Communications for Statistical Applications and Methods
2015, Vol. 22, No. 5, 435–443

DOI: http://dx.doi.org/10.5351/CSAM.2015.22.5.435
Print ISSN 2287-7843 / Online ISSN 2383-4757

A Note on Complex Two-Phase Sampling with Different Sampling Units of Each Phase

Sang Eun Lee[a], Young Jin[b], Key-Il Shin[1,c]

[a]Department of Applied Statistics and Information, Kyonggi University, Korea;
[b]Sampling Division, Statistics Korea, Korea;
[c]Department of Statistics, Hankuk University of Foreign Studies, Korea

Abstract

Two phase sampling design is useful to increase estimation efficiency using deep stratification, improved non-response adjustment and reduced coverage bias. The same sampling units are commonly used for the first and the second phases in complex two-phase sampling design. In this paper we consider a sampling scheme where the first phase sampling units are clusters and the second phase sampling units are list samples. Using selected clusters in first phase requires that we list up elements in the selected clusters from the first phase and then use the list as a secondary sampling frame for the second phase sampling design. Then we select second phase samples from the listed sampling frame. We suggest an estimator based on the complex two-phase sampling design with different sampling units of each phase. Also the estimated variances of the estimator obtained by using classic and replication variance methods are considered and compared using simulation studies. For real data analysis, 2010 Korea Farm Household Economy Survey (KFHES) and 2011 Korea Agriculture Survey (KAS) are used.

Keywords: Jackknife variance estimation, agriculture sampling, cluster sampling

1. Introduction

Two phase sampling design, also known as double sampling design, can be a cost effective method for large scale survey. For a study variable with high variability in the population, it may be expensive to obtain a reasonable parameter estimate because of the need of a large sample size. The success of two phase sampling is because it may be inexpensive to collect data on auxiliary variables well correlated with the study variable in a large first phase sample. Frequently the auxiliary variable is used for stratification.

Many studies have been devoted to the two phase sampling design. Fuller (2000) investigated the regression estimation for two phase sampling design. Hidiroglou (2001) studied the regression estimation for nest and non-nested two phase sampling design; in addition the problem of variance estimation for two phase sampling design has generated considerable interest. Kott and Stukel (1997) compared performance of the re-weight expansion estimator (REE) and double expansion estimator (DEE) using jackknife variance estimation. After that many researches such as Kim and Sitter (2003), Kim *et al.* (2006), Robotham *et al.* (2008) studied replication variance estimation for two-phase sampling. Also Kim and Yu (2011) suggested the bias correction method for the jackknife variance

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education (NRF-2014R1A1A2056857).
[1] Corresponding author: Department of Statistics, Hankuk University of Foreign Studies, Yongin 17035, Korea.
E-mail: keyshin@hufs.ac.kr

Published 30 September 2015 / journal homepage: http://csam.or.kr

estimation under two-phase sampling. Most of studies considered the same sampling units for the first and the second phases; however, Singh (2008) studied different sampling units for each phase. For the first phase he used clusters as samples and for the second phase, elements are sampled. He used an underlying assumption of conditional unbiased (CU) and conditional independent (CI).

In this paper we consider a two phase sampling design with cluster level sampling at the first phase followed by element level sampling at the second phase. An unbiased estimator is suggested and the estimated variances of the estimator obtained by using classic and replication variance methods are considered for this double sampling design.

The rest of this paper is organized as follows. In Section 2, we explain the two phase sampling set up used in this study and suggest an estimator of population mean. We also investigate the estimated variances of the suggested estimator. In Section 3, we perform limited simulation studies. In Section 4, the real data analysis is conducted using Korea Farm Household Economy Survey (KFHES). Here Korea Census of Agriculture (KCA) is used as the first phase sampling frame and Korea Agriculture Survey (KAS) is used as the second phase sampling frame. In Section 5, concluding remarks are made.

2. Proposed Method

2.1. Basic setup

In this section, we consider a two phase design with cluster level sampling at the first phase followed by element level sampling at the second phase. Here we adopt the notations and set up used by Singh (2008). Let U_1 denote the first phase finite population of clusters with total number of clusters N_1 and s_1 denote the first phase sample of size n_1, and $\{\pi_{1(i)}\}_{1\le i\le N_1}$ denote the sample inclusion probability at the first phase. Here we use SRS method for the first phase. The sample s_1 with (random) total number of elementary units N_2 is stratified into L strata at phase two based on the auxiliary variable information from s_1 where it is assumed that the stratum definition does not depend on s_1. Here $N_2 = \sum_{s_1} N_{2(i)}$ and $N_{2(i)}$ is the size of element in the i^{th} cluster selected. Now conditional on s_1, from each stratum h, $(1 \le h \le L)$, a second phase sample s_{2h} of size n_{2h} is selected by simple random sample with sample inclusion probabilities $\{\pi_{2hk}\}_{1\le k\le N_{2h}}$. Thus the total sample size of the resulting two phase sample is $n_2 = \sum_h n_{2h}$. The parameter of interest is the population mean $\bar{Y}_M$ is defined as $M^{-1}\sum_{i=1}^{N_1} T_{Y_{(i)}}$ where $T_{Y_{(i)}} = \sum_{k=1}^{N_{2(i)}} y_{k(i)}$ which also equals $\sum_{h=1}^{L}\sum_{k=1}^{N_{2h(i)}} y_{hk(i)}$, $y_{hk(i)}$ being the y observation on the k^{th} unit in the h^{th} stratum subgroup of the i^{th} cluster and M is the total number of population elements. Here $N_{2h(i)}$ is the size of element in the h^{th} stratum of the i^{th} cluster. We now use index h for stratum, i for cluster and k for element. Then the standard unbiased estimator is given by

$$\hat{\bar{Y}}_M = \frac{1}{M}\sum_h \sum_{k\in s_{2h}} \frac{y_{hk}}{\pi_{1hk}\pi_{2hk}} = \frac{1}{M}\sum_h \hat{t}_{y_{(h)}}, \tag{2.1}$$

which can alternatively be expressed in terms of estimated cluster totals as in two stage design by

$$\hat{\bar{Y}}_M = \frac{1}{M}\sum_{i\in s_1} \frac{\hat{t}_{y(i)}}{\pi_{1(i)}}, \tag{2.2}$$

where $\hat{t}_{y(i)} = \sum_h \sum_{s_{2h(i)}} y_{hk(i)}\pi_{2hk}{}^{-1}$. Note that the second phase sample size $n_{2(i)}$ for the i^{th} cluster is random. However the total sample size in the second phase $n_2 = \sum_{i\in s_1} n_{2(i)}$ is fixed under this design. Here for some i, we have $n_{2(i)} = 0$.

Also the variance of $\hat{\bar{Y}}_M$ under two phase design is given by

$$V\left(\hat{\bar{Y}}_M\right) = \frac{1}{M^2}\left[V_1\left(\sum_{s_1}\frac{T_{y_{(i)}}}{\pi_{1i}}\right) + E_1 V_2\left(\sum_h \hat{t}_{y_{(h)}}\right)\right], \tag{2.3}$$

where V_1 is the phase one component and V_2 is the phase two component of the total variance. Now under the assumption of the independence between strata, (2.3) can be expressed by

$$V(\hat{\bar{Y}}_M) = \frac{1}{M^2}\left[V_1\left(\sum_{s_1}\frac{T_{y_{(i)}}}{\pi_{1(i)}}\right) + E_1\left\{\sum_h V_2(\hat{t}_{y_{(h)}})\right\}\right]. \tag{2.4}$$

2.2. The proposed estimator and variance estimator

In this section we consider an unbiased estimator of population mean and the variance estimator. We consider two methods (standard variance estimation and replication variance estimation) for the variance estimation, especially the jackknife method.

2.2.1. An unbiased estimator

First let the i^{th} cluster be selected by simple random sample with probability, ${N_1}^{-1}n_1$ for the first phase sampling. So that we have $\{\pi_{1(i)}\}_{1\le i\le N_1} = {N_1}^{-1}n_1$. Also for the second phase sampling, we use $\{\pi_{2hk}\}_{k\in h} = {N_{2h}}^{-1}n_{2h}$ where N_{2h} and n_{2h} are the first and second phase numbers of sample sizes in stratum h respectively. We then have an unbiased estimator by plugging these weights into (2.1).

$$\hat{\bar{Y}}_M = \frac{1}{M}\sum_h^L \hat{t}_{y_{(h)}} = \frac{1}{M}\sum_h\sum_{s_{2h}}\frac{y_{hk}}{\pi_{1hk}\pi_{2hk}} = \frac{1}{M}\frac{N_1}{n_1}\sum_h^L\frac{N_{2h}}{n_{2h}}\sum_{s_{2h}} y_{hk}. \tag{2.5}$$

Also from (2.2) and $\pi_{2hk} = {N_{2h}}^{-1}n_{2h}$, we have

$$\hat{t}_{y(i)} = \sum_h \hat{t}_{yh(i)} = \sum_h \frac{N_{2h}}{n_{2h}}\sum_{s_{2h(i)}} y_{hk(i)} \tag{2.6}$$

and so we have

$$\hat{\bar{Y}}_M = \frac{1}{M}\sum_{s_1}\frac{\hat{t}_{y(i)}}{\pi_{1(i)}} = \frac{1}{M}\sum_{s_1}\frac{N_1}{n_1}\sum_h\frac{N_{2h}}{n_{2h}}\sum_{s_{2h(i)}} y_{hk(i)}. \tag{2.7}$$

2.2.2. Standard variance estimator for two phase sampling

In (2.4) we have two terms. The first term in (2.4) can be easily calculated using the one-stage cluster sampling variance with $\pi_{1(i)} = {N_1}^{-1}n_1$. That is

$$V_1\left(\sum_{s_1}\frac{T_{y_{(i)}}}{\pi_{1(i)}}\right) = \frac{{N_1}^2}{{n_1}^2}\sum_{s_1} V_1(T_{y(i)}) = {N_1}^2\frac{1-f_1}{n_1}\frac{\sum_{i=1}^{N_1}\left(T_{y(i)} - \bar{T}_{y(i)}\right)^2}{N_1 - 1}, \tag{2.8}$$

where $\bar{T}_{y(i)} = {N_1}^{-1}\sum_{i=1}^{N_1} T_{y(i)}$ and $f_1 = {N_1}^{-1}n_1$.

For the second term in (2.4), we re-arrange the summations in (2.6). Then we have

$$\hat{t}_{y(h)} = \frac{N_1}{n_1}\frac{N_{2h}}{n_{2h}}\sum_{k=1}^{n_{2h}} y_{hk}. \tag{2.9}$$

Therefore we have

$$V_2(\hat{t}_{y(h)}) = \frac{{N_1}^2}{{n_1}^2}N_{2h}^2\frac{S_{1h}^2}{n_{2h}}(1-f_{2h}), \tag{2.10}$$

where $S_{1h}^2 = (N_{2h}-1)^{-1}\sum_{k=1}^{N_{2h}}(y_{hk}-\bar{y}_{1h})^2$, $\bar{y}_{1h} = {N_{2h}}^{-1}\sum_{k=1}^{N_{2h}} y_{hk}$ and $f_{2h} = {N_{2h}}^{-1}n_{2h}$. Since N_{2h} and n_{2h} are random, we can assume that $N_{2h} = N_{2h}^c + O_P(1)$ and $n_{2h} = n_{2h}^c + O_P(1)$ where $N_{2h}^c = E(N_{2h})$ and $n_{2h}^c = E(n_{2h})$. Then we have

$$N_{2h}^2\frac{S_{1h}^2}{n_{2h}}(1-f_{2h}) = {N_{2h}^c}^2\frac{S_{1h}^2}{n_{2h}^c}\left(1-f_{2h}^c\right) + O_P\left(N_{2h}^c\right). \tag{2.11}$$

Now by taking expectation, we have

$$E_1\left[N_{2h}^2\frac{S_{1h}^2}{n_{2h}}(1-f_{2h})\right] = E_1\left[{N_{2h}^c}^2\frac{S_{1h}^2}{n_{2h}^c}\left(1-f_{2h}^c\right)\right] + O\left(N_{2h}^c\right) = {N_{2h}^c}^2\frac{S_h^2}{n_{2h}^c}\left(1-f_{2h}^c\right) + O\left(N_{2h}^c\right),$$

where $S_h^2 = (M_h-1)^{-1}\sum_{i=1}^{M_h}(y_{hk}-\bar{Y}_h)^2$, $\bar{y}_h = {M_h}^{-1}\sum_{k=1}^{M_h} y_{hk}$ and M_h is the total number of elements in the h^{th} stratum.

Therefore we have

$$E_1\left(V_2\left(\hat{t}_{y(h)}\right)\right) = \frac{N_1^2}{n_1^2}{N_{2h}^c}^2\frac{S_h^2}{n_{2h}^c}\left(1-f_{2h}^c\right) + O\left(N_{2h}^c\right). \tag{2.12}$$

Finally combining (2.8) and (2.12) we have

$$V\left(\hat{\bar{Y}}_M\right) \approx \frac{1}{M^2}\left[{N_1}^2\frac{1-f_1}{n_1}\frac{\sum_{i=1}^{N_1}\left(T_{y(i)}-\bar{T}_{y(i)}\right)^2}{N_1-1} + \sum_{h=1}^{L}\frac{N_1^2}{n_1^2}{N_{2h}^c}^2\frac{S_h^2}{n_{2h}^c}\left(1-f_{2h}^c\right)\right] \tag{2.13}$$

by ignoring $O((M^2)^{-1}N_{2h}^c)$ term. Also an estimator $\hat{V}(\hat{\bar{Y}}_M)$ of $V(\hat{\bar{Y}}_M)$ can be obtained

$$\hat{V}\left(\hat{\bar{Y}}_M\right) = \frac{1}{M^2}\left[{N_1}^2\frac{1-f_1}{n_1}\frac{\sum_{i=1}^{n_1}\left(\hat{t}_{y(i)}-\bar{\hat{\tau}}_{y(i)}\right)^2}{n_1-1} + \sum_{h=1}^{L}\frac{N_1^2}{n_1^2}N_{2h}^2\frac{s_{2h}^2}{n_{2h}}(1-f_{2h})\right], \tag{2.14}$$

where $\bar{\hat{\tau}}_{y(i)} = {n_1}^{-1}\sum_{i=1}^{n_1}\hat{t}_{y(i)}$ and $s_{2h}^2 = (n_{2h}-1)^{-1}\sum_{k=1}^{n_{2h}}(y_{hk}-\bar{y}_{2h})^2$.

2.2.3. Replication variance estimation for two phase sampling

In this section, we consider the standard replication variance estimation for REE because (2.14) is not an unbiased variance estimator. This method is studied by Kott and Stukel (1997), Kim *et al.* (2006)

Table 1: Mean and variance used for population generation

Population	Strata																			
	ST1		ST2		ST3		ST4		ST5		ST6		ST7		ST8		ST9		ST10	
	M	S	M	S	M	S	M	S	M	S	M	S	M	S	M	S	M	S	M	S
A, C	20	20	29	50	40	20	30	20	25	50	-	-	-	-	-	-	-	-	-	-
B, D	20	20	29	50	40	20	30	20	25	50	55	25	15	10	35	40	5	30	60	20

and Kim and Yu (2011). Let jackknife sample weights be $w_1 = (n_1 - 1)^{-1}N_1$, $w_{2h}^{(j)}$ where j is the index of the cluster deleted in the jackknife replicate. Then from (2.9), we have

$$t_{y(h)}^{(j)} = w_1 w_{2h}^{(j)} t_{2h}^{(j)}, \tag{2.15}$$

where

$$w_{2h}^{(j)} = \begin{cases} \dfrac{N_{2h} - N_{2h(j)}}{n_{2h} - n_{2h(j)}}, & \text{if } j \in s_1 \text{ and } j \in s_2, \\ \dfrac{N_{2h} - N_{2h(j)}}{n_{2h}}, & \text{if } j \in s_1 \text{ and } j \notin s_2, \\ \dfrac{N_{2h}}{n_{2h}}, & \text{if } j \notin s_1 \text{ and } j \notin s_2, \end{cases} \quad \text{and} \quad t_{2h}^{(j)} = \begin{cases} \displaystyle\sum_{k=1}^{n_{2h}-n_{2h(j)}} y_{hk}, & \text{if } j \in s_2, \\ \displaystyle\sum_{k=1}^{n_{2h}} y_{hk}, & \text{if } j \notin s_2. \end{cases}$$

Let the jackknife replicate be $\hat{\bar{Y}}_M^{(j)} = M^{-1} \sum_h^L t_{y(h)}^{(j)}$. Then the full jackknife variance estimator is defined by

$$\hat{JV} = \frac{n_1 - 1}{n_1}(1 - f_1) \sum_{j=1}^{n_1} \left(\hat{\bar{Y}}_M^{(j)} - \overline{\hat{\bar{Y}}_M^{(j)}} \right)^2, \tag{2.16}$$

where $f_1 = {N_1}^{-1} n_1$.

3. Simulation Study

Small simulation studies are conducted for the comparison of the variance estimates calculated by (2.14) and the jackknife variance estimates (2.16). For the population, we generate the population data as follows. First we have four populations: population A and C have 5 strata and population B and D have 10 strata. Each stratum has different mean μ_j and variance σ_j^2. Using the given mean μ_j and variance σ_j^2, we generated random numbers from $N(\mu_j, \sigma_j^2)$. Table 1 shows the μ_j and σ_j^2 for populations. In this table M means the average and S stands for standard deviation.

For population A, we have 5 strata and the cluster sizes are exactly 30 and about 30±5. For population B, we have 10 strata and the cluster sizes are exactly 30 and about 30±5. Also for population C, we have 5 strata and the cluster sizes are exactly 60 and about 60±5. For population D, we have 10 strata and the cluster sizes are exactly 60 and about 60±5.

For each population, we randomly select n_1 clusters for first phase samples and make a sampling frame for the second phase sampling. Based on this sampling frame, we select n_2 samples proportional to the each stratum size. Then finally we calculated (2.14) and (2.16).

For the comparison statistics, we consider mean squared error (MSE), absolute relative error (ARE) and bias for the check of precision of the suggested estimator of (2.7). Also MVE, the mean of variance estimates from (2.14), and the MJVE, the mean of jackknife variance estimates are calculated and the MDP-VE, the mean difference percentage of two variance estimates are obtained for

Table 2: Results of comparison for population A with cluster size 30 with 5 strata

cluster size	n_1	n_2	MSE	Bias	ARE	MVE	MJVE	MDP-VE
30	700	5000	0.8868	0.0143	2.10	0.9068	0.8846	4.08
		7000	0.8233	0.0298	2.01	0.8704	0.8562	3.00
		10000	0.7922	0.0243	1.97	0.8431	0.8364	2.13
	1000	5000	0.5991	−0.0051	1.72	0.6124	0.5818	5.61
		7000	0.5401	−0.0111	1.60	0.5758	0.5562	4.11
		10000	0.5209	−0.0115	1.58	0.5481	0.5355	3.05
30±5	700	5000	0.9388	−0.0707	2.12	0.9404	0.9203	3.95
		7000	0.9280	−0.0619	2.12	0.9032	0.8910	2.88
		10000	0.9095	−0.0444	2.13	0.8754	0.8679	2.14
	1000	5000	0.6230	0.0129	1.75	0.6226	0.5899	5.65
		7000	0.5935	−0.0027	1.69	0.5856	0.5638	4.50
		10000	0.5852	0.0155	1.68	0.5580	0.5454	3.05

MSE = mean squared error; ARE = absolute relative error; MVE = mean of variance estimate;
MJVE = mean of jackknife variance estimate; MDP-VE = mean difference percentage of two variance estimate.

direct comparison of two variance estimates.

$$\text{MSE} = \frac{1}{R}\sum_{r=1}^{R}\left(\hat{\bar{Y}}_M^{(r)} - \bar{Y}_M\right)^2,$$

$$\text{Bias} = \frac{1}{R}\sum_{r=1}^{R}\left(\bar{Y}_M - \hat{\bar{Y}}_M^{(r)}\right),$$

$$\text{ARE} = \frac{1}{R}\sum_{r=1}^{R}\left|\frac{\bar{Y}_M - \hat{\bar{Y}}_M^{(r)}}{\bar{Y}_M}\right| \times 100,$$

$$\text{MVE} = \frac{1}{R}\sum_{r=1}^{R}\hat{V}\left(\hat{\bar{Y}}_M\right),$$

$$\text{MJVE} = \frac{1}{R}\sum_{r=1}^{R}JV\left(\hat{\bar{Y}}_M\right),$$

$$\text{MDP-VE} = \frac{1}{R}\sum_{r=1}^{R}\left|\frac{\hat{V}\left(\hat{\bar{Y}}_M\right) - JV\left(\hat{\bar{Y}}_M\right)}{\hat{V}\left(\hat{\bar{Y}}_M\right)}\right| \times 100.$$

As you can see the results from the Tables 2–5, MSE, MVE, the average of estimated variance and MJVE, the average of jackknife variance, are almost the same. Therefore we can use the variance estimator developed in this study or the jackknife variance estimator even though the developed variance estimator is not unbiased. We can also confirm that the estimator developed in this paper is an unbiased estimator.

4. Real Data Analysis

4.1. Data description

For real data analysis, Korea Census of Agriculture (KCA) and Korea Agriculture Survey (KAS) are used for the first and second phase sampling frame respectively. KCA is the census and is performed in every 5 years and mostly used as sampling frame for other sample survey. From KCA, each enumerated district (ED) which includes about 30 farm households, is the sampling unit for the first phase

Table 3: Results of comparison for population B with cluster size 30 with 10 strata

cluster size	n_1	n_2	MSE	Bias	ARE	MVE	MJVE	MDP-VE
30	700	5000	0.8832	−0.0551	1.95	0.9488	0.9321	3.10
		7000	0.9171	−0.0309	2.00	0.9152	0.9032	2.55
		10000	0.8717	−0.0341	1.95	0.8901	0.8829	1.80
	1000	5000	0.5808	−0.0290	1.57	0.6311	0.6078	4.28
		7000	0.5850	−0.0336	1.59	0.5975	0.5794	3.55
		10000	0.5394	−0.0344	1.54	0.5723	0.5610	2.54
30±5	700	5000	0.9243	−0.0010	1.96	0.9645	0.9432	3.22
		7000	0.9292	−0.0385	1.95	0.9313	0.9207	2.42
		10000	0.8715	−0.0239	1.90	0.9063	0.8996	1.86
	1000	5000	0.7022	0.0148	1.70	0.6394	0.6146	4.32
		7000	0.6536	0.0251	1.66	0.6058	0.5856	3.74
		10000	0.6534	0.0151	1.68	0.5808	0.5697	2.41

MSE = mean squared error; ARE = absolute relative error; MVE = mean of variance estimate;
MJVE = mean of jackknife variance estimate; MDP-VE = mean difference percentage of two variance estimate.

Table 4: Results of comparison for population C with cluster size 60 with 5 strata

cluster size	n_1	n_2	MSE	Bias	ARE	MVE	MJVE	MDP-VE
60	500	5000	1.1658	0.0248	2.39	1.1168	1.0797	4.47
		7000	1.1781	−0.0087	2.40	1.0801	1.0593	3.27
		10000	1.1194	0.0047	2.34	1.0524	1.0401	2.49
	700	5000	0.6860	−0.0005	1.76	0.7103	0.6629	6.93
		7000	0.6725	−0.0123	1.78	0.6734	0.6414	4.98
		10000	0.6458	−0.0067	1.73	0.6456	0.6239	3.78
60±5	500	5000	1.1789	−0.0162	2.39	1.1270	1.0939	4.22
		7000	1.1214	−0.0433	2.36	1.0902	1.0702	3.33
		10000	1.1232	−0.0273	2.34	1.0623	1.0462	2.70
	700	5000	0.6771	−0.0246	1.79	0.7142	0.6649	7.32
		7000	0.6476	−0.0189	1.77	0.6774	0.6447	5.18
		10000	0.6566	−0.0280	1.79	0.6498	0.6297	3.72

MSE = mean squared error; ARE = absolute relative error; MVE = mean of variance estimate;
MJVE = mean of jackknife variance estimate; MDP-VE = mean difference percentage of two variance estimate.

Table 5: Results of comparison for population D with cluster size 60 with 10 strata

cluster size	n_1	n_2	MSE	Bias	ARE	MVE	MJVE	MDP-VE
60	500	5000	1.1034	−0.0333	2.16	1.1690	1.1403	3.67
		7000	1.0725	−0.0370	2.10	1.1353	1.1149	2.87
		10000	1.0638	−0.0323	2.08	1.1101	1.0976	2.20
	700	5000	0.7341	−0.0201	1.75	0.7338	0.6911	6.05
		7000	0.6748	−0.0195	1.70	0.7000	0.6700	4.53
		10000	0.6469	−0.0067	1.67	0.6749	0.6553	3.19
60±5	500	5000	1.0861	−0.0019	2.10	1.1500	1.1221	3.47
		7000	1.0951	0.0100	2.12	1.1171	1.0983	2.71
		10000	1.0110	0.0056	2.03	1.0922	1.0798	2.03
	700	5000	0.6486	0.0286	1.66	0.7243	0.6814	6.10
		7000	0.5973	0.0384	1.57	0.6911	0.6614	4.51
		10000	0.6156	0.0440	1.57	0.6664	0.6476	3.15

MSE = mean squared error; ARE = absolute relative error; MVE = mean of variance estimate;
MJVE = mean of jackknife variance estimate; MDP-VE = mean difference percentage of two variance estimate.

sampling. KAS is a yearly sample survey with samples from KCA. With the results of KAS, we have the list of household sampling frame for the second phase. Korea Farm Household Economic Survey (KFHES) with about 2,600 farm households samples from that list sampling frame are conducted.

Table 6: SE and RSE using proposed and jackknife methods

Variables	Variance Estimators			
	Proposed		Jackknife	
	S.E.	RSE	S.E.	RSE
1. Farm Household Income	1,119,232	3.34	949,831	2.84
2. Agriculture Income	601,736	5.07	788,195	6.65
3. Gross Farm Income	1,529,599	4.62	1,853,044	5.61
4. Assets	12,881,457	3.39	12,098,540	3.19

Stratified cluster sampling is used for the first phase. That means we stratified the population by region and then in each region, cluster samples which are constructed with about 30 farm-households each, are obtained. For the second phase, these farm households from the clustered sampling units are restructured by list sampling frame and stratified by farming types. Then final samples of farm households are selected and surveyed for incomes and assets using a stratified systematic sampling method.

4.2. Estimators for real data analysis

For real data analysis, we use the following estimator suggested in this paper

$$\hat{\bar{Y}}_M = \frac{1}{M}\sum_{h}^{L} t_{y(h)} \equiv \frac{1}{M}\frac{N_1}{n_1}\sum_{h}^{L}\frac{N_{2h}}{n_{2h}}\sum_{s_{2h}} y_{hk}$$

and the variance estimator $\hat{V}(\hat{\bar{Y}}_M)$ calculated by

$$\hat{V}\left(\hat{\bar{Y}}_M\right) = \frac{1}{M^2}\left[{N_1}^2\frac{1-f_1}{n_1}\frac{\sum_{i=1}^{n_1}\left(t_{y(i)}-\bar{t}_{y(i)}\right)^2}{n_1-1} + \sum_{h=1}^{L}\frac{N_1^2}{n_1^2}N_{2h}^2\frac{s_{2h}^2}{n_{2h}}(1-f_{2h})\right],$$

where M is 1,185,550, the total number of households and N_1 is 72,907, the total number of clusters, n_1= 3,375, the number of sample clusters in phase one and n_2 = 2,646, the number of samples in phase two and all other notations are the same as in (2.14). We finally use the jackknife variance estimator in (2.16) defined by

$$\hat{JV} = \frac{n_1-1}{n_1}(1-f_1)\sum_{j=1}^{n_1}\left(\hat{\bar{Y}}_M^{(j)} - \overline{\hat{\bar{Y}}_M^{(j)}}\right)^2.$$

4.3. The results of real data analysis

For this analysis, we consider 4 study variables: farm household income, agriculture income, gross farm income, assets. Then we calculate proposed variance and jackknife variance. We also calculate the relative standard error (RSE) for each variance estimation defined by the following.

$$\text{RSE} = \frac{\hat{\sigma}_{\hat{\bar{Y}}}}{\hat{\bar{Y}}}.$$

The results of SE and RSE are presented in Table 6.

From Table 6 the proposed and jackknife variances are similar on RSE; in addition, also the values of RSEsare all less than 10%. That means those results are announceable as an official statistics for national values.

5. Conclusion

This study suggests an estimator for complex two-phase sampling scheme that has different sampling units in each phase; subsequently, proposed and jackknife variance estimates of the estimator are suggested and evaluated. As a result, we confirm that the estimator suggested in this study is unbiased and through the real data analysis as well as prove that this estimator can improve the precision of the estimates of KFHES. The two variance estimates are almost the same even though the developed variance estimator is not unbiased; however, we suggest the jackknife variance estimate as an official statistic since the proposed variance estimator is complicated and biased. The jackknife method gives the unbiased estimate and is easy to calculate.

References

Fuller, W. A. (2000). Two-phase sampling, In *Proceedings of the Survey Methods Section: SSC Annual Meeting*, Ottawa, Canada, 23–30.

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

Kim, J. K., Navapro, A. and Fuller, W. A. (2006). Replication variance estimation for two-phase stratified sampling, *Journal of the American Statistical Association*, **101**, 312–320.

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

Kim, J. K. and Yu, C. L. (2011). Replication variance estimation under two-phase sampling, *Survey Methodology*, **37**, 67–74.

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

Robotham, H., Young, Z. I. and Saavedra-Nievas, J. C. (2008). Jackknife method for estimating the variance of the age composition using two-phase sampling with an application to commercial catches of swordfish (Xiphias gladius), *Fisheries Research*, **93**, 135–139.

Singh, A. C. (2008). Single phase simplified variance estimation approach to two phase-stage hybrid designs, In *Proceedings of the Joint Statistical Meetings: Section on Survey Research Methods*, Denver, CO, 2501–2508.

Received April 2, 2015; Revised June 26, 2015; Accepted August 12, 2015