A Study of the Properties of a Bootstrap Variance Estimator Under Sampling Without Replacement

Lenka Mach, Jean Dumais and Lauriane Robidou

Statistics Canada
Tunney's Pasture, R. H. Coats Building, Ottawa (Ontario), K1A 0T6, Canada
lenka.mach@statcan.ca, jean.dumais@statcan.ca, lauriane_robidou@hotmail.com

Abstract

This paper evaluates the performance of a bootstrap variance estimation approach in the case of a two-stage survey design with high sampling fractions at the first-stage. Multi-stage sampling is commonly used by statistical agencies, for example for household or education surveys. Typically, samples are selected without replacement (WOR) at each stage but the standard bootstrap method assumes that PSUs are selected with replacement (WR) or that the first-stage sampling fractions are negligible. Thus variance overestimation is suspected with the bootstrap method when high fractions of PSUs are sampled WOR. Modified versions of the bootstrap for designs with WOR sampling have been proposed, but these are restricted to single-stage sampling of clusters or two-stage designs with equal-probability sampling of PSUs.

We use a simulation study, based on data from the Statistics Canada Youth in Transition Survey (YITS) of 15-year-olds, to illustrate the extent of the bias as well as the stability of the Rao-Wu bootstrap variance estimator. At the first stage of YITS, a stratified sample of schools is selected and, at the second stage, students within selected schools are surveyed. Due to requirements for precise estimates in small sub-populations, PSU sampling rates in some strata are as high as 60%. So far, we have studied the properties of the bootstrap variance estimator for simple statistics - totals and means - for small, medium and large strata of schools. Our results suggest that the first-stage sampling rate is not the only factor determining the bias; the second stage sampling rate seems to play a role as well.

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 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 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 (totals, means, quantiles etc.). It measures the amount of *sampling error* in the estimate due to observing a sample instead of the whole population. 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, that is when the sample design is *informative* (Binder and Roberts 2001, 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. as discussed in 1.1 above. There are two basic approaches to variance estimation: i) an analytical approach using the *linearization* method, and ii) *resampling* and *replication* methods (jackknifing, balanced repeated replication, bootstrapping). The description of these methods can be found in many books on survey sampling, see for example Chapter 9 in Lohr (1999). Recently, there has been a move away from the analytical approach and towards the resampling approach, for a combination of reasons. The increase in computing power has made the use of the resampling techniques feasible for large survey samples. These methods are also relatively easy to implement because, regardless of the point estimator, a resampling method always uses the same procedure, replicated many times, while the linearization approach requires a development of a new formula for every estimator and weight adjustment and usually still requires additional simplifying assumptions (Binder, Kovacevic, and Roberts, 2004).

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*. Since then, there has been much theoretical and empirical research examining properties of the bootstrap estimators in the i.i.d. case and bootstrapping has become a popular tool for classical statistical analysis. 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 point 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. The following bootstrap methods that were modified for survey samples are discussed in Chapter 6 of Shao and Tu (1996):

I) The with-replacement bootstrap (McCarthy and Snowden, 1985),
II) the rescaling bootstrap (Rao and Wu, 1988, Rao, Wu and Yue, 1992),
III) the mirror-match bootstrap (Sitter, 1992a), and
IV) the without-replacement bootstrap (Gross, 1980, Chao and Lo, 1985, Bickel and Freedman, 1984, Sitter, 1992b).

At Statistics Canada, we use the rescaling bootstrap, referred to as the Rao-Wu bootstrap in this article. We describe this method in detail in Section 2.2.

2.2 Rao-Wu Bootstrap Variance Estimator

Rao and Wu (1988) proposed a bootstrap method for stratified multi-stage designs with WR sampling of PSUs that 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.

Here we describe the modified rescaling bootstrap method proposed by Rao, Wu and Yue (1992):

To estimate the variance of the estimator $\hat{\theta}$, the following steps (i) to (iv) are independently replicated *B* times, where B is quite large (typically, *B*=500 for Statistics Canada surveys).

(i) Independently in each stratum *h*, select a bootstrap sample by drawing a simple random sample of $n_h^{(b)}$ primary sampling units (PSUs) with replacement from the n_h sample PSUs. Let $t_{hi}^{(b)}$ be the number of times that PSU *hi* is selected in the bootstrap sample *b*, *b*=1,2…*B*.

(ii) For each secondary sampling unit (SSU) *k* in PSU *hi*, calculate the initial bootstrap weight by rescaling its initial sampling weight:

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

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

(iii) To obtain the final bootstrap weight $fw_{hik}^{(b)}$, adjust the initial bootstrap weight $w_{hik}^{(b)}$ by using all the same weight adjustments (e.g. non-response and calibration) that were applied to the initial sampling weight w_{hik} to produce the final survey weight fw_{hik} .

(iv) Calculate $\hat{\theta}^{(b)}$, the bootstrap replicate of estimator $\hat{\theta}$ by replacing the final survey weights fw_{hik} with the final bootstrap weights $fw_{hik}^{(b)}$ in the formula for $\hat{\theta}$. This step is discussed for example by Mantel, Nadon and Yeo (2000).

The bootstrap variance estimator of $\hat{\theta}$ is then given by

$$v_{BS}(\hat{\theta}) = \frac{1}{B}\sum_{b=1}^{B}(\hat{\theta}^{(b)} - \hat{\theta})^2 \text{, or} \tag{2a}$$

$$v_{BS}^{*}(\hat{\theta}) = \frac{1}{B}\sum_{b=1}^{B}(\hat{\theta}^{(b)} - \hat{\theta}_{BS}^{*})^2 \text{, with } \hat{\theta}_{BS}^{*} = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^{(b)} \ . \tag{2b}$$

The estimators (2a) and (2b) are Monte Carlo approximations of the bootstrap estimator of $V\left(\hat{\theta}\right)$ given by

$$\hat{V}_{BS}(\hat{\theta}) = E_{BS}\left[\hat{\theta}^{(b)} - E_{BS}\left(\hat{\theta}^{(b)}\right)\right]^2 \text{, where} \tag{2}$$

E_{BS} denotes the expectation with respect to bootstrap sampling.

Rao and Wu (1988) show that, in the case of $\hat{\theta}$ being a linear estimator, their bootstrap variance estimator (2) reduces to the standard unbiased linear estimator for WR sampling. For the nonlinear case, they show that $\hat{V}_{BS}(\hat{\theta}) = \hat{V}_L\left(\hat{\theta}\right) + O_p\left(n^{-2}\right)$, where $\hat{V}_L\left(\hat{\theta}\right)$ is the linearization variance estimator for WR sampling. Because $\hat{V}_L\left(\hat{\theta}\right)$ is a consistent estimator of $V\left(\hat{\theta}\right)$, the Rao-Wu bootstrap variance estimator $\hat{V}_{BS}\left(\hat{\theta}\right)$ is also consistent for $V\left(\hat{\theta}\right)$ when PSUs in the original design are sampled WR.

Both (2a) and (2b) are used in practice and they usually produce very similar values. However, the variance estimate (2a) is always larger than (2b), i.e. $v_{BS}(\hat{\theta}) \geq v_{BS}^{*}(\hat{\theta})$:

$$v_{BS}(\hat{\theta}) = \frac{1}{B}\sum_{b=1}^{B}(\hat{\theta}^{(b)} - \hat{\theta}_{BS}^{*} + \hat{\theta}_{BS}^{*} - \hat{\theta})^2 = \frac{1}{B}\sum_{b=1}^{B}(\hat{\theta}^{(b)} - \hat{\theta}_{BS}^{*})^2 + \left(\hat{\theta}_{BS}^{*} - \hat{\theta}\right)^2 = v_{BS}^{*}(\hat{\theta}) + \left(\hat{\theta}_{BS}^{*} - \hat{\theta}\right)^2 \ . \tag{3}$$

We see that the estimator (2a) includes a positive quantity $\left(\hat{\theta}_{BS}^{*} - \hat{\theta}\right)^2$ which converges to zero for all consistent estimators $\hat{\theta}$.

2.2.1 Size of the bootstrap sample. If $n_h^{(b)} \leq n_h - 1$, then the bootstrap weights are never negative. Usually surveys use $n_h^{(b)} = n_h - 1$, mainly because it greatly simplifies the calculation of the bootstrap weights; the rescaling formula given in (1) becomes:

$$w_{hik}^{(b)} = w_{hik}\left\{\cdot\frac{n_h}{n_h-1}\cdot t_{hi}^{(b)}\right\}. \tag{4}$$

Note that if $t_{hi}^{(b)}$ =0, i.e. PSU *hi* is not selected in the bootstrap replicate *b*, its bootstrap weight (4) is zero.
Empirical studies by Kovar, Rao and Wu (1988) demonstrated that Rao-Wu rescaling bootstrap performs well for smooth functions when $n_h^{(b)} = n_h - 1$.

2.2.2 Mean bootstrap weights. For the outside analysts, some Statistics Canada surveys produce a Public Use Microdata File (PUMF) constructed in such a way that no confidential information can be disclosed. As a result, some information about the sample design, like the stratum and cluster identifiers, is not included on the PUMF files and thus the analysts cannot use them to produce design-based variance estimates. Including the bootstrap weights on the PUMF files had been considered but it was realized that, for stratified multi-stage designs and $n_h^{(b)} = n_h - 1$, the cluster membership can be identified when the final bootstrap weights are combined over all *B* bootstrap samples. This happens because, for each bootstrap sample, the final bootstrap weights are zero for all members of at least one cluster per stratum.

To avoid breach of confidentiality, Yung (1997) proposed a modification of the Rao-Wu bootstrap procedure in which the individual bootstrap weights are replaced by the mean bootstrap weights as follows:
For each *b*, *b*=1,2…*B*, repeat the step (i) above *Q* times, i.e. independently in each stratum *h*, select a bootstrap sample by drawing an SRSWR of $n_h - 1$ PSUs from the n_h sample PSUs and repeat this selection *Q* times. Let $t_{hi(q)}^{(b)}$ be the number of times that PSU *hi* is selected in the repetition *q*, *q* =1,2…*Q*, of the bootstrap sample *b*. Calculate the average number of times the PSU *hi* is selected over the *Q* repetitions, $\bar{t}_{hi}^{(b)} = \frac{1}{Q}\sum_{q=1}^{Q} t_{hi(q)}^{(b)}$.

a) For each *b*, *b*=1,2…*B*, perform the steps (ii), (iii) and (iv) above but replace $t_{hi}^{(b)}$ in the weight rescaling formula by $\bar{t}_{hi}^{(b)}$.

The *mean bootstrap variance estimator* is given by $v_{MBS}(\hat{\theta}) = \frac{Q}{B}\sum_{b=1}^{B}(\hat{\theta}^{(b)} - \hat{\theta})^2$ or $v_{MBS}^{*}(\hat{\theta}) = \frac{Q}{B}\sum_{b=1}^{B}(\hat{\theta}^{(b)} - \hat{\theta}_{BS}^{*})^2$. (5)

Unfortunately, even with the mean bootstrap weights, the cluster membership can sometimes still be identified and therefore the mean bootstrap weights are only provided on a few PUMF files where there is no risk of disclosure. However, even if it does not fully resolve the confidentiality problem, the use of the mean bootstrap weights may be an effective bootstrapping alternative since it eliminates zero bootstrap weights and thus lets each survey observation contribute to every bootstrap estimate. In fact, some Statistics Canada surveys, that implemented bootstrapping for variance estimation, use the mean bootstrap weights and the estimator (5) rather than the estimators (2a) and (2b).

3. Simulation Study

The objective of our study is to examine the properties of the Rao-Wu rescaling bootstrap variance estimator (i.e. the modified version proposed by Rao, Wu and Yue (1992) and described in Section 2.2) when sampling of PSUs is done WOR and the first-stage sampling fractions are not negligible, say > 10%. We also plan to include in our study a limited examination of the mean bootstrap variance estimator introduced in 2.2.2 above. Since such an evaluation cannot be done theoretically, we conduct a simulation study.

We used data from the 2000 YITS/PISA survey of 15-year-olds to create a population file for our simulation. YITS is the Youth in Transition Survey, a relatively new longitudinal survey, developed by Statistics Canada in partnership with Human Resources Development Canada. It is designed to collect information that analysts and policy makers can use to better understand the experiences of youth and young adults in the education system and in the labour market. The 2000 YITS cycle of the 15-year-old cohort was integrated with PISA, the Programme for International Student Achievement launched by the Organisation for Economic Cooperation and Development (OECD). The sample for the 15-year-old cohort was selected in two stages. 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. Three different sampling plans were used, depending on the school size: i) SRSWOR sampling in the strata of small schools, ii) probability-proportional-to-size (PPS) sampling WOR in the medium-school strata, and iii) a census of the largest schools. 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 born in 1984.

3.1 Creation of Population File

We used the 2000 YITS/PISA survey frame, sample counts and survey files to create a population file for our study that simulates the actual survey population. The survey school file contained data for 1,117 participating schools and the student file contained 29,330 records with students' responses, representing the population of about 4,000 schools and close to

400,000 15-year-old students. From the many variables collected by the survey, we selected a few for our study and imputed their values when missing as follows:

First, we applied weighted hot deck imputation (the version proposed by Rao and Shao, 1992) to replace missing values on the survey school and student files (partial nonresponse), and to impute data for the schools and students that had been selected for the survey but did not respond (total nonresponse). Then, data for schools and students that were on the survey frame but not selected for the survey were mass-imputed using the hot deck method and thus the size of the simulated population is equal to the size of the actual survey population. The imputation was done within homogeneous imputation classes. Finally, we compared the distributions of different variables in our artificial population with the YITS/PISA sample estimates and concluded that our simulated population resembles well the actual survey population. (More details are available from the authors.)

3.2 Methodology Used

In our study, we examine the properties of the Rao-Wu bootstrap estimator for the two following two-stage designs:

A) Stage 1: SRSWOR of n_h schools from the population of N_h schools in each stratum *h*.

Stage 2: SRSWOR or equal-probability systematic sample of m_{hi} students from the population of M_{hi} students in each selected school *hi*.

Note: Plan A is the 2000 YITS/PISA design used for the strata of small schools.

B) Stage 1: PPSWOR of n_h schools from the population of N_h schools in each stratum *h*.

We use the school size measure available on the frame, X_{hi}, for calculating the probability of selecting school *hi*: $p_{hi} = X_{hi} / X_h$, where $X_h = \sum_{i=1}^{N_h} X_{hi}$. The sample is selected using the SAS procedure SURVEYSELECT with the option METHOD=PPS, which is based on the Hanurav-Vijayan algorithm.

Stage 2: Same as in design A.

Note: Plan B is similar to the design used by the 2000 YITS/PISA survey for the strata of medium schools. The survey applied the systematic method to select the PPSWOR sample of schools.

We vary the first-stage and the second-stage sampling fractions to study their impact on the performance of the studied variance estimators. Different variables (categorical and continuous) and estimators $\hat{\theta}$ of various parameters (total, mean, proportion, ratio and median) will be included in the study.

For each studied sampling scenario, we repeat the sample selection *R* times and for each of these simulations we select *B* bootstrap samples of schools following the method described in 2.2 above with $n_h^{(b)} = n_h - 1$ and calculate the bootstrap weights given in (4).

For each simulation *r*, *r* =1,2,…*R*, we calculate the two bootstrap variance estimates (2a) and (2b). For the case when $\hat{\theta}$ is a linear estimator, we also calculate the standard unbiased variance estimators for WOR and WR sampling. For example, for design B and $\hat{\theta} = \sum_{hik\varepsilon s} w_{hik} y_{hik}$ we obtain the Sen-Yates-Grundy variance estimator for WOR sampling

$$v_{SYG}\left(\hat{\theta}\right) = \sum_{h=1}^{H} \left[\frac{1}{2} \sum_{i=1}^{n_h} \sum_{\substack{j=1 \\ j \neq i}}^{n_h} \frac{\pi_{hi}\pi_{hj} - \pi_{h,ij}}{\pi_{h,ij}} \left(\frac{\hat{\theta}_{hi}}{\pi_{hi}} - \frac{\hat{\theta}_{hj}}{\pi_{hj}} \right)^2 + \sum_{i=1}^{n_h} \frac{\hat{V}\left(\hat{\theta}_{hi}\right)}{\pi_{hi}} \right] \tag{6}$$

and the usual estimator for WR sampling

$$v_{WR}\left(\hat{\theta}\right) = \sum_{h \in H} \left[\frac{1}{n_h (n_h - 1)} \sum_{i=1}^{n_h} \left(\frac{\hat{\theta}_{hi}}{p_{hi}} - \hat{\theta}_h \right)^2 \right]. \tag{7}$$

In both (6) and (7), $\hat{\theta}_{hi}$ is an unbiased estimator of the school *hi* total, $\theta_{hi} = \sum_{k=1}^{M_{hi}} y_{hik}$. In (6), $\hat{V}\left(\hat{\theta}_{hi}\right)$ is an unbiased estimator of the second-stage variance $V\left(\hat{\theta}_{hi}\right)$, π_{hi} is the probability of selecting school *hi* in the PPSWOR sample, equal to $n_h p_{hi}$, and $\pi_{h,ij}$ is the joint selection probability for schools *hi* and *hj*. The joint probability $\pi_{h,ij}$ is calculated for all pairs of PSUs selected in the stratum *h* sample by the SURVEYSELECT SAS procedure. In (7), $\hat{\theta}_h$ is an unbiased estimator of the stratum *h* total, calculated as $\sum_{i=1}^{n_h} \frac{\hat{\theta}_{hi}}{p_{hi}}$. Note that we use (7) to estimate the variance for WOR sampling and thus each selected school contributes one unique value of $\hat{\theta}_{hi}$ to the sum within the square brackets. When used for WR sampling, an estimate $\hat{\theta}_{hi}$ is included as many times as this school was selected in the first-stage sample, each time based on a different second-stage sample of m_{hi} students. The different student samples selected within one school must be independent of each other.

When $\hat{\theta}$ is a nonlinear estimator, we can obtain analytical variance estimators by linearization and examine how their properties compare to those of the bootstrap estimators.

To examine the *accuracy* and *stability* of the rescaling bootstrap estimator and to compare these properties with those of the analytical methods, we obtain the empirical (Monte Carlo) expectations by averaging various types of estimates over all *R* simulations. The *empirical sampling variance* of $\hat{\theta}$ is

$$V_R\left(\hat{\theta}\right)=\frac{1}{R}\sum_{r=1}^{R}\left[\hat{\theta}_r - E_R\left(\hat{\theta}\right)\right]^2 \text{, where } E_R\left(\hat{\theta}\right)=\frac{1}{R}\sum_{r=1}^{R}\hat{\theta}\,. \qquad (8)$$

The empirical expectation and variance for each studied variance estimator (*type* = BS, SYG, WR) are respectively

$$E_R\left[v_{type}\left(\hat{\theta}\right)\right]=\frac{1}{R}\sum_{r=1}^{R}v_{type}\left(\hat{\theta}\right) \text{ and} \qquad (9)$$

$$V_R\left[v_{type}\left(\hat{\theta}\right)\right]=\frac{1}{R}\sum_{r=1}^{R}\left[v_{type}\left(\hat{\theta}\right)-V_R\left(\hat{\theta}\right)\right]^2 . \qquad (10)$$

To evaluate the accuracy of the studied variance estimators, we calculate *relative bias*

$$RB_R\left[v_{type}\left(\hat{\theta}\right)\right]=\frac{E_R\left[v_{type}\left(\hat{\theta}\right)\right]}{V_R\left(\hat{\theta}\right)}-1 \ . \qquad (11)$$

To evaluate the stability, we use *relative root mean square error*

$$RS_R\left[v_{type}\left(\hat{\theta}\right)\right]=\frac{\sqrt{V_R\left[v_{type}\left(\hat{\theta}\right)\right]}}{V_R\left(\hat{\theta}\right)} \ , \qquad (12)$$

and also examine the empirical distributions for the studied variance estimators.

3.3 Results

Here we present results for the estimate of a population total. The variable of interest *Y* is the student's agreement with the statement "Most of the time, I would like to be any place other than in school." Thus *Y* is dichotomous variable with $y_{hik} = 1$ for student *hik* that agrees with the statement and $y_{hik} = 0$ otherwise. The parameter of interest, t_Y, is the total number of students who agree with the statement and its unbiased estimate is

$$\hat{\theta} \ = \ \hat{t}_Y \ = \ \sum_{hik \in s} w_{hik}\, y_{hik} \ . \qquad (13)$$

We examine the variance estimation for plan B in a medium-size-school stratum. Its characteristics are shown in Table 1.

Table 1: Stratum characteristics

Stratum ID	N_h	X_h	Range of X_{hi}	M_h	Range of M_{hi}	$t_{Y,h}$	$p_{Y,h}$
58	25	1,328	35 – 78	1,316	15 - 91	667	51%

In the first-stage, we select a PPSWOR of n_h schools (n_h = 3, 5, 10), and in the second-stage an SRSWOR of m_{hi} students (m_{hi} = 5, 10, 30). If $m_{hi} \geq M_{hi}$, all students in school *hi* are included in the sample, i.e. $m_{hi} = M_{hi}$. For each combination of n_h and m_{hi} that yields a total sample m_h of at least 30 students, we use *R* = 10,000 to obtain the empirical sampling variance

$V_R\left(\hat{\theta}\right)$, and R = 5,000 to get the empirical expectation $E_R\left[v_{type}\left(\hat{\theta}\right)\right]$ and variance $V_R\left[v_{type}\left(\hat{\theta}\right)\right]$, given in (9) and (10) above, for each studied variance estimator (*type* = BS, SYG, WR). For r = 1,…, 5,000, 100 bootstrap samples are selected to calculate $v_{BS}^{*}\left(\hat{\theta}\right)$ defined in (2b). (We decided that using B=100 was sufficient for our population and simple estimator (13), after comparing $v_{BS}^{*}\left(\hat{\theta}\right)$ based on 100 and 500 bootstrap samples for some combinations of n_h and m_{hi}.) The relative bias is given in Table 2 and the relative stability in Table 3 below.

Table 2: Empirical relative bias

n_h	m_{hi}	$V_R\left(\hat{\theta}\right)$	Relative Bias RB_R				
			$v_{SYG}\left(\hat{\theta}\right)$	$v_{WR}\left(\hat{\theta}\right)$	$v_{BS}^{*}\left(\hat{\theta}\right)$	$\left(1-n_h/N_h\right)v_{BS}^{*}\left(\hat{\theta}\right)$	$\left(1-m_h/M_h\right)v_{BS}^{*}\left(\hat{\theta}\right)$
3	10	22, 638	-0.000	0.058	0.049	-0.077	0.025
3	30	12,605	-0.006	0.096	0.088	-0.042	0.014
5	10	13,206	-0.036	0.057	0.046	-0.163	0.007
5	30	6,979	-0.026	0.154	0.139	-0.088	0.010
10	5	10,476	-0.016	0.099	0.088	-0.347	0.047
10	10	5,672	0.029	0.256	0.246	-0.253	0.151
10	30	2,819	-0.004	0.458	0.447	-0.132	0.117

Table 3: Empirical relative root mean square error

n_h	m_{hi}	$V_R\left(\hat{\theta}\right)$	Relative RMSE RS_R		
			$v_{SYG}\left(\hat{\theta}\right)$	$v_{WR}\left(\hat{\theta}\right)$	$v_{BS}^{*}\left(\hat{\theta}\right)$
3	10	22, 638	0.965	1.098	1.107
3	30	12,605	1.056	1.189	1.199
5	10	13,206	0.620	0.776	0.783
5	30	6,979	0.689	0.858	0.864
10	5	10,476	0.303	0.530	0.543
10	10	5,672	0.368	0.658	0.679
10	30	2,819	0.421	0.813	0.841

4. Conclusion

In this paper, we present initial results of our investigation of the properties of the Rao-Wu rescaling bootstrap variance estimator when sampling of PSUs is done WOR and the first-stage sampling fractions are not negligible. We use a simulation study based on the data from the 2000 YITS/PISA survey. The results for one stratum, PPSWOR of PSUs and the simple linear estimator of the population total, $\hat{\theta} = \hat{t}_Y = \sum_{hik \in s} w_{hik}\, y_{hik}$, are presented in Section 3.3.

The statistics in Table 2 above indicate that the bootstrap variance estimator developed for WR sampling of PSUs overestimates the sampling variance when used for WOR sampling and the first-stage fractions exceed 10%. For a given second-stage sample size m_{hi}, the relative bias of $v_{BS}^{*}\left(\hat{\theta}\right)$ increases as the first-stage fraction n_h/N_h increases. Our results also suggest that the first-stage sampling rate is not the only factor determining the bias; the second stage sampling rate seems to play a role as well. For a given first-stage sample size n_h, the relative bias of $v_{BS}^{*}\left(\hat{\theta}\right)$ increases as the second-stage sample size m_{hi} increases. The simple adjustment of multiplying the stratum sampling variance estimate by $\left(1-n_h/N_h\right)$ severely "over-corrects" the estimate, thus yielding a negative bias, and should never be used for multi-stage designs. On the other hand, the adjustment based on the stratum student sampling fraction $\left(1-m_h/M_h\right)$ seems to work reasonably well in our example. The values of RB_R for $v_{BS}^{*}\left(\hat{\theta}\right)$ and $v_{WR}\left(\hat{\theta}\right)$ are very similar but not identical because $v_{BS}^{*}\left(\hat{\theta}\right)$ is a Monte Carlo approximation, based on $B = 100$, of the true bootstrap estimator given in (2) which, for a linear $\hat{\theta}$, reduces to $v_{WR}\left(\hat{\theta}\right)$. The empirical values of RB_R for $v_{SYG}\left(\hat{\theta}\right)$ are all very small as expected for an unbiased estimator; the true $RB\left[v_{SYG}\left(\hat{\theta}\right)\right]$ is zero and the empirical values are its Monte Carlo approximations.

The empirical relative root mean square error values in Table 3 show that the Sen-Yates-Grundy variance estimator for WOR sampling is less variable and thus more stable than both the usual estimator for WR sampling given in (7) and the Rao-Wu bootstrap estimator; the relative stability of the latter two is almost the same. In fact, $RS_R\left[v_{WR}\left(\hat{\theta}\right)\right]$ is always just slightly smaller than $RS_R\left[v_{BS}^{*}\left(\hat{\theta}\right)\right]$ because $v_{BS}^{*}\left(\hat{\theta}\right)$ is an approximation of the true bootstrap sampling variance estimator based on B=100. For a given second-stage sample size m_{hi}, the empirical RS_R decreases for all three variance estimators, and hence their relative stability improves, as the first-stage fraction n_h/N_h increases. On the other hand, for a given first-stage sample size n_h, the empirical RS_R increases for all three variance estimators as the second-stage sample size m_{hi} increases.

We plan to use the population file and the methodology described in Section 3 to investigate the accuracy and stability of the Rao-Wu rescaling bootstrap variance estimator for other strata, larger domains, different variables, estimators and designs.

5. Acknowledgements

The authors would like to thank Georgia Roberts for her invaluable advice and comments, Zdenek Paták for his review of the paper, Joanne Moloney for her contribution in the planning of the study and the paper discussant, Phil Kott, for his questions and comments.

6. References

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

Binder, D. A. and Roberts G.R. (2001). Can informative designs be ignorable? *Newsletter of the Survey Research Methods Section,* Issue 12, American Statistical Association.

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.

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.

Chao, M.T. and Lo, S.H. (1985). A bootstrap method for finite populations. *Sankhya A*, 47, 399-405.

Efron, B. (1979). Bootstrap methods: Another look at the Jackknife. *The Annals of Statistics*, 7, 1-26.

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

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.

Mantel, H.J., Nadon, S. and Yeo, D. (2000). Effect of nonresponse adjustments on variance estimates for the National Population Health Survey. *Proceedings of the Joint Statistical Meetings, Section on Survey Research Methods,* 521-226.

McCarthy, P.J. and Snowden, C.B. (1985). The bootstrap and finite population sampling. *Vital and Health Statistics*, 2-95, Public Health Service Publication 85-1369, U.S. Government Printing Office, Washington, D.C.

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

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

Rao, J. N. K., Wu, C. F. J. and Yue, K. (1992). Some recent work on resampling methods for complex surveys, *Survey Methodology*, 18, pp.209-217.

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.

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

Sitter, R.R. (1992b). Comparing three bootstrap methods for survey data. *Canadian Journal of Statistics*, 20, 135-154.

Statistics Canada (2003). *Survey methods and practices*. Catalogue No. 12-587-XPE.

Yung, W. (1997). Variance estimation for public use microdata files. *Proceedings of the Statistics Canada Symposium,* 91-95.