

STUDY OF THE PROPERTIES OF THE RAO-WU BOOTSTRAP VARIANCE ESTIMATOR: WHAT HAPPENS WHEN ASSUMPTIONS DO NOT HOLD?

Lenka Mach¹, Abdelnasser Saïdi², and Rob Pettapiece³

ABSTRACT

The Rao-Wu bootstrap variance estimation method is frequently used at Statistics Canada. It is simple to implement, the same formula applies to all estimators, including non-smooth statistics, and it enables the analysts to use the design-based approach. The method assumes primary sampling units are selected with replacement or the first-stage sampling fractions are negligible. We examine the properties of the Rao-Wu bootstrap variance estimator for a two-stage design and a variety of scenarios: different sampling methods, varying sampling fractions, several variables and estimators. We also compare these properties with those of the analytical variance estimators.

KEY WORDS: Complex surveys, Estimation, Finite population, Two-stage design.

RÉSUMÉ

La méthode d'estimation de la variance bootstrap Rao-Wu est fréquemment utilisée à Statistique Canada. La méthode est simple à mettre en oeuvre, la même formule s'applique à tous les estimateurs incluant les statistiques non lisses et elle permet aux analystes d'utiliser l'approche selon le plan d'enquête. La méthode suppose que les unités primaires d'échantillonnage sont sélectionnées avec remise ou que les fractions d'échantillonnage au premier degré sont négligeables. Nous examinons les propriétés de l'estimateur de la variance bootstrap de Rao-Wu pour un plan à deux degrés et sous une diversité de scénarios: différentes méthodes d'échantillonnage, différentes fractions de sondage, plusieurs variables et estimateurs ponctuels. Nous comparons aussi ces propriétés à celles des estimateurs analytiques de la variance.

MOTS CLÉS : Enquêtes complexes; estimation; plan à deux degrés; population finie.

1. INTRODUCTION

1.1 Complex Surveys

The need for reliable estimates, often for relatively small sub-populations, on the one hand, and limited survey resources as well as the types of frame and sampling methods that are feasible, on the other hand, lead to complex survey designs. These designs typically use some of the following sampling techniques: sampling without replacement (WOR) from a finite population, systematic sampling, stratification, clustering, unequal probabilities of selection, multi-stage or multi-phase sampling. As a consequence, the values of the variables of interest observed for units selected in a complex survey sample are neither independent nor identically distributed.

In addition, survey processing, aimed at improving the quality and usability of survey data, reducing the bias of the estimates, satisfying the confidentiality requirements etc. further increase the complexity of the survey data. For example, imputation for missing data produces a complete file for analytical use but introduces an additional source of variation. As another example, the various weight adjustments (for unit non-response, post-stratification, benchmarking etc.), typically required to reduce the bias or improve efficiency and consistency with other data sources, lead to complex estimators. See

¹ Lenka Mach, Statistics Canada, R.H. Coats 15-I, 100 Tunney's Pasture Driveway, Ottawa, Canada, K1A 0T6, Lenka.Mach@statcan.ca

² Abdelnasser Saïdi, Statistics Canada, R.H. Coats 15-K, 100 Tunney's Pasture Driveway, Ottawa, Canada, K1A 0T6, Abdelnasser.Saidi@statcan.ca

³ Rob Pettapiece, University of Waterloo (Work was done while he worked as a co-op student at Statistics Canada.)

Lohr (1999), Särndal, Swensson and Wretman (1992) and Statistics Canada (2003) for more discussion of complex surveys.

1.2 Variance Estimation

Sampling variance is an important measure of the quality of estimates of finite population parameters such as totals, means, quantiles, etc. It measures the variability among the different values of a given estimator obtained by the process of taking all possible samples under a given sample design. Estimates of sampling variance are needed to produce the *coefficients of variation* (cv) that are disseminated along with the survey estimates and to construct confidence intervals for finite population parameters of interest. Sampling variance is also used as the variability measure for inferences about super-population models when the *design-based approach* is recommended (Binder and Roberts 2003). Estimation of the sampling variance can become very complicated due to the complex sample design, use of non-linear estimators, impact of survey processing etc. There are two basic approaches to variance estimation in complex surveys: analytical methods and *resampling* methods (jackknife, balanced repeated replication, bootstrap). For their description see for example Chapter 9 in Lohr (1999).

1.3 Objective and Organization of This Paper

The Rao-Wu bootstrap method is used to estimate variances for many Statistics Canada surveys. It became popular because its implementation is simple and it facilitates the use of design-based analysis. The method assumes that the primary sampling units (PSUs) are selected with replacement (WR) or the first-stage sampling fractions are negligible. However, surveys use WOR sampling and the PSU sampling fractions may be high. In this paper we report some findings of our investigation of the Rao-Wu bootstrap variance estimator. For a two-stage design, we examine its accuracy and stability when PSUs are sampled WOR and the PSU sampling fractions are not negligible, say $> 10\%$. We also compare these properties with the ones of common analytical variance estimators. Bootstrap methods are discussed in Section 2, our study and its results in Section 3, followed by conclusion in Section 4.

2. BOOTSTRAP

The bootstrap was first introduced by Efron (1979) for samples of independent and identically distributed (i.i.d.) observations from some distribution F . An overview of the bootstrap theory and applications in the i.i.d. case can be found in Shao and Tu (1996). A bootstrap method modified for survey samples is now frequently chosen as the variance estimation method for surveys conducted by Statistics Canada because it seems to perform well for most estimators, is relatively easy to implement and enables researchers to more readily perform design-based analysis.

2.1 Bootstrap Variance Estimator for Complex Surveys

The survey samplers started to study the use of bootstrapping for variance estimation in the mid eighties. A direct extension to surveys samples of the standard bootstrap method developed for i.i.d. samples is to apply the standard bootstrap independently in each stratum. This methodology is often referred to as the *naïve bootstrap*. Because the naïve bootstrap variance estimator is inconsistent in the case of bounded stratum sample sizes, several modified bootstrap methods were proposed. Shao and Tu in Chapter 6 of their 1996 book discuss a number of bootstrap methods that were modified for survey samples. For more recent development see, for example, Binder, Kovacevic and Roberts (2004), Fuaoka, Saigo, Sitter and Toida (2006) and Saigo (2007).

2.2 Rao-Wu Bootstrap Variance Estimator

At Statistics Canada, we use the rescaling bootstrap, referred to as the Rao-Wu bootstrap in this article. Rao and Wu (1988) proposed a bootstrap method for stratified multi-stage designs and with-replacement sampling of PSUs. The method applied a scale adjustment directly to the survey data values. Rao, Wu and Yue (1992) presented a modification of the 1988 method where the scale adjustment is applied to the survey weights rather than to the data values. This modification increases the applicability of the method, from variance estimation for smooth statistics to the inclusion of non-smooth statistics as well. We describe the method in this Section.

Let θ denote a finite population parameter and $\hat{\theta}$ its estimator based on the full survey sample. To estimate the variance of $\hat{\theta}$, we repeatedly select bootstrap samples from the full survey sample and apply the procedure given below. Let $\hat{\theta}^*$ denote a bootstrap replicate of $\hat{\theta}$.

For $b=1, \dots, B$, where B is large (typically, $B=500$ for Statistics Canada surveys), repeat independently steps (i) to (iv):

(i) Independently in each stratum h , $h=1, \dots, H$, select a bootstrap sample by drawing a simple random sample of n_h^* PSUs with replacement from the sample of n_h PSUs. Let $t_{hi,b}^*$ be the number of times that PSU hi is selected in the bootstrap sample b $\left(\sum_{i=1}^{n_h} t_{hi,b}^* = n_h^* \right)$.

(ii) For each secondary sampling unit (SSU) k in PSU hi , calculate the initial bootstrap weight

$$d_{hik,b}^* = d_{hik} \left\{ \left(1 - \sqrt{\frac{n_h^*}{n_h - 1}} \right) + \sqrt{\frac{n_h^*}{n_h - 1}} \cdot \frac{n_h}{n_h^*} \cdot t_{hi,b}^* \right\}, \quad (1)$$

where d_{hik} is the initial sampling weight of the SSU hik , equal to the inverse of its selection probability, i.e. $d_{hik} = 1/\pi_{hik}$.

(iii) For each SSU, calculate the final bootstrap weight $w_{hik,b}^*$ by applying, to the initial bootstrap weight $d_{hik,b}^*$, the same weight adjustment procedures (e.g., re-weighting for non-response or calibration) that were applied to the initial sampling weight d_{hik} to obtain the final survey weight w_{hik} .

(iv) Calculate $\hat{\theta}_b^*$, the b -th bootstrap replicate of $\hat{\theta}$, by replacing w_{hik} with $w_{hik,b}^*$ in the formula for $\hat{\theta}$.

The bootstrap variance estimator of $\hat{\theta}$ is then given by $v_{BS}(\hat{\theta}) = \frac{1}{B} \sum_{b=1}^B (\hat{\theta}_b^* - \hat{\theta})^2$. (2)

The full sample estimator $\hat{\theta}$ in (2) can be replaced by $\hat{\theta}_{(.)}^* = \frac{1}{B} \sum_{b=1}^B \hat{\theta}_b^*$ in which case the sum of squares is divided by $(B-1)$.

The estimator $v_{BS}(\hat{\theta})$ in (2) is a Monte Carlo approximation of $\hat{V}_{BS}(\hat{\theta}) = E_* \left[\hat{\theta}^* - E_*(\hat{\theta}^*) \right]^2$, (3)

where E_* denotes the expectation with respect to bootstrap sampling. Rao and Wu (1988) show: a) For a linear $\hat{\theta}$, the bootstrap variance estimator (3) reduces to the standard unbiased variance estimator for WR sampling. b) For any $\hat{\theta}$, (3) is a consistent estimator of the true sampling variance $V_{WR}(\hat{\theta})$ for WR sampling of PSUs.

2.2.1 Size of the bootstrap sample. If $n_h^* \leq n_h - 1$ then the bootstrap weights are never negative. Usually $n_h^* = n_h - 1$,

which simplifies bootstrap weights given in (1) to: $d_{hik,b}^* = d_{hik} \left\{ \frac{n_h}{n_h - 1} \cdot t_{hi,b}^* \right\}$.

Empirical studies by Kovar, Rao and Wu (1988) demonstrated that (2) performs well for smooth $\hat{\theta}$ when $n_h^* = n_h - 1$.

3. SIMULATION STUDY

3.1 Creation of Population File

We used data from the 2000 Youth in Transition Survey (YITS) of 15-year-old students to create a population file for our simulation. YITS is a longitudinal survey, developed by Statistics Canada in partnership with Human Resources Development Canada. It is designed to examine the major transitions in the lives of youth, particularly between education, training and work. The 2000 YITS sample was selected in two stages: 1) First, a sample of 1,241 schools was selected from the population of 3,997 eligible schools stratified by province, language of instruction, and size (defined as the number of 15-year-olds) into 49 strata. 2) At the second stage, an equal-probability systematic sample of students was selected from the sampled schools that had agreed to participate, yielding a total sample of 37,568 students.

The survey school file contained data for 1,117 participating schools and the student file contained 29,330 records with students' responses. From the many variables collected by YITS, we selected a few for our study and imputed their values when missing due to nonresponse. Then, the data for schools and students present on the survey frame but not on the survey files were imputed using random hot deck imputation within imputation classes. That is each record with missing data was assigned data values of a donor record that was selected at random from the same class. Classes of similar records, recipients and donors, were formed using the available frame information. The size of the simulated population is equal to the size of the actual survey population (3,997 schools and close to 400,000 15-year-old students) and the distributions of the different study variables in our simulated population resemble well the actual survey population.

3.2 Methodology

We study many different scenarios that include: different variables (dichotomous, continuous), estimators (of totals, proportions, means, ratios), estimation domains and sampling designs (varying selection methods and sampling fractions for each stage). In this paper we present results for the following design:

Stage 1: Independently, within each school stratum h , a probability-proportional-to-size (PPS) sample of n_h schools is selected WOR with the Hanurav-Vijayan algorithm (using SAS PROC SURVEYSELECT, METHOD=PPS).

Stage 2: From each sampled school hi , $h = 1, \dots, H$, $i = 1, \dots, n_h$, a simple random sample (SRS) of m_{hi} students is selected WOR (using SAS PROC SURVEYSELECT, METHOD=SRS).

This two-stage sample selection is repeated $R=10,000$ times for each scenario. For each simulation r , $r = 1, \dots, R$, we select B (100 or 500) bootstrap samples as described in Section 2.2 with $n_h^* = n_h - 1$. For each r , $\hat{\theta}$ and $v_{BS}(\hat{\theta})$ are calculated as well as two non-resampling variance estimators of $\hat{\theta}$: The Sen-Yates-Grundy estimator for WOR sampling of PSUs, $v_{SYG}(\hat{\theta})$, and the usual estimator for WR sampling of PSUs, $v_{WR}(\hat{\theta})$ (see pp. 197 and 192 respectively in Lohr, 1999).

3.3 Empirical Expectations and Measures

We obtain the empirical Monte Carlo expectation and variance of $\hat{\theta}$ as $E_R(\hat{\theta}) = \frac{\sum_{r=1}^R \hat{\theta}_r}{R}$ and $V_R(\hat{\theta}) = \frac{\sum_{r=1}^R [\hat{\theta}_r - E_R(\hat{\theta})]^2}{R}$.

To evaluate *accuracy* of a particular variance estimator $v(\hat{\theta})$ ($= v_{BS}(\hat{\theta}), v_{SYG}(\hat{\theta}), v_{WR}(\hat{\theta})$), we calculate *relative bias*

$RB_R[v(\hat{\theta})] = \frac{E_R[v(\hat{\theta})]}{V_R(\hat{\theta})} - 1$, where $E_R[v(\hat{\theta})] = \frac{\sum_{r=1}^R v_r(\hat{\theta})}{R}$ is the empirical expectation of $v(\hat{\theta})$. To evaluate *stability*, we use

relative root mean square error (RMSE), also referred to as *coefficient of variation* (CV),

$CV_R[v(\hat{\theta})] = \sqrt{V_R[v(\hat{\theta})]} / V_R(\hat{\theta})$, where $V_R[v(\hat{\theta})] = \frac{\sum_{r=1}^R \{v_r(\hat{\theta}) - E_R[v(\hat{\theta})]\}^2}{R}$.

3.4 Results

For the scenarios described in Section 3.2, we observed that the empirical properties of the studied variance estimators depend on the size and composition of the estimation domain. The selected results for a small and a large domain (shown in Figures 1-3) are representative of the results obtained for all the described scenarios.

Small domain: In our study, a small domain is composed of a single selection stratum. Figures 1 and 2 show results for the domain of French schools of medium size (35 to 85 15-year old students) in Ontario that consists of $N=25$ schools with a total of $M=1,316$ 15-year old students. PPSWOR samples of $n=5, 10, 15$ schools are selected, followed by SRSWOR of $m_i=10$ or 30 students (the subscript h is dropped here because only a single stratum is used).

Figure 1 shows RB and CV of $v_{BS}(\hat{\theta})$, where $\hat{\theta}$ is an estimated total of a dichotomous variable. It demonstrates that:

- RB increases as n and m_i increase. The bias is not too serious for $n=5$ ($n/N=0.2$), but it is close to 90% overestimation for $n=15$ ($n/N=0.6$) and $m_i=30$.
- CV decreases (i.e. the stability improves) as n increases. However, for a given n , CV increases as m_i increases. (It should be noted that $V[v_{BS}(\hat{\theta})]$ decreases as m_i increases, but not relative to $V(\hat{\theta})$ that decreases faster.)

Figure 2 compares the properties of $v_{BS}(\hat{\theta})$, $v_{SYG}(\hat{\theta})$ and $v_{WR}(\hat{\theta})$. We observe:

- RB and CV of $v_{BS}(\hat{\theta})$ and $v_{WR}(\hat{\theta})$ are almost identical.
- As expected, RB of $v_{SYG}(\hat{\theta})$ is close to zero and, rather surprisingly, CV of $v_{SYG}(\hat{\theta})$ is the smallest. $v_{SYG}(\hat{\theta})$ can be negative and thus could be unstable; however, in this particular case, $v_{SYG}(\hat{\theta})$ was positive for all 10,000 simulations.

Large domain: The population of 15-year olds in all Ontario schools, composed of six selection strata, with the total of $N=898$ schools and $M=149,124$ 15-year old students. The first-stage sample of 5, 10, and 17.5% of N schools respectively was allocated to the 6 strata using the \sqrt{N} -proportional allocation: $n_h = n \left(\frac{\sqrt{N_h}}{\sum_h \sqrt{N_h}} \right)$. This lead to varying stratum fractions. For example, for $n/N=0.175$, n_h/N_h ranged from 0.107 (largest N_h) to 0.531 (smallest N_h). In each stratum, PPSWOR sample of n_h schools was selected, followed by SRSWOR of $m_{hi} = 35$ students from each selected school. This sampling plan is similar to the one used by the 2000 YITS survey. (Note that we were not able to select more than 17.5% of schools because the Hanurav-Vijayan algorithm requires the relative size of each school hi to be less than $1/n_h$.)

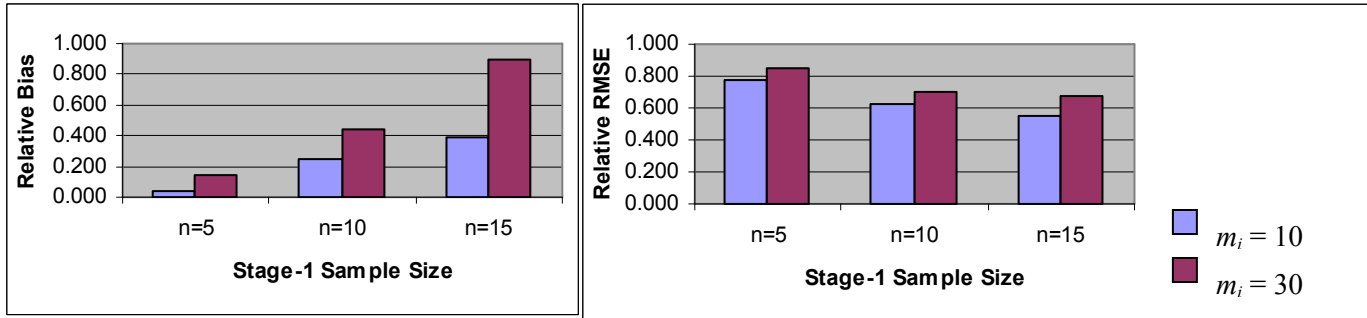


Figure 1: Relative bias and relative RMSE of the Rao-Wu bootstrap variance estimator for a small domain d .

$$\hat{\theta} = \hat{Y} = \sum_{hik\epsilon s} w_{hik} y_{hik}, \text{ where } y_{hik} \text{ is a dichotomous variable, and the population total } Y = 667.$$

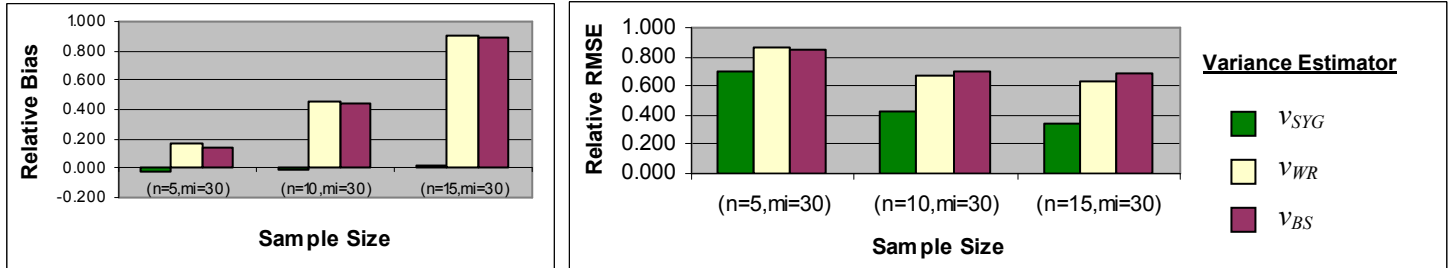


Figure 2: Relative bias and relative RMSE - Comparison of variance estimators for a small domain d .

$$\hat{\theta} = \hat{Y} = \sum_{hik\epsilon s} w_{hik} y_{hik}, \text{ where } y_{hik} \text{ is a dichotomous variable, and the population total } Y = 667.$$

Figure 3 compares RB and CV of $v_{BS}(\hat{\theta})$, $v_{SYG}(\hat{\theta})$ and $v_{WR}(\hat{\theta})$, where $\hat{\theta}$ is a ratio estimator of mean parent income. The non-resampling formulas are estimating the variance of a linear approximation of the ratio $\hat{\theta}$. We observe:

- RB of $v_{BS}(\hat{\theta})$ is basically negligible for all three sampling fractions. Surprisingly, it is higher for $n/N=0.1$ (0.032) than $n/N=0.175$ (0.003). The likely cause of this fluctuation is the variability due to Monte Carlo sampling.
- $RB_R(v_{WR}(\hat{\theta})) < RB_R(v_{BS}(\hat{\theta}))$. This could be explained by the tendency of $v_{WR}(\hat{\theta})$, based on linearization of $\hat{\theta}$, to underestimate $V_{WR}(\hat{\theta})$ for small to moderate sample sizes.
- $v_{SYG}(\hat{\theta})$ is underestimating the true variance for WOR sampling of PSUs, $V_{WOR}(\hat{\theta})$, for $n/N=0.05$ and 0.10. It is known that variance estimators based on linearization underestimate the true variances unless the sample is large.
- $v_{SYG}(\hat{\theta})$ is the least stable variance estimator while $v_{BS}(\hat{\theta})$ and $v_{WR}(\hat{\theta})$ have basically identical CV s. Among the 10,000 simulations for each fraction, there were 92, 187 and 326 negative estimates $v_{SYG}(\hat{\theta})$.

- In this case, $v_{BS}(\hat{\theta})$ and $v_{WR}(\hat{\theta})$ are better estimators of $V_{WOR}(\hat{\theta})$ than $v_{SYG}(\hat{\theta})$.

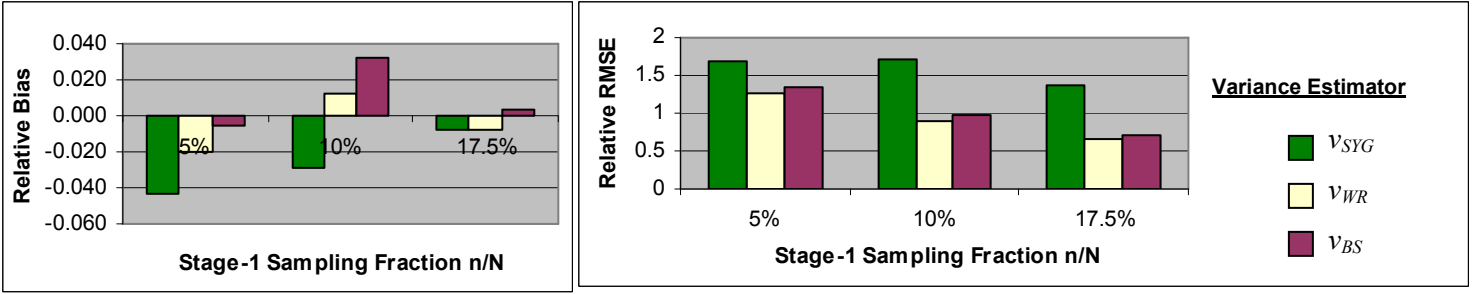


Figure 3: Relative bias and relative RMSE - Comparison of variance estimators for a large domain d .

$$\hat{\theta} = \frac{\hat{Y}}{\hat{M}} = \frac{\sum_{hik\epsilon s} w_{hik} y_{hik}}{\sum_{hik\epsilon s} w_{hik}}, \text{ where } y_{hik} \text{ is parents' income, and the population mean is } \$69,928.$$

The properties of $v_{BS}(\hat{\theta})$ presented in Figures 1-3 are based on $B=100$. Some simulations were also done for $B=500$ and we investigated the impact of B on the bias and stability. In summary, the value of B does not impact RB but, as B increases, CV decreases and hence $v_{BS}(\hat{\theta})$ becomes more stable.

4. CONCLUSION

We examined how well the Rao-Wu bootstrap variance estimator performs when the assumption of WR sampling of PSUs is violated, i.e. when PSUs are sampled WOR and the first-stage sampling fractions are not negligible. For a variety of smooth estimators we found that the amount of overestimation of $V_{WOR}(\hat{\theta})$ by $v_{BS}(\hat{\theta})$ depends on the size and composition of the estimation domain. For the extreme case when the domain is composed of a single stratum, RB can be large for $n/N > 0.1$. On the other hand, when the domain is composed of a number of strata with varying fractions n_h/N_h , RB of $v_{BS}(\hat{\theta})$ could be negligible even when $n/N > 0.1$. We also compared the Rao-Wu bootstrap variance estimator with the non-resampling variance estimators for both WR and WOR sampling of PSUs. We observed that, for the large domain, $v_{BS}(\hat{\theta})$ outperforms the Sen-Yates-Grundy estimator for WOR sampling. Since, in this paper, only a sample of typical results is presented to illustrate these findings, the authors plan to include a more complete set of results in an internal report.

5. ACKNOWLEDGEMENTS

5.

The authors would like to thank Milorad Kovacevic and Claude Girard for reviewing the manuscript and providing valuable comments which lead to a number of improvements in the paper. The authors would also like to thank Georgia Roberts for her very helpful advice during the project.

REFERENCES

- Binder, D.A., Kovacevic, M.S., and Roberts, G. (2004). Design-based methods for survey data: Alternative uses of estimating functions. *Proceedings of the Joint Statistical Meetings, Section on Survey Research Methods*, 3301-3312.
- Binder, D. A. and Roberts G.R. (2003). Statistical inference for survey data analysis. *Proceedings of the Joint Statistical Meetings, Section on Survey Research Methods*, 568-572.
- Efron, B. (1979). Bootstrap methods: Another look at the Jackknife. *The Annals of Statistics*, 7, 1-26.
- Fuanoka, F., Saigo, H., Sitter, R.R., and Toida, T. (2006). Bernoulli Bootstrap for Stratified Multistage Sampling. *Survey Methodology*, 32, 151-156.

- Kovar, J.G., Rao, J. N. K. and Wu, C. F. J. (1988). Bootstrap and other methods to measure errors in survey estimates. *Canadian Journal Statistics*, 16, 25-45.
- Lohr, S. (1999). *Sampling: Design and Analysis*. Duxbury Press.
- 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.
- Saigo, H. (2007). Mean-Adjusted Bootstrap for Two-Phase Sampling. *Survey Methodology*, 33, 61-68.
- Särndal C. E., Swensson, B. and Wretman J. (1992). *Model assisted survey sampling*. Springer-Verlang. New-York, Inc.
- Shao, J. and Tu, D. (1996). *The Jackknife and Bootstrap*. Springer series in statistics. Springer-Verlag, New York.
- Statistics Canada (2003). *Survey methods and practices*. Catalogue No. 12-587-XPE.