Statistica Sinica **13**(2003), 641-653

EFFICIENT REPLICATION VARIANCE ESTIMATION FOR TWO-PHASE SAMPLING

J. K. Kim and R. R. Sitter

Hankuk University of Foreign Studies and Simon Fraser University

Abstract: Variance estimation for the regression estimator for a two-phase sample is investigated. A replication variance estimator with number of replicates equal to or slightly larger than the size of the second-phase sample is developed. In surveys where the second-phase sample is much smaller than the first-phase sample, the procedure has practical advantages. The method is similar in spirit to one of Fuller (1998), but is more easily applied and avoids some of that method's difficulties. The proposed method can be directly applied to variance estimation for the double expansion estimator and the reweighted expansion estimator. In these cases, the proposed method is asymptotically equivalent to the full jackknife, but uses a smaller number of replications.

Key words and phrases: Double sampling, jackknife, regression estimator.

1. Introduction

It is common in surveys to use *two-phase* or *double* sampling when it is relatively inexpensive to draw a large first-phase sample for which a vector auxiliary variate, $\boldsymbol{x}$, correlated with the characteristic of interest $\boldsymbol{y}$, alone is observed. A second-phase subsample of the initial first-phase sample is then drawn and both $\boldsymbol{y}$ and $\boldsymbol{x}$ are measured. Various estimation strategies exist for combining the information from both phases of sampling to estimate characteristics of the population based on $\boldsymbol{y}$ or $(\boldsymbol{y}, \boldsymbol{x})$. In this article, we will focus specifically on two-phase regression estimation. For more general discussion see Cochran (1977), Wolter (1985), Särndal, Swensson and Wretman (1992) and Lohr (1999). These references also give descriptions of linearization variance estimators for this setting.

Jackknife variance estimators for two-phase sampling have been developed in Rao and Sitter (1995, 1997) and Sitter (1997). These procedures create jackknife replicates for each unit in the first-phase sample. Kott (1990) and Kott and Stukel (1997) consider variance estimation when the second-phase stratification differs from the first-phase stratification, and suggest a jackknife variance estimator for a particular regression estimator which they term the reweighted expansion estimator (REE). Kim, Navarro and Fuller (2000) develop a jackknife

variance estimator to handle what Kott and Stukel (1997) term the double expansion estimator (DEE). It should be noted that some of the ideas for these jackknife estimators are implicit in Rao and Shao (1992), though they consider the jackknife for random imputation. For discussion on the connection between their imputation cells and the second-phase strata, see Kott and Stukel (1997, p.83). Breidt and Fuller (1993) propose a replication method for multi-phase sampling.

A key feature of these full jackknife variance estimators is that replicates are formed for each unit in the first-phase sample. When the first-phase sample is very large, as is common, and in particular much larger than the second-phase sample, there are practical reasons why having so many replicates may be undesirable. When the final user is different than the data provider, it is common practice to include the set of replicate weights in the data set. Thus a large number of replicates in a large survey with many measured characteristics causes what turns out to be unnecessary, computational and storage burdens on the end user.

Fuller (1998) recognizes this practical issue and proposes a creative solution whereby the required number of jackknife replicate weights can be reduced in some cases by considering a decomposition of the variance of the regression estimator into two parts. He then argues that if one has a replication variance estimator for estimating the first term with fewer replicates than the full jackknife, then one can intelligently adjust these replicate weights so as to add back in the second term of the decomposition.

We propose instead to consider the replicates of the full jackknife directly by viewing those obtained by deleting units in the second phase separately from those obtained from the first phase but which were not then subsampled in the second. We can then use a strategy different from that of Fuller to form fewer replicate weights to capture the second term. The result is a jackknife with fewer replicate weights which retains the efficiency of the full jackknife. In this way we can "piggy-back" on the full jackknife. In some common settings this turns out to be very simple.

In Section 2, we develop the proposal for a general two-phase regression estimator. In Section 3, we highlight its application by considering the DEE when the second-phase strata are nested within the first-phase strata. Section 4 discusses the impact of relaxing the nested structure. In Section 5 we consider cluster sampling at the first phase where the clusters are ignored at the second phase to illustrate a simple situation where the proposed method applies but it is difficult to know what one should do in using the Fuller (1998) strategy. A limited simulation is provided in Section 6.

2. Reducing the Number of Replicate Weights in the Jackknife for the Two-phase Regression Estimator

Consider estimation of the population total, $\boldsymbol{Y}$, of vector $\boldsymbol{y}$ from a two-phase sample. Let $\hat{\boldsymbol{X}}_1$ be an unbiased estimator of the population total, $\boldsymbol{X}$, of vector $\boldsymbol{x}$ constructed from the first-phase sample, A_1, $\hat{\boldsymbol{X}}_2$ be an unbiased estimator of the population total of $\boldsymbol{x}$ constructed from the second-phase sample, A_2, and $\hat{\boldsymbol{Y}}_2$ be an unbiased estimator of the population total of $\boldsymbol{y}$ constructed from A_2. Write

$$\hat{\boldsymbol{X}}_1 = \sum_{i \in A_1} w_i \boldsymbol{x}_i \quad \text{and} \quad \left(\hat{\boldsymbol{X}}_2, \hat{\boldsymbol{Y}}_2\right) = \sum_{i \in A_2} w_i w_{i2} \left(\boldsymbol{x}_i, \boldsymbol{y}_i\right). \tag{2.1}$$

The first-phase sampling weight, w_i, is often the inverse of the inclusion probability for the first-phase sampling. The second-phase sampling weight, w_{i2}, is often the inverse of the conditional selection probability for the second-phase sample given the first-phase sample.

For simplicity, consider a scalar y variable. The regression estimator of Y takes the form

$$\hat{Y}_{reg} = \hat{Y}_2 + (\hat{\boldsymbol{X}}_1 - \hat{\boldsymbol{X}}_2)' \hat{\boldsymbol{\beta}}_{(2)}, \tag{2.2}$$

where $\hat{\boldsymbol{\beta}}_{(2)} = (\sum_{j \in A_2} w_j w_{j2} \boldsymbol{x}_j \boldsymbol{x}_j')^{-1} \sum_{j \in A_2} w_j w_{j2} \boldsymbol{x}_j y_j$, and $w_j w_{j2}$ are the two-phase weights used in (2.1). If we include 1 as the first component of $\boldsymbol{x}$, i.e. an intercept, then we can rewrite $\hat{Y}_{reg}$ as

$$\hat{Y}_{reg} = \hat{\boldsymbol{X}}'_1 \hat{\boldsymbol{\beta}}_{(2)}. \tag{2.3}$$

Let us imagine we had the observed y on the entire first-phase sample. Then the full first-phase sample variance of $\hat{Y}_1 = \sum_{i \in A_1} w_i y_i$ can be estimated by a full jackknife estimator of the form

$$v_J(\hat{Y}_1) = \sum_{k \in A_1} c_k (\hat{Y}_1^{(k)} - \hat{Y}_1)^2, \tag{2.4}$$

where $\hat{Y}_1^{(k)} = \sum_{i \in A_1} w_i^{(k)} y_i$ and c_k is a factor associated with the sampling design. For example, in an unstratified setting commonly used replication weights are $w_i^{(k)} = (n-1)^{-1} n w_i$ for $k \neq i$ and $w_i^{(i)} = 0$, and the factor c_k is equal to $n^{-1}(n-1)$.

Let the two-phase sample variance of $\hat{Y}_2 = \sum_{i \in A_2} w_i w_{i2} y_i$ also be estimated by a full jackknife estimator of the form

$$v_J(\hat{Y}_2) = \sum_{k \in A_1} c_k (\hat{Y}_2^{(k)} - \hat{Y}_2)^2, \tag{2.5}$$

where $\hat{Y}_2^{(k)} = \sum_{i \in A_2} w_i^{(k)} w_{i2}^{(k)} y_i$ and $w_{i2}^{(k)}$ is the k-th replicate of the second phase weighting factor w_{i2} (see Kim, Navarro and Fuller (2000)).

Now, given the variance estimators for the first-phase estimator of the form (2.4) and for the two-phase direct estimator of the form (2.5), replication variance estimators are available for the two-phase regression estimators because the regression estimator is a smooth function of the direct estimators on the first-phase sample and the second-phase sample. Thus, one can define for $k \in A_1$,

$$\hat{Y}_{reg}^{(k)} = \sum_{i \in A_2} \alpha_i^{(k)} y_i = \hat{\boldsymbol{X}}_1^{(k)'} \hat{\boldsymbol{\beta}}_{(2)}^{(k)}, \tag{2.6}$$

where $\hat{\boldsymbol{X}}_1^{(k)} = \sum_{i \in A_1} w_i^{(k)} \boldsymbol{x}_i$ and $\hat{\boldsymbol{\beta}}_{(2)}^{(k)} = (\sum_{j \in A_2} w_j^{(k)} w_{j2}^{(k)} \boldsymbol{x}_j \boldsymbol{x}_j')^{-1} \sum_{j \in A_2} w_j^{(k)} w_{j2}^{(k)} \times \boldsymbol{x}_j y_j$. The full jackknife variance estimator would then take the form $v_J(\hat{Y}_{reg}) = \sum_{k \in A_1} c_k (\hat{Y}_{reg}^{(k)} - \hat{Y}_{reg})^2$, and would require the formation of n_1 sets of replicate weights, for n_2 records.

When the second-phase sample is much smaller than the first-phase sample, we may wish to reduce the total number of replicates. Having a smaller number of replicates is particularly important in practice not only because of faster computation but also because of the smaller storage needed. When the final user is different from the data provider, it is a common practice to include the replication weights in the data set.

Fuller (1998) recognizes this problem and is able to reduce the number of required replicates in such cases by considering the regression estimator in two parts, namely as,

$$\hat{Y}_{reg} \doteq (\hat{Y}_2 - \hat{\boldsymbol{X}}_2' \boldsymbol{\beta}) + \hat{\boldsymbol{X}}_1' \boldsymbol{\beta},$$

and decomposing the variance into that corresponding to each term. He then shows that, if one has a replication method for estimating the variance of the first term, one can use a simple method to adjust it to add back the variance for the second term provided the covariance between the terms is negligible. There are two difficulties in some cases: (i) the covariance between the terms may not be negligible; (ii) it can be difficult to obtain a replication method for estimating the unconditional variance of the first term that requires a relatively small number of replicates.

Instead, note that

$$\begin{aligned} v_J(\hat{Y}_{reg}) &= \sum_{k \in A_2} c_k (\hat{Y}_{reg}^{(k)} - \hat{Y}_{reg})^2 + \sum_{k \in A_1 \cap A_2^c} c_k (\hat{Y}_{reg}^{(k)} - \hat{Y}_{reg})^2 \\ &= v_{J,2} + v_{J,1-2}, \end{aligned} \tag{2.7}$$

and consider the second term. It turns out to be possible in many situations to create fewer replicates than the number of elements in $A_1 \cap A_2^c$ to capture the

second term of (2.7). To see this, rewrite

$$\begin{aligned} v_{J,1-2} &= \hat{\boldsymbol{\beta}}'_{(2)}\Big[\sum_{k\in A_1\cap A_2^c} c_k(\hat{\boldsymbol{X}}_1^{(k)} - \hat{\boldsymbol{X}}_1)(\hat{\boldsymbol{X}}_1^{(k)} - \hat{\boldsymbol{X}}_1)'\Big]\hat{\boldsymbol{\beta}}_{(2)} \\ &\quad + \sum_{k\in A_1\cap A_2^c} c_k\Big[\hat{\boldsymbol{X}}_1^{(k)'}(\hat{\boldsymbol{\beta}}_{(2)}^{(k)} - \hat{\boldsymbol{\beta}}_{(2)})\Big]^2 \\ &\quad + 2\sum_{k\in A_1\cap A_2^c} c_k\Big[(\hat{\boldsymbol{X}}_1^{(k)} - \hat{\boldsymbol{X}}_1)'\hat{\boldsymbol{\beta}}_{(2)}\Big]\Big[\hat{\boldsymbol{X}}_1^{(k)'}(\hat{\boldsymbol{\beta}}_{(2)}^{(k)} - \hat{\boldsymbol{\beta}}_{(2)})\Big] \\ &= \hat{\boldsymbol{\beta}}'_{(2)}\tilde{\boldsymbol{V}}_x\hat{\boldsymbol{\beta}}_{(2)} + V_2 + 2V_{12}. \end{aligned}$$

Then, in some common cases

$$\hat{\boldsymbol{\beta}}_{(2)}^{(hi)} \doteq \hat{\boldsymbol{\beta}}_{(2)} \tag{2.8}$$

for all units (hi) that belong to the first-phase sample but not the second, and thus $V_2 \doteq V_{12} \doteq 0$. For example, if the first-phase sampling design is simple random sampling and the second phase sampling design is stratified random sampling, then including the second phase stratum vector in the column space of $\boldsymbol{x}$ will make condition (2.8) hold. We will discuss other such situations in the sequel.

If (2.8) holds, we can employ a tactic similar to that of Fuller (1998) but applied to this portion of v_J. That is, let $\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_{n_2}$ be a set of m-dimensional vectors, where $m < n_2$ is the dimension of $\boldsymbol{x}$, and $\sum_{j=1}^{n_2} \boldsymbol{\delta}_j\boldsymbol{\delta}'_j = \tilde{\boldsymbol{V}}_x$. For example, let $\boldsymbol{\gamma}_j$ be the characteristic vectors of $\tilde{\boldsymbol{V}}_x$ and λ_j their corresponding roots. Then define $\boldsymbol{\delta}_j = \lambda_j^{1/2}\boldsymbol{\gamma}_j$ for $j = 1, \ldots, m$, and $\boldsymbol{\delta}_j = \boldsymbol{0}$ for $j = m+1, \ldots, n_2$.

We obtain a set of $2n_2$ adjusted replicate weights, $\tilde{\alpha}_i^{(k)}$, for the set $k \in A_2$ such that $\tilde{Y}_{reg}^{(k)} = \sum_{i\in A_2} \tilde{\alpha}_i^{(k)} y_i = \hat{\boldsymbol{X}}_1^{(k)'}\hat{\boldsymbol{\beta}}_{(2)}^{(k)} + c_k^{-1/2}\boldsymbol{\delta}'_k\hat{\boldsymbol{\beta}}_{(2)}$. That is,

$$\tilde{\alpha}_i^{(k)} = \alpha_i^{(k)} + c_k^{-1/2}\boldsymbol{\delta}'_k\Big(\sum_{j\in A_2} w_j w_{j2}\boldsymbol{x}_j\boldsymbol{x}'_j\Big)^{-1} w_i w_{i2}\boldsymbol{x}_i, \tag{2.9}$$

for $k \in A_2$, where $\alpha_i^{(k)}$ is given in (2.6). If one repeats this entire process creating $\tilde{Y}_{reg2}^{(k)} = \hat{\boldsymbol{X}}_1^{(k)'}\hat{\boldsymbol{\beta}}_{(2)}^{(k)} - c_k^{-1/2}\boldsymbol{\delta}'_k\hat{\boldsymbol{\beta}}_{(2)}$ by subtracting $c_k^{-1/2}\boldsymbol{\delta}_k$ in the right hand side of (2.9), for $k \in A_2$, then we can define a new jackknife variance estimator as

$$\tilde{v}_J = \sum_{k\in A_2} c_k(\tilde{Y}_{reg}^{(k)} - \hat{Y}_{reg})^2 + \sum_{k\in A_2} c_k(\tilde{Y}_{reg2}^{(k)} - \hat{Y}_{reg})^2.$$

If so it follows that $\tilde{v}_J \doteq v_J$ and, even though only $2n_2$ replicates are needed, the efficiency of the full jackknife variance estimator is retained.

Though this is rather cute, it is sometimes advantageous to not try and combine the replications in this way. Instead, if m is much less than n_2 one could

use the usual n_2 replicate weights, $\tilde{\alpha}_i^{(k)} = \alpha_i^{(k)}$, for the set $k \in A_2$ to form $v_{J,2}$, that is, $\tilde{Y}_{reg}^{(k)} = \sum_{i\in A_2} \tilde{\alpha}_i^{(k)} y_i = \hat{\boldsymbol{X}}_1^{(k)\prime} \hat{\boldsymbol{\beta}}_{(2)}^{(k)}$, for $k \in A_2$ and form another $m < n_2$ replicate weights to capture $v_{J,1-2}$, that is, $\tilde{Y}_{reg}^{(k)} = \sum_{i\in A_2} \tilde{\alpha}_i^{(k)} y_i = c_k^{-1/2} \boldsymbol{\delta}_k' \hat{\boldsymbol{\beta}}_{(2)}$. This turns out to be particularly advantageous in some cases. In the next section, we will consider one such case, where the first-phase sample is a stratified sample and is used to form sub-strata for the second-phase sample.

3. Illustration via Nested-strata Two-phase Estimator

In this section, we illustrate the potential of the proposed method in a simple situation. Consider stratified simple random sampling, where n_h units are selected with equal probability without replacement from a population of size N_h, independently across H strata. Let y_{hi} be the value of the study variable of unit i in stratum h. Instead of observing the y_{hi}'s directly, assume that we observe $\boldsymbol{x}_{hi} = (x_{hi1}, x_{hi2}, \cdots, x_{hiG_h})$, where x_{hig} takes the value 1 if unit i in stratum h belongs to group g, and takes the value 0 otherwise. Each unit belongs to one and only one group. We call group g in stratum h sub-stratum (hg). There are $n_{hg} = \sum_{\{i:(hi)\in A_1\}} x_{hig}$ units in sub-stratum (hg).

For the second-phase sampling, we assume that $r_{hg} \geq 2$ elements are selected without replacement with equal probability independently across the sub-strata. From the selected elements, we observe y_{hig}, where the subscript g is used to emphasize that unit (hi) belongs to group g. Then, an unbiased estimator for the total of the y-variable is

$$\hat{Y}_2 = \sum_{(hi)\in A_2} \sum_{g=1}^{G_h} \frac{N_h}{n_h} \frac{n_{hg}}{r_{hg}} y_{hig}. \tag{3.1}$$

The first factor $n_h^{-1} N_h$ is the inverse of the inclusion probability for the first-phase sampling and the second factor $r_{hg}^{-1} n_{hg}$ is the inverse of the inclusion probability for the second-phase sampling. The variance of $\hat{Y}_2$ can be written as

$$\begin{aligned} \text{Var}\left(\hat{Y}_2\right) = {} & E\Big\{ \sum_{h=1}^{H} \left(\frac{N_h}{n_h}\right)^2 (1-f_{1h}) \frac{n_h}{n_h-1} \sum_{g=1}^{G_h} \Big[n_{hg} \left(\bar{y}_{hg} - \bar{y}_h\right)^2 + (n_{hg}-1) s_{hg}^2 \Big] \Big\} \\ & + E\Big\{ \sum_{h=1}^{H} \left(\frac{N_h}{n_h}\right)^2 \sum_{g=1}^{G_h} \frac{n_{hg}^2}{r_{hg}} \left(1 - \frac{r_{hg}}{n_{hg}}\right) s_{hg}^2 \Big\}, \end{aligned} \tag{3.2}$$

where $f_{1h} = N_h^{-1} n_h$ is the first phase sampling rate, $\bar{y}_{hg} = n_{hg}^{-1} \sum_{\{i:(hi)\in A_1\}} y_{hig}$ is the first-phase sample mean of sub-stratum (hg), $s_{hg}^2 = (n_{hg}-1)^{-1} \sum_{\{i:(hi)\in A_1\}} (y_{hig} - \bar{y}_{hg})^2$ is the first-phase sample variance of sub-stratum (hg), $s_h^2 = (n_h - 1)^{-1}$

$\sum_{\{i:(hi)\in A_1\}} (y_{hi} - \bar{y}_h)^2$ is the first-phase sample variance of stratum h, and $\bar{y}_h = n_h^{-1} \sum_{g=1}^{G_h} n_{hg} \bar{y}_{hg}$ is the first-phase sample mean of stratum h. If $f_{1h} \to 0$ and $0 < n_{hg}^{-1} r_{hg} < 1$ for all h and g, then (3.2) is dominated by

$$E\Big\{ \sum_{h=1}^{H} N_h^2 n_h^{-2} \sum_{g=1}^{G_h} n_{hg} (\bar{y}_{hg} - \bar{y}_h)^2 + \sum_{h=1}^{H} N_h^2 n_h^{-2} \sum_{g=1}^{G_h} r_{hg}^{-1} n_{hg}^2 s_{hg}^2 \Big\}. \tag{3.3}$$

A variance estimator can be easily derived from (3.2) by replacing $\bar{y}_{hg}$ and s_{hg}^2 by their estimates $\bar{y}_{hg2} = r_{hg}^{-1} \sum_{\{i:(hi)\in A_2\}} y_{hig}$ and $s_{hg2}^2 = (r_{hg} - 1)^{-1} \sum_{\{i:(hi)\in A_2\}} (y_{hig} - \bar{y}_{2hg})^2$, respectively. That is, ignoring the f_{1h} terms, a consistent variance estimator is

$$\hat{V} = \sum_{h=1}^{H} N_h^2 n_h^{-2} \sum_{g=1}^{G_h} n_{hg} (\bar{y}_{hg2} - \bar{y}_{h2})^2 + \sum_{h=1}^{H} N_h^2 n_h^{-2} \sum_{g=1}^{G_h} r_{hg}^{-1} n_{hg}^2 s_{hg2}^2. \tag{3.4}$$

Kim, Navarro and Fuller (2000) develop a jackknife variance estimator by successively deleting units from the entire first-phase sample and then adjusting the weights. The weights of the two-phase estimator in (3.1) are products of $w_{hi} = n_h^{-1} N_h$, the first-phase sampling weight, and $w_{hgi2} = r_{hg}^{-1} n_{hg}$, the second-phase sampling weight. The full jackknife replicate weights are given by

$$w_{hi}^{(h'i')} = \begin{cases} 0 & \text{if } h = h', i = i' \\ (n_h - 1)^{-1} n_h w_{hi} & \text{if } h = h', i \neq i' \\ w_{hi} & \text{if } h \neq h', \end{cases} \tag{3.5}$$

$$w_{hgi2}^{(h'i')} = \begin{cases} 0 & \text{if } h = h', i = i' \\ (r_{hg}-1)^{-1}(n_{hg}-1) & \text{if } h = h', i \neq i', x_{h'i'g} = 1, \text{ and } (h'i') \in A_2 \\ r_{hg}^{-1}(n_{hg} - 1) & \text{if } h = h', i \neq i', x_{h'i'g} = 1, \text{ and } (h'i') \notin A_2 \\ r_{hg}^{-1} n_{hg} & \text{otherwise.} \end{cases} \tag{3.6}$$

The full jackknife variance estimator of the form $\hat{V}_J = \sum_{(hi)\in A_1} [(n_h - 1)/n_h] (\hat{Y}_2^{(hi)} - \hat{Y}_2)^2$, where $\hat{Y}_2^{(h'i')} = \sum_{(hi)\in A_2} \sum_{g=1}^{G_h} w_{hi}^{(h'i')} w_{hgi2}^{(h'i')} y_{hig}$, is asymptotically equivalent to the variance estimator in (3.4), with total number of replicates $n = \sum_{h=1}^{H} \sum_{g=1}^{G_h} n_{hg}$ for r records. To apply the idea proposed in the previous section, note that

$$\hat{Y}_2^{(hi)} - \hat{Y}_2 = \begin{cases} \dfrac{N_h}{n_h - 1} (\bar{y}_{h2} - \bar{y}_{hg2}) + \dfrac{N_h}{n_h - 1} \dfrac{n_{hg} - 1}{r_{hg} - 1} (\bar{y}_{hg2} - y_{hig}) & \text{if } (hi) \in A_2 \\ \dfrac{N_h}{n_h - 1} (\bar{y}_{h2} - \bar{y}_{hg2}) & \text{if } (hi) \notin A_2 \end{cases}$$

for unit (hi) in group g, which makes the decomposition of the full jackknife variance estimator as in (2.7) particularly simple:

$$\begin{aligned}\hat{V}_J &= v_{J,2} + v_{J,1-2} \\ &= \sum_{(hi)\in A_2} \frac{n_h - 1}{n_h}(\hat{Y}_2^{(hi)} - \hat{Y}_2)^2 + \sum_{(hi)\in A_1\cap A_2^c} \frac{n_h - 1}{n_h}(\hat{Y}_2^{(hi)} - \hat{Y}_2)^2 \\ &\doteq \sum_{(hi)\in A_2} \frac{n_h - 1}{n_h}(\hat{Y}_2^{(hi)} - \hat{Y}_2)^2 + \sum_{h=1}^{H}\left(\frac{N_h}{n_h}\right)^2 \sum_{g=1}^{G_h}(n_{hg} - r_{hg})\left(\bar{y}_{hg2} - \bar{y}_{h2}\right)^2.\end{aligned}$$

Thus, deleting a unit which is in the first-phase sample but not in the second-phase sample does not contribute to the method's capturing of the second component of (3.4).

Using the full jackknife method directly uses r replicates to calculate $v_{J,2}$ and $n-r$ replicates to calculate $v_{J,1-2}$. Our proposed method amounts to calculating $v_{J,2}$ from the full jackknife method using the same r replicates, but calculating $v_{J,1-2}$ using a smaller number of replicates. In this simple setting, it is quite easy to create replicates for $v_{J,1-2}$. To see this, note that we can write $v_{J,1-2} = \sum_{h=1}^{H}\sum_{g=1}^{G_h} c_{hg}(\hat{Y}_2^{(hg)} - \hat{Y}_2)^2$, where $\hat{Y}_2^{(hg)} = \hat{Y}_2 + c_{hg}^{-1/2}(n_{hg} - r_{hg})^{1/2} N_h(\bar{y}_{hg2} - \bar{y}_{h2})/n_h$ for any $c_{hg} \geq (n_{hg} - r_{hg})$, where condition $c_{hg} \geq (n_{hg} - r_{hg})$ guarantees nonnegative replication weights for all records. Therefore, the total number of replicates is reduced to $r + G$, where the first r replicates are used to estimate $v_{J,2}$ and the last $G < r$ replicates are used to estimate $v_{J,1-2}$.

4. Non-nested-strata Two-phase Estimator

We now consider the case where the second-phase strata can cut across the first-phase strata. Assume that, in the first phase sample, we observe the group vector $\boldsymbol{x}_{hi} = (x_{hi1}, x_{hi2}, \cdots, x_{hiG})$, where x_{hig} takes the value 1 if unit i in stratum h belongs to group g, and takes the value 0 otherwise. The group is used to form the stratum variable for the second-phase sampling and may cut across the first-phase sampling strata. There are $m_g = \sum_{h=1}^{H}\sum_{\{i:(hi)\in A_1\}} x_{hig}$ units in group g.

For the second-phase sampling, we assume that $r_g \geq 2$ elements are selected without replacement with equal probability independently across the groups. From the selected elements, we observe y_{hig}. Then, an unbiased estimator for the total of y is $\hat{Y}_2 = \sum_{(hi)\in A_2}\sum_{g=1}^{G}(N_h/n_h)(m_g/r_g)y_{hig} \overset{(let)}{=} \sum_{(hi)\in A_2}\sum_{g=1}^{G} w_{hi}w_{hig2}\, y_{hig}$, where the first factor $w_{hi} = n_h^{-1}N_h$ is the inverse of the inclusion probability for the first-phase sampling and the second factor $r_g^{-1}m_g$ is the inverse of the inclusion probability for the second-phase sampling. Kott and Stukel (1997) called $\hat{Y}_2$ the double expansion estimator. Another commonly used estimator, termed the reweighted expansion estimator by Kott and Stukel (1997), is essentially the

regression estimator using the indicator vector for the groups as the auxiliary variable.

Now consider the full jackknife variance estimator. Let $a_{hig} = 1$ if unit (hi) in group g is selected in the second-phase sample and $a_{hig} = 0$ otherwise. According to Kim, Navarro and Fuller (2000), the full jackknife replicate weights for the second-phase sample are

$$\begin{aligned} w_{hig2}^{(h'i')} &= x_{hig} \frac{\sum_{(hi)\in A_1} w_{hi}^{(h'i')} w_{hi}^{-1} x_{hig}}{\sum_{(hi)\in A_1} w_{hi}^{(h'i')} w_{hi}^{-1} x_{hig} a_{hig}} \\ &= x_{hig} \frac{m_g - m_{h'g} + \left(m_{h'g} - x_{h'i'g}\right) \dfrac{n_{h'}}{n_{h'}-1}}{r_g - r_{h'g} + \left(r_{h'g} - x_{h'i'g} a_{h'i'g}\right) \dfrac{n_{h'}}{n_{h'}-1}}, \end{aligned}$$

where $m_{hg} = \sum_{\{i:(hi)\in A_1\}} x_{hig}$ and $r_{hg} = \sum_{\{i:(hi)\in A_2\}} x_{hig}$. Thus, if $n_{h'}$ is large for all (h', i'), we have

$$w_{hig2}^{(h'i')} \doteq \begin{cases} x_{hig}\left(r_g - 1\right)^{-1}\left(m_g - 1\right) & \text{if } a_{h'i'g} = 1 \text{ and } x_{h'i'g} = 1 \\ x_{hig} r_g^{-1}\left(m_g - 1\right) & \text{if } a_{h'i'g} = 0 \text{ and } x_{h'i'g} = 1 \\ x_{hig} r_g^{-1} m_g & \text{if } a_{h'i'g} = 0 \text{ and } x_{h'i'g} = 0, \end{cases} \tag{4.1}$$

which is exactly the same as the replicate weights in (3.6) except for the case of $(h', i') = (h, i)$. Since $w_{hi}^{(hi)} = 0$, the replicate weights in (4.1) have the same effect as the replicate weights in (3.6), and therefore, provided we have a large enough sample size in the first-phase strata, we can still apply the methods discussed in Section 3.

That is, under the assumption of $n_h^{-1}(n_h - 1) \to 1$, we have $\hat{Y}_2^{(hi)} - \hat{Y}_2 \doteq w_{hi}$ $(\bar{y}_{h2} - \bar{y}_{hg2})$ if $a_{hig} = 0$, for unit (hi) in group g, where $\bar{y}_{hg2} = r_{hg}^{-1} \sum_{\{i;(hi)\in A_2\}} y_{hig}$ and $\bar{y}_{h2} = r_h^{-1} \sum_{g=1}^{G} r_{hg} \bar{y}_{hg2}$. Thus, the total contribution of the full jackknife variance estimator for deleting the units not in the second phase sample is

$$\begin{aligned} v_{J,1-2} &= \sum_{(hi)\in A_1 \cap A_2^c} \frac{n_h - 1}{n_h} \left(\hat{Y}_2^{(hi)} - \hat{y}_2\right)^2 \\ &\doteq \sum_{h=1}^{H} \left(\frac{N_h}{n_h}\right)^2 \sum_{g=1}^{G_h} \left(m_{hg} - r_{hg}\right) \left(\bar{y}_{hg2} - \bar{y}_{h2}\right)^2, \end{aligned} \tag{4.2}$$

and a set of G replicates to estimate this component can be constructed in the same way as in Section 3. The variance estimator for REE can be constructed easily because the REE is a function of several DEEs.

5. Cluster Sampling in the First Phase

So far, we have covered the cases in which both the method of Fuller (1998) and the proposed method are equally applicable. In this section, we consider

an example where the Fuller method is hard to construct, while the proposed method is relatively easy to apply, that is, when the primary sampling unit for the first-phase sample is different from the primary sampling unit for the second-phase sample. Often clusters are selected in the first-phase sample and units are selected in the second-phase sample. Then, condition (2.10) in Fuller (1998) is not satisfied and hence we cannot directly apply his method. Furthermore, it is hard to conceptualize the unconditional variance in this situation. On the other hand, the full jackknife method still applies to this case and we can expect to reduce the number of replicates in the full jackknife method if some of the clusters selected in the first-phase sample do not have any units selected in the second-phase sample.

For simplicity of presentation, we assume that n clusters are selected with equal probability from the N clusters in the first-phase sample. Let y_{ij} be the value of the y-variable associated with element j in cluster i and M_i be the size of cluster i. From the selected first-phase sample, we observe $\mathbf{x}_{ij} = (x_{ij1}, x_{ij2}, \cdots, x_{ijG})$, where x_{ijg} takes the value 1 if element j in cluster i belongs to group g, and takes the value 0 otherwise. There are $m_g = \sum_{(ij)\in A_1} x_{ijg}$ elements in group g.

For the second-phase sample, we assume that $r_g \geq 2$ elements are selected without replacement with equal probability independently across the groups. From the selected elements, we observe y_{ijg}. An unbiased estimator for the total of the y-variable is

$$\hat{Y}_2 = \sum_{(ij)\in A_2} \sum_{g=1}^{G} \frac{N}{n} \frac{m_g}{r_g} y_{ijg} = \sum_{(ij)\in A_2} \sum_{g=1}^{G} w_i w_{ijg2} y_{ijg}, \tag{5.1}$$

where w_i is the first-phase sampling weight of cluster i and $w_{ijg2} = r_g^{-1} m_g$ is the second-phase weighting factor in group g. Note that the estimator (5.1) is both a DEE and REE because the first-phase sampling weight w_i is constant.

The full jackknife method for the first-phase sample deletes one cluster successively and increases the weight of the remaining clusters by $n/(n-1)$. The jackknife replicate weights for the second-phase are

$$w_{ijg2}^{(k)} = x_{ijg} \frac{\sum_{(ij)\in A_1} w_i^{(k)} x_{ijg}}{\sum_{(ij)\in A_2} w_i^{(k)} x_{ijg}} = x_{ijg} \frac{m_g - m_{kg}}{r_g - r_{kg}},$$

where $m_{kg} = \sum_{j=1}^{M_k} x_{kjg}$ is the number of first-phase sample elements in cluster k that belong to group g and $r_{kg} = \sum_{j;(kj)\in A_2} x_{kjg}$ is the number of second-phase sample elements in cluster k that belong to group g.

Let $A_1^a = \{k : r_{kg} > 0 \text{ for some } g\}$ and $A_1^b = \{k : r_{kg} = 0 \text{ for all } g\}$. Then, for $k \in A_1^b$, we have $\hat{Y}_2^{(k)} - \hat{Y}_2 = N(\bar{y}_2 - \tilde{y}_k^*)/(n-1)$ where $\hat{y}_k^* = \sum_{g=1}^{G} m_{kg} \bar{y}_{g2}$,

$\bar{y}_{g2} = r_g^{-1} \sum_{(ij)\in A_2} y_{ijg}$, and $\bar{y}_2 = n^{-1} \sum_{k\in A_1} \hat{y}_k^*$. Thus, for sufficiently large n, $v_{J,1-2} \doteq \sum_{k\in A_1^b} N^2 (\bar{y}_2 - \hat{y}_k^*)^2 / n^2 = \hat{\boldsymbol{\beta}}' \tilde{V}_x \hat{\boldsymbol{\beta}}$, where $\hat{\boldsymbol{\beta}} = (\bar{y}_{12}, \bar{y}_{22}, \cdots, \bar{y}_{G2})$. Thus, we can apply the tactics discussed in Section 2 to get G replicates for estimating $v_{J,1-2}$.

6. Simulation Studies

Here are some results from a limited simulation study. We consider a stratified population with two strata and two groups, where the groups cut across the strata. Each group is 50% of the population in each stratum. From an artificial population of size N=50,000, a simple random sample of size $n_h = 500$ is selected as a first-phase sample from the N_h population elements in stratum h for $h = 1, 2$, independently across the strata. We used several values of N_h's but reported only the case of $N_1 = 40,000$ and $N_2 = 10,000$ for brevity. Other simulations have similar results. A single set of study variables is generated from a normal distribution, where the parameters for a population of size $N = 50,000$ are given in Table 6.1. From the generated population, $B = 5,000$ two-phase samples are independently drawn. Instead of observing the study variable directly, the first-phase samples are re-stratified according to group and a simple random sample of size $r_g = 30$ is selected as a second-phase sample from the first-phase sample within group $g = 1, 2$, independently across the groups.

Table 6.1. Parameter set.

	Stratum	Group One		Group Two	
Stratum	Weight	Mean	Variance	Mean	Variance
1	0.8	7.0	1.0	12.0	1.0
2	0.2	12.0	1.0	17.0	1.0

Two types of point estimators, REE and DEE, are calculated and two types of variance estimators, full jackknife and the proposed new method with fewer replicates described in Section 4, are calculated separately for each of the point estimators. The full jackknife method uses 1,000 replicates because we have $n = 1,000$ first-phase sample units. The new method uses 64 replicates, where the first 60 represent the size of the second-phase sample and the other 4 represent the number of groups by strata.

The Monte Carlo result for $5,000$ samples generated using the parameters in Table 6.1 are given in Table 6.2 and Table 6.3. Table 6.2 shows the mean and variance of the two point estimators and Table 6.3 shows the relative bias (RB) and coefficient of variation (CV) of the two variance estimators. The RB of $\hat{V}$ as an estimator of the mean squared error of $\bar{y}$ is calculated by $[MSE_B(\bar{y}_I)]^{-1}[E_B(\hat{V}) - MSE_B(\bar{y}_I)]$ and the CV of $\hat{V}$ is calculated by

$[MSE_B(\bar{y}_I)]^{-1}\{\text{Var}_B(\hat{V}) + [E_B(\hat{V}) - MSE_B(\bar{y}_I)]^2\}^{1/2}$, where the subscript B denotes the distribution generated by the Monte Carlo simulation.

The following remarks can be made about Table 6.2 and Table 6.3.

Remark 1. For point estimation, the REE is significantly more efficient than the DEE for this population. The REE will be more efficient than the DEE if the study variables are more homogenous within each group. On the other hand, the REE has a slight bias for the population mean. The t-statistics for the significance of the bias are 0.65 for DEE and 3.26 for REE, respectively. The bias for REE is essentially the ratio bias and will be reduced for large second-phase sample sizes.

Remark 2. For variance estimation, the new variance estimator performs slightly better than the full jackknife variance estimator for DEE. For variance estimation of the REE variance, there is no significant difference.

Table 6.2. Mean and variance of the point estimators (5,000 samples).

Estimator	Mean	Variance
DEE	12.008	0.7591
REE	12.018	0.1522

Table 6.3. Relative bias (RB) and coefficient of variation (CV) for the variance estimators (5,000 samples).

Estimator	Method	RB (%)	CV(%)
DEE	Full jackknife	6.88	9.69
	New method	4.01	7.64
REE	Full jackknife	2.96	9.99
	New method	2.46	9.95

Acknowledgements

The research of the first author was mostly done when he was working for Westat Inc, and supported partially by a grant from U.S. Bureau of Census and by Hankuk University of Foreign Studies Research Fund of 2002. The second author was support by a grant from the Natural Science and Engineering Research Council of Canada and by Westat Inc.

References

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

Cochran, W. G. (1977). *Sampling Techniques*. 3rd edition. John Wiley and Sons, New York.

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

Kim, J. K., Navarro, A. and Fuller, W. A. (2000). Variance estimation for 2000 census coverage estimates. *Proc. ASA Section on Survey Research Methods*, 515-520.

Kott, P. S. (1990). Variance estimation when a first-phase area sample is restratified. *Survey Methodology* **16**, 99-103.

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

Lohr, S. L. (1999). *Sampling: Design and Analysis*. Duxbury Press, Pacific Grove, CA.

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.

Rao, J. N. K. and Sitter, R. R. (1997). Variance estimation under stratified two-phase sampling with applications to measurement bias. In *Survey Measurement and Process Quality: Wiley Series in Probability and Statistics* (Edited by L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz and D. Trewin), 753-768. John Wiley, New York.

Särndal, C. E., Swensson, B. and Wretman, J. (1992). *Model-Assisted Survey Sampling*, Springer-Verlag, New York.

Sitter, R. R. (1997). Variance estimation for the regression estimator in two-phase sampling. *J. Amer. Statist. Assoc.* **92**, 780-787.

Wolter, K. M. (1985). *Introduction to Variance Estimation*. Springer-Verlag, New York.

Department of Statistics, Hankuk University of Foreign Studies, Yongin, Kyungki-Do 449-791, Korea.

E-mail: kimj@hufs.ac.kr

Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, BC, V5A 1S6, Canada.

E-mail: sitter@stat.sfu.ca

(Received June 2002; accepted April 2003)