Component of Statistics Canada
Catalogue no. 12-001-X Business Survey Methods Division

Article

Replication variance estimation under two-phase sampling

by Jae Kwang Kim and Cindy Long Yu

June 2011

Statistics Canada Statistique Canada

Canada

Survey Methodology, June 2011
Vol. 37, No. 1, pp. 67-74
Statistics Canada, Catalogue No. 12-001-X

Replication variance estimation under two-phase sampling

Jae Kwang Kim and Cindy Long Yu [1]

Abstract

In two-phase sampling for stratification, the second-phase sample is selected by a stratified sample based on the information observed in the first-phase sample. We develop a replication-based bias adjusted variance estimator that extends the method of Kim, Navarro and Fuller (2006). The proposed method is also applicable when the first-phase sampling rate is not negligible and when second-phase sample selection is unequal probability Poisson sampling within each stratum. The proposed method can be extended to variance estimation for two-phase regression estimators. Results from a limited simulation study are presented.

Key Words: Double sampling; Jackknife; Regression estimator; Reweighted expansion estimator.

1. Introduction

Two-phase sampling, first introduced by Neyman (1938) and sometimes called double sampling, is a cost effective technique in survey sampling. It is typically used when it is very expensive to collect data on the variables of interest, but it is relatively inexpensive to collect data on variables that are correlated with the variables of interest. Two-phase sampling has application in different forms (*e.g.*, Rao 1973; Cochran 1977; Breidt and Fuller 1993; Rao and Sitter 1995; Hidiroglou and Särndal 1998; Fuller 1998; Hidiroglou 2001; Fuller 2003). Two-phase sampling for stratification refers to the situation where the observation from the first-phase sample is used to make a stratification for the second-phase sampling. By selecting the first-phase sample for stratification purpose, two-phase sampling is a useful tool when there is no sampling frame available for stratification at the beginning. For example, in forest surveys, it is very difficult and expensive to travel to remote areas to make on-ground determinations. However, aerial photographs are relatively inexpensive, and determinations on, say, forest type from aerial photos are strongly correlated with ground determinations and can be used to stratify the first phase sample.

Replication variance estimation is very popular in complex surveys. Rust and Rao (1996) and Wolter (2007) provide comprehensive overviews on this topic. The replication method does not require the computation of the partial derivative of the Taylor expansion and the user can easily produce variance estimates without knowing the sampling design that was used to collect the data. Furthermore, this tendency is increasing because of confidentiality issues (Lu and Sitter 2006). Once the replication weights are provided, the design information such as stratum identifier is not needed for the user's analysis.

There are two commonly used estimators of the population mean under two phase sampling: the double expansion estimator (DEE) and the reweighted expansion estimator (REE), named by Kott and Stukel (1997). In general the REE is more efficient than the DEE in the situation of two-phase sampling for stratification when the y's within a stratum are homogeneous. Variance estimation for two-phase sampling is a challenging practical problem, and replication variance estimation is of interest among practitioners. Rao and Shao (1992) proposed a consistent jackknife variance estimator for the REE in the context of hot deck imputation treating the respondents as the second-phase sample. Kott and Stukel (1997) considered the same problem and concluded that the jackknife variance estimator works well for the REE if the first-phase sampling rate is negligible. The sampling rate, or the sampling fraction, $f_1 = nN^{-1}$ is called negligible if f_1 converges to zero under the asymptotic setup described in Section 2. Binder, Babyak, Brodeur, Hidiroglou and Jocelyn (2000) studied variance estimation for a similar two-phase sample design using the Taylor linearization method. Kim *et al.* (2006, KNF) provided a rigorous investigation of the replication method and considered replication for other types of estimators. The KNF method has been developed mainly under the situation where the first-phase sampling rate is negligible and the second-phase sampling is a stratified random sampling. If the first-phase sampling rate is not negligible, additional replicates are needed to get consistent variance estimates.

In this paper, we propose a new replication method for variance estimation under two-phase sampling. The proposed method is an extension of the KNF method to cover the situation where the first-phase sampling rate is not necessarily negligible. Unlike the KNF method, the proposed method does not require additional replicates for bias correction in the variance estimation, but does require adjustments in the replication weights. Also, the proposed method is applicable to unequal probability Poisson sampling within

1. Jae Kwang Kim, Department of Statistics, Iowa State University, Ames, Iowa 50011, U.S.A.; Cindy Long Yu, Department of Statistics, Center for Survey Statistics and Methodology, Iowa State University, Ames, Iowa 50011, U.S.A. E-mail: cindyyu@iastate.edu.

second-phase strata, which was not discussed in KNF. Because the proposed method is a replication-based method, it is very easy to implement and can be applied to various types of estimators.

The rest of the paper is organized as follows. In Section 2, the basic setup is introduced, and in Section 3, the proposed method is described. In Section 4, the proposed method is extended to other estimators in two-phase sampling. In Section 5, results from a limited simulation study are presented. Concluding remarks are made in Section 6.

2. Basic setup

For better motivation, in this section we simply assume the situation where the first phase is a simple random sample of size n from a finite population of size N and the second phase sampling is a stratified random sample. In section 3, the setup is extended to include any arbitrary measurable sampling in the first phase and unequal probability Poisson sampling within each stratum in the second phase. Using the information obtained from the first-phase sample, it is stratified into H strata for second-phase sampling. In stratum h, we have n_h first-phase sample elements and let A_{h1} be the set of indices for the first-phase sample elements in stratum h. In the second-phase sampling, a stratified random sample of size r is selected with sample size $r_h (\le n_h)$ in stratum h, where $r = \sum_{h=1}^{H} r_h$ and the sampling rate r_h/n_h is fixed for each stratum. To formally discuss the asymptotic theory, we assume a sequence of finite populations, a sequence of first-phase samples, and a sequence of second-phase samples, as described in KNF. In this asymptotic setup, we allow that the second-phase sample size r goes to infinity at the same rate as the first phase sample size n, *i.e.*, $r = O(n)$ and $r^{-1} = O(n^{-1})$, and H is fixed. Thus, in the setup of fixed H, $r_h^{-1} = O(n^{-1})$.

When the study variable y_i is observed in the second phase sample, the population mean of y is estimated by

$$\bar{y}_{tp} = \frac{1}{n}\sum_{h=1}^{H}\sum_{i\in A_{h2}}\frac{n_h}{r_h}\, y_i,$$

where A_{h2} is the set of indices for the second-phase sample elements that belong to stratum h. The variance of $\bar{y}_{tp}$ can be written as

$$\mathrm{Var}(\bar{y}_{tp}) = \left(\frac{1}{n}-\frac{1}{N}\right)S^2 + E\left\{\sum_{h=1}^{H}\left(\frac{n_h}{n}\right)^2\left(\frac{1}{r_h}-\frac{1}{n_h}\right)s_{h1}^2\right\} \quad (1)$$

where $\bar{y}_1 = n^{-1}\sum_{h=1}^{H}\sum_{i\in A_{h1}} y_i$, $S^2 = (N-1)^{-1}\sum_{i=1}^{N}(y_i - \bar{y}_N)^2$, $s_{h1}^2 = (n_h-1)^{-1}\sum_{i\in A_{h1}}(y_i - \bar{y}_{h1})^2$, and $\bar{y}_{h1} = n_h^{-1}\sum_{i\in A_{h1}} y_i$. Using

$$n^{-1}S^2 \doteq E\left\{n^{-1}\sum_{h=1}^{H} w_h[(\bar{y}_{h1}-\bar{y}_1)^2 + s_{h1}^2]\right\}$$

where $w_h = n^{-1}n_h$ and $\doteq$ indicates an approximation ignoring the terms of order $o(n^{-1})$, the variance term (1) is approximated by

$$\mathrm{Var}(\bar{y}_{tp}) \doteq E\left\{n^{-1}(1-f_1)\sum_{h=1}^{H} w_h\,(\bar{y}_{h1}-\bar{y}_1)^2 + \sum_{h=1}^{H}(r_h^{-1}-n_h^{-1}f_1)\,w_h^2\,s_{h1}^2\right\}, \quad (2)$$

where $f_1 = nN^{-1}$.

A consistent estimator of the variance of $\bar{y}_{tp}$ can be derived from (2) by replacing $\bar{y}_{h1}$ and s_{h1}^2 by their estimates $\bar{y}_{h2} = r_h^{-1}\sum_{i\in A_{h2}} y_i$ and $s_{h2}^2 = (r_h-1)^{-1}\sum_{i\in A_{h2}}(y_i - \bar{y}_{h2})^2$, respectively. That is, a consistent variance estimator is

$$\hat{V} = n^{-1}(1-f_1)\sum_{h=1}^{H} w_h(\bar{y}_{h2}-\bar{y}_2)^2 + \sum_{h=1}^{H}(r_h^{-1}-n_h^{-1}f_1)\,w_h^2\,s_{h2}^2, \quad (3)$$

where $\bar{y}_2 = \sum_{h=1}^{H} w_h\,\bar{y}_{h2}$. The variance estimator (3) is a linearized variance estimator.

Kott and Stukel (1997) and KNF developed a jackknife variance estimator by successively deleting units from the entire first-phase sample and then adjusting the weights. The full jackknife replicates are

$$\bar{y}_{tp}^{(k)} = \frac{1}{N}\sum_{h=1}^{H}\hat{N}_{h1}^{(k)}\,\bar{y}_{h2}^{(k)} \quad (4)$$

where k is the index of the unit deleted in the jackknife replicate,

$$\frac{1}{N}\hat{N}_{h1}^{(k)} = \sum_{i\in A_{h1}} w_i^{(k)} = \begin{cases}(n-1)^{-1}(n_h-1) & \text{if } k\in A_{h1}\\ (n-1)^{-1}n_h & \text{if } k\notin A_{h1}\end{cases}$$

and

$$\bar{y}_{h2}^{(k)} = \frac{\sum_{i\in A_{h2}} w_i^{(k)} y_i}{\sum_{i\in A_{h2}} w_i^{(k)}} = \begin{cases}(r_h-1)^{-1}(r_h\bar{y}_{h2}-y_k) & \text{if } k\in A_{h2}\\ \bar{y}_{h2} & \text{if } k\notin A_{h2}.\end{cases} \quad (5)$$

The full jackknife variance estimator of the form

$$\hat{V}_J = \sum_{k\in A_1}\frac{n-1}{n}(1-f_1)\,(\bar{y}_{tp}^{(k)}-\bar{y}_{tp})^2, \quad (6)$$

where $\bar{y}_{tp}^{(k)}$ is defined in (4), is asymptotically equivalent to

$$\hat{V}_J \doteq n^{-1}(1-f_1)\sum_{h=1}^{H} w_h(\bar{y}_{h2}-\bar{y}_2)^2 + (1-f_1)\sum_{h=1}^{H} r_h^{-1}\,w_h^2\,s_{h2}^2. \quad (7)$$

Thus, comparing (7) with (2), the bias of the jackknife variance estimator (6) is

$$\text{Bias}(\hat{V}_J) \doteq -E\left\{f_1 \sum_{h=1}^{H} (r_h^{-1} - n_h^{-1})\ s_{h2}^2\right\}.$$

Therefore, if the first-phase sampling rate is negligible in the sense of $f_1 \doteq 0$, the bias is negligible, *i.e.*, the bias = $o(n^{-1})$. Otherwise, the variance estimator underestimates the variance.

To consider a bias-corrected jackknife method, instead of (5), we consider

$$\bar{y}_{h2}^{(k)} = \begin{cases} (r_h - \delta_h)^{-1}(r_h \bar{y}_{h2} - \delta_h y_k) & \text{if } k \in A_{h2} \\ \bar{y}_{h2} & \text{if } k \notin A_{h2}, \end{cases} \quad (8)$$

where δ_h is to be determined. In (5), $\delta_h = 1$ was used. The jackknife variance estimator using (8) instead of (5) is asymptotically equivalent to

$$\hat{V}_J \doteq n^{-1}(1 - f_1) \sum_{h=1}^{H} w_h (\bar{y}_{h2} - \bar{y}_2)^2 + (1 - f_1) \sum_{h=1}^{H} \frac{(r_h - 1)\ \delta_h^2}{(r_h - \delta_h)^2}\ w_h^2\ s_{h2}^2.$$

Thus, the asymptotic bias is

$$\text{Bias}(\hat{V}_J) \doteq E\left[\sum_{h=1}^{H}\left\{(1 - f_1)\ \frac{(r_h - 1)\ \delta_h^2}{(r_h - \delta_h)^2} - \frac{1}{r_h}\left(1 - f_1 \frac{r_h}{n_h}\right)\right\} w_h^2\ s_{h2}^2\right].$$

The asymptotic bias is zero if

$$\delta_h = \frac{r_h}{1 + \sqrt{r_h (r_h - 1)}/d_h}$$

where $d_h = \sqrt{(1 - f_1 r_h n_h^{-1})/(1 - f_1)}$. Hence, with such determined δ_h in equation (8), the resulting jackknife variance estimator is approximately unbiased without assuming $f_1 \doteq 0$.

3. Proposed method

The proposed method in Section 2 is now extended to a more general first-phase sampling design. To do this, we need to assume that the replication variance estimator of the form

$$\hat{V}_1 = \sum_{k=1}^{L} c_k (\hat{\theta}^{(k)} - \hat{\theta})^2,$$

where $\hat{\theta} = \sum_{i \in A_1} w_i\ y_i$, and $\hat{\theta}^{(k)} = \sum_{i \in A_1} w_i^{(k)} y_i$, is consistent for the variance of $\hat{\theta}$ under the single (first) stage sampling design. That is,

$$\frac{\hat{V}_1}{\text{Var}(\hat{\theta})} - 1 = o_p(1). \quad (9)$$

Here L is the number of replicates. For most of the measurable designs, which are designs with all positive joint inclusion probabilities, we can construct a replication variance estimator satisfying (9) even when the sample rate $f = n/N$ is large. For example, see Fay (1984) and Flyer (1987). Brick and Morganstein (1996) describes the basic algorithm for WesVar, a commercially available software for replication variance estimation in survey sampling.

In this section, we also consider a more challenging case of stratified unequal probability sampling for the second phase. More specifically, the second phase sampling considered is unequal probability Poisson sampling within the second-phase strata. Fuller (1998) also considered Poisson sampling in the second phase and argued that Poisson sampling in the second phase sampling is a good approximation. An example of this in the context of forest surveys is that, in addition to forest types, the photo-interpreters can also identify tree density and tree height from the aerial photos taken in the first phase, which can be used to construct the second phase selection probabilities within each stratum (forest type).

In this section, we will focus on the REE-type estimator first since it is more efficient than the DEE-type, and extension to the DEE is discussed in Section 4. Let w_i be the first-phase sampling weight and let w_{i2} be the inverse of the conditional probability in the second-phase. That is, $w_{i2} = \pi_{i2}^{-1}$ where $\pi_{i2} = \Pr(i \in A_{h2} \mid i \in A_{h1})$. The REE-type estimator can be written as

$$\bar{y}_{tp} = \frac{1}{N} \sum_{h=1}^{H} \hat{N}_{h1}\, \bar{y}_{h2} \quad (10)$$

where $\hat{N}_{h1} = \sum_{i \in A_{h1}} w_i$ and $\bar{y}_{h2} = (\sum_{i \in A_{h2}} w_i \pi_{i2}^{-1})^{-1} \sum_{i \in A_{h2}} w_i \pi_{i2}^{-1} y_i$. In KNF, π_{i2} is assumed to be constant within the second-phase stratum.

We consider a replication-based approach for variance estimation of the REE-type estimator (10) when π_{i2} is not necessarily constant within the second-phase stratum. We consider the special case when the second-phase sampling design is Poisson sampling. Using the replication method satisfying (9), the KNF-type variance estimator can be applied to estimate the variance of $\bar{y}_{tp}$ in this situation. That is,

$$\hat{V}_{\text{KNF}} = \sum_{k=1}^{L} c_k (\bar{y}_{tp}^{(k)} - \bar{y}_{tp})^2, \quad (11)$$

where

$$\bar{y}_{tp}^{(k)} = \frac{1}{N} \sum_{h=1}^{H} \hat{N}_{h1}^{(k)}\, \bar{y}_{h2}^{(k)} \quad (12)$$

with $\bar{y}_{h2}^{(k)} = (\sum_{i\in A_{h2}} w_i^{(k)}\pi_{i2}^{-1})^{-1}\sum_{i\in A_{h2}} w_i^{(k)}\pi_{i2}^{-1}y_i$ and $\hat{N}_{h1}^{(k)} = \sum_{i\in A_{h1}} w_i^{(k)}$, and c_k is a factor associated with replicate k determined by the replication method. Under Poisson sampling in the second phase, we have the following asymptotic bias:

$$\text{Bias}(\hat{V}_{\text{KNF}}) = -\frac{1}{N^2}\sum_{h=1}^{H}\sum_{i\in U_h}\pi_{i2}^{-1}(1-\pi_{i2})(y_i-\bar{Y}_h)^2, \quad (13)$$

where U_h is the set of indices of population elements in stratum h and $\bar{Y}_h = N_h^{-1}\sum_{i\in U_h} y_i$. A sketched proof of (13) is presented in Appendix A.

An asymptotically unbiased estimator of the bias (13) is

$$\hat{V}_{\text{bias}} = -\frac{1}{N^2}\sum_{h=1}^{H}\sum_{i\in A_{h2}} w_i\,\pi_{i2}^{-2}(1-\pi_{i2})(y_i-\bar{y}_{h2})^2. \quad (14)$$

The bias is negligible if $n/N \doteq 0$. Thus, we can safely ignore the bias of the KNF-type variance estimator when the first-phase sampling rate is negligible. The bias can be arbitrarily large if the first-phase sampling rate n/N is not negligible. KNF also discuss a bias-correction replication method using additional replicates, which can lead to a large number of replicates. Creating additional replicates for bias-correction can be cumbersome for large scale surveys.

We consider an alternative bias-corrected replication variance estimator that does not require creating additional replicates. To develop a replication-based bias-corrected variance estimator, define a random variable

$$\delta_{ki} \overset{\text{indep}}{\sim} \text{Bernoulli}(p_k), \quad (15)$$

where p_k is to be determined. Let

$$\hat{V}_{\text{KNF}}^{*} = \sum_{k=1}^{L} c_k(\bar{y}_{tp}^{*(k)} - \bar{y}_{tp})^2 \quad (16)$$

where

$$\bar{y}_{tp}^{*(k)} = \frac{1}{N}\sum_{h=1}^{H}\hat{N}_{h1}^{(k)}\,\bar{y}_{h2}^{*(k)} \quad (17)$$

with $\hat{N}_{h1}^{(k)} = \sum_{i\in A_{h1}} w_i^{(k)}$,

$$\bar{y}_{h2}^{*(k)} = \frac{\sum_{i\in A_{h2}} w_i^{(k)} M_{i2}^{(k)}\pi_{i2}^{-1}y_i}{\sum_{i\in A_{h2}} w_i^{(k)} M_{i2}^{(k)}\pi_{i2}^{-1}} \quad (18)$$

with

$$M_{i2}^{(k)} = 1 + (\delta_{ki} - p_k)b_i \quad (19)$$

and b_i is also to be determined. By construction, $E_*(\delta_{ki} - p_k) = 0$, where E_* denotes that the expectation is taken with respect to the mechanism in (15). Thus, the replicates (18) create additional variation in the replication weights, where the additional variation in (18) comes from the distribution (15). A suitable choice of p_i and b_i can make the resulting variance estimator consistent.

Under the regularity conditions discussed in KNF, we have

$$E_*(\hat{V}_{\text{KNF}}^{*}) = \hat{V}_{\text{KNF}} + N^{-2}\sum_{h=1}^{H}\sum_{i\in A_{h2}} w_i^2\,b_i^2\,\pi_{i2}^{-2}\,u\,(y_i-\bar{y}_{h2})^2 + o_p(n^{-1}), \quad (20)$$

where $u = \sum_{k=1}^{L} c_k p_k(1-p_k)$. A sketched proof of (20) is presented in Appendix B. If b_i are determined by

$$b_i = \sqrt{(1-\pi_{i2})w_i^{-1}u^{-1}}, \quad (21)$$

the variance estimator (16) is consistent because the second term in (20) cancels out $\hat{V}_{\text{bias}}$ in (14). This is true even when the first-phase sampling rate n/N is not negligible. To guarantee nonnegative replication weights in (18), we require that b_i in (19) is ≤ 1. If we set $p_k = 0.5$, then

$$b_i = \sqrt{\frac{4\,(1-\pi_{i2})\,w_i^{-1}}{\sum_{k=1}^{L} c_k}},$$

which is less than or equal to 1 if $\sum_{k=1}^{L} c_k \ge 4$. In fact, the p_k's can be chosen to be any number between 0 and 1 as long as the resulting b_i in (21) is less than or equal to 1.

4. Extensions

In this section, we consider some extensions of the proposed replication method to types of two-phase estimators other than the REE in (10).

4.1 Double expansion estimator

In two-phase sampling, the double expansion estimator, termed by Kott and Stukel (1997), is also used. The double expansion estimator (DEE) has the simple form

$$\bar{y}_{\text{DEE}} = \frac{1}{N}\sum_{h=1}^{H}\sum_{i\in A_{h2}} w_i\,\pi_{i2}^{-1}\,y_i. \quad (22)$$

When the second-phase sample is a stratified random sample, $\pi_{i2} = r_h/n_h$ and the KNF method can be applied using the replicate

$$\bar{y}_{\text{DEE}}^{(k)} = \frac{1}{N}\sum_{h=1}^{H}\left(\frac{\sum_{i\in A_{h1}} w_i^{(k)}\,w_i^{-1}}{\sum_{i\in A_{h2}} w_i^{(k)}\,w_i^{-1}}\right)\sum_{i\in A_{h2}} w_i^{(k)}\,y_i.$$

The KNF variance estimator for DEE is consistent when the first-phase sampling rate is negligible. When the first-phase sampling rate is not negligible, we can use the replication method proposed in Section 3. The proposed replication method for the DEE creates replicates,

$$\bar{y}_{\text{DEE}}^{*(k)} = \frac{1}{N}\sum_{h=1}^{H}\sum_{i\in A_{h2}} w_i^{(k)} w_{i2}^{*(k)} y_i, \tag{23}$$

where

$$w_{i2}^{*(k)} = M_{i2}^{(k)} \frac{\sum_{i\in A_{h1}} w_i^{(k)} w_i^{-1}}{\sum_{i\in A_{h2}} w_i^{(k)} w_i^{-1} M_{i2}^{(k)}},$$

and $M_{i2}^{(k)}$ is the replication factor defined in (19). The bias of the replication variance estimator using replicate (23) is negligible if the replicates are constructed to satisfy (21).

If the second-phase sample is an unequal probability sample within each stratum, the replication method such as (23) is not directly applicable. The DEE in (22) is generally less efficient than the REE in (10). Note that the REE in (10) can also be expressed as

$$\bar{y}_{\text{REE}} = \frac{1}{N}\sum_{h=1}^{H}\sum_{i\in A_{h2}} w_i w_{i2}^{*} y_i, \tag{24}$$

where

$$w_{i2}^{*} = \pi_{i2}^{-1} \frac{\sum_{i\in A_{h1}} w_i}{\sum_{i\in A_{h2}} w_i \pi_{i2}^{-1}}. \tag{25}$$

The replicates (17) can be written

$$\bar{y}_{\text{REE}}^{*(k)} = \frac{1}{N}\sum_{h=1}^{H}\sum_{i\in A_{h2}} w_i^{(k)} w_{i2}^{*(k)} y_i, \tag{26}$$

where

$$w_{i2}^{*(k)} = M_{i2}^{(k)}\pi_{i2}^{-1} \frac{\sum_{i\in A_{h1}} w_i^{(k)}}{\sum_{i\in A_{h2}} w_i^{(k)} M_{i2}^{(k)} \pi_{i2}^{-1}} \tag{27}$$

and $M_{i2}^{(k)}$ is defined in (19).

4.2 Regression estimator

In two-phase sampling, auxiliary variables that are observed in the first-phase sample can be further used at the estimation stage. The two-phase regression estimator of the population total can be written in the form

$$\hat{Y}_{t,\text{REG}} = \hat{\mathbf{T}}_{x,1}' \hat{\boldsymbol{\beta}}_2 \tag{28}$$

where $\hat{\mathbf{T}}_{x,1} = \sum_{i\in A_1} w_i \mathbf{x}_i$ is the vector of estimated population totals of the control variable $\mathbf{x}_i$ estimated with the first-phase sample and $\hat{\boldsymbol{\beta}}_2 = (\sum_{i\in A_2} w_i w_{i2}^{*} \mathbf{x}_i \mathbf{x}_i')^{-1} \sum_{i\in A_2} w_i w_{i2}^{*} \mathbf{x}_i y_i$ is a vector of estimated regression coefficients estimated with the second-phase sample and w_{i2}^{*} is given by (25). Note that the regression estimator in (28) can incorporate the stratified sampling design in the second-phase if $\mathbf{x}_i$ includes the vector of stratum indicators.

Using the arguments of Section 3, the k^{th} replicate for $\hat{Y}_{t,\text{REG}}$ can be constructed by

$$\hat{Y}_{t,\text{REG}}^{(k)} = \hat{\mathbf{T}}_{x,1}^{(k)\prime} \hat{\boldsymbol{\beta}}_2^{(k)}, \tag{29}$$

where

$$\hat{\mathbf{T}}_{x,1}^{(k)} = \sum_{i\in A_1} w_i^{(k)} \mathbf{x}_i$$

$$\hat{\boldsymbol{\beta}}_2^{(k)} = \left(\sum_{i\in A_2} w_i^{(k)} w_{i2}^{*(k)} \mathbf{x}_i \mathbf{x}_i'\right)^{-1} \sum_{i\in A_2} w_i^{(k)} w_{i2}^{*(k)} \mathbf{x}_i y_i$$

and $w_{i2}^{*(k)}$ is defined in (27).

The replication method (29) can be directly applicable to the two-phase calibration estimator that was discussed in Hidiroglou and Särndal (1998). If $H = 1$, then the replicate of $\hat{\boldsymbol{\beta}}_2$ in (29) reduces to

$$\hat{\boldsymbol{\beta}}_2^{(k)} = \left(\sum_{i\in A_2} w_i^{(k)} M_{i2}^{(k)} \pi_{i2}^{-1} \mathbf{x}_i \mathbf{x}_i'\right)^{-1} \sum_{i\in A_2} w_i^{(k)} M_{i2}^{(k)} \pi_{i2}^{-1} \mathbf{x}_i y_i.$$

5. Simulation study

To study the finite sample performance of the proposed estimators, we conducted a limited simulation study. In the simulation, we first generated an artificial finite population of size $N = 1{,}000$ with five variables $(z_i, q_i, x_i, y_i, u_i)$, where the population elements are independently generated from $z_i \sim \exp(1) + 2$; $q_i \sim \chi^2(1) + 2$; $x_i \sim N(2, 1)$; $u_i \sim \text{Unif}\{1, 2, 3, 4\}$, where $\text{Unif}\{1, ..., G\}$ denotes a discrete uniform distribution with support $\{1, ..., G\}$; and

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \beta_3 q_i + e_i$$

with $(\beta_0, \beta_1, \beta_2, \beta_3) = (0, 2, 1, 1)$ and $e_i \sim N(0, 1)$. The variables z_i, q_i, x_i, u_i, and e_i are mutually independent. The stratum for the second-phase sampling was defined using variable u_i. Variable x_i was used to compute the two-phase regression estimator (28) with $\mathbf{x}_i = (1, x_i)'$, variable z_i was used as a size measure for the unequal probability sampling in the first phase sampling, and variable q_i was used as a size measure for the unequal probability sampling in the second phase sampling.

To obtain unequal probability samples for this simulation study, we used either Poisson sampling or Rao-Sampford sampling (Rao 1965 and Sampford 1967), with selection probabilities proportional to the measure of the size variable. Note that the final sample size is random under Poisson sampling but is fixed under Rao-Sampford sampling.

The simulation setup employed a $2 \times 3 \times 2$ factorial structure with three factors. The factors are

1. Sampling for the first-phase sample (2): Simple random sampling of size $n = 200$ versus the Rao-Sampford sampling of size $n = 200$ using z_i as the measure of size.

2. Sampling for the second-phase sample (3): Stratified random sampling of size $r_h = 25$, stratified Poisson sampling with expected sample size $r_h = 25$ using q_i as the size measure for the unequal probability sampling, and stratified Rao-Sampford sampling of size $r_h = 25$ using q_i as the size measure for the unequal probability sampling.
3. Variance estimation methods (2): The KNF estimator (11) without additional replication versus the proposed variance estimator using (16) were computed based on the jackknife method.

From the finite population generated above, we generated $B = 5{,}000$ independent Monte Carlo samples for simulation. For the designs with Rao-Sampford sampling in the first phase, we used the jackknife variance estimation method proposed by Berger (2007), which gives a consistent estimator of the first phase sampling variance. The parameter of interest is the population mean of the y variable. From each Monte Carlo sample, we computed two point estimators, the REE in (24) and the regression estimator (REG) in (28) using the auxiliary variable $(1, x_i)$. Relative biases of the variance estimators were computed by dividing the Monte Carlo bias of the variance estimator by the Monte Carlo variance of the point estimator.

Table 1 shows the mean and variance of the two point estimators. For point estimation, the regression estimator is significantly more efficient than the REE for this population because the auxiliary variable x is correlated with the study variable y. The theoretical asymptotic variance of the regression estimator under simple random sampling in the first phase and stratified random sampling in the second phase is approximately equal to

$$\left(\frac{1}{200}-\frac{1}{1{,}000}\right)8+\left(\frac{1}{100}-\frac{1}{200}\right)4 = 0.052$$

and the theoretical asymptotic variance of the REE under the same design is, approximately, $(1/100-1/1{,}000)\,8 = 0.072$, which is consistent with the numerical results in Table 1. The Rao-Sampford sampling in the second phase is slightly more efficient than the Poisson sampling because of the fixed sample size in the Rao-Sampford sampling.

Table 2 shows the relative bias (RB) and coefficient of variation (CV) of the two variance estimators. Relative biases of the variance estimators were computed by dividing the Monte Carlo bias of the variance estimator by the Monte Carlo variance of the point estimator. Coefficients of variation of the variance estimator were computed by dividing the Monte Carlo standard error of the variance estimator by the Monte Carlo average of the variance estimator.

Table 1
Mean and variance of the point estimators (5,000 samples)

Estimator	First-phase Sampling	Second-Phase Sampling	Mean	Variance
REE	SRS	St. SRS	10.0	0.0749
		St. Poi	10.0	0.0784
		St. RS	10.0	0.0754
	RS	St. SRS	10.0	0.0768
		St. Poi	10.0	0.0827
		St. RS	10.0	0.0781
REG	SRS	St. SRS	10.0	0.0540
		St. Poi	10.0	0.0510
		St. RS	10.0	0.0495
	RS	St. SRS	10.0	0.0551
		St. Poi	10.0	0.0531
		St. RS	10.0	0.0515

REE: reweighted expansion estimator (23),
REG: regression estimator (27),
SRS: Simple random sampling,
RS: Rao-Sampford sampling,
St. SRS: Stratified simple random sampling,
St. Poi: Stratified Poisson sampling,
St. RS: Stratified Rao-Sampford sampling.

Table 2
Relative bias (RB) and coefficient of variation (CV) for the variance estimators (5,000 samples)

Method	Estimator	First-phase Sampling	Second-Phase Sampling	RB (%)	CV(%)
KNF	REE	SRS	St. SRS	-11.25	18.22
			St. Poi	-9.56	18.67
			St. RS	-7.75	15.35
		RS	St. SRS	-8.05	18.61
			St. Poi	-9.03	20.84
			St. RS	-5.73	17.27
	REG	SRS	St. SRS	-6.76	22.32
			St. Poi	-6.06	15.81
			St. RS	-3.26	12.82
		RS	St. SRS	-4.17	21.74
			St. Poi	-3.64	16.92
			St. RS	-3.20	13.78
New	REE	SRS	St. SRS	0.09	18.23
			St. Poi	-1.23	19.70
			St. RS	-0.04	16.06
		RS	St. SRS	0.78	19.78
			St. Poi	-2.07	21.26
			St. RS	1.00	17.67
	REG	SRS	St. SRS	-0.61	22.00
			St. Poi	-0.57	16.55
			St. RS	-0.08	13.36
		RS	St. SRS	0.67	22.86
			St. Poi	-0.01	16.97
			St. RS	0.59	14.02

KNF: Kim *et al.* (2006) variance estimator without additional replicates for bias correction,
New: the proposed variance estimator (16),
REE: reweighted expansion estimator (23),
REG: regression estimator (27),
SRS: Simple random sampling,
RS: Rao-Sampford sampling,
St. SRS: Stratified simple random sampling,
St. Poi: Stratified Poisson sampling,
St. RS: Stratified Rao-Sampford sampling.

In this simulation, because the first-phase sampling fraction is not negligible $(n/N = 0.2)$, the KNF variance estimator without additional replicates underestimates the true variance and the proposed variance estimator estimates the variance with smaller bias, less than 3% in absolute values in all cases, which is consistent with the theory in Section 3 and Section 4. The absolute value of the relative biases in the KNF variance estimator are big because, although in (29) the variance due to $\hat{\mathbf{T}}_{x,1}$ is consistently estimated, the variance due to $\hat{\boldsymbol{\beta}}_2$ is underestimated without additional replicates. The relative biases in our proposed variance estimator are reduced because replicates (18) create additional variation in the replication weights through additional perturbation δ_k drawn from a properly chosen distribution. The proposed variance estimator shows slightly bigger CVs than the KNF method because it involves extra randomness due to generating δ_{ki} from (15).

6. Concluding remarks

Replication variance estimation under two-phase sampling is an importance practical problem in survey sampling and the KNF method is a useful tool in this direction. In this article, we propose an extension of the KNF method in that it can be directly applicable when the first-phase sampling rate is non-negligible, without increasing the number of replicates. The proposed method is also applicable to unequal probability Poisson sampling within each stratum in the second-phase sample. Although the theory has been developed only under Poisson sampling in the second phase, the simulation results in section 5 show that the proposed method works reasonably well for other unequal probability sampling designs, such as the Rao-Sampford sampling design. Since the proposed replication method provides consistent variance estimators for population means, it can be readily applied to other finite population parameters which are smooth functions of population means.

In some large scale surveys, the number of replicates can be quite large because it uses the same number of replicates for the first-phase sample. If one wishes to reduce the number of replicates further, the method of Fuller (1998) or Kim and Sitter (2003) can be considered. Further investigation in this direction will be a topic of future study.

Acknowledgements

The research was supported by a Cooperative Agreement No. 68-3A75-4-122 between the USDA Natural Resources Conservation Service and the Center for Survey Statistics and Methodology at Iowa State University. The authors wish to thank Wayne Fuller and two anonymous referees for helpful comments.

Appendix

A. Proof of (13)

Let $\mathbf{a} = (a_1, ..., a_N)$ where a_i is the extended version of the second-phase sampling indicator as discussed in Kim *et al.* (2006). That is, $a_i = 1$ if unit i is selected for the second-phase sample once it is in the first-phase sample and $a_i = 0$ otherwise.

By assumption (9), conditional on $\mathbf{a}$, we have

$$\sum_{k=1}^{L} c_k (\bar{y}_{h2}^{(k)} - \bar{y}_{h2})^2 = \mathrm{Var}(\bar{y}_{h2} \mid \mathbf{a}) + o_p(n^{-1}).$$

Thus, the bias of $\sum_{k=1}^{L} c_k (\bar{y}_{h2}^{(k)} - \bar{y}_{h2})^2$ as an estimator for $\mathrm{Var}(\bar{y}_{h2})$ is then equal to, ignoring $o(n^{-1})$ terms,

$$E\{\mathrm{Var}(\bar{y}_{h2} \mid \mathbf{a})\} - \mathrm{Var}(\bar{y}_{h2}) = \mathrm{Var}\{E(\bar{y}_{h2} \mid \mathbf{a})\}.$$

Using the extended definition of a_i, we have

$$E(\bar{y}_{h2} \mid \mathbf{a}) = \frac{\sum_{i \in U_h} \pi_{i2}^{-1} a_i y_i}{\sum_{i \in U_h} \pi_{i2}^{-2} a_i}$$

and, by the Poisson sampling assumption of a_i's,

$$\mathrm{Var}\left(\frac{\sum_{i \in U_h} \pi_{i2}^{-1} a_i y_i}{\sum_{i \in U_h} \pi_{i2}^{-1} a_i} \right) = N_h^{-2} \sum_{i \in U_h} \pi_{i2}^{-1} (1 - \pi_{i2}) (y_i - \bar{Y}_h)^2 + o(N^{-1}). \quad \text{(A.1)}$$

Thus, the bias of the KNF variance estimator is of the form (13) under the Poisson sampling assumption of a_i.

B. Proof of (20)

For each k,

$$\bar{y}_{tp}^{*(k)} - \bar{y}_{tp} = \bar{y}_{tp}^{*(k)} - \bar{y}_{tp}^{(k)} + \bar{y}_{tp}^{(k)} - \bar{y}_{tp},$$

where $\bar{y}_{tp}^{(k)}$ is defined in (12). Thus,

$$\hat{V}_{\mathrm{KNF}}^{*} = \sum_{k=1}^{L} c_k (\bar{y}_{tp}^{*(k)} - \bar{y}_{tp})^2 = \sum_{k=1}^{L} c_k (\bar{y}_{tp}^{(k)} - \bar{y}_{tp})^2 + 2 \sum_{k=1}^{L} c_k (\bar{y}_{tp}^{(k)} - \bar{y}_{tp})(\bar{y}_{tp}^{*(k)} - \bar{y}_{tp}^{(k)}) + \sum_{k=1}^{L} c_k (\bar{y}_{tp}^{*(k)} - \bar{y}_{tp}^{(k)})^2. \quad \text{(B.1)}$$

By the construction of $\bar{y}_{tp}^{*(k)}$, we have

$$E_*(\bar{y}_{tp}^{*(k)}) = \bar{y}_{tp}^{(k)} + o_p(n^{-1}). \quad \text{(B.2)}$$

Also, writing $q_{ki} = M_{i2}^{(k)} - 1$, we have $q_{ki} = O_p(n^{-1/2})$ and we can apply a Taylor expansion to get

$$\bar{y}_{h2}^{*(k)} = \bar{y}_{h2}^{(k)} + \frac{\sum_{i\in A_{h2}} w_i^{(k)} \pi_{i2}^{-1} q_{ki} (y_i - \bar{y}_{h2}^{(k)})}{\sum_{i\in A_{h2}} w_i^{(k)} \pi_{i2}^{-1}} + o_p(n^{-1}). \quad \text{(B.3)}$$

Also, because

$$\frac{1}{N_h} \sum_{i\in A_{h2}} w_i^{(k)} \pi_{i2}^{-1} z_i - \frac{1}{N_h} \sum_{i\in A_{h2}} w_i \pi_{i2}^{-1} z_i = O_p(n^{-1})$$

for any z variable with bounded fourth moments, it can be shown that (B.3) reduces to

$$\bar{y}_{h2}^{*(k)} = \bar{y}_{h2}^{(k)} + \frac{\sum_{i\in A_{h2}} w_i \pi_{i2}^{-1} q_{ki} (y_i - \bar{y}_{h2})}{\sum_{i\in A_{h2}} w_i \pi_{i2}^{-1}} + o_p(n^{-1}).$$

Hence, we can write

$$\sum_{k=1}^{L} c_k (\bar{y}_{tp}^{*(k)} - \bar{y}_{tp}^{(k)})^2 =$$

$$\sum_{k=1}^{L} c_k \left\{ N^{-1} \sum_{h=1}^{H} \sum_{i\in A_{h2}} w_i \pi_{i2}^{-1} q_{ki} (y_i - \bar{y}_{h2}) \right\}^2 + o_p(n^{-1}). \quad \text{(B.4)}$$

Inserting (B.2) and (B.4) into (B.1), we have

$$E_*(\hat{V}_{\text{KNF}}^*) = \hat{V}_{\text{KNF}} + \frac{1}{N^2} \sum_{k=1}^{L} c_k \sum_{h=1}^{H} \sum_{i\in A_{h2}} w_i^2 E_*(q_{ki}^2) \pi_{i2}^{-2} (y_i - \bar{y}_{h2})^2 + o_p(n^{-1}),$$

and because $E_*(q_{ki}^2) = p_k(1 - p_k)\, b_i^2$, we have (20).

References

Berger, Y.G. (2007). A jackknife variance estimator for unistage stratified samples with unequal probabilities. *Biometrika*, 94, 953-964.

Binder, D.A., Babyak, C., Brodeur, M., Hidiroglou, M. and Jocelyn, W. (2000). Variance estimation for two-phase stratified sampling. *The Canadian Journal of Statistics*, 28, 751-764.

Breidt, F.J. and Fuller, W.A. (1993). Regression weighting for multipurpose samplings. *Sankhyā*, B, 55, 297-309.

Brick, J.M., and Morganstein, D. (1996). WesVarPC: Software for computing variance estimates from complex designs. *Proceedings of the 1996 Annual Research Conference*, U.S. Bureau of the Census, 861-866.

Cochran, W.G. (1977). *Sampling Techniques*. New York: John Wiley & Sons, Inc.

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

Flyer, P. (1987). Finite population correction for replication estimates of variance. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 732-736.

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

Fuller, W.A. (2003). Estimation for multiple phase samples. In *Analysis of Survey Data*, (Eds., R.L. Chambers and C.J. Skinner). Wiley, Chichester, England, 307-322.

Hidiroglou, M.A. (2001). Double sampling. *Survey Methodology*, 27, 143-154.

Hidiroglou, M.A., and Särndal, C.-E. (1998). Use of auxiliary information for two-phase sampling. *Survey Methodology*, 24, 11-20.

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

Kim, J.K., and Sitter, R.R. (2003). Efficient variance estimation for two-phase sampling. *Statistica Sinica*, 13, 641-653.

Kott, P.S., and Stukel, D.M. (1997). Can the jackknife be used with a two-phase Sample? *Survey Methodology*, 23, 81-89.

Lu, W., and Sitter, R.R. (2006). Disclosure risk and variance estimation. *Proceedings of Statistics Canada international symposium series*, 11-522-XIE.

Neyman, J. (1938). Contribution to the theory of sampling human populations. *Journal of the American Statistical Association*, 33, 101-116.

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

Rao, J.N.K. (1973). On double sampling for stratification and analytical surveys. *Biometrika*, 60, 125-133.

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 Sitter, R.R. (1995). Variance estimation under two-phase sampling with application to imputation for missing data. *Biometrika*, 82, 453-460.

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

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

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