Article
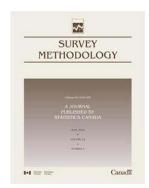
Sparse and efficient replication variance estimation for complex surveys

by Jae Kwang Kim and Changbao Wu

Statistics Canada   Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

email at infostats@statcan.gc.ca,

telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following toll-free numbers:

• Statistical Information Service	1-800-263-1136
• National telecommunications device for the hearing impaired	1-800-363-7629
• Fax line	1-877-287-4369

Depository Services Program

• Inquiries line	1-800-635-7943
• Fax line	1-800-565-7757

To access this product

This product, Catalogue no. 12-001-X, is available free in electronic format. To obtain a single issue, visit our website, www.statcan.gc.ca, and browse by "Key resource" > "Publications."

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "About us" > "The agency" > "Providing services to Canadians."

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Standard symbols

The following symbols are used in Statistics Canada publications:

.	not available for any reference period
..	not available for a specific reference period
...	not applicable
0	true zero or a value rounded to zero
0^s	value rounded to 0 (zero) where there is a meaningful distinction between true zero and the value that was rounded
p	preliminary
r	revised
x	suppressed to meet the confidentiality requirements of the *Statistics Act*
E	use with caution
F	too unreliable to be published
*	significantly different from reference category ($p < 0.05$)

Sparse and efficient replication variance estimation for complex surveys

Jae Kwang Kim and Changbao Wu[1]

Abstract

It is routine practice for survey organizations to provide replication weights as part of survey data files. These replication weights are meant to produce valid and efficient variance estimates for a variety of estimators in a simple and systematic manner. Most existing methods for constructing replication weights, however, are only valid for specific sampling designs and typically require a very large number of replicates. In this paper we first show how to produce replication weights based on the method outlined in Fay (1984) such that the resulting replication variance estimator is algebraically equivalent to the fully efficient linearization variance estimator for any given sampling design. We then propose a novel weight-calibration method to simultaneously achieve efficiency and sparsity in the sense that a small number of sets of replication weights can produce valid and efficient replication variance estimators for key population parameters. Our proposed method can be used in conjunction with existing resampling techniques for large-scale complex surveys. Validity of the proposed methods and extensions to some balanced sampling designs are also discussed. Simulation results showed that our proposed variance estimators perform very well in tracking coverage probabilities of confidence intervals. Our proposed strategies will likely have impact on how public-use survey data files are produced and how these data sets are analyzed.

Key Words: Bootstrap; Calibration; Jackknife; Linearization method; Replication weights; Sampling design; Spectral decomposition.

1 Introduction

Variance estimation is an important practical problem in sample surveys. In addition to analytic use of variances such as testing statistical hypotheses and constructing confidence intervals, variance estimation can also be used to provide descriptive measures on the accuracy of survey estimates and the efficiency of the given sampling design. There are two types of commonly used techniques for variance estimation under the design-based framework. The first is called the linearization method, which uses the standard variance formula applied either directly to the estimator if the parameter is a population total or to the linearized one-step Taylor

1. Jae Kwang Kim, Department of Statistics, Iowa State University, Ames IA 50011-1210. E-mail: jkim@iastate.edu; Changbao Wu, Department of Statistics and Actuarial Science, University of Waterloo, Waterloo ON N2L 3G1. E-mail: cbwu@uwaterloo.ca.

series expansion of the estimator if the parameter is a nonlinear function of one or several population totals. The second is called the replication method, which constructs variance estimators in a simple systematic way using multiple sets of replication weights along with the original survey data set.

Replication variance estimation techniques have become very popular for design-based inferences using complex survey data. Some early practices using replication weights go back to 1970s at the U.S. Bureau of the Census, Bureau of Labor Statistics and Westat (Dippo, Fay and Morganstein 1984). It is now a routine practice for survey organizations to provide replication weights together with survey data. The most attractive feature of this approach is that it works the same way regardless of the complexity of the parameter. For parameters that are smooth functions of population means or totals, the "linearization" step has been automatically built into the estimation process and computation of partial derivatives involved in the Taylor series expansion is not required. It is extremely user-friendly for multi-purpose data analyses once the survey data set is released together with replication weights. Furthermore, the use of replication methods reduces concerns on confidentiality issues since detailed design information such as stratum or cluster identifier is not released (Lu and Sitter 2008).

Replication weights are typically constructed by the bootstrap, the jackknife or the balanced repeated replication (BRR) methods. Rust and Rao (1996), Shao (1996, 2003) and Wolter (2007) provided excellent overviews on the topic. There are three major issues in the construction of replication weights: *validity, efficiency* and *sparsity*. Validity refers to the asymptotic unbiasedness of replication variance estimators under the given sampling design. The asymptotic unbiasedness of an estimator is generally a weaker concept than the estimator being consistent. If the coefficient of variation of the variance estimator goes to zero, then the asymptotically unbiased variance estimator is also consistent. Efficiency is measured by the relative performance of the replication variance estimator to the standard linearization variance estimator which is

viewed as fully efficient. Sparsity refers to the number of sets of replication weights required to achieve fully efficient variance estimation.

Validity of replication variance estimators was discussed by Krewski and Rao (1981), Shao and Tu (1995) and Fuller (2009a), among others. Efficiency and stability of replication variance estimators were discussed by Rust and Kalton (1987) and Jang and Eltinge (2009). For sparsity, Kott (2001) considered using delete-a-group jackknife to achieve sparsity under certain designs, and Lu, Brick and Sitter (2006) also discussed combining strata for sparse replication variance estimation.

Most replication methods discussed in the literature are only valid for certain sampling designs. For example, the jackknife method is commonly used for stratified random sampling (Krewski and Rao 1981). The bootstrap method has several popular procedures, including the without-replacement bootstrap method (Gross 1980; McCarthy and Snowden 1985), the re-scaling bootstrap method (Rao and Wu 1988; Preston 2009) and the mirror-match bootstrap method (Sitter 1992). These procedures, however, are only applicable for certain types of sampling designs.

The sparsity of a replication method depends on how the replication weights are constructed. The number of sets of the jackknife replication weights is related to the number of units in the sample and can be very large if the sample size is large. Bootstrap methods typically require at least 1,000 sets of replication weights in order to achieve the desired level of efficiency. As a compromise, most survey organizations provide 500 sets of bootstrap weights alongside the main survey variables. The resulting data sets are still too big for data users to have visual checks and can be very cumbersome to manipulate in practice.

This paper presents methods for constructing efficient and sparse replication weights for variance estimation under the design-based framework. By maintaining full efficiency of the resulting variance estimator for key variables with a smaller number of sets of replication weights, our methods address one of the major tasks at the data file preparation stage and can

easily be applied by survey runners to reduce the burden of data users in dealing with excessively large data files. A major limitation of our proposed method is that it does not directly handle situations where design weights are adjusted for nonresponse or calibrated to known auxiliary population information.

In Section 2, we present a general procedure for constructing replication weights based on the method of Fay (1984) and Fay and Dippo (1989), which provides fully efficient replication weights for arbitrary sampling designs. In Section 3, we discuss two strategies, random sampling and calibration weighting, for constructing sparse replication weights. By using a novel application of the calibration technique, our proposed methods allow the use of a small number of sets of replication weights while the resulting replication variance estimators remain efficient. In Section 4, some asymptotic theory for the validity of the replication variance estimator is presented. In Section 5, extensions to some balanced sampling designs are discussed. In Section 6, we report results from a simulation study, using real data from Statistics Canada's Family Expenditure Survey, to evaluate the effectiveness of the proposed strategies for replication variance estimation. Some concluding remarks are given in Section 7.

2 A general procedure for constructing fully efficient replication weights

In principle, we can construct replication weights for any measurable sampling design, using the method outlined in Fay (1984) and Fay and Dippo (1989), such that the resulting replication variance estimators are algebraically equivalent to the standard linearization variance estimators.

Let $\mathcal{U} = \{1, 2, \ldots, N\}$ be the set of N units in the finite population and $\mathcal{S} = \{1, 2, \ldots, n\}$ be the set of n units in the sample, selected according to a probability sampling design. Let $w_i = 1 / \pi_i$ be the basic design weight, where $\pi_i = P(i \in \mathcal{S})$ is the first order inclusion probability for unit i.

Let y_i be the value of the study variable y for unit i and $t_y = \sum_{i=1}^{N} y_i$ be the population total of interest. The Horvitz-Thompson estimator of t_y is given by

$$\hat{t}_y = \sum_{i \in \mathcal{S}} w_i y_i. \tag{2.1}$$

The estimator $\hat{t}_y$ given in (2.1) is also called the expansion estimator, with the basic design weight w_i denoting the number of units in the population represented by unit i in the sample. The standard variance estimator of $\hat{t}_y$ can be written as

$$v = \sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{S}} \Omega_{ij} y_i y_j, \tag{2.2}$$

where $\Omega_{ij} = (\pi_{ij} - \pi_i \pi_j) / (\pi_{ij} \pi_i \pi_j)$ and $\pi_{ij} = P(i, j \in \mathcal{S})$ is the second order joint inclusion probability for (ij). It is assumed that $\pi_{ij} > 0$ for all (ij). Note that $\pi_{ii} = \pi_i$. The standard variance estimator v is often viewed as fully efficient since it is the Horvitz-Thompson estimator of the design-based variance $V(\hat{t}_y)$.

Let $\Delta = (\Omega_{ij})$ be an $n \times n$ matrix. We can re-write (2.2) as $v = \mathbf{y}' \Delta \mathbf{y}$, where $\mathbf{y} = (y_1, y_2, \ldots, y_n)'$ is the vector of sampled y_i's. The matrix Δ is nonnegative definite and can be decomposed as

$$\Delta = \sum_{k=1}^{p} \lambda_k \delta_k \delta_k' \tag{2.3}$$

for some $\lambda_k > 0$ and some n-dimensional vectors $\delta_k, k = 1, 2, \ldots, p$. The most well-known decomposition (2.3) is given by the spectral decomposition where δ_k is the eigenvector associated with the eigenvalue λ_k. In practice, very small eigenvalues are often ignored for computational reasons. For stratified sampling, the matrix Δ is block-diagonal so the computational burden may be alleviated. However, we do not restrict (2.3) to the spectral decomposition. Any decomposition satisfying (2.3) can be used.

Suppose that we want to express the fully efficient variance estimator v given by (2.2) as a replication variance estimator in the form of

$$v_R = \sum_{k=1}^{L} c_k (\hat{t}_y^{(k)} - \hat{t}_y)^2, \tag{2.4}$$

where $\hat{t}_y^{(k)} = \sum_{i \in \mathcal{S}} w_i^{(k)} y_i$, $\mathbf{w}^{(k)} = (w_1^{(k)}, \ldots, w_n^{(k)})'$ is the kth set of replication weights, $c_k > 0$ is the factor associated with the kth set of replication weights and L is the total number of replications; see Kim, Navarro and Fuller (2006) for further discussion.

The form given by (2.4) does not include all replication variance estimators. For instance, Campbell (1980) provided a jackknife variance estimator where the pseudovalues are derived based on the von Mises approximation to the parameter of interest. Nevertheless, most replication variance estimators can be put in this form.

We have the following result on the construction of $\mathbf{w}^{(k)}$ for v_R based on the decomposition (2.3).

Theorem 1. The fully efficient variance estimator v and the replication variance estimator v_R are algebraically identical if we let $L = p$ and $\mathbf{w}^{(k)} = \mathbf{w} + (\lambda_k / c_k)^{1/2} \boldsymbol{\delta}_k$, where $\mathbf{w} = (w_1, \ldots, w_n)'$ is the set of original basic design weights.

Proof. The proof follows directly from the fact that $v = \mathbf{y}' \Delta \mathbf{y} = \sum_{k=1}^{p} \lambda_k (\boldsymbol{\delta}_k' \mathbf{y})^2$ and that $\hat{t}_y^{(k)} - \hat{t}_y = (\mathbf{w}^{(k)} - \mathbf{w})' \mathbf{y} = (\lambda_k / c_k)^{1/2} \boldsymbol{\delta}_k' \mathbf{y}$.

The choices of c_k's can be arbitrary and bear no impact on the validity and efficiency of the replication variance estimators. However, certain choices of c_k will result in replication weights with negative values, which is undesirable as it may produce negative replicates for the parameters that are always positive. In practical situations one can always choose relatively large c_k to avoid negative values for replication weights. In our simulation study (*Case I*) reported in Section 5, the problem of negative replication weights can be eliminated with the choice of $c_k = 1$.

The replication variance estimator $v_R = \sum_{k=1}^{L} c_k (\hat{t}_y^{(k)} - \hat{t}_y)^2$ with $L = p$ and replication weights $\boldsymbol{w}^{(k)} = \boldsymbol{w} + (\lambda_k / c_k)^{1/2} \boldsymbol{\delta}_k$ is fully efficient for an arbitrary variable y. It also provides fully efficient variance estimator for $\hat{\theta}$ when θ is a smooth function of population means or totals. Practical implementation of the method depends crucially on two related issues: (i) the feasibility of the decomposition of the $n \times n$ matrix $\boldsymbol{\Delta}$ specified in (2.3); and (ii) the number of sets of replication weights required to achieve the full efficiency determined by $p = \text{rank}(\boldsymbol{\Delta})$.

As for the first issue, modern advances in computational power and improved performances of available software packages make it possible to do the spectral decomposition with relatively large n. For instance, on a 12-CPU unix machine with 96 gigabytes of memory, the R function *eigen*() can handle matrices of sizes at least as large as $n = 4,000$. Note that the computational task involved here is for survey runners at the data preparation stage and is not for users of the data files. As for the second issue, the value of p is related to the given sampling design. For simple random sampling without replacement, we have

$$\boldsymbol{\Delta} = N^2 (1 - n / N)(n(n-1))^{-1}(\boldsymbol{I}_n - \boldsymbol{1}_n \boldsymbol{1}'_n / n),$$

where $\boldsymbol{I}_n$ is the $n \times n$ identity matrix and $\boldsymbol{1}_n = (1, 1, \ldots, 1)'$ is the $n \times 1$ vector of 1's. It follows that $p = \text{rank}(\boldsymbol{\Delta}) = \text{trace}(\boldsymbol{I}_n - \boldsymbol{1}_n \boldsymbol{1}'_n / n) = n - 1$. This is typically the case for single stage unequal probability sampling designs. For stratified simple random sampling, we have $p = n - H$, where H is the total number of strata.

It should be noted that $p \leq n$ for any sampling design and the exact value of p is not required for the proposed procedure to be implemented. However, since the values of p and n have the same order of magnitude, the proposed method requires a large number of replicates whenever n is large. Under the current practices in sample surveys, the fully efficient replication weights described above become immediately implementable if $p \leq 500$ and the second order inclusion probabilities π_{ij} are available. When p is large, a two-stage procedure to be described

in Section 3 can be used to produce a small number L_0 sets of replication weights for public-use data files.

In some cases, the spectral decomposition (2.3) can be avoided. For example, Deville (1999) argued that the variance estimator of $\hat{t}_y$ under unequal probability sampling designs with fixed sample size can be approximated by

$$v \doteq c \sum_{i \in \mathcal{S}} (1 - \pi_i) \left(\frac{y_i}{\pi_i} - \tilde{t}_y \right)^2 \tag{2.5}$$

where $c = \left(1 - \sum_{i \in \mathcal{S}} b_i^2 \right)^{-1}, b_i = (1 - \pi_i) / \sum_{k \in \mathcal{S}} (1 - \pi_k)$ and $\tilde{t}_y = \sum_{i \in \mathcal{S}} b_i (y_i / \pi_i)$. More generally, we consider the following form of matrix Δ in $v = y'\Delta y$, where

$$\Delta = \Delta_0 - \Delta_0 X (X'\Delta_0 X)^{-1} X'\Delta_0 \tag{2.6}$$

where $\Delta_0 = \text{diag}\{\lambda_1, \ldots, \lambda_n\}, \lambda_i > 0$ for all $i = 1, 2, \ldots, n, X' = (x_1, \ldots, x_n)$ and x_i is a vector of design and auxiliary variables. Many elementary single-stage sampling designs take the form (2.6) for variance estimation. In particular, Deville's formula in (2.5) can be expressed as $v \doteq y'\Delta y$ with Δ given by (2.6), where $\lambda_i = c\pi_i^{-2}(1 - \pi_i)$ in Δ_0 and $x_i = \pi_i$. The conditional Poisson sampling design to be discussed in Section 5 also takes the form (2.6) where x_i are the design variables in the design constraint $\sum_{i \in \mathcal{S}} \pi_i^{-1} x_i = \sum_{i=1}^{N} x_i$.

For the matrix given by (2.6), it can be shown that

$$y'\Delta y = (y - X\hat{\beta})'\Delta_0 (y - X\hat{\beta}),$$

where $\hat{\beta} = (X'\Delta_0 X)^{-1} X'\Delta_0 y$. Thus, we have

$$y'\Delta y = \sum_{k=1}^{n} \lambda_k (y_k - x_k'\hat{\beta})^2, \tag{2.7}$$

which is useful in deriving an expression for replication variance estimator in the form given by (2.4). The fully efficient variance estimator v in (2.7) and the replication variance estimator v_R

in (2.4) are algebraically identical if we let $L = n$ and $\boldsymbol{w}^{(k)} = \boldsymbol{w} + (\lambda_k / c_k)^{1/2} \boldsymbol{\delta}_k$, where $\boldsymbol{w} = (w_1, \ldots, w_n)'$ is the set of original basic design weights and $\boldsymbol{\delta}_k = (\delta_{1k}, \ldots, \delta_{nk})'$ with

$$
\delta_{ik} = \begin{cases} -1 + \boldsymbol{x}_k'(\boldsymbol{X}'\boldsymbol{\Delta}_0\boldsymbol{X})^{-1}\boldsymbol{x}_i\lambda_i & \text{if } i = k \\ \boldsymbol{x}_k'(\boldsymbol{X}'\boldsymbol{\Delta}_0\boldsymbol{X})^{-1}\boldsymbol{x}_i\lambda_i & \text{otherwise.} \end{cases}
$$

The proof follows directly from the fact that $\boldsymbol{\delta}_k'\boldsymbol{y} = -y_k + \boldsymbol{x}_k'\hat{\boldsymbol{\beta}}$, $\boldsymbol{y}'\boldsymbol{\Delta}\boldsymbol{y} = \sum_{k=1}^{n} \lambda_k (\boldsymbol{\delta}_k'\boldsymbol{y})^2$ and that $\hat{t}_y^{(k)} - \hat{t}_y = (\boldsymbol{w}^{(k)} - \boldsymbol{w})'\boldsymbol{y} = (\lambda_k / c_k)^{1/2} \boldsymbol{\delta}_k'\boldsymbol{y}$.

3 Sparse and efficient replication weights

Large-scale complex surveys usually have a relatively large sample size ranging from a few hundreds to many thousands. The fully efficient replication weights described in Section 2 or replication weights constructed by some existing methods such as the jackknife or the bootstrap methods would involve a very large number of sets of weights. Although valid replication weights provide enormous convenience to the users of survey data, who are not necessarily the survey runners, the burden of manipulating a data set with hundreds or even thousands of replicate weights can be enormous. As a result, how to achieve efficient replication variance estimation with a relatively small number of replicate weights is a question with both theoretical and practical value.

We propose two strategies to construct sparse and efficient replication weights. We start with a large number L sets of replication weights. These initial weights may be produced using the general method described in Section 2 or by existing methods. Suppose they can be viewed as fully efficient. The first strategy is to select a small number L_0 sets of weights from the initial large number L sets of weights using a probability sampling method. The small number L_0 satisfies the desired *sparsity* and the random selection procedure guarantees *validity* of the resulting variance estimators. The second strategy is to achieve *efficiency* through a novel

weight-calibration procedure. The L_0 sets of calibrated replication weights provide fully efficient variance estimators for variables used in the calibration and also highly efficient variance estimators for variables related to calibration variables.

3.1 Achieve sparsity and efficiency through random sampling

Suppose that the fully efficient replication variance estimator is given by $v_R = \sum_{k=1}^{L} c_k (\hat{t}_y^{(k)} - \hat{t}_y)^2$, with replication weights constructed by using Theorem 1. Observe that v_R can be viewed as a finite population total. If we want to use $L_0 (< L)$ sets of replication weights to provide valid inference for variance estimation, the following simple strategy can be used. First, select L_0 sets of weights from the original L sets of weights by simple random sampling without replacement. For notational simplicity and without loss of generality, we denote the selected sets of weights by $w^{(j)}, j = 1, 2, \ldots, L_0$. Then, calculate the replication variance estimator of $\hat{t}_y$ based on the L_0 sets of weights as

$$v_R^{(1)} = \frac{L}{L_0} \sum_{j=1}^{L_0} c_j (\hat{t}_y^{(j)} - \hat{t}_y)^2. \tag{3.1}$$

The variance estimator $v_R^{(1)}$ is still unbiased for an arbitrary variable y, since $E^*(v_R^{(1)}) = \sum_{k=1}^{L} c_k (\hat{t}_y^{(k)} - \hat{t}_y)^2 = v_R$, where $E^*(\cdot)$ denotes the expectation under the random selection of L_0 sets of weights.

An alternative form of the replication variance estimator based on the L_0 sets of weights can be derived as follows. Noting that $\hat{t}_y^{(k)} - \hat{t}_y = (\lambda_k / c_k)^{1/2} \delta_k' y$, we can re-write the fully efficient variance estimator as

$$v_R = m \left\{ \sum_{k=1}^{L} \lambda_k (\delta_k' y)^2 \bigg/ \sum_{k=1}^{L} \lambda_k \right\},$$

where $m = \sum_{k=1}^{L} \lambda_k$. The δ_k's are orthogonal eigenvectors satisfying $|\delta_k| = 1$ under spectral decomposition and $\delta_k' y$ are projections of y onto the n-dimensional unit-sphere. It is very

natural to use the following weighted version for the variance estimator of $\hat{t}_y$ based on the L_0 randomly selected sets of weights:

$$v_R^{(2)} = m \left\{ \sum_{j=1}^{L_0} \lambda_j (\boldsymbol{\delta}_j' \boldsymbol{y})^2 \bigg/ \sum_{j=1}^{L_0} \lambda_j \right\} = \frac{m}{m_0} \sum_{j=1}^{L_0} c_j (\hat{t}_y^{(j)} - \hat{t}_y)^2, \tag{3.2}$$

where $m_0 = \sum_{j=1}^{L_0} \lambda_j$. Noting that $v_R^{(2)}$ is a ratio estimator of v_R, it is usually more efficient than $v_R^{(1)}$.

A third version of the replication variance estimator can be constructed by first selecting L_0 sets of weights with unequal probabilities and then using a Horvitz-Thompson estimator of v_R. Note that we can view $v_R = \sum_{k=1}^{L} \lambda_k (\boldsymbol{\delta}_k' \boldsymbol{y})^2$ as a population total and λ_k as a size variable. We select L_0 sets of weights from the original L sets of weights with inclusion probabilities η_k proportional to λ_k. The resulting variance estimator of $\hat{t}_y$ is given by

$$v_R^{(3)} = \sum_{j=1}^{L_0} \frac{c_j}{\eta_j} (\hat{t}_y^{(j)} - \hat{t}_y)^2, \tag{3.3}$$

where $\eta_j = L_0 \lambda_j \big/ \sum_{k=1}^{L} \lambda_k$. It turns out that the eigenvalues λ_k differ substantially in magnitude and using λ_k as size measure leads to a large portion of the L sets of weights being selected with probability one. In the simulation study described in Section 5, we also included a fourth version of the replication variance estimator, denoted as $v_R^{(4)}$, with $\lambda_k^{1/2}$ as the size measure and $\eta_j = L_0 \lambda_j^{1/2} \big/ \sum_{k=1}^{L} \lambda_k^{1/2}$.

Another possible version of the replication variance estimator is to simply select L_0 sets of weights corresponding to the L_0 largest values of λ_k and then use $v_R^{(2)}$. Simulation results, not reported here, showed that the resulting variance estimator is severely biased and shouldn't be used in practice.

3.2 Achieve sparsity and efficiency through weight-calibration

We now discuss a novel approach of achieving sparsity without losing the efficiency of the variance estimators for some key variables. Suppose L_0 is the desired replication size, which is

much smaller than the sample size n. For example, the Natural Resources Inventory Survey (sponsored by the US department of Agriculture) used $L_0 = 29$ while the PSU sample size can be as large as $n = 300,000$. We present a weight-calibration technique that not only allows the use of a small L_0, but also provides fully efficient variance estimators for key population parameters. Our proposed strategy for constructing the smaller number L_0 sets of replication weights consists of the following four steps:

Step 1. Choose a set of key variables for which full efficiency of the variance estimator is desired. Let $z_i = (z_{i1}, \ldots, z_{im})'$ be the vector of key variables for the ith unit included in the survey data file, where $m \le L_0$. Among them can be important auxiliary variables and study variables as well as design variables. Let $\hat{t}_z = \sum_{i \in S} w_i z_i$. Let $v_1(\hat{t}_z)$ be an $m \times m$ estimated variance-covariance matrix for $\hat{t}_z$ computed by the standard linearization method or by a replication method that is fully efficient.

Step 2. Construct an initial L_0 sets of replication weights that produce an asymptotically unbiased variance estimator. These initial replicates can be obtained by a bootstrap method with L_0 replicates, or by the delete-a-group jackknife method of Kott (2001), or by the sampling method described in Section 3.1. Let $w_0^{(k)} = (w_{10}^{(k)}, \ldots, w_{n0}^{(k)})', k = 1, \ldots, L_0$ be the initial sets of weights. Denote $\hat{t}_{y0}^{(k)} = \sum_{i \in S} w_{i0}^{(k)} y_i$ and let

$$v_0 = \sum_{k=1}^{L_0} c_{k0} (\hat{t}_{y0}^{(k)} - \hat{t}_y)^2 \tag{3.4}$$

be the replication variance estimator based on the L_0 sets of weights.

We can apply the variance formula (3.4) to the vector of key variables z to get $v_0(\hat{t}_z) = \sum_{k=1}^{L_0} c_{k0} (\hat{t}_{z0}^{(k)} - \hat{t}_z)(\hat{t}_{z0}^{(k)} - \hat{t}_z)'$, where $\hat{t}_{z0}^{(k)} = \sum_{i \in S} w_{i0}^{(k)} z_i$. Note that $v_0(\hat{t}_z)$ is not as efficient as $v_1(\hat{t}_z)$ obtained in *Step* 1.

Step 3. Decompose the nonnegative definite variance-covariance matrix $v_1(\hat{t}_z)$ as

$$v_1(\hat{t}_z) = \sum_{k=1}^{m} \alpha_k \boldsymbol{q}_k \boldsymbol{q}_k' \tag{3.5}$$

using the spectral decomposition or any other suitable methods. Let $\alpha_k = 0$ for $k = m + 1, \ldots, L_0$ and define

$$\hat{t}_z^{(k)} = \hat{t}_z + (\alpha_k / c_{k0})^{1/2} \boldsymbol{q}_k, k = 1, 2, \ldots, L_0.$$

It follows that the L_0 pseudo-replicates $\hat{t}_z^{(k)}$ defined above satisfy

$$\sum_{k=1}^{L_0} c_{k0} (\hat{t}_z^{(k)} - \hat{t}_z)(\hat{t}_z^{(k)} - \hat{t}_z)' = v_1(\hat{t}_z), \tag{3.6}$$

due to the decomposition to $v_1(\hat{t}_z)$ given in (3.5). It should be noted that (3.5) bears no relation to the decomposition to Δ described in Section 2 and the condition $m \leq L_0$ is required to make (3.6) possible.

Step 4. Improve the efficiency of v_0 computed from (3.4) for an arbitrary y variable through a weight-calibration procedure. For the kth set of initial weights $\boldsymbol{w}_0^{(k)} = (w_{10}^{(k)}, \ldots, w_{n0}^{(k)})'$, the calibrated weights $\boldsymbol{w}_c^{(k)} = (w_{1c}^{(k)}, \ldots, w_{nc}^{(k)})'$ minimize the chi-square distance measure

$$\Phi(\boldsymbol{w}_c^{(k)}, \boldsymbol{w}_0^{(k)}) = \sum_{i \in \mathcal{S}_k} \tau_i (w_{ic}^{(k)} - w_{i0}^{(k)})^2 / w_{i0}^{(k)} \tag{3.7}$$

subject to the constraint

$$\sum_{i \in \mathcal{S}} w_{ic}^{(k)} \boldsymbol{z}_i = \hat{t}_z^{(k)}, \tag{3.8}$$

where $\mathcal{S}_k = \{i | i \in \mathcal{S}; w_{i0}^{(k)} > 0\}$, the τ_i's are known constants, and $\hat{t}_z^{(k)}$ is the kth pseudo replicate of $\hat{t}_z$ defined in *Step* 3.

The calibrated weights $\boldsymbol{w}_c^{(k)} = (w_{1c}^{(k)}, \ldots, w_{nc}^{(k)})', k = 1, 2, \ldots, L_0$ are used in (3.4) to compute the final replication variance estimator $v_C(\hat{t}_y) = \sum_{k=1}^{L_0} c_{k0} (\hat{t}_{yc}^{(k)} - \hat{t}_y)^2$, where $\hat{t}_{yc}^{(k)} = \sum_{i \in \mathcal{S}} w_{ic}^{(k)} y_i$.

The proposed weight-calibration procedure ensures that the replication estimator $v_C(\hat{t}_z)$ based on the L_0 sets of calibrated weights matches exactly the fully efficient estimator $v_1(\hat{t}_z)$, due to the calibration constraints (3.8) and the equation (3.6). Furthermore, the calibrated replication weights provide more efficient variance estimators for an arbitrary y that is related to z. To see this, we re-write $\hat{t}_{y0}^{(k)} = \sum_{i \in \mathcal{S}} w_{i0}^{(k)} y_i$ as

$$\hat{t}_{y0}^{(k)} = \hat{t}_{e0}^{(k)} + (\hat{t}_{z0}^{(k)})'\hat{\boldsymbol{\beta}}, \tag{3.9}$$

where $\hat{t}_{e0}^{(k)} = \sum_{i \in \mathcal{S}} w_{i0}^{(k)} \hat{e}_i$, $\hat{e}_i = y_i - z_i'\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}} = \left\{ \sum_{i \in \mathcal{S}} w_i z_i z_i' / \tau_i \right\}^{-1} \sum_{i \in \mathcal{S}} w_i z_i y_i / \tau_i$. Let $\hat{t}_e = \sum_{i \in \mathcal{S}} w_i \hat{e}_i$. The variance estimator of $\hat{t}_y$ based on the initial L_0 sets of weights can be expressed as

$$
\begin{aligned}
v_0(\hat{t}_y) &= \sum_{k=1}^{L_0} c_{k0} (\hat{t}_{y0}^{(k)} - \hat{t}_y)^2 \\
&= \sum_{k=1}^{L_0} c_{k0} (\hat{t}_{e0}^{(k)} - \hat{t}_e)^2 + \sum_{k=1}^{L_0} c_{k0} ((\hat{t}_{z0}^{(k)})'\hat{\boldsymbol{\beta}} - (\hat{t}_z)'\hat{\boldsymbol{\beta}})^2 + 2\sum_{k=1}^{L_0} c_{k0} (\hat{t}_{e0}^{(k)} - \hat{t}_e)((\hat{t}_{z0}^{(k)})'\hat{\boldsymbol{\beta}} - (\hat{t}_z)'\hat{\boldsymbol{\beta}}) \\
&= v_0(\hat{t}_e) + \hat{\boldsymbol{\beta}}' v_0(\hat{t}_z)\hat{\boldsymbol{\beta}} + 2\hat{\boldsymbol{\beta}}' \operatorname{cov}_0(\hat{t}_e, \hat{t}_z),
\end{aligned}
$$

where $\operatorname{cov}_0(\hat{t}_e, \hat{t}_z)$ is the estimated covariance between $\hat{t}_e$ and $\hat{t}_z$ based on the initial L_0 sets of replication weights. In many designs, we can choose a suitable τ_i such that $\operatorname{Cov}(\hat{t}_e, \hat{t}_z) \doteq \mathbf{0}$. This is the case, for instance, with the choice of $\tau_i = (w_i - 1)^{-1}$ or $\tau_i = w_i^{-1}$ under Poisson sampling. Fuller (1998) discussed the required conditions in the context of two-phase sampling. It follows that

$$v_0(\hat{t}_y) \doteq v_0(\hat{t}_e) + \hat{\boldsymbol{\beta}}' v_0(\hat{t}_z)\hat{\boldsymbol{\beta}}. \tag{3.10}$$

Using similar argument, it can be shown that the variance estimator based on the L_0 sets of calibrated weights satisfies

$$v_C(\hat{t}_y) \doteq v_0(\hat{t}_e) + \hat{\boldsymbol{\beta}}' v_1(\hat{t}_z)\hat{\boldsymbol{\beta}}. \tag{3.11}$$

The variance estimator $v_C(\hat{t}_y)$ given by (3.11) is generally more efficient than $v_0(\hat{t}_y)$ given by (3.10), due to the use of $v_1(\hat{t}_z)$ instead of $v_0(\hat{t}_z)$. The gain of efficiency depends on the

relative magnitude of $\hat{\boldsymbol{\beta}}' v_1(\hat{t}_z) \hat{\boldsymbol{\beta}}$ over $v_0(\hat{t}_e)$. If y is highly correlated with $\hat{y} = \boldsymbol{z}' \hat{\boldsymbol{\beta}}$, the variance of the residual term $v_0(\hat{t}_e)$ will be relatively small. In this case $v_C(\hat{t}_y)$ will be highly efficient. On the other hand, if there is no correlation between y and $\hat{y}$, then no improvement will be achieved by using the calibrated weights $\boldsymbol{w}_c^{(k)}$; see also Theorem 3 in Section 4.

One of the drawbacks of the chi-square distance $\Phi(\boldsymbol{w}_c^{(k)}, \boldsymbol{w}_0^{(k)})$ in *Step* 4 is that some of the resulting calibrated weights could take negative values. To avoid negative weights, we propose replacing the chi-square distance in (3.7) by the following minimum entropy distance

$$D(\boldsymbol{w}_c^{(k)}, \boldsymbol{w}_0^{(k)}) = -\sum_{i \in \mathcal{S}_k} \tau_i^{-1} \left\{ w_{i0}^{(k)} \log\left(\frac{w_{ic}^{(k)}}{w_{i0}^{(k)}}\right) - w_{ic}^{(k)} + w_{i0}^{(k)} \right\} \tag{3.12}$$

for two reasons. First, the calibrated weights $w_{ic}^{(k)}$ are guaranteed to be positive. Second, there exists a well-behaved computational algorithm for this constrained minimization problem. It can be shown that $\boldsymbol{w}_c^{(k)}$ minimizing $D(\boldsymbol{w}_c^{(k)}, \boldsymbol{w}_0^{(k)})$ subject to (3.8) are given by

$$w_{ic}^{(k)} = \frac{w_{i0}^{(k)} / \tau_i}{1 + \boldsymbol{\lambda}' \boldsymbol{z}_i}, \tag{3.13}$$

where the Lagrange multiplier $\boldsymbol{\lambda}$ is the solution to

$$g(\boldsymbol{\lambda}) = \sum_{i \in \mathcal{S}} \frac{w_{i0}^{(k)} \boldsymbol{z}_i / \tau_i}{1 + \boldsymbol{\lambda}' \boldsymbol{z}_i} - \hat{t}_z^{(k)} = \boldsymbol{0}. \tag{3.14}$$

An efficient computational algorithm for finding the solution $\boldsymbol{\lambda}$ to (3.14) can be found in Wu (2004) and a related R function can be obtained by a minor modification of the R function presented in Wu (2005).

4 Validity

In this section we provide some general discussion on the validity of the replication variance estimator. Let $\theta = f(t_y)$ be a finite population parameter, which is a smooth function of the population total $t_y = \sum_{i=1}^{N} y_i$. We assume that $\hat{\theta} = f(\hat{t}_y)$ is used to estimate θ, where $\hat{t}_y$ is the

Horvitz-Thompson estimator of t_y defined in (2.1). The replication variance estimator of $\hat{\theta}$ is constructed by

$$v_R(\hat{\theta}) = \sum_{k=1}^{L} c_k (\hat{\theta}^{(k)} - \hat{\theta})^2, \tag{4.1}$$

where $\hat{\theta}^{(k)} = f(\hat{t}_y^{(k)})$ and $\hat{t}_y^{(k)}$ is the kth replicate of $\hat{t}_y$.

To explore the asymptotic properties of the replication variance estimator (4.1), we assume a sequence of the finite populations and the survey samples, as described in Isaki and Fuller (1982). The finite populations and the sampling designs satisfy following regularity conditions.

C1. For any population characteristics u_i with bounded second moments,

$$\sum_{i \in \mathcal{S}} w_i u_i u_i' - \sum_{i=1}^{N} u_i u_i' = O_p(n^{-1/2} N).$$

C2. The design weights are uniformly bounded. That is, $K_1 < N^{-1} n w_i < K_2$ for all i and any n, where K_1 and K_2 are fixed constants.

C3. $n V(N^{-1} \hat{t}_y)$ is bounded.

C4. For any y with bounded fourth moments, the replication variance estimator $v_R(\hat{t}_y) = \sum_{k=1}^{L} c_k (\hat{t}_y^{(k)} - \hat{t}_y)^2$ satisfies

$$E[\{c_k (\hat{t}_y^{(k)} - \hat{t}_y)^2\}^2] < K L^{-2} \{V(\hat{t}_y)\}^2 \tag{4.2}$$

for some K, uniformly in $k = 1, \ldots, L$,

$$\max_k c_k^{-1} = O(L), \tag{4.3}$$

and

$$E\left[\left\{\frac{v_R(\hat{t}_y)}{V(\hat{t}_y)} - 1\right\}^2\right] = o(1). \tag{4.4}$$

Condition (4.2) ensures that no single replicate dominate the others. Condition (4.3) controls the order of the factor c_k. Condition (4.4) implies that $v_R(\hat{t}_y)$ is a consistent estimator of $V(\hat{t}_y)$. Conditions (4.2) - (4.4) were also used in Kim, Navarro and Fuller (2006).

Using the above regularity conditions, the following theorem proves the consistency of the replication variance estimator in the form of (4.1).

Theorem 2. Let $\theta = f(t_y)$ be the parameter of interest and $\hat{\theta} = f(\hat{t}_y)$, where $f(\cdot)$ is a smooth function with derivative continuous at t_y. Under the regularity conditions described above, the variance estimator $v_R(\hat{\theta})$ in (4.1) satisfies

$$\frac{v_R(\hat{\theta})}{V(\hat{\theta})} = 1 + o_p(1). \tag{4.5}$$

Proof. See Appendix A.

We now prove the validity of the improved variance estimator $v_C(\hat{t}_y)$ proposed in Section 3.2. For simplicity, we assume that $v_1(\hat{t}_y)$ is a fully efficient estimator of the variance $V(\hat{t}_y)$ for $\hat{t}_y = \sum_{i \in \mathcal{S}} w_i y_i$. We also assume that $v_0(\hat{t}_y)$, defined in (3.4), satisfies

$$E^*\{v_0(\hat{t}_y)\} = v_1(\hat{t}_y), \tag{4.6}$$

where $E^*(\cdot)$ denotes expectation under the random selection of the L_0 replicates from the L sets of fully efficient replication weights, as discussed in Section 3.1. If $v_1(\hat{t}_y)$ is asymptotically unbiased, then $v_0(\hat{t}_y)$ is also asymptotically unbiased by (4.6). For the delete-a-group jackknife, condition (4.6) can be understood as $E\{v_0(\hat{t}_y)\} = E\{v_1(\hat{t}_y)\}$ and $V\{v_0(\hat{t}_y)\} \geq V\{v_1(\hat{t}_y)\}$.

Theorem 3. Assume that the initial variance estimator $v_0(\hat{t}_y)$ defined in (3.4) satisfies (4.6). Assume that the improved variance estimator $v_C(\hat{t}_y) = \sum_{k=1}^{L_0} c_{k0}(\hat{t}_{yc}^{(k)} - \hat{t}_y)^2$ is computed using the calibrated replication weights as described in Section 3.2, with the choice of τ_i satisfying $\mathrm{Cov}(\hat{t}_e, \hat{t}_z) \doteq \mathbf{0}$. By ignoring smaller order terms, we have

$$E\{v_C(\hat{t}_y)\} = E\{v_1(\hat{t}_y)\} \tag{4.7}$$

and

$$V\{v_1(\hat{t}_y)\} \leq V\{v_C(\hat{t}_y)\} \leq V\{v_0(\hat{t}_y)\}. \tag{4.8}$$

Proof. See Appendix B.

For a general parameter $\theta = f(t_y)$, we let $\hat{\theta}_c^{(k)} = f(\hat{t}_{yc}^{(k)})$ and compute $v_C(\hat{\theta}) = \sum_{k=1}^{L_0} c_{k0}(\hat{\theta}_c^{(k)} - \hat{\theta})^2$. Validity of $v_C(\hat{\theta})$ can be established by combining results from Theorem 2 and Theorem 3.

5 Extension to some balanced sampling designs

We now consider sampling designs which are balanced in x_i in the sense that $\hat{t}_x = \sum_{i \in \mathcal{S}} x_i / \pi_i = t_x$ holds exactly or nearly exactly, where x_i is a q-dimensional vector and $t_x = \sum_{i \in \mathcal{U}} x_i$ is known. We assume that the first element of x_i is equal to π_i, which implicitly assumes that the survey design has fixed sample size. Tillé (2006) provides a comprehensive account of balanced sampling designs.

Deville and Tillé (2005) argue that $\hat{t}_y = \sum_{i \in \mathcal{S}} y_i / \pi_i$ under balanced sampling has a variance that can be approximated by its variance under conditional Poisson sampling. Breidt and Chauvet (2011) using the same approximation derived

$$v(\hat{t}_y) = \frac{n}{n-q} \sum_{i \in \mathcal{S}} (1 - \pi_i) \left(\frac{y_i}{\pi_i} - \frac{\tilde{y}_i}{\pi_i} \right)^2, \tag{5.1}$$

where $\tilde{y}_i = x_i' \hat{\beta}_P$ and $\hat{\beta}_P = \left\{ \sum_{i \in \mathcal{S}} (1 - \pi_i) \pi_i^{-2} x_i x_i' \right\}^{-1} \sum_{i \in \mathcal{S}} (1 - \pi_i) \pi_i^{-2} x_i y_i$. Roughly speaking, the variance formula (5.1) can be interpreted as approximating $\hat{t}_y$ under the balanced sampling design by a generalized regression estimator under Poisson sampling. That is, $V(\hat{t}_y) \doteq V(\hat{t}_P)$, where $\hat{t}_P = \hat{t}_y + (t_x - \hat{t}_x)' \hat{\beta}_P$. For a formal justification on this approximation, see Fuller (2009b).

The variance formula (5.1) can be used to derive replication weights. To see this, we re-express (5.1) as a jackknife replication variance estimator

$$v_J(\hat{t}_y) = \sum_{k=1}^{n} c_k (\tilde{t}_y^{(k)} - \hat{t}_y)^2, \tag{5.2}$$

where $\tilde{t}_y^{(k)} = \hat{t}_y^{(k)} + (t_x - \hat{t}_x^{(k)})' \hat{\beta}_P^{(k)}, (\hat{t}_x^{(k)}, \hat{t}_y^{(k)}) = \sum_{i \in \mathcal{S}^{(k)}} \pi_i^{-1}(x_i', y_i),$

$$\hat{\beta}_P^{(k)} = \left\{ \sum_{i \in \mathcal{S}^{(k)}} (1 - \pi_i) \pi_i^{-2} x_i x_i' \right\}^{-1} \sum_{i \in \mathcal{S}^{(k)}} (1 - \pi_i) \pi_i^{-2} x_i y_i,$$

$c_k = (1 - \pi_k)n / (n - q)$, and $\mathcal{S}^{(k)} = \mathcal{S} \setminus \{k\}$. To show the asymptotic equivalence between (5.1) and (5.2), we first note that

$$\tilde{t}_y^{(k)} - \hat{t}_y = (\hat{t}_y^{(k)} - \hat{t}_y) + (t_x - \hat{t}_x^{(k)})'\hat{\boldsymbol{\beta}}_P + (t_x - \hat{t}_x^{(k)})'(\hat{\boldsymbol{\beta}}_P^{(k)} - \hat{\boldsymbol{\beta}}_P).$$

Under certain regularity conditions, we have $\hat{\boldsymbol{\beta}}_P^{(k)} = \hat{\boldsymbol{\beta}}_P + O_p(n^{-1})$ and $t_x - \hat{t}_x^{(k)} = x_k / \pi_k = O_p(n^{-1}N)$. Here we used the condition $t_x = \hat{t}_x$ under the balanced sampling design. It follows that $\tilde{t}_y^{(k)} - \hat{t}_y = -\pi_k^{-1}(y_k - x_k'\hat{\boldsymbol{\beta}}_P) + O_p(n^{-2}N)$, and $v_J(\hat{t}_y)$ in (5.2) is asymptotically equivalent to $\sum_{k=1}^{n} c_k \pi_k^{-2}(y_k - \tilde{y}_k)^2$, which equals $v(\hat{t}_y)$ given by (5.1). The variance formula (5.2) is quite useful because it makes the construction of the replication weights quite straightforward for balanced sampling designs. When n is large, the number of replicates can be reduced by using the weight-calibration method described in Section 3.2. Simulation results based on the rejective Poisson sampling of Fuller (2009b), not reported here to save space, showed that the proposed replication variance estimator performs very well.

6 Simulation study

In this section we report results from a simulation study. We consider a synthetic finite population of size $N = 2{,}248$ families using a real data set of Statistics Canada's 2000 Family Expenditure Survey for the province of Ontario. For the ith selected family, the data set contains observations on several variables, including x_{i1}: the number of persons in the family; x_{i2}: the number of children (age < 15); x_{i3}: the number of youths (age 15 - 24); x_{i4}: the total annual income after taxes; y_{i1}: the total annual expenditure; y_{i2}: the annual expenditure on clothing; y_{i3}: the annual expenditure on furnishings and equipment.

We consider three population parameters for comparing different versions of replication variance estimators. The first is the population total of overall annual expenditures, *i.e.*, $\theta_1 = t_{y1} = \sum_{i=1}^{N} y_{i1}$. The second is the ratio of population totals of expenditures on clothing and on furnishings and equipment, *i.e.*, $\theta_2 = t_{y2} / t_{y3} = \left(\sum_{i=1}^{N} y_{i2}\right) / \left(\sum_{i=1}^{N} y_{i3}\right)$. Note that $\theta_2 = \mu_{y2} / \mu_{y3}$. The third is the population correlation coefficient $\theta_3 = \rho(y_1, y_2)$ between the

overall annual expenditure (y_1) and the annual expenditure on clothing (y_2). For each parameter, several replication variance estimators were evaluated through simulation.

We investigate two approaches of replication variance estimation. For the first one, the initial L sets of the replication weights are constructed using the general method described in Section 2. For the second one, the $L = n$ sets of standard delete-1 jackknife replication weights are used to produce fully efficient variance estimators.

Case I. Unequal probability samples are selected by the Rao-Sampford PPS sampling method (Rao 1965; Sampford 1967), with inclusion probabilities π_i proportional to the total annual income x_{i4}. One of the attractive features of the Rao-Sampford method is that the second order inclusion probabilities π_{ij} can be computed exactly. The general procedure described in Section 2 is used to create $L = n$ sets of fully efficient replication weights, and the corresponding variance estimator is denoted as v_R. Those weights are used as the basis to compute and compare different versions of variance estimators $v_R^{(l)}, l = 1, 2, 3, 4$ described in Section 3.1, based on a smaller number L_0 sets of weights. We restrict L_0 to be 25 or 50.

The calibrated replication variance estimator described in Section 3.2 is denoted as v_C. The initial L_0 sets of weights are selected from the original L sets of weights by simple random sampling, and $z_i = (x_{i1}, x_{i2}, x_{i3}, x_{i4}, y_{i2}, y_{i3})'$ is used as calibration variables. Under this setting, the first parameter θ_1 is not directly related to z but the second parameter θ_2 is defined as a nonlinear but smooth function of t_z. The third parameter $\rho(y_1, y_2)$ is more complex and involves population quantities not included in t_z.

Case II. The population of $N = 2,248$ units is first duplicated 10 times, to create a larger population with 22,480 units. Simple random samples of $n = 100$, 200 or 400 are selected from the population. The sampling fractions are less than 2%. Under such scenarios the standard n sets of delete-1 jackknife weights provide fully efficient variance estimator v_J. Let $v_J^{(1)}$ be the variance estimator using L_0 sets of weights, randomly selected from the n sets of jackknife

weights. Let $v_J^{(C)}$ be the variance estimator using the L_0 sets of weights plus calibration over the z variables.

For each simulated sample of size n and a particular population parameter θ, we compute design-based estimator $\hat{\theta}$ and different versions of variance estimators. The process is repeated B times, independently, with $B = 5{,}000$ for *Case I* and $B = 10{,}000$ for *Case II*. The true variance $V = V(\hat{\theta})$ is approximated by $V \doteq B^{-1} \sum_{b=1}^{B} (\hat{\theta}_b - \theta)^2$, where $\hat{\theta}_b$ is calculated from the bth simulated sample, using another independent B simulated samples. Simulation results show that the bias of $\hat{\theta}$ is negligible for all three parameters. Performances of a variance estimator v are measured by the simulated coverage probability of the 95% normal theory confidence interval, computed as $\text{CP} = B^{-1} \sum_{b=1}^{B} I[\hat{\theta}_b - 1.96(v_b)^{1/2} \le \theta \le \hat{\theta}_b + 1.96(v_b)^{1/2}]$, the average length of the interval $\text{AL} = B^{-1} \sum_{b=1}^{B} 2 \times 1.96(v_b)^{1/2}$, and the Relative Root Mean Square Error (RRMSE), computed as $\text{RRMSE} = \{\text{MSE}(v)\}^{1/2} / V$, where v_b is the variance estimator v computed from the bth simulated sample, and $\text{MSE}(v) = B^{-1} \sum_{b=1}^{B} (v_b - V)^2$.

The simulated coverage probabilities are reported in Tables 6.1 and 6.2. The fully efficient variance estimator v_R and v_J provides good coverage for all scenarios except for $\rho(y_1, y_2)$ with *Case I* where the coverage is a bit low. The variance estimators $v_R^{(l)}, l = 1, 2, 3, 4$ and $v_J^{(1)}$ based on L_0 sets of weights seem to work for θ_1, to certain degree for θ_2 as well, but none is working for $\theta_3 = \rho(y_1, y_2)$. The calibrated estimator v_C provides satisfactory coverage for all scenarios for *Case I*. As for the calibrated estimator $v_J^{(C)}$ with *Case II*, it works very well for θ_1 and θ_2, but none are working well for $\theta_3 = \rho(y_1, y_2)$.

It should be noted that the definition of $\rho(y_1, y_2)$ involves population means over three derived variables y_1^2, y_2^2 and $y_1 y_2$. When those three variables are also included at the calibration stage, in addition to z, the resulting variance estimator is denoted as $v_J^{(C+)}$ for *Case II*. It turns out that $v_J^{(C+)}$ provides much better results for θ_3 and also improved results for θ_1 and θ_2.

Also included in Tables 6.1 and 6.2 are the average length of the confidence intervals using $v_C, v_J^{(C)}$ and $v_J^{(C+)}$. The results (AL, in parentheses) are relative to the length of the interval using v_R (Table 6.1) or v_J (Table 6.2), with a value (say) 1.05 indicating 5% increase in length. It can be seen that the calibrated variance estimators produce confidence intervals which are either comparable in length to the intervals using $v_R(v_J)$ or slightly wider, depending on the parameter and/or sample sizes.

The relative root mean square errors (RRMSE) of variance estimators are presented in Tables 6.3 and 6.4. The results seem to depend not only on the parameter and its estimator but also the sampling design and the replication method. For *Case I*, the variance estimator v_C, which is of primary interest, is more stable than v_R for θ_1, almost the same for θ_2, and is less stable for θ_3. Because y_{i1} is well explained by z_i, v_C is quite efficient for estimating the variance of $\hat{\theta}_1 = \hat{t}_{y1}$. For *Case II*, $v_J^{(C)}$ and $v_J^{(C+)}$ are similar to each other but both are less stable than v_J.

Table 6.1
Coverage probabilities of 95% confidence intervals (*Case I*)

θ	L_0	n	v_R	$v_R^{(1)}$	$v_R^{(2)}$	$v_R^{(3)}$	$v_R^{(4)}$	v_C (AL)
t_{y1}	25	50	93.9	92.9	93.1	92.4	93.1	94.3 (1.03)
		100	94.4	92.0	92.4	91.9	93.0	93.4 (1.01)
		150	95.1	91.5	91.9	92.1	93.2	93.7 (0.99)
	50	100	94.5	93.2	93.2	93.4	93.6	94.1 (1.01)
		150	95.1	93.0	93.3	93.5	93.8	94.5 (0.99)
μ_{y2} / μ_{y3}	25	50	92.6	91.0	91.2	90.6	91.0	92.9 (1.02)
		100	93.7	91.1	91.2	89.6	90.8	93.7 (1.01)
		150	94.3	91.1	90.7	89.5	90.8	94.3 (1.00)
	50	100	93.6	92.6	92.5	91.9	92.5	93.8 (1.01)
		150	94.2	92.7	92.6	91.9	92.9	94.3 (1.00)
$\rho(y_1, y_2)$	25	50	89.0	85.7	85.6	79.3	81.9	91.3 (1.14)
		100	90.5	85.4	85.3	78.6	81.5	92.1 (1.17)
		150	90.7	84.6	84.5	76.9	81.6	91.9 (1.17)
	50	100	90.4	88.2	88.2	83.2	85.8	92.9 (1.18)
		150	90.7	87.5	87.6	81.7	84.8	93.4 (1.18)

v_R: The fully efficient replication variance estimator (Section 2); $v_R^{(l)}$, $l = 1, 2, 3, 4$: replication variance estimators based on L_0 sets of weights (Section 3.1); v_C: replication variance estimator based on L_0 sets of calibrated weights (Section 3.2); AL: average length of the confidence interval relative to the one using v_R.

Table 6.2
Coverage probabilities of 95% confidence intervals (*Case II*)

θ	L_0	n	v_J	$v_J^{(1)}$	$v_J^{(C)}$ (AL)	$v_J^{(C+)}$ (AL)
t_{y1}	25	100	94.4	92.0	94.4 (1.02)	94.9 (1.07)
		200	95.0	92.4	95.0 (1.01)	95.2 (1.03)
		400	95.3	92.5	95.1 (0.99)	95.3 (1.01)
	50	100	94.1	93.1	94.2 (1.02)	94.8 (1.07)
		200	94.7	93.3	94.8 (1.01)	95.0 (1.04)
		400	94.7	93.4	94.5 (0.99)	94.8 (1.02)
$\mu_{y2}\,/\,\mu_{y3}$	25	100	92.6	87.3	92.6 (1.05)	93.3 (1.11)
		200	93.6	86.8	93.3 (1.02)	93.7 (1.07)
		400	94.1	86.8	93.8 (0.99)	94.1 (1.04)
	50	100	92.8	90.3	92.9 (1.06)	94.2 (1.11)
		200	93.9	89.8	93.8 (1.03)	94.3 (1.08)
		400	94.1	89.6	93.8 (1.00)	94.1 (1.05)
$\rho(y_1, y_2)$	25	100	92.5	78.0	89.4 (1.06)	91.7 (1.09)
		200	92.7	72.3	86.3 (1.00)	91.4 (1.05)
		400	93.2	71.2	84.5 (0.95)	92.1 (1.04)
	50	100	92.2	84.5	92.2 (1.09)	92.5 (1.11)
		200	92.8	80.5	90.3 (1.05)	92.2 (1.08)
		400	93.1	77.4	88.1 (1.00)	92.6 (1.06)

v_J: The delete-1 jackknife variance estimator; $v_J^{(1)}$: replication variance estimator based on L_0 sets of jackknife weights; $v_J^{(C)}$: replication variance estimator based on L_0 sets of calibrated jackknife weights; $v_J^{(C+)}$: replication variance estimator based on L_0 sets of calibrated jackknife weights, with added variables for weight-calibration; AL: average length of the confidence interval relative to the one using v_J.

Table 6.3
Relative Root Mean Square Errors (RRMSE, *Case I*)

θ	L_0	n	v_R	$v_R^{(1)}$	$v_R^{(2)}$	$v_R^{(3)}$	$v_R^{(4)}$	v_C
t_{y1}	25	50	1.84	2.76	2.24	1.99	1.86	1.43
		100	1.32	2.34	1.67	1.89	1.40	0.83
		150	1.19	1.99	1.34	1.37	1.46	0.87
	50	100	1.32	1.91	1.69	1.63	1.35	0.92
		150	1.19	1.81	1.50	1.62	1.24	0.73
$\mu_{y2}\,/\,\mu_{y3}$	25	50	0.72	0.89	0.88	1.07	0.89	0.74
		100	0.45	0.78	0.77	0.99	0.72	0.46
		150	0.41	0.93	0.87	1.01	0.74	0.41
	50	100	0.46	0.60	0.60	0.77	0.56	0.46
		150	0.41	0.70	0.68	0.70	0.54	0.41
$\rho(y_1, y_2)$	25	50	0.65	0.79	0.83	1.45	1.26	0.96
		100	0.65	1.12	1.16	2.20	1.37	1.24
		150	0.59	1.29	1.34	2.27	1.43	1.50
	50	100	0.65	0.84	0.88	1.63	0.95	1.03
		150	0.59	0.88	0.94	1.48	1.05	1.12

v_R: The fully efficient replication variance estimator (Section 2); $v_R^{(l)}$, $l = 1, 2, 3, 4$: replication variance estimators based on L_0 sets of weights (Section 3.1); v_C: replication variance estimator based on L_0 sets of calibrated weights (Section 3.2).

Table 6.4
Relative Root Mean Square Errors (RRMSE, *Case II*)

θ	L_0	n	v_J	$v_J^{(1)}$	$v_J^{(C)}$	$v_J^{(C+)}$
t_{y1}	25	100	0.29	0.56	0.56	0.66
		200	0.21	0.57	0.47	0.53
		400	0.15	0.56	0.20	0.41
	50	100	0.29	0.41	0.50	0.57
		200	0.21	0.41	0.39	0.44
		400	0.15	0.41	0.17	0.35
μ_{y2} / μ_{y3}	25	100	0.78	1.58	1.90	1.98
		200	0.56	1.44	1.41	1.57
		400	0.39	1.54	0.87	1.40
	50	100	0.81	1.12	1.61	1.67
		200	0.57	1.10	1.22	1.32
		400	0.38	1.04	0.72	1.00
$\rho(y_1, y_2)$	25	100	1.02	2.43	2.71	2.72
		200	0.74	2.44	2.64	2.57
		400	0.47	2.51	2.52	2.42
	50	100	1.04	1.64	1.97	2.12
		200	0.71	1.75	2.01	1.96
		400	0.48	1.76	1.83	1.73

v_J: The delete-1 jackknife variance estimator; $v_J^{(1)}$: replication variance estimator based on L_0 sets of jackknife weights; $v_J^{(C)}$: replication variance estimator based on L_0 sets of calibrated jackknife weights; $v_J^{(C+)}$: replication variance estimator based on L_0 sets of calibrated jackknife weights, with added variables for weight-calibration.

7 Some concluding remarks

Replication methods offer an asymptotically equivalent alternative to linearization methods but are operationally more convenient and flexible. We focused on population parameters that are smooth functions of means or totals. Our theoretical results and limited simulation studies showed that the proposed strategies for constructing sparse and efficient replication weights work well for variance estimation and confidence intervals. Nevertheless, there are a number of issues which require further investigation. First, for complex parameters such as population correlation coefficients, sparse replication variance estimators are not very stable. Second, further evidences on the effectiveness of the proposed strategies for large complex surveys in conjunction to the use of general bootstrap or jackknife weights are needed. Third, it is not clear whether the sparse replication weights will be efficient for parameters that are not smooth functions of means or

totals, such as population quantiles, for which normal theory confidence intervals are known to be inefficient (Sitter and Wu 2001).

Another important issue is the potential application of the proposed methods for parameters and estimators defined through estimating equations. Let $\boldsymbol{\theta}$ be defined as the solution to

$$U_N(\boldsymbol{\theta}) = \sum_{i=1}^{N} u_i(y_i, \boldsymbol{x}_i; \boldsymbol{\theta}) = \boldsymbol{0}. \tag{7.1}$$

Let $\hat{\boldsymbol{\theta}}$ be obtained by solving a sample-based version of (7.1) given by

$$U_n(\boldsymbol{\theta}) = \sum_{i \in \mathcal{S}} w_i u_i(y_i, \boldsymbol{x}_i; \boldsymbol{\theta}) = \boldsymbol{0}. \tag{7.2}$$

Regression or logistic regression analyses using complex survey data can both be viewed special cases of the general forms given by (7.1) and (7.2). The usual sandwich-type variance of $\hat{\boldsymbol{\theta}}$ is given by

$$V(\hat{\boldsymbol{\theta}}) \doteq \left\{ \frac{\partial U_N(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right\}^{-1} V\{U_n(\boldsymbol{\theta})\} \left\{ \frac{\partial U_N(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right\}^{-1} \tag{7.3}$$

A variance estimator $v(\hat{\boldsymbol{\theta}})$ can now be obtained if we substitute $\partial U_N(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}$ by $\partial U_n(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}$ at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$ and estimate $V\{U_n(\boldsymbol{\theta})\}$ by applying replication variance estimation method to $\hat{U}_n = \sum_{i \in \mathcal{S}} w_i \boldsymbol{u}_i$ with $\boldsymbol{u}_i = u_i(y_i, \boldsymbol{x}_i; \hat{\boldsymbol{\theta}})$. For detailed discussions on estimating equations and survey sampling, see, for instance, Binder (1983), Skinner (1989), and Godambe and Thompson (2009), among others.

Achieving efficient variance estimation using a limited number of sets of replication weights is an important research problem with both theoretical and practical significance. The fully efficient replication weights constructed using the procedure described in Section 2 can be treated as initial sets of weights if the sample size n is large. In principle, our proposed strategies in Section 3 for producing sparse and efficient replication weights can be combined with other initial sets of replication weights, including bootstrap weights (Shao 1996) or delete-a-group

jackknife (Kott 2001). One should also include as many relevant variables as possible in the calibration step, so that the final calibrated replication weights are not only sparse but also efficient in providing variance estimators for a large class of estimators. Extensions of the proposed methods to handle calibration weights or nonresponse adjustment are currently under investigation.

Acknowledgements

Appendix

A Proof of Theorem 2

By assumption (4.2), we have

$$\max_{1 \le k \le L} c_k (\hat{t}_y^{(k)} - \hat{t}_y)^2 = O_p(L^{-1} n^{-1} N^2),$$

which, combined with (4.3), implies that

$$\max_{1 \le k \le L} (\hat{\mu}_y^{(k)} - \hat{\mu}_y) = o_p(1), \tag{A.1}$$

where $\hat{\mu}_y^{(k)} = N^{-1} \hat{t}_y^{(k)}$ and $\hat{\mu}_y = N^{-1} \hat{t}_y$. Let $g(\mu_y) = f(N\mu_y)$. We can write

$$\hat{\theta}^{(k)} - \hat{\theta} = g(\hat{\mu}_y^{(k)}) - g(\hat{\mu}_y) = \dot{g}(\hat{\mu}_y)(\hat{\mu}_y^{(k)} - \hat{\mu}_y) + Q_{nk}(\hat{\mu}_y^{(k)} - \hat{\mu}_y),$$

where $\dot{g}(\mu) = \partial g(\mu) / \partial \mu, Q_{nk} = \dot{g}(\mu_k^*) - \dot{g}(\hat{\mu}_y)$, and μ_k^* is an inner point on the line segment between $\hat{\mu}^{(k)}$ and $\hat{\mu}$. By (A.1), we have

$$\max_{1 \le k \le L}(\mu_k^* - \hat{\mu}_y) = o_p(1). \tag{A.2}$$

Define

$$D_\delta = \left\{ \mu \,\Big|\, \max_k \left\| \mu_k^* - \mu \right\| < \delta \text{ and } \max_k \left\| \dot{g}(\mu_k^*) - \dot{g}(\mu) \right\| > \epsilon \right\}.$$

By construction, we have, for any $\epsilon > 0$ and $\delta > 0$,

$$P\left\{ \max_k \left\| \dot{g}(\mu_k^*) - \dot{g}(\hat{\mu}_y) \right\| > \epsilon \right\} \le P(\hat{\mu}_y \in D_\delta) + P\left(\max_k \left\| \mu_k^* - \hat{\mu}_y \right\| \ge \delta \right).$$

By the continuity of $\dot{g}(\mu)$ at $\mu = \mu_y$ and the fact that $\hat{\mu}_y = \mu_y + o_p(1)$, we have that, for any $\epsilon > 0$, there exists a $\delta = \delta(\epsilon) > 0$ such that $P(\hat{\mu}_y \in D_\delta) = o(1)$. This, together with (A.2), implies that

$$\max_k \left\| \dot{g}(\mu_k^*) - \dot{g}(\hat{\mu}_y) \right\| = o_p(1). \tag{A.3}$$

Now, we have

$$\sum_{k=1}^{L} c_k (\hat{\theta}^{(k)} - \hat{\theta})^2 = A_n + B_n + 2C_n, \tag{A.4}$$

where

$$A_n = \sum_{k=1}^{L} c_k \{ \dot{g}(\hat{\mu}_y)(\hat{\mu}_y^{(k)} - \hat{\mu}_y) \}^2,$$
$$B_n = \sum_{k=1}^{L} c_k \{ Q_{nk}(\hat{\mu}_y^{(k)} - \hat{\mu}_y) \}^2, \text{ and}$$
$$C_n = \sum_{k=1}^{L} c_k \dot{g}(\hat{\mu}_y)(\hat{\mu}_y^{(k)} - \hat{\mu}_y)^2 Q_{nk}.$$

Note that (4.4) implies

$$\sum_{k=1}^{L} c_k (\hat{\mu}_y^{(k)} - \hat{\mu}_y)^2 / V(\hat{\mu}_y) = 1 + o_p(1). \tag{A.5}$$

By standard linearization arguments, we have $A_n / V(\hat{\theta}) \to 1$ in probability. Furthermore, by (A.3) and (A.5), we have $B_n / V(\hat{\theta}) = o_p(1)$ and $C_n / V(\hat{\theta}) = o_p(1)$. This establishes (4.5).

B Proof of Theorem 3

Combining (3.10) and (3.11) and ignoring terms of smaller order, we have

$$v_0(\hat{t}_y) - v_C(\hat{t}_y) \doteq \hat{\boldsymbol{\beta}}' v_0(\hat{t}_z)\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}' v_1(\hat{t}_z)\hat{\boldsymbol{\beta}} \doteq \boldsymbol{\beta}' v_0(\hat{t}_z)\boldsymbol{\beta} - \boldsymbol{\beta}' v_1(\hat{t}_z)\boldsymbol{\beta}.$$

where $\boldsymbol{\beta}$ is the probability limit of $\hat{\boldsymbol{\beta}}$. By (4.6), we have

$$E^*\{v_0(\hat{t}_z)\} = v_1(\hat{t}_z), \tag{B.1}$$

where $E^*(\cdot)$ denotes expectation under random selection of the L_0 sets of weights conditional on the L sets of weights. Similarly, by (3.11), we have

$$v_1(\hat{t}_y) - v_C(\hat{t}_y) \doteq v_1(\hat{t}_e) - v_0(\hat{t}_e).$$

By (4.6) again, we have

$$E^*\{v_0(\hat{t}_e)\} = v_1(\hat{t}_e). \tag{B.2}$$

Let $\hat{d}_1 = v_C(\hat{t}_y) - v_1(\hat{t}_y)$, we have $E(\hat{d}_1) = 0$ by (B.2), which proves (4.7). Furthermore, by (B.2) again, we have $\text{Cov}\{\hat{d}_1, v_1(\hat{t}_y)\} = 0$. Thus, we have

$$V\{v_C(\hat{t}_y)\} = V\{v_1(\hat{t}_y)\} + V(\hat{d}_1) \geq V\{v_1(\hat{t}_y)\}. \tag{B.3}$$

Similarly, we can also prove that $V\{v_0(\hat{t}_y)\} \geq V\{v_C(\hat{t}_y)\}$.

References

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

Breidt, F.J., and Chauvet, G. (2011). Improved variance estimation for balanced samples drawn via the cube method. *Journal of Statistical Planning and Inference*, 141, 411-425.

Campbell, C. (1980). A different view of the finite population estimation. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 319-324.

Deville, J.-C. (1999). Variance estimation for complex statistics and estimators: Linearization and residual techniques. *Survey Methodology*, 25, 2, 193-203.

Deville, J.-C., and Tillé, Y. (2005). Variance approximation under balanced sampling. *Journal of Statistical Planning and Inference*, 128, 411-425.

Dippo, C.S., Fay, R.E. and Morganstein, D.H. (1984). Computing variances from complex samples with replicate weights. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, Washington, DC, 489-494.

Fay, R.E. (1984). Some properties of estimators of variance based on replication methods. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, Washington, DC, 495-500.

Fay, R.E., and Dippo, C.S. (1989). Theory and application of replicate weighting for variance calculations. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, Washington, DC, 212-217.

Fuller, W.A. (1998). Replication variance estimation for two phase samples. *Statistica Sinica*, 8, 1153-1164.

Fuller, W.A. (2009a). *Sampling Statistics*. Hoboken, New Jersey: John Wiley & Sons, Inc.

Fuller, W.A. (2009b). Some design properties of a rejective sampling procedure. *Biometrika*, 96, 933-944.

Godambe, V.P., and Thompson, M.E. (2009). Estimating functions and survey sampling. In *Handbook of Statistics*, (Eds., D. Pfeffermann and C.R. Rao), Sample Surveys: Inference and Analysis, North Holland, Vol. 29B, 83-101.

Gross, S. (1980). Median estimation in sample surveys. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, Washington, DC, 181-184.

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

Jang, D., and Eltinge, J.L. (2009). Use of within-primary-sample-unit variances to assess the stability of a standard design-based variance estimator. *Survey Methodology*, 35, 2, 235-245.

Kim, J.K., Navarro, A. and Fuller, W.A. (2006). Replication variance estimation for two-phase stratified sampling. *Journal of the American Statistical Association*, 101, 312-320.

Kott, P.S. (2001). The delete-a-group jackknife. *Journal of Official Statistics*, 17, 521-526.

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.

Lu, W.W., Brick, J.M. and Sitter, R.R. (2006). Algorithms for constructing combining strata variance estimators. *Journal of the American Statistical Association*, 101, 1680-1692.

Lu, W.W., and Sitter, R.R. (2008). Disclosure risk and replication-based variance estimation. *Statistica Sinica*, 18, 1669-1687.

McCarthy, P.J., and Snowden, C.B. (1985). *The Bootstrap and Finite Population Sampling*. Vital and Health Statistics, Ser. 2, No. 95, Public Health Service Publication 85-1369, U.S. Government Printing Office, Washington, DC.

Preston, J. (2009). Rescaled bootstrap for stratified multistage sampling. *Survey Methodology*, 35, 2, 227-234.

Rao, J.N.K. (1965). On two simple schemes of unequal probability sampling without replacement. *Journal of the Indian Statistical Association*, 3, 173-180.

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.

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

Rust, K.F., and Rao, J.N.K. (1996). Variance estimation for complex surveys using replication techniques. *Statistical Methods in Medical Research*, 5, 283-310.

Sampford, M.R. (1967). On sampling without replacement with unequal probabilities of selection. *Biometrika*, 54, 499-513.

Shao, J. (1996). Resampling methods in sample surveys (with discussion). *Statistics*, 27, 203-254.

Shao, J. (2003). Impact of the bootstrap on sample surveys. *Statistical Science*, 18, 191-198.

Shao, J., and Tu, D. (1995). *The Jackknife and Bootstrap*. New York: Springer.

Sitter, R.R. (1992). A resampling procedure for complex survey data. *Journal of the American Statistical Association*, 87, 755-765.

Sitter, R.R., and Wu, C. (2001). A note on Woodruff confidence intervals for quantiles. *Statistics and Probability Letters*, 52, 353-358.

Skinner, C.J. (1989). Domain means, regression and multivariate analysis. In *Analysis of Complex Surveys*, (Eds., C.J. Skinner, D. Holt and T.M. Smith), New York: John Wiley & Sons, Inc., 59-88.

Tillé, Y. (2006). *Sampling Algorithms*. Springer Science + Business Media, Inc.

Wolter, K.M. (2007). *Introduction to Variance Estimation* (2nd Edition). New York: Springer-Verlag.

Wu, C. (2004). Some algorithmic aspects of the empirical likelihood method in survey sampling. *Statistica Sinica*, 14, 1057-1067.

Wu, C. (2005). Algorithms and R codes for the pseudo empirical likelihood method in survey sampling. *Survey Methodology*, 31, 2, 239-243.