Communications for Statistical Applications and Methods
2023, Vol. 30, No. 4, 403–410

https://doi.org/10.29220/CSAM.2023.30.4.403
Print ISSN 2287-7843 / Online ISSN 2383-4757

Variance estimation of a double expanded estimator for two-phase sampling

Mingue Park[1,a]

[a]Department of Statistics, Korea University, Korea

Abstract

Two-Phase sampling, which was first introduced by Neyman (1938), has various applications in different forms. Variance estimation for two-phase sampling has been an important research topic because conventional variance estimators used in most softwares are not working. In this paper, we considered a variance estimation for two-phase sampling in which stratified two-stage cluster sampling designs are used in both phases. By defining a conditionally unbiased estimator of an approximate variance estimator, which is calculable when all elements in the first phase sample are observed, we propose an explicit form of variance estimator of the double expanded estimator for a two-phase sample. A small simulation study shows the proposed variance estimator has a negligible bias with small variance. The suggested variance estimator is also applicable to other linear estimators of the population total or mean if appropriate residuals are defined.

Keywords: double sampling, DEE, replicate variance estimation, linearization method

1. Introduction

Two-Phase sampling, also known as double sampling, was first introduced by Neyman (1938). He suggested to use two-phase sampling for stratification, which refers to a situation where the observations from the first phase sample is used to make a stratification for the second phase sampling.

Two-Phase sampling has various applications in different forms. Rao (1973) and Cochrane (1977) considered the case where the first phase sample is a simple random sample and the second phase sample is selected using the information obtained in the first phase. Breidt and Fuller (1993) suggested a regression type estimator for multiple-phase sampling that is applied to analyze a natural resource inventory data. Rao and Sitter (1995) considered a ratio estimator when a two-phase simple random sample is selected and proposed a new linearization variance estimator. Hidiroglou and Särndal (1998) studied the way of using different types of auxiliary information, which is obtained from the population or the first phase sample, to develop the calibration weights. Hidiroglou (2001) considered a situation in which the second phase sample is not nested in the first phase sample.

Variance estimation for two-phase sampling has been an important research topic because conventional variance estimators used in most softwares are not working. Unlike two-stage cluster sampling, two-phase sampling does not have invariant and independent property, which makes it more complicate to estimate the variance of estimators of population mean or total. Especially in the case of complex sampling designs, stratification and clustering are applied and are used in both phases, it

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. 2022M3J6A1063595).

[1] Corresponding author: Department of Statistics, Korea University, 145 Anam-Ro, Seoul 02841, Korea. E-mail: mpark2@korea.ac.kr

Published 31 July 2023 / journal homepage: http://csam.or.kr

is hard to derive an explicit form of the variance estimator. Therefore, many replicate variance estimators have been suggested where relatively simple element designs are used in the first and(or) second phase. Even though replicate variance estimation shows a way to define an asymptotically unbiased variance estimator, calculating replicate weights is a cumbersome process because the number of replications, which is proportional to first phase sample size, is huge in general. Rao and Shao (1992) proposed a jackknife variance estimator in the context of hot-deck imputation in which responses correspond to the second phase. Fuller (1998) proposed a replicate variance estimator for the two-phase regression estimator. Kim *et al.* (2006) proposed a consistent replicate variance estimator applicable to two-phase sampling for stratification. Kim and Yu (2011) developed a bias-adjusted replicate variance estimator that extends the result of Kim *et al.* (2006).

In this paper, we provided an explicit formula of an asymptotically unbiased variance estimator which is applicable when stratified two-stage cluster sampling is used in both phases. The variance estimator is based on the linearization and assumes a relationship between inclusion probability and selection probability. Under the multi-stage cluster sampling, we also assume that most of the variance is explained by the variability among primary sampling units.

2. Double expanded estimator

Let the parameter of interest be the population total $t_y = \sum_{k\in U} y_k$, where y_k is the value of a variable y for the k^{th} element in the population. The number of elements in the population is either known or not. Let s_a be the first phase sample that is selected using a probability sampling design $p_a(s_a)$. If every element in s_a is observed, an unbiased estimator of the population total is

$$\hat{t}_{s_a} = \sum_{k\in s_a} w_{ak} y_k,$$

and an unbiased variance estimator of $\hat{t}_{s_a}$, suggested by Horvitz and Thompson (1952), is

$$\hat{V}_{unb}\left(\hat{t}_{s_a}\right) = \sum_{k\in s_a}\sum_{l\in s_a} \frac{\Delta_{akl}}{\pi_{akl}} \frac{y_k}{\pi_{ak}} \frac{y_l}{\pi_{al}}, \tag{2.1}$$

where $w_{ak} = \pi_{ak}^{-1} = [Pr(k \in s_a)]^{-1}$ is the sampling weights defined by $p_a(s_a)$, $\Delta_{akl} = \pi_{akl} - \pi_{ak}\pi_{al}$, and $\pi_{akl} = Pr[(k \in s_a) \cap (l \in s_a)]$.

A second phase sample s is selected from the first phase sample s_a through a conditional sampling design $p(s|s_a)$ conditioning on s_a. The double expanded estimator (DEE) of the population total using a sample s is

$$\hat{t}_{DEE} = \sum_{k\in s} w_{ak} w_{k|s_a} y_k = \sum_{k\in s} w_k^* y_k, \tag{2.2}$$

where $w_k^* = w_{ak} w_{k|s_a}$ and $w_{k|s_a} = \pi_{k|s_a}^{-1} = [Pr(k \in s|s_a)]^{-1}$. By applying the following conditional variance result

$$V\left(\hat{t}_{DEE}\right) = V_{p_a}\left[E_p\left(\hat{t}_{DEE} \mid s_a\right)\right] + E_{p_a}\left[V_p\left(\hat{t}_{DEE} \mid s_a\right)\right],$$

we have the variance of $\hat{t}_{DEE}$ as shown below.

$$V\left(\hat{t}_{DEE}\right) = \sum_{k\in U}\sum_{l\in U} \Delta_{akl} \frac{y_k}{\pi_{ak}} \frac{y_l}{\pi_{al}} + E_{p_a}\left(\sum_{k\in s_a}\sum_{l\in s_a} \Delta_{kl|s_a} \frac{y_k}{\pi_k^*} \frac{y_l}{\pi_l^*}\right), \tag{2.3}$$

where $\Delta_{kl|s_a} = \pi_{kl|s_a} - \pi_{k|s_a}\pi_{l|s_a}$, $\pi_{kl|s_a} = Pr[(k \in s) \cap (l \in s)|s_a]$, $\pi_k^* = \pi_{ak}\pi_{k|s_a}$ and U is the set of elements in the population.

Särndal *et al.* (1991) provided the unbiased variance estimator when general sampling designs are used in both phases as shown below.

$$\hat{V}\left(\hat{t}_{DEE}\right) = \sum_{k \in s}\sum_{l \in s} \frac{\Delta_{akl}}{\pi_{kl}^*} \frac{y_k}{\pi_{ak}} \frac{y_l}{\pi_{al}} + \sum_{k \in s}\sum_{l \in s} \frac{\Delta_{kl|\, s_a}}{\pi_{kl|\, s_a}} \frac{y_k}{\pi_k^*} \frac{y_l}{\pi_l^*}, \tag{2.4}$$

where $\pi_{kl}^* = \pi_{akl}\pi_{kl|s_a}$.

The variance estimator given in (2.4) depends on double sums and second order inclusion probabilities, which make the calculation impossible or cumbersome. An asymptotically unbiased and applicable variance estimator corresponding to the second term of (2.4) is relatively easy to derive. However, it is difficult to approximate the first term of (2.4) due to its dependency of the conditional inclusion probability, conditioning on all possible s_a. In this paper, we provide an asymptotically equivalent and calculable formula for the first term of the variance estimator (2.4) by finding an approximation of (2.1).

3. Proposed variance estimator

For a single-phase sampling design, most survey software use the following approximately unbiased variance estimator (3.1) to avoid the calculation of double sum and second order inclusion probabilities.

$$\hat{V}\left(\sum_{k \in s} \frac{y_k}{\pi_k}\right) = \left(1 - \frac{n}{N}\right) \frac{1}{n(n-1)} \sum_{k \in s} \left(\frac{ny_k}{\pi_k} - \frac{1}{n} \sum_{l \in s} \frac{ny_l}{\pi_l}\right)^2, \tag{3.1}$$

where $\pi_k = Pr(k \in s)$, n and N are the number of primary sampling units (PSU) in the sample and population, and $\sum_{k \in s} y_k/\pi_k$ is the Horvitz-Thompson (1952) estimator of the population total. The variance estimator given in (3.1) is obtained under the assumption $\pi_k \approx np_k$, so that

$$\frac{1}{n} \sum_{k \in s} \frac{y_k}{p_k} \approx \frac{1}{n} \sum_{k \in s} \frac{ny_k}{\pi_k},$$

where p_k is a selection probability used to define an unbiased estimator when with-replacement sampling is used. Note that the variance estimator (3.1) depends on neither double sum nor second order inclusion probability. The variance estimator in (3.1) works well when the number of PSU is large enough so that the inverse of the number of PSU is close to zero and the chance of selecting a PSU more than one time in the selection process is near zero. The variance estimator (3.1) is not directly applicable for two-phase sampling because $\pi_k = Pr(k \in s) = Pr(k \in s_a) \times Pr(k \in s|k \in s_a) \neq \pi_k^*$ and the first phase sampling is not independent from the second phase sampling.

To derive an applicable approximate variance estimator of $\hat{t}_{DEE}$ under a two-phase sampling, we assume that a stratified two-stage cluster sampling is used in both phases. The result can be directly extended to multi-stage cluster sampling. In the first phase, a stratified sample of clusters is selected at the first stage and, at the second stage, samples of elements are selected from each cluster selected in the first stage. If all elements selected in the first phase are observed, an approximate variance estimator of $E(\hat{t}_{DEE}|s_a) = \hat{t}_{s_a}$, which is asymptotically unbiased to the expectation of (2.1) as given in

Woodruff (1971), is

$$\hat{V}_a\left(\hat{t}_{s_a}\right) = \sum_{h=1}^{H}\left(1 - \frac{n_{ah}}{N_{ah}}\right)\frac{n_{ah}}{n_{ah}-1}\sum_{i=1}^{n_{ah}}\left(y_{hi\cdot} - \bar{y}_{h\cdot\cdot}\right)^2, \tag{3.2}$$

where $y_{hi\cdot} = \sum_{j=1}^{m_{ahi}} w_{ahij}y_{hij}$, $y_{h\cdot\cdot} = n_{ah}^{-1}\sum_{i=1}^{n_{ah}} y_{hi\cdot}$, $w_{ahij} = \pi_{ahij}^{-1}$ and the subscript hij means the j^{th} element in cluster i of stratum h defined for selecting a first phase sample. For example, m_{ahi} and n_{ah} denote the number of elements in cluster i, it belongs to stratum h, and the number of clusters selected from stratum h.

At the second phase, strata and clusters are redefined which are either the same as the ones defined in the first phase or not. To obtain a sample s, a cluster sample of size n_g is selected from stratum g in the first stage, and a sample of elements of size m_{gk} is selected from cluster k in stratum g that is selected in the first stage. Note that the variance estimator in (3.2) is approximately unbiased to the first term of (2.3), if all elements in the first phase sample are observed. Thus, by finding a conditionally unbiased estimator of (3.2), we obtain an approximately unbiased variance estimator corresponding to the first term of (2.4).

We propose an estimator of (3.2) based on the second phase sample s as

$$\hat{V}_1 = \sum_{h=1}^{H}\left(1 - \frac{n_{ah}}{N_{ah}}\right)\frac{n_{ah}}{n_{ah}-1}\sum_{i=1}^{n_{ah}}\left(\hat{y}_{hi\cdot} - \bar{\hat{y}}_{h\cdot\cdot}\right)^2, \tag{3.3}$$

where

$$\hat{y}_{hi\cdot} = n_{ah}\frac{\sum_{g=1}^{G}\sum_{k=1}^{n_g}\sum_{l=1}^{m_{gk}} w^*_{gkl} \times y_{gkl} \times z_{gkl}(hi)}{\sum_{g=1}^{G}\sum_{k=1}^{n_g}\sum_{l=1}^{m_{gk}} w_{agkl|s_a} \times z_{gkl}(hi)},$$

and $z_{gkl}(hi)$ is one if element l in cluster k of stratum g in the second phase belongs to cluster i in stratum h defined in the first phase, and is zero otherwise. Note that $\hat{y}_{hi\cdot}$ in (3.3) is a ratio estimator of the cluster total where $\sum_{g=1}^{G}\sum_{k=1}^{n_g}\sum_{l=1}^{m_{gk}} w_{agkl|s_a} \times z_{gkl}(hi)$ is an estimator of n_{ah}. We use a ratio estimator because a conditionally unbiased estimator of n_{ah} based on s usually has large variability. Because $\hat{y}_{hi\cdot}$ is approximately unbiased to $y_{hi\cdot}$, the variance estimator (3.3) is also conditionally unbiased and consistent to $\hat{V}_a(\hat{t}_{s_a})$ in (3.2).

By applying the equation (3.1) with $y_k =: \pi_{ak}^{-1}y_k$ and $\pi_k =: \pi_{k|s_a}$, we obtain a conditionally and approximately unbiased estimator for the second term of variance estimator (2.4) as

$$\hat{V}_2 = \sum_{g=1}^{G}\left(1 - \frac{n_g}{n_{ag}}\right)\frac{n_g}{n_g-1}\sum_{k=1}^{n_g}\left(\dot{y}_{gk\cdot} - \bar{\dot{y}}_{g\cdot\cdot}\right)^2, \tag{3.4}$$

where

$$\dot{y}_{gk\cdot} = \sum_{l=1}^{m_{gk}}\frac{y_{gkl}}{\pi_{agkl}\pi_{gkl|s_a}}, \quad \bar{\dot{y}}_{g\cdot\cdot} = \frac{1}{n_g}\sum_{k=1}^{n_g}\dot{y}_{gk\cdot}.$$

Thus, the final variance estimator is obtained by adding (3.3) and (3.4) as shown below.

$$\hat{V} = \hat{V}_1 + \hat{V}_2. \tag{3.5}$$

The proposed variance estimator does not depend on double sum or second inclusion probability and is easy to use because it only requires the first order inclusion probability and the definition of z. Most single-phase complex designs are formulated as a form of stratified multi stage cluster sampling, thus variance estimator of (3.5) is directly extended to the general sampling designs in which stratification and clustering are applied in both phases. Also, the proposed variance estimator can be applied immediately in conventional survey software using the given formula.

3.1. Two-Phase sampling for stratification

Consider a two-phase sampling for stratification problem in which a simple random sample is selected in the first phase. Then the first phase sample is stratified using the information observed in the first phase. A two-phase sample is selected using a stratified simple random sampling by selecting a simple random sample from each stratum independently.

If we apply equation (3.2) to get an asymptotically unbiased variance estimator based on the first phase sample s_a, we have

$$\begin{aligned}\hat{V}_a\left(\hat{t}_{s_a}\right) &= N^2\left(1-\frac{n_a}{N}\right)\frac{1}{n_a}\frac{1}{n_a-1}\sum_{k\in s_a}\left(y_k-\bar{y}_{s_a}\right)^2\\ &= N^2\left(1-\frac{n_a}{N}\right)\frac{1}{n_a}\sum_{g=1}^{G} w_{ag}\left[S^2_{y_{s_{ag}}}+\left(\bar{y}_{s_{ag}}-\bar{y}_{s_a}\right)^2\right],\end{aligned} \tag{3.6}$$

by partitioning the sum of squares such that

$$\sum_{k\in s_a}\left(y_k-\bar{y}_{s_a}\right)^2=\sum_{g=1}^{G}\left(n_{ag}-1\right)S^2_{y_{s_{ag}}}+\sum_{g=1}^{G}n_{ag}\left(\bar{y}_{s_{ag}}-\bar{y}_{s_a}\right)^2,$$

where $\bar{y}_{s_{ag}} = n_{ag}^{-1}\sum_{k=1}^{n_{ag}} y_{gk}$, $S^2_{y_{s_{ag}}} = (n_{ag}-1)^{-1}\sum_{k=1}^{n_{ag}}(y_{gk}-\bar{y}_{s_{ag}})^2$, $w_{ag} = (n_a-1)^{-1}(n_{ag}-1)$ and n_{ag} is a number of elements in the stratum g defined in the first phase sample. By replacing $\bar{y}_{s_{ag}}$ and $S^2_{y_{ag}}$ with their estimators based on the final samples, we obtain an estimator corresponding to $\hat{V}_1$ of (3.7) which is equivalent to the one suggested by Särndal *et al.* (1991),

$$\hat{V}_1 = N^2\left(1-\frac{n_a}{N}\right)\frac{1}{n_a}\sum_{g=1}^{G} w_{ag}\left[S^2_{ys_g}+\left(\bar{y}_{s_g}-\bar{y}_{DEE}\right)^2\right], \tag{3.7}$$

where $S^2_{ys_g} = (n_g-1)^{-1}\sum_{k=1}^{n_g}(y_{gk}-\bar{y}_{s_g})^2$, $\bar{y}_{s_g} = n_g^{-1}\sum_{k=1}^{n_g} y_{gk}$, $\bar{y}_{DEE} = \sum_{g=1}^{G} w_{ag}\bar{y}_{s_g}$ and n_g is the number of elements in the stratum g of sample s.

By applying equation (3.4), we obtain the second term of the variance estimator of $\hat{t}_{DEE}$ as shown below.

$$\hat{V}_2 = N^2\sum_{g=1}^{G}\left(1-\frac{n_g}{n_{ag}}\right)w_{ag}^2\frac{S^2_{ys_g}}{n_g}. \tag{3.8}$$

4. Simulation study

To investigate the performance of the suggested variance estimator, a small simulation study was done. For a simulation study, we generate a finite population of (y, x) from the following model. For

Table 1: Mote Carlo properties of variance estimator

Scenario	Variance	Relative bias (%)	CV (%)	Coverage rate (%)
1	7,738,707,503	5.307	1.176	95.6
2	14,818,176,130	−0.390	4.662	94.8
3	9,857,814,309	3.881	4.265	95.2

$c = 1, \ldots, 10$,

$$y_{cij} = (10 + c) + \eta_{ci} + \epsilon_{cij}, \quad x_{cij} = 15 + 0.7\left(y_{cij} - 15\right) + \delta_{cij}$$
$$\eta_{ci} \sim N(0,2), \quad \epsilon_{cij} \sim N(0,1), \quad \delta_{cij} \sim N(0,1).$$

Note that y is generated independently in each stratum and each observation is nested within cluster and stratum. Also y and x are correlated in the population. The number of clusters and elements in the population are 1,977 and 397,678, respectively. The population mean of x and y are 1.54 and 1.55.

To select a two-phase sample, we considered three sampling scenarios. In the first one, we select a simple random sample of size 100,000 elements in the first phase. For the second scenario, we select a two-stage simple random sample in the first phase where 500 clusters are selected in the first stage and 50% of the elements are selected from each cluster selected in the first stage. In the final scenario, a two-stage πps cluster sample of size 500 clusters was selected using the number of elements in the cluster as a size variable in the first stage, and 100 elements were selected from each selected cluster using a simple random sampling.

In all three scenarios, a stratified simple random sample is selected at the second phase. For stratification, we use an information on x variable obtained in the first phase. We stratify the first phase sample that results in 10 strata, where the stratum boundaries are 11.96, 13.09, 13.95, 14.72, 15.44, 16.16, 16.92, 17.79 and 18.94. As an allocation method, we use a proportional allocation having a proportion 20%. The number of replications for the simulation is 5,000. No stratification of clusters in selecting the first phase sample is considered because sample selection across strata are independent and thus the results of a simulation study are directly applied to the case where more than one stratum is used.

Table 1 shows the Monte Carlo properties of the suggested variance estimator. The variance in Table 1 is the Monte Carlo variance of $\hat{t}_{DEE}$ and the relative bias and CV are given as shown below.

$$\text{Relative bias}\,(\%) = \frac{\text{Monte Carlo mean of variance estimator} - \text{Variance}}{\text{Variance}} \times 100,$$

$$\text{CV}\,(\%) = \frac{\sqrt{\text{Monte Carlo variance of variance estimator}}}{\text{Monte Carlo mean of variance estimator}} \times 100.$$

Coverage rate shows a relative frequency of confidence intervals containing the true population total among 5,000 confidence intervals having a 95% confidence level.

Properties commonly required for a variance estimator are 1) non-negativity 2) computability 3) approximately unbiased 4) small variance, and 5) a pivotal property, which means that the distribution of the normalized quantity defined from using the variance estimator does not depend on parameters. By its definition, the suggested variance estimator provides non-negative values and is also computable if we have the first order inclusion probability which is known. Also, it is known that the variance estimator given in (3.4) is asymptotically unbiased and thus its conditionally unbiased one is also asymptotically unbiased. The simulation result shows that the relative bias in the three scenarios is negligible, being less than 6%. The coefficient of variation is only about 1.2 – 4.7% and the coverage rate of the 95% confidence interval constructed using the suggested variance estimator is near the nominal one at 95%.

5. Conclusion

Variance estimation is a possible hindrance in using two-phase sampling due to its complexity. Even though replicate variance estimation is often used in practice, the calculation is cumbersome if the first-phase sample size is huge, which is common. In this paper, we provided an explicit formula for a variance estimator that satisfies the properties required for a variance estimator. Intuition of defining the variance estimator is simple, that is, finding a conditionally unbiased estimator of an approximate variance estimator that is applicable if all elements in the first-phase sample are observed.

A small simulation study shows the suggested variance estimator works reasonably well with respect to bias and variance. The variance estimator can be also used for linear estimators by defining appropriate residuals. The suggested variance estimator performs poorly if the number of elements, which belong to a first phase cluster and are selected in the second phase, is small. To reduce the variability of the variance estimator in such a case, increasing the second phase sample size or applying an appropriate adjustment coefficient to n_{ah} in (3.3) is necessary.

References

Cochran WG (1977). *Sampling Techniques* (3rd ed), Wiley, New York.

Breidt FJ and Fuller WA (1993). Regression weighting for multipurpose samplings, *Sankhyā B*, **55**, 297–309.

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

Hidiroglou MA and Särndal CE (1998). Use of auxiliary information for two phase sampling, *Survey Methodology*, **24**, 11–20.

Hidiroglou MA (2001). Double sampling, *Survey Methodology*, **27**, 143–154.

Horvitz DG and Thompson DJ (1952). A generalization of sampling without replacement from a finite universe, *Journal of the American Statistical Association*, **47**, 663–685.

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

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

Kim JK and Yu CL (2011). Replication variance estimation under two-phase sampling, *Survey Methodology*, **37**, 67–74.

Rao JNK (1973). On double sampling for stratification and analytical surveys, *Biometrika*, **60**, 125–133.

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

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

Särndal CE, Swensson B, and Wretman J (1991). *Model assisted Survey Sampling*, Springer, New York.

Woodruff RS (1971). A simple method for approximating the variance of a complicated estimate, *Journal of the American Statistical Association*, **66**, 411–414.

Received January 10, 2023; Revised February 15, 2023; Accepted February 27, 2023