Variance Estimation in Complex Surveys with One PSU per Stratum

Harold Mantel and Suzelle Giroux
Statistics Canada, 16th floor, R.H. Coats Building, Tunney's Pasture,
Ottawa, Ontario, Canada K1A 0T6

Abstract
In this paper we discuss variance estimation for surveys where the number of first-stage sample units in one or more strata is one. Unbiased variance estimation in such cases is not possible. We review relevant ideas that have been discussed in the literature, particularly collapsing of strata. We then propose a new approach based on components of variance from different stages of sampling. The motivation and illustrations come from Statistics Canada's Canadian Health Measures Survey, in which the number of PSUs per stratum is very small (in fact, just one for the Atlantic region) due to operational and financial constraints.

Key Words: surveys, variance estimation, one PSU per stratum, stratum collapsing

1. Introduction

In this paper we consider the problem of variance estimation for surveys where the number of first-stage sample units in one or more strata is one. When only one PSU is selected in some strata, design unbiased estimation of the variance is not possible, since we have no information about differences between the PSUs in such strata.

This paper is motivated by the new Canadian Health Measures Survey (CHMS) which samples only a single PSU in one of its five strata. Estimates from CHMS are required at the national level. We also want to produce preliminary estimates based on the first half of the data collection, for which the problem appears in three strata.

Variance estimation for CHMS is also complicated by the fact that exact second-order inclusion probabilities for PSUs are not easily calculated. Some approximation to these probabilities is needed in order to estimate the variance.

CHMS has a 3-stage design. The first stage is an area frame with 257 collection sites (PSUs) stratified into 5 geographic regions. A total of 15 PSUs are selected using PPS systematic sampling from ordered lists within strata (size measure is census population), which makes unbiased estimation of the first-stage variance impossible in principle. However, we will assume that randomized PPS systematic (RPPSS) or some similar sampling design was used. This assumption seems not unreasonable, but does introduce an unknown bias into the variance estimation.

At the second stage, a list frame of dwellings is constructed for each site and stratified into 5 strata based on presence of individuals in different age groups from 2006 census data, and a sixth stratum for dwellings with no individuals in the target group at the time of the census. The sample size is around 600 dwellings per site using SRS within strata.

At the third stage individuals are selected within households. If children aged 6 to 11 are present then one of them is selected randomly. In addition, whether or not there are children aged 6 to 11 one person aged 12 to 79 is selected using probabilities that depend on age group: 12 to 19, 20 to 39, 40 to 59, and 60 to 79. The age-group specific probabilities are designed to achieve equal respondent sample sizes in each age-sex group – see Dion and Giroux (2009) for more details. The target sample size of respondent individuals is 350 to 375 per PSU.

A significant challenge for variance estimation is the small number of PSUs selected. In fact, for one of the strata (Atlantic region) only one PSU is selected. Thus unbiased estimation of the first-stage design variance is *a priori* impossible, even under the RPPSS assumption. Preliminary estimates are also required based on the first half of the data collected. For operational reasons all of the data collection in a PSU is completed before moving on to the next PSU, so that having only half the data means having only half of the PSUS. The first-stage strata and sample sizes are given in Table 1.

Table 1: Strata and Sample PSUs for CHMS

Region (Stratum)	No. PSUs selected	Preliminary PSUs
Atlantic	1	1
Quebec	4	2
Ontario	6	3
Prairies	2	1
British Columbia	2	1

The aim of this paper is to present different variance estimation methods for the CHMS, and to compare the resultant estimates. It is motivated by the question of whether standard replication methods, such as the Rao-Wu bootstrap for complex surveys, give acceptable estimates of variance for this survey, in view of the small number of sample PSUs. The answer seems to be a cautious “yes”.

In the remainder of this paper we first introduce useful notation in Section 2. In Section 3 we develop variance estimation assuming at least two PSUs per stratum are selected. In this section we also consider the problem of approximating the second-order inclusion probabilities under the PPS design. Then in Section 4 we discuss how the variance estimation can be adapted to the problematic case of strata with only one PSU selected. Finally, in Section 5 we compare some alternative variance estimators empirically using data from the preliminary sample.

2. Notation

In this section we define some notation and quantities of interest that will be used in the rest of the paper. All of the variables, totals, means and estimates are in terms of person level variates. Y is the population total of a person level variate y.

Regions – index r (*i.e.* first-stage stratum).

$$Y = \sum_r Y_r \qquad \hat{Y} = \sum_r \hat{Y}_r$$

PSUs (sites) within regions – index p, s_r is the sample of m_r PSUs selected from the M_r PSUs within region r.

w_p is the weight for PSU p in the sample = the inverse probability of selection of PSU p $= 1/\pi_p$.

$$Y_r = \sum_{p \in r} Y_p \qquad \hat{Y}_r = \sum_{p \in s_r} w_p \hat{Y}_p$$

Strata within PSUs – index u,

$$Y_p = \sum_{u \in p} Y_u \qquad \hat{Y}_p = \sum_{u \in p} \hat{Y}_u$$

Households – index h, s_u is the sample of m_u households h selected from the M_u households in stratum u.

w_h is weight for household h conditional on the PSU containing h being in the sample

$w_h = M_{u(h)} / m_{u(h)}$ where $u(h)$ is the stratum containing h, $m_{u(h)}$ is the number of households selected in stratum $u(h)$, and $M_{u(h)}$ is the total number of households in stratum $u(h)$. Note that $m_{u(h)}$ may be adjusted for household level non-response.

$$Y_u = \sum_{h \in u} Y_h \qquad \hat{Y}_u = \sum_{h \in s_u} w_h \hat{Y}_h$$

$$\bar{Y}_u = \sum_{h \in u} Y_h / M_u \qquad \bar{\hat{Y}}_u = \sum_{h \in s_u} \hat{Y}_h / m_u$$

Individuals – index j, s_h is the sample of individuals j selected from household h.

ω_j is weight for individual j (may be adjusted for non-response and calibrated, but we ignore that). Y_h is the total of variate y for household h

Define w_j by $\omega_j = w_p w_h w_j$, so w_j is something like the inverse probability weight of individual j given household h is in the sample, but may be distorted by calibration or other adjustments to the weights. In subsequent developments we will implicitly assume that $\hat{Y}_h$ remains unbiased for Y_h despite the effects of these adjustments.

$$Y_h = \sum_{j \in h} y_j \qquad \hat{Y}_h = \sum_{j \in s_h} w_j y_j$$

Putting all of this together we can write

$$Y = \sum_r \sum_{p \in r} \sum_{u \in p} \sum_{h \in u} \sum_{j \in h} y_j \text{, and}$$

$$\hat{Y} = \sum_r \sum_{p \in s_r} w_p \sum_{u \in p} \sum_{h \in s_u} w_h \sum_{j \in s_h} w_j y_j = \sum_r \sum_{p \in s_r} \sum_{u \in p} \sum_{h \in s_u} \sum_{j \in s_h} \omega_j y_j .$$

3. Variance Estimation

Now we can write

$$\hat{Y} = \sum_r \sum_{p \in s_r} w_p \sum_u \hat{Y}_u = \sum_r \sum_{p \in s_r} w_p \hat{Y}_p .$$

It is useful to break down the variance as the variance due to the first-stage sampling (first-stage variance) and the variance conditional on the first-stage sample (variance due to second and subsequent stages of sampling). Estimation procedures can then be developed for these two terms separately.

$$Var\left(\hat{Y}\right) = Var\left\{E\left(\hat{Y}\,|s_r\right)\right\} + E\left\{Var\left(\hat{Y}\,|s_r\right)\right\}$$

$$= Var\left\{\sum_r \sum_{p \in s_r} w_p Y_p\right\} + E\left\{\sum_r \sum_{p \in s_r} w_p^2 \sum_u Var\left(\hat{Y}_u\right)\right\} \qquad (1a)$$

$$= Var\left\{\sum_r \sum_{p \in s_r} w_p Y_p\right\} + E\left\{\sum_r \sum_{p \in s_r} w_p^2 Var(\hat{Y}_p)\right\} \qquad (1b)$$

It is also possible to break the variance down into components representing each stage of sampling by writing

$$Var\left(\hat{Y}_u\right) = Var\left\{E\left(\hat{Y}_u\,|s_u\right)\right\} + E\left\{Var\left(\hat{Y}_u\,|s_u\right)\right\}$$

and substituting this expression into (1a) above. However, this further breakdown is not very useful. In particular, the third stage variance, $Var\left(\hat{Y}_u\,|s_u\right)$, cannot be directly estimated since at the third stage of sampling only one unit is selected per stratum (the strata for the third stage of sampling consist of the sets of 6 to11 year-olds and the 12 to 79 year-olds, and one individual is selected from each non-empty stratum within a household). As we will see later, this further breakdown of the variance is also not necessary.

3.1. First-Stage Variance

The first term in (1) is problematical because only a small number of PSUs is selected, only one for Atlantic region, only two in each of BC and prairie regions. However, ignoring that problem for the moment we can write the first stage variance of $\hat{Y}$ as

$$Var\left\{\sum_r \sum_{p\in s_r} w_p Y_p\right\} = \sum_r \sum_{p\in r} \sum_{q\in r} (\pi_{pq} - \pi_p \pi_q) \frac{Y_p}{\pi_p} \frac{Y_q}{\pi_q} \quad (2a)$$

$$= -\frac{1}{2} \sum_r \sum_{p\in r} \sum_{q\in r} (\pi_{pq} - \pi_p \pi_q) \left(\frac{Y_p}{\pi_p} - \frac{Y_q}{\pi_q} \right)^2, \quad (2b)$$

where π_p is the probability that site (PSU) p is included in sample s_r, and π_{pq} is the probability that both p and q are in s_r. Form (2a) is the general form of the variance under sampling without replacement; form (2b) is the Sen-Yates-Grundy form and applies only if the sample size in each stratum is fixed.

Based on (2a) the Horvitz-Thompson unbiased estimator of variance is

$$\hat{Var}_{HT}\left\{\sum_r \sum_{p\in s_r} w_p Y_p\right\} = \sum_r \sum_{p\in s_r} \sum_{q\in s_r} \frac{\pi_{pq} - \pi_p \pi_q}{\pi_{pq}} \frac{Y_p}{\pi_p} \frac{Y_q}{\pi_q},$$

and from (2b) the Sen-Yates-Grundy variance estimator is given by

$$\hat{Var}_{SYG}\left\{\sum_r \sum_{p\in s_r} w_p Y_p\right\} = -\frac{1}{2} \sum_r \sum_{p\in s_r} \sum_{q\in s_r} \frac{\pi_{pq} - \pi_p \pi_q}{\pi_{pq}} \left(\frac{Y_p}{\pi_p} - \frac{Y_q}{\pi_q} \right)^2.$$

Neither of these two estimators is available since Y_p is not observed, only $\hat{Y}_p$. We may define

$$\hat{Var}_{HT}\left\{\sum_r \sum_{p\in s_r} w_p \hat{Y}_p\right\} = \sum_r \sum_{p\in s_r} \sum_{q\in s_r} \frac{\pi_{pq} - \pi_p \pi_q}{\pi_{pq}} \frac{\hat{Y}_p}{\pi_p} \frac{\hat{Y}_q}{\pi_q} \quad (3)$$

and show that

$$E\left[\hat{Var}_{HT}\left\{\sum_r \sum_{p\in s_r} w_p \hat{Y}_p\right\} \middle| s_r\right]$$
$$= \hat{Var}_{HT}\left\{\sum_r \sum_{p\in s_r} w_p Y_p\right\} + \sum_r \sum_{p\in s_r} \frac{1-\pi_p}{\pi_p^2} Var(\hat{Y}_p).$$

For the Sen-Yates-Grundy form we may define

$$\hat{Var}_{SYG}\left\{\sum_r \sum_{p\in s_r} w_p \hat{Y}_p\right\} = -\frac{1}{2} \sum_r \sum_{p\in s_r} \sum_{q\in s_r} \frac{\pi_{pq} - \pi_p \pi_q}{\pi_{pq}} \left(\frac{\hat{Y}_p}{\pi_p} - \frac{\hat{Y}_q}{\pi_q} \right)^2 \quad (4)$$

and it can be shown that

$$E\left[\hat{V}ar_{SYG}\left\{\sum_r \sum_{p\in s_r} w_p \hat{Y}_p\right\}\middle| s_r\right]$$

$$= \hat{V}ar_{SYG}\left\{\sum_r \sum_{p\in s_r} w_p Y_p\right\} - \sum_r \sum_{p\in s_r} \frac{Var(\hat{Y}_p)}{\pi_p^2} \sum_{q\in s_r, q\neq p} \frac{\pi_{pq} - \pi_p \pi_q}{\pi_{pq}}.$$

In either case, whether we use the Horvitz-Thompson estimator (3) or the Sen-Yates-Grundy estimator (4), we have an estimator whose expectation, conditional on the first stage sample s_r, is equal to an unbiased estimator of the first term of (1) plus a bias term involving variances from the later stages of sampling. We could estimate the variances in the bias term in order to adjust for it.

3.1.1. Calculation of π_{pq} **s**

Estimation of the first-stage variance using either equation (3) or (4) requires the first- and second-order inclusion probabilities for the PSUs in the sample. The first-order inclusion probabilities, π_p, are simply equal to $m_r N_p / N_r$, where m_r is the number of PSUs selected in stratum *r*, and N_r and N_p are, respectively, the census populations of region *r* and PSU *p*. This is true since the PSUs are selected using PPS within regions with size measure N_p. Exact calculation of the second-order inclusion probabilities, π_{pq}, is practically impossible under the RPPSS approximation to the first-stage design, and various approximations have been suggested in the literature. One possible approach is to simply substitute approximations for π_{pq} directly into (3) or (4). A different approach taken in the literature is to use approximations to π_{pq} to derive approximations to the actual variance in (2a) or (2b), and then to derive estimates of these approximations.

Hartley and Rao (Annals, 1962) derive asymptotic approximations to the π_{pq} for single-stage RPPSS sampling using Edgeworth series. They then derive expressions for the variance and for variance estimates that are accurate to order $O(N^0)$, and simplified expressions accurate to $O(N^1)$. To order $O(N^0)$ their estimator takes the form:

$$\hat{V}ar_{HR,0}\left\{\sum_r \sum_{p\in s_r} w_p Y_p\right\} =$$

$$\frac{1}{2}\sum_r \frac{1}{m_r - 1} \sum_{p\in s_r} \sum_{q\in s_r} \left[1 - (\pi_p + \pi_q) + \frac{\sum_{d\in r} \pi_d^2}{m_r} + \mathrm{K}_r\right]\left(\frac{Y_p}{\pi_p} - \frac{Y_q}{\pi_q}\right)^2$$

where

$$\mathrm{K}_r = -\frac{1}{m_r}(\pi_p^2 + \pi_q^2) - \frac{2}{m_r^3}\left(\sum_{d\in r} \pi_d^2\right)^2 + \frac{1}{m_r^2}(\pi_p + \pi_q)\sum_{d\in r} \pi_d^2 + \frac{2}{m_r^2}\sum_{d\in r} \pi_d^3$$

To order $O(N^1)$ they obtain the simplified estimator

$$\hat{V}ar_{HR,1}\left\{\sum_r \sum_{p\in s_r} w_p Y_p\right\} =$$

$$\frac{1}{2}\sum_r \frac{1}{m_r - 1}\sum_{p\in s_r}\sum_{q\in s_r}\left[1-(\pi_p+\pi_q)+\frac{\sum_{d\in r}\pi_d^2}{m_r}\right]\left(\frac{Y_p}{\pi_p}-\frac{Y_q}{\pi_q}\right)^2 \qquad (5)$$

where the K_r term is dropped.

Brewer and Donadio (2003) consider the problem from a different perspective; namely, they try to find useful approximations to the π_{pq} that depend only on the first-order inclusion probabilities, π_p. They consider high entropy designs, that is, designs for which knowledge that unit *p* is in the sample gives very little knowledge of what other units are in the sample - RPPSS is one example of a high-entropy design. For such high entropy designs they propose approximations of the form

$$\pi_{pq} \cong \tilde{\pi}_{pq} = \pi_p \pi_q (c_p + c_q)/2 \qquad (6)$$

and consider different possible values of c_p suggested by various identities that second-order inclusion probabilities must satisfy. One of the values of c_p that they consider, namely

$$c_p = \frac{m_{r(p)} - 1}{m_{r(p)} - 2\pi_p + m_{r(p)}^{-1}\sum_{q\in r}\pi_q^2}\ , \qquad (7)$$

is an approximation to asymptotic expressions derived by Hartley and Rao (1962) and by Asok and Sukhatme (JASA, 1976). Under SRSWOR this value of c_p, when substituted in (6), yields the exact value for π_{pq}. It is also interesting to note that if $m_{r(p)} = 1$ then $c_p = 0$. In fact, this is true for all of the values of c_p considered by Brewer and Donadio. Through some rather ingenious algebraic manipulations, and using the approximation (6) they derive the following approximation to the first-stage variance:

$$\tilde{V}ar_{BD}\left\{\sum_r \sum_{p\in s_r} w_p Y_p\right\} = \sum_r \sum_{p\in r} \pi_p (1 - c_p \pi_p)\left(\frac{Y_p}{\pi_p} - \frac{Y_r}{m_r}\right)^2 .$$

They then derive an approximate design-unbiased estimator for this expression:

$$\hat{\tilde{V}}ar_{BD}\left\{\sum_r \sum_{p\in s_r} w_p Y_p\right\} = \sum_r \sum_{p\in s_r} (c_p^{-1} - \pi_p)\left(\frac{Y_p}{\pi_p} - \frac{\tilde{Y}_r}{m_r}\right)^2 . \qquad (8)$$

where $\tilde{Y}_r = \sum_{p\in s_r} Y_p/\pi_p$. Finally, considering model properties of the estimator (8) under a ratio-type regression model in which the expected value of Y_p is proportional π_p, they propose a slightly modified version of c_p, namely

$$c_p = \frac{m_{r(p)} - 1}{m_{r(p)} - (2m_{r(p)} - 1)(m_{r(p)} - 1)^{-1}\pi_p + (m_{r(p)} - 1)^{-1}\sum_{q \in r}\pi_q^2} \qquad (9)$$

With this value of c_p the model bias of the variance estimator (8) is of order $O(N n^{-2})$, while with other proposed values of c_p it is of order $O(N n^{-1})$.

If either (5) or (8) is used to estimate the first-stage variance, then the quantities Y_p and $\widetilde{Y}_r$ would need to be estimated based on subsequent stages of sampling. Replacing Y_p by $\hat{Y}_p$ and $\widetilde{Y}_r$ by $\hat{Y}_r$ we get, from (5), the Hartley-Rao variance estimator as

$$\hat{Var}_{HR,1}\left\{\sum_r \sum_{p \in s_r} w_p \hat{Y}_p\right\} =$$

$$\frac{1}{2}\sum_r \frac{1}{m_r - 1}\sum_{p \in s_r}\sum_{q \in s_r}\left(1 - (\pi_p + \pi_q) + \frac{\sum_{d \in r}\pi_d^2}{m_r}\right)\left(\frac{\hat{Y}_p}{\pi_p} - \frac{\hat{Y}_q}{\pi_q}\right)^2 \qquad (10)$$

and it can be shown that

$$E\left[\hat{Var}_{HR,1}\left\{\sum_r \sum_{p \in s_r} w_p \hat{Y}_p\right\} \middle| s_r\right] = \hat{Var}_{HR,1}\left\{\sum_r \sum_{p \in s_r} w_p Y_p\right\}$$

$$+ \sum_r \frac{1}{m_r - 1}\sum_{p \in s_r}\frac{Var(\hat{Y}_p)}{\pi_p^2}\left\{(m_r - 1)\left(1 - \pi_p + \frac{\sum_{d \in r}\pi_d^2}{m_r}\right) - \sum_{q \in s_r, q \neq p}\pi_q\right\}. \qquad (11)$$

Similarly from (8), defining the Brewer-Donadio estimator as

$$\hat{\widetilde{Var}}_{BD}\left\{\sum_r \sum_{p \in s_r} w_p \hat{Y}_p\right\} = \sum_r \sum_{p \in s_r}(c_p^{-1} - \pi_p)\left(\frac{\hat{Y}_p}{\pi_p} - \frac{\hat{Y}_r}{m_r}\right)^2 \qquad (12)$$

it can be shown that

$$E\left[\hat{\widetilde{Var}}_{BD}\left\{\sum_r \sum_{p \in s_r} w_p \hat{Y}_p\right\} \middle| s_r\right] = \hat{\widetilde{Var}}_{BD}\left\{\sum_r \sum_{p \in s_r} w_p Y_p\right\}$$

$$+ \sum_r \sum_{p \in s_r}\frac{Var(\hat{Y}_p)}{\pi_p^2}\left(\left(1 - \frac{1}{m_r}\right)^2(c_p^{-1} - \pi_p) + \frac{1}{m_r^2}\sum_{q \in s_r, q \neq p}(c_q^{-1} - \pi_q)\right) \qquad (13)$$

Again, in the case of the Hartley-Rao estimator (10) or the Brewer-Donadio estimator (12), we have an estimator whose expectation, conditional on the first stage sample s_r, is equal to an approximately unbiased estimator of the first-stage variance plus a bias term involving variances from the later stages of sampling. We could estimate the variances in the bias term in order to adjust for them.

It should also be noted that both the Hartley-Rao and the Brewer-Donadio estimators are based on the Sen-Yates-Grundy form of the variance, so they assume implicitly that the first stage sample sizes are fixed within strata.

3.2. Second and Third Stage Variance

The second term in (1) represents the variance due to second and subsequent stages of sampling. It can be estimated by $\sum_r \sum_{p\in s_r} w_p^2 \sum_u \hat{Var}\left(\hat{Y}_u\right)$ where $\hat{Var}\left(\hat{Y}_u\right)$ is any suitable estimator of $Var\left(\hat{Y}_u\right)$.

Note that $Var(\hat{Y}_p) = \sum_u Var(\hat{Y}_u)$ so that if we can estimate $Var\left(\hat{Y}_u\right)$, then we can also estimate and adjust for the bias of the first-stage variance estimators in (3), (4), (10) and (12).

Now $Var\left(\hat{Y}_u\right) = Var\left\{\sum_{h\in s_u} w_h \hat{Y}_h\right\}$

$$= Var\left\{\sum_{h\in s_u} w_h Y_h\right\} + E\left\{\sum_{h\in s_u} w_h^2 Var\left(\hat{Y}_h\right)\right\}$$

$$= \left(\frac{1}{m_u} - \frac{1}{M_u}\right)\frac{M_u^2}{M_u - 1}\sum_{h\in u}\left(Y_h - \overline{Y}_u\right)^2 + \frac{M_u}{m_u}\sum_{h\in u} Var\left(\hat{Y}_h\right)$$

It can be shown that

$$E\left[\frac{1}{m_u - 1}\sum_{h\in s_u}\left(\hat{Y}_h - \overline{\hat{Y}}_u\right)^2\right] = \frac{1}{M_u - 1}\sum_{h\in u}\left(Y_h - \overline{Y}_u\right)^2 + \frac{1}{M_u}\sum_{h\in u} Var(\hat{Y}_h)$$

so that

$$E\left[\left(\frac{1}{m_u} - \frac{1}{M_u}\right)\frac{M_u^2}{M_u - 1}\sum_{h\in s_u}\left(\hat{Y}_h - \overline{\hat{Y}}_u\right)^2\right]$$

$$= \left(\frac{1}{m_u} - \frac{1}{M_u}\right)\frac{M_u^2}{M_u - 1}\sum_{h\in u}\left(Y_h - \overline{Y}_u\right)^2 + \left(\frac{M_u}{m_u} - 1\right)\sum_{h\in u} Var(\hat{Y}_h)$$

$$= Var\left(\hat{Y}_u\right) - \sum_{h\in u} Var(\hat{Y}_h)$$

which is almost equal to $Var\left(\hat{Y}_u\right)$. If the second stage sampling fraction $\frac{m_u}{M_u}$ is small then the difference, $\sum_{h\in u} Var(\hat{Y}_h)$, is negligible. In CHMS the sampling fraction would be quite small, since the average PSU size must be around 50,000 dwellings, from which about 600 are selected at the second stage, so a typical second-stage sampling fraction would be around 1.2%. In any case, the difference is impossible to estimate since, as

noted above, for sampling within households *h* we select only one individual *j* per stratum. Let us therefore define

$$\hat{V}ar(\hat{Y}_u) = \left(\frac{1}{m_u} - \frac{1}{M_u}\right)\frac{M_u^2}{m_u - 1}\sum_{h\in s_u}\left(\hat{Y}_h - \bar{\hat{Y}}_u\right)^2 \tag{14}$$

and

$$\hat{V}ar(\hat{Y}_p) = \sum_u \hat{V}ar(\hat{Y}_u)\,. \tag{15}$$

3.3. Overall Variance Estimation

Assuming that the first-stage sample size in each stratum is at least 2, for overall variance estimation we may use any of the estimators (3), (4), (10) or (12) for the first-stage variance, and then use (14) and (15) to adjust for the bias in first-stage variance estimation and to estimate the variance from subsequent stages of sampling. However, it should be noted that the Horvitz-Thompson estimator (3) and the Sen-Yates-Grundy estimator (4) require the second-order inclusion probabilities π_{pq} which are generally unknown.

Based on the Horvitz-Thompson estimator (3) and its bias, the overall variance in (1) can be estimated by

$$\hat{V}ar_{HT}\left\{\sum_r\sum_{p\in s_r} w_p\hat{Y}_p\right\} - \sum_r\sum_{p\in s_r}\frac{1-\pi_p}{\pi_p^2}\hat{V}ar(\hat{Y}_p) + \sum_r\sum_{p\in s_r} w_p^2\hat{V}ar\left(\hat{Y}_p\right) \tag{16}$$

where $\hat{V}ar(\hat{Y}_p)$ is as defined in equations (14) and (15).

Based on the Sen-Yates-Grundy estimator (4) and its bias, the overall variance can be estimated by

$$\hat{V}ar_{SYG}\left\{\sum_r\sum_{p\in s_r} w_p\hat{Y}_p\right\} + \sum_r\sum_{p\in s_r}\frac{\hat{V}ar(\hat{Y}_p)}{\pi_p^2}\sum_{q\in s_r, q\neq p}\frac{\pi_{pq}-\pi_p\pi_q}{\pi_{pq}} + \sum_r\sum_{p\in s_r} w_p^2\hat{V}ar\left(\hat{Y}_p\right). \tag{17}$$

Based on the Hartley-Rao estimator (10) and its approximate bias in (11), the overall variance can be estimated by

$$\hat{V}ar_{HR}\left\{\sum_r\sum_{p\in s_r} w_p\hat{Y}_p\right\} + \sum_r\sum_{p\in s_r} w_p^2\hat{V}ar\left(\hat{Y}_p\right) - \sum_r\frac{1}{m_r-1}\sum_{p\in s_r}\frac{\hat{V}ar(\hat{Y}_p)}{\pi_p^2}\left\{(m_r-1)\left(1-\pi_p+\frac{\sum_{d\in r}\pi_d^2}{m_r}\right) - \sum_{q\in s_r, q\neq p}\pi_q\right\} \tag{18}$$

Finally, based on the Brewer-Donadio estimator (12) and its approximate bias in (13), the overall variance can be estimated by

$$\hat{\hat{V}}ar_{BD}\left\{\sum_r \sum_{p\in s_r} w_p \hat{Y}_p\right\} + \sum_r \sum_{p\in s_r} w_p^2 \hat{V}ar\left(\hat{Y}_p\right)$$

$$-\sum_r \sum_{p\in s_r} \frac{\hat{V}ar(\hat{Y}_p)}{\pi_p^2}\left\{\left(1-\frac{1}{m_r}\right)^2 (c_p^{-1}-\pi_p) + \frac{1}{m_r^2}\sum_{q\in s_r, q\neq p}(c_q^{-1}-\pi_q)\right\} \qquad (19)$$

where c_p is as given in equation (9). Because of the superior model-based properties under a reasonable ratio-type regression model, as noted in Section 2.1, this is the recommended estimator.

4. Variance Estimation When Some Strata Have Only One PSU Selected

4.1. Variance Formulas with Collapsed Strata

When one or more of the first stage strata have only one PSU selected, a popular approach to variance estimation is to collapse the problem strata with other strata and estimate the variance as if the sample had been selected from the combined strata. All of the collapsed strata will have at least two PSUs selected so that the estimators (3), (4), (10) or (12) can be calculated, provided that suitable values for the joint inclusion probabilities, π_{pq}, can be defined for the collapsed strata.

Since we are supposing that sampling is done using RPPSS within strata (regions), it would be good to define the π_{pq} as if RPPSS sampling had been done in the collapsed strata. We should also define the PPS size measure for the collapsed strata in such a way that the first order inclusion probabilities are preserved, since we want to estimate the variance of $\hat{Y} = \sum_r \sum_{p\in s_r} w_p \hat{Y}_p = \sum_r \sum_{p\in s_r} \hat{Y}_p / \pi_p$. The original selection was done using population as a size measure, but using this size measure in the collapsed strata will probably not preserve the first order inclusion probabilities, because of small differences in the average sizes of PSUs in different strata. However, if we take the original π_ps as the size measure for RPPSS sampling in the collapsed strata then the first order inclusion probabilities are preserved. The second order inclusion probabilities π_{pq} can then be calculated as if the samples in the collapsed strata had been selected using RPPSS with size measure proportional to π_p.

The most serious problem with collapsed stratum estimates of variance is that they are usually biased, since the mean of the variable of interest is usually different for different strata. In the context of multistage sampling that we have here, with collapsing of first stage strata for estimation of the first stage variance using either equation (3) or (4), it is the differences among the stratum means of $N_p Y_p = E\left\{\hat{Y}_p / \pi_p\right\}$ that lead to the bias. It has therefore been suggested that this bias be reduced by combining "similar" strata.

The approximate first variance estimators in (10) and (12) depend only on first-order inclusion probabilities, and do not involve the second-order inclusion probabilities. Nevertheless, these estimators cannot be calculated if one or more of the strata have $m_r = 1$, since the estimator in (10) has a factor $(m_r - 1)^{-1}$ while the estimator in (12) involves c_p^{-1} and $c_p = 0$ for strata with $m_r = 1$. Collapsing strata with $m_r = 1$ with other strata will fix this problem; however, it will introduce some bias as discussed above.

Another idea developed by Hartley, Rao and Keifer (JASA, 1969) is to reduce the bias of the collapsed stratum estimator of variance by replacing the common stratum mean in a standard variance formula by a regression predictor based on some concomitant variables. However, they develop this in the context of SRS within strata, where the variance estimator is expressed as a weighted sum of squared residuals (observations minus mean) in which the mean can be replaced by an alternative predictor. Extending this idea to PPS sampling requires that the variance estimator can be written in this form.

The Brewer-Donadio estimator in (12) is a weighted sum of squared residuals, $\left(\frac{Y_p}{\pi_p} - \frac{\hat{Y}_r}{m_r} \right)^2$. However, the value of c_p as given in either (7) or (9) is 0 for strata *r* in which only one PSU is selected (*i.e.*, $m_r = 1$), so the estimator in (12) could still not be computed. This difficultly could perhaps be dealt with by some ad hoc fix, but we will not pursue it further here.

4.2. Resampling With Collapsed Strata

The development in Section 4.1 is based on approximating exact expressions for variance under the collapsed stratum design, as developed in Section 3. Another general approach to variance estimation is based on resampling methods, with the Jackknife method or the Rao-Wu bootstrap (Rao, Wu and Yue, 1992) methods being quite popular when the sampling design is multi-stage stratified PPS. It is quite straighforward to apply these resampling methods after collapsing of strata. It is simply a matter of assuming that the sample PSUs were drawn from the collapsed strata rather than the original strata, and calculating the resulting resampling estimate of the variance.

4.3. Variance Components Modelling Approach

A third approach to variance estimation for strata with only one PSU selected is to model the first stage variance as a function of known characteristics of the design and other quantities that are known or that can be estimated from the sample. For example, for CHMS it may be reasonable to assume that the proportion of total variance due to the first stage of sampling is constant across strata, since the PSU sizes, the first-stage sampling rates and the within-PSU sampling designs are all similar across strata.

To be more specific, we can get a synthetic estimator for the total variance of strata in which only one PSU is selected by modeling the ratio of the total variance to the within-PSU variance as a constant across strata. This is equivalent to assuming that the ratio of the first-stage variance to the within-PSU variance is constant. In the empirical

comparisons in Section 5 we use the estimator (19) based on Brewer and Donadio (2003) for the total variance, but we could also have used (16), (17) or (18).

Let

$$\hat{A}_r = \hat{V}ar\left(\hat{Y}_r\right) \Big/ \hat{V}ar_{within-PSU}\left(\hat{Y}_r\right) \tag{20}$$

where $\hat{V}ar\left(\hat{Y}_r\right)$ is calculated from (19) but without summing over strata *r* and is calculated only for those strata *r* for which more than one PSU is selected, and $\hat{V}ar_{within-PSU}\left(\hat{Y}_r\right) = \sum_{p \in s_r} w_p^2 \sum_u \hat{V}ar\left(\hat{Y}_u\right)$ with $\hat{V}ar\left(\hat{Y}_u\right)$ as given in (14). $\hat{A}_r$ can be calculated for each region *r* for which more than one PSU is selected. We may then take $\hat{A}$ to be either the average of $\hat{A}_r$ or (more conservatively) as the maximum observed $\hat{A}_r$. For the empirical comparisons in Section 5 we use the maximum. The synthetic estimator of the total variance is then defined as

$$\hat{V}ar_{syn}\left(\hat{Y}_r\right) = \hat{A}\hat{V}ar_{within-PSU}\left(\hat{Y}_r\right). \tag{21}$$

5. Empirical Comparison of Variance Estimation Methods

Using the preliminary dataset, in which data from only about half of the sample PSUs is available, as described in Table 1 of Section 1, we calculated variance estimates for various characteristics using the methods described in the previous sections. Since the preliminary dataset has data from only one sample PSU in three of the five strata, some form of collapsing or modelling is required in order to produce variance estimates at the Canada level. For the collapsed-stratum methods, we collapsed the Atlantic stratum with Quebec and combined the Prairies and British Columbia strata.

This empirical test had two aims: (1) to compare estimates of variance given by the different methods, and (2) to investigate the appropriateness of the variance components modelling approach described in Section 4.3. All of the methods apart from this modelling approach are based on the collapsed strata.

Table 2 shows various estimates of variance, as described in the previous sections, for three CHMS variables –blood cadmium, blood lead and blood mercury. For the Horvitz-Thompson estimators and the Sen-Yates-Grundy estimators, the first-stage joint inclusion probabilities are replaced by approximations based on (6) with c_p taking either the value in (7) or (9).

The different variance estimation methods produce largely comparable results. The estimates based on the Horvitz-Thompson estimator, given in (16), tend to be somewhat larger; however, with a single sample and such a small number of PSUs it is not clear that this difference means anything. The fact that some quite different approaches all yield comparable results should increase our confidence in those results. The estimates from the resampling methods, bootstrap and jackknife, are in good agreement with the other estimates, confirming that these methods are as valid as any.

Table 2: Variance Estimates from Preliminary CHMS Data Using Different Methods (Numbers in parentheses refer to equations in the text)

Method\Variable	Total Blood Cadmium ($\times 10^{14}$)	Total Blood Lead ($\times 10^{10}$)	Total Blood Mercury ($\times 10^{15}$)
HT1 – (16), (6), (7)	9.06	5.37	1.54
HT2 – (16), (6), (9)	12.05	9.89	1.92
SYG1 – (17), (6), (7)	6.24	2.06	1.28
SYG2 – (17), (6), (9)	6.20	2.05	1.23
HR – (18)	6.24	2.07	1.28
BD – (19), (9)	6.20	2.12	1.27
Bootstrap	6.86	3.14	1.53
Jackknife	6.90	2.40	1.43
Variance Components	6.20	2.12	1.27

The main misgivings about the variance estimates in Table 2 arise from the small number of sample PSUs, and in particular from the three out of five design strata for which the preliminary sample had data from only one sample PSU. It will be important, once the full dataset is available, to redo these analyses to confirm that the different methods yield consistent estimates. This would strengthen the mutual validation of the different approaches and, in particular, confirm the validity of the bootstrap method for this survey.

The full dataset could also be used to calculate variances both with and without collapsing of the Prairies and British Columbia strata. This would allow some assessment of the effect of stratum collapsing, though it would still not be possible to calculate an uncollapsed variance estimate for the Atlantic.

References

Asok, C., and Sukhatme, B.V. (1976). On Sampford's Procedure of Unequal Probability Sampling Without Replacement. *Journal of the American Statistical Association*, 71, 912-918.

Brewer, K.R., and Donadio, M.E. (2003) The High Entropy Variance of the Horvitz-Thompson Estimator, *Survey Methodology*, 29, 189-196.

Dion, S.-M., and Giroux, S. (2009). Achieving the Unique Objectives of the Canadian Health Measures Survey. *Proceedings of the Survey Research Methods Section*, American Statistical Association.

Hartley, H.O., and Rao, J.N.K. (1962). Sampling With Unequal Probabilities and Without Replacement. *The Annals of Mathematical Statistics*, 33, 350-374.

Hartley, H.O., and Rao, J.N.K., and Kiefer, G. (1969). "Variance Estimation with One Unit per Stratum". *Journal of the American Statistical Association*, 64, 841-851.

Rao, J.N.K., Wu., C.J.F., and Yue, K. (1992). Some Recent Work on Resampling Methods for Complex Surveys. *Survey Methodology*, 18, 209-217.