Resampling for Estimation and Inference for Variances

Thompson, Mary (1st author)
University of Waterloo, Department of Statistics and Actuarial Science
200 University Avenue West
Waterloo N2L 3G1, Canada
E-mail: methomps@uwaterloo.ca

Wang, Zilin (2nd author)
Wilfrid Laurier University, Department of Mathematics
75 University Avenue West
Waterloo N2L 3C5, Canada
E-mail: zwang@wlu.ca

1 Introduction

In survey sampling, estimating and making inferences on population variances can be challenging when sampling designs are complicated. Most of the literature limits the investigation to the case where a random sample is obtained without replacement (For example, Thompson (1997) and Cho and Cho (2008)). In this paper, we pursue a computational approach to inference for population variances and, ultimately, variance components with complex survey data.

This computational approach is based on a resampling technique, artificial population bootstrapping (APB), introduced in Wang and Thompson (2010). The APB procedure departs from the common application of resampling procedures in survey sampling. It belongs to the category of bootstrapping without replacement, with unequal probability sampling being used to select the resamples. In particular, using a set of complex survey data, many artificial populations are built, and within each of the those populations, resamples are selected by using the same probability sampling design as the original sample. The APB procedure entails not only mimicking the sampling design, but also creating artificial populations to resemble as closely as possible the population from which the original sample is selected. As a result, inferences on parameters for the artificial population using resamples would resemble inferences on parameters from the finite population using the original sample. Because full information on the artificial populations and the resamples is available, point and interval estimates of the artificial population parameters are computable. That is, we can use the empirical inferences on the artificial population to solve inferential issues for finite population parameters.

Based on this principle, we can estimate the variance of an estimator, for example an estimator of a total or variance, and estimate percentiles of a point estimator so as to obtain interval estimates. Wang and Thompson (2010) used the APB procedures to correct the bias of variance components estimators in a multilevel model. To solve a potential bias problem caused in part by reduced variability of artificial populations, we have also investigated a "double APB" procedure Wang and Thompson (2011) for use in interval estimation and when estimating the variance of a sample variance.

The survey sampling literature on inference for variance components of a multilevel model includes other methods, for example, the Taylor linearization approach in Pfeffermann et al. (1998) and Rabe-Hesketh and Skrondal (2006) and a computational approach in Kovacević et al. (2006).

Note that the computational approach in Kovačević et al. (2006) is based on bootstrapping with replacement.

In unequal probability sampling designs, both single stage and multistage, the inclusion probabilities of the units are often informative (see for example Pfeffermann et al. (1998)). When the aim is analysis this makes it important to consider a conceptual two-phase structure in which the finite population under study is generated from a superpopulation, the parameters of which are the objects of inference, and the sample is taken from a probability sampling design. However, in this paper, we consider for simplicity that finite population quantities such as totals and variances are to be estimated.

The paper is organized as follows. Section 2 describes the artificial population bootstrap algorithm for probability sampling; Section 3 implements the APB in estimation of variance of a Horvitz-Thompson estimator, with some explanation of the justification. Section 4 describes a way to set confidence intervals for population variances using the APB. Four approaches are proposed in Section 5 to estimate the variance of sample a variance. Section 6 describes the simulation studies, and Section 7 presents conclusions.

2 The Artificial Population Bootstrap (APB)

Suppose there is a set $U = \{1, \cdots, N\}$ indexing the finite population, and y_i is a real number associated with the i^{th} population unit. The sample is denoted by $\mathbf{s} \subset U$, and is chosen by a probability sampling design $p(\mathbf{s})$ with fixed size n, using a method which produces specified inclusion probabilities. Denote by π_i the inclusion probability of unit i. Let $k_i = \lfloor 1/\pi_i \rfloor$, where $\lfloor \rfloor$ signifies the greatest integer less than or equal to its argument. The proposed method, which is illustrated in Figure 1, follows these steps:

1. For $i \in \mathbf{s}$, make k_i copies of (y_i, π_i) to form a partial artificial population, U_1^a.
2. Use Bernoulli sampling to select a sample, U_2^a, so that i from the sample (with (y_i, π_i)) is included again in the artificial population with probability $r_i = (1/\pi_i) - \lfloor 1/\pi_i \rfloor$.
3. Combine U_1^a and U_2^a to create an artificial population U^a.
4. Within the artificial population U^a, compute inclusion probabilities π_i^* proportional to the π_i and summing to m; select R random samples, $\mathbf{s}^*$, using the probability design $p(\mathbf{s})$ with the new inclusion probabilities and fixed sample size m.
5. Repeat steps 1 to 4 Q times.

Note that the size of the artificial population from Step 3 will not be N in general.

In an artificial population, if the sample unit i occurs a_i times, the sample pair (i, j) occurs $a_i a_j$ times. The variables a_i and a_j are independent, conditional on the sample $\mathbf{s}$. Given that $a_i = k_i + \zeta_i$ where $\zeta_i = 1$ with probability r_i and $\zeta_i = 0$ with probability $1 - r_i$, we have $E_{art} a_i = k_i + r_i = w_i = 1/\pi_i$ and $Var_{art}(a_i) = r_i(1 - r_i)$, where E_{art} and Var_{art} denote moments with respect to the generation of the artificial population, conditional on the sample $\mathbf{s}$. The first order inclusion probabilities in the artificial population are $\pi_i^a = m\pi_i / (\sum_{i \in \mathbf{s}} a_i \pi_i)$, and higher order inclusion probabilities such as π_{ij}^a are functions of the design and the π_i^a's.

Figure 1: The Artificial Population Bootstrapping Procedure

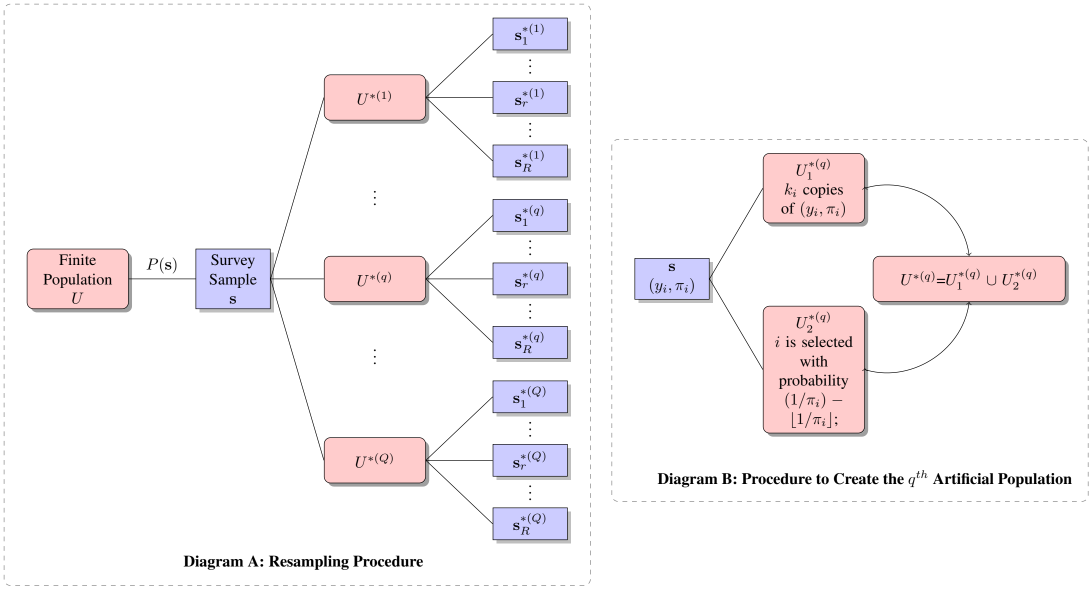


Diagram A: Resampling Procedure

Diagram B: Procedure to Create the q^{th} Artificial Population

3 Estimation of variance of a Horvitz-Thompson estimator

The Horvitz-Thompson estimator estimates a population total $Y = \sum_{i=1}^{N} y_i$. It takes the form

$$\hat{Y}_{HT} = \sum_{i \in \mathbf{s}} \frac{y_i}{\pi_i}.$$

Its sampling variance is

$$(1) \quad Var_p(\hat{Y}_{HT}) = \frac{1}{2} \sum_{i}^{N} \sum_{j}^{N} (\pi_i \pi_j - \pi_{ij}) (\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j})^2,$$

which has as unbiased estimator

$$(2) \quad v(\hat{Y}_{HT}) = \frac{1}{2} \sum_{i \in \mathbf{s}} \sum_{j \in \mathbf{s}} \frac{(\pi_i \pi_j - \pi_{ij})}{\pi_{ij}} (\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j})^2.$$

Conditional on the sample $\mathbf{s}$, the expectation of the artificial population total, $Y_a = \sum_{i \in \mathbf{s}} a_i y_i$, is $\hat{Y}_{HT}$.

In resamples of size m from the artificial population, the corresponding Horvitz-Thompson estimators $\hat{Y}^*_{HT}$ will be unbiased estimators of Y_a. If we took one artificial population and many resamples, the empirical variance of $\hat{Y}^*_{HT}$ would estimate

$$(3) \quad Var_{p^a}(\hat{Y}^*_{HT}) = \frac{1}{2} \sum_{i \in \mathbf{s}} \sum_{j \in \mathbf{s}} a_i a_j (\pi_i^a \pi_j^a - \pi_{ij}^a) (\frac{y_i}{\pi_i^a} - \frac{y_j}{\pi_j^a})^2,$$

which we would like to be close to the unbiased estimator $v(\hat{Y}_{HT})$.

Simulation studies not shown here confirm that (3) has typically only a small relative bias as an estimator of (1) for both systematic pps and Rao-Sampford designs provided the sample size is not too small and the inclusion probabilities are not informative and highly variable.

Using the with replacement approximation $\pi^a_{ij} \simeq (m-1)\pi^a_i\pi^a_j/m$, we have

$$(4) \quad Var_{p^a}(\hat{Y}^*_{HT}) \simeq \frac{1}{2}\sum_{i\in\mathbf{s}}\sum_{j\in\mathbf{s}} a_i a_j \frac{\pi^a_i\pi^a_j}{m}(\frac{y_i}{\pi^a_i}-\frac{y_j}{\pi^a_j})^2.$$

The right hand side of (4) is

$$(5) \quad \frac{1}{2}\sum_{i\in\mathbf{s}}\sum_{j\in\mathbf{s}} a_i a_j \frac{m^2\pi_i\pi_j}{m(\sum_{i\in\mathbf{s}} a_i\pi_i)^2}\frac{(\sum_{i\in\mathbf{s}} a_i\pi_i)^2}{m^2}(\frac{y_i}{\pi_i}-\frac{y_j}{\pi_j})^2,$$

and E_{art} of (5) is approximately equal to $(n-1)v(\hat{Y}_{HT})/m$. Thus if we take $m = n-1$, the empirical variance of $\hat{Y}^*_{HT}$ will be an approximately unbiased estimator of $Var_p(\hat{Y}_{HT})$, although there may be a small bias associated with the particular artificial population generated.

If we take many artificial populations with many resamples of size m each, the average (over the artificial populations) of the empirical variance of $\hat{Y}^*_{HT}$ over the number of resamples within an artificial population will estimate $Var_p(\hat{Y}_{HT})$ better.

Under regularity conditions benefits of using the APB for estimating the variance of a Horvitz-Thompson estimator will extend to estimating variances of estimators which are solutions of sample estimating equations and estimating equation systems – one such case being an inflation estimator of a population variance.

4 Estimation of confidence intervals for S^2

With the APB algorithm, we have the variance of y in the artificial population and the resample inflation estimator of this quantity, as follows:

$$(6) \quad S^2_a = \frac{1}{N_a-1}\sum_{i=1}^{N_a}(y_i-\bar{Y}_a)^2,$$

and

$$(7) \quad s^{2*} = \sum_{i\in\mathbf{s}^*} w^a_i(y_i-\bar{y}_{\mathbf{s}^*})^2/\sum_{i\in\mathbf{s}^*} w^a_i,$$

where $\bar{y}_{\mathbf{s}^*} = \sum_{i\in\mathbf{s}^*} w^a_i y_i/\sum_{i\in\mathbf{s}^*} w^a_i$ is the resample mean of y.

If $s^2_\pi/S^2_y \sim G(\xi)$ then a two-sided $(1-\alpha)\%$ confidence interval is expressed as

$$(8) \quad \frac{s^2_\pi}{\xi_{1-\alpha/2}} \le S^2_y \le \frac{s^2_\pi}{\xi_{\alpha/2}}$$

where $\xi_{1-\alpha/2}$ and $\xi_{\alpha/2}$ are the $(1-\alpha/2)^{th}$ and $(\alpha/2)^{th}$ quantiles of G. To estimate the interval, conditional on the sample, we assume that $E_{art}G^a(\xi)$, where $G^a(\xi)$ is the distribution (conditional on the sample and artificial population) of s^{2*}/S^2_a, resembles $G(\xi)$. Hence, if $\hat{\xi}_{1-\alpha/2}$ and $\hat{\xi}_{\alpha/2}$ are the $(1-\alpha/2)^{th}$ and $(\alpha/2)^{th}$ quantiles of this distribution, the interval $s^2_\pi/\hat{\xi}_{1-\alpha/2} \le S^2_y \le s^2_\pi/\hat{\xi}^{\alpha/2}$ may be close to the confidence interval in (8). To estimate the β^{th} percentile, $\hat{\xi}_\beta$, we follow the following procedures.

1. From the original sample, generate Q artificial populations and R resamples within each artificial population;

2. In the q^{th} artificial population, compute the artificial population variance, S_{aq}^2, and resample variance estimates, $s_{q1}^{2*}, \cdots, s_{qR}^{2*}$, using (6) and (7); order the ratios, $s_{q1}^{2*}/S_{aq}^2, \cdots, s_{qR}^{2*}/S_{aq}^2$ to get the empirical percentiles $\hat{\xi}_\beta^q$ for the q^{th} artificial population;

3. Repeat step 2 for all the artificial populations to provide $\hat{\xi}_\beta^1, \cdots, \hat{\xi}_\beta^Q$; the average of the empirical percentiles could be taken as the estimate $\hat{\xi}_\beta$ of the β^{th} percentile ξ_β. That is,

$$\hat{\xi}_\beta = \frac{1}{Q}\sum_{q=1}^{Q}\hat{\xi}_\beta^q.$$

Using $\hat{\xi}_{\alpha/2}$ and $\hat{\xi}_{1-\alpha/2}$, we have the estimated confidence interval for S_y^2 as

$$s_\pi^2/\hat{\xi}_{1-\alpha/2} \leq S_y^2 \leq s_\pi^2/\hat{\xi}_{\alpha/2}.$$

5 Estimating the variance of a bias-corrected variance estimator

There are several resampling approaches to this problem, involving either the APB as already described, or a double APB bootstrap. In the double APB resampling procedure, as illustrated in Figure 2, the same APB procedures as described in Figure 1 are repeated in two generations. For $q = 1, \cdots, Q$, for $q^* = 1, \cdots, Q^*$, for $r = 1, \cdots, R$ and for $r^* = 1, \cdots, R^*$:

1. Create artificial populations, $U_1^a, \cdots, U_Q^a$, from the original sample and within the q^{th} population, construct R resamples, $\mathbf{s}_{q1}^*, \cdots, \mathbf{s}_{qR}^*$.

2. In the q^{th} artificial population, for the r^{th} resample, $\mathbf{s}_{qr}^*$, create artificial populations $U_{qr1}^{a*}, \cdots, U_{qrQ^*}^{a*}$ and within the q^{*th} artificial population, construct R^* resamples, $\mathbf{s}_{qrq^*1}^{**}, \cdots, \mathbf{s}_{qrq^*R^*}^{**}$.

   We have

$$S_y^2 = \frac{1}{N-1}\left[\sum_{i=1}^{N} y_i^2 - \frac{(\sum_{i=1}^N y_i)^2}{N}\right] = \frac{1}{2N(N-1)}\sum_{i}^{N}\sum_{j}^{N}(y_i - y_j)^2.$$

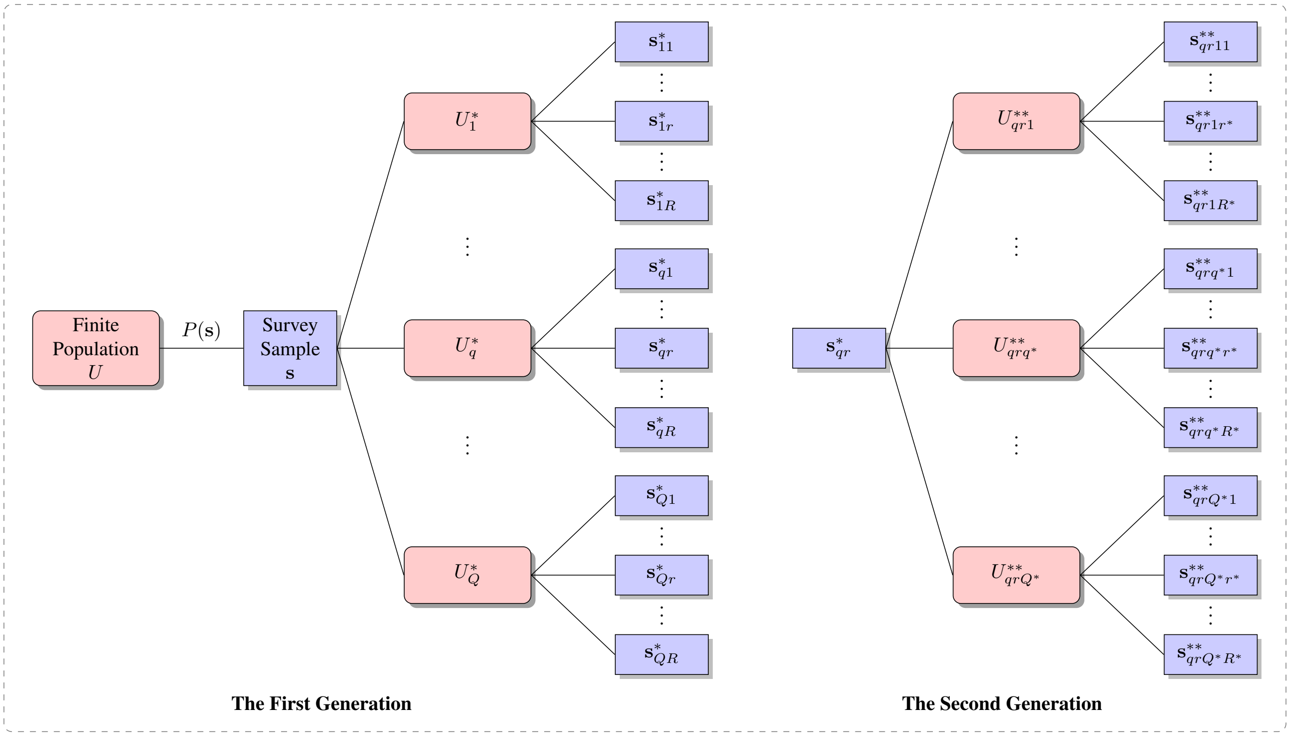
The inflation estimator of variance s_π^2 can also be written as

$$s_1^2 = \frac{\sum_{i\in\mathbf{s}}\frac{y_i^2}{\pi_i}}{\sum_{i\in\mathbf{s}}\frac{1}{\pi_i}} - \frac{(\sum_{i\in\mathbf{s}}\frac{y_i}{\pi_i})^2}{(\sum_{i\in\mathbf{s}}\frac{1}{\pi_i})^2}.$$

An unbiased estimator is

$$s_2^2 = \frac{1}{2N(N-1)}\sum_{i\in\mathbf{s}}\sum_{j\in\mathbf{s}}\frac{(y_i - y_j)^2}{\pi_{ij}}.$$

Figure 2: The Double Resampling Procedure



It is possible to show that $s_1^2 \simeq (n-1)s_2^2/n$ under the with replacement approximation, consistently with the bias correction factor for s_1^2 used in Wang and Thompson (2010) being approximately $n/(n-1)$. For

$$s_1^2 = \frac{1}{2(\sum_{i\in\mathbf{s}} \frac{1}{\pi_i})^2} \sum_{i\in\mathbf{s}} \sum_{j\in\mathbf{s}} \frac{1}{\pi_i}\frac{1}{\pi_j}(y_i - y_j)^2,$$

and $\sum_{i\in\mathbf{s}} 1/\pi_i$ is an unbiased estimator of N, while $\pi_{ij} \simeq \frac{n-1}{n}\pi_i\pi_j$.

5.1 Linearization of inflation estimator

In the linearization approach, we note that

$$s_1^2 = \frac{\hat{T}_3}{\hat{T}_1} - \frac{\hat{T}_2^2}{\hat{T}_1^2},$$

estimating

$$\frac{T_3}{T_1} - \frac{T_2^2}{T_1^2} = \frac{1}{N}\left(\sum_{i=1}^{N} y_i^2 - \frac{(\sum_{i=1}^{N} y_i)^2}{N}\right) = \frac{N-1}{N}S_y^2.$$

Then

$$s_1^2 - \frac{(N-1)}{N}S_y^2 \simeq \frac{1}{T_1}\left[(\hat{T}_3 - T_3) - 2\bar{Y}(\hat{T}_2 - T_2) - \frac{T_3}{T_1}(\hat{T}_1 - T_1) + 2\bar{Y}^2(\hat{T}_1 - T_1)\right],$$

and

$$Var(s_1^2) \simeq \frac{1}{T_1^2} Var(\sum_{i \in \mathbf{s}} w_i z_i) = \frac{1}{N^2} Var(\sum_{i \in \mathbf{s}} w_i z_i)$$

where

$$z_i = y_i^2 - 2\bar{Y} y_i - \frac{1}{N} \sum_{k=1}^{N} y_k^2 + 2\bar{Y}^2 = (y_i - \bar{Y})^2 - \frac{N-1}{N} S_y^2.$$

This variance could be estimated analytically if $\bar{Y}$ and S_y^2 were known; in the estimator so formed, estimates of these population quantities could be substituted. Alternatively, we could construct an artificial population and resample from it many times, and calculate the bootstrap variance as the value for the artificial population of the empirical variance of

$$\frac{1}{N_a} \sum_{i \in \mathbf{s}^*} w_i^a z_i^a$$

where $w_i^a = 1/\pi_i^a$ and $z_i^a = (y_i - \bar{Y}_a)^2 - \frac{N_a - 1}{N_a} S_{y,art}^2$. It is expected that a bias correction might be required, using double resampling. This estimator without the bias correction is denoted by $v_{plin}(s^2)$ in Section 6.3.

5.2 Modified linearization of inflation estimator

From the new form of s_1^2 above, we have

$$s_1^2 = \frac{1}{2} \sum_{i \in \mathbf{s}} \sum_{j \in \mathbf{s}} w_i w_j (y_i - y_j)^2 / \sum_{i \in \mathbf{s}} \sum_{j \in \mathbf{s}} w_i w_j.$$

Define S_π^2 by

$$Es_1^2 \simeq \frac{1}{2} \sum_{i}^{N} \sum_{j}^{N} \frac{\pi_{ij}}{\pi_i \pi_j} (y_i - y_j)^2 / \sum_{i}^{N} \sum_{j}^{N} \frac{\pi_{ij}}{\pi_i \pi_j} = S_\pi^2.$$

Then

$$s_1^2 - Es_1^2 \simeq \frac{1}{2 \sum_{i \in \mathbf{s}} \sum_{j \in \mathbf{s}} w_i w_j} [\sum_{i \in \mathbf{s}} \sum_{j \in \mathbf{s}} w_i w_j \{(y_i - y_j)^2 - S_\pi^2\}].$$

We can estimate the variance of s_1^2 as $(\sum_{i \in \mathbf{s}} w_i)^{-4}$ times an estimate of the variance of

$$\frac{1}{2} \sum_{i \in \mathbf{s}} \sum_{j \in \mathbf{s}} w_i w_j \{(y_i - y_j)^2 - S_\pi^2\}.$$

An analytic expression for an unbiased estimator of this variance would use joint inclusion probabilities up to fourth order. Alternatively, we could estimate the second part (numerator) using an APB. This would require that the empirical variance of the statistic in the artificial population be close to an unbiased estimator of its variance in the original population.

5.3 Estimating moments of unbiased estimator

Alternatively, we could work with the unbiased estimator s_2^2.

$$Es_2^2 = \frac{1}{2N(N-1)} \sum_i^N \sum_j^N (y_i - y_j)^2.$$

$$Var_p(s_2^2) = \frac{1}{4N^2(N-1)^2} \sum_i^N \sum_j^N \sum_k^N \sum_l^N C_{ij,kl}(y_i - y_j)^2(y_k - y_l)^2$$

where

$$C_{ij,kl} = \frac{\pi_{ijkl}}{\pi_{ij}\pi_{kl}} - 1.$$

If $u_{ij} = (y_i - y_j)^2$, so that $u_{ii} = 0$, then

$$(9)\quad Var_p(s_2^2) = \frac{1}{4N^2(N-1)^2}[2\sum_{i,j}^N C_{ij,ij}u_{ij}^2 + \sum_{i\neq}^N \sum_{j\neq}^N \sum_{k\neq}^N \sum_l^N C_{ij,kl}u_{ij}u_{kl} + 4\sum_i^N \sum_j^N \sum_k^N C_{ij.ik}u_{ij}u_{ik}].$$

The equation (9) reduces for SRS to

$$(10)\quad Var_p(s^2) = \frac{f_0}{2} \cdot A_U + \frac{f_1 - 3f_0}{2} \cdot B_U,$$

where s^2 is the sample variance, $f_0 = 1/n - 1/N$, $f_1 = 1/(n-1) - 1/(N-1)$, $A_U = \sum\sum_{i\neq j\in U} u_{ij}^2/[N(N-1)]$ and $B_U = \sum\sum\sum\sum_{i\neq j\neq k\neq l\in U} u_{ij}u_{kl}/[N(N-1)(N-2)(N-3)]$.

The key to proceeding analytically is to note that in (9) or(10) each of the major terms has an unbiased estimator from the sample. For example, for SRS, we can obtain an unbiased estimator of $Var_p(s^2)$ as

$$(11)\quad v_{pu}(s^2) = \frac{f_0}{2} \cdot A_{\mathbf{s}} + \frac{f_1 - 3f_0}{2} \cdot B_{\mathbf{s}}.$$

where $A_{\mathbf{s}} = \sum\sum_{i\neq j\in\mathbf{s}} (y_i - y_j)^4/[n(n-1)]$

and $B_{\mathbf{s}} = \sum\sum\sum\sum_{i\neq j\neq k\neq l\in\mathbf{s}} (y_i - y_j)^2(y_k - y_l)^2/[n(n-1)(n-2)(n-3)]$. Exploring an APB approach in the SRS case, we note that analogous to the original population quantities A_U and B_U, we have the following artificial population quantities A^a and B^a, which can be expressed as

$$A^a = \sum\sum_{i\neq j\in\mathbf{s}} a_i a_j u_{ij}^2/[N(N-1)]$$

and

$$\begin{aligned} B^a \;\; = \;\; & \{\sum\sum\sum\sum_{i\neq j\neq k\neq l\in\mathbf{s}} a_i a_j a_k a_l u_{ij}u_{kl} + 2\sum\sum_{i\neq j\in\mathbf{s}} a_i(a_i - 1)a_j(a_j - 1)u_{ij}^2 \\ & +4\sum\sum\sum_{i\neq \jmath\neq k\in\mathbf{s}} a_i(a_i - 1)a_j a_k u_{ij}u_{ik}\}/[N(N-1)(N-2)(N-3)]. \end{aligned}$$

If $N = nk$, then after some algebra we obtain

$$(12)\quad E_{art}[A^a] = bA_{\mathbf{s}}$$

where $b = (n-1)N/(N-1)n$ and

$$(13)\quad E_{art}[B^a] = dA_{\mathbf{s}^*} + cB_{\mathbf{s}^*}$$

where $d = 2k(k-1)^2n(n-1)((n+1)k-2n-2k+4)/N^{(4)}$ and $c = k^4n^{(4)}/N^{(4)}$.

Suppose that in a double APB , $\mathbf{s}^*$ is a resample from artificial population U^q, and U^* is s corresponding artificial population from $\mathbf{s}^*$. We consider the following two regression models:

$$(14)\quad E_{art^*}[A^{a*}] = bA_{\mathbf{s}^*}$$

and

$$(15)\quad E_{art^*}[B^{a*}] = dA_{\mathbf{s}^*} + cB_{\mathbf{s}^*}.$$

Using least squares estimation, we can estimate d, b and c using the regression models on (12) and (13), $\hat{d}$, $\hat{b}$ and $\hat{c}$, and obtain the following approximately unbiased estimates of $Var_p(s^2)$:

$$(16)\quad v_{pr}(s^2) = \frac{f_0}{2}\hat{b}A^a + \frac{f_1 - 3f_0}{2\hat{c}}(B^a - \frac{\hat{d}}{\hat{b}}A^a).$$

More simply, because we have its explicit decomposed expression, we could estimate $V_p(s^2)$ using the double APB as:

$$(17)\quad v_{pd}(s^2) = \frac{\left[\frac{f_0}{2}\cdot A^a + \frac{f_1-3f_0}{2}\cdot B^a\right]^2}{\frac{1}{R}\sum_{r=1}^{R}(\frac{f_0}{2}\cdot A^{a*} + \frac{f_1-3f_0}{2}\cdot B^{a*})_r}.$$

This APB approach is unnecessary with SRS, but it can be generalized with more effort to the case of pps sampling, where it is necessary to consider a decomposition of the variance into three parts, as indicated in (9).

5.4 Full resampling approach

In the full resampling approach, we would consider a bias-corrected estimator Cs_1^2. From each artificial population, we would resample, and compute the empirical variance $V\hat{a}r_{art}(C_as_1^{*2})$of $C_as_1^{*2}$; this will estimate $Var_{art}(C_as_1^{*2})$, and it will also estimate $Var_p(Cs_1^2)$ with some multiplicative bias as expressed by

$$(18)\quad E_p[E_{art}(V\hat{a}r_{art}(C_as_1^{*2})|\mathbf{s})] = Var_p(Cs_1^2)/D.$$

To correct the bias of the empirical variance of $C_{art}s_1^{*2}$, we might try to estimate D by a double APB procedure, assuming that

$$(19)\quad E_{p^*}[E_{art^*}(V\hat{a}r_{art^*}(C_{a^*}s_1^{**2})|\mathbf{s}^*)] \simeq Var_{art}(C_as_1^{*2})/D,$$

where E_{p^*} is the expectation due to the sampling design of resamples selected from the artificial population, U^a, and E_{art^*} denotes expectation over the generation of artificial populations, U^{a*}, from the resample $\mathbf{s}^*$. We denote the resulting "naive" estimator as $v_{pn}(s^2)$.

6 Simulation studies

In the following simulation studies, we examine the performance of the APB procedures in the estimation of variances. Values used for examining the performance of point estimators are defined in the corresponding tables. We use the coverage probability to examine the interval estimators of variance.

6.1 Estimation of variance of Horvitz-Thompson estimator

We consider a finite population of size N where units y_i for $i = 1 \cdots , N$ are generated from a $N(3, 1)$ distribution. We examined the performance of estimation of the variance of Horvitz-Thompson estimators computed from samples selected employing SRS and a probability proportional to size (pps) sampling design.

We set the population size to $N = 500$ and various sample sizes. For the pps sampling design, we selected our samples with the Rao-Sampford method, which provides calculable secondary order inclusion probabilities. The size variable was also a N(3,1) random variable. Using (2), we computed $v(\hat{Y}_{HT})$ from the sample, the average over artificial populations of $Var_{p^a}(\hat{Y}^*_{HT})$ defined in (3), and $v^*(\hat{Y}_{HT})$, the average over the artificial populations of the empirical variance of $\hat{Y}^*_{HT}$ from resamples of size $m = n-1$ using the APB. We simulated $I = 1000$ samples from the population and for each sample selected with SRS, we generated $R = 50, 100, 500$ resamples within the artificial population, whereas for the pps design, we created $Q = 25$ and $Q = 100$ artificial populations with $R = 100$ resamples within each of the artificial populations. For each sample, we computed the relative biases with respect to the true variance $Var(\hat{Y}_{HT})$ defined in (1).

Results of the simulations appear in Tables 1 and 2. Table 1 suggests that under the SRS design the empirical variance from the APB procedure works as well as the unbiased estimator of the variance of the Horvitz-Thompson estimator for various sample sizes. The resampling procedure does not require a very large number of resamples. As expected, $Var_{p^a}(\hat{Y}^*_{HT})$, calculated assuming a resample size of n, is biased for small sample sizes, but it improves greatly assuming a resample size of $m = n-1$. Table 2 for pps sampling indicates that the empirical variance $v^*(\hat{Y}_{HT})$ performs as well as the unbiased estimates of variance $\nu(\hat{Y}_{HT})$ for various sample sizes. Increases in the number of artificial populations did not reduce the bias appreciably.

6.2 APB for estimating confidence intervals for variance

In this study, a finite population of size N was generated from the following model, $Y = \beta + \varepsilon$ where $\varepsilon \sim N(0, 1)$. With $N = 1000$, the finite population parameter S^2 is 0.977.

We considered two sampling schemes: an SRS scheme and a pps sampling scheme. The size variable, V_i, was an exponential function of a normally distributed random variable whose mean and variance were zero and 1, respectively, and which was truncated to be in the range -1.5 to 1.5. In a range of sample sizes, $n = 10, 20, 40$ and 100 population units were sampled with probability proportional to the size variable, V_i, and hence the inclusion probability π_i was $nV_i / \sum_i V_i$ for $i = 1, \cdots , N$. We denote by I the number of samples, by Q the number of artificial populations, and by R the number of resamples (of size $n-1$) from each artificial population.

The results are reported in Tables 3 and 4, where rBias(up) and rBias(low) denote relative biases for the estimators of the 0.975 and 0.025 percentiles, respectively. Coverage refers to the coverage percentages of the estimated confidence interval. Note that the relative bias of the sample

Table 1: Estimation of Variance of HT Estimates for SRS Samples

n	values	$v^*(\hat{Y}_{HT})$	$v^*(\hat{Y}_{HT})$	$v^*(\hat{Y}_{HT})$	$v(\hat{Y}_{HT})$	$Var_{p^a}(\hat{Y}_{HT})$	$Var_{p^a}(\hat{Y}_{HT})$
		$R = 50$	$R = 100$	$R = 500$		$m = n - 1$	$m = n$
5	$rbias$	0.0159	0.0180	0.0184	0.0107	0.0147	-0.1861
	rSD	0.7211	0.7138	0.7098	0.6968	0.6995	0.5693
	$rMCE$	0.0161	0.0160	0.0159	0.0156	0.0156	0.0135
10	$rbias$	0.0141	0.0104	0.0096	0.006	0.0101	-0.1138
	rSD	0.5121	0.4632	0.4595	0.452	0.453	0.4171
	$rMCE$	0.0115	0.0104	0.0103	0.0101	0.0101	0.0095
25	$rbias$	0.0116	0.0114	0.0113	0.0084	0.0126	-0.0333
	rSD	0.3023	0.2959	0.2908	0.2784	0.2795	0.2683
	$rMCE$	0.0068	0.0066	0.0065	0.0062	0.0062	0.0061
50	$rbias$	0.0095	0.0080	0.0077	0.0033	0.0075	-0.0148
	rSD	0.2128	0.2030	0.1976	0.1857	0.1865	0.1851
	$rMCE$	0.0048	0.0045	0.0044	0.0042	0.0042	0.0041
100	$rbias$	0.0029	0.0018	0.0012	0.0025	0.0035	-0.0086
	rSD	0.1896	0.1624	0.1500	0.1266	0.1272	0.1256
	$rMCE$	0.0042	0.0036	0.0034	0.0028	0.0028	0.0028

rbias=$\sum(\hat{\theta}_r/\theta - 1)/I$, rSD=$\sqrt{\sum(\hat{\theta}_r - \bar{\hat{\theta}})^2/(I-1)/\theta}$, $rMCE = \sqrt{\sum(\hat{\theta}_r - \bar{\hat{\theta}})^2/I(I-1)/\theta}$

estimator $\hat{\xi}_\beta$ is calculated according to $(\sum_{j=1}^{I}(\hat{\xi}_{\beta j}/\xi_\beta) - 1)/I$ where $\hat{\xi}_{\beta j}$ is the estimate of the β^{th} percentile from the j^{th} simulated sample for $j = 1, \cdots, I$ and ξ_β is the finite population parameter. In this case, since we do not know the true value for ξ_β, the β^{th} percentile of the distribution of s^2/S^2, we use its empirical value. That is, for each sample, we compute s^2/S^2 and of the $I = 1000$ simulated values, we find the β^{th} percentile, and use it as our true value of the percentile. Table 3 suggests that for the SRS case, relative biases improve with increasing sample size, but not with an increase in the number of resamples. Coverage probability improves to some extent. The relative bias of the estimated upper percentile indicates that it underestimates the true percentile, which is consistent with the finding of undercoverage by the confidence intervals. Table 4 suggests for the pps case, relative biases and coverage probability improve with increasing sample size, and with increasing number of resamples, R, but are insensitive to the number of artificial populations, Q. A double APB resampling procedure might be implemented to reduce the bias of the upper and lower limits such as to improve the coverage probability.

6.3 Estimation of variance of sample variance estimator

In this study, a finite population of size N was generated from the following model, $Y = \beta + \varepsilon$ where $\varepsilon \sim N(0, 1)$. With $N = 1000$, the finite population parameter S^2 is 1.009. We consider two sampling schemes here. One is a SRS scheme and the other is a pps sampling scheme.

For the SRS scheme, we examined the performance of five different estimators of the variance of sample variance. Those estimates are $v_{pu}(s^2)$ from (11) , $v_{pd}(s^2)$ from (17), $v_{pr}(s^2)$ from (16), $v_{pn}(s^2)$ from Section 5.4, an inflation estimator $v_{p^*}(s^2) = f_0 A^a/2 + (f_1 - 3f_0)B^a/2$, an adjusted

Table 2: Estimation of Variance of HT Estimates for PPS Samples

n	values	$v^*(\hat{Y}_{HT})$	$v^*(\hat{Y}_{HT})$	$\bar{V}ar_{p^a}(\hat{Y}^*_{HT})$	$\bar{V}ar_{p^a}(\hat{Y}^*_{HT})$	$\nu(\hat{Y}_{HT})$
		$Q = 25$	$Q = 100$	$Q = 25$	$Q = 100$	
5	$rbias$	0.0235	0.0243	0.0299	0.0299	0.0202
	rSD	0.8489	0.8467	0.8498	0.8499	0.8434
	$rMCE$	0.0269	0.0269	0.0270	0.0270	0.0268
10	$rbias$	-0.0120	-0.0149	0.0152	0.0152	0.0130
	rSD	0.6420	0.6411	0.6504	0.6505	0.6426
	$rMCE$	0.0177	0.0177	0.0179	0.0179	0.0176
25	$rbias$	-0.0109	-0.0101	0.0350	0.0350	-0.0108
	rSD	0.3928	0.3928	0.4073	0.4075	0.3951
	$rMCE$	0.0176	0.0176	0.0182	0.0182	0.0177

rbias=$\sum(\hat{\theta}_r/\theta - 1)/I$, rSD=$\sqrt{\sum(\hat{\theta}_r - \bar{\hat{\theta}})^2/(I-1)/\theta}$, $rMCE = \sqrt{\sum(\hat{\theta}_r - \bar{\hat{\theta}})^2/I(I-1)/\theta}$

Table 3: Performance of Interval and Quantile Estimates for SRS Samples

n	Coverage	rBias(up)	rBias(low)	Coverage	rBias(up)	rBias(low)
		$R = 100$			$R = 200$	
10	90.10	-0.1214	-0.0239	90.65	-0.1118	-0.0535
20	90.25	-0.0728	0.0219	90.70	-0.0409	0.0840
40	92.25	-0.0542	0.0232	91.20	-0.0248	0.0357
100	91.70	-0.0137	0.0152	91.85	-0.0154	0.0229
200	91.95	-0.0036	0.0165	92.95	-0.0073	0.0133

rbias=$\sum(\hat{\theta}_r/\theta - 1)/I$

estimator $v^c_{p^*} = f_0 A^c_{\mathbf{s}}/2 + (f_1 - 3f_0)B^c_{\mathbf{s}}/2$, where $A^c_{\mathbf{s}}$ and $B^c_{\mathbf{s}}$ are bias-corrected versions with a single APB of $A_{\mathbf{s}}$ and $B_{\mathbf{s}}$ respectively, and the linearization estimator $\hat{v}_{plin}(s^2)$ as described in Section 5.1. Using $Var_p(s^2)$ defined in (10) as the true population parameter, we calculated the relative bias of each estimate, and the results are reported in Table 5. When the sample size is small ($n = 10$), we found that the empirical variance ($v_{p^*}(s^2)$) based on the artificial population was biased by -29%. Using the double APB estimate, $v_{pd}(s^2)$, we can reduce the bias of $v_{p^*}(s^2)$ to -5%. The regression-based estimate, $v_{pr}(s^2)$, is exactly unbiased in this SRS case. The estimated variances ($v_{pn}(s^2)$) obtained directly from the double resampling procedure are found to overestimate the true values and work well only when the sample size is large. This suggests that when the bias structure is not "simple", using the APB on the decomposition of the variance expression will provide more accurate estimates. The linearization estimator $v_{plin}(s^2)$ has larger relative biases for all sample sizes. The same design as in section 6.2 was used to generate the pps sample. For $Q = 25$ and $R = 25$, we computed the estimated variance of $s^2_1 = s^2_\pi$ using the naive double resampling procedures described in section 5.4. Results (not shown here) are not very promising. A decomposition approach will be investigated along with the APB procedures.

Table 4: Performance of Interval and Quantile Estimates for PPS Samples

	$n = 20$			$n = 40$			$n = 100$		
(Q, R)	Coverage	RBias (up)	RBias (low)	Coverage	RBias (up)	RBias (low)	Coverage	RBias (up)	RBias (low)
(50,50)	79.5	-0.225	0.208	83.6	-0.169	0.122	86.5	-0.076	0.045
(100,50)	79.6	-0.225	0.208	83.4	-0.169	0.121	86.9	-0.076	0.045
(50,100)	80.5	-0.219	0.187	84.0	-0.165	0.109	87.4	-0.072	0.039
(100,100)	80.8	-0.219	0.187	84.2	-0.165	0.109	87.3	-0.072	0.039

rbias=$\sum(\hat{\theta}_r/\theta - 1)/I$

Table 5: Performance of Estimation of Variance of Sample Variance for SRS Samples

n	values	$v_{pu}(s^2)$	$v_{p^*}(s^2)$	$v^c_{p^*}$	$v_{pd}(s^2)$	$v_{pr}(s^2)$	$v_{plin}(s^2)$	$v_{pn}(s^2)$
	rbias	-0.0330	-0.2856	-0.0295	-0.0488	0.0162	-0.3093	0.1939
10	*SD*	0.3310	0.2211	0.3339	0.3217	0.3523	0.2288	0.2959
	MCE	0.0134	0.0120	0.0134	0.0133	0.0136	0.0118	0.0094
	rbias	-0.0261	-0.1659	-0.0227	-0.0274	-0.0167	-0.1675	0.0425
20	*SD*	0.1010	0.0827	0.1026	0.1020	0.1022	0.0856	0.0825
	MCE	0.0099	0.0093	0.0100	0.0099	0.0100	0.0093	0.0076
	rbias	0.0152	-0.0596	0.0173	0.0162	0.0173	-0.0621	0.0360
40	*SD*	0.0354	0.0322	0.0358	0.0357	0.0355	0.0337	0.0294
	MCE	0.0072	0.0070	0.0072	0.0072	0.0072	0.0070	0.0060
	rbias	-0.0026	-0.0308	-0.0008	-0.0007	-0.0023	-0.0265	0.0424
100	*SD*	0.0080	0.0077	0.0081	0.0081	0.0080	0.0084	0.0078
	MCE	0.0044	0.0044	0.0044	0.0044	0.0044	0.0044	0.0020

rbias=$\sum(\hat{\theta}_r/\theta - 1)/I$, SD=$\sqrt{\sum(\hat{\theta}_r - \bar{\hat{\theta}})^2/(I-1)/\theta}$, $MCE = \sqrt{\sum(\hat{\theta}_r - \bar{\hat{\theta}})^2/I(I-1)/\theta}$

7 Conclusions

We used APB procedures to estimate and make inferences about population variances. Because analytic expressions for variance are often more complex to calculate than the design is to implement, using an APB algorithm to estimate variances has some practical value. Simulation studies show that APB performs well for an SRS design. It is speculated that the APB procedure along with decomposition of the theoretical variance will outperform other resampling-based estimates of variance of the sample variance, since the results for SRS are exact. Further work is needed to apply decomposition-assisted APB procedures to the pps sampling design because the “naive” double APB procedure does not perform well with small sample sizes. A double resampling procedure is suggested for interval estimation for complex designs.

Acknowledgement

This research was funded by grants from the Natural Sciences and Engineering Research Council (NSERC) of Canada.

References

Cho, E. and M. Cho (2008). Variance of sample variance. In Proceedings of the Survey Research Methods Section, *pp. 1291–1293. JMS.*

Kovacević, M., R. Huang, and Y. You (2006). Bootstrapping for variance estimation in multi-level models fitted to survey data. In 2006 Joint Statistical Meetings-Section on Survey Research Methods*.*

Pfeffermann, D., C. J. Skinner, D. J. Holmes, H. Goldstein, and J. Rasbash (1998). Weighting for unequal selection probabilities in multilevel models. Journal of Royal Statistical Society, Series B 60*, 23–40.*

Rabe-Hesketh, S. and A. Skrondal (2006). Multilevel modelling of complex survey data. Journal of Royal Statistical Society, Series A 169*, 805–827.*

Thompson, M. E. (1997). Theory of Sample Surveys *(first ed.). Chapman and Hall.*

Wang, Z. and M. E. Thompson (2010). A resampling approach to estimate variance components of multilevel models. Working Paper*.*

Wang, Z. and M. E. Thompson (2011). Empirical inference on variance components. Working Paper*.*