_1}{x_1}\cdots\binom{n_4}{x_4}}. \quad (4)$$

The probability of selecting any given sample S_m using the inverse algorithm is the product of the two probabilities given in equations (2) and (4). It is straightforward to show that this product is equal to

$$\frac{1}{\binom{N}{m}}.$$

Therefore this procedure reproduces a simple random sampling mechanism unconditionally, *i.e.*, when taken over all possible stratified samples. Note that in order to generate all possible SRS's from this population, the entire sequence must be repeated, starting with selecting a stratified sample and proceeding through steps 1 - 3.

2.4 Inverting a One Stage Cluster Sample

In this subsection, we consider three special cases. To begin with, we examine cluster samples where the clusters are of equal size. This is followed by the more usual case where the clusters are of unequal size. In both of these settings we assume the clusters are sampled by a simple random sampling mechanism and without replacement. The third case studied is that of sampling unequal clusters by a probability proportional to size (PPS) mechanism. In this last instance we assume that the sampling is with replacement.

2.4.1 One Stage Cluster Sampling With Equal Cluster Sizes, Sampled With Equal Probability

Assume we have a population of N clusters where all clusters are of size M and k of them are selected by a simple random sampling mechanism without replacement.

To construct an inverse algorithm, we need to decide what the largest element subsample might be. It is immediate that the largest SRS of elements that can be selected is k. Incidentally, the cluster size is not a constraint on the size of the subsample.

For a given sample S_k, let q denote the number of clusters represented in S_k; $0 < q \le k$. Then the probability that S_k is contained in the cluster sample is equal to the number of cluster samples containing these q clusters divided by the total number of possible cluster samples, *i.e.*

$$\Pr(S_k \subset S_D) = \frac{\binom{N-q}{k-q}}{\binom{N}{k}}. \quad (5)$$

As for the stratified sample, the algorithm first determines the number of units to be chosen from each cluster, $(m_1, m_2, ..., m_k)$. The probability distribution to be used to select the m_i's is

$$\Pr(m_1 = i_1, ..., m_k = i_k) = \frac{\binom{M}{i_1}\cdots\binom{M}{i_k}}{\binom{NM}{k}} * \frac{N(N-1)...(N-q+1)}{k(k-1)...(k-q+1)} \quad (6)$$

where $0 \le i_j \le k, i_1 + i_2 + ... + i_k = k$, and q is the number of nonzero i_j's. For example, with $M = 100$, $N = 6$, and $k = 3$

$$\Pr(m_1 = 1, m_2 = 0, m_3 = 2) = \frac{\binom{100}{1}\binom{100}{0}\binom{100}{2}}{\binom{600}{3}} * \frac{6*5}{3*2}$$

$$\Pr(m_1 = 3, m_2 = 0, m_3 = 0) = \frac{\binom{100}{3}}{\binom{600}{3}} * \frac{6}{3}.$$

Once the m_i's are determined, a simple random sample of size m_i is selected from cluster $i, i = 1, 2, ..., k$. Therefore the conditional probability of selecting S_k is

$$\Pr(\text{select } S_k \mid S_D) = \frac{1}{\binom{NM}{k}} * \frac{N(N-1)...(N-q+1)}{k(k-1)...(k-q+1)}. \quad (7)$$

The probability of selecting a particular sample S_k is found by multiplying equation (5) times equation (7). It is routine to verify that this gives the correct probability of selecting an SRS.

Unlike the stratified example, where the function for selecting the values of m_i was a known probability function, it is not immediately obvious that equation (6) describes a probability distribution. Since the values generated by this function are all nonnegative, it need only be shown that they sum to one over the space of possible values. The first factor in the equation has the form of a hypergeometric distribution, except that the numerator is constrained to only k out of the N clusters, while the denominator still reflects the total N clusters. It is useful to define a partition of k as a combination of positive integers that adds to k, without regard to order. For example, the partitions of $k = 3$ are {3}, {1,2}, and {1,1,1}. Because the clusters are all of the same size, M, all patterns of selection that correspond to the same partition have the same probability of occurring. Take, for example, $N = 6$, and $k = 3$. In the full hypergeometric distribution, with equal cluster size, each of the following combinations has the same probability of occurring

$$(0,0,0,0,1,2),(0,0,0,0,2,1),(0,0,0,1,2,0),\ ...,\ (2,1,0,0,0,0).$$

The total number of such combinations is $N(N-1)...(N-q+1)$, where q is the size of the partition, that is the number of (nonzero) values in the partition. In the example above, $q = 2$. For a given partition, if the nonzero counts can only be put into k specific cells, then there are $k(k-1)...(k-q+1)$ such orderings. Therefore, summing the distribution over all values of $(i_1, ..., i_k)$ can be done by first summing over all partitions of k and then for each partition, summing over all possible orderings of that partition in k cells. Because all orderings associated with a particular partition share a common probability of occurrence, this results in a summation that is equivalent to summing the hypergeometric over the correct space, and therefore expression (6) sums to one.

The probability distribution needed for this simple cluster design (equation 6) is noticeably more difficult to generate than the hypergeometric distribution in the case of the stratified sample. However, as the sampling fraction k/N decreases, the probability is often contained in only two of the partitions: $q = k$ and $q = k - 1$. (These probabilities are calculated in the Appendix). Indeed, the probability may be concentrated in just the pattern with $q = k$ (A special case of this is also shown in the Appendix).

Given the results in the Appendix, it may be possible to approximate the exact inverse by selecting one case from each cluster, using systematic sampling from the original cluster sample. This approach is of real value because the probability distribution calculations become unwieldy as the number of clusters in the sample grows large. For a systematic inverse to work, however, the "step" would naturally have to be at least as large as M or maybe even greater, depending on the number of clusters in the population. To carry out this subsampling repeatedly, for each systematic sample inverse, the units within each cluster would be reordered randomly before the next selection and the clusters resorted randomly as well - then another random start obtained before stepping again through the original sample.

2.4.2 One Stage Cluster Sampling with Unequal Clusters, Sampled With Equal Probability

The inverse sampling algorithm for a sample of clusters of equal size does not generalize readily when a sample of unequal sized clusters is drawn. This is so despite the fact that it would appear to be straightforward to generalize this approach in an obvious way. In particular, it does not seem difficult to generalize the previous method so that the "probabilities" would multiply out successfully to give the "correct" probability of selection, *i.e.*

$$\frac{1}{\binom{M_+}{k}}, \quad \text{where} \quad M_+ = \sum_1^N M_i. \quad (8)$$

However, generalizing to unequal cluster sizes M_i by selecting the m_i as

$$\Pr(m_1=i_1,...,m_k=i_k) = \frac{\binom{M_1}{i_1}...\binom{M_k}{i_k}}{\binom{\sum_1^N M_i}{k}} * \frac{N(N-1)...(N-q+1)}{k(k-1)...(k-q+1)} \quad (9)$$

does not result in a valid probability distribution. We will again assume, by the way, that the original clusters are being sampled with equal probability and without replacement, as was the case in subsection 2.4.1. Later (Subsection 2.4.3), as already noted, we will look at original samples which employ some form of Probability Proportional to Size (PPS) selection.

To see that it is not straightforward to simply generalize equation (6) into the form in equation (9), consider the following counter-example where the "probability" calculated using (9) is greater than one. Suppose $N = 4$ with cluster sizes; $M_1 = 4$, $M_2 = 6$, $M_3 = 8$, and $M_4 = 10$. Suppose further that we draw a cluster sample with $k = 2$ and that just by chance the two clusters picked are the largest – *i.e.*, $M_3 = 8$ and $M_4 = 10$. It is immediate that with these selections, equation (9) would generate a probability of selecting one unit from each cluster that was greater than one.

Can this difficulty be fixed? Yes, although not perhaps in an entirely satisfactory way. One method is to employ a

hypergeometric that assumes all the clusters were as large as the largest cluster in the population. The price paid is that the inverse sample size achieved is no longer fixed, and the resulting subsample is only conditionally SRS given the achieved sample size, denoted, say, as k_0. That is, for a given sample size k_0, $k_0 \le k$, all samples of size k_0 have the same probability of being selected using the inverse algorithm.

Let M_* denote the maximum cluster size, $M_* = \text{Max}\{M_1, M_2, ..., M_N\}$. Create a population by filling out each original cluster with "dummy" units or placeholders, $j = M_i + 1, M_i + 2, ..., M_*$. Then using a method similar to Lahiri's (1951) for PPS sampling, the inverse algorithm selects units from the population consisting of N clusters each of size M_*, and then discards any element not in the "subpopulation" consisting of the original clusters of size M_i.

Specifically, given a cluster sample consisting of k clusters, select the vector $\boldsymbol{m}$ from the probability distribution

$$\Pr(m_1 = i_1, ..., m_k = i_k) = \frac{\binom{M_*}{i_1}\binom{M_*}{i_2}\cdots\binom{M_*}{i_k}}{\binom{NM_*}{k}} * \frac{N(N-1)...(N-q+1)}{k(k-1)...(k-q+1)} \quad (10)$$

where the components of $\boldsymbol{m}$ sum to k, and q of the components m_i are nonzero. This is now a proper probability distribution. Given the selected value of m_i, select a random sample of m_i units from cluster i, where the cluster contains M_i units from the population and $M_* - M_i$ "placeholders." Discard any selected units that are placeholders, in the set of $j = M_i + 1, M_i + 2, ..., M_*$. Therefore the final sample size will not necessarily be equal to k, but may be smaller, say k_0.

The resulting sample is conditionally a SRS from the population, in the sense that for a given value of k_0, all samples of size k_0 have the same probability of being selected using this inverse algorithm. To see this, continue to view the problem as a subpopulation, P, of N clusters of size M_i, $i = 1, ..., N$, within a population P_* of N clusters each of size M_*. Note that for any sample, S_*, of size k selected from the population P_*, the probability of selecting S_* using the inverse algorithm is

$$\frac{1}{\binom{NM_*}{k}}. \quad (11)$$

If $k_0 = k$ then this is the probability of selecting this sample using the inverse algorithm. For a fixed $k_0 < k$, let S_0 denote any given sample of size k_0 contained in P. We can generate a sample S_* containing S_0 by starting with S_0 and adding to it $k - k_0$ elements from the $N * M_* - M_+$ placeholders in P_*. The number of such samples S_*, that result in selecting S_0, is

$$\binom{NM_* - M_+}{k - k_0} \quad \text{where} \quad M_+ = \sum_{i=1}^{N} M_i. \quad (12)$$

Therefore, the probability of selecting S_0 using the inverse algorithm is equal to the probability of selecting S_* using the inverse algorithm, given in (11), summed over all samples S_* constructed as described above, where the number of such samples is given by (12). This probability equals

$$\frac{\binom{NM_* - M_+}{k - k_0}}{\binom{NM_*}{k}}$$

and all samples of size k_0 have the same probability of being selected using the inverse algorithm.

There is a positive probability, unfortunately, that a sample might be selected with this approach that has no elements. This could occur if there were a large difference in the cluster sizes. However, if the number of clusters k in the original sample is large, this is unlikely to be a problem.

Again, as in the case of equal cluster sizes, an approximation is available using a systematic subsample as an inverse. This time we would want a step at least as large as the maximum cluster size. Using a systematic inverse, by the way, would have the advantage of controlling better the actual subsample size drawn.

2.4.3 One Stage Cluster Sampling with Unequal Clusters, Sampled With Unequal Probability

If a sample of k clusters is selected with PPS, an inverse algorithm may exist. Suppose the samples are selected with replacement from a population consisting of N clusters, with unequal cluster sizes, $M_1, M_2, ..., M_N$. Suppose, further, that the measure of size is either equal to M_i or proportional to M_i. Then at each draw,

$$\Pr(\text{select cluster } j) = \frac{M_j}{M_+} \quad \text{where } M_+ = \sum_{i=1}^{N} M_i. \quad (13)$$

Finally, since a one stage sample is being taken, once cluster j is selected, then all M_j units from that cluster are included in the sample.

An inverse algorithm in this case should result in a SRSWR. That is, for any vector $\boldsymbol{S}$ resulting from k independent selections from the population, the probability of selecting the ordered vector is

$$\Pr(\text{select } \boldsymbol{S}) = \left(\frac{1}{M_+}\right)^k. \quad (14)$$

An inverse algorithm is to simply randomly select one unit from each cluster in the cluster sample. Because the clusters were chosen with replacement, one should think of the sampled clusters as being ordered, by the order in which they were selected, or in any fixed order. For example, if the population contained 20 clusters, a possible cluster sample of size $k = 5$ is (7, 5, 7, 18, 6), *etc.*

The population consists of M_+ units, denoted as $u_1, u_2, ..., u_{M+}$. Let S denote a given sample, with replacement, $S = (s_1, s_2, ..., s_k)$, and let $c = (c_1, c_2, ..., c_k)$ denote the associated cluster for each unit. For example, suppose the population is:

Cluster	Units
1	$u_1\ u_2\ u_3\ u_4$
2	$u_5\ u_6\ u_7\ u_8$
3	$u_9\ u_{10}\ u_{11}$
4	$u_{12}\ u_{13}\ u_{14}$
5	$u_{15}\ u_{16}\ u_{17}$
6	$u_{18}\ u_{19}\ u_{20}$

and $k = 3$. Then the sample $(s_1 = u_2, s_2 = u_4, s_3 = u_{17})$ corresponds to $c = (1, 1, 5)$. The sample $(s_1 = u_{18}, s_2 = u_{19}, s_3 = u_{18})$ corresponds to $c = (6, 6, 6)$. Note that this second sample can only be selected if cluster 6 is the only cluster chosen in the cluster sample.

For a given sample S of size k, and the corresponding vector c of cluster membership, the unconditional probability of selecting S using the inverse algorithm is

$$\Pr(\text{select } S \mid \text{cluster sample } c) * \Pr(\text{select } c) = \left(\prod_{i=1}^{k} \frac{1}{M_{c(i)}}\right)\left(\prod_{i=1}^{k} \frac{M_{c(i)}}{M_+}\right) \qquad (15)$$

which is equal to the desired probability, equation (14).

Note that this same inverse algorithm works in the case where k clusters are selected with ppswr, but a sample of fixed size m is selected (srswor) from the chosen cluster, assuming that $M_i > m$ for all clusters i.

2.4.4 Some Comments On One Stage Designs.

We have seen that, with care, inverse algorithms can be constructed for several special cases where the original sample has a one stage cluster design. Two of our results are for cluster samples drawn with equal probability without replacement. The third is a ppswr design.

A convenient systematic inverse may even be workable as an approximation to the correct inverse algorithm when we have a cluster sample. The approximation works when using SRSWR is "close to" SRSWOR – *i.e.*, in our notation when k/NM is very small so that $1/(NM - k + 1)$ is approximately equal to $1/NM$. So everything seems intuitively to be consistent, across the cases studied.

Many cluster designs do not fall into any of the special cases examined. For some of them we conjecture that exact inverse algorithms may not exist. In particular, the general case of PPSWOR sampling seems to be one of these, including the frequently used variant of systematic PPSWOR. This may, or may not be a problem for practitioners who often employ the (usually) conservative practice of assuming that the sampling was with replacement – in which case an inverse algorithm would exist to the same order of approximation as was being assumed to estimate variances.

2.5 Multistage Cluster Designs

What about multistage designs? Can they be inverted? In some cases, we believe the answer is "Yes." Three designs will be looked at: (1) a two-stage design with simple random sampling at the first and second stages (Subsection 2.5.1); then, (2) a design which employed probability proportional to size (PPS) sampling at the first stage and simple random sampling at the second (Subsection 2.5.2). Finally, (3) the very important stratified multistage design with two PSUs per stratum deserves at least a brief comment.

As will be seen, the stratified and one stage results extend fairly readily. To demonstrate this, our basic strategy is to repeatedly apply the approaches already discussed earlier.

2.5.1 Multistage Designs With Simple Random Sampling at Both Stages

Suppose, first, that originally a simple random sample of k clusters, all of size M, was drawn at the first stage and a simple random subsample of size "r" was drawn at the second stage, within each cluster selected at the first stage.

As earlier, our inverse sample can be no larger than k. Suppose first that $1/(NM - k + 1)$ is approximately equal to $1/NM$, then we can employ an srswr inverse algorithm, since SRSWR and SRSWOR are very close. Using the results in Subsection 2.4.3, we would take a SRSWR sample of k clusters and then within each selected cluster take one observation at random. Alternatively, we could as in Subsection 2.4.1, first determine the number of units to be chosen from each cluster, $(m_1, m_2, ..., m_k)$. Once the m_i's are determined, a simple random sample without replacement of size m_i is selected from cluster i, $i = 1, 2, ..., k$. This may be a nearly exact result, except for the possibility that the inverse second stage sample size m_i may be larger than the original second stage sample size "r." When this occurs, we still can appeal to the results in Subsection 2.4.2 and draw our second stage sample with "placeholders." In this second instance, the resulting actual sample would no longer be fixed; but still would be conditionally SRS. If the first stage clusters are unequal in size but sampled with replacement, then we can again employ the trick used in Subsection 2.4.2 of creating "placeholders." The sample sizes are random and only conditionally do we achieve an SRS inverse.

Another way to approach this problem is to note that the largest SRS that can be selected using an inverse algorithm is

of size $k_0 = \min\{k, r\}$. This is done by first determining the number of units to select from each cluster, $(m_1, m_2, ..., m_k)$, where now the m_i's must sum to k_0 rather than k. Once the m_i's are determined, a simple random sample of size m_i is selected from cluster i, $i = 1, 2, ..., k$. The probability distribution to be used to select the m_i's is

$$\Pr(m_1 = i_1, ..., m_k = i_k) = \frac{\binom{M}{i_1} \cdots \binom{M}{i_k}}{\binom{NM}{k_0}} * \frac{N(N-1)...(N-q+1)}{k(k-1)...(k-q+1)}$$

where $0 \le i_j \le k_0$, $i_1 + i_2 + ... + i_k = k_0$, and q is the number of nonzero i_j's.

One final comment, for both equal and unequal cluster sizes, the possibility of an approximate systematic inverse seems available – with essentially the same caveats, of course, as noted above.

2.5.2 Multistage Designs With PPS Sampling at the First Stage and SRS Sampling at the Second

Again, our inverse sample can be no larger than k. It is immediate that one way to construct an inverse would be to use the results in Subsection 2.4.3. Specifically, we would take a srswr sample of k clusters and then within each selected cluster take one observation at random. Other inverse algorithms may exist too. A systematic inverse seems reasonable, provided the probability of selecting the same cluster more than once is small to vanishing.

2.5.3 Stratified Multistage Designs With Two PSU's Per Stratum

Can two Primary Sampling Unit (PSU) designs be inverted? Our answer is "Yes," if the within stratum selections are made in one of the ways we discussed in detail earlier. This is basically the only case we will cover.

From our results in Subsections 2.3 and 2.4, it is immediate that if an inverse is to exist, then the sample size m cannot be any larger than $m = 2$. Depending on the sampling within each strata, we could employ one or more of the exact or approximate inverses to obtain two SRS selections within each stratum. To obtain an overall SRS sample, we would employ the inverse algorithm of Subsection 2.3 on these two selections and end up, finally, with just two selections overall.

2.5.4 Some Comments On Multistage Designs

In this Subsection, we have quickly covered a few multistage designs and provided exact or approximate inverses. The results were derived by appealing to earlier results in Subsections 2.3 and mainly 2.4. Of course, many multistage designs do not fall into any of the special cases examined - notably those with systematic selections at the last stage.

One last observation, many readers may wonder, at this point, how a method that selects only a sample of size two (as we did in Subsection 2.5.3) can be of any practical value. Perhaps the next section will help.

3. RESAMPLING TO INCREASE POWER

3.1 General Setting

Drawing a single, smaller simple random sample from a larger, more complex sample might be adequate for some users in some settings. However, for most users, the loss in power between the estimate based on the complex sample and the estimate based on a simple random sample would not be acceptable.

In order to increase the power of our approach, it was natural to consider resampling techniques. We are limited in the size of the SRS that can be drawn, but we can repeat the process. By repeating the entire subsampling procedure, we can generate g simple random samples each of size m, where each SRS is selected independently from the overall original sample. Each repetition must include all steps of the subsampling procedure. For example, in the stratified case, the stratum subsample sizes must be redrawn using the hypergeometric distribution.

In this section, conditions are given under which the precision of the estimates using multiple SRSs can be made arbitrarily close to the precision of the original estimates. We will begin our discussion by first defining some notation.

Let D denote any invertible design (such as a design of the type covered in Section 2). Let T be the population quantity of interest (say, a population total); and let T_D be an unbiased estimator of T calculated from the sample S_D. Suppose g simple random samples are independently drawn from the given sample S_D and let t_i denote the estimator from the i-th simple random sample. Then it can be shown that

$$\text{if } E(t_i \mid S_D) = T_D$$

$$\text{then } \operatorname{Var}\left(\frac{1}{g}\sum_{i=1}^{g} t_i\right) = \operatorname{Var}(T_D) + \frac{1}{g}(\operatorname{Var}(t_1) - \operatorname{Var}(T_D)).$$

Proof: Because the g replications of the simple random sampling process are conditionally independent, then

$$\text{for } i \ne j,\ E(t_i t_j \mid S_D) = T_D^2.$$

Therefore, unconditionally, for i not equal to j,

$$\begin{aligned}\operatorname{Cov}(t_i, t_j) &= E(t_i t_j) - T^2 \\ &= \operatorname{Var}(T_D).\end{aligned}$$

And the result follows directly.

Some of the conditions in this proof can be relaxed; if T_D is biased, then similar results can be obtained for MSE instead of variance. However, the condition that

$$E(t_i \mid S_D) = T_D$$

is necessary. And this condition is not met for ratio estimators. But, if the condition is met separately for the numerator and for the denominator of the ratio estimate and if the final size of the *combined* sample is sufficiently large so that a Taylor Series approximation is acceptable, then similar results can be found for approximations to the variance for ratios in the usual manner. Incidentally, even in the two PSU per stratum design, this approach works – provided we can obtain an unbiased estimate from each individual sample of size 2. And for estimates of totals, this can be the case – assuming at each stage of sampling that an inverse can be constructed.

3.2 Estimating The Sampling Error for Means or Totals

By resampling, one can achieve almost the same precision as the original design estimator. But because the resampled srs's are only conditionally independent, the estimation of the standard error is not as simple as if only one srs had been drawn. However the estimation remains relatively straightforward.

Let S^2 denote the population variance for the variable X and let T be its population total. For the sample means, totals and variances calculated from the generated simple random samples, let

$$t_{\cdot\cdot} = \frac{1}{g}\sum_{j=1}^{g} t_j = \frac{1}{g}\sum_{j=1}^{g} N\bar{x}_j = \frac{1}{g}\sum_{j=1}^{g}\frac{N}{m}\sum_{i=1}^{m} x_{ji}$$

$$s_j^2 = \left(\frac{1}{m-1}\right)\sum_{i=1}^{m}(x_{ji} - \bar{x}_j)^2$$

$$s_{\cdot}^2 = \left(\frac{1}{gm-1}\right)\sum_{j=1}^{g}\sum_{i=1}^{m}(x_{ji} - \bar{x}_{\cdot\cdot})^2$$

where $\bar{x}_{\cdot\cdot} = \frac{t_{\cdot\cdot}}{N} = \frac{1}{gm}\sum_j\sum_i x_{ji}$.

Note that the sample variance using all gm units can be expressed as

$$s_{\cdot}^2 = \frac{1}{mg-1}\left[(m-1)\sum_{j=1}^{g} s_j^2 + \frac{m}{N^2}\sum_{j=1}^{g}(t_j - T)^2 - \frac{mg}{N^2}(t_{\cdot\cdot} - T)^2\right].$$

Hence

$$E(s_{\cdot}^2) = \frac{1}{mg-1}\left[g(m-1)S^2 + \frac{m}{N^2}\sum_{j=1}^{g}\mathrm{Var}(t_j) - \frac{mg}{N^2}\mathrm{Var}(t_{\cdot\cdot})\right].$$

Rewriting this gives

$$\mathrm{Var}(t_{\cdot\cdot}) = N^2\left(\frac{m-1}{m}\right)S^2 + \left(\frac{1}{g}\right)\sum_{j=1}^{g}\mathrm{Var}(t_j) - N^2\left(\frac{mg-1}{mg}\right)E(s_{\cdot}^2).$$

Therefore, by replacing S^2 and $\mathrm{Var}(t_j)$ with unbiased estimates and replacing $E(s_{\cdot}^2)$ with $s_{\cdot}^2$, we can generate approximately unbiased estimates of $\mathrm{Var}(t_{\cdot\cdot})$.

It may be worth emphasizing that this result does not require the user to know anything about the original sample design. If users are given a way to invert the original design, then they can, by repeated subsampling, achieve nearly the efficiency of the original design and readily estimate the appropriate sampling errors. There is one condition on this result, namely that the subsample size be such that $m \geq 2$. Incidentally, for $m = 2$, the variance expression becomes

$$\mathrm{Var}(t_{\cdot\cdot}) = \frac{N^2}{2}S^2 + \left(\frac{1}{g}\right)\sum_{j=1}^{g}\mathrm{Var}(t_j) - N^2\left(\frac{2g-1}{2g}\right)E(s_{\cdot}^2).$$

Based on this, as above, a variance estimator could be built for two PSU per stratum designs.

3.3 An SOI Illustration

In this subsection we consider an example of an inverse algorithm and how well it works. The Statistics of Income (SOI) corporate sample will be our starting point. Now, as noted earlier, the SOI sample has essentially a stratified SRS design and so can be inverted (subsection 2.2).

It is our belief that many SOI users might find a full SRS inverse sample more valuable and easier to employ than the complete, stratified sample data base. An interim goal could be to provide them with a set of simple random samples. A more flexible system would be to provide the interactive software to allow the user to designate the simple random samples of interest, to be selected from the complete data base.

In our simulations we used four of the strata in the SOI sample of corporate returns, namely the strata representing the smallest regular corporations (Hughes *et al.* 1994). As can be seen from table 1, the stratified sample (of four strata) consisted of 15,618 units, and the largest SRS that can be selected is $m = 2{,}224$. The table also shows the population sizes and the estimated variance of the variable Total Assets, within each stratum.

Table 1
Corporate Population and Sample Size, plus Estimated Stratum Variances, For Four SOI Stratum

Strata (h)	N_h	n_h	S_h^2 (in 1000's)
1	1,376,801	3,889	222,808
2	552,909	2,224	670,162
3	678,371	4,005	12,796,578
4	436,023	5,500	14,984,753

The variable total assets was used because it is the primary stratifying variable; and, therefore, the loss in precision due to removing the stratification should be relatively large. Indeed, this proved to be the case.

Shown below is the ratio of the variance of the estimated total using g simple random samples, of 2,224 each, divided by the variance of the total based on the stratified sample. The table displays values of g from 1 to 1,000. For example, if only one SRS is selected the variance of the estimated total is 29 times larger than the variance of the stratified total.

g	Relative Variance Increase
1	29.31
2	15.16
10	3.83
100	1.28
500	1.06
1000	1.03

By resampling 500 to 1,000 times, the variance has been reduced to the same order of magnitude as the stratified sample. Even at 100 subsamples good results exist here, suggesting that the use of an inverse algorithm could work well for strata such as these. This is not to recommend that an inverse algorithm be employed in general with so few resamples. Doubtless, in highly skewed populations a much larger number would be required.

4. POTENTIAL APPLICATIONS AND NEXT STEPS

In this paper we have shown that inverse sample design algorithms exist in a few special cases. We do not, as yet, have a general result – if, indeed, there is one. This is clearly a part of the problem that needs more work. Like most tools, an inverse sampling algorithm may not be the best choice in certain cases; it may not be even a reasonable alternative in some circumstances. But there are applications where it appears to have advantages and so should be considered. In this section we both briefly suggest areas where this methodology may be useful and also mention some of the limitations and problems that remain.

Customer-Driven Perspective – It is worth emphasizing the customer-driven nature of our approach. Even if it could not be justified on other grounds, inverse algorithms might be advocated as a part of "reinvention" (*e.g.*, Osborne and Gaebler 1992). Right now many large complex surveys may not be sufficiently benefiting society, because they are so badly under-analyzed or even misanalyzed:

- For the long run, we must work towards increasing the survey and general quantitative literacy of existing and potential customers – *e.g.*, as with the new series *What Is a Survey?* (Scheuren (ed.) 1995).
- In the short run, we need to start where our customers are – giving due respect to the often small part that survey data may add to their decision making. Certainly it is worth thinking about ways to lower the cognitive costs customers bear when using our complex survey "products."

A "Sample" of Possible Opportunities – There is an increasing awareness of the weaknesses within the traditional randomization paradigm (*e.g.*, Särndal and Swensson 1993). Of particular concern here is all the fiddling we have to do when trying to correct for nonsampling errors. Some of this flavour is evident in Rao and Shao (1993). By putting the possible adjustments for these nonsampling errors back into a simple random sampling framework, we may, indeed, be able to make more progress.

For decades, survey practitioners have elaborated exceedingly complex sample designs; and, then, made efficient point and confidence interval estimates from them. On the other hand, how much do we really understand about the distributions that our sample estimators generate when effective sample sizes are small to moderate? Will we be able to fully capitalize on the "visualization revolution" now occurring (*e.g.*, Cleveland 1993)? Particularly in the presence of nonsampling error? Maybe we should be building in a way to always look at distributions. The use of an inverse sampling algorithm might be one possibility (See also Pfeffermann and Nathan 1985). In any case, stronger visualization tools for complex surveys could help, even the very experienced among us, deepen our intuitions and connect them better to the particular population under study. Obviously, visualization efforts also pay off by lowering the price customers pay to use survey data.

An intriguing problem where the inverse sampling algorithm may have an application is the case where we have a two PSU per stratum design with L strata where L is small, say less than 30. Suppose further that for some of the variables in the survey the stratification and clustering are unimportant – *i.e.*, the design effect is $\delta = 1$, approximately. For these variables, would it not be possible for the stability of the variance estimate to be greater with the resampled method than with the Balanced Repeated Replication (BRR) approach to variance estimation that is usually employed?

Another example that we are considering is the case where the user is interested in tests of independence in 2×2 tables, based on stratified sample data (Hinkins, Oh and Scheuren 1995). For the chi-square test statistic we are now in the midst of comparing our results with the approach suggested by Scheuren (1972) and Fellegi (1980). So far it appears that the power of our method is comparable to these more familiar approaches (as might be expected from, say, Westfall and Young (1993)). This may be an instance where the extra work involved in the inverse sampling algorithm may have real benefits – beyond just making it easier for users to employ familiar tools – by allowing the user to look at the distribution rather than just one p-value.

A "Sample" of Problems Remaining – A "sample" of the problems that remain with our inverse algorithm might be given here. For example, what happens when we do not know what the population size is? What happens when the population has more than one elementary unit – persons, say, for one analysis; households for another; neighbourhoods for still a third? Answers exist for these difficulties but they have

an *ad hoc* flavour to us. In many surveys, for instance, we guess about N and use that guess in poststratification. That degree of approximation for an inverse might be acceptable. For the problem of multiple analysis units, we could do several inverses. While potentially workable, this seems exceedingly awkward.

We have indicated that in some cases it may not be too difficult to resample multiple times using the inverse algorithm in order to reproduce reasonable efficiency. But what about the case where the user of a stratified sample is interested in subpopulations. If the domains of interest are in fact the strata, then the user does not gain any benefits by using the SRS's produced using the inverse algorithm. If the domains of interest cut across the strata and they are small, then the number of samples required using the inverse algorithm may be very large in order to maintain reasonable estimation for the domains.

Finally, we briefly mention one more problem that we have thought about. Many multistage designs actually select only one PSU per stratum. The strata are then paired for variance estimation purposes. We have already noted that an inverse to this approximation is available which can be made about as good as that approximation is to begin with. Is there a way to get a better approximation using the inverse approach directly?

Last Words – Many things are changing in our profession. The worldwide quality revolution certainly has had an impact (Mulrow and Scheuren 1996). We are remaking the way surveys are done – from design, to data capture, to the way customers use them. This paper may be a small contribution to that process.

ACKNOWLEDGEMENTS

We wish to express our particular appreciation to the referees and associate editor for their insightful prodding and scholarship. The original submission we sent in was only a sketch of what is now included. We also owe a debt of gratitude to Phil Kott, who has been discussing our ongoing work at various Washington Statistical Society meetings.

APPENDIX

Suppose one has a cluster sample of k clusters from a population of N clusters, where each cluster has the same number of units, M. In the inverse sampling algorithm, the first step is to choose the vector $(m_1, m_2, ..., m_k)$ containing the number of units to be chosen from each cluster. Let q indicate the number of nonzero values of m_i. The probability of selecting the one pattern with $q = k$, that is the pattern with $m_i = 1$, for all $i = 1, 2, ..., k$, is

$$\Pr(q = k) = M^{k-1}\frac{(N-1)(N-2)...(N-k+1)}{(NM-1)(NM-2)...(NM-k+1)}.$$

Call this probability P_1. If $NM>>>k$ then P_1 can be approximated by

$$\prod_{i=1}^{k-1}\frac{(N-i)}{N} = \frac{(N-1)(N-2)...(N-k+1)}{N^{k-1}}.$$

Consider next the partition of k corresponding to $q = k - 1$; this corresponds to exactly one partition of k, namely $\{1, 1, ..., 1, 2\}$. There are $k(k-1)$ equally likely possible patterns of $(m_1, ..., m_k)$ with $q = k - 1$. The probability of selecting a vector $\boldsymbol{m}$ with $q = k - 1$, is

$$\Pr(q = k-1) = \frac{k(k-1)(M-1)}{2M(N-k+1)}P_1.$$

Therefore it is not difficult to calculate the probability that the selected $\boldsymbol{m}$ has either $q = k$ or $q = k - 1$. The following table shows some examples for two values of M.

Table A
$\Pr(q = k - 1 \text{ or } q = k)$

k	N	$M = 10$	$M = 100$
4	8	.92	.90
4	20	.99	.98
10	20	.38	.34
10	30	.63	.59
10	50	.83	.80
10	200	.99	.98
50	500	.35	.30
50	1000	.70	.66
50	5000	.98	.98

For small k, it is not difficult to calculate the entire probability distribution needed to generate m. But as k increases, the number of partitions increases, and this calculation becomes difficult or at least tedious. For $k = 4$, there are only 4 partitions; for $k = 10$ there are 39 possible partitions. One can see from Table A, that as the cluster sample becomes "larger," if the sampling rate is small enough, *i.e.*, if $k<<N$, then one might only need to calculate the probabilities for these two partitions in order to approximately invert the cluster sample. For $k = 10$ and $N = 200$, these two partitions essentially account for all of the probability distribution.

The probability of selecting just one unit per cluster $(q = k)$ is smaller than the values in Table A; so, in order to use a systematic inverse, we would want $k<<<N$. This can be obtained in some settings when the number of clusters is large and we are willing to take k very small, relying on repeatedly resampling the original survey, as described in Section 3.

To illustrate, assume a sample of size k_0 where, of course, $k_0 < k$, so that an inverse is possible; Further, to see if a systematic inverse would work, let $k_0 <<< N$. This is the case we illustrate in table B. In table B, we have confined

attention to just one value of N, $N = 5000$ clusters, although the results could be extended readily.

Table B
Pr{inverse sample picks the pattern (1,1, ..., 1)}

k_0	k_0/N	$M = 10$	$M = 100$
2	.0004	.9998	.9998
5	.001	.9982	.9980
10	.002	.9919	.9911
20	.004	.9663	.9627
30	.006	.9245	.9166
40	.008	.8687	.8553
50	.01	.8015	.7821

Clearly, as k/N gets small, a systematic sample becomes a better and better approximate inverse. Only experience would confirm if the approximation at $k_0 = 20$ and $k_0/N = .004$, say, is adequate. We think it might be, especially since the effect of using a systematic inverse usually is to make the variance calculations more conservative (since typically the intracluster correlation $\rho > 0$).

REFERENCES

BELLHOUSE, D. (1988). A brief history of random sampling methods. *Handbook of Statistics*, 6, 1-14.

CLEVELAND, W. (1993). *Visualizing Data.* Summit, NJ: Hobart Press.

COCHRAN, W. (1977). *Sampling Techniques.* New York: Wiley.

EFRON, B. (1979). Bootstrap methods: another look at the jackknife. *Annals of Statistics, 7, 139-172.*

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

HANSEN, M. (1987). Some history and reminiscences on survey sampling. *Statistical Science,* 2, 162-179.

HINKINS, S., OH, H.L., and SCHEUREN, F. (1995). Using an Inverse Algorithm for Testing of Independence Based on Stratified Samples. George Washington University Technical Report.

HUGHES, S., MULROW, J., HINKINS, S., COLLINS, R., and UBERALL, B. (1994). Section 3, *Statistics of Income – 1991, Corporation Income Tax Returns,* 9-17. Washington, DC: Internal Revenue Service.

KISH, L. (1995). The Hundred Years Wars of Survey Sampling. Centennial representative Sampling Conference, Rome, May 31, 1995.

LAHIRI, D. (1951). A method for sample selection providing unbiased ratio estimates, *Bulletin of the International Statistical. Institute,* 34, 72-86.

McCARTHY, P., and SNOWDEN, C. (1985). The bootstrap and finite population sampling. *Vital and Health Statistics.* Series 2, No. 95, DHHS Pub. No. (PHS) 85-1369. Washington, DC: Public Health Service.

MULROW, J., and SCHEUREN, F. (1996). Measuring to improve quality and productivity in a processing environment. *Data Quality,* 2, 11-20.

OSBORNE, D., and GAEBLER, T. (1992). *Reinventing Government.* New York: Plume.

PFEFFERMANN, D., and NATHAN, G. (1985). Problems in model identification based on data from complex samples. *Bulletin of the International Statistical Institute,* 68.

RAO, J.N.K., and SHAO, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika,* 79, 811-822.

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.

SÄRNDAL, C.-E., and SWENSSON, B. (1993). Washington Statistical Society talk on the shifting nature of the survey sampling paradigm.

SCHEUREN, F. (1972). Topics in Multivariate Finite Population Sampling and Data Analysis. George Washington University Doctoral Dissertation.

SCHEUREN, F. (Ed.) (1995). *What is a Survey?* One of a series of pamphlets published by the American Statistical Association to increase survey literacy.

SKINNER, C., HOLT, D., and SMITH, T., (Eds.) (1989). *Analysis of Complex Surveys.* New York: Wiley.

WESTFALL, P., and YOUNG, S. (1993). *Resampling-Based Multiple Testing.* New York: Wiley.

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

Survey Methodology, June 1997
Vol. 23, No. 1, pp. 23-32
Statistics Canada

Variable Selection for Regression Estimation in Finite Populations

PEDRO L.D. NASCIMENTO SILVA and CHRIS J. SKINNER[1]

ABSTRACT

The selection of auxiliary variables is considered for regression estimation in finite populations under a simple random sampling design. This problem is a basic one for model-based and model-assisted survey sampling approaches and is of practical importance when the number of variables available is large. An approach is developed in which a mean squared error estimator is minimised. This approach is compared to alternative approaches using a fixed set of auxiliary variables, a conventional significance test criterion, a condition number reduction approach and a ridge regression approach. The proposed approach is found to perform well in terms of efficiency. It is noted that the variable selection approach affects the properties of standard variance estimators and thus leads to a problem of variance estimation.

KEY WORDS: Auxiliary information; Calibration; Sample surveys; Subset selection; Ridge regression.

1. INTRODUCTION

Regression estimation is widely used in sample surveys for incorporating auxiliary population information (Cochran 1977, chap. 7). For the basic case when the population mean $\bar{X}$ of a vector of variables x_i is known and simple random sampling is used, the regression estimator of the population mean $\bar{Y}$ of a survey variable y_i takes the form

$$\bar{y}_r = \bar{y} + (\bar{X} - \bar{x})'b \qquad (1)$$

where $\bar{y}$ and $\bar{x}$ are the sample means of y_i and x_i respectively, and b is the sample vector of linear regression coefficients of y_i on x_i.

Regression estimation is useful for at least three reasons. First, it is flexible. Any number of population means of continuous or binary variables can, in principle, be incorporated into $\bar{X}$. In particular, poststratification arises as a special case (Särndal, Swensson and Wretman 1992, sec. 7.6). The procedure also extends to handle complex sampling designs. Second, regression estimation has certain optimal efficiency properties. See, for example, Isaki and Fuller (1982, Theorem 3). Third, $\bar{y}_r$ has the "calibration" property that if y_i is one of the variables of x_i so that $\bar{Y}$ is known then $\bar{y}_r = \bar{Y}$ (Deville and Särndal 1992).

In this paper we consider the question of how to select the x variables for use in the regression estimator. This question is of interest for at least two reasons. First, there is simply the practical reason that in some circumstances the number of potential variables in x_i may be very large. For example, in population censuses in a number of countries values of some variables are recorded on a "short form" for all individuals and values of other variables are collected on a "long form" for a sample. The population means of the short form variables together with their squares, cubes, products and so forth will thus be known. Small area identification will also typically be available. Thus the dimension of x_i as a vector containing functions of the short form variables together with dummy variables representing each small area could easily run into the thousands. In such cases, the selection of x variables becomes a practical necessity.

A second reason is more fundamental for a model-assisted or model-based approach to survey sampling. These approaches may be characterised as follows in the context of regression estimation. First a regression model is selected which has "good predictive power", so that the regression estimator will have "good efficiency". Then, either a design-based approach to inference is adopted in the model-assisted approach (Särndal *et al.* 1992) or model-based prediction is employed in the model-based approach. Although the literature on the latter problem of inference is vast, there seems remarkably little formal attention devoted to the former model selection problem. In practice, the most that seems to happen is that the "main" x variables which account for "most of" the sample R^2 are chosen (*cf.* Särndal *et al.* 1992, sec. 7.9.1). However, more theoretical guidance seems needed, especially when a large number of x variables is available.

A further reason for considering the variable selection problem more formally is that it may help clarify the issue of the impact of variable selection on inference. The problem that sample-based selection of estimators may affect the properties of the selected estimator has long been recognized (Hansen and Tepping 1969, App.) but little study seems to have been made of what the effects may be.

In this paper we consider a variable selection approach aimed at minimising the mean squared error of $\bar{y}_r$. First, however, we study the dependence of the mean squared error of $\bar{y}_r$ on the number of x variables in section 2 and then consider alternative estimators of the mean squared error of $\bar{y}_r$

[1] Pedro L.D. Nascimento Silva, IBGE-Departamento de Metodologia, Avenida Chile 500, Rio de Janeiro-RJ, Brasil; and Professor Chris J. Skinner, Department of Social Statistics, University of Southampton, Southampton, SO17 1BJ, United Kingdom.

in section 3. Variable selection procedures based on these estimators are then proposed in section 4.

We contrast our variable selection approach with four existing approaches. First, we consider the traditional approach of using a fixed subset of auxiliary variables regardless of the observed sample. Next, we consider a "condition number reduction procedure" inspired by work of Bankier (1990), in which auxiliary variables are discarded in order to reduce the condition number of a certain cross-products matrix of the x variables.

Third, we follow Bardsley and Chambers (1984) and consider a ridge regression approach. This does not involve variable selection but instead addresses the possible problem of multicollinearity in the regression estimator by modifying the estimator, allowing for some calibration error. Both the ridge regression and condition number reduction procedures have the advantage that they do not require specification of a response variable y, because they aim to provide a single set of "calibration" weights to be used for all survey variables. However, they do not guarantee gains in efficiency. Their results are separated by a line from the results for the other procedures in the tables presented in section 6 to indicate that they differ.

Fourth, we consider variable selection following conventional significance test criteria. Our general view is that the objective of variable selection in regression estimation for finite populations is quite different from the objective of parameter estimation or prediction of y values for single observations in classical regression (Miller 1990). However, it seems desirable to treat such an approach as one benchmark for comparison.

In section 5 we consider properties of the regression estimator following variable selection on the basis of estimated variances. Section 6 describes an empirical study carried out to compare our proposed variable selection procedures with the competing procedures described above. This study used data from a test census carried out in the municipality of Limeira, Brasil, as part of the preparation for the 1991 Brazilian Population Census. Section 7 presents our conclusions and some directions for further research.

2. THE DEPENDENCE OF THE VARIANCE OF THE REGRESSION ESTIMATOR ON THE NUMBER OF x VARIABLES

We begin by defining some notation. Let $U = \{1,...,N\}$ denote a finite population of N distinguishable elements and let $s \subset U$ denote a sample of n distinct elements drawn from U according to a simple random sampling without replacement design. Let $x_i = (x_{i1},...,x_{iq})'$ be the $q \times 1$ vector of auxiliary variables associated with the i-th population element. It is assumed that the sample values of $x_i (i \in s)$, together with the population mean vector $\bar{X} = N^{-1}\sum_{i \in U} x_i$ are known. The vector of sample means is denoted $\bar{x} = n^{-1}\sum_{i \in s} x_i$.

Let y_i denote the value of a survey variable y for the i-th population element and suppose the values of y_i are only observed for $i \in s$. The aim is to estimate the population mean $\bar{Y} = N^{-1}\sum_{i \in U} y_i$.

The regression estimator of $\bar{Y}$ is given by $\bar{y}_r$ in equation (1), where $\bar{y} = n^{-1}\sum_{i \in s} y_i$, $b = \hat{S}_x^{-1}\hat{S}_{xy}$, $\hat{S}_x = n^{-1}\sum_{i \in s}(x_i - \bar{x})(x_i - \bar{x})'$, and $\hat{S}_{xy} = n^{-1}\sum_{i \in s}(x_i - \bar{x})(y_i - \bar{y})$.

This estimator may be motivated by the underlying linear model

$$y_i = \beta_0 + x_i'\beta + \epsilon_i \qquad (2)$$

where the ϵ_i are independent disturbances with zero means and common variance σ^2, since we may write $\bar{y}_r = \hat{\beta}_0 + \bar{X}'\hat{\beta}$, where $\hat{\beta}_0 = \bar{y} - \bar{x}'b$ and $\hat{\beta} = b$ are the least squares estimators of β_0 and β, respectively. Under this model the variance of $\bar{y}_r - \bar{Y}$ conditional on the x_i may be written

$$\mathrm{Var}_M(\bar{y}_r - \bar{Y} \mid x_i) = \sigma^2 n^{-1}[1 - n/N + (\bar{X} - \bar{x})'\hat{S}_x^{-1}(\bar{X} - \bar{x})]. \qquad (3)$$

The final term may be interpreted as the effect of estimating β by b. As the number q of x variables increases the residual variance σ^2 may be expected to decrease, but the term $(\bar{X} - \bar{x})'\hat{S}_x^{-1}(\bar{X} - \bar{x})$ may increase as $\hat{S}_x^{-1}$ becomes more unstable. An alternative way to interpret this term is to write $\bar{y}_r$ as a weighted estimator $\bar{y}_r = n^{-1}\sum_{i \in s} g_i y_i$, where $g_i = 1 + (\bar{X} - \bar{x})'\hat{S}_x^{-1}(x_i - \bar{x})$. Then we may write (3) alternatively as

$$\mathrm{Var}_M(\bar{y}_r - \bar{Y} \mid x_i) = \sigma^2 n^{-1}(1 - n/N + c_g^2) \qquad (4)$$

where c_g is the sample coefficient of variation of the g_i.

To study the expected dependence of c_g^2 on q we now extend the model by supposing that the x_i are independently and identically normally distributed. Noting the independence of $(\bar{x} - \bar{X})$ and $\hat{S}_x$ and also that $E_M(\bar{y}_r - \bar{Y} \mid x_i) = 0$, we obtain the unconditional variance

$$\begin{aligned}\mathrm{Var}_M(\bar{y}_r - \bar{Y}) &= \sigma^2 n^{-1}\{1 - n/N + \mathrm{tr}[E_M[(\bar{X} - \bar{x})(\bar{X} - \bar{x})']E_M(\hat{S}_x^{-1})]\} \qquad (5)\\ &= \sigma^2 n^{-1}(1 - n/N)[1 + q/(n - q - 2)]\end{aligned}$$

using the fact that $n^{-1}\hat{S}_x^{-1}$ has an inverse Wishart distribution (Mardia, Kent and Bibby 1979, p. 69 and 85). This result holds for large n even without normality, in the sense that $[1 - n/N + c_g^2]/(1 - n/N)[1 + q/(n - q - 2)]$ still converges to 1 as n increases for fixed q (under weak conditions).

Expression (5) makes the dependence on q explicit. As q increases we may expect σ^2 to decrease but $E_M(c_g^2)$ to increase. The reduction of σ^2 may be expected to be small after a few important x variables are included and thus the variance may be expected to start increasing at some point where the number of x variables is a nonnegligible fraction of the sample size.

Results (4) and (5) are based on strong modelling assumptions and hence provided us only with motivation. In the general case $\bar{x} - \bar{X} = O_p(n^{-1/2})$ (under the randomization distribution with standard regularity conditions) so that the

last term of (3) is of $O_p(n^{-2})$. A more general second order asymptotic approximation for the design mean squared error of $\bar{y}_r$ when model (2) need not hold may be obtained by generalising Theorem 4.1 of Deng and Wu (1987). Details are given in Silva (1996).

Our aim is to develop a variable selection procedure that minimizes the estimated mean squared error of $\bar{y}_r$, and estimators of this mean squared error are considered next.

3. ESTIMATION OF THE MEAN SQUARED ERROR OF THE MULTIPLE REGRESSION ESTIMATOR

A simple estimator of the mean squared error of $\bar{y}_r$ is obtained by generalizing expression (7.29) of Cochran (1977, p. 195) to the case of several auxiliary variables:

$$v_s = \frac{1-f}{n}\hat{S}_e \tag{6}$$

where $\hat{S}_e = (n-q-1)^{-1}\sum_{i\in s}\hat{e}_i^2$ and $\hat{e}_i = (y_i - \bar{y}) - (x_i - \bar{x})'b$.

This estimator makes no allowance for the $O(n^{-2})$ component of the mean squared error, however. Thus, as a second mean squared error estimator, we generalize the estimator v_d studied in Deng and Wu (1987) to the case of general q. This is a special case of the model-based, bias-robust variance estimator G_2 originally proposed by Royall and Cumberland (1978), for the case where the residual variances in the model (2) are constant. This estimator is given by

$$v_d = \frac{1-f}{n(n-1)}\sum_{i\in s}\alpha_i\hat{e}_i^2 \tag{7}$$

where

$$\alpha_i = (g_i^2 - 2g_i f + f)/\{(1-f)[1-(x_i-\bar{x})'\hat{S}_x^{-1}(x_i-\bar{x})/(n-1)]\}.$$

We originally conjectured that v_d would be second order unbiased, as Deng and Wu (1987, eq. 4.4) show that it is for the case of $q = 1$. However this turns out not to be the case for general $q > 1$, although it may be expected that the bias of v_d is smaller than that of v_s, as indicated by the second order bias expressions for v_s and v_d obtained by Silva (1996).

A difficulty with v_d as a variance estimator is that it does not generalize easily to complex survey designs. Thus we consider as a third variance estimator a modified version of an estimator proposed by Särndal, Swensson and Wretman (1989), defined as:

$$v_g = \frac{1-f}{n(n-q-1)}\sum_{i\in s} g_i^2\hat{e}_i^2. \tag{8}$$

This estimator may be expected to behave similarly to v_d since $\alpha_i = g_i^2 + O_p(n^{-1/2})$. In the terminology of Särndal *et al.* (1992, p. 232), the g_i are the appropriate *g-weights* under simple random sampling if (2) is adopted as the underlying model. Expression (8) differs from the corresponding estimator proposed by Särndal *et al.* (1989, example 4.4) in that we use the denominator $(n-q-1)$ instead of the original $(n-1)$.

4. VARIABLE SELECTION PROCEDURES

We consider two basic variable selection procedures. First, an *all subsets* approach that involves computing one of the mean squared error estimators v_s, v_d or v_g of section 3 for all 2^q possible subsets of the q auxiliary variables (always including the intercept) and choosing that subset corresponding to the smallest mean squared error estimate. This procedure can clearly involve considerable computation if q is large. Thus as a second procedure, we consider a *forward selection* approach which starts with the sample mean as an estimator, then adds that variable which minimizes the mean squared error estimate. The procedure is repeated until the mean squared error estimate starts to increase, at which point the subset of variables which gave the minimum mean squared error estimate is selected.

These procedures may be contrasted with an approach inspired by the work of Bankier and his associates – see Bankier (1990) and Bankier, Rathwell and Majkowski (1992). We call this a *condition number reduction approach.* To describe the approach, first note that the regression estimator in (1) can alternatively be expressed as

$$\bar{y}_r = [n\bar{y} + (N\bar{X}^* - n\bar{x}^*)'(X_s^{*\prime}X_s^*)^{-1}X_s^{*\prime}y_s]/N \tag{9}$$

where X_s^* is the $n\times(q+1)$ matrix with $x_i^{*\prime} = (1, x_{i1}, \ldots, x_{iq})' = (1 : x_i')$ as its i-th row, $\bar{x}^* = (1 : \bar{x}')'$ and $\bar{X}^* = (1 : \bar{X}')'$ are the sample and population mean vectors of x_i^* respectively, and y_s is the $n\times 1$ vector with the sample observations of the response.

The regression estimator thus depends on the inversion of the cross-products matrix $X_s^{*\prime}X_s^*$, a matrix which can sometimes become ill-conditioned and thereby inflate the variance of the regression estimator.

Bankier (1990) proposed a two-step procedure for computing regression estimators of means (or totals) in which columns of the auxiliary data matrix X_s^* were eliminated in order to reduce the condition number of the cross-products matrix $X_s^{*\prime}X_s^*$, as well as to avoid undesirable situations (negative or outlying weights, rare characteristics, or exact linear dependence between columns). Bankier *et al.* (1992) describe in detail the procedure as applied to the 1991 Canadian Population Census. It is worth noting that the approach developed by Bankier and associates, although incorporating variable selection, is not targeted at achieving efficiency for a particular survey variable. Its main focus is on calibration, while at the same time providing a single set of weights that are used for all survey variables.

The condition number reduction approach that we consider can be described by the algorithm below, which adopts a backward elimination procedure to discard auxiliary variables generating large condition numbers for the cross-products matrix $CP = X_s^{*\prime} X_s^*$, instead of the forward inclusion of variables described by Bankier *et al.* (1992).

1) Compute the cross-products matrix $CP = X_s^{*\prime} X_s^*$ considering all the columns initially available (saturated subset).

2) Compute the Hermite canonical form of CP, say H (see Rao 1973, p.18), and check for singularity by looking at the diagonal elements of H. Any zero diagonal elements in H indicate that the corresponding columns of $X_s^{*\prime} X_s^*$ (and X_s^*) are linearly dependent on other columns (see Rao 1973, p. 27). Each of these columns is eliminated by deleting the corresponding rows and columns from $X_s^{*\prime} X_s^*$.

3) After removing any linearly dependent columns, the condition number $c = \lambda_{max}/\lambda_{min}$ of the reduced CP matrix is computed, where λ_{max} and λ_{min} are the largest and smallest of the eigenvalues of CP, respectively. If $c < L$, a specified value, stop and use all the auxiliary variables remaining.

4) Otherwise perform backward elimination as follows. For every k, drop the k-th row and column from CP, and recompute the eigenvalues and the condition number of the reduced matrix. Compute the condition number reductions $r_k = c - c_k$, where c_k is the condition number after dropping the k-th row and column from CP. Determine $r_{max} = \max_k(r_k)$ and $k_{max} = \{k: r_{max} = r_k\}$ and eliminate the column k_{max} by deleting the k_{max} row and column from CP. Make $c = c_{k_{max}}$ and iterate while $c \geq L$ and $q \geq 2$, starting each new iteration with the reduced CP matrix resulting from the previous one.

One further approach that we consider is the ridge regression estimator of Bardsley and Chambers (1984). It does not rely on selecting subsets from the auxiliary variables available, but rather on relaxing the calibration properties of the regression estimator in favour of more stable estimates. The ridge regression estimator is given by

$$\bar{y}_{BC} = [n\bar{y} + (N\bar{X}^* - n\bar{x}^*)'(\lambda C^{-1} + X_s^{*\prime} X_s^*)^{-1} X_s^{*\prime} y_s]/N \qquad (10)$$

where λ is a scalar ridging parameter and C is a diagonal matrix of "cost" coefficients associated with the calibration errors tolerated when estimating totals of the auxiliary variables using $\bar{y}_{BC}$.

Bardsley and Chambers (1984) suggested that the specification of the matrix C could be used to control the influence of each auxiliary variable on the resulting estimator of the response mean, thus imitating the subset selection process. As for the ridging parameter λ, they suggested taking the smallest value such that all the implicit case weights are not smaller than $1/N$ (or 1 for estimating totals).

5. PROPERTIES OF REGRESSION ESTIMATORS AFTER VARIABLE SELECTION

For our basic variable selection procedures, a set of estimation strategies $S = \{(\bar{y}_r^\gamma, v^\gamma); \gamma \in \Gamma\}$ is considered, where $\bar{y}_r^\gamma$ and v^γ are the regression estimator and an estimator of its variance respectively for a subset γ of the q auxiliary variables available, and Γ is the set of all subsets. The variable selection procedure selects a subset γ^* from Γ according to a rule which is determined by the data and by S, and the resulting point estimator is $\bar{y}_r^{\gamma^*}$.

For each fixed subset γ, it follows under standard regularity conditions (Isaki and Fuller 1982) that $\bar{y}_r^\gamma$ is consistent for the population mean $\bar{Y}$, that is $\bar{y}_r^\gamma - \bar{Y} = o_p(1)$. Now, for given $\delta > 0$, $|\bar{y}_r^{\gamma^*} - \bar{Y}| > \delta$ implies $|\bar{y}_r^\gamma - \bar{Y}| > \delta$ for some γ, and so we have

$$\Pr(|\bar{y}_r^{\gamma^*} - \bar{Y}| > \delta) \leq \sum_{\gamma \in \Gamma} \Pr(|\bar{y}_r^\gamma - \bar{Y}| > \delta) \qquad (11)$$

and because Γ is finite, the right hand side of (11) converges to zero, and it follows that $\bar{y}_r^{\gamma^*}$ is also consistent.

The distribution of $\bar{y}_r^{\gamma^*}$ will, however, depend on the selection rule in a complex way. See Grimes and Sukhatme (1980) for an investigation of the efficiency of $\bar{y}_r^{\gamma^*}$ in the simplest case when there are just two possible estimators: a regression estimator with one x variable and a difference estimator (a special case of which is the mean) and the variables are jointly normally distributed.

In contrast to the consistency of $\bar{y}_r^{\gamma^*}$, there is no reason why v^{γ^*} should be consistent for $\mathrm{Var}(\bar{y}_r^{\gamma^*})$, even if v^γ is consistent for $\mathrm{Var}(\bar{y}_r^\gamma)$ for each fixed γ. In particular we may expect v^{γ^*} to underestimate $\mathrm{Var}(\bar{y}_r^{\gamma^*})$ if the selection rule is such that v^{γ^*} is the minimum of the v^γ. This effect is similar to the well known overestimation of R^2 after subset selection in standard multiple linear regression (Miller 1990, p. 7-10).

6. A SIMULATION STUDY

In this section we present a small simulation study carried out to evaluate the performance of the alternative variable selection procedures considered. We took as our simulation population a data set comprising 426 records for heads of household surveyed using the sample (long) questionnaire during the 1988 Test Population Census of Limeira, in São Paulo state, Brasil.

This test was carried out as a pilot survey during the preparation for the 1991 Brazilian Population Census. The test consisted of two rounds of data collection. In the first round, each enumerator would visit all the occupied households in a given enumeration area (an area with between 200 and 300 households on average) and would fill in a short questionnaire. This form contained a few questions about characteristics of the household and about each member of the household (sex, age, relationship to head of household

and literacy). For heads of household only, a question on education and another about monthly total income were also included. The reported monthly total income for heads of household provides only a proxy to the actual income, due to the limitations of the interviewing process in this first round of data collection.

Then a second round of data collection was undertaken in each enumeration area. The same enumerators would visit a sample of 1 in 10 of the households (selected systematically from the list of occupied households compiled in the first round of data collection) to obtain information using a long (more detailed) questionnaire, which contained all the questions asked in the short form plus many other questions.

The size of the surveyed population was approximately 44,000 households with 188,000 individuals. The sample size was roughly 10% of the population size. For reasons of computational cost, we used in our simulation study a sub-population comprising all the sample records for 426 heads of household living in 20 of the 170 enumeration areas. We chose these records as our simulation population because they contain all the detailed information provided in the sample questionnaire, as well as the proxy information available from the first round interviews using the short form.

We considered total monthly income, as obtained from the long form, as the main response variable (y) together with 11 potential auxiliary variables, namely:

x_1 = indicator of sex of head of household equal male;
x_2 = indicator of age of head of household less than or equal to 35;
x_3 = indicator of age of head of household greater than 35 and less than or equal to 55;
x_4 = total number of rooms in household;
x_5 = total number of bathrooms in household;
x_6 = indicator of ownership of household;
x_7 = indicator that household type is house;
x_8 = indicator of ownership of at least one car in household;
x_9 = indicator of ownership of colour TV in household;
x_{10} = years of study of head of household;
x_{11} = proxy of total monthly income of head of household.

From these 11 variables, we constructed two alternative sets of auxiliary variables for our simulations. The first set was defined by taking five auxiliary variables, namely $x_1,...,x_4$ and x_{11}, that have reasonable explanatory power in predicting y, especially due to the presence of the proxy income x_{11}. The second set we considered contained ten auxiliary variables, namely $x_1,...,x_{10}$, which due to the exclusion of x_{11}, has smaller predictive power than the previous one. For reference, the population correlation matrix for the survey variable y and the 11 auxiliary variables in the population is given in Table 3.

We then selected 1,000 samples of size 100 from this simulation population by simple random sampling without replacement.

Before proceeding to examine the detailed simulation results, we first consider the potential for gains from variable selection following the motivating model-based discussion of section 2. Recall from equation (4) that under model (2) the conditional variance of $\bar{y}_r$ is inflated by a term c_g^2 because of estimation of β. We evaluated the distribution of c_g^2 over the 1,000 samples for both the cases of five and ten auxiliary variables. For the case of five auxiliary variables, the median value of c_g^2 was 0.036, with upper quartile of 0.056 and maximum 0.255. This accords roughly with equation (5) which implies that under the model the expected value of c_g^2 is $(1 - n/N)q/(n - q - 2) = 0.041$. Note that the wide variation of c_g^2 across samples suggests that it may be sensible to adopt a procedure which selects a different set of variables for each sample. The variation of c_g^2 is even greater for the case of ten auxiliary variables, when the median was 0.078, the upper quartile was 0.107 and the maximum was 0.329, which also accords roughly with the expected value under the model of 0.087, according to equation (5). This interpretation clearly depends on the validity of the model (2), which is doubtful for these data, but it does suggest that there are potential efficiency gains to be made from variable selection.

Another way to assess the potential for efficiency gains from variable selection is to compute approximations to the variance of the regression estimator considering various subsets of the auxiliary variables available, using all the population records. Figure 1 displays a plot of the approximation given by a finite population version of equation (5) computed for increasing subsets of the ten auxiliary variables, where the variable added at each step is the one yielding the biggest decrease in the approximation. The values of the standard first order design-based approximation $(1 - f)S_e/n$ are also plotted for reference, although as has already been noted, this approximation is monotone non-increasing when new auxiliary variables are added. Simulation estimates of the mean squared error for the regression estimator corresponding to each subset are also plotted. The plot shows clearly that if a standard regression estimator with a fixed set of auxiliary variables is to be used, the subset with five predictors would be the best choice when

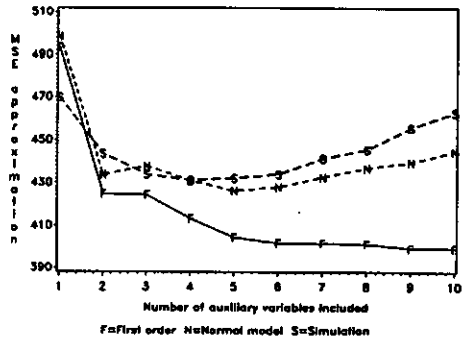


Figure 1. Finite population approximations and simulation estimations for the MSSE of the regression estimator computed for increasing subsets of the ten auxiliary variables.

the normal approximation for the variance based on expression (5) was considered, whereas the saturated subset would be chosen in case the standard design-based approximation for the variance was considered. The plot also reveals that the simulation estimates of the mean squared error agree more closely with the normal model approximation than with the standard first order approximation, especially for larger subsets of auxiliary variables. Similar results are achieved when corresponding variance approximations are computed given the set of five auxiliary variables.

Hence both the simulation distributions of c_g^2 and the finite population approximations to the variance of the regression estimator indicate that there are potential efficiency gains to be made from variable selection for this population. To investigate this for our data we now proceed to describe the details of the simulation study.

For each sample replicate (say s) and for each of the two alternative sets of auxiliary variables considered, estimates of the population mean of total monthly income were computed, as well as corresponding variance estimates, using a number of estimation strategies. Each estimation strategy is defined as a combination of a subset selection procedure, an estimator for the mean and a corresponding variance estimator. The list of all strategies considered follows.

SM) Sample mean estimator, with no auxiliary variables $(\bar{y}, v_s)$. This strategy provides the standard against which all the others will be compared.
Fs) Forward selection of auxiliary variables with $(\bar{y}_r, v_s)$.
Fd) Forward selection of auxiliary variables with $(\bar{y}_r, v_d)$.
Fg) Forward selection of auxiliary variables with $(\bar{y}_r, v_g)$.
Bs) Best subset selection from all subsets of auxiliary variables with $(\bar{y}_r, v_s)$.
Bd) Best subset selection from all subsets of auxiliary variables with $(\bar{y}_r, v_d)$.
Bg) Best subset selection from all subsets of auxiliary variables with $(\bar{y}_r, v_g)$.
FI) Fixed subset of auxiliary variables with $(\bar{y}_r, v_s)$.
SS) Saturated subset of auxiliary variables with $(\bar{y}_r, v_s)$.
FR) Forward subset selection using SAS PROC REG, with $(\bar{y}_r, v_s)$.
CN) Condition number reduction subset selection procedure with $(\bar{y}_r, v_s)$.
RI) Ridge regression estimator with saturated subset of auxiliary variables and a variance estimator that we denote v_{DC}, proposed by Dunstan and Chambers (1986), $(\bar{y}_{BC}, v_{DC})$.

Strategies Fs to Bg are variations of the two procedures we proposed for subset selection arising from the use of the three mean squared error estimators considered in section 3. Strategies FI and SS use the same set of auxiliary variables irrespective of the sample selected. In SS the saturated subset including all auxiliary variables available is always used. In FI a subset was chosen from each of the two sets with five (x_1, x_4, x_{11} chosen) or ten ($x_1, x_2, x_5, x_8, x_{10}$ chosen) auxiliary variables considered, by applying a standard forward subset selection regression procedure to the population dataset. The selected subsets were then used for every sample, thus the name "fixed subset" strategy for FI. This strategy would not be feasible in practice because the population information would not be available for the response, but it was considered as a theoretical "best possible scenario" under the traditional approach.

For the strategy FR, SAS PROC REG was used "naively" to perform a standard forward subset selection for each sample. The *p-value* used to decide whether a new variable should be included was the default of the procedure, namely 0.50. For more details, see SAS (1990, p. 1397).

For the condition number reduction subset selection strategy CN, the value used for the parameter L that controls the method was 1,000. For the ridge regression estimator strategy RI, the cost coefficients associated with calibration errors for different variables were all set equal to 1. After having chosen the value of λ that guarantees all the weights are not less than $1/N$, the weights were rescaled such that they sum to exactly 1, in order to ensure exact calibration when estimating the population size.

For any estimation strategy, the estimates of the population mean and its mean squared error for the sample s are denoted by $\bar{y}(s)$ and $v[\bar{y}(s)]$ respectively. The simulation results for each estimation strategy were summarised by computing estimates of the bias, mean squared error (MSE), and average of mean squared error estimates (AVMSE) from the set of 1,000 sample replicates, given respectively by

$$\text{BIAS} = \sum_s [\bar{y}(s) - \bar{Y}]/1{,}000 \tag{12}$$

$$\text{MSE} = \sum_s [\bar{y}(s) - \bar{Y}]^2/1{,}000 \tag{13}$$

$$\text{AVMSE} = \sum_s v[\bar{y}(s)]/1{,}000. \tag{14}$$

A measure of efficiency was also calculated for each strategy by dividing the corresponding simulation mean squared error by the simulation mean squared error for the sample mean (strategy SM) and multiplying the result by 100. Empirical coverage rates for 95% confidence intervals based on asymptotic normal theory were also computed for each estimation strategy and these rates, expressed as percentages, are presented in the last columns of Tables 1 and 2.

Table 1 displays the simulation results for estimation of the population mean of the response variable given the set of five auxiliary variables ($x_1 - x_4, x_{11}$) with larger predictive power. In this case, the use of the regression estimator greatly improves precision for every estimation strategy employed, except for subset selection using condition number reduction (CN). The bias was negligible (less than 1% in terms of the absolute relative bias) for all estimation strategies (the population mean of y is 194.34) except perhaps RI, which displayed a slight bias.

Table 1
Bias, Mean Squared Error, Average of Mean Squared Error Estimates, Efficiency and Empirical Coverage of Alternative Estimation Strategies for the Mean of Response Variable y with Five Auxiliary Variables ($x_1 - x_4, x_{11}$) Available

Estimation strategy	BIAS	MSE	AVMSE	Efficiency over SM (%)	Empirical[1] Coverage (%)
SM) Sample mean $(\bar{y}, v_s)$	0.25	620.09	619.05	100.00	91.8
Fs) Forward $(\bar{y}_r, v_s)$	0.40	233.78	239.62	37.70	82.7
Fd) Forward $(\bar{y}_r, v_d)$	-1.25	188.08	196.88	30.33	82.0
Fg) Forward $(\bar{y}_r, v_g)$	-1.28	188.38	192.73	30.38	81.1
Bs) Best $(\bar{y}_r, v_s)$	0.44	236.90	239.49	38.20	82.7
Bd) Best $(\bar{y}_r, v_d)$	-1.22	190.52	196.84	30.72	82.0
Bg) Best $(\bar{y}_r, v_g)$	-1.24	190.83	192.71	30.77	81.1
FI) Fixed $(\bar{y}_r, v_s)$	0.29	227.90	241.24	36.75	83.3
SS) Saturated $(\bar{y}_r, v_s)$	0.30	233.58	242.32	37.67	82.5
FR) PROC REG $(\bar{y}_r, v_s)$	0.38	235.86	240.26	38.04	82.5
CN) Cond. num. red. $(\bar{y}_r, v_s)$	0.34	507.33	483.63	81.82	89.8
RI) Ridge $(\bar{y}_{BC}, v_{DC})$	2.12	304.95	257.07	49.18	82.5

[1] Nominal 95% coverage.

There was no difference between the results for strategies based on forward selection (Fs-Fg) and corresponding strategies based on selection from all possible subsets (Bs-Bg). Hence the faster and cheaper forward selection procedures are preferable.

Amongst the strategies using forward subset selection, Fd and Fg (with v_d and v_g as the mean squared error estimators respectively) yielded greater efficiency, and performed very similarly. Note also that Fd and Fg performed better than FI and SS, the strategies that adopted the regression estimator with a fixed subset of the five auxiliary variables for every sample. This is true both for the saturated subset (SS) and when the fixed subset was chosen using information from the whole population (FI). This shows that one can do better than the traditional approach of using the regression estimator with a fixed set of auxiliary variables, by using an adaptive procedure that chooses the "best" regression estimator (subset) for each given sample, at least when the target response variable is the one considered for subset selection. This property was suggested by the wide variation in the values of c_g^2 between samples, where we may expect to benefit from a strategy which selects fewer x variables for samples with the largest values of c_g^2.

Comparison with the adaptive strategy FR, which used the standard subset selection available in PROC REG of SAS, shows that a criterion using an appropriate estimator of the mean squared error of the regression estimator makes some difference. FR yielded similar efficiency to that of traditional fixed subset strategies (FI-SS).

A more striking result is the low efficiency achieved by the subset selection procedure based on condition number reduction (CN) compared to all the other strategies based on the regression estimator. This was not unexpected, because that procedure did not take the response variable into account. This favours the argument that when the mean of some specified response variable is the main target for inference, this should be taken into account when selecting the auxiliary variables to use in connection with the regression estimator.

When the set of five auxiliary variables was considered, we also observed that, for every sample, the first variable eliminated to reduce the condition number was proxy income (x_{11}). This happened because eigenvalues (and hence condition numbers) of the CP matrix are dependent on the units of measurement of the auxiliary variables. Because all other auxiliary variables are counts of some kind, proxy income is the variable with the largest variance by far. Its exclusion for every sample provides some explanation for the poor performance of this approach, because it is the best single predictor for the response.

This difficulty was not apparent in Bankier's work, because in the target application of his procedure, the sample data from the 1991 Canadian Population Census, all the auxiliary variables considered were counts of persons, families or households, thus measured in similar units.

Unlike the eigenvalues of the CP matrix, the regression estimator is invariant to location and scale transformation of the auxiliary variables. To remove the arbitrary dependence of the condition number approach on the units of the auxiliary variables, it is therefore natural to standardise these variables first and to compute the condition number of the sample correlation matrix $\hat{R}_x$ rather than $X_s^{*\prime} X_s^*$. However this was tried and even modest values of L (100) failed to cause elimination of any auxiliary variables, which resulted in the saturated set being used every time, so that CN reduced to SS.

The strategy based on the ridge regression estimator (RI) performed worse than the saturated subset strategy (SS) in terms of efficiency. It also displayed some bias for estimating the mean squared error. This loss of efficiency is due to the

requirement that all the weights should be greater than or equal to $1/N$, which was imposed only under this strategy. On the other hand, it performed much better than the condition number reduction strategy CN in terms of efficiency.

In terms of the empirical coverage rates, only the condition number reduction strategy CN performed close to SM (sample mean), both leading to modest undercoverage. All the other strategies based on regression estimation yielded similar coverage rates, well below the target of 95%.

Results for the simulation carried out with the set of ten auxiliary variables ($x_1 - x_{10}$) are displayed in Table 2 below. As expected, these results show that the strategies that use the regression estimator still provide some gain in efficiency over the sample mean. However these gains are not as large as those reported in Table 1, when there are five auxiliary variables with higher explanatory power. As before, adaptive strategies based on forward subset selection performed similarly to their counterparts based on best subset selection from all possible subsets. Adaptive strategies using v_d or v_g as the estimator of the mean squared error were again slightly more efficient than the corresponding strategies based on v_s, although in this case at the expense of larger undercoverage of the corresponding nominal 95% confidence intervals.

The more efficient adaptive estimation strategies (Fd, Fg, Bd and Bg) display nonnegligible bias for both the population mean and for the mean squared error. In contrast, strategies FI and SS present no significant bias for the mean, although there is some bias in the mean squared error estimation under strategy SS. Note particularly the large negative bias of the estimators of the mean squared error, as indicated by the differences between the columns labelled MSE and AVMSE in Table 2. This appears to be worse for strategies Fd, Fg, Bd and Bg, followed by Fs and Bs, and not so bad for SS, FR and CN.

Comparing Fd and Fg with CN, there is a moderate gain in efficiency over the condition number reduction procedure, at the expense of some increased bias in both the mean and mean squared error estimators. Thus, even when the predictive power of the available auxiliary variables is not large, it is still possible to gain efficiency over strategy CN.

A bad choice of fixed subset (as for example, the saturated subset used in strategy SS) could yield poor results in terms of efficiency and also some bias in the mean squared error estimation. However, if for example v_d was used as the estimator for the mean squared error under strategy SS instead of v_s, there would be no apparent bias (the AVMSE observed in that case was 459.67, hence much closer to the estimated simulation mean squared error of 462.71).

The ridge regression estimator was again slightly inferior to the saturated subset strategy (SS), but now without any apparent bias in estimating the mean or the mean squared error. It outperformed the condition number reduction strategy CN once again in terms of efficiency, albeit by a smaller margin. It also performed well in terms of empirical coverage.

Strategy FR performed similarly to the fixed subset strategies FI and SS again, and so was outperformed by strategies using a specialized criterion based on an estimator of the mean squared error of the regression estimator such as v_d or v_g.

These results suggest that, when estimating the population mean of a single response, the proposed adaptive procedures combining the regression estimator with some form of subset selection based on an appropriate mean squared error estimator can offer some useful improvements in efficiency against its competitors. However such strategies may introduce some bias when the predictive power of the auxiliary variables available is not large, and the corresponding MSE estimators may be substantially biased, leading to poor coverage.

Table 2
Bias, Mean Squared Error, Average of Mean Squared Error Estimates, Efficiency and Empirical Coverage of Alternative Estimation Strategies for the Mean of Response Variable y with Ten Auxiliary Variables (x_1–x_{10}) Available

Estimation strategy	BIAS	MSE	AVMSE	Efficiency over SM (%)	Empirical[1] Coverage (%)
SM) Sample mean $(\bar{y}, v_s)$	0.25	620.09	619.05	100.00	91.8
Fs) Forward $(\bar{y}_r, v_s)$	0.06	468.46	397.99	75.55	86.7
Fd) Forward $(\bar{y}_r, v_d)$	-8.12	434.27	338.90	70.03	81.7
Fg) Forward $(\bar{y}_r, v_g)$	-7.90	433.71	328.46	69.94	81.6
Bs) Best $(\bar{y}_r, v_s)$	-0.00	466.16	397.59	75.18	86.6
Bd) Best $(\bar{y}_r, v_d)$	-7.90	434.54	336.88	70.08	81.5
Bg) Best $(\bar{y}_r, v_g)$	-7.60	433.26	326.05	69.87	81.6
FI) Fixed $(\bar{y}_r, v_s)$	0.45	490.49	461.86	79.10	89.0
SS) Saturated $(\bar{y}_r, v_s)$	-0.20	462.71	413.17	74.62	86.9
FR) PROC REG $(\bar{y}_r, v_s)$	-0.07	466.13	399.34	75.17	86.4
CN) Cond. num. red. $(\bar{y}_r, v_s)$	3.49	562.91	450.36	90.78	87.3
RI) Ridge $(\bar{y}_{BC}, v_{DC})$	1.05	480.18	472.82	77.44	89.4

[1] Nominal 95% coverage.

Table 3
Correlation Matrix for Variables Used in the Simulation Study with the 1988 Census Population

Variable	y	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
x_1	0.23										
x_2	-0.04	0.20									
x_3	0.17	0.07	-0.40								
x_4	0.47	0.13	-0.15	0.12							
x_5	0.48	0.09	-0.11	0.15	0.83						
x_6	0.05	-0.09	-0.32	-0.03	0.22	0.20					
x_7	-0.17	0.01	-0.12	-0.01	-0.17	-0.31	0.16				
x_8	0.38	0.29	0.07	0.17	0.44	0.41	0.13	-0.20			
x_9	0.20	0.08	-0.06	0.04	0.30	0.25	0.16	-0.13	0.37		
x_{10}	0.43	0.23	0.33	0.17	0.39	0.39	-0.10	-0.30	0.49	0.26	
x_{11}	0.78	0.23	-0.00	0.22	0.54	0.54	0.01	-0.19	0.41	0.21	0.49

7. CONCLUSIONS AND FUTURE DIRECTIONS

Our results suggest that, when using regression estimation, there is potential for some gain in efficiency by adopting a variable selection procedure based on one of the mean squared error estimators v_d or v_g. Under SRS, and considering the limited simulation evidence, there seems little to choose between these two mean squared error estimators.

Forward subset selection procedures were as effective as those based on searches carried out considering all possible subsets, which involve much more computation. Our results also indicate that it is possible to improve over subset selection procedures based on condition number reduction whenever a specific response variable is of interest.

One problem with a variable selection approach is that the associated variance estimation is likely to become biased for the estimation of the overall mean squared error of the regression estimator following variable selection, thus leading to poor coverage of standard confidence interval procedures. Further research is necessary to investigate possible alternative variance estimation procedures.

This paper has focused on the use of regression estimation to reduce sampling variance in the classical sampling context. In practice, regression estimation is widely used to correct for biases arising from non-sampling errors. In such applications the question of how many auxiliary variables to use is also an important one. Some variables might be included for reasons unrelated to sampling error, for example because they are known to be important determinants of nonresponse. Nevertheless, as the number of auxiliary variables increases the sampling variance may also eventually increase and we suggest that a decision rule to limit the number of auxiliary variables employed might still usefully be based on sampling variance considerations. In the presence of nonsampling bias, the difference between $\bar{x}$ and $\bar{X}$ will generally be of $O_p(1)$ not $O_p(n^{-1/2})$ and so the results of this paper are not directly applicable. Further research is therefore needed to consider the extension of our approach to this case.

Further research is also necessary to extend our approach to complex sampling designs. One possible approach for the general regression estimators, considered *e.g.* by Särndal *et al.* (1992, sec. 6.4), would be to replace the weights g_i by the "generalized" weights, described by Särndal *et al.* (1992, eq. 6.5.9), and to base variable selection on the minimization of the generalized version of v_g given by Särndal *et al.* (1992, eq. 6.6.4).

ACKNOWLEDGEMENTS

Pedro L.D. Nascimento Silva is grateful to CVCP-UK, CNPq-Brasil and IBGE-Brasil for financial support. The authors are grateful to Ray Chambers, Danny Pfeffermann, Jon Rao, Michael Bankier and two anonymous referees for comments. Michael Bankier was also very helpful for providing documentation and software about his GLSEP procedure.

REFERENCES

BANKIER, M.D. (1990). Two Step Generalized Least Squares Estimation. Ottawa: Statistics Canada, Social Survey Methods Division, internal report.

BANKIER, M.D., RATHWELL, S., and MAJKOWSKI, M. (1992). Two Step Generalized Least Squares Estimation in the 1991 Canadian Census. Methodology Branch Working Paper, SSMD, 92-007E, Statistics Canada.

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

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

DENG, L.Y., and WU, C.F.J. (1987). Estimation of variance of the regression estimator. *Journal of the American Statistical Association*, 82, 568-576.

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

DUNSTAN, R., and CHAMBERS, R.L. (1986). Model-based confidence intervals in mltipurpose surveys. *Applied Statistics*, 35, 276-280.

GRIMES, J.E., and SUKHATME, B.V. (1980). A regression-type estimator based on preliminary test of significance. *Journal of the American Statistical Association*, 75, 957-962.

HANSEN, M.H., and TEPPING, B.J. (1969). Progress and problems in survey methods and theory illustrated by the work of the United States Bureau of the Census. *New Developments in Survey Sampling*, (N.L. Johnson and H. Smith Jr., Eds.). New York: John Wiley & Sons.

ISAKI, C.T., and FULLER, W.A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77, 89-96.

MARDIA, K.V., KENT, J.T., and BIBBY, J.M. (1979). *Multivariate Analysis*. London: Academic Press.

MILLER, A.J. (1990). *Subset Selection in Regression*. London: Chapman and Hall.

RAO, C.R. (1973). *Linear Statistical Inference and its Applications* (2nd ed.). New York: John Wiley & Sons.

ROYALL, R.M., and CUMBERLAND, W.G. (1978). Variance estimation in finite population sampling. *Journal of the American Statistical Association*, 73, 351-358.

SAS INSTITUTE INC. (1990). *SAS/STAT User's Guide* (Version 6, Vol. 2, 4th ed.). Cary, NC: SAS Institute Inc.

SÄRNDAL, C.-E., SWENSSON, B., and WRETMAN, J. (1989). The weighted residual technique for estimating the variance of the general regression estimator of the finite population total. *Biometrika*, 76, 527-537.

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

SILVA, P.L.D.N. (1996). Some Asymptotic Results on the Mean Squared Error of the Regression Estimator Under Simple Random Sampling Without Replacement. Southampton: University of Southampton, Centre for Survey Data Analysis Technical Report 96-2.

Survey Methodology, June 1997
Vol. 23, No. 1, pp. 33-40
Statistics Canada

Diagnostics for Formation of Nonresponse Adjustment Cells, With an Application to Income Nonresponse in the U.S. Consumer Expenditure Survey

JOHN L. ELTINGE and IBRAHIM S. YANSANEH[1]

ABSTRACT

This paper discusses the use of some simple diagnostics to guide the formation of nonresponse adjustment cells. Following Little (1986), we consider construction of adjustment cells by grouping sample units according to their estimated response probabilities or estimated survey items. Four issues receive principal attention: assessment of the sensitivity of adjusted mean estimates to changes in k, the number of cells used; identification of specific cells that require additional refinement; comparison of adjusted and unadjusted mean estimates; and comparison of estimation results from estimated-probability and estimated-item based cells. The proposed methods are motivated and illustrated with an application involving estimation of mean consumer unit income from the U.S. Consumer Expenditure Survey.

KEY WORDS: Incomplete data; Missing data; Quasi-randomization; Response propensity; Sensitivity analysis; Weighting adjustment.

1. INTRODUCTION

1.1 Problem Statement

Survey analysts often use adjustment cell methods to account for nonresponse. The main idea is to define groups, or "cells", of sample units which are believed to have approximately equal response probabilities, or approximately equal values of a specific survey item, *e.g.*, income. Weighting adjustment or simple hot-deck imputation then is carried out separately within each adjustment cell. The resulting adjusted estimator of a population mean or total will have a nonresponse bias approximately equal to zero, provided the within-cell covariances between survey items and response probabilities are approximately equal to zero.

Some previous nonresponse-adjustment work formed adjustment cells through combinations of simple demographic or geographical classificatory variables. However, Little (1986) and others considered formation of cells by direct grouping of sample units according to their estimated response probabilities or estimated item values. The present paper discusses some simple diagnostics that are useful in implementing these cell-formation ideas. Principal attention is directed to the sensitivity of results to the number of cells used; identification of specific cells that require additional refinement; comparison of adjusted and unadjusted mean estimates; and comparison of estimation results from estimated-probability and estimated-item based cells. These diagnostics are illustrated with income data collected in the U.S. Consumer Expenditure Survey.

1.2 Notation, Nonresponse Bias, and Adjustment Cells

Let U be a fixed population of size N with survey items $Y_i, i \in U$; and consider estimation of the population mean $\bar{Y} = N^{-1}\sum_{i\in U} Y_i$. A sample s of size n is selected from U, and π_i is the probability that unit i is included in the sample.

Nonresponse is assumed to satisfy the following quasi-randomization model (Oh and Scheuren 1983). Let R_i be an indicator variable equal to 1 if the selected sample unit i is a respondent and equal to 0 otherwise. Assume that the R_i are mutually independent Bernoulli (η_i) random variables, where the fixed response probabilities η_i are allowed to differ across units. In addition, define the survey weights $\lambda_i = \pi_i^{-1}$ and the unadjusted survey-weighted mean response

$$\hat{\bar{Y}}_1 \stackrel{\text{def}}{=} \left(\sum_{i\in s}\lambda_i R_i\right)^{-1}\sum_{i\in s}\lambda_i R_i Y_i. \qquad (1.1)$$

Because of differences among the η_i, the unadjusted estimator $\hat{\bar{Y}}_1$ has a nonresponse bias approximately equal to $N^{-1}\bar{\eta}^{-1}\sum_{i\in U}\eta_i(Y_i - \bar{Y})$, where $\bar{\eta} = N^{-1}\sum_{i\in U}\eta_i$ and expectations are taken over both the original sample design and the quasi-randomization model. To reduce this bias, one often partitions the population into k "adjustment cells" U_h, partitions the sample s into corresponding groups s_h, and then uses the adjusted estimator

$$\hat{\bar{Y}}_k \stackrel{\text{def}}{=} \sum_{h=1}^{k} w_h \bar{Y}_{hR}, \qquad (1.2)$$

where $w_h = (\sum_{i\in s}\lambda_i)^{-1}\sum_{i\in s_h}\lambda_i$ and $\bar{Y}_{hR} = (\sum_{i\in s_h}\lambda_i R_i)^{-1}$ $\sum_{i\in s_h}\lambda_i R_i Y_i$. Note that if $k = 1$, then estimators (1.1) and (1.2) are identical. For some general discussion of adjustment cell methods see, *e.g.*, Cassel, Särndal and Wretman (1983), Oh and Scheuren (1983), and Kalton and Maligalig (1991).

The adjusted estimator $\hat{\bar{Y}}_k$ has remaining nonresponse bias approximately equal to

$$N^{-1}\sum_{h=1}^{k}\bar{\eta}_h^{-1}\sum_{i\in U_h}(\eta_i - \bar{\eta}_h)(Y_i - \bar{Y}_h), \qquad (1.3)$$

[1] John L. Eltinge, Department of Statistics, Texas A&M University, College Station, TX 77843-3143, U.S.A.; Ibrahim S. Yansaneh, Westat, 1650 Research Blvd., Rockville, MD 20850-3195, U.S.A.

where N_h is the number of units in U_h and $(\bar{\eta}_h, \bar{Y}_h) = N_h^{-1} \sum_{i \in U_h} (\eta_i, Y_i)$. Consequently, one prefers to construct cells such that the population covariance between η_i and Y_i is approximately equal to zero within each cell. In practice, one attempts to accomplish this by constructing cells that are approximately homogeneous in the response probabilities η_i or in the items Y_i, or both. In some cases, "natural" sets of cells are defined *a priori* through combinations of classificatory variables that are available for both respondents and nonrespondents. For example, Ezzati and Khare (1992) used 72 cells defined by age, race, region, urbanization status, and household size to perform nonresponse adjustments for part of the National Health and Nutrition Examination Survey. In many practical cases, however, the list of reasonable candidate variables for cell formation is fairly large, and may produce a substantial number of cells that contain few, if any, respondents. Consequently, several authors have developed methods to screen out the less important classificatory variables and to collapse sparse adjustment cells in a way that preserves a reasonable degree of homogeneity within each of the remaining cells. See, *e.g.*, Tremblay (1986); Lepkowski, Kalton and Kasprzyk (1989); Kalton and Maligalig (1991); Göskel, Judkins and Mosher (1991); and the related discussion of pooling of poststrata in Little (1993). In addition, adjustment cell methods are related to other methods like regression-based adjustments (*e.g.*, Rao 1996, Section 2.4 and references cited therein) and generalized raking (Deville, Särndal and Sautory 1993).

1.3 Adjustment Cells Based on Estimated Response Propensities or Predicted Items

Adjustment cells are expected to be approximately homogeneous, so one may argue that such cells implicitly define a model for either the η_i or Y_i values, or both. More explicit modeling leads to two related cell formation methods. First, let X_i be a vector of auxiliary variables observed for both responding and nonresponding sample units i, and use the sample (R_i, X_i) values to fit a model for $\eta_i = \eta(X_i)$ through linear, logistic, or probit regression. The sample cells s_h are then formed by grouping the sample units according to their estimated response probabilities $\hat{\eta}_i$. As a second alternative, consider regression of responses Y_i on an auxiliary vector X_i to produce estimated items $\hat{Y}_i$ for both responding and nonresponding sample units. The sample cells s_h are then formed by grouping units according to the values $\hat{Y}_i$.

These two methods were suggested by Little (1986), extending the observational-data propensity-score work of Rosenbaum and Rubin (1983, 1984). See also David, Little, Samuhel and Triest (1983). These ideas were developed originally in a model- based context, but extend directly to the current framework. Little (1986) argued that use of cells based on either the $\hat{\eta}_i$ or $\hat{Y}_i$ values could reduce nonresponse bias, and that the $\hat{Y}_i$-based cells could also control variance. Also, in some cases the $\hat{\eta}_i$ and $\hat{Y}_i$-based cells can be more flexible than cells defined *a priori*. In addition, the $\hat{Y}_i$-based adjustment cells are conceptually related to optimum stratification ideas (*e.g.*, Cochran 1977, Sections 5A.7-5A.8).

Little (1986) did not propose a specific rule to determine cell divisions. However, in keeping with related observational-data work by Cochran (1968) and by Rosenbaum and Rubin (1984), one may consider cell divisions defined by the estimated $k^{-1}j$ quantiles of the $\hat{\eta}_i$ or $\hat{Y}_i$ populations, $j = 1, 2, ..., k - 1$. This equal-quantile method gives some control over the expected number of respondents in each cell. In addition, review of the preceding two references suggests that, for a given set of predictors X_i, most of the feasible bias reduction may be achieved with a relatively small number of cells, say $k = 5$. A case study by Czajka, Hirabayashi, Little and Rubin (1992) used $k = 6$ $\hat{\eta}_i$-based adjustment cells within each of several strata, using cell-formation rules that were somewhat more complex than the equal-quantile rule considered here. However, the potential adequacy of a small number of cells should not be over-interpreted. For example, if an important regressor is omitted, then the resulting cell-based adjusted estimators may retain a substantial amount of bias, regardless of the specific number of estimated-probability or estimated-item based cells used.

Finally, an important alternative to weighting adjustment is imputation. For example, simple hot-deck imputation replaces a missing value within a given adjustment cell by randomly selecting respondent donors from the same cell. In parallel with (1.1) and (1.2), the resulting mean estimator is $\bar{Y}_{imp} = (\sum_{i \in s} \lambda_i)^{-1} \sum_{i \in s} \lambda_i Y_i^*$, where Y_i^* is either an observed or imputed value, as appropriate. Practical applications often use weighting adjustment for unit nonresponse and imputation for item nonresponse. However, for a given set of cells, both the weighting adjustment point estimator (1.2) and the imputation estimator $\bar{Y}_{imp}$ have the same approximate bias (1.3). For simplicity, the remainder of this paper will focus on weighting adjustment, but one should bear in mind that for a given set of cells, the same bias-reduction issues arise regardless of whether those cells are used for weighting adjustment or simple hot deck imputation.

1.4 Outline of the Present Paper

This paper discusses some implementation details of the estimated-probability and estimated-item methods of cell formation. We devote special attention to diagnostics to identify problems in a specific set of cells, and motivate and illustrate these diagnostics with an extended example involving income nonresponse in the U.S. Consumer Expenditure Survey. Section 2 gives some general background on this income nonresponse problem. Section 3 describes and applies several diagnostics, including comparison of $\bar{Y}_k$ estimates and standard errors for several values of k (Section 3.1); partial assessment of within-cell bias (Section 3.2.1); assessment of cell widths relative to the precision of $\hat{\eta}_i$ estimates (Section 3.2.2); and comparison of the adjusted and unadjusted mean estimates $\bar{Y}_k$ and $\bar{Y}_1$ (Section 3.3). Section 4 shows that similar diagnostics can be applied to adjustment cells based on predicted incomes $\hat{Y}_i$,

and also compares the mean income estimates computed from estimated-probability and estimated-income based cells. Section 5 summarizes the main ideas used in this paper, and notes some areas for future research.

2. INCOME NONRESPONSE IN THE U.S. CONSUMER EXPENDITURE SURVEY

2.1 The Consumer Expenditure Survey, Weighting Methods and Variance Estimation

The U.S. Consumer Expenditure Survey (CE) is a stratified multistage rotation sample survey conducted by the Census Bureau for the Bureau of Labor Statistics. Sample elements are "consumer units", roughly equivalent to households. In the interview component of this survey, each selected sample unit is asked to participate in five interviews. The current CE weighting procedure accounts for initial selection probabilities, a noninterview adjustment, post-stratification based on several demographic variables, and additional refinements; see Zieschang (1990) and United States Bureau of Labor Statistics (1992). The complexity of the CE weighting work has led the BLS to use variance estimators based on pseudo-replication methods with 44 replicates. This pseudo-replication is approximately equivalent to standard balanced repeated replication (Wolter 1985, Ch. 3). All standard errors reported here are based on this pseudo-replication method, with all additional parameter estimation and weighting adjustment steps performed separately within each replicate.

2.2 Income Nonresponse

The noninterview adjustment in the current CE weighting procedure is generally considered to account adequately for unit nonresponse, *e.g.*, noncontact or refusal to participate in a specific interview. Thus, unit nonresponse in the CE will not be considered further here. However, the BLS has had concerns about possible bias in mean income estimates due to item nonresponse that occurs with income questions in the CE; some background is as follows.

Detailed income data are collected in the second and fifth interviews of the CE, and are used to produce estimates of mean consumer unit income (U.S. Bureau of Labor Statistics 1991) and other parameters. CE income data are collected through a complex set of questions, and nonresponse rates for these questions are relatively high. To provide a summary indication of response or nonresponse to the full set of income questions, the BLS classifies each second- or fifth-interview consumer unit as a complete or incomplete reporter of income. The formal definition of "complete income reporter" status is fairly complex; Garner and Blanciforti (1994) give a detailed discussion. Current BLS procedure estimates mean income with the unadjusted mean response $\hat{\bar{Y}}_1$ defined by (1.1), with the R_i equal to indicators of complete income reporting, Y_i equal to income, and weights λ_i as described in Section 2.1. The weighted mean $\hat{\bar{Y}}_1$ uses both second- and fifth-interview data from a specified time period, but does not make direct use of the CE panel-data structure. In parallel with this, the present paper will distinguish between second- and fifth-interview data only in the construction of $\hat{\eta}_i$ and $\hat{Y}_i$ models.

Here, we used data from the second and fifth interview reports from all consumer units that had a second interview scheduled during 1990. The second-interview data involved 5,125 interviewed units and the fifth-interview data involved 5,093 interviewed units. For each interviewed unit (both the complete and the incomplete income reporters), BLS records provided a large number of demographic and expenditure variables; these were used as auxiliary variables in the modeling work described in Sections 3 and 4 below. For both the second and the fifth interviews, approximately 14 percent of the interviewed consumer units were incomplete income reporters.

3. CELLS BASED ON ESTIMATED RESPONSE PROBABILITIES

We first considered construction of adjustment cells based on estimated response probabilities. Logistic regression models for the complete-income-reporter probabilities $\eta_i = \eta(X_i)$ were fit separately for the second and fifth interview data described in Section 2. Model fitting details, including model parameter estimates and standard errors, are reported in Yansaneh and Eltinge (1993). All variance estimates were computed by the pseudo-replication method described in Section 2. The final model fits were used to estimate complete-reporter probabilities $\hat{\eta}_i$ for each second- and fifth-interview unit. Following the strategy in Section 1.3, units were grouped according to their $\hat{\eta}_i$ values into a total of k cells, with cell boundaries defined by the equal-quantile method.

3.1 Initial Sensitivity Analysis for the Number of Cells Used

The first three columns of Table 1 report the adjusted point estimates $\hat{\bar{Y}}_k$ of mean income, and associated standard errors, for several values of k. Comparisons of these point estimates indicate the extent to which the adjusted estimates are sensitive to a specific choice of k. For $k \geq 5$, the reported point estimates are relatively stable, varying between \$32,630 and \$32,664. This is consistent with the suggestion in Section 1.3 that $k = 5$ cells may provide most of the effective bias reduction to be obtained from a given cell-formation method; see Rosenbaum and Rubin (1984, Section 1 and Appendix A) for some related mathematical background.

In addition, note that for $k \geq 3$, the standard errors of $\hat{\bar{Y}}_k$ are also relatively stable, ranging from \$508 to \$530. This is in partial contrast with the general idea that selection of an

appropriate number of cells hinges on a bias-variance trade-off. For the present dataset, it appears that the effective bias reduction occurs fairly quickly (at $k = 5$, say), while substantial variance inflation does not occur until some point beyond $k = 20$. This is not unreasonable, since even for $k = 20$, the number of income responses per cell remained fairly large (ranging from 461 to 569), and thus avoided the general unstable-estimator problem associated with increasing numbers of sparse cells. Conversely, bias-variance tradeoff problems may be more severe for moderate k in applications involving smaller effective sample sizes, *e.g.*, estimation for small subpopulations.

Table 1
Adjusted Estimates of Mean Income with Cell Boundaries Determined by Estimated Response Probability Quantiles

Number of Cells	Point Estimate	Standard Error	$SE(\hat{\bar{Y}}_k - \hat{\bar{Y}}_1)$	MSE Ratio $(\hat{\bar{Y}}_k)$
Unadjusted ($k = 1$)	32,967	569	N/A	N/A
$k = 3$ cells	32,736	530	112	1.30
$k = 4$ cells	32,779	518	122	1.28
$k = 5$ cells	32,630	523	138	1.53
$k = 6$ cells	32,664	515	122	1.51
$k = 10$ cells	32,640	514	116	1.58
$k = 15$ cells	32,638	515	118	1.58
$k = 20$ cells	32,634	508	118	1.63

3.2 Two Simple Cell Diagnostics

To complement the preceding sensitivity analysis, it is useful to study some sets of adjustment cells in additional detail. Let $C_1 = \{s_1, ..., s_k\}$ be a given candidate set of adjustment cells, *e.g.*, the $k = 3$ or $k = 5$ equal-quantile- division cells in Section 3.1. The cells in C_1 can be refined by using equal-quantile divisions with a larger value of k; or by directly splitting one or more of the cells in C_1. This refinement may be worthwhile if there are empirical indications: (1) that the within-cell mean estimator $\bar{Y}_{hR}$ may be substantially biased; or (2) that a cell is wide relative to the precision with which the η_i values are estimated. Subsections 3.2.1 and 3.2.2 use two simple diagnostic methods to address issues (1) and (2), respectively. In each subsection, the proposed diagnostics lead to identification of potential "problem cells", and to construction of a refined set of adjustment cells, C_2, say. Comparisons of estimates of $\bar{Y}$ based on C_1 and C_2 then lead to some conclusions regarding the preferred set of $\hat{\eta}_i$-based adjustment cells.

3.2.1 Assessment of Within-Cell Bias

As noted in Section 1.2, a given adjusted estimator $\hat{\bar{Y}}_k$ reduces, but may not entirely eliminate, nonresponse bias; and the residual bias of $\hat{\bar{Y}}_k$ depends on the biases of the within-cell mean estimates $\bar{Y}_{hR}$. Consider the alternative within-cell mean estimator

$$\bar{Y}_{h\eta} = \left(\sum_{i \in s_h} \hat{\eta}_i^{-1} \lambda_i R_i \right)^{-1} \sum_{i \in s_h} \hat{\eta}_i^{-1} \lambda_i R_i Y_i. \qquad (3.1)$$

If the $\hat{\eta}_i$ estimates were equal to the true response probabilities η_i, then (3.1) would be an approximately unbiased estimator of the true subpopulation mean $\bar{Y}_h$. In that case, an estimator of the within-cell bias $E(\bar{Y}_{hR} - \bar{Y}_h)$ would be $\hat{B}_h = \bar{Y}_{hR} - \bar{Y}_{h\eta}$, and the corresponding estimator of the overall bias $E(\bar{Y}_k - \bar{Y})$ would be $\hat{B} = (\sum_{h=1}^{k} \sum_{j \in s_h} \lambda_j)^{-1} \sum_{h=1}^{k} (\sum_{j \in s_h} \lambda_j) \hat{B}_h$.

Because the $\hat{\eta}_i$ values are subject to estimation error, the terms $\hat{B}_h$ and $\hat{B}$ give only a partial indication of potential bias problems. For example, a large value of $\hat{B}_h$ may reflect a substantial bias in $\bar{Y}_{hR}$, or may reflect biases in the alternative estimator $\bar{Y}_{h\eta}$ due to the errors $\hat{\eta}_i - \eta_i$; *cf.* the cautionary remarks in Little (1986, p. 146) regarding direct use of the weights $\hat{\eta}_i^{-1}$ in adjusted estimation of $\bar{Y}$. Thus, if one observes a large value of $\hat{B}_h$, it is worthwhile to consider refinement of cell h; but the final decision of whether to use the resulting refined set of cells will depend on whether the refined set produces a substantially different estimate of the overall mean $\bar{Y}$.

Tables 2 and 3 present $\hat{B}_h$ values and associated standard errors and t statistics for equal-quantile-division cells with $k = 3$ and $k = 5$, respectively. Note that for the case $k = 3$, the $\hat{B}_h$ diagnostics indicate a possible bias contribution from the lowest cell. This is consistent with the suggestion from Section 3.1 that $k = 3$ cells may not provide a satisfactory nonresponse adjustment. In addition, the corresponding value of $\hat{B}$ was 111, with a standard error of 75; this value of $\hat{B}$ is very close to the difference $\hat{\bar{Y}}_3 - \hat{\bar{Y}}_5 = 106$ of the estimates $\hat{\bar{Y}}_3$ and $\hat{\bar{Y}}_5$ from Table 1.

Table 2
Within-Cell $\hat{B}_h$ Statistics for Probability-Based Cells, $k = 3$

h	$\hat{B}_h$	$se(\hat{B}_h)$	$t = \hat{B}_h / se(\hat{B}_h)$
1	269	136	1.98
2	-19	43	-0.44
3	84	45	1.87

Table 3
Within-Cell $\hat{B}_h$ Statistics for Probability-Based Cells, $k = 5$

h	$\hat{B}_h$	$se(\hat{B}_h)$	$t = \hat{B}_h / se(\hat{B}_h)$
1	96	217	0.44
2	-72	116	-0.62
3	-52	56	-0.93
4	-16	27	-0.59
5	98	50	1.96

In light of the preceding results, the low-$\hat{\eta}_i$ cell from the $k = 3$ case was split in half. The upper bounds for the two new cells ($h = 1'$ and $h = 1''$, say) were determined by the

0.167 and 0.333 estimated quantiles of the $\hat{\eta}_i$ population. The resulting $\hat{B}_h$ values and standard errors were 90 and 197 for cell 1′, and −42 and 79 for cell 1″. In addition, the refined set of four cells had $\hat{B} = 30$, with a standard error of 75; and the adjusted estimate of $\bar{Y}$ equal to $32,652 and standard error of $518 were close to those obtained from the equal-quantile-division method with $k = 5$.

In contrast with the results for $k = 3$, the $\hat{B}_h$ results for $k = 5$ indicated relatively little basis for concern, with the possible exception of cell $h = 5$, which had a t statistic of 1.96. For $k = 5$, the value of $\hat{B}$ was 11, with a standard error of 93. Additional splitting of cell $h = 5$ did not lead to notable changes in either the estimate of $\bar{Y}$ or the associated standard errors. The $\hat{B}_h$ results, for equal-quantile-division cells with larger values of k showed even fewer indications of within-cell bias. For example, for $k = 6$ all six cells had $\hat{B}_h$ values with t statistics less than or equal to 1.65; and for $k = 10$, all cells had $\hat{B}_h$ values with t statistics less than or equal to 1.54.

3.2.2 Relation of Cell Widths to Precision of η_i Estimates

The relationship between the widths of adjustment cells and the widths of confidence intervals for the response probabilities η_i leads to another diagnostic for identification of potential problem cells. First, define $a_h = (\sum_{i \in s_h} \lambda_i R_i)^{-1} \sum_{i \in s_h} \lambda_i$, the nonresponse-adjustment factor used for responding units in cell h. Second, following standard results for logistic regression, note that an approximate 95% confidence interval for η_i is

$$(LB_i, UB_i) = ([1 + \exp\{-X_i'\hat{\theta} + 1.96 D_i^{1/2}\}]^{-1}$$

$$[1 + \exp\{-X_i'\hat{\theta} - 1.96 D_i^{1/2}\}]^{-1}),$$

where $\hat{\theta}$ is the vector of logistic regression parameter estimates, $D_i = X_i' \hat{V}_\theta X_i$, and $\hat{V}_\theta$ is the pseudo-replicate-based estimated covariance matrix for $\hat{\theta}$. Let $\bar{d}_h$ be the λ_i-weighted sample mean of the confidence interval widths $UB_i - LB_i$ for units i in cell h, and consider a comparison of $\bar{d}_h$ to the width of cell h. If cell h is relatively wide, both on an absolute scale and relative to $\bar{d}_h$, then division of this cell may produce two new cells with two substantially different weight factors a_h. Conversely, if $\bar{d}_h$ is substantially larger than the width of cell h, then differences among $\hat{\eta}_i$ in that cell may result more from estimation error than from differences in the true η_i. In that case, additional division of cell h is unlikely to produce much useful change in weight factors a_h; and thus there will be relatively little change in the resulting nonresponse-adjusted estimator of $\bar{Y}$.

Tables 4 and 5 report cell boundaries, cell widths, $\bar{d}_h$, and a_h values for $k = 5$ and $k = 10$, respectively. For $k = 5$, the widths of cells 2 through 5 were not large relative to the $\bar{d}_h$ values. Each of these cells is essentially split in half to produce the $k = 10$ cell case. The resulting pairs of a_h for $k = 10$ are relatively close to the corresponding a_h values in cells 2 through 5 with $k = 5$.

By contrast, with $k = 5$, cell 1 is over twice as wide as $\bar{d}_1$. When $k = 10$, this cell is divided into cells with somewhat different nonresponse adjustment weight factors a_h: 1.45 and 1.27, respectively. However, the corresponding cell-mean estimates are relatively close: $\bar{Y}_{1R}$ = $24,045 and $\bar{Y}_{2R}$ = $24,582 for $k = 10$. Thus, in this example, the nonresponse-adjusted estimates $\hat{\bar{Y}}_5$ and $\hat{\bar{Y}}_{10}$ are relatively close because four of the five cell divisions produced relatively small changes in weights, and because the other cell division produced two cells with similar cell means.

Table 4
Estimated-Probability Cell Boundaries, Cell Widths, Mean Confidence Interval Widths and Nonresponse Adjustment Factors, $k = 5$

h	Lower Bound	Upper bound	Cell Width	$\bar{d}_h$	a_h
1	0.384	0.810	0.426	0.197	1.35
2	0.810	0.861	0.051	0.139	1.20
3	0.861	0.894	0.033	0.110	1.13
4	0.894	0.924	0.030	0.088	1.08
5	0.924	0.994	0.070	0.067	1.07

Finally, the a_h factors in Table 5 indicate that mean response rates in the $k = 10$ cells fall in a moderate range, from $(1.45)^{-1} = 0.69$ to $(1.06)^{-1} = 0.94$. Some other nonresponse datasets involve a wider range, and thus are more likely to produce more pronounced cell-splitting results. Conversely, other nonresponse datasets may display a tighter distribution of response probabilities, and thus are less likely to display notable cell-splitting effects.

Table 5
Estimated-Probability Cell Boundaries, Cell Widths, Mean Confidence Interval Widths and Nonresponse Adjustment Factors, $k = 10$

h	Lower Bound	Upper Bound	Cell Width	$\bar{d}_h$	a_h
1	0.384	0.762	0.378	0.220	1.45
2	0.762	0.810	0.048	0.174	1.27
3	0.810	0.840	0.030	0.146	1.21
4	0.840	0.861	0.021	0.132	1.19
5	0.861	0.878	0.017	0.111	1.14
6	0.878	0.894	0.016	0.108	1.11
7	0.894	0.908	0.014	0.093	1.09
8	0.908	0.924	0.016	0.083	1.08
9	0.924	0.944	0.020	0.072	1.08
10	0.944	0.994	0.050	0.062	1.06

3.3 Comparison of Cell-Based Estimates to the Unadjusted Estimate

To conclude the assessment of $\hat{\eta}_i$-based cells, we compared the adjusted estimates $\hat{\bar{Y}}_k$ with the unadjusted

estimate $\hat{\bar{Y}}_1$. First, Table 1 indicates that for the reported values of $k \geq 5$, the differences $\hat{\bar{Y}}_1 - \hat{\bar{Y}}_k$ are greater than or equal to \$303. Second, for $k \geq 5$, the estimated standard errors of the differences $\hat{\bar{Y}}_1 - \hat{\bar{Y}}_k$ are all less than or equal to \$138, and the corresponding t statistics are all greater than 2.44. Thus, for $k = 5$, say, a formal test of the hypothesis H_0: $E(\hat{\bar{Y}}_1 - \hat{\bar{Y}}_5) = 0$ would be rejected at standard significance levels; *i.e.*, the adjustment-cell method has produced a significant change in the mean income estimate.

In addition, a rough comparison of the efficiencies of $\hat{\bar{Y}}_1$ and $\hat{\bar{Y}}_k$ follows from the estimated mean squared error ratio

$$\hat{\gamma}_k = \{\hat{V}(\hat{\bar{Y}}_k)\}^{-1}[\hat{V}(\hat{\bar{Y}}_1) + \max\{0, (\hat{\bar{Y}}_1 - \hat{\bar{Y}}_k)^2 - \hat{V}(\hat{\bar{Y}}_1 - \hat{\bar{Y}}_k)\}]$$

where $\hat{V}(\hat{\bar{Y}}_1)$, $\hat{V}(\hat{\bar{Y}}_k)$, and $\hat{V}(\hat{\bar{Y}}_1 - \hat{\bar{Y}}_k)$ are the pseudo-replicate-based variance estimates for the indicated means. To interpret this ratio, assume for the moment that $\hat{\bar{Y}}_k$ is an approximately unbiased estimator of $\bar{Y}$. Then $\hat{\gamma}_k$ is an estimator of the mean squared error of the unadjusted estimator $\hat{\bar{Y}}_1$, relative to the mean squared error of $\hat{\bar{Y}}_k$. Consequently, $\hat{\gamma}_k$ reflects the loss of efficiency incurred by using the biased, unadjusted estimator $\hat{\bar{Y}}_1$ instead of the adjusted, unbiased estimator $\hat{\bar{Y}}_k$. However, this interpretation should be viewed with some caution, since it depends on the assumption that $\hat{\bar{Y}}_k$ is approximately unbiased for $\bar{Y}$, and since the $\hat{\gamma}_k$ are functions of the random terms $\hat{\bar{Y}}_1 - \hat{\bar{Y}}_k$, $\hat{V}(\hat{\bar{Y}}_1)$, $\hat{V}(\hat{\bar{Y}}_k)$, and $\hat{V}(\hat{\bar{Y}}_1 - \hat{\bar{Y}}_k)$.

As suggested by a referee, one could also consider a mean squared error ratio

$$\{\hat{V}(\hat{\bar{Y}}_\eta)\}^{-1}[\hat{V}(\hat{\bar{Y}}_k) + \max\{0, (\hat{\bar{Y}}_k - \hat{\bar{Y}}_\eta)^2 - \hat{V}(\hat{\bar{Y}}_k - \hat{\bar{Y}}_\eta)\}]$$

where $\hat{\bar{Y}}_\eta$ equals expression (1.1) with λ_i replaced by $(\hat{\eta}_i)^{-1}\lambda_i$. This would amount to comparing each cell-based estimate $\hat{\bar{Y}}_k$ to $\hat{\bar{Y}}_\eta$. This is appropriate if $\hat{\bar{Y}}_\eta$ is approximately unbiased, but this unbiasedness may be problematic in some cases; *cf.* Little (1986, p. 146).

The final column of Table 1 reports the estimated ratios $\hat{\gamma}_k$ for specified values of k. For $k \geq 5$, each reported $\hat{\gamma}_k$ is greater than 1.5. Finally, note that each adjusted estimate $\hat{\bar{Y}}_k$ fell *below* the unadjusted estimate $\hat{\bar{Y}}_1$. This occurred because, for a given k, cells associated with larger response probabilities tended to have larger mean estimates $\bar{Y}_{hR}$. For example, for $k = 5$, the $\bar{Y}_{hR}$ values were \$24,333, \$33,729, \$33,398, \$34,620, and \$37,057 for $h = 1$ (the low $\hat{\eta}_i$ cell) through $h = 5$ (the high $\hat{\eta}_i$ cell), respectively.

4. CELLS BASED ON ESTIMATED INCOME VALUES

The general diagnostic ideas of Section 3 also apply to $\hat{Y}_i$ based cells. To illustrate this idea, we fit separate weighted regressions of Y_i = reported income for second- and fifth-interview respondents. Yansaneh and Eltinge (1993) report details of the work, including parameter estimates and standard errors. The resulting regression models were used to compute estimated incomes $\hat{Y}_i$ for both complete and incomplete income reporters. Units were then grouped into cells according to their $\hat{Y}_i$ values, with cell boundaries determined by the equal-quantile method.

Table 6 reports the basic sensitivity-analysis and efficiency results for the $\hat{Y}_i$ based cells; the organization of this table is the same as in Table 1. The sensitivity-analysis results are qualitatively similar, but not identical, to those reported for the $\hat{\eta}_i$-based cells. In additional work not detailed here, we considered splitting individual equal-quantile $\hat{Y}_i$-based cells. For $k \geq 4$, the resulting mean estimates and associated standard errors did not differ notably from those reported in Table 6.

Table 6
Adjusted Estimates of Mean Income with Cell Boundaries Determined by Estimated Income Quantiles

Adjustment Method	Point Estimate	Standard Error	$SE(\hat{\bar{Y}}_k - \hat{\bar{Y}}_1)$	MSE Ratio
Unadjusted ($k = 1$)	32,967	569	N/A	N/A
$k = 3$ cells	32,512	509	106	2.01
$k = 4$ cells	32,468	512	108	2.14
$k = 5$ cells	32,473	511	115	2.12
$k = 6$ cells	32,492	508	117	2.08
$k = 10$ cells	32,488	510	119	2.07
$k = 15$ cells	32,478	504	124	2.16
$k = 20$ cells	32,495	513	124	2.02

The final two columns of Table 6 permit comparison of $\hat{\bar{Y}}_k$ to the unadjusted estimate $\hat{\bar{Y}}_1$. For $k \geq 4$, the differences $\hat{\bar{Y}}_1 - \hat{\bar{Y}}_k$ are greater than or equal to \$472, with estimated standard errors less than or equal to \$124. The associated t statistics are all greater than 3.80. In addition, the estimated mean squared error ratios $\hat{\gamma}_k$ are all greater than 2.0.

Also, the $\hat{\eta}_i$ and $\hat{Y}_i$-based cells produced somewhat different adjusted estimates of mean income, but the observed differences were not statistically significant at customary α levels. For example, with $k = 5$, the difference between the $\hat{\eta}_i$- and $\hat{Y}_i$-based cell estimates is \$32,630 - \$32,473 = \$157, with a standard error of \$122 and a t statistic of 1.29. Similarly, for $k = 10$, the difference between the $\hat{\eta}_i$- and $\hat{Y}_i$-based estimates is \$152, with a standard error of \$104. Thus, the data provide relatively little power to distinguish between results of the two general cell-formation methods.

Finally, note that a given set of $\hat{Y}_i$-based cells are fundamentally linked with a particular Y variable, *e.g.*, consumer unit income. Consequently, that set of cells will not necessarily work well for estimation of the mean of a different Y variable.

5. DISCUSSION

5.1 Summary of Methods

This paper has discussed some simple diagnostics for formation of nonresponse adjustment cells. The methodology may be summarized as follows.

1. Based on preliminary modeling work and observed auxiliary variables X_i, compute an estimated response probability $\hat{\eta}_i$ for each sample unit (respondents and nonrespondents).
2. Construct k adjustment cells with boundaries determined by the estimated $k^{-1}j$ quantiles of the $\hat{\eta}_i$ population, $j = 1, 2, ..., k - 1$. Compute the resulting adjusted mean estimate, $\bar{\hat{Y}}_k$.
3. Repeat (2) for several integers $k > 1$. As k increases, identify the point at which the $\bar{\hat{Y}}_k$ become approximately constant. In keeping with Rosenbaum and Rubin (1984) and the empirical results discussed here, values of k near 5 may be of special interest.
4. Use simple screening diagnostics (*e.g.*, $\hat{B}_h$ and $\bar{d}_h$ in Section 3.2) to check for potential problems in the equal-quantile-division adjustment cells. If the diagnostics identify potential "problem cells," then try additional refinements of these cells. Compute estimates of $\bar{Y}$ based on these refined sets of cells, and compare these new estimates to the $\bar{\hat{Y}}_k$ from (3).
5. Assess the overall effect of adjustment by comparing the differences $\bar{\hat{Y}}_1 - \bar{\hat{Y}}_k$ to the standard errors $se(\bar{\hat{Y}}_1 - \bar{\hat{Y}}_k)$; and by computing the estimated mean squared error ratios $\hat{\gamma}_k$.
6. Repeat steps (1) through (5), as appropriate, for $\hat{Y}_i$-based adjustment cells. Compare the final estimates of $\bar{Y}$ obtained from the $\hat{\eta}_i$ and $\hat{Y}_i$-based cell methods.

5.2 Areas for Future Research

The results of this work suggest two potentially useful areas for future research. First, the CE income nonresponse problem is similar to nonresponse problems in some other large-scale surveys, but as with any case study one should not over-generalize the empirical results reported here. It would be useful to apply these diagnostics to problems involving different estimands (*e.g.*, cross-class means) or involving nonresponse datasets with somewhat different characteristics, *e.g.*, larger or smaller effective sample sizes; or wider or narrower distributions of $\hat{\eta}_i$ estimates. This in turn would offer additional insight into the operating characteristics of $\hat{\eta}_i$ and $\hat{Y}_i$-based adjustment cell methods in practical applications. Second, extensions to multivariate problems (*e.g.*, relationships involving second-interview and fifth-interview CE income data) also would be of interest.

ACKNOWLEDGEMENTS

The authors thank Richard Dietz, Thesia Garner, Paul Hsen, Eva Jacobs, Geoffrey Paulin, Stuart Scott, and Stephanie Shipp for many helpful discussions of the Consumer Expenditure Survey; and Wayne Fuller, Steve Miller, Geoff Paulin, Stuart Scott, three referees and the editor for helpful comments on earlier versions of this paper. This work was carried out while the authors were visiting the Bureau of Labor Statistics through the ASA/NSF/BLS Research Fellow Program, and was supported by a grant from the National Science Foundation (SES-9022443). Eltinge's research was also supported in part by a grant from the National Institutes of Health (CA 57030-04). The views expressed in this paper are those of the authors and do not necessarily represent the policies of the Bureau of Labor Statistics.

REFERENCES

CASSEL, C.-M., SÄRNDAL, C.-E., and WRETMAN, J.H. (1983). Some uses of statistical models in connection with the nonresponse problem. In *Incomplete Data in Sample Surveys*, (Vol. 3), (Eds. W.G. Madow, I. Olkin, and D. Rubin). New York: Academic Press, 143-160.

COCHRAN, W.G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*, 24, 205-213.

COCHRAN, W.G. (1977). *Sampling Techniques*. New York: Wiley.

CZAJKA, J.L., HIRABAYASHI, S.M., LITTLE, R.J.A., and RUBIN, D.B. (1992). Projecting from advance data using propensity modeling: An application to income and tax statistics. *Journal of Business and Economic Statistics*, 10, 117-131.

DAVID, M.H., LITTLE, R.J.A., SAMUHEL, M., and TRIEST, R. (1983). Imputation models based on the propensity to respond. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 168-173.

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.

EZZATI, T., and KHARE, M. (1992). Nonresponse adjustments in a national health survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 339-344.

GARNER, T.I., and BLANCIFORTI, L.A. (1994). Household income reporting: An analysis of U.S. Consumer Expenditure Survey data. *Journal of Official Statistics* 10, 69-91.

GÖKSEL, H., JUDKINS, D.R., and MOSHER, W.D. (1991). Nonresponse adjustments for a telephone follow-up to a national in-person survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 581-586.

KALTON, G., and MALIGALIG, D.S. (1991). A comparison of methods of weighting adjustment for nonresponse. *Proceedings of the 1991 Annual Research Conference*, U.S. Bureau of the Census, 409-428.

LEPKOWSKI, J., KALTON, G., and KASPRZYK, D. (1989). Weighting adjustments for partial nonresponse in the 1984 SIPP panel. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 296-301.

LITTLE, R.J.A. (1986). Survey nonresponse adjustments for estimates of means. *International Statistical Review*, 54, 139-157.

LITTLE, R.J.A. (1993). Post-stratification: A modeler's perspective. *Journal of the American Statistical Association*, 88, 1001-1012.

OH, H.L., and SCHEUREN, F.J. (1983). Weighting adjustment for unit nonresponse. In *Incomplete Data in Sample Surveys*, (Vol. 2), (Eds. W.G. Madow, I. Olkin and D.B. Rubin). New York: Academic Press, 143-184.

RAO, J.N.K. (1996). On variance estimation with imputed survey data. *Journal of the American Statistical Association*, 91, 499-506.

ROSENBAUM, P.R., and RUBIN, D.B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 41-55.

ROSENBAUM, P.R., and RUBIN, D.B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, 79, 516-524.

TREMBLAY, V. (1986). Practical criteria for definition of weighting classes. *Survey Methodology*, 12, 85-97.

UNITED STATES BUREAU OF LABOR STATISTICS (1991). News: Consumer Expenditures in 1990. Publication USDL 91-607, United States Department of Labor, Washington, DC.

UNITED STATES BUREAU OF LABOR STATISTICS (1992). BLS Handbook of Methods. Bulletin 2414, United States Department of Labor, Washington, DC.

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

YANSANEH, I.S., and ELTINGE, J.L. (1993). Construction of adjustment cells based on surrogate items or estimated response propensities. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 538-543.

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

Survey Methodology, June 1997
Vol. 23, No. 1, pp. 41-52
Statistics Canada

Variance Estimation for Measures of Income Inequality and Polarization – An Empirical Study

MILORAD S. KOVAČEVIĆ and WESLEY YUNG[1]

ABSTRACT

Measures of income inequality and polarization are fundamental to the discussions of many economic and social issues. Most of these measures are non-linear functions of the distribution function and/or the quantiles and thus their variances are not expressible by simple formulae and one must rely on approximate variance estimation techniques. In this paper, several methods of variance estimation for six particular income inequality and polarization measures are summarized and their performance is investigated empirically through a simulation study based on the Canadian Survey of Consumer Finance. Our findings indicate that for the measures studied here, the bootstrap and the estimating equations approach perform considerably better than the other methods.

KEY WORDS: Gini index; Lorenz curve ordinate; Low income proportion; Polarization index; Quantile share; Resampling variance estimation; Linearization method.

1. INTRODUCTION

Analyses of the distribution of income are fundamental to the discussions of important economic and social issues such as the extent of inequality, poverty, the size of the middle class, *etc.* There exists extensive statistical and econometric literature on this subject, especially on different measures of income inequality and their properties (Sen 1973, Kakwani 1980, Nygärd and Sandström 1981). However, seldom is there any attempt to produce information regarding the sampling variability associated with the estimates used to assess the magnitude of inequality or polarization. Such information is necessary for two reasons: i) as a measure of the precision of the estimates obtained from survey data and ii) to provide a basis for formal statistical inference on income distributions, particularly when income distributions are compared over different regions or across time.

Measures of income inequality and polarization are finite population parameters expressible as functions of the ordered population values, thus their variances are not obtainable in simple formulae and one has to rely on approximate variance estimation techniques. Generally, inference about these measures, based on a complex sample design, embodies point estimation and confidence intervals. We investigate variance estimation for some of these measures such as quantiles, low income line, low income proportion, Lorenz curve ordinates, quantile shares, Gini index, and the polarization index.

Throughout this paper we assume a fixed finite population framework, that is, we assume that associated with each population unit is a fixed but unknown real number: the value of income earned by the unit. We assume that the population is stratified into L strata with N_h primary sampling units (PSU's) in the h-th stratum. In the first stage sample, $n_h (\geq 2)$ PSU's are selected from stratum h (independently across strata). We assume that subsampling within sampled PSU's is performed to ensure unbiased estimation of PSU totals, $Y_{hc}, c = 1,...,n_h; h = 1,...,L$. Attached to the (hci)-th ultimate unit, along with the observed variable of interest, y_{hci}, is the sampling weight w_{hci}. We use $\sum_s = \sum_h \sum_c \sum_i$ to denote summation over all ultimate units in the sample, incorporating all stages of sampling.

After reviewing the basic definitions of these measures, we give their point estimates under our sample design in section 2. Section 3 deals with variance estimation of these measures. Existing methods are reviewed and five methods, jackknifing, grouped and repeatedly grouped balanced half-sample, bootstrap and linearization via the estimating equations approach are summarized in detail. Section 4 contains the description of the simulation study based on data collected in the 1988 Canadian Survey of Consumer Finance. The empirical study is aimed at comparisons of the variance estimation methods for a number of income inequality measures. Various results are presented, summarized and interpreted. Our conclusions are presented in section 5.

2. ESTIMATION OF INCOME INEQUALITY MEASURES

The simplest measures of inequality between two distributions are the cumulative distribution function (CDF) and the quantiles of the two distributions. We start this section by defining the CDF and the finite population quantiles. The remaining measures studied in this paper are functions of the CDF or a fixed number of quantiles and are introduced in section 2.1.

For a variable Y defined over a finite population $U = \{1,...,N\}$, we define the CDF as

[1] Milorad S. Kovačević, Senior Methodologist, Household Survey Methods Division, and Wesley Yung, Senior Methodologist, Business Survey Methods Division, Statistics Canada, R.H. Coats Building, Tunney's Pasture, Ottawa, Ontario, Canada, K1A 0T6.

$$F_N(y) = \sum_{i \in U} I\{Y_i \le y\} \frac{1}{N},$$

where $I\{a\}$ is an indicator function taking on a value of 1 if a is true and 0 otherwise. A design unbiased estimator of $F_N(y)$ is

$$\tilde{F}(y) = \sum_{i \in s} I\{y_i \le y\} \frac{w_i}{N}$$

where the sampling weights, w_i, are obtained from the sample design and are equal to the inverse of the first order inclusion probabilities. This estimator may not be a CDF since $\tilde{F}(\infty) = \hat{N}/N$ may not necessarily be equal to 1. Thus we would rather use the possibly design-biased estimator:

$$\hat{F}(y) = \sum_{i \in s} I\{y_i \le y\} w_i \Big/ \sum_{i \in s} w_i = \sum_{i \in s} I\{y_i \le y\} \tilde{w}_i, \qquad (2.1)$$

where $\tilde{w}_i = w_i / \sum_s w_i$, $i \in s$. The estimator (2.1) is design unbiased when $\sum_s w_i = N$ which can occur under simple random sampling or if the weights, w_i, are benchmarked to known population totals. In general, the estimator (2.1) uses final weights which usually involve poststratification, nonresponse adjustment, some iterative calibrations and so on. In this paper, we consider only the case where the design weights are used.

Turning to the quantiles, we define the finite population quantiles as

$$\xi_N(p) = \inf_{i \in U} \{Y_i \mid F_i \ge p\} \quad \text{for} \quad 0 < p \le 1,$$

where $F_i = F_N(Y_i)$. The population quantiles are estimated by the sample quantiles

$$\hat{\xi}_p = \inf_{i \in s} \{y_i \mid \hat{F}_i \ge p\} \quad \text{for} \quad 0 < p \le 1,$$

where $\hat{F}_i = \hat{F}(y_i)$. If a parameter is a function of quantiles, say $\theta_N = g(\xi_N)$ with $\xi_N = \{\xi_N(p_1), ..., \xi_N(p_k)\}$, then it is estimated by $\hat{\theta} = g(\hat{\xi})$ where $\hat{\xi} = (\hat{\xi}_{p_1}, ..., \hat{\xi}_{p_k})$.

2.1 Income Inequality and Polarization Measures as Finite Population Parameters

In this section we present some frequently used income inequality and polarization measures. They are the low income line, the low income proportion, the Lorenz Curve and its related statistics, the quantile shares, the Gini index and finally the polarization curve and the polarization index. Our intention is to briefly introduce these measures, not to discuss them in detail. For more details, we refer the readers to Nygård and Sandström (1981) and Wolfson (1994).

The *low income line*, or the *poverty line*, is defined as a fraction of the median, $\lambda_\alpha = \alpha\, \xi_N(0.5)$, where $0 < \alpha \le 1$ is a given constant and $\xi_N(0.5)$ is the finite population median. Its estimate is simply $\hat{\lambda}_\alpha = \alpha\, \hat{\xi}_{0.5}$.

The *low income proportion (LIP)* is the percentage of units (individuals, families, households) in the population falling below the low income line λ_α and is given by $\Lambda_\alpha = F_N(\lambda_\alpha)$. The estimate of the low income proportion involves the estimation of both the distribution function and the low income line, $\hat{\Lambda}_\alpha = \hat{F}(\hat{\lambda}_\alpha) = \sum_s I\{y_{hci} \le \alpha\, \hat{\xi}_{0.5}\} \tilde{w}_{hci}$.

The finite population *Lorenz curve ordinate (LCO)* gives the share of income received by the poorest $100p$ percent of the population and is defined as a function of p ($0 \le p \le 1$). It simply depicts the cumulative income against the population share. As a parameter it is defined as

$$L(p) = \frac{1}{\mu_Y} \int_0^p \xi_q \, dq$$

where μ_Y is the population mean, and ξ_q is the quantile function. For a large population without ties the expression above is approximated by

$$L_N(p) \approx \sum_U \frac{I\{F_i \le p\}\, Y_i}{\mu_N} \frac{1}{N}$$

and estimated as

$$\hat{L}(p) = \sum_s \frac{I\{\hat{F}_{hci} \le p\}\, y_{hci}}{\hat{\mu}} \tilde{w}_{hci}$$

where $\hat{\mu} = \sum_s \tilde{w}_{hci}\, y_{hci}$ and $\hat{F}_{hci} = \hat{F}(y_{hci})$.

The *quantile share (QS)* is defined as the proportion of total income shared by the population allocated to a quantile interval $[\xi_{p_1}, \xi_{p_2})$:

$$Q_N(p_1, p_2) \approx \sum_U \frac{I\{ep_1 \le F_i \le p_2\}\, Y_i}{\mu_N} \frac{1}{N} = L_N(p_2) - L_N(p_1)$$

For $0 \le p_1 < p_2 \le 1$ it is estimated by replacing the parameters with their estimates.

The most popular measure of aggregate inequality of income distribution, the *Gini index*, is defined as the area between the Lorenz curve and the 45° line, normalized to lie between 0 and 1: $G = 1 - 2\int_0^1 L(p)\,dp$. Its finite population version is estimated by

$$\hat{G} = \sum_s \frac{[2\hat{F}_{hci} - 1]\, y_{hci}}{\hat{\mu}} \tilde{w}_{hci}.$$

For more about the Gini index we refer the reader to Nygård and Sandström (1985).

Using the analogy of the Lorenz curve and the Gini index, Foster and Wolfson (1992) defined the *polarization curve* as

$$B(p) = \int_{0.5}^{p} \frac{F^{-1}(q) - \xi_{0.5}}{\xi_{0.5}}\, dq,$$

or in the finite population form

$$B(p) = \begin{cases} 0.5 - p - \dfrac{1}{\xi_{0.5}} \displaystyle\sum_U I\{p < F_i < 0.5\}\, Y_i \dfrac{1}{N}, & 0 < p \le 0.5, \\ 0.5 - p + \dfrac{1}{\xi_{0.5}} \displaystyle\sum_U I\{0.5 \le F_i < p\}\, Y_i \dfrac{1}{N}, & 0.5 < p \le 1. \end{cases}$$

The polarization curve shows, for any population percentile, how far its income is from the median. The area below the polarization curve is considered as a summary measure of the polarization. A version of it, normalized to lie between 0 and 1, is named the *polarization index* (PI):

$$PI_N = \sum_U \frac{[2 - 2I\{F_i \le 0.5\} - 2F_i]Y_i}{\xi_N(0.5)} \frac{1}{N}$$

where $\xi_N(0.5), \mu_N$ and F_i were previously defined. The estimate of the polarization index is obtained by replacing the parameters with their estimates.

3. VARIANCE ESTIMATION

The estimation of the variance of non-smooth statistics like quantiles, as well as quantile based functions like the low income proportion or the polarization index, is not straightforward especially when the assumption of simple random sampling is untenable and there is a need to take into account the complex sample design. In the first part of this section we review some results on variance estimation for quantiles as a starting point for understanding the complexity of variance estimation for income inequality measures. We also review results on variance estimation for some measures like the Lorenz curve ordinates. The second part describes the methods of variance estimation that are used in this study.

Woodruff (1952) proposed a method to obtain confidence intervals for individual quantiles. These intervals were used by Francisco and Fuller (1986) and Rao and Wu (1987) to derive variance estimators. Though the estimator depends on the confidence coefficient, Rao and Wu (1987) established its asymptotic consistency for any significance level α. Using Monte Carlo simulations, they studied the standard errors of quantiles for cluster samples estimated in this manner. Their results suggest that a 95% confidence interval works well as a basis for extracting the standard error. Binder (1991) obtained a similar form of the variance estimator by using the linearization method.

Jackknife variance estimators have become extremely popular for smooth functions of totals and means with the increase in computing power. Standard asymptotic theory applied to the median of a distribution with bounded continuous density, f, shows that $nE(\hat{\xi}_{0.5} - \xi_{0.5})^2 \to 1/[4f^2(\xi_{0.5})]$ as $n \to \infty$. Efron (1979) pointed out that the jackknife method applied to the sample median gives a variance estimate which is asymptotically inconsistent since

$$n \operatorname{var}_{JK}(\hat{\xi}_{0.5}) \to \frac{1}{4f^2(\xi_{0.5})} [\chi_2^2/2]^2$$

where $[\chi_2^2/2]^2$ has mean of 2 and variance of 20 which means that the jackknife variance estimator tends to over estimate, on the average, the correct asymptotic variance by 100%. Kovar (1987) confirmed empirically the inconsistency of the delete-one-unit jackknife estimators for a stratified sample design. In a simulation study using a stratified population, he showed that the delete-one-unit jackknife estimators (he considered six of them) performed poorly, over estimating the true variance by 30-70% in the design with two units per stratum and performed even worse in the five units per stratum design. Shao and Wu (1989), however, have shown that under certain conditions, the delete-d jackknife method has desirable asymptotic properties for variance estimation of non-smooth statistics. This result has motivated Rao, Wu and Yue (1992) to apply the delete-one-PSU jackknife for stratified multistage sampling. In a limited simulation study they found that both bias and relative bias of the jackknife variance estimator of the median decrease as the cluster size increases for a fixed intracluster correlation.

Bootstrap variance estimation for the median was first reported by Efron (1979), and in the case of independent and identically distributed observations the bootstrap provides consistent results, (see also Babu 1986). Rao and Wu (1988) gave a modified bootstrap method for variance estimation in stratified designs. Kovar (1987) and Kovar, Rao and Wu (1988) reported good performance for medians when the size of the bootstrap sample is $n_h^* = n_h - 1$.

In the grouped balanced half-sample method (GBHS) of variance estimation, the sampled clusters in each stratum are randomly divided into two groups (halves) and the balanced repeated replication method is applied to the groups. Rao and Shao (1996) showed that this method is asymptotically incorrect in the sense that the associated t-pivotal does not converge in distribution to a standard normal distribution and that the associated confidence intervals are asymptotically incorrect. To overcome this difficulty they proposed independently repeating the grouping T times and then taking the average of the resulting T variance estimates. They showed the asymptotic correctness of such an estimator for a stratified random sampling design as $\min n_h \to \infty$ and $T \to \infty$. In a small simulation study they found that the method performs well for T as small as 15 in the case of smooth estimators. For a variance estimator of the population median, the RGBHS method performed better than the jackknife and GBHS in the sense that the RGBHS had a smaller relative bias and a smaller coefficient of variation. Recently, McCarthy (1993) discussed and compared a variety of procedures for variance estimation of the median based on simple random samples drawn from a finite population without replacement. His study includes most resampling procedures.

Although, the linearization methods useful for nonlinear statistics are difficult to implement for quantiles since density estimation is involved, Binder (1991), Binder and Kovačević (1995) and Kovačević and Binder (1997) obtained consistent estimators for the variance of some non-smooth measures of income inequality and polarization using the linearization method within the estimating equation framework. Estimators obtained using this method are computationally simpler than the resampling estimators but require theoretical derivation.

Variance estimation of the Gini Index has been studied by several authors under the assumption of simple random sampling, Glasser (1962), Sendler (1979), Sandström, Wretman and Waldén (1985) and Yitzhaki (1991). In the case of a complex design, Love and Wolfson (1976) proposed a 'crude half-sample replication' method. Sandström, Wretman and Waldén (1988) compared approximate variance techniques with the delete-one-unit jackknife for three sampling designs, two of which were complex.

Estimation of the variance of the Lorenz curve ordinates and the corresponding quantile shares has received less attention. The derivation of their asymptotic variances is quite complicated. There is the pioneering work of Beach and Davidson (1983) and Beach and Kaliski (1986). Their work is based on the superpopulation framework in which the survey weights are seen as constants in the construction of estimates. This approach, due to its model-based nature, may have its limitations in applications to data obtained from sample surveys where the sample design is deemed to be significant.

In the following subsections we review the variance estimation methods used in this study.

3.1 Delete-one-PSU Jackknife

This method is based on the sequential exclusion (deletion) of one PSU at a time from the computation of the estimate. After deletion, the weights of the remaining units in the sample are modified in such a manner that the deleted weights are compensated and that the CDF estimated from the remaining sample has the same properties of the original CDF. Let $\hat{F}_{(gj)}(y)$ denote the estimate of the CDF based on a sample without the gj-th PSU, that is

$$\hat{F}_{(gj)}(y) = \hat{G}_{(gj)}(y)/\hat{N}_{(gj)}$$

where

$$\hat{G}_{(gj)}(y) = \sum_{h\neq g}\sum_{c}\sum_{i} w_{hci} I\{y_{hci} \le y\} + \frac{n_g}{n_g - 1}\sum_{c\neq j}\sum_{i} w_{gci} I\{y_{gci} \le y\}$$

and

$$\hat{N}_{(gj)} = \sum_{h\neq g}\sum_{c}\sum_{i} w_{hci} + \frac{n_g}{n_g - 1}\sum_{c\neq j}\sum_{i} w_{gci}.$$

The 'delete-one-PSU' jackknife variance estimator of $\hat{F}(y)$ is

$$v_{J1}(\hat{F}(y)) = \sum_{g=1}^{L}\frac{n_g - 1}{n_g}\sum_{j=1}^{n_g}(\hat{F}_{(gj)}(y) - \hat{F}(y))^2.$$

Asymptotic consistency of $v_{J1}(\hat{F}(y))$ can be established using results from Krewski and Rao (1981).

For convenience, we note that all measures considered here can be written in the general form

$$\theta_N = \sum_U J(F_N, Y_i, \boldsymbol{\beta})\frac{1}{N},$$

where $J(\cdot)$ is a real-valued function possibly dependent on the nuisance parameter, $\boldsymbol{\beta}$. The finite population parameter θ_N is then estimated by

$$\hat{\theta} = \sum_s J(\hat{F}, y_{hci}, \hat{\boldsymbol{\beta}})\,\tilde{w}_{hci} \tag{3.1}$$

where $\hat{\boldsymbol{\beta}}$ denotes the estimated vector of nuisance parameters and $\tilde{w}_{hci}$ are the standardized weights. Using this general form, the estimate of an income inequality measure computed from the sample after omitting PSU gj, is

$$\hat{\theta}_{(gj)} = \sum_s J(\hat{F}_{(gj)}, y_{hci}, \hat{\boldsymbol{\beta}}_{(gj)})\,\tilde{w}_{hci(gj)}$$

where $\hat{F}_{(gj)}$ and $\hat{\boldsymbol{\beta}}_{(gj)}$ are the values of the distribution function and the nuisance parameter estimated from the sample with the gj-th PSU deleted and

$$\tilde{w}_{hci(gj)} = \begin{cases} w_{hci}/\hat{N}_{(gj)}, & \text{if } h \neq g, \\ \dfrac{n_g}{n_g - 1} w_{gci}/\hat{N}_{(gj)}, & \text{if } h = g, c \neq j, \\ 0, & \text{if } h = g,\ c = j. \end{cases}$$

The resulting 'delete-one-PSU' jackknife variance estimator of $\hat{\theta}$ is

$$v_{J1}(\hat{\theta}) = \sum_{g=1}^{L}\frac{n_g - 1}{n_g}\sum_{j=1}^{n_g}(\hat{\theta}_{(gj)} - \hat{\theta})^2. \tag{3.2}$$

If $\hat{\theta}$ is substituted by $\hat{\theta}_{\cdot\cdot} = \sum_g\sum_j \hat{\theta}_{(gj)}/n$ a variant of the jackknife variance estimate is obtained. We denote it by $v_{J2}(\hat{\theta})$. Obviously $v_{J2}(\hat{\theta}) \le v_{J1}(\hat{\theta})$. The consistency of (3.2) for smooth statistics has been established by Krewski and Rao (1981).

In the case of variance estimation for quantiles and functions of quantiles, we first compute the quantiles based on the sample with the gj-th PSU deleted,

$$\hat{\xi}_{(gj)}(p) = \inf\{y_{hci} \mid \hat{F}_{(gj)}(y_{hci}) \ge p, hci \in s\backslash(gj)\},$$

compute $\hat{\theta}_{(gj)} = g(\hat{\xi}_{(gj)})$ and then use equation (3.2) to obtain a jackknife variance estimator.

3.2 Grouped Balanced Half-Sample (GBHS) Method and Repeatedly Grouped Balanced Half-Sample (RGBHS) Method

Originally, the balanced half-sample method was proposed for the two clusters-per-stratum designs. The case that we are interested in is when there are more than 2 clusters per stratum. This situation is usually handled by grouping the clusters (primary stage units) in each stratum into two groups. We explore the idea given by Wu (1991) and simplify its application for the variance estimation of the CDF. First, in each stratum $h, (h = 1, ..., L)$, the PSU's are grouped at random

into two halves, h_1 and h_2, containing $m_{h_1} = [n_h/2]$ and $m_{h_2} = n_h - m_{h_1}$ PSU's, respectively. Setting the group indicator to

$$\delta_h^{(r)} = \begin{cases} 1, & h_1 \in r \\ -1, & h_2 \in r \end{cases}$$

where $r = 1, \ldots, R$ denotes a half-sample (replicate), the half-samples are balanced on the groups if $\sum_{r=1}^{R} \delta_h^{(r)} = 0$ and $\sum_{r=1}^{R} \delta_h^{(r)} \delta_{h'}^{(r)} = 0, (h \neq h')$. A minimal set of balanced half-samples can be obtained from a Hadamard matrix of order $R (L + 1 \leq R \leq L + 4)$.

The estimator of the distribution function based on the r-th half-sample is

$$\hat{F}^{(r)}(y) = \frac{\hat{G}^{(r)}(y)}{\hat{N}^{(r)}}$$

where

$$\hat{G}^{(r)}(y) = \sum_h \sum_c A_{hc}^{(r)} \sum_i w_{hci} I\{y_{hci} \leq y\}, \hat{N}^{(r)} = \sum_h \sum_c A_{hc}^{(r)} \sum_i w_{hci}$$

and $A_{hc}^{(r)}$ is the weight modifier and is constant for all clusters in the same half-sample. We assume that the weights of all units (households) in a cluster are rescaled equally by the modifier $A_{hc}^{(r)}$.

The standard GBHS method, when n_h is even, uses

$$A_{hc}^{(r)} = \begin{cases} 1 + \delta_h^{(r)}, & c \in h_1, \\ 1 - \delta_h^{(r)}, & c \in h_2 \end{cases} \tag{3.3}$$

which means that the weights are modified either by 2 or 0 depending on whether a unit is in the replicate or not. When n_h is odd, a number of different modifications have been considered (see Shao 1993 and Sitter 1993).

The method that we are using is based on the standard balanced replication resampling plan and a variant of the rescaling method proposed by Shao (1993):

$$A_{hc}^{(r)} = \begin{cases} 1 + (1 - a_h)\delta_h^{(r)}, & c \in h_1; \\ 1 - (1 - b_h)\delta_h^{(r)}, & c \in h_2. \end{cases}$$

The maintenance of the stratum sample size in any of the half-sample replicates means that

$$\sum_{c \in h1} [1 + (1 - a_h)\delta_h^{(r)}] + \sum_{c \in h2} [1 - (1 - b_h)\delta_h^{(r)}] = n_h,$$

which results in

$$A_{hc}^{(r)} = \begin{cases} 1 + (1 - a_h)\,\delta_h^{(r)}, & c \in h_1; \\ 1 - (1 - a_h)\dfrac{m_{h_1}}{m_{h_2}}\delta_h^{(r)}, & c \in h_2. \end{cases} \tag{3.4}$$

To ensure the non-negativity of the modified weights, a_h should satisfy $0 \leq a_h < 1$. When n_h is even we would like (3.4) to reduce to (3.3). Following Shao's idea (1993), we want the GBHS variance estimator to agree with a consistent estimator of the variance in the case of linear statistics. This leads to the following requirements for the stratum-specific perturbation factors $1 - a_h$:

For all h: (i) $0 < 1 - a_h \leq 1$; (ii) $(1 - a_h)^2 (m_{h_1}/m_{h_2})^2 \approx 1$; (iii) $(1 - a_h)^2 m_{h_1}/m_{h_2} \approx 1$. For the even n_h's we simply let $1 - a_h = 1$. However, keeping $1 - a_h = 1$ for odd n_h's would exclude any contribution from the clusters in the first half-sample when $\delta_h^{(r)} = -1$, see equation (3.4). For the purpose of the simulation study we chose

$$1 - a_h = \sqrt{\frac{n_h}{2\, m_{h2}}} \tag{3.5}$$

which reduces to 1 for an even n_h. In the case of an odd stratum sample size it is equal to $\sqrt{1 - 1/(n_h + 1)}$. In our simulation study very few strata have an odd n_h and we obtain $v_{GB1}(\hat{\mu}_Y) = v_{GB2}(\hat{\mu}_Y) \approx v_L(\hat{\mu}_Y)$ where $\hat{\mu}_Y$ is the sample mean and $v_L(\hat{\mu}_Y)$ is the commonly used linearization variance estimator. However, it is felt that more research is needed into modifying the GBHS method to handle many strata containing an odd number of PSU's.

As in the case of the jackknife method, the estimate of the income inequality measure computed from the r-th half-sample is $\hat{\theta}^{(r)} = \sum_s J(\hat{F}^{(r)}, y_{hci}, \hat{\beta}^{(r)}) \tilde{w}_{hci}^{(r)}$ where $\hat{\beta}^{(r)}$ is an estimate of the nuisance parameter based on the r-th half-sample and $\tilde{w}_{hci}^{(r)} = \tilde{w}_{hci} A_{hc}^{(r)}$. The resulting GBHS variance estimator of $\hat{\theta}$ is

$$v_{GB1}(\hat{\theta}) = \frac{1}{R} \sum_{r=1}^{R} (\hat{\theta}^{(r)} - \hat{\theta})^2. \tag{3.6}$$

By repeating the random grouping of units within each stratum T times, computing $v_{GB1}(\hat{\theta})$ each time and averaging over the T repetitions we obtain the Repeatedly Grouped Balanced Half Sample (RGBHS) variance estimator

$$v_{RG1}(\hat{\theta}) = \frac{1}{T} \sum_{t=1}^{T} v_{GB1_t}(\hat{\theta}).$$

A variant of the GBHS estimator (and RGBHS) is obtained by replacing $\hat{\theta}$ by $\hat{\theta} = \sum_r \hat{\theta}^{(r)}/R$, and will be denoted by $v_{GB2}(\hat{\theta})$ (and $v_{RG2}(\hat{\theta})$).

Needless to say that when weights are calibrated they have to be properly modified for each GBHS replication using the same balanced half sample procedure.

3.3 Bootstrap Method

We also investigated the performance of the bootstrap method for variance estimation of different income statistics. We adopted the bootstrap resampling scheme for the stratified cluster sample as given by Rao, Wu and Yue (1992). Briefly, draw a simple random sample of $n_h - 1$ clusters with replacement (from the n_h sample clusters) independently in

Table 1
Definition of u^*_{hci} Variates for the EE Approach

Measure	u^*_{hci}
Gini Index	$2[\hat{A}(y_{hci})y_{hci} + \hat{B}(y_{hci}) - \hat{\mu}(\hat{G}+1)/2]/\hat{\mu}$ where $A(y) = \hat{F}(y) - \frac{\hat{G}+1}{2}$ and $B(y) = \sum_s w_{hcj} y_{hcj} I\{y_{hcj} \geq y\}$.
Lorenz Curve	$[(y_{hci} - \hat{\xi}_p) I\{y_{hci} \leq \hat{\xi}_p\} + p\hat{\xi}_p - y_{hci}\hat{L}(p)]/\hat{\mu}$
Quantile Share	$\frac{1}{\hat{\mu}}[(y_{hci} - \hat{\xi}_{p_2}) I\{y_{hci} \leq \hat{\xi}_{p_2}\} - (y_{hci} - \hat{\xi}_{p_1}) I\{y_{hci} \leq \hat{\xi}_{p_1}\} + p_2\hat{\xi}_{p_2} - p_1\hat{\xi}_{p_1} - y_{hci}\hat{Q}(p_1, p_2)]$
Quantile	$-[I\{y \leq \hat{\xi}_p\} - p]/\hat{f}(\hat{\xi}_p)$, $\hat{f}(.)$ is the finite population density estimator
Low Income Proportion	$-\frac{\hat{f}(\hat{\xi}_{0.5}/2)}{2\hat{f}(\hat{\xi}_{0.5})}[I\{y_{hci} \leq \hat{\xi}_{0.5}\} - 1/2] + [I\{y_{hci} \leq \hat{\xi}_{0.5}/2\} - \hat{\Lambda}_{0.5}]$
Polarization Index	$\frac{2}{\hat{\xi}_{0.5}}[(\hat{\xi}_{0.5} - y_{hci})(I\{y_i \leq \hat{\xi}_{0.5}\} - 0.5) - (A(y_{hci})y_{hci} + B(y_{hci}) - (\hat{G}+1)\hat{\xi}_{0.5}/2 + \hat{G}y_{hci}/2)] + \frac{P\hat{I}}{\hat{\xi}_{0.5}\hat{f}(\hat{\xi}_{0.5})}(I\{y_{hci} \leq \hat{\xi}_{0.5}\} - 0.5) - P\hat{I}$

each stratum. The bootstrap weight, w^*_{hci}, is obtained by modifying the original weight w_{hci} as follows:

$$w^*_{hci} = A_{hc} w_{hci}$$

where

$$A_{hc} = \frac{n_h}{n_h - 1} m^*_{hc}$$

and m^*_{hc} is the number of times the hc-th cluster is selected. Note that $\sum_c m^*_{hc} = n_h - 1$. This procedure is repeated independently B times; for each bootstrap sample, we calculate $\hat{\theta}^* = \sum_s J(\hat{F}^*, y_{hci}, \hat{\beta}^*)\bar{w}^*_{hci}$ where $\hat{\beta}^*$ is an estimate of the nuisance parameter based on the bootstrap sample and $\bar{w}^*_{hci} = w^*_{hci}/\sum_s w^*_{hci}$. The bootstrap estimate of the variance of $\hat{\theta}$ is then given by

$$v_{B1}(\hat{\theta}) = \frac{1}{B}\sum_{b=1}^{B}(\hat{\theta}^*_{(b)} - \hat{\theta})^2.$$

Another variance estimate is obtained by substituting $\hat{\theta}$ with the mean of bootstrap replicates.

3.4 Linearization via the Estimating Equations Approach

The estimating equations (EE) approach of Binder (Binder 1991, Binder and Patak 1994, Binder and Kovačević 1995), unlike the resampling methods, is not computationally intensive. This method, based on linearization, provides formulae for asymptotic variances which are easy to program despite their complicated appearance.

Applying the EE methodology as given in Binder and Patak (1994), Binder and Kovačević (1995) and Kovačević and Binder (1997) one obtains expressions for the approximate variance estimators of the studied measures as

$$v_{EE} = \sum_h \frac{n_h}{n_h - 1}\sum_c \left(u^*_{hc} - \bar{u}^*_h\right)^2 \qquad (3.7)$$

where $u^*_{hc} = \sum_i \bar{w}_{hci} u^*_{hci}$, $\bar{u}^*_h = \sum_c u^*_{hc}/n_h$, and $\bar{w}_{hci}$ is a normalized weight. For more on the EE approach, in particular the relationship between the u^*_{hci} variates and the J function, we refer the reader to Binder and Kovačević (1995). The u^*_{hci} variates for the considered measures are given in Table 1.

The expressions for the u^*_{hci} variates for the low income proportion and polarization index depend on the estimate of the density function at the median, $\hat{f}(\hat{\xi}_{0.5})$, and half of the median, $\hat{f}(\hat{\xi}_{0.5}/2)$. An appropriate method for estimating these quantities is given in Binder and Kovačević (1995).

4. SIMULATION STUDY

4.1 Data and the Design of the Simulation Study

The Ontario sample from the 1988 Canadian Survey of Consumer Finance (SCF) was used as the underlying population of the study. The SCF is an annual supplement to the monthly Canadian Labour Force Survey. The population contained 7474 households in 525 PSU's from 40 strata. Originally, the Ontario sample was taken from 91 strata which we collapsed to form sufficiently large strata. For each household a nonnegative value of the total annual income was available. The distribution of the income on this micro population was highly skewed to the right with coefficients of skewness and kurtosis obtained as 4.5 and 89.5, respectively. The true values of the parameters of interest (measures of income inequality and polarization) were computed from this population. Neyman allocation was used to assign 108 sample clusters (PSU's) to the 40 strata. A one-stage cluster design with the strata samples sizes between 2 and 6 clusters, selected with probability proportional to size and with replacement was used. In a selected cluster all households (6 to 20) were enumerated.

We considered the following measures in the study: Gini Index, Low Income Proportion, Polarization Index, a set of

Quantile Shares, a set of Lorenz Curve Ordinates and the corresponding quantiles. The MSE's of the estimates of these measures were approximated by the empirical mean squared error (EMSE), computed over 10,000 independent samples drawn by the design explained above. These EMSE's were used as 'true' MSE's for comparison with the estimated variances.

From each of the 10,000 samples, along with the estimates of the parameters, we computed estimates of the sampling variances using the following methods: the delete-one-PSU jackknife (JK), the grouped balanced half-sample (GBHS) and the repeatedly grouped balanced half-sample (RGBHS), the bootstrap (BS) and the linearization method via estimating equations (EE). For all resampling methods two different estimators were used, one using the 'full sample' estimate and another one using the mean over all replicates. The jackknife variance estimators were based on 108 jackknife replicates while the bootstrap method was based on 100 replicates. The GBHS and RGBHS were based on 44 balanced replicates obtained from a 44 by 44 Hadamard matrix and 3 repetitions for RGBHS, totalling 132 half-sample replicates for this method. Note that the number of jackknife replicates is non-arbitrary and is determined by the number of clusters in the sample. Similarly, the number of GBHS replicates is determined by the number of strata. In order to make the number of replicates comparable over all methods, we decided to have 100 ($\approx$ 108) bootstrap replicates and 3 repetitions of the GBHS resulting in 132 replicates for RGBHS.

In order to evaluate the accuracy and the precision of the considered methods we computed their relative biases and relative variance (instability) over the $A = 10{,}000$ simulations:

$$\text{rel.bias}(v_M) = \frac{\sum_a v_M(a)/A - \text{EMSE}}{\text{EMSE}}$$

$$\text{rel.var.}(v_M) = \frac{\sqrt{\sum_a [v_M(a) - \text{EMSE}]^2/A}}{\text{EMSE}}$$

To evaluate the effectiveness of normal-theory confidence intervals, empirical coverage rates were computed for nominal confidence coefficients of $100(1 - \alpha)\% = 90, 95$ and 99 percent,

$$\text{cov. prob.}(v_M) = \frac{\sum_a I\{|\hat{\theta}_a - \theta| \Big/ \sqrt{v_M(a)} \le z_{\alpha/2}\}}{A}$$

where $z_{\alpha/2}$ is the upper $\alpha/2$-th standard normal percental. Upper and lower tailed error rates were also calculated as follows,

$$\text{err\_}L(v_M) = \frac{\sum_a I\{(\hat{\theta}_a - \theta) \Big/ \sqrt{v_M(a)} < -z_{\alpha/2}\}}{A}$$

$$\text{err\_}U(v_M) = \frac{\sum_a I\{(\hat{\theta}_a - \theta) \Big/ \sqrt{v_M(a)} > z_{\alpha/2}\}}{A}.$$

The large set of results obtained from the simulation study are summarized separately for each income inequality measure.

4.2 Summary of Findings

Gini Index

Concerning the accuracy of the variance estimators for the Gini index, all methods performed similarly, with very small negative relative biases ranging between -2.2 and -0.6 percent. Of all the estimators, the RGBHS estimators had the smallest relative bias.

All estimators were found to be of approximately the same stability, in the range of 87-99%. The grouped balanced half-sample methods (GBHS and RGBHS) perform slightly worse than other methods.

The coverage probabilities for the 95% confidence intervals were in the range of 92.6 (for GBHS) to 93.9 (for RGBHS). The lower tail error rates were understated by the nominal 2.5% rate for all methods considered. We found that the lower tails were more than 100% heavier than the nominal 2.5%, ranging between 4.6 and 5.4%. The upper tail error rates were overstated by the nominal rate for all methods. (See Table 2). We also computed the coverage rates for the 90% and 99% confidence intervals and they were in the range of 87.2 (for GBHS) to 88.5 (for RGBHS) and in the range of 97.7 (for GBHS) to 98.5 (for RGBHS), respectively. Similarly, the tail rates for the nominal 5% and 1% followed the pattern of 2.5%.

Overall, for variance estimation of the Gini index it is difficult to say which method is the best since all compared methods performed similarly. There is a slight trade off between accuracy and stability in the case of the balanced half-sample methods which give the most accurate estimates of the variance but at the same time the least stable. The empirical coverage probabilities for all of the estimators are also very similar. The realized values of the tail error rates suggest that the use of asymmetric confidence intervals is more appropriate.

Low Income Proportion (LIP)

All methods considered tended to overestimate the variance of the LIP. However, the difference in the magnitude of overestimation was large, and ranged between 1.1% for the EE and 76.9% for the JK1. The best performer among resampling methods was the bootstrap, where the relative bias for the BS1 estimator was 8.9% and for BS2 3.8%.

The jackknife estimate of the variance of the LIP was very unstable. The GBHS estimators also had increased instability. The bootstrap and EE estimators performed similarly with relative variation between 31 and 45%.

Table 2
Values of the Evaluation Statistics for the Variance Estimators of the Gini Index

		Jackknife		GBHS		RGBHS		Bootstrap		Estimating Equations
		v_{J1}	v_{J2}	v_{GB1}	v_{GB2}	v_{RG1}	v_{RG2}	v_{B1}	v_{B2}	v_{EE}
Relative Bias (%)		-1.3	-1.3	-0.9	-1.1	-0.6	-0.7	-1.2	-2.2	-1.5
Relative Variation (%)		87.1	87.1	99.4	99.2	95.2	95.1	88.5	87.6	87.0
Coverage Probability (95%)		93.8	93.8	92.6	92.6	93.9	93.9	93.5	93.4	93.7
Tail Error Rates (2.5%)	L	4.8	4.8	5.4	5.4	4.6	4.6	5.0	5.1	4.9
	U	1.4	1.4	2.0	2.0	1.5	1.5	1.5	1.5	1.4

Table 3
Values of the Evaluation Statistics for the Variance Estimators of the Low Income Proportion

		Jackknife		GBHS		RGBHS		Bootstrap		Estimating Equations
		v_{J1}	v_{J2}	v_{GB1}	v_{GB2}	v_{RG1}	v_{RG2}	v_{B1}	v_{B2}	v_{EE}
Relative Bias (%)		76.9	58.4	25.8	21.0	26.8	21.9	8.9	3.8	1.1
Relative Stability (%)		113.1	81.0	62.5	61.0	40.8	39.5	35.1	33.5	31.0
Coverage Probability (95%)		97.4	96.9	94.6	94.1	96.2	95.7	93.9	93.3	93.2
Tail Error Rates (2.5%)	L	2.1	2.6	3.3	3.5	2.4	2.6	4.6	5.0	5.0
	U	0.5	0.6	2.0	2.4	1.4	1.7	1.5	1.7	1.7

The 95% confidence interval for the LIP based on the JK variance estimates had higher than nominal coverage rates, 97.4 and 96.9%, consequences of the overestimation of the variance. The other methods had slightly lower coverage rates than nominal. The tail error rates showed that all methods resulted in heavier lower tails, indicating a skewed distribution of the LIP with a long tail to the right. For the cases of 90% and 99% confidence intervals we obtained exactly the same pattern for the coverage and the tail error rates.

Overall, for variance estimation of the LIP, the bootstrap and the EE method show supremacy over the other methods considered.

Polarization Index

The evaluation statistics for the variance estimators of the polarization index showed a high level of agreement in performance with variance estimation for the low income proportion. Again, the bootstrap and EE method were the best.

Table 4
Values of the Evaluation Statistics for the Variance Estimators of the Polarization Index

		Jackknife		GBHS		RGBHS		Bootstrap		Estimating Equations
		v_{J1}	v_{J2}	v_{GB1}	v_{GB2}	v_{RG1}	v_{RG2}	v_{B1}	v_{B2}	v_{EE}
Relative Bias (%)		95.4	56.5	13.9	11.2	14.7	12.1	6.0	2.9	4.2
Relative Stability (%)		138.7	78.5	77.5	75.9	60.0	58.6	48.4	47.0	50.0
Coverage Probability (95%)		98.6	98.0	94.2	93.8	95.4	95.2	95.0	94.7	94.4
Tail Error Rates (2.5%)	L	0.7	0.8	2.2	2.4	1.4	1.4	1.8	2.0	2.0
	U	0.8	1.1	3.6	3.9	3.2	3.4	3.2	3.4	3.6

Lorenz Curve Ordinates and Quantile Shares

The full results for the Lorenz Curve Ordinates and Quantile Shares are given in Kovačević, Yung and Pandher (1995). We present here a graphical summary of the results in Figures 1a-1c. The jackknife method (both estimators) significantly overestimates the variances of all considered Lorenz Curve Ordinates (LCO) and Quantile Shares (QS). The relative bias of the JK1 estimator for the LCO ranged between 15 and 45% and between 9 and 27% for the JK2 estimator. The relative bias was smaller in the middle of the interval ($0 \le p \le 1$) and almost three times larger at the tails (for small and large values of p). The relative bias of the JK1 estimator was about 50% larger than the relative bias of the JK2 estimator for the LCO. The difference can be attributed to the significant difference between the full sample estimate of the LCO and the average taken over jackknife replicates.

Similar findings held for the performance of the JK variance estimators for QS's which overestimated the variance between 26-237%, depending on the population share. The largest overestimation appeared in the middle. Again, the JK1 was larger than JK2 by about 75%.

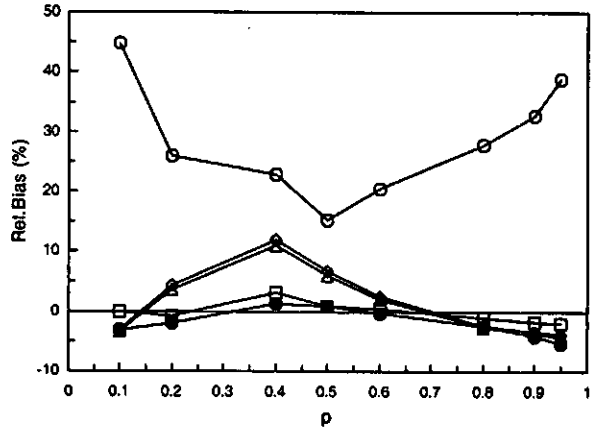
The magnitude of the relative bias was very small for the other two methods. However, there was no clear pattern about the direction of bias – sometimes it was positive, but often it was negative. The bootstrap estimators and the EE estimator outperformed the other methods, especially around the LCO corresponding to $p = 0.5$ (see Figure 2a). For clarity of the graphical presentation the JK methods are not shown in Figures 2a and 2b.

The variance of the QS's is estimated similarly. The bootstrap and EE provided the most accurate estimates of the variances of LCO and QS. For the LCO the relative bias ranged between -2 and +3% for bootstrap and -5 to +1% for EE. At the same time, for the QS, the bootstrap estimates had relative biases between -3 and +8% and EE estimates between -3 and +5%.

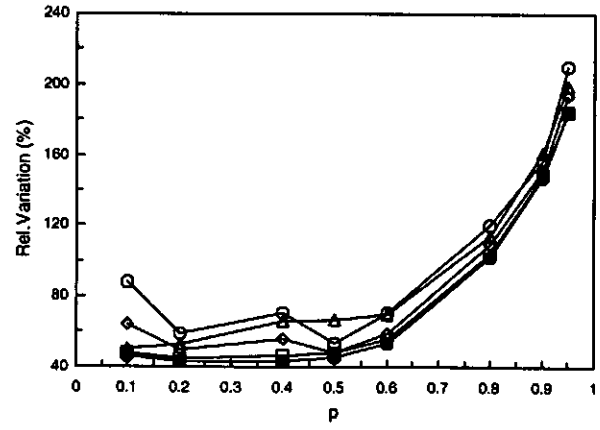
Concerning the stability of the different variance estimators we found that all methods perform similarly with a slight advantage for the EE method. Also, there is an obvious direct dependence of the relative variation measure and the value of p.

When we compared the methods according to the coverage properties of the variance estimators for the LCO and QS we found that for the nominal 95% confidence interval, the JK method gave empirical coverage rates between 94.5 and 96.5% for the LCO and 94.5 to 99% for the QS. Other methods performed similarly with coverage rates between 88 and 94%. Better coverage was found for the LCO and QS with smaller value of p (see Figure1c). In contrast to findings for the Gini index, the lower tail error rates were about twice the upper tail error rates for all methods and for both LCO and QS. A similar pattern was observed for 90% and 99% confidence intervals.

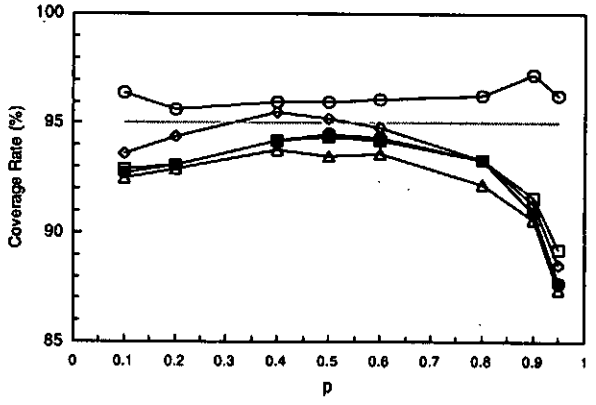
Our empirical findings suggest that the jackknife method is not a good choice for the variance estimation of the LCO and QS especially for small and large values of p. Much better alternatives are the GBHS or the RGBHS. However, the best choice is either the EE method or the bootstrap.



a) Relative Bias



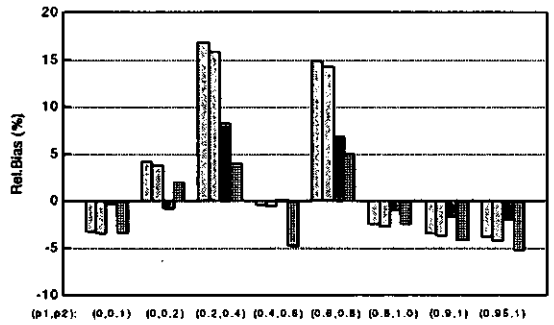
b) Relative Variation



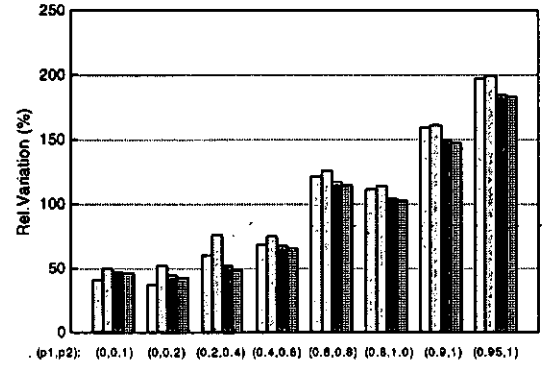
c) Coverage Rate (for Nominal 95%)

o-JK1 ◇-GBHS1 △-RGBHS1 □-BS1 ●-EE

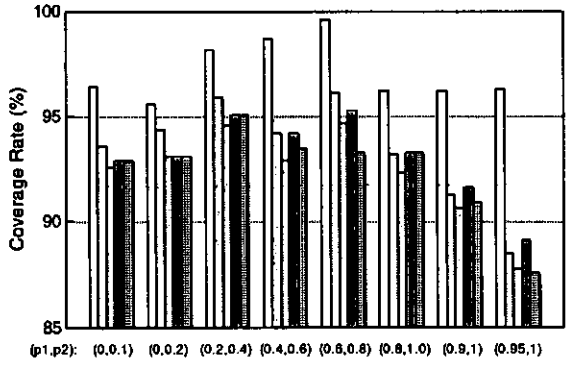
Figure 1. Properties of the Variance Estimators of Lorenz Curve Ordinates



a) Relative Bias (JK methods are not shown)



b) Relative Variation (JK methods are not shown)

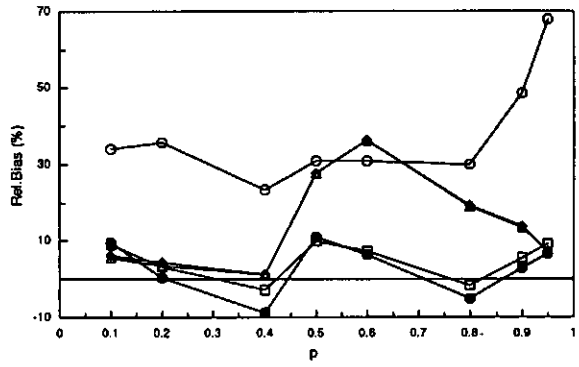


c) Coverage Rate (for nominal 95%)

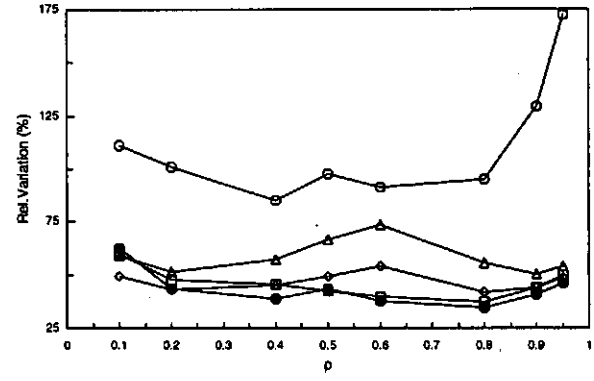
Figure 2. Properties of the Variance Estimators of Quantile Shares

Quantiles

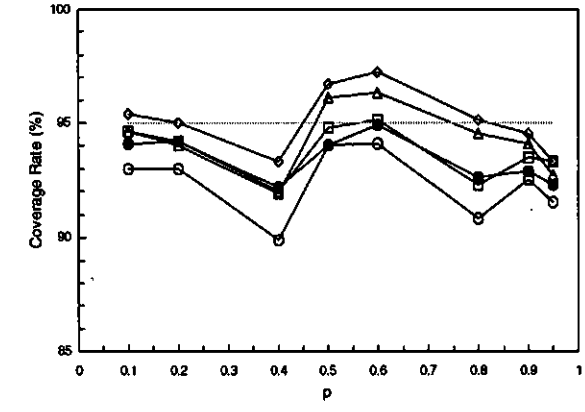
The full results obtained for the quantiles are presented in Kovačević, Yung and Pandher (1995) and are summarized graphically here. The relative bias of the JK1 estimate of the variance for the quantiles was between 23 and 67% and for JK2 between 17 and 52%. The largest overestimation occurred for the variances of $\hat{\xi}_{0.90}$ and $\hat{\xi}_{0.95}$. The RGBHS and GBHS show quite a different picture. The variance of the median was overestimated by 27% but the variances of tail quantiles were obtained very accurately, with the relative bias between 3 and 7%. Other methods also performed much better for the tail quantiles and moderately better for the median and quantiles around it. In particular, the bootstrap and the EE method produced estimates with the smallest relative biases, although without clear pattern about the direction of the bias. For the bootstrap estimators, the relative bias was in the interval (-5%,+9%), and for EE (-8%, +9%) (see Figure 3a).



a) Relative Bias



b) Relative Variation



c) Coverage Rate (for nominal 95%)

○-JK1 ◇-GBHS1 △-RGBHS1 □-BS1 ●-EE

Figure 3. Properties of the Variance Estimators of Quantiles

Table 5
Rankings of methods by relative bias, relative stability and empirical coverage probability

	Jackknife	GBHS	RGBHS	Bootstrap	EE (Taylor)	Best methods
Gini Index	All procedures performed similarly					–
Quantiles	5, 5, 5	3, 4, 4	4, 3, 1	1, 2, 3	2, 1, 2	EE, BS
Lorenz Curve	5, 5, 5	3, 4, 4	4, 3, 1	1, 2, 3	2, 1, 2	EE, BS
Quantile Shares	5, 5, 5	3, 4, 4	4, 3, 1	1, 2, 2	2, 1, 3	BS, EE
Low Income	5, 5, 5	3, 4, 2	4, 3, 1	2, 2 ,3	1, 1, 4	EE, BS
Polarization Index	5, 5, 5	3, 4, 4	4, 3, 2	2, 1, 1	1, 2, 3	BS, EE

The jackknife estimators were the least stable. The RGBHS, bootstrap and EE showed similar stability which, on average over all quantiles, was about three times higher than the stability of JK estimators. The highest stability was attained around the median (see Figure 3b).

In general, the coverage probabilities for the quantiles were less than nominal for all of the methods considered, with some exceptions for the GBHS and RGBHS methods (see Figure 3c). When we compared the observed tail error rates, it seemed that all methods exhibited similar behaviour, for the lower quantiles ($p = 0.1, 0.2$) the upper (right) tails were heavier; for others it was opposite, the lower tails were heavier. Similar results were obtained for the 90% and 99% confidence intervals.

The findings from this empirical study confirm that for variance estimation of quantiles, the jackknife method should be avoided. For the variance of the median, in particular, the best choice seems to be either the EE or the bootstrap. For other quantiles the RGBHS showed very good performance as well.

We condense our findings in Table 5 where the relative bias, relative variation and the coverage probabilities for the methods considered were ranked from 1 to 5 (1 = the best). For the resampling methods we averaged the values over both estimators. For the quantiles, LCO and QS we averaged the values over all p's. The last column contains the choice of the two best performing methods.

5. DISCUSSION AND CONCLUSION

The linearization method via EE has shown the best overall performance, the smallest relative bias, the smallest relative variation and relatively good coverage properties. Next to the EE method is the bootstrap method, as the best resampling method considered. The RGBHS and GBHS method performed comparably well for the Lorenz Curve ordinates, quantile shares and some of the quantiles, in the sense of the small relative bias and relative stability comparable with the bootstrap method. The jackknife method has performed poorly for all measures except the Gini index.

It is well known that the jackknife variance estimator performs poorly for non-smooth functions. The smoothness of the J function defined in (3.1) is an essential determinant of the asymptotic properties of its variance estimator. Classifying our measures as smooth or non-smooth on the basis of the J functions, we see that the only smooth estimator considered here was the Gini index. Not surprisingly, the Gini index was the only measure for which the jackknife performed well. However, when considering the jackknife variance estimator, care must be taken to ensure that the assumptions under which the jackknife is valid are fulfilled.

If the goal is to provide one method for variance estimation for the large list of different income statistics, our empirical study has shown that the bootstrap is the best resampling choice, and that the linearization via the estimating equations approach is the best computationally non-intensive method, which however, requires some preparatory algebraic work, different for each measure.

It should be emphasized that the empirical study was based on an one-stage cluster sampling design, with the clusters selected proportionally to their size, so the intracluster variability was not accounted for. Some other limited studies have shown similar behaviour of these methods in the case of two stage sampling plans (see Binder and Kovačević 1995, and Kovačević and Binder 1997).

ACKNOWLEDGMENTS

The authors would like to thank G.S. Pandher for his fruitful participation at the beginning of the project, J. Gambino for his thorough reading of an earlier version of the paper, H. Mantel, associate editor, anonymous referees and the editor, for valuable comments that significantly improved quality of the paper.

REFERENCES

BABU, G.J. (1986). A note on bootstrapping the variance of sample quantile. *Annals of the Institute of Statistical Mathematics*,38-A, 439-443.

BEACH, C.M., and DAVIDSON, R. (1983). Distribution-free statistical inference with Lorenz curves and income shares. *Review of Economic Studies*, 50, 723-735.

BEACH, C.M., and KALISKI, S.F. (1986). Lorenz curve inference with sample weights: an application to the distribution of unemployment experience. *Applied Statistics*, 35, 38-45.

BINDER, D.A. (1991). Use of estimating functions for interval estimation from complex surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 34-42.

BINDER, D.A., and KOVAČEVIĆ, M.S. (1995). Estimating some measures of income inequality from survey data: an application of the estimating equation approach. *Survey Methodology*, 21, 137-145.

BINDER, D.A., and PATAK, Z. (1994). Use of estimating functions for interval estimation from complex surveys. *Journal of the American Statistical Association*, 89, 1035-1043.

EFRON, B. (1979). Bootstrap method: another look at the jackknife. *Annals of Statistics* 7, 1-26.

FOSTER, J.E., and WOLFSON, M.C. (1992). Polarization and the decline of the middle class: Canada and the U.S. (Manuscript).

FRANCISCO, C.A., and FULLER, W.A. (1991). Quantile estimation with a complex survey design. *Annals of Statistics*, 19, 454-469.

GLASSER, G.J. (1962). Variance formulas for the mean difference and coefficient of concentration. *Journal of the American Statistical Association*, 57, 648-654.

KAKWANI, N.C. (1980). *Income Inequality and Poverty*. Washington, D.C.: World Bank.

KOVAČEVIĆ, M.S., and BINDER, D.A. (1997). Variance estimation for measures of income inequality and polarization-the estimating equations approach. (To appear in *Journal of Official Statistics*).

KOVAČEVIĆ, M.S., YUNG, W., and PANDHER, G.S. (1995). Estimating the Sampling Variances of Income Inequality and Polarization – An Empirical Study. Methodology Branch Working Paper, HSMD-95-007-E. Statistics Canada.

KOVAR, J.G. (1987). Variance Estimation of Medians in Stratified Samples. Methodology branch working paper, BSMD-87-004-E. Statistics Canada.

KOVAR, J.G., RAO, J.N.K., and WU, C.F.J. (1988). Bootstrap and other methods to measure errors in survey estimates. *The Canadian Journal of Statistics* 16, 25-45.

KREWSKI, D., and RAO, J.N.K. (1981). Inference from stratified samples: Properties of the linearization, jackknife and balanced repeated replication methods. *Annals of Statistics*, 9, 1010-1019.

LOVE, R., and WOLFSON, M.C. (1976). Income inequality: statistical methodology and Canadian illustrations. Ottawa, Statistics Canada.

McCARTHY, P.J. (1993). Standard error and confidence interval estimation for the median. *Journal of Official Statistics*, 9, 673-689.

NYGÄRD, F., and SANDSTRÖM, A. (1981). *Measuring Income Inequality*. Stockholm: Almqvist & Wiksell International.

NYGÄRD, F., and SANDSTRÖM, A. (1985). The estimation of the Gini and the entropy inequality parameters in finite populations. *Journal of Official Statistics*, 1, 399-412.

RAO, J.N.K., and SHAO, J. (1996). On balanced half-sample variance estimation in stratified random sampling. *Journal of the American Statistical Association*, 91, 343-348.

RAO, J.N.K., and WU, C.F.J. (1987). Methods for standard errors and confidence intervals from survey data: Some recent work. *Proceedings of the 46th Session of International Statistical Institute*, 3, 5-19.

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., WU, C.F.J., and YUE, K. (1992). Some recent work on resampling methods for complex surveys. *Survey Methodology*, 18, 209-217.

SANDSTRÖM, A., WRETMAN, J.H., and WALDÉN, B. (1985). Variance estimators of the Gini coefficient, simple random sampling. *Metron*, 43, 41-70.

SANDSTRÖM, A., WRETMAN, J.H., and WALDÉN, B. (1988). Variance estimators of the Gini coefficient – probability sampling. *Journal of Business and Economic Statistics*, 6, 113-119.

SEN, A.K. (1973). *On Economic Inequality*. London: Oxford University Press.

SENDLER, W. (1979). On statistical inference in concentration measurement. *Metrika*, 26, 119-122.

SHAO, J. (1993). Balanced repeated replication. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 544-549.

SHAO, J., and WU, C.F.J. (1989). A general theory for jackknife variance estimation. *Annals of Statistics*, 17, 1176-1197.

SITTER, R.R. (1993). Balanced repeated replications based on orthogonal multi-arrays. *Biometrika*, 80, 211-221.

WOLFSON, M.C. (1994). When inequalities diverge. *American Economic Review*, 84, 353-358.

WOODRUFF, R.S. (1952). Confidence intervals for medians and other position measures. *Journal of the American Statistical Association*, 47, 635-646.

WU, C.F.J. (1991). Balanced repeated replications based on mixed orthogonal arrays. *Biometrika*, 78, 181-188.

YITZHATI, S. (1991). Calculating jackknife variance estimators for parameters of the Gini method. *Journal of Business and Economic Statistics*, 9, 235-239.

Survey Methodology, June 1997
Vol. 23, No. 1, pp. 53-60
Statistics Canada

Instrumental Variable Estimation of Gross Flows in the Presence of Measurement Error

K. HUMPHREYS and C. J. SKINNER[1]

ABSTRACT

The problem of estimating transition rates from longitudinal survey data in the presence of misclassification error is considered. Approaches which use external information on misclassification rates are reviewed, together with alternative models for measurement error. We define categorical instrumental variables and propose methods for the identification and estimation of models including such variables by viewing the model as a restricted latent class model. The numerical properties of the implied instrumental variable estimators of flow rates are studied using data from the Panel Study of Income Dynamics.

KEY WORDS: Latent class; Longitudinal; Misclassification; Transition rate.

1. INTRODUCTION

One of the major benefits of longitudinal surveys is that they permit the estimation of gross flows, for example flows out of unemployment into employment (see *e.g.*, Hogue and Flaim 1986). A key problem when estimating flows is the bias induced by measurement error. For the estimation of cross-sectional proportions, misclassification into and out of states may tend to cancel out (Chua and Fuller 1987). Such compensation tends not to occur, however, when estimating longitudinal flows.

The first response to the problem of measurement error should clearly be to attempt to reduce the error in the survey measurement procedures. Relevant approaches are discussed by Biemer, Groves, Lyberg, Mathiowetz and Sudman (1991), but will not be considered here. Even with the "best" survey procedures, however, some measurement error will inevitably arise and there will remain a need to compensate for the effect of error in the survey analysis.

Methods for compensating for measurement error are generally based on some assumed model of the error process. Some models which have been proposed in the literature will be referred to in Section 2. In order to identify and estimate these models it is generally necessary to use additional auxiliary information, such as provided by reinterview studies (*e.g.*, Meyer 1988). Since reinterview studies are costly, however, and since in practice their aim is often not to estimate the characteristics of the measurement error distribution (Forsman and Schreiner 1991), there remains a need for alternative procedures which may be used when no reinterview data is available. For measurement error on continuous variables, a common approach employed in the absence of auxiliary information about the measurement error distribution is the method of instrumental variable estimation (*e.g.*, Fuller 1987, Sect. 1.4). An instrumental variable is a variable included in the survey dataset which is related to the true variable measured with error but is uncorrelated with the measurement error. These and associated assumptions supply information which replaces that provided by reinterview studies and enables parameters of the model involving the true variable to be identified and estimated. The aim of this paper is to investigate how the instrumental variable estimation method may be adapted to estimate flows among discrete states. We find that latent class models (*e.g.*, Bartholomew 1987, Ch. 2) provide a general framework within which the assumptions about the instrumental variable correspond to certain restrictions on the model parameters. Our approach is thus related to other approaches which impose restrictions on latent class models (*e.g.*, van de Pol and de Leeuw 1986; van de Pol and Langeheine 1990).

2. MODELS

We consider only the case of two occasions $t = 1$ and $t = 2$. Let the number of states into which each individual can be classified at each occasion be r. Denote the classified states at $t = 1$ and $t = 2$ by X and Y respectively and the corresponding true states by x and y. We assume a model in which the vectors of values of (X, Y, x, y) are generated as independent outcomes of a common random vector with distribution $\text{pr}(X = i, Y = j, x = u, y = v)$.

The first assumption about this distribution, made by a number of authors (*e.g.*, Abowd and Zellner 1985; Poterba and Summers 1986 and Chua and Fuller 1987) and which we shall also make, is that the classification errors on the two occasions are conditionally independent given the true states, that is

$$\text{pr}(X = i, Y = j \mid x = u, y = v) = \text{pr}(X = i \mid x = u, y = v)\,\text{pr}(Y = i \mid x = u, y = v). \quad \text{(A1)}$$

[1] K. Humphreys, Department of Psychology, Stockholm University, S-106 91 Stockholm, Sweden; C.J. Skinner, Department of Social Statistics, University of Southampton, Southampton, SO17 1BJ, United Kingdom.

Such an assumption is common in general latent variable models (*e.g.*, Anderson 1959). It seems a reasonable initial assumption when the survey measurement procedures are independent on the two occasions. On the other hand, if X is obtained retrospectively from the same interview in which Y is measured then it seems likely that the tendency for respondents to give over-consistent responses in a single interview may tend to induce positive association between classification errors. See, for example, Marquis and Moore (1990) on evidence from the Survey of Income and Program Participation. A further reason for doubting the conditional independence assumption is the possibility of individual heterogeneity in misclassification probabilities, for example some respondents may be more reliable than others. See Skinner and Torelli (1993) and Singh and Rao (1995). In Section 4 we shall allow for heterogeneity by assuming only that the model holds within cells of a cross-classification of observed variables.

Our next basic assumption is that classification error only depends on current true state so that

$$\text{pr}(X = i \mid x = u, y = v) = \text{pr}(X = i \mid x = u) = K_{xiu}, \text{ say,}$$

$$\text{pr}(Y = j \mid x = u, y = v) = \text{pr}(Y = j \mid y = v) = K_{yjv}, \text{ say.} \quad \text{(A2)}$$

The K_{xiu} and K_{yjv} define $r \times r$ misclassification matrices $\boldsymbol{K}_x = [K_{xiu}]$ and $\boldsymbol{K}_y = [K_{yjv}]$. Letting $\boldsymbol{P}$ denote the $r \times r$ matrix with ij-th element $\text{pr}(X = i, Y = j)$ and Π the $r \times r$ matrix with uv-th element $\text{pr}(x = u, y = v)$ we have the matrix equation

$$\boldsymbol{P} = \boldsymbol{K}_x \Pi \boldsymbol{K}_y'. \quad (1)$$

The matrix Π contains the parameters of interest, whereas it is the matrix $\boldsymbol{P}$ which may be estimated consistently from sample X and Y values. If auxiliary estimates of $\boldsymbol{K}_x$ and $\boldsymbol{K}_y$ are available and these are non-singular then we can solve equation (1) to obtain estimates of Π. If it is possible to ascertain the true states in reinterview studies then $\boldsymbol{K}_x$ and $\boldsymbol{K}_y$ may be estimated directly (Abowd and Zellner 1985). On the other hand, if the reinterview study only provides independent reclassifications then it is only possible to estimate the interview-reinterview matrices

$$\boldsymbol{K}_x \Delta_x \boldsymbol{K}_x' \text{ and } \boldsymbol{K}_y \Delta_y \boldsymbol{K}_y'$$

where $\Delta_x = \text{diag}[\text{pr}(x = u)]$, $\Delta_y = \text{diag}[\text{pr}(y = v)]$ (Chua and Fuller 1987). Each interview-reinterview matrix is symmetric with elements summing to one and so only contains $r(r+1)/2 - 1$ "independent" items of information. Since each column of each $\boldsymbol{K}$ matrix and the diagonal of each Δ matrix sum to one, the number of unknown parameters on each occasion is $r(r-1) + r - 1 = r^2 - 1$. The excess of parameters over items of information is therefore $r^2 - 1 - r(r+1)/2 + 1 = r(r-1)/2$ at each occasion and so the model is underidentified for $r \geq 2$. Chua and Fuller (1987) suggest that a natural extra assumption to make to help achieve identification is to suppose that the measurement errors are unbiased on each occasion in the sense that

$$\text{pr}(x = i) = \text{pr}(X = i), \ \text{pr}(y = i) = \text{pr}(Y = i) \ \ i = 1, \ldots, r. \quad (2)$$

In this case false positives and false negatives tend to compensate for each other in cross-sectional estimates of proportions. This assumption reduces the number of parameters by $r - 1$ on each occasion. Even under this assumption the model remains underidentified for $r \geq 3$ and Chua and Fuller (1987) have to introduce further assumptions.

Let us now consider how the model might be identified when no reinterview data is available. For simple linear regression with measurement error in the covariate, the instrumental variable approach (Fuller 1987, Sect. 1.4) assumes the availability of an observed "instrumental" variable W, which is correlated with the covariate, but is independent of the measurement error and independent of the error in the regression equation. We extend this assumption to our framework by defining W to be an *instrumental variable* if it is not independent of x and if

W and (X, Y) are conditionally independent given (x,y), (A3)

W and y are conditionally independent given x. (A4)

In general we shall allow W to be a categorical variable with an arbitrary number s of categories, although since we shall desire W to be closely related to x, we shall usually have $s = r$ in practice. One specific possibility is to take W as the classified state at time $t - 1$. This use of a lagged value of a "covariate" as an instrumental variable may be traced back to the earliest discussions of instrumental variable estimation (*e.g.*, Reiersol 1941; Durbin 1954). In this case, assumption A4 follows if the true states obey a Markov process and the classification errors are conditionally independent, as in A1.

The model resulting from assumptions (A1)-(A4) may be represented by the conditional independence graph in Figure 1. Each vertex in the graph represents a variable. Edges between pairs of vertices are absent if the corresponding variables are conditionally independent given the remaining variables.

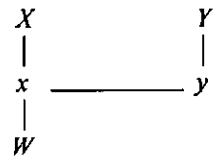


Figure 1. Conditional Independence Graph of Basic Model

The model is an example of a restricted latent class model (Goodman 1974), where the observed variables X, Y and W are conditionally independent given the latent variables x and y, that is they are independent within the r^2 latent classes defined by the pairs of values of (x,y). There are $2(r-1)r^2 + (s-1)r^2 + (r^2 - 1)$ parameters of this model given by the $(r-1)r^2$ parameters $\text{pr}(X = i \mid x = u, y = v)$, the $(r-1)r^2$ parameters $\text{pr}(Y = j \mid x = u, y = v)$, the $(s-1)r^2$ parameters $\text{pr}(W = k \mid x = u, y = v)$ and the $r^2 - 1$ free

parameters $\text{pr}(x = u, y = v)$. These parameters are subject to the $2r(r-1)^2$ restrictions in (A2) and the $(s-1)r(r-1)$ restrictions implied by (A4). We first restrict attention to the case $r = 2$. In this case there are $4s + 7$ parameters subject to $2s + 2$ restrictions, leaving $2s + 5$ free parameters

$$\{K_{x2u}, K_{y2u}, \varphi_{2u}, \ldots, \varphi_{su}, \theta_u, \pi;\ u = 1,2\ v = 1,2\},$$

where $\varphi_{ku} = \text{pr}(W = k \mid x = u)$, $\theta_u = \text{pr}(y = 2 \mid x = u)$, and $\pi = \text{pr}(x = 2)$. The number of "free" cell probabilities in the observed table of X by Y by W is $r^2s - 1$, or $4s - 1$ when $r = 2$. Hence a necessary condition for identification when $r = 2$ is that $4s - 1 \geq 2s + 5$ or $s \geq 3$. Unfortunately, this is not a sufficient condition. For let

$$R_u = \text{pr}(Y = 2 \mid x = u) = \sum_{v=1}^{2} K_{y2v}\theta_u^{v-1}(1 - \theta_u)^{2-v}. \qquad (3)$$

Then

$$\text{pr}(X = i, Y = j, W = k) = \sum_{u=1}^{2} K_{xiu}\varphi_{ku} R_u^{j-1}(1 - R_u)^{2-j}\pi^{u-1}(1 - \pi)^{2-u}. \qquad (4)$$

Hence the $4s - 1$ free cell probabilities are determined by just the $2s + 3$ parameters

$$\{K_{x2u}, \varphi_{2u}, \ldots, \varphi_{su}, R_u, \pi;\ u = 1, 2\}$$

so a necessary condition for identification of these parameters is that $4s - 1 \geq 2s + 3$ or $s \geq 2$. In fact this is also a sufficient condition for identification of these parameters, except for certain exceptional combinations of these parameters. (See Madansky (1960) for the case $s = 2$ and Goodman (1974) for the case of general $s \geq 2$.)

However, even though the above $2s + 3$ parameters are in general identified for $s \geq 2$ it is not possible to determine the 4 parameters $K_{y21}, K_{y22}, \theta_1$ and θ_2 since they are related to only two identified parameters, R_1 and R_2, via equation (3). In particular the key parameters of interest θ_1 and θ_2 remain underidentified whatever the value of s.

It is therefore necessary to impose at least 2 further restrictions on the model to identify θ_1 and θ_2. Following Chua and Fuller (1987), one idea would be to assume unbiased measurement errors as in (2) which imposes the two constraints

$$\pi = K_{x21}(1 - \pi) + K_{x22}\pi \qquad (5)$$

$$\theta_1(1 - \pi) + \theta_2\pi = R_1(1 - \pi) + R_2\pi. \qquad (6)$$

Unfortunately the first constraint only applies to the parameters which are already identified for $s \geq 2$ so these constraints are insufficient to identify θ_1 and θ_2. An alternative assumption which we shall make is that the error process is constant over time so that

$$K_{xiu} = K_{yiu} = K_{iu}, \quad \text{say,} \quad \text{for } i, u = 1, 2, \ldots, r. \qquad (A5)$$

This seems a natural basic assumption if the same survey measurement procedure is used over time. The underidentification problem for the case $r = 2$ discussed above is removed by this assumption since, given the identification of $K_{xiu} = K_{iu}$ and R_u, we can determine θ_u from (3) by

$$\theta_u = (R_u - K_{21})/(K_{22} - K_{21}) \qquad (7)$$

(excluding the trivial case when the measured variables are independent of the true variables so that $K_{22} = K_{21}$).

In summary, when assumptions (A1) - (A5) hold and $r = 2$, our model has $2s + 3$ free parameters $\{K_{2u}, \varphi_{2u}, \ldots, \varphi_{su}, \theta_u, \pi;\ u = 1,2\}$ which are identified if $s \geq 2$, except in exceptional cases such as discussed by Madansky (1960).

Finally, let us return to the case of general r. Since (A5) imposes $(r - 1)r$ restrictions, the number of free parameters becomes $2(r-1)r^2 + (s-1)r^2 + (r^2 - 1) - [2r(r-1)^2 + (s-1)r(r-1)] - (r-1)r = 2r^2 + sr - 2r - 1$. There are $r^2s - 1$ free cell probabilities in the table of X by Y by W so the model will in general be identified if $r(r - 1)(s - 2) \geq 0$. Thus the condition for identification of these parameters remains $s \geq 2$, for any value of $r \geq 2$. Furthermore we can write

$$R_{ju} = \text{Pr}(Y = j \mid x = u) = \sum_{v=1}^{r} K_{jv}\theta_{uv}$$

where $\theta_{uv} = \text{pr}(y = v \mid x = u)$. Hence, provided the matrix $[K_{iu}]$ is non-singular, the θ_{uv} may be determined from the R_{ju} and K_{jv} and hence are also identified. Thus for general r, the model is identified under assumptions (A1)-(A5), except for exceptional cases as discussed by Goodman (1974).

3. ESTIMATION

We shall suppose that for a sample of size n we observe counts n_{ijk} in the cells of the $r \times r \times s$ contingency table of $X \times Y \times W$, and that these are multinomially distributed with parameters n and $p_{ijk} = \text{pr}(X = i, Y = j, W = k)$. The implied log likelihood is

$$l = \sum_i \sum_j \sum_k n_{ijk} \log p_{ijk}.$$

Under a complex sampling design, we may take the n_{ijk} to be weighted counts, giving a pseudo log likelihood (Skinner 1989). The estimators of the parameters obtained by maximising l will be called *instrumental variable* (IV) estimators.

For the remainder of this paper we shall only consider the case $r = s = 2$ when the model is just identified (except for exceptional values of the parameters). In this case we might

attempt to set $p_{ijk} = n_{ijk}/n$ and then solve equations (6) and (7) for the unknown parameters. If the resulting solutions lie within the feasible parameter space, that is probabilities lie in the range [0,1], then these solutions will be the IV estimates. However, in practice we have found that, for moderate sample sizes, infeasible solutions can often arise. Furthermore the solution of these equations is not computationally straightforward. Hence we have found it easier to maximise l directly using the numerical procedures in the package GAUSS (Edlefsen and Jones 1984) or else by using packages which fit latent class models using the EM algorithm such as PANMARK (van de Pol, Langeheine and de Jong 1991). For a latent class package it would be possible to fit an unrestricted two class model and then to estimate θ_1 and θ_2 via (7). However, there would be no guarantee that the resulting estimates would lie in the feasible range [0,1] with this approach. Furthermore there would be the additional complication of determining standard errors for the estimates of θ_1 and θ_2 from the covariance matrix of the estimates of $(R_1, R_2, K_{21}, K_{22})$. Hence we have found it more convenient to fit the model directly as a restricted latent class model. A further advantage of this approach is that it extends naturally to the fitting of similar models across subgroups subject to possible constraints that some parameters are constant across subgroups. This possibility is explored further in Section 4.

Under multinomial assumptions, standard errors may be based on the second derivatives of the log-likelihood evaluated at the IV estimates. This approach becomes problematic, however, if the maximum of l is at the boundary of the parameter space. One approach then is simply to treat the values of the parameters at the boundary as known. However, this is likely to lead to underestimation of uncertainty. Baker and Laird (1988) consider two alternative approaches to obtaining interval estimates for individual parameters in such circumstances: a bootstrap method and a profile likelihood method. The bootstrap method involves drawing repeated multinomial samples with p_{ijk} set equal to n_{ijk}/n and recording the distribution of parameter estimates across repeated bootstrap samples. Interval estimates for given parameters are obtained by the profile likelihood methods as the sets of values of the parameter which are not rejected by a likelihood ratio test. These methods are illustrated at the end of Section 4.

4. NUMERICAL ILLUSTRATIONS

For the purpose of numerical illustration we use data from the equal probability subsample of the US Panel Study of Income Dynamics (PSID). See Hill (1992). We consider the two states employed and not employed, coded 1 and 2 respectively, thus restricting attention again to the binary variable case. For simplicity, we ignore non-response and consider the sample of 5,357 individuals aged 18-64 in 1986 with complete values on the variables: employment status in 1985, 1986 and 1987, car ownership, age, sex and education.

We assess the properties of the IV estimator in two ways. First, in Section 4.1, we compare the bias and standard error of the IV estimator with the "unadjusted" estimator for hypothetical instrumental variables, with a range of different associations with x. Second, in Section 4.2, we consider the impact of using different actual PSID variables as instrumental variables.

4.1 Bias and Standard Error Properties of Estimators for Hypothetical Instrumental Variables

The parameters of primary interest are the joint probabilities $\text{pr}(x = i, y = j)$ or the conditional probabilities $\text{pr}(y = j \mid x = i)$ derived from these. The simple "unadjusted" estimators of these parameters are based on the corresponding sample proportions for the classified variables X and Y and have expectations $\text{pr}(X = i, Y = j)$ under multinomial sampling. Since $\text{Pr}(X = i, Y = j)$ differs in general from $\text{pr}(x = i, y = j)$ the unadjusted estimators are typically biased. Provided the model assumptions (A1)-(A5) hold, the IV estimators of $\text{pr}(x = i, y = j)$ will be asymptotically unbiased although their variances may be larger than those of the unadjusted estimators. The aim of this section is to investigate the extent to which there exists a trade-off in practice between the bias of the unadjusted estimators and the increased variance of the IV estimators. It will be assumed that the model assumptions (A1)-(A5) hold and that the sample is large enough for the IV estimator to be treated as unbiased.

For the numerical investigation in this section we wish to use some "realistic" parameter values. These were determined by rounding the values of estimates for annual flows between the years 1986 and 1987 from analyses in Section 4.2 (reported in Table 3). The values of the five free model parameters not involving W were set to be $K_{21} = 0.03$, $K_{22} = 0.94$, $\text{pr}(x = 2) = \pi = 0.22$, $\text{pr}(y = 2, x = 1) = \theta_1(1 - \pi) = 0.03$ and $\text{pr}(y = 2, x = 2) = \theta_2\pi = 0.19$. Different values of the remaining two free parameters $\varphi_{11} = \text{pr}(W = 1 \mid x = 1)$ and $\varphi_{12} = \text{pr}(W = 1 \mid x = 2)$ are set in the different columns of Table 1. Cramér's V statistic, which measures the association between two binary variables, essentially by scaling the chi-square statistic to a [0,1] interval, is provided as a summary of the strength of association between the variables W and x. For each of the choices of parameter values, Table 1 displays the estimated standard errors of the IV estimators for the PSID sample size $n = 5{,}357$. Table 1 also contains the biases and standard errors of the unadjusted estimator for the same parameter values K_{21}, K_{22}, π, θ_1 and θ_2 and the same sample size.

To illustrate the calculation of the biases of the unadjusted estimators, consider $\text{pr}(x = 1, y = 1)$. The expectation of the unadjusted estimator of this parameter is $\text{pr}(X = 1, Y = 1)$, which is calculated from the given values of K_{21}, K_{22}, π, θ_1 and θ_2 and assumptions (A1)-(A5) as 0.71. This compares with the assumed value of $\text{pr}(x = 1, y = 1)$ of 0.75. The bias is thus $0.71 - 0.75 = -0.04$. The biases of the IV estimators are, as noted above, assumed to be zero. The standard errors of the unadjusted estimators are obtained from standard binomial

Table 1
Biases and Standard Errors under Alternative Hypothetical IVs

		Parameter Values Assumed for IV estimator						
$pr(W=1 \mid x=1)$		1.0	0.1	0.1	0.1	0.3	0.1	0.5
$pr(W=1 \mid x=2)$		0.0	0.9	0.7	0.5	0.7	0.3	0.3
Cramér's V		1.0	0.74	0.59	0.42	0.34	0.24	0.17
		Standard Errors (× 100)						
Parameter Estimated	Bias (× 100) of Unadjusted Estimator	Unadjusted Estimator	IV Estimator					
$pr(x=1, y=1)$	-4.0	0.62	0.68	0.75	0.88	1.13	1.16	1.82	2.05
$pr(x=1, y=2)$	3.0	0.32	0.39	0.43	0.51	0.64	0.69	1.03	1.24
$pr(x=2, y=1)$	3.0	0.32	0.32	0.37	0.44	0.57	0.66	0.95	1.27
$pr(x=2, y=2)$	-2.0	0.51	0.59	0.65	0.73	0.89	1.06	1.42	1.99
$pr(y=1 \mid x=1)$	-3.9	0.37	0.50	0.55	0.64	0.81	0.88	1.30	1.58
$pr(y=1 \mid x=2)$	12.4	0.60	1.40	1.63	1.95	2.56	2.90	4.30	5.55

Note: 1 = employed, 2 = not employed; $n = 5{,}357$; multinomial sampling assumed; biases of IV estimators are zero.

formulae. For example, the standard error of the unadjusted estimator of $pr(x=1, y=1)$ is $\sqrt{0.71 \times 0.29/5{,}357} = 0.0062$, where 0.71 is the value of $\Pr(X=1, Y=1)$. The standard errors of the IV estimators are obtained from the inverse of the expected information matrix, which is given by $n\sum p_{ijk} H_{ijk}$, where H_{ijk} is the 7×7 matrix of second derivatives of $\log p_{ijk}$ with respect to the seven free parameters. Following differentiation, these parameters are set equal to their assumed values, as indicated above. Note that the standard errors obtained from the multinomial information matrix are likely to be under-estimates because of the complex sampling design employed in the PSID.

There is a clear pattern of the standard errors of the IV estimator increasing as the association between W and x decreases. The amount of increase is fairly similar across all parameters, for example the ratio for V = 0.20 versus V = 1.00 lies between 3 and 4 for all parameters. In all cases the standard error of the IV estimator is greater than that of the unadjusted estimator. The loss of efficiency of the "best" IV estimator (with perfect association between W and x) compared to the adjusted estimator varies between parameters. Roughly speaking, the loss is greater for the conditional parameters than for the unconditional parameters. This loss of efficiency might be interpreted as the effect of adjusting for measurement error in y, which is still necessary even when x is perfectly measured by W. Under this interpretation, the greater relative loss of efficiency for the conditional parameters seems plausible since these are "less dependent" on the parameters of the marginal x distribution which the W information helps to estimate.

To examine the trade-off between the bias of the unadjusted estimator and the increased variance of the IV estimator we have calculated the minimum value of the sample size n necessary for the MSE of the IV estimator to be less than that of the unadjusted estimator. For complex designs the sample sizes should be interpreted as effective sample sizes. Table 2 gives these minimum values under a variety of strengths of association between W and x. If there were no misclassification the entries would all be infinity since the unadjusted estimators would always be more efficient than the IV estimators. For the assumed amount of misclassification given by $K_{21} = 0.03$ and $K_{12} = 0.06$, the sample size required increases rapidly as V decreases. The differences between the rows of Table 2 are partly accounted for by the differences between the rows of Table 1 and partly by differences between the biases of the unadjusted estimator. Thus, the bias of the unadjusted estimator of $pr(x=2, y=2)$ is relatively small and this leads to the large values in the corresponding row of Table 2. Note that the value of 1 for $pr(x=2, y=1)$ and Cramér's V = 1 arises because in this case the standard errors of the two estimators are equal (see Table 1) and so the bias of the unadjusted estimators implies that the IV estimator has smaller MSE for any $n \geq 1$.

The main conclusion we wish to draw from Table 2, however, is simply that we may expect there to be a number of practical situations where IV estimation will be worthwhile provided the model assumptions hold, even if the necessary sample sizes are inflated somewhat to allow for complex sampling designs.

4.2 Results for Actual Instrumental Variables

The results in the previous section were based on hypothetical instrumental variables. To provide a more realistic illustration we now consider possible real instrumental variables. The key problem is how to choose a variable W which obeys (A3) and (A4). It seems easier to find a variable which satisfies (A3) than (A4), in particular

Table 2
Sample Size Necessary for MSE of IV Estimator to be less than that of Unadjusted Estimator (Multinomial Sampling)

	Value of Cramér's V assumed for IV estimators						
	1.0	0.74	0.59	0.42	0.34	0.24	0.17
Parameter Estimated	Sample size n required						
$pr(x = 1, y = 1)$	28	59	132	300	320	971	1273
$pr(x = 1, y = 2)$	31	50	91	184	219	573	843
$pr(x = 2, y = 1)$	1	20	51	129	198	476	811
$pr(x = 2, y = 2)$	112	227	366	720	1184	2397	5070
$pr(y = 1 \mid x = 1)$	42	60	97	183	219	541	818
$pr(y = 1 \mid x = 2)$	57	81	121	216	281	633	1061

measured without error obey (A3). However, it seems more difficult to find variables which one is sure are not related to change in employment status and hence obey (A4).

For illustration, we have considered two possibilities. First we have taken W as car ownership ($W = 2$ if the individual owns a car, $W = 1$ if not). This variable is likely to be measured with some error but it seems a reasonable first assumption that this error is unrelated to errors in measuring employment status. For example, in an analysis of errors in recording car ownership in the 1981 British Census, Britton and Birch (1985, p. 67) conclude that "the main problems associated with the small number of discrepancies were those connected with either vehicles out of use or vehicles temporarily available – for example, those hired..." and it seems at least plausible that such errors need have little relation to the kinds of errors in recording employment status. On the other hand, it is plausible that car ownership acts as a proxy for some kind of social or economic status which is related to change in employment status so assumption (A4) seems more questionable. However, for our illustrative purpose we assume (A3) and (A4) hold.

As a second illustration we have taken W to be the lagged employment status in 1985. A problem here is that (A4) effectively implies that individual employment histories follow Markov processes with common transition rates. In fact, transition rates will vary among individuals and this will invalidate assumption (A4) (*e.g.*, van de Pol and Langeheine 1990). Therefore, to allow for departures from assumption (A4), we disaggregated the sample into 16 groups defined by cross-classifying age (4 groups), sex and education (up to college level or not). We then assumed the model held within subgroups and used likelihood ratio tests to assess what parameters were constant across subgroups. These tests only provide a very rough guide since they ignore the complex sampling design of the PSID. There was no significant evidence of differences in the misclassification probabilities K_{ij} across subgroups. Furthermore, within each of the 8 subgroups defined by age × sex there was no significant evidence of differences in $\Pr(W \mid x, \text{subgroup})$ between the 2 education subgroups. Assuming equality of these parameters gave a non-significant likelihood-ratio goodness-of-fit chi-squared value of 52.9 on 46 df (46 is obtained as the number of cells = $16 \times 8 = 128$, less $2K_{ij}$ parameters, less $16 \times 4 = 64$ $pr(x, y, \text{subgroup})$ parameters, less $8 \times 2 = 16$ $pr(W \mid x, \text{subgroup})$ parameters). Combining the parameter estimates for the disaggregated model appropriately gives estimates of the overall flows $pr(x, y)$.

Table 3 contains estimates of the key parameters for the two choices of instrumental variable and for the disaggregated version of the second choice. We note first that the standard errors for the IV estimator based on car ownership are relatively high. This may be expected from Table 1 since the association between x and W is low (Cramér's V is 0.12). Even so, the resulting adjustments increasing the estimates for the diagonal entries are plausible and the confidence intervals resulting from this IV estimator seem more realistic than those for the unadjusted estimator.

Table 3
Unadjusted and IV Estimates for PSID Data

Parameter	Unadjusted Estimates	IV Estimates		
		IV = Car Ownership	IV = Lagged Employment	IV = Lagged Employment (Disaggregated)
$pr(x = 1, y = 1)$	0.719	0.773	0.766	0.757
	(0.006)	(0.033)	(0.008)	(0.007)
$pr(x = 1, y = 2)$	0.055	0.011	0.017	0.025
	(0.003)	(0.020)	(0.005)	(0.003)
$pr(x = 2, y = 1)$	0.061	0.018	0.024	0.032
	(0.003)	(0.019)	(0.004)	(0.003)
$pr(x = 2, y = 2)$	0.166	0.198	0.193	0.186
	(0.005)	(0.027)	(0.007)	(0.006)

Note: Standard errors under multinomial assumptions in parentheses. Disaggregation is by age (4 groups), sex and education (2 groups).

The standard errors for the second choice of instrumental variable are smaller, as expected since the association with X is now higher (Cramér's V is 0.73). Indeed these standard errors are not much larger than those for the unadjusted estimator. The (2 standard error) confidence intervals now do not overlap with the corresponding intervals for the unadjusted estimator for any of the four parameters.

As noted earlier, assumption (A4) is questionable for the lagged employment variable. The disaggregated version of this estimator makes "weaker" assumptions by only requiring (A4) to hold within subgroups. The resulting estimates are seen to be fairly close to the original IV estimator and to have slightly smaller standard errors, perhaps attributable to the use of the additional information on sex, age and education (but see later discussion). It is interesting that the effect of the disaggregation is to diminish the effect of adjustment by a relatively small amount in each case. It seems plausible that departures from (A4) may tend to lead to overadjustment in the IV estimator and that the disaggregation approach here helps to overcome this bias and, for alternative choices of disaggregating variables, enables an assessment of the sensitivity of results to the model specification.

As noted in Section 3 we have often come across IV estimates on the boundary of the interval [0,1]. Of the analyses reported in Table 3 in fact only the disaggregated analysis involved boundary estimates. For the 64 parameters $\text{pr}(x = i, y = j, \text{subgroup})$ for $i, j = 1,2$, subgroup = 1, ..., 16, five of the estimates were on the boundary (none of the estimates of the remaining 18 parameters, $\text{pr}(W = 1 \mid X = 1)$ and so forth, were). The standard errors reported in Table 3 treat these parameters as known and hence may underestimate the uncertainty in the estimates of the aggregate $\text{pr}(x = i, y = j)$ parameters.

Table 4
Alternative Estimates of Standard Errors
for Males Aged 26-35 with no College Education

Parameter	IV estimates	Estimated Standard Error	
		Standard	Bootstrap
$\text{pr}(W = 1 \mid x = 1)$	0.947	0.011	0.011
$\text{pr}(W = 1 \mid x = 2)$	0.107	0.089	0.091
$\text{pr}(X = 1 \mid x = 1)$	0.969	0.006	0.007
$\text{pr}(X = 1 \mid x = 2)$	0.084	0.088	0.075
$\text{pr}(x = 1, y = 1)$	0.953	0.011	0.012
$\text{pr}(x = 1, y = 2)$	0	*	*
$\text{pr}(x = 2, y = 1)$	0.006	0.007	0.006
$\text{pr}(x = 2, y = 2)$	0.041	0.012	0.011
$\text{pr}(x = 1)$	0.953	0.011	0.011
$\text{pr}(y = 1 \mid x = 1)$	1	*	*
$\text{pr}(y = 1 \mid x = 2)$	0.128	0.139	0.117

Note: $n = 455$; "standard" estimators based on observed information matrix, treating parameters estimated at the boundary as known; 10,000 replications of bootstrap; multinomial assumptions.

Table 4 presents alternative estimates of the standard errors for one subgroup, males aged 26-35 with no college education. The estimate of $\text{pr}(x = 1, y = 2)$ as well as derived estimates, such as $\text{pr}(y = 1 \mid x = 1)$ lie on the boundary. The "standard" estimates of the standard errors are, as in Table 3, based on the observed information matrix, treating parameters estimated at the boundary as known. Bootstrap standard error estimates (for 10,000 replications) are found to be very close to these standard estimates for parameters with estimates not on the boundary. For the IV estimate of $\text{pr}(x = 1, y = 2)$ at the boundary no standard estimate of the standard error is available. Indeed it seems to make little sense to estimate the standard deviation of the sampling distribution in this case. It seems more sensible to derive a one-sided confidence interval which may be done either using the profile likelihood method, which gives [0, .016], or using the bootstrap percentile method, which gives [0, .009]. The corresponding intervals for $\text{pr}(y = 1 \mid x = 1)$ are [.983, 1] and [.990, 1].

5. CONCLUSION

The presence of measurement error can induce substantial bias into standard estimates of transition rates from longitudinal data. If external estimates of misclassification rates are available then a variety of adjustment methods exist. If no such information is available then this paper shows how adjustment for measurement error alternatively can be carried out using instrumental variable estimation.

The main problem, as in conventional instrumental variable estimation, is finding a variable which one can be confident satisfies the conditions required of an instrumental variable. Even if the conditions are satisfied then it is desirable, in order to obtain reasonable precision, that there be a fairly strong association between this variable and the true state. If such a variable can be found then instrumental variable estimation may be useful.

ACKNOWLEDGEMENTS

We are grateful to Wayne Fuller for suggesting the basic idea underlying this paper. Research was supported by grant number H519 25 5005 from the Economic and Social Research Council under its Analysis of Large and Complex Datasets programme.

REFERENCES

ABOWD, J.M., and ZELLNER, A. (1985). Estimating gross labor force flows. *Journal of Business and Economic Statistics*, 3, 254-283.

ANDERSON, T.W. (1959). Some scaling models and estimation procedures in the latent class model. *Probability and Statistics*, (Ed. U. Grenander). Stockholm: Wiksell and Almquist.

BAKER, S.G., and LAIRD, N.M. (1988). Regression analysis for categorical variables with outcome subject to nonignorable nonresponse. *Journal of the American Statistical Association*, 83, 62-69.

BARTHOLOMEW, D.J. (1987). *Latent Variable Models and Factor Analysis*. London: Griffin.

BIEMER, P.P., GROVES, R.M., LYBERG, L.E., MATHIOWETZ, N.A., and SUDMAN, S. (1991). *Measurement Errors in Surveys* New York: Wiley.

BRITTON, M., and BIRCH, F. (1985). *1981 Census Post-Enumeration Survey*. London: Her Majesty's Stationery Office.

CHUA, T., and FULLER, W.A. (1987). A model for multinomial response error applied to labor flows. *Journal of the American Statistical Association*, 82, 46-51.

DURBIN, J. (1954). Errors in variables. *Review of the International Statistical Institute*, 22, 23-31.

EDLEFSEN, L.E., and JONES, S.D. (1984). Reference Guide to GAUSS. Applied Technical Systems.

FORSMAN, G., and SCHREINER, I. (1991). The design and analysis of reinterview: an overview. In *Measurement Errors in Surveys*. (Eds. Biemer, P.P., Groves, R.M., Lyberg, L.E., Mathiowetz, N.A., and Sudman, S.). New York: Wiley.

FULLER, W.A. (1987). *Measurement Error Models*. New York: Wiley.

GOODMAN, L.A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, 61, 215-231.

HILL, M.S. (1992). *The Panel Study of Income Dynamics: A User's Guide*. Newbury Park, CA: Sage.

HOGUE, C.R., and FLAIM, P.O. (1986). Measuring gross flows in the labor force: an overview of a special conference. *Journal of Business and Economic Statistics*, 41, 111-21.

MADANSKY, A. (1960). Determinental methods in latent class analysis. *Psychometrika*, 25, 183-198.

MARQUIS, K.H., and MOORE, J.C. (1990). Measurement errors in the Survey of Income and Program Participation (SIPP): Program Reports. *Proceedings of the 1990 Annual Research Conference*. US Bureau of the Census, 721-745.

MEYER, B.D. (1988). Classification-error models and labor-market dynamics. *Journal of Business and Economic Statistics*, 6, 385-390.

POTERBA, J.M., and SUMMERS, L.H. (1986). Reporting errors and labor market dynamics. *Econometrica*, 54, 1319-1338.

REIERSOL, D. (1941). Confluence analysis by means of lag moments and other methods of confluence analysis. *Econometrica*, 9, 1-24.

SINGH, A.C., and RAO, J.N.K. (1995). On the adjustment of gross flow estimates for classification error with application to data from the Canadian Labour Force Survey. *Journal of the American Statistical Association*, 90; 478-488.

SKINNER, C.J. (1989). Domain means, regression and multivariate analysis. In *Analysis of Complex Surveys*, (Ch. 3) (Eds. Skinner, C.J., Holt, D., and Smith, T.M.F.). Chichester: Wiley.

SKINNER, C.J., and TORELLI, N. (1993). Measurement error and the estimation of gross flows from longitudinal economic data. *Statistica*, 53, 391-405.

VAN DE POL, F., and DE LEEUW, J. (1986). A latent Markov model to correct for measurement error. *Sociological Methods and Research*, 15, 118-141.

VAN DE POL, F., and LANGEHEINE, R. (1990). Mixed Markov latent class models. In *Sociological Methodology 1990*, (Ed. C.C. Clogg). Oxford: Basil Blackwell, 213-247.

VAN DE POL, F., LANGEHEINE, R., and DE JONG, W. (1991). PANMARK User Manual. Panel analysis using Markov chains. Version 2.2. Netherlands Central Bureau of Statistics.

Survey Methodology, June 1997
Vol. 23, No. 1, pp. 61-71
Statistics Canada

Geographic-Based Oversampling in Demographic Surveys of the United States

JOSEPH WAKSBERG, DAVID JUDKINS and JAMES T. MASSEY[1]

ABSTRACT

Often one of the key objectives of multi-purpose demographic surveys in the U.S. is to produce estimates for small domains of the population such as race, ethnicity, and income. Geographic-based oversampling is one of the techniques often considered for improving the reliability of the small domain statistics using block or block group information from the Bureau of the Census to identify areas where the small domains are concentrated. This paper reviews the issues involved in oversampling geographical areas in conjunction with household screening to improve the precision of small domain estimates. The results from an empirical evaluation of the variance reduction from geographic-based oversampling are given along with an assessment of the robustness of the sampling efficiency over time as information for stratification becomes out of date. The simultaneous oversampling of several small domains is also discussed.

KEY WORDS: Sample design; Stratification; Rare populations.

1. INTRODUCTION

The sponsors of many broad multi-purpose demographic surveys require separate analyses of domains defined by race, ethnicity and income. Equal probability samples generally do not provide sufficient sample sizes for some of these domains to yield the precision needed, making some form of oversampling necessary. This requirement poses interesting methodological problems since there is no registry of the U.S. population from which samples stratified by these domains can be drawn. Housing lists containing identifiers for these domains are maintained at the Bureau of the Census, but they are not available to researchers outside of the Bureau. For surveys requiring face-to-face interviews, outside researchers are thus forced to use area sampling techniques. Even within the Bureau, geography is sometimes used as the basis of oversampling since the lists are only updated once every ten years. This paper describes efficient methods for oversampling the aforementioned domains in the context of area sampling.

Data from the U.S. Decennial Census on concentrations of various demographic domains are publicly available for small geographic units; race and ethnicity are reported for every block and income for every block group. (A "block" is an area bounded on all sides by roads and not transected by any roads. Block groups are combinations of several neighbouring blocks.) These data may be used to inexpensively improve the precision of statistics about rare domains by oversampling blocks or block groups that contain higher than average concentration of members of rare domains and then dropping or subsampling screened persons not in the targeted rare domains. The general theory for this type of sample design was worked out by Kish (1965, Section 4.5). An independent presentation of the theory with examples from the 1960 Decennial Census was given by Waksberg (1973). Further examples and a discussion of alternative methods are given by Kalton and Anderson (1986) and by Kalton writing for the United Nations (1993). In this paper, we extend prior illustrations to cover more domains, update results to 1990, and evaluate empirically the robustness of these methods over time.

We first briefly review the issues involved with screening and subsampling persons not in the targeted domains. Then we review the theory for optimal allocation where the strata are defined in terms of the density of rare populations and apply this theory to several rare populations. The main part of the paper is an empirical evaluation of the reduction in variance reduction from the geographic oversampling of various minority and other rare populations as well as how robust the variance reductions are over time. We also discuss the special problems involved with simultaneous targeting of several rare populations before summarizing our conclusions.

2. SURVEY COST STRUCTURE AND THE SCREENING DECISION

Let U stand for some target universe such as persons or households for which a sampling frame exists. Let D stand for some small domain of particular interest such as black persons that cannot be separately identified from the balance of U at the time of sampling. Let Y be a vector of characteristics of interest such as annual income, employment status, and number of doctors' visits in the last year. In some surveys, the only objective is estimation of the distribution of Y on D. In such surveys, members of U-D that are discovered in the course of screening sampled members of U will be dropped from the sample. A general inexpensive interview

[1] Joseph Waksberg, Westat Inc., 1650 Research Blvd., Rockville, MD 20850, U.S.A.; David Judkins, Research Triangle Institute, 5901-B Peachtree-Dunwoody Road, Suite 500, Atlanta, GA 30325, U.S.A.; James T. Massey, formerly of Westat Inc., now deceased.

questionnaire is used for the screening to determine who is eligible for a full questionnaire.

In other surveys, estimation of the distribution of Y on D and on U are both important objectives. For such a survey, at least some of the members of U-D that are discovered in the course of screening interviews will be retained for full interviews. If geographic-based oversampling is used, the initial sample will contain an oversample of those members of U-D who happen to reside in areas with heavy concentrations of D. Even when U-D is of interest, this oversampling of U-D in areas with high concentrations of D is usually undesirable since resulting variation in probabilities of selection for U-D leads to unnecessarily large design effects for statistics both about U and about U-D. These larger design effects mean that the extra sample size for U-D will usually result in only a trivial decrease in variances for statistics about U-D. Generally, the funds expended on the extra interviews with U-D would be better spent on increasing the total initial sample size.

It is fairly easy to set up subsampling procedures that result in an equi-probability sample of U-D. The subsampling can be done centrally after the completion of the entire screening operation, or it can be done by the interviewer while still in the sample household after obtaining data on household composition. Techniques have been developed that make the subsampling process very easy for the interviewer (Waksberg and Mohadjer 1991). Interviewers do not need to be trained to carry out random draws. With paper and pencil survey instruments, interviewers are given house-by-house pre-interview instructions about which domains can be interviewed at which households. These instructions are randomized centrally prior to screening to yield the desired sampling rates. Alternatively, with CAPI, the subsampling can be programmed and carried out automatically in the laptop computer used for CAPI; the computer notifies the interviewer which households are to be retained for the full interview and which ones to reject as a result of subsampling.

Whether it is better to keep all sampled members of U-D or to subsample them depends on the relative sizes of U and U-D, the precision requirements for both and on the relative costs of full interviews and the shorter screening interviews. Let $c\star$ be the variable cost associated with sampling a single member of U and collecting and processing all data of interest about that member. Let c' be the variable cost associated with sampling, screening, and then dropping a single member of U. Let $c = c\star/c'$, be the ratio of the cost of a full interview to the cost of a screening interview. If c is much greater than 1, then subsampling should be considered for the survey that has interest in U-D even though subsampling of U-D will introduce some additional complexity into survey operations. Given that the full interview is by definition longer that the screening interview, it should always be the case that c is at least slightly greater than 1. On panel and longitudinal surveys, the cost of all follow-back interviews should be counted as part of $c\star$, typically making the cost of a full interview many times larger than the cost of a screening interview; *i.e.*, $c >> 1$. The same will be true of surveys that involve the collection of physical specima requiring expensive laboratory work and of surveys that require expensive experts (such as medical doctors) to participate in the primary data collection. For such surveys, we would highly recommend that geographic-based oversampling not be employed by itself, but rather, in conjunction with screening and subsampling. For a door-to-door survey with a single interview by a standard grade interviewer (trained to ask questions and record answers but not to make any technical or anthropological assessments), c is frequently in the range of 3 to 5. This is large enough in many applications to justify the complication of subsampling U-D in oversampled areas.

3. FORMING THE STRATA

We assume that even though D cannot be separated from U at the time of sampling, there is some information available about the distribution of D and U across a set of geographically defined entities. In the United States, the natural entities are blocks or block groups (BGs) and information for these entities is supplied by the decennial census. (Prior to the 1990 decennial census, blocks were not defined in rural areas; larger entities called "enumeration districts" were used for oversampling.) The U.S. Bureau of the Census makes data on the racial and ethnic composition of blocks publicly available along with mapping information so that these blocks can be identified years later by any survey organization. Income data are only made available at the BG level.

Standard practice calls for the stratification of the blocks or BGs by the local concentration of D. Thus, all blocks where D constitutes less that 10 percent of the block's total population might constitute one stratum. Further cutpoints for defining the strata might be 30 percent, and 60 percent, yielding a total of four strata. There has been little empirical study of the optimal number of strata nor of the optimal cutpoints. In general, more strata will yield more efficient designs, but, at some point, the operational complexities of a large number of strata outweigh the gains in efficiency. Conventional wisdom dating back to Kish (1965) holds that a fairly small number of strata will achieve most of the gains attainable through stratification.

4. OPTIMAL ALLOCATION FOR A SINGLE DOMAIN

Our objective is to adapt the general formulas for optimum allocation of a stratified sample to apply to the reduction in variance due to geographic-based oversampling. The derivations are essentially those given by Kish (1965) using the notation of Kalton in United Nations (1993). Let the population be divided into a number of strata as discussed above. Let N be the size of the total population and N_h be the

size of the total population within the h-th stratum. Let P_h be the proportion of the h-th stratum that consists of members of D. Let P be the overall proportion of the population that belongs to D. We may use the prior decennial census to estimate P_h and P, or we may use some more recent large survey that carried block and/or BG codes for every sample household/person so that matching to the last decennial census will yield the stratum identification for every sample household/person.

We assume that c is constant across the strata even though this may sometimes not be very accurate. For example, interviewing in blocks with high concentrations of American Indians, Eskimos or Aleuts almost always means interviewing in remote locations with difficult transportation issues. However, estimation of even a national average for c is difficult for most survey operations. It will not generally be possible to get estimates by stratum.

We also assume that the distribution of Y on D is constant across the strata. More specifically, we assume that

$$E(Y \mid D \text{ and } h) \equiv E(Y \mid D) \quad \text{and that}$$

$$\text{Var}(Y \mid D \text{ and } h) \equiv \text{Var}(Y \mid D),$$

where the expected value and variance are with respect to the population, not the sample design. This is usually not a very good assumption, but given a vector of characteristics of interest, the components of the vector will usually behave differently across the strata so there is no point in trying to be more exact. Lastly, we assume that the sampling fractions are small enough in all the strata to make the finite population correction factors ignorable.

Given these assumptions, the optimal sampling fraction for the h-th stratum for a survey where all screened members of U-D are dropped is

$$f_h = k\sqrt{\frac{P_h}{P_h(c-1)+1}}, \tag{1}$$

where k is a constant determined by either precision requirements or budget constraints. (For a proof of (1), see either of the sources referenced above. This allocation rule is an application of Neyman allocation.) If $c = 1$, (*i.e.*, screening is as expensive as interviewing), then this proportionality reduces to $f_h \propto \sqrt{P_h}$, which can yield allocations quite different from an equi-probability sample across strata. However, if the cost of screening is far less than the cost of interviewing (*i.e.*, $c >> 1$) and D is not extremely rare (*i.e.*, P_h is not close to zero), then this relationship results in close to a flat set of sampling intervals, which is equivalent to allocation in proportion to total population.

Given a fixed budget of B, k is determined by the cost equation

$$B = \sum_h N_h f_h c' \left[P_h c + (1 - P_h)\right]. \tag{2}$$

To obtain a simple random sample of size n from domain D would require selecting a screening sample of size n/P, resulting in a total cost of

$$B = ncc' + \left(\frac{n}{P} - n\right)c'. \tag{3}$$

By equating these two costs, we can solve for the constant of proportionality in (1) and get:

$$k = \frac{n\left(c - 1 + \frac{1}{P}\right)}{\sum_h N_h P_h \sqrt{c - 1 + \frac{1}{P_h}}}. \tag{4}$$

To calculate the benefits of this allocation realistically, it is necessary to acknowledge the fact that the estimates of P_h that are used to guide the allocation will be somewhat out of date by the time that the survey is actually conducted. Let A_h be the proportion of D actually to be found within the h-th stratum at the time of sampling and data collection. It is assumed that P is unchanged even though the distribution across strata changes according to A_h. By letting $NP = N_D$ and $N_D A_h = N_{Dh}$ it can readily be shown that the actual sample size, n_D, that will be achieved on D is given by

$$n_D = \sum_h NPA_h f_h. \tag{5}$$

From Kish (1965), this sample will have higher variance than a simple random sample of the same size on D. The variance inflation factor or design effect associated with the differential sampling rates across strata is the well-known

$$deff = \left(\sum_h A_h f_h\right)\left(\sum_h A_h / f_h\right). \tag{6}$$

Thus, the *effective* sample size associated with the geographic-based oversampling is

$$\frac{n_D}{deff} = \frac{NP}{\left(\sum_h A_h / f_h\right)}. \tag{7}$$

Substitution of formulae (1) and (4) into (7) yields

$$\frac{n_D}{deff} = \frac{n\left(c - 1 + \frac{1}{P}\right)}{\left(\sum_h A_h \sqrt{c - 1 + \frac{1}{P_h}}\right)\left(\sum_h \frac{N_h P_h}{NP}\sqrt{c - 1 + \frac{1}{P_h}}\right)}. \tag{8}$$

This formula allows us to compare the variance for an arbitrary statistic on domain D given geographic-based oversampling with the variance for the same statistic given a simple random sample of D of the same total cost B. Formula (8) can be rewritten algebraically such that the proportion of simple random sample variance that is eliminated by the geographic-based oversampling is given by

$$\frac{\frac{\sigma^2}{n} - \frac{\sigma^2 deff}{n_D}}{\frac{\sigma^2}{n}} = 1 - \frac{\left(\sum_h A_h \sqrt{c - 1 + \frac{1}{P_h}}\right)\left(\sum_h \frac{N_h P_h}{NP} \sqrt{c - 1 + \frac{1}{P_h}}\right)}{\left(c - 1 + \frac{1}{P}\right)}. \quad (9)$$

It is definitely possible for this reduction to be negative, meaning that a simple random sample would have provided lower variance for the same cost. This is most likely to happen when there exists a stratum for which $NPA_h >> N_h P_h$, meaning that there exists a stratum which was thought to have a very small portion of D but, in fact, has quite a significant portion of D. Note that if $P_h \equiv P$, then no variance reduction can be expected from geographic-based oversampling. Also, as c goes to infinity for fixed P (equivalent to screening becoming cheaper and cheaper relative to full interviews), the variance reduction approaches zero. Given the extra complication of a stratified sample, this means that for large c and moderate P, the sample designer should consider drawing a simple random sample instead of a stratified sample. Geographic-based oversampling increases in value as P approaches zero, c approaches 1, and D becomes more concentrated in a single stratum. As the small domain of interest, D, becomes more concentrated in a single stratum the sample becomes more efficient, since there are fewer cases from D in the remaining strata with large differential. The potential reductions in variance due to geographic-based oversampling under a number of conditions are shown empirically for several demographic domains in the section below.

5. EMPIRICAL EVALUATION

Equation (9) is quite difficult to evaluate for domains of interest. Data on P_h can be obtained from summary tapes from the decennial censuses that are published at the block, block group, and enumeration district levels by the Bureau of the Census. This allows one to define reasonable strata and to evaluate equations (1) through (4). If one were to assume that the P_h are static over time, then the rest of the equations could also be evaluated. However, Americans tend to move frequently, and the racial and ethnic composition of many blocks change in that process (Judkins, Massey and Waksberg 1992). To the extent that members of D move into areas where they were previously not common, the benefits of the geographic-based oversampling diminish. Not wishing to overstate the benefits of the procedure, we searched for some method to get reasonable estimates of the A_h at postcensal time points. Matching block- or BG-level data for two consecutive censuses might appear to be a good solution but is not possible. Up to now, blocks have been defined and labelled independently from census to census with no attempt to preserve definitions for longitudinal. Thus, alternate information sources are required to estimate A_h.

For the analysis of the benefits of geographic-based oversampling for the black and Hispanic populations, micro-level data from current household surveys conducted by the Census Bureau turned out to be a good source of information on the A_h. Specifically, we used data from the 1988 National Health Interview Survey (NHIS). Staff at the Census Bureau prepared a special tape for us that gave the 1980 block group or enumeration district code for almost all households interviewed in the 1988 NHIS in residences built prior to 1980. (Residences constructed during the 1980s would have been sampled for the NHIS from building permits rather than by area sampling. Due to technical difficulties, block and block group labels are not attached to such sample dwellings.) We then matched the 1988 NHIS against 1980 Census summary files by block group or enumeration district in order to classify NHIS households into strata defined by concentrations of blacks and Hispanics in 1980. Using survey weights, we were then able to estimate the distribution of various domains across those strata. (Housing built during the 1980s was assumed to be in the stratum with the lowest concentration of the rare domains.) Similar operations could have been carried out for Asians, Pacific Islanders, American Indians, Eskimos, Aleuts, and persons with low income but were not.

Tables and charts in the balance of the paper will refer to data at several points in time and from several sources. It is useful to bear in mind that the data used to form the strata do not have to be the same as the data used to allocate the sample, and that the data used to evaluate the sample may be from a third point in time or source. We have the following combinations in this paper:

Label	Source of stratification data	Source of allocation data	Source of evaluation data
80/80/80 BG	1980 Census (BG level)	1980 Census	1980 Census
80/80/88 BG	1980 Census (BG level)	1980 Census	1988 NHIS
80/88/88 BG	1980 Census (BG level)	1988 NHIS	1988 NHIS
90/90/90 BG	1990 Census (BG level)	1990 Census	1990 Census
90/90/90 blk	1990 Census (block level)	1990 Census	1990 Census

Table 1
Residential Clustering of Blacks

Density stratum (Blacks as a percent of the stratification unit in the year of stratification)	Percentage of blacks living in the stratum in the indicated year				Percentage of the total population living in the stratum in the indicated year			
Measurement year	1980	1988	1990	1990	1980	1988	1990	1990
Stratification year	1980	1980	1990	1990	1980	1980	1990	1990
Stratification unit	BG/ED	BG/ED	BG	Block	BG/ED	BG/ED	BG	Block
< 10%	9.7	20.5	12.0	8.5	78.2	81.4	75.7	77.5
10-30%	13.5	13.2	16.8	13.9	8.9	7.1	11.4	9.6
30-60%	18.9	20.4	20.3	16.2	5.1	5.1	5.7	4.5
60-100%	57.9	45.9	51.0	61.4	7.8	6.4	7.2	8.4
Total populations (1000s)	26,495	29,380	29.986	29,986	226,546	240,876	248,710	248,710
Blacks as percent of nation in measurement year	11.7	12.0	12.1	12.1				

Sources: 1980 Decennial Census (Westat tabulation)
1988 National Health Interview Survey (Westat tabulation)
1990 Decennial Census (Westat tabulation)

6. OVERSAMPLING THE BLACK POPULATION

Table 1 shows various aspects of residential segregation for the black population in the U.S. that are important to know about when designing a population survey. Although the percentage of blacks living in densely black (60+ percent) block groups declined between 1980 and 1990, it is clear that blacks were still strongly segregated. The columns about the population in 1988 are particularly important since they show the dynamics of the stratification data over time. By 1988, the percentage of the black population living in the block groups that were less than 10 percent black in 1980 had doubled, from just 9.7 percent of blacks to 20.5 percent. This has major implications for the efficacy of geographic-based oversampling as will be shown below. It is also interesting to note that the total population in the block groups that were densely black (*i.e.*, over 60% black) in 1980 actually declined by about 2 million persons between 1980 and 1988. At least part of this shift came from abandonment of some old housing and neighbourhoods. Concentration levels are sharper at the block level than at the block group level in 1990, as would be expected. (Block level data are not available for the whole nation from 1980.) Although sampling blocks is slightly more costly than sampling block groups (due to the larger number of blocks and the need to make provisions for blocks that have fewer inhabitants than the desired sample cluster size), it does allow sharper focus on the targeted domain.

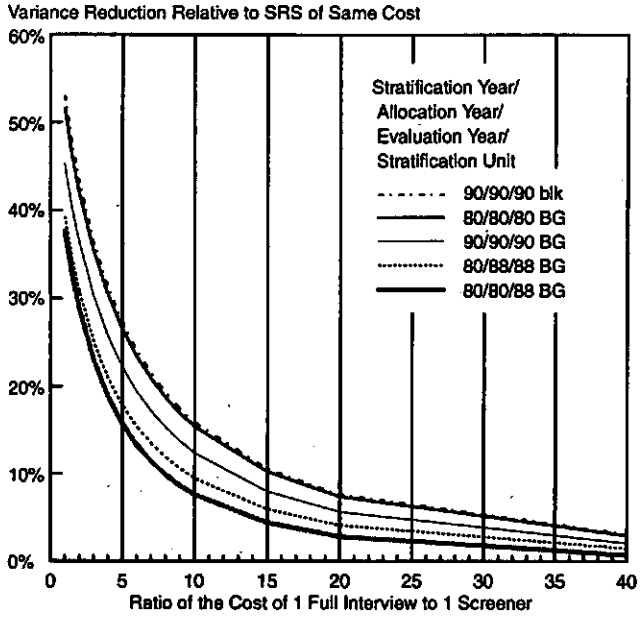


Figure 1. Variance Reduction from Geographic-based Oversampling for Blacks

Figure 1 summarizes the implications of the density data shown in Table 1 for oversampling blacks. This figure shows the substantial effect of c on the efficiency of geographic-based oversampling. For values of c beyond 20, the best way to sample the black population is probably just to screen an equi-probability sample.

The figure also illustrates the danger of relying upon the stratification data to evaluate the benefits of geographic-based oversampling. The 80/80/80 line shows the variance reductions that could be made if there were no change over time in the distribution of the black population across the density strata defined in terms of 1980 block group data. The 80/80/88 line shows the actual variance reductions that are possible in 1988 for the same strata and allocation. At $c = 5$, the variance reduction given a static distribution is 26 percent, while the variance reduction given observed changes in the distribution is just 16 percent. We examined whether allocating the sample across the old strata according to new distribution data could improve the actual variance reduction in 1988. The answer is yes, but not by much. The 80/88/88 shows the variance reductions that are possible using the 1988

distribution across the 1980 strata to guide the allocation for a survey conducted in 1988. At $c = 5$, the variance reduction given this allocation is 18 percent, a very modest improvement over the 16 percent variance reduction possible with the allocation guided by the old distribution. This led us to conclude that the major problem was the old stratification itself. By 1988, the extent of migration by the black population from block groups that were densely black in 1980 into block groups that had lower concentrations of black populations in 1980 was so great as to cut the variance reduction achievable through oversampling almost in half. The shift of the black population into block groups with lower concentrations of blacks in 1980 results in more sample blacks with large weights thus increasing the variability among weights which increases the variance. Nonetheless, the variance reductions indicated by the 80/80/88 line for $c < 10$ are certainly large enough to be useful.

Turning attention to the 1990 data in Figure 1, we observe that the 90/90/90 BG line is consistently several points below the 80/80/80 line, indicating that geographic oversampling at the block group level is likely to be slightly less useful during the 1990s than it was during the 1980s. This is a reflection of the slight reduction in segregation of the American black population in 1990 compared to 1980 noted above. On the other hand, the 90/90/90 blk line is almost exactly the same as the 80/80/80 line, indicating that the geographic oversampling at the block level can be expected to be as effective during the 1990s as it was at the block group level in the 1980s. Although data have not yet been collected on the distribution of the black population in the late 1990s across 1990 density strata, we would expect that migration has continued and that therefore the gains indicated by the 1990 lines should probably be reduced (along the general trend indicated by the 80/80/88 line) when projecting savings into the late 1990s and the first few years after 2000.

7. OVERSAMPLING HISPANICS

Table 2 shows various aspects of residential segregation for Hispanics in the U.S. that are important to know about when designing a population survey. Several points are interesting to note. First, it appears that Hispanics (unlike blacks) became slightly more segregated between 1980 and 1990. Other patterns, however, are similar for the black and Hispanic populations. In 1980, 30 percent of the Hispanic population lived in block groups that were 60 percent or more Hispanic. By 1988 these same block groups contained only

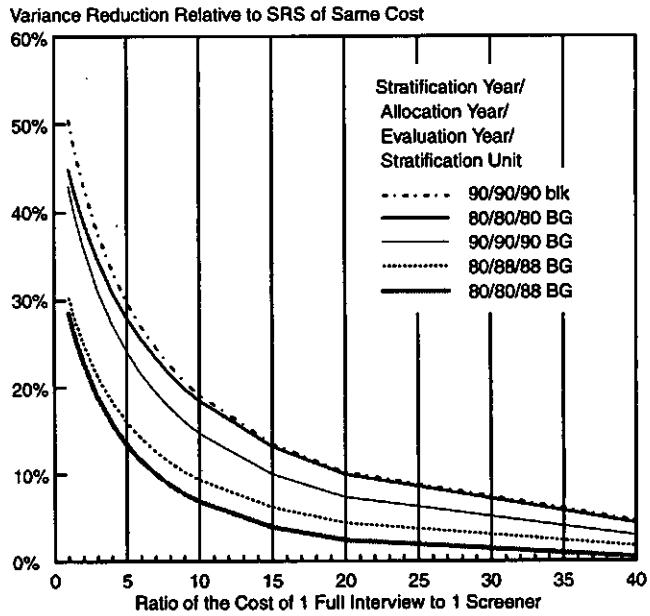


Figure 2. Variance Reduction from Geographic-based Oversampling for Hispanics

Table 2
Residential Clustering of Hispanics

Density stratum (Hispanics as a percent of the stratification unit in the year of stratification)	Percentage of Hispanics living in the stratum in the indicated year				Percentage of the total population living in the stratum in the indicated year			
Measurement year	1980	1988	1990	1990	1980	1988	1990	1990
Stratification year	1980	1980	1990	1990	1980	1980	1990	1990
Stratification unit	BG/ED	BG/ED	BG	Block	BG/ED	BG/ED	BG	Block
< 5%	14.8	29.3	10.6	6.6	76.8	79.8	68.4	68.9
5-10%	9.6	9.5	8.7	8.1	8.8	7.7	10.9	10.3
10-30%	22.6	21.2	22.8	22.1	8.5	7.4	11.8	11.5
30-60%	23.1	18.8	24.1	23.3	3.5	3.0	5.1	4.9
60-100%	30.0	21.2	33.9	39.8	2.4	2.0	3.8	4.4
Total populations (1000s)	14,609	19,393	22,354	22,354	226,546	240,876	248,710	248,710
Hispanics as percent of nation in measurement year	6.4	8.1	9.0	9.0				

Sources: 1980 Decennial Census (Westat tabulation)
1988 National Health Interview Survey (Westat tabulation)
1990 Decennial Census (Westat tabulation)

about 21 percent of the Hispanic population. In contrast, the percent of Hispanic population living in the 1980 block groups that were less than 5 percent Hispanic increased from 15 percent in 1980 to 29 percent in 1988. These changes reflect both a shift of the Hispanic between areas and the increase in the Hispanic population coming into the United States. The restratification of the Hispanic population using 1990 data shows patterns similar to the 1980 distribution patterns.

Figure 2 summarizes the implications of these segregation data on oversampling schemes. The curves show the same general patterns as the black curves. Geographic-based oversampling appears to be a useful tool for values of $c < 10$. Again though, it is important to be mindful of the effect of migration on the variance reduction. The gap between the 80/80/80 and 80/80/88 lines is greater for Hispanics than for blacks, particularly for $c < 5$. At present, we do not have a good basis for predicting whether this will be as true in the 1990s as it was in the 1980s.

8. OVERSAMPLING OTHER RACIAL MINORITIES

Tables 3 and 4 show segregation data for Asians and Pacific Islanders and for American Indians, Eskimos and Aleuts, respectively. Figures 3 and 4 show corresponding implications for oversampling these domains. Data from 1980 and 1988 were not tabulated for this work because the 1990 data are not encouraging for the inexpensive oversampling of these populations even with the use of stratification by density. The percent reductions in variance are quite large, greater than those for the black and Hispanic populations, since the amount of screening that would otherwise be required is much larger. However, the rarity of these populations in the U.S. means that very large screening samples are still required in order to get respectable interviewed sample sizes. For example, with a cost ratio of 3, even with geographic-based oversampling, it is necessary to screen 61,000 persons (or about 24,000 households) in order

Table 3
Residential Clustering of Asians and Pacific Islanders

Density stratum (Asians and Pacific Islanders as a percent of the 1990 block or block group in 1990)	Percentage of Asians and Pacific Islanders living in the stratum in 1990		Percentage of the total population living in the stratum in 1990	
Stratification unit:	BG	Block	BG	Block
< 5%	30.5	19.4	86.4	85.2
5-10%	17.2	17.7	7.2	7.4
10-30%	27.8	32.1	5.0	5.7
30-60%	14.6	18.0	1.0	1.3
60-100%	9.8	13.0	0.4	0.5
Total population (1000s)	6,968	6,968	248,710	248,710
Asians and Pacific Islanders as percent of nation in measurement year	2.8	2.8		

Sources: 1990 Decennial Census (Westat tabulation)

Table 4
Residential Clustering of American Indians, Eskimos and Aleuts

Density stratum (American Indians, Eskimos and Aleuts as a percent of the 1990 block or block group in 1990)	Percentage of American Indians, Eskimos and Aleuts living in the stratum in 1990		Percentage of the total population living in the stratum in 1990	
Stratification unit:	BG	Block	BG	Block
< 5%	50.3	34.6	98.3	97.4
5-10%	7.4	12.1	0.8	1.4
10-30%	12.4	15.9	0.6	0.8
30-60%	6.0	7.7	0.1	0.1
60-100%	23.8	29.6	0.2	0.2
Total population (1000s)	1,793	1,793	248,710	248,710
American Indians, Eskimos and Aleuts as percent of nation in measurement year	0.7	0.7		

Sources: 1990 Decennial Census (Westat tabulation)

to obtain a sample of American Indians, Eskimos and Aleuts with precision equal to a (theoretical) simple random sample of 1,000 persons from this domain. (Of course, to successfully screen 24,000 households, more housing units would have to be selected to allow for vacants and nonresponse). The comparable number for Asians and Pacific Islanders is 18,000 persons or roughly 7,000 households.

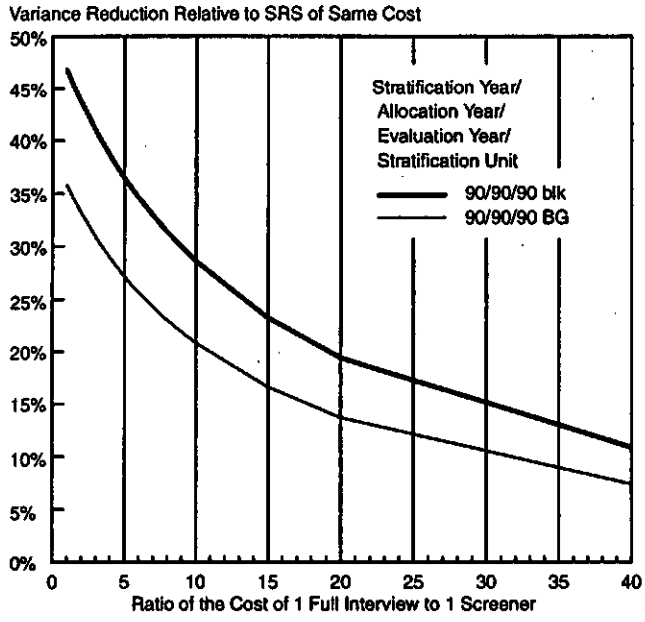


Figure 3. Variance Reduction from Geographic-based Oversampling for Asians and Pacific Islanders

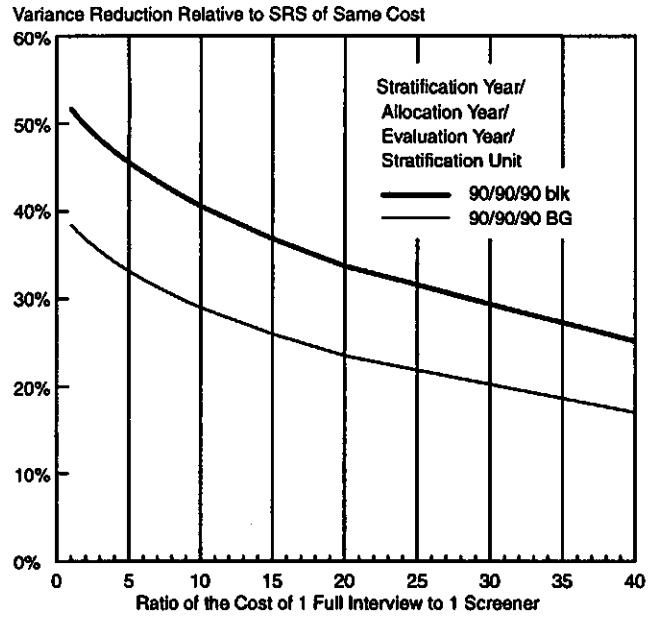


Figure 4. Variance Reduction from Geographic-based Oversampling for American Indians, Eskimos and Aleuts

9. OVERSAMPLING THE POOR

Table 5 shows the 1990 distribution of the low income population by block groups classified according to the proportion of low-income population in the BG. The BGs in each of the classes depends on the definition of low income. The figures shown in the table are the percentages of low-income persons in each class. Table 5 shows a rather flat distribution of low income among the classes for all three definitions in 1990. Data (not shown) from the 1970 decennial census and the Current Population Survey indicate that segregation of persons below the poverty level increased between 1970 and 1990 (Waksberg 1995), but the segregation is still far less than the segregation of racial and ethnic groups. The concentrations are somewhat greater for persons under 150 percent than for the other two definitions but, even for this group, it is considerably less than for racial and ethnic groups. As can be seen, with this definition, only about 25 percent of the poor live in BGs where 50 percent or more of the population is poor. The comparable percentages are 19 percent for persons below 125 percent of poverty and only 13 percent for persons below 100 percent of poverty. Such distributions imply that oversampling households in the strata with relatively high percentages of low-income persons will not be much better than oversampling and screening the entire sampling frame unless the full interview costs are only slightly higher than screening costs.

Figure 5 shows the ratio of the variance of the optimum sample to an SRS at the same cost, for statistics relating to the low-income populations. Interestingly, despite the greater concentration associated with the broadest definition of low

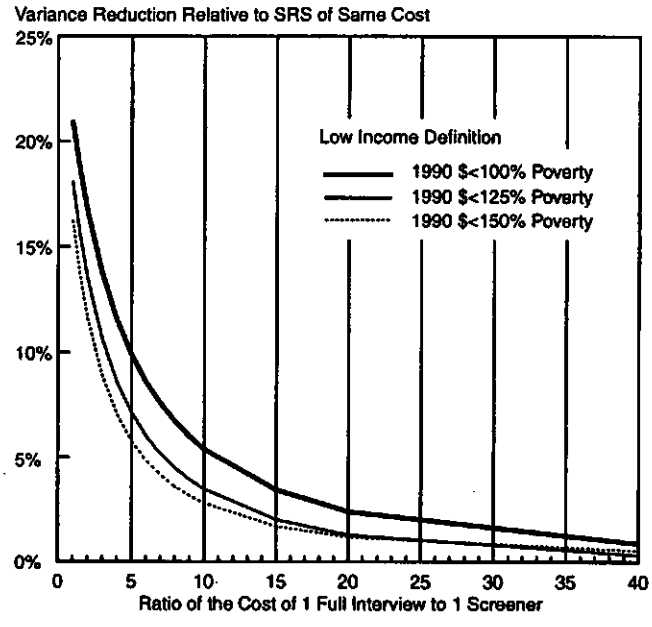


Figure 5. Variance Reduction from Geographic-based Oversampling for Persons with Low Income

Table 5
Residential Clustering of the Low Income Population

Density stratum (Persons with low income as a percent of 1990 block group in 1990 according to various definitions of low income)	Percentage of persons with low income living in the stratum in 1990			Percentage of the total population living in the stratum in 1990		
Low income definition:	$ < Poverty	$ < 125% of Poverty	$ < 150% of Poverty	$ < Poverty	$ < 125% of Poverty	$ < 150% of Poverty
< 5%	5.8	3.2	1.8	33.3	22.4	15.4
5-10%	12.3	8.3	5.7	22.1	19.7	16.7
10-20%	24.8	21.0	16.8	22.8	25.2	24.8
20-30%	19.8	20.2	19.2	10.7	14.4	16.8
30-40%	14.3	15.9	17.0	5.4	8.1	10.7
40-50%	10.0	12.2	13.7	2.9	4.8	6.7
50-100%	13.0	19.3	25.7	2.8	5.4	8.8
Total populations (1000s)	31,797	42,316	52,521	248,710	248,710	248,710
Persons with low income as percent of nation in measurement year	12.8	17.0	21.1			

Sources: 1990 Decennial Census (Westat tabulation of STF-3)

income, the reduction in variance for geographic-based oversampling is strongest for the narrowest definition because it requires more screening and thus has more to gain from a sampling strategy that reduces screening. For all three definitions, there appear to be moderate advantages to oversampling when c is under 3 or 4, about a 10 or 15 percent reduction in variances. When c is as large as 10, the gains are very slight, and there is virtually no advantage to oversampling BGs with high levels of poverty when c is 20 or larger. Of course, migration must be taken into account here as well, but we did not obtain the necessary data. Due to the effects of migration, the actual variance reductions will almost certainly be smaller than those shown in the chart. Furthermore, the income data in the 1990 Census are based on a one-sixth sample. The sample size in a typical block group was a little under 100 households. The classification of blocks according to percentage of low-income persons therefore has a fair amount of fuzziness to it, and many block groups will not be in the categories that Census data assign them, but in neighbouring classes, further weakening the variance reductions that can be achieved with geographic-based oversampling. As a result of these factors, it is unlikely that geographic-based oversampling will improve the efficiency. In fact, by mid-decade or later, it may actually result in an increase in variance. A related unpublished study by Waksberg in 1989 showed similar results when considering the possibility of merging ZIP-code level summary income data onto banks of telephone numbers used in RDD sampling. The gains achievable through stratification appear quite limited.

An examination of more detailed tables (not shown) indicates that the effectiveness is about the same for various types of geographic breakdowns, *e.g.*, states, large or small MSAs, central cities, suburban areas, and nonmetropolitan areas. Conclusions drawn from this analysis will thus approximately apply to subnational surveys.

However, geographic-based oversampling is an extremely effective tool for the low-income black and Hispanic populations. As shown in Table 6, blacks and Hispanics living in poverty are highly concentrated and others living in poverty are not. The left-hand side of Table 6 indicates the distribution of the poor black, Hispanic, and other populations across density strata defined in terms of poverty rates specific to the domain of interest. Interpreting one example from the left side, 32 percent of poor Hispanics lived in 1990 in block groups where the poverty rate for Hispanics was over 50 percent. The right hand side indicates the distribution of the poor black and Hispanic populations across density strata defined just in terms of the local concentrations of blacks or Hispanics without regard to income levels. Interpreting one example from the right side, 44.8 percent of poor Hispanics lived in 1990 in block groups where Hispanics constituted over 60 percent of the local population. From these numbers, we infer that over 90 percent of both poor blacks and poor Hispanics live in areas with above average concentrations of their respective racial/ethnic groups. This means that a sampling strategy that oversamples blocks with high black or Hispanic concentrations will automatically yield disproportionately large numbers of poor blacks and Hispanics. Furthermore, almost no poor blacks or poor Hispanics live in areas with low poverty rates for their groups. This stands in marked contrast to the patterns for poor people who are neither black nor Hispanic. It appears that many poor nonhispanic whites live in close proximity to more well-off whites, possibly because poverty tends to be a transitory phenomenon for them, or perhaps because they are retired and purchased their homes when they were in better circumstances.

Table 6
Residential Clustering of the Low Income Population by Race and Ethnicity

Density stratum (Poverty rate in 1990 for persons of the indicated race/ethnicity within the block group in 1990)	Percentage of persons with the indicated race/ethnicity and income below the poverty line living in the stratum in 1990			Density stratum (Indicated minority as a percent of 1990 block in 1990)	Percentage of persons with the indicated race/ethnicity and income below the poverty line living in the stratum in 1990		
	Domain				Domain		
	Blacks	Hispanics	Others		Blacks	Hispanics	Others
< 5%	0.6	0.6	10.4	< 5%	4.0	4.6	n/a
5-10%	2.2	2.4	19.6	5-10%	3.7	5.1	n/a
10-20%	8.8	11.0	32.6	10-30%	13.2	19.9	n/a
20-30%	13.8	17.0	18.1	30-60%	19.0	25.5	n/a
30-40%	17.0	19.3	9.0	60-100%	60.0	44.8	n/a
40-50%	17.3	17.7	4.6				
50-100%	40.4	32.0	5.6				
Total populations (1000s)	8,557	5,536	17,975	Total populations (1000s)	8,557	5,536	17,975

Sources: 1990 Decennial Census (Westat tabulation of STF-3)

10. SIMULTANEOUS OVERSAMPLING OF SEVERAL RACE-ETHNIC DOMAINS

In general, geographic-based oversampling can be used as easily and effectively for targeting multiple race-ethnic domains as for a single race-ethnic domain. In fact, the optimal sampling rates for the strata with high concentrations of each of the targeted domains will be about the same as if only it were being targeted. However, the overall level of screening will be increased since the number of areas with high sampling rates will increase with the number of targeted domains. Both these observations are due to the limited overlap between the highly segregated areas of the examined racial and ethnic minorities.

Table 7 presents some data on this subject from the 1990 Decennial Census. The only domains that overlap significantly in their concentrated areas are Hispanics and Asians and Pacific Islanders, and even that overlap only works one way. Since there are so many more Hispanics in the U.S. than Asians and Pacific Islanders, the proportion of Hispanics that live in blocks with Asian /Pacific Islander populations over 10 percent of the local population is only 13.7 percent while the percent of Asians and Pacific Islanders that live in blocks with Hispanic populations over 10 percent of the local population is a high 40.8 percent. The practical significance of this particular overlap is probably slight, however, since it would take such a large screening sample (both in and out of highly concentrated areas) to find enough Asians and Pacific Islanders to meet moderate precision requirements that such

Table 7
Residential Mixing of Minorities

Density stratum (Indicated minority as a percent of 1990 block in 1990)	Percentage of blacks living in the stratum in 1990			Percentage of Hispanics living in the stratum in 1990			Percentage of Asians and Pacific Islanders living in 1990			Percentage of American Indians, Eskimos and Aleuts living in 1990		
	Stratification domain			Stratification domain			Stratification domain			Stratification domain		
	Hispanic	Asian and Pacific Islander	American Indian, Eskimo and Aleut	Black	Asian and Pacific Islander	American Indian, Eskimo and Aleut	Black	Hispanic	American Indian, Eskimo and Aleut	Black	Hispanic	Asian and Pacific Islander
< 10%	79.2	95.4	99.6	73.4	86.3	99.1	78.9	59.2	99.6	85.9	81.4	95.1
10-30%	12.7	3.8	0.3	15.5	10.7	0.8	15.2	26.9	0.4	8.2	12.3	3.9
30-60%	5.8	0.7	0.0	7.4	2.5	0.1	4.2	10.8	0.0	3.3	4.5	0.8
60-100%	2.2	0.1	0.0	3.6	0.5	0.1	1.6	3.2	0.0	2.5	1.8	0.2

Sources: 1990 Decennial Census (Westat tabulation)

a screening sample would probably find enough Hispanics without resorting to disproportionate allocation of the sample to blocks with higher concentrations of Hispanics.

11. CONCLUSIONS

For household surveys in the U.S., geographic-based oversampling using data from the most recent decennial census is a useful sampling strategy for improving the precision of statistics about the black and Hispanic populations provided that the cost of full interviews is less than 5 to 10 times the cost of screener interviews. It is also a useful strategy for improving the precision of statistics about the Asian/Pacific Islander and American Indian/Eskimo/Aleut populations, even at very high ratios of the cost of full interviews to the cost of screener interviews.

However, this does not mean that a survey of reasonable cost can be designed to simultaneously provide highly precise statistics about all these domains while maintaining desired precision levels for the total population. Most demographic surveys require reasonable precision for both targeted domains and for the total population. Shifting some portion of the full interviews from the white nonhispanic population to the other domains is bound to decrease the precision of statistics about the total population. It is generally useful to strike a balance between precision attained for subpopulations and the total population. The point of this observation is merely that geographic-based oversampling does not obviate the need to select very large samples and conduct many screening interviews when trying to obtain precise statistics about rare domains at the lowest possible cost. Furthermore, precise statistics about rare domains will continue to be expensive even when using geographic-based oversampling.

For surveys of low-income persons, only small gains are possible with geographic-based oversampling, and those only when the cost of a full interview is only a few times larger than the cost of screening and dropping a household. Most of these gains are likely to disappear when deterioration over time is taken into account. In fact, by the middle of a decade or later, when Census data become seriously outdated, there is the distinct possibility that geographic-based oversampling could reduce efficiency rather than improve it because of migration of the poor and sampling error in measuring poverty at the block group level. Geographic-based oversampling is a useful tool, however, when the focus of interest is on the black or Hispanic poor.

ACKNOWLEDGMENTS

This research was conducted by Westat Inc. under contract 200-89-7021 sponsored by the National Center for Health Statistics, Centers for Disease Control and Prevention. David Judkins and James Massey participated in the project while they were with Westat and NCHS, respectively. The authors would like to gratefully acknowledge the programming contributions of John Edmonds and Robert Dymowski of Westat and to thank the referees for their useful comments and suggestions on an earlier version of the paper.

REFERENCES

JUDKINS, D., MASSEY, J., and WAKSBERG, J. (1992). Patterns of residential concentration by race and Hispanic origin. *Proceedings of the Social Statistics Section, American Statistical Association,* 51-60.

KALTON, G., and ANDERSON, D.W. (1986) Sampling rare populations. *Journal of the Royal Statistical Society, Series A,* 149, 1, 65-82.

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

MASSEY, J., JUDKINS, D., and WAKSBERG. J. (1993). Collecting health data on minority populations in a national survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association,* 75-84.

UNITED NATIONS (1993). Sampling Rare and Elusive Populations. Department for Economic and Social Information and Policy Analysis, Statistical Division, National Household Survey Capability Programme. New York.

WAKSBERG, J. (1973). The effect of stratification with differential sampling rates on attributes of subsets of the population. *Proceedings of the Social Statistics Section, American Statistical Association,* 429-434.

WAKSBERG, J. (1995). Distribution of poverty in Census block groups (BG's) and implications for sample design. *Proceedings of the Section on Survey Research Methods, American Statistical Association,* 497-502.

WAKSBERG, J., and MOHADJER, L. (1991). Automation of within-household sampling. *Proceedings of the Section on Survey Research Methods, American Statistical Association,* 350-355.

Survey Methodology, June 1997
Vol. 23, No. 1, pp. 73-75
Statistics Canada

A Modified Random Groups Standard Error Estimator

WILLARD C. LOSINGER[1]

ABSTRACT

The standard error estimation method used for sample data in the U.S. Decennial Census from 1970 through 1990 yielded irregular results. For example, the method gave different standard error estimates for the "yes" and "no" response for the same binomial variable, when both standard error estimates should have been the same. If most respondents answered a binomial variable one way and a few answered the other way, the standard error estimate was much higher for the response with the most respondents. In addition, when 100 percent of respondents answered a question the same way, the standard error of this estimate was not zero, but was still quite high. Reporting average design effects which were weighted by the number of respondents that reported particular characteristics magnified the problem. An alternative to the random groups standard error estimate used in the U.S. census is suggested here.

KEY WORDS: Census; Variance estimation; Random groups; Design effect.

1. INTRODUCTION

During the 1990 Decennial Census, all respondents were asked to provide information on certain data items (called 100-percent data). Most respondents provided this information on the census short form. In addition, a systematic sample (ranging from one-eighth to one-half, but averaging about one-sixth) of respondents provided information for more data items (sample data) on the census long form.

Rather than providing standard error estimates for each published sample data estimate, the Census Bureau published tables of generalized design effects. For any sample data estimate, data users were instructed to create a standard error assuming simple random sampling (either using the standard formula or from a table) and a one-in-six sampling rate. Then, data users were to multiply this standard error by a generalized design effect (provided in another table). The table of generalized design effects listed design effects by data item type and percent of persons or housing units included in the sample (Table 1 provides the design effects published for 1990 U.S. census sample data for Vermont). For example, for all published sample estimates that dealt with occupation, a data user would find four generalized design effects for occupation: one for each of four sampling rate categories for persons in the report. To estimate the standard error for the number of teachers in a published report, a data user would multiply the simple-random-sampling standard error (assuming a one-in-six sampling rate, derived from the formula or table of standard errors) by the design effect for occupation data items for the reported sampling rate. The data user could then use the estimated number of teachers and standard error to construct a confidence interval. More details on the use of the table of design effects are available in the Accuracy of the Data section for any sample data product (U.S. Bureau of the Census 1993, for example).

2. ESTIMATION OF STANDARD ERRORS

A random-groups approach was used to estimate standard errors for the census sample data. The United States was divided into just over 60,000 distinct areas (called weighting areas--areas for which sample weights were derived). For each weighting area, sample units (a sample unit being either a housing unit or a person residing in a group quarter) were assigned systematically among 25 random groups. Thus, it was thought that each random group so formed met the requirement of having approximately the same sampling design as the parent sample (Wolter 1985).

For each of the 25 random groups, a separate estimate of the total for each of 1,804 sample data items was computed by multiplying the weighted count for the sample data item within the random group by 25. For each data item for which the total number of people with a particular characteristic was estimated from the sample data, the random-groups standard error estimate was then computed from the 25 different estimates of the total from the random groups:

$$S_{RG} = \sqrt{(1 - n/N)\sum_{i=1}^{25} \frac{(\hat{Y}_i - \hat{Y})^2}{24}}$$

where n represents the unweighted number of persons in the sample within the weighting area; N represents the census count of persons within the weighting area; $\hat{Y}_i$ represents the estimate of the total for the data item achieved by multiplying the weighted count for the data item within the i-th random group by 25; and $\hat{Y}$ is the weighted count for the data item (*i.e.*, the sample estimate) within the weighting area.

[1] Willard C. Losinger, U.S. Department of Agriculture: APHIS:VS, CEAH, 555 South Howes Street, Suite 200, Fort Collins, CO 80521, U.S.A.

Table 1
Design Effects Published for 1990 U.S. Census Sample Data for Vermont

Characteristic	Percent of persons or housing units in sample			
	< 15%	15 - 30%	30 - 45%	≥ 45%
Age	1.2	1.0	0.6	0.5
Sex	1.2	1.0	0.6	0.5
Race	1.2	1.0	0.6	0.5
Hispanic origin (of any race)	1.2	1.0	0.6	0.5
Marital status	1.1	0.9	0.6	0.5
Household type and relationship	1.2	1.0	0.6	0.5
Children ever born	2.5	2.2	1.3	1.2
Work disability and mobility limitation status	1.2	1.0	0.6	0.5
Ancestry	1.8	1.5	1.0	0.8
Place of birth	1.9	1.6	1.0	0.9
Citizenship	1.7	1.4	1.0	0.8
Residence in 1985	1.9	1.7	1.0	0.9
Year of entry	1.3	1.0	0.6	0.5
Language spoken at home and ability to speak English	1.6	1.3	0.9	0.7
Educational attainment	1.3	1.1	0.6	0.5
School enrollment	1.6	1.4	1.0	0.8
Type of residence (urban/rural)	1.7	1.7	1.4	1.4
Household type	1.2	1.0	0.6	0.5
Family type	1.1	1.0	0.6	0.5
Group quarters	1.0	1.1	0.9	0.8
Subfamily type and presence of children	1.1	0.9	0.5	0.5
Employment status	1.2	1.0	0.6	0.5
Industry	1.2	1.0	0.6	0.5
Occupation	1.2	1.0	0.6	0.5
Class of worker	1.2	1.0	0.6	0.5
Hours per week and weeks worked in 1989	1.4	1.2	0.7	0.6
Number of workers in family	1.3	1.1	0.7	0.6
Place of work	1.4	1.2	0.8	0.6
Means of transportation to work	1.4	1.2	0.7	0.6
Travel time to work	1.3	1.1	0.6	0.5
Private vehicle occupancy	1.4	1.2	0.7	0.6
Time leaving to go to work	1.2	1.0	0.6	0.5
Type of income in 1989	1.3	1.1	0.6	0.5
Household income in 1989	1.1	1.0	0.6	0.5
Family income in 1989	1.1	1.0	0.6	0.5
Poverty status in 1989 (persons)	1.5	1.2	0.7	0.7
Poverty status in 1989 (families)	1.1	0.9	0.5	0.5
Armed forces and veteran status	1.4	1.1	0.7	0.6

Source: U.S. Bureau of the Census (1993). 1990 Census of Population: Social and Economic Characteristics: Vermont. Report Number 1990 CP-2-47. Page C-11.

A standard error based upon simple random sampling and a one-in-six sampling rate was computed thus:

$$S_{SRS} = \sqrt{5\ \hat{Y}(1 - \hat{Y}/N)}$$

developed from standard formulas displayed in Cochran (1977).

For each data item within the weighting area, a design effect was computed as the ratio of the S_{RG} to S_{SRS}:

$$F = \frac{S_{RG}}{S_{SRS}}.$$

For a state report of sample data, the design effects for each data item were averaged across the weighting areas in the state. Then, a generalized design effect for each data item type (for example, all data items that dealt with occupation) was computed. The generalized design effect was weighted in favor of data items that had higher population estimates. Details on most of the procedures followed are available in a Census Bureau document (U.S. Bureau of the Census 1991). The same basic method was also used for sample data products in both the 1970 and 1980 census.

3. A HYPOTHETICAL EXAMPLE OF RANDOM GROUPS

Table 2 presents a hypothetical example of data that might have arisen from the random-groups method. For a weighting area in Vermont, weighted counts of whites and blacks are listed for the 25 random groups. In this hypothetical weighting area, there are no persons of other race. The standard errors assuming simple random sampling are the same for whites and blacks (as one would expect for a binomial variable). However, S_{RG} is much higher for the estimate of whites than the estimate of blacks. And, the design effect is nearly five times higher for the estimate of whites than the estimate of blacks. Since the generalized design effect computed for groups of data items was weighted in favor of data items that had higher population estimates, the generalized design effect computed for race for the state of Vermont was quite high.

Data on race were frequently included in 1990 U.S. census sample data products. Because race was asked of every census respondent (*i.e.*, it was a census 100-percent data item), and because the weighting process used by the Census Bureau effectively forced the sample estimates by race to match the 100-percent Census counts by race, the standard errors for estimates of race probably should have been considered to be zero. However, generalized design effects were still published by race, although set to arbitrary constants for all reports (rather than as computed by this method).

4. A MODIFIED APPROACH TO THE RANDOM GROUPS METHOD

A slight modification of the random groups method (essentially applying a ratio-estimation technique) can achieve much more satisfactory results in the estimation of standard errors. Rather than using $\hat{Y}_i$ as defined above for the estimate of the total for the i-th random group, one could instead use

$$\hat{L}_i = N X_i / W_i$$

Table 2
Hypothetical example of data that could have resulted from the Random Groups method used to estimate standard errors for census sample data.
For a weighting area in Vermont, people are asked their race. A few (110) are black; most (2,518) are white. A sampling rate of one-in-six is assumed ($N = 2{,}628$, $n = 438$).

Random Group	Weighted count of blacks*	Weighted count of whites*	Total weighted population count #
1	10	90	100
2	0	100	100
3	0	110	110
4	0	140	140
5	5	70	75
6	8	50	58
7	12	103	115
8	20	60	80
9	0	65	65
10	0	100	100
11	0	125	125
12	0	130	130
13	10	90	100
14	0	100	100
15	0	110	110
16	0	140	140
17	5	70	75
18	8	52	60
19	12	103	115
20	20	160	180
21	0	65	65
22	0	100	100
23	0	125	125
24	0	130	130
25	0	130	130
Sum of weighted counts ($\hat{Y}$)	110	2,518	2,628
S_{RG}	145.98	687.96	
S_{SRS}	22.96	22.96	
F	6.36	29.96	

* The first 25 figures in this column represent X_i for the i-th random group under the modified random groups method. Multiplying the figure by 25 yields $\hat{Y}_i$ for the random groups method employed by the U.S. Bureau of the Census.

\# The first 25 figures in this column represent W_i under the modified random groups method.

where X_i represents the weighted count for the data item within the i-th random group, W_i is the weighted count of all persons in the i-th random group, and N represents the census count of persons in the weighting area. The modified random groups standard error estimate is then

$$S_L = \sqrt{(1 - n/N)\sum_{i=1}^{25} \frac{(\hat{L}_i - \hat{Y})^2}{24}}.$$

Using this method, S_L is 160.78 for both blacks and whites in the hypothetical weighting area of Table 1 (close to the value of S_{RG} for blacks). In this case, the requirement for standard error estimates for both responses for a binomial variable to be identical is met. Moreover, if all sample units have the same response for some variable, S_L becomes zero, whereas S_{RG} only becomes zero when each random group has the same weighted count.

This modified standard error estimation procedure could be useful for researchers who do not have access to any of the many computer programs now available for computing estimates from sample data (such as SUDAAN, STATA, PC-CARP, VPLX, *etc.*). In addition, the U.S. Bureau of the Census ought to consider modifying its approach for estimating standard errors for sample data from the 2000 census. Moreover, with the U.S. Bureau of the Census' current emphasis on quality management, the U.S. Bureau of the Census may wish to poll users of sample data products to determine how useful the presentation of standard errors (through design effects) was to them, and involve a number of the data users in improving the presentation of standard errors for the next census.

REFERENCES

COCHRAN, W.G. (1977). *Sampling Techniques* (third edition). New York: John Wiley & Sons.

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

U.S. BUREAU OF THE CENSUS (1991). Computer Specifications for the 1990 Decennial Census Variance Estimation Operation. STSD Decennial Census Memorandum Series #Z-65.

U.S. BUREAU OF THE CENSUS (1993). Appendix C. Accuracy of the Data. Pp. C-1 to C-11 in 1990 Census of Population: Social and Economic Characteristics: Vermont. Bureau of the Census Document 1990 CP-2-47.

Survey Methodology, June 1997
Vol. 23, No. 1, pp. 77-78
Statistics Canada

A Simple Derivation of the Linearization of the Regression Estimator

KEES ZEELENBERG[1]

ABSTRACT

We show how the use of matrix calculus can simplify the derivation of the linearization of the regression coefficient estimator and the regression estimator.

KEY WORDS: Matrix calculus; Regression estimator; Taylor expansion.

1. INTRODUCTION

Design-based sampling variances of non-linear statistics are often calculated by means of a linear approximation obtained by a Taylor expansion; examples are the variances of the general regression coefficient estimator and the regression estimator. The linearizations usually need some complicated differentiations. The purpose of this paper is to show how matrix calculus can simplify these derivations, to the extent that even the Taylor expansion of the regression coefficient estimator can be derived in one line, which should be compared with the nearly one page that Särndal *et al.* (1992, p. 205-206) need. To be honest, the use of matrix calculus requires some more machinery to be set up, which is not needed for traditional methods. However this set-up can be regarded as an investment; once it has been learned, it can be used fruitfully in many other applications. After this paper had been written, Binder (1996) appeared, in which similar techniques are used to derive variances by means of linearization. The present paper can be seen as a pedagogical note, in which the use of differentials is exposed.

2. MATRIX DIFFERENTIALS

2.1 Introduction

We will use the matrix calculus by means of differentials, as set out by Magnus and Neudecker (1988); this calculus differs somewhat from the usual methods, which focus on derivatives instead of differentials. Therefore in this section we will briefly describe the definitions and properties of differentials (see Zeelenberg 1993, for a more extensive survey). We first define differentials for vector functions, and then generalize to matrix functions.

2.2 Vector Functions

Let f be a function from an open set $S \subset \mathbb{R}^m$ to $\mathbb{R}^n$; let x_0 be a point in S. The function f is *differentiable* at x_0 if there exists a real $n \times m$-matrix A, depending on x_0, such that for any $u \in \mathbb{R}^m$ for which $x_0 + u \in S$, there holds

$$f(x_0 + u) = f(x_0) + A_{x_0} u + o(u), \tag{1}$$

where $o(u)$ is a function such that $\lim_{|u| \to 0} |o(u)|/|u| = 0$; the matrix A is called the *first derivative* of f at x_0; it is denoted as $Df(x_0)$ or $\partial f/\partial(x')|_{x=x_0}$. The derivative Df is equal to the matrix of partial derivatives, *i.e.*, $Df(x)_{ij} = \partial f_i/\partial x_j$. The linear function $df_{x_0} : \mathbb{R}^m \mapsto \mathbb{R}^n$ defined by $df_{x_0} : u \mapsto A_{x_0} u$ is called the *differential* of f at x_0. Usually we write dx instead of u so that $df_{x_0}(dx) = A_{x_0} dx$. From (1) we see that the differential corresponds to the linear part of the function, which can also be written as

$$y - y_0 = A_{x_0}(x - x_0),$$

where $y_0 = f(x_0)$. Therefore the differential of a function is the linearization of the function: it is the equation of the hyperplane through the origin that is parallel to the hyperplane tangent to the graph of f at x_0; so the linearized function can be written as

$$f(x) \doteq f(x_0) + A_{x_0}(x - x_0). \tag{2}$$

Alternatively, if B is a matrix such that $df_{x_0}(dx) = B dx$, then B is the derivative of f at x_0 and contains the partial derivatives of f at x_0. This one-to-one relationship between differentials and derivatives is very useful, since differentials are easy to manipulate.

Finally, we usually omit the subscript 0 in x_0, so that we write $df = A_x dx$.

2.3 Matrix Functions

A matrix function F from an open set $S \subset \mathbb{R}^{m \times n}$ to $\mathbb{R}^{p \times q}$ is differentiable if vec F is differentiable. The derivative DF is the derivative of vec F with respect to vec X, and is also denoted by $\partial \text{ vec } F/\partial(\text{vec } X)'$. The differential dF is the matrix function defined by vec $dF_{X_0}(U) = A_{X_0}$ vec U.

[1] Kees Zeelenberg, Department of Statistical Methods, Statistics Netherlands, P.O. Box 4000, 2270 JM Voorburg, The Netherlands.

2.4 Properties of Differentials

Let A be a matrix of constants, F and G differentiable matrix functions, and α a real scalar. Then the following properties are easily proved:

$$dA = 0, \tag{3}$$

$$d(\alpha F) = \alpha\, dF, \tag{4}$$

$$d(F + G) = dF + dG, \tag{5}$$

$$d(FG) = (dF)G + F(dG), \tag{6}$$

$$dF^{-1} = -F^{-1}(dF)F^{-1}. \tag{7}$$

The last property can be proved by taking the differential of $FF^{-1} = I$ and rearranging.

3. LINEARIZATION OF THE REGRESSION COEFFICIENT ESTIMATOR

The π-estimator (Horvitz-Thompson estimator) of the finite population regression coefficient (*cf.* Särndal *et al.* 1992, section 5.10) is

$$\hat{B} = \hat{T}^{-1}\hat{t}, \tag{8}$$

where

$$\hat{T} = \sum_{k \in s} \frac{x_k x_k'}{\pi_k},$$

$$\hat{t} = \sum_{k \in s} \frac{x_k y_k}{\pi_k},$$

y_k is the variable of interest for individual k, x_k is the vector with the auxiliary variables for individual k, π_k is the inclusion probability for individual k, and s denotes the sample.

Taking the total differential of (8), using properties (6) and (7), and evaluating at the point where $\hat{T} = T$, $\hat{t} = t$, we get

$$d\hat{B} = -T^{-1}(d\hat{T})T^{-1}t + T^{-1}(d\hat{t}). \tag{9}$$

Because of the connection between differentials and linear approximation, as given in equation (2), it immediately follows that (9) corresponds to the linearization of the regression coefficient estimator:

$$\hat{B} \doteq B - T^{-1}(\hat{T}-T)T^{-1}t + T^{-1}(\hat{t}-t) = B + T^{-1}(\hat{t} - \hat{T}B),$$

where $B = T^{-1}t$.

4. LINEARIZATION OF THE REGRESSION ESTIMATOR

The regression estimator of a population total is (*cf.* Särndal *et al.* 1992, section 6.6)

$$\hat{t}_{yr} = \hat{t}_{y\pi} + (t_x - \hat{t}_{x\pi})'\hat{B}, \tag{10}$$

where $\hat{t}_{y\pi}$ is the π-estimator of the variable of interest, t_x is the vector with the population totals of the auxiliary variables, $\hat{t}_{x\pi}$ is the vector with the π-estimators of the auxiliary variables, and $\hat{B}$ is the estimator of the regression coefficient of the auxiliary variables on the variable of interest. Taking the total differential of (10), using properties (3) and (6), and evaluating at the point where $\hat{t}_{y\pi} = t_y$, $\hat{t}_{x\pi} = t_x$, and $\hat{B} = B$, we get the linear approximation of the regression estimator

$$d\hat{t}_{yr} = d\hat{t}_{y\pi} - (d\hat{t}_{x\pi})'B,$$

so that

$$\hat{t}_{yr} \doteq t_y + \hat{t}_{y\pi} - t_y + (t_x - \hat{t}_{x\pi})'B = \hat{t}_{y\pi} + (t_x - \hat{t}_{x\pi})'B.$$

Note that for the linearization of the regression estimator we do not need that of the regression coefficient estimator B.

ACKNOWLEDGEMENTS

I wish to thank Jeroen Pannekoek, Jos de Ree, Robbert Renssen, two referees, and an Associate Editor for their comments. The views expressed in this article are those of the author and do not necessarily reflect the policy of Statistics Netherlands.

REFERENCES

BINDER, D.A. (1996). Linearization methods for single phase and two-phase samples: a cookbook approach. *Survey Methodology*, 22, 17-22.

MAGNUS, J.R., and NEUDECKER, H. (1988). *Matrix Differential Calculus*. New York: Wiley.

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

ZEELENBERG, C. (1993). *A Survey of Matrix Differentiation*. Research Paper, Department of Statistical Methods, Statistics Netherlands, Voorburg.

CONTENTS TABLE DES MATIÈRES

Volume 25, No. 1, March/mars 1997

Volume 25, No. 2, June/juin 1997

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 4, 1996

Volume 13, Number 1, 1997

Volume 13, Number 2, 1997

All inquires about submissions and subscriptions should be directed to the Chief Editor:
Lars Lyberg, R&D Department, Statistics Sweden, Box 24 300, S - 104 51 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.