Communications for Statistical Applications and Methods
2016, Vol. 23, No. 1, 47–55

http://dx.doi.org/10.5351/CSAM.2016.23.1.047
Print ISSN 2287-7843 / Online ISSN 2383-4757

A composite estimator for stratified two stage cluster sampling

Sang Eun Lee[a], Pu Reum Lee[b], Key-Il Shin[1,b]

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

Abstract

Stratified cluster sampling has been widely used for effective parameter estimations due to reductions in time and cost. The probability proportional to size (PPS) sampling method is used when the number of cluster element are significantly different. However, simple random sampling (SRS) is commonly used for simplicity if the number of cluster elements are almost the same. Also it is known that the ratio estimator produces a good performance when the total number of population elements is known. However, the two stage cluster estimator should be used if the total number of elements in population is neither known nor accurate. In this study we suggest a composite estimator by combining the ratio estimator and the two stage cluster estimator to obtain a better estimate under a certain population circumstance. Simulation studies are conducted to compare the superiority of the suggested estimator with two other estimators.

Keywords: post weight adjustment, jackknife method, linear combination, ratio estimator, enumerated district

1. Introduction

Stratified two stage cluster sampling is widely used to reduce time and cost for the effective estimation of total and variance. However, when reducing the cost, the accuracy of estimates will not be good if a small number of clusters are selected and many units in each selected cluster are sampled. Furthermore, the accuracy of estimates deteriorates when cluster sizes are very different and simple random sampling method is used for cluster selection. For instance, in a household survey, if we set the cluster with district named Dong, Yup, Myun and select a small number of clusters using a SRS method, the accuracy of the estimates is not guaranteed due to the differences in cluster sizes. A common solution for this issue is to use a PPS method with known cluster sizes and the total number of population elements. However, a stratified two stage cluster sampling method with SRS is frequently adopted if the differences of sizes of each cluster are small. For instance, in household survey, census is mostly used as a sampling frame which is formed by clusters called as enumerated district (ED) with including about sixty households and the stratified two stage cluster sampling method with SRS is commonly used.

However, the stratified two stage cluster sampling with SRS is not an appropriate method if we select a rather small number of clusters with different sizes because there may exist a difference between the total sum of design weights and the total number of population elements. Therefore,

[1] Corresponding author: Department of Statistics, Hankuk University of Foreign Studies, 81 Oedae-ro, Mohyeon-myeon, Cheoin-gu, Yongin-si, Gyeonggi-do 17035, Korea. E-mail: keyshin@hufs.ac.kr

Published 31 January 2016 / journal homepage: http://csam.or.kr

we need to eliminate the difference for more accurate estimates. For that adjustment, the size of the population needs to be known and the calibration adjustment can be applied. Here the obtained estimator using the calibration adjustment is known as the ratio estimator. For this reason, the ratio estimator which utilizes the exact number of population elements has better results than the other estimator obtained by a usual two stage cluster estimation.

However, getting the exact size of population is very difficult in practice. For instance, Census on Establishment in Korea is commonly used as a sampling frame. However, the information from that census was generally obtained two years prior. Hence, we suggest a composite estimator when the cluster sizes are not exact but with minor differences and a simple random sampling used to select clusters. A composite estimator is obtained by combining the two stage cluster estimator and the ratio estimator. To calculate the composite estimator, two weights on each estimator need to be calculated and obtained by one of the popular methods in Rao (2003). The variance estimate is known for the two stage cluster estimation method. The different clusters sizes then indicate an approximate variance estimator of the ratio estimator as illustrated in Cochran (1977).

This study used a Jackknife method (one of the popular replication variance estimation methods) for the ratio estimator since the variance estimate of the ratio estimator is approximately obtained. The weights which are the coefficients of the composite estimator are calculated using variances estimated in each estimation method.

In this paper, the two stage cluster sampling method will be explicitly mentioned and the composite estimator, which combines the ratio and the two stage cluster estimator as stated in Section 2. Also calculating the coefficients of the composite estimator (called as weights) is explained, especially the delete one cluster Jackknife variance estimation method for the ratio estimator is focused in Section 3. In Section 4, some simulation results are compared and a real data analysis is conducted using Taxi Company data in Section 5. The summary and conclusion are stated in Section 6.

2. Two stage cluster sampling

2.1. Stratified two stage cluster sampling

Stratified two stage cluster sampling method is a sampling technique to obtain an efficient estimation by selecting a part of elements in selected clusters. Here are the notations in this study. L is the number of strata in population, N_h is the number of clusters of population and n_h is the number of sample clusters in stratum h, M_{hi} is the number of elements in i^{th} cluster, h^{th} stratum in population and m_{hi} is the size of sample units in i^{th} sampled cluster, h^{th} stratum. Therefore the total number of clusters in population is $N = \sum_{h=1}^{L} N_h$, the total number of elements of population is $M_0 = \sum_{h=1}^{L} M_h$ and the total number of elements of population in stratum h is $M_h = \sum_{i=1}^{N_h} M_{hi}$. Following estimators are well explained in Cochran (1977).

2.1.1. Stratified two stage cluster estimation

In stratified two stage cluster sampling with L strata, the estimator of total, $\hat{Y}^C$ is defined by

$$\hat{Y}^C = \sum_{h=1}^{L} \frac{N_h}{n_h} \sum_{i=1}^{n_h} \frac{M_{hi}}{m_{hi}} \sum_{j=1}^{m_{hi}} y_{hij}, \tag{2.1}$$

where y_{hij} is the j^{th} observation in i^{th} cluster, stratum h. Now the unbiased variance estimate of the estimator, $\hat{Y}^C$ is as follows.

$$\hat{V}\left(\hat{Y}^C\right) = \sum_{h=1}^{L}\left[\frac{N_h^2}{n_h}(1-f_{1h})\frac{\sum_{i=1}^{n_h}\left(\hat{Y}_{hi}-\bar{\hat{Y}}_h^C\right)^2}{n_h-1} + \frac{N_h}{n_h}\sum_{i=1}^{n_h} M_{hi}^2(1-f_{2hi})\frac{s_{2hi}^2}{m_{hi}}\right], \tag{2.2}$$

where $\hat{Y}_{hi} = M_{hi}\bar{y}_{hi}$, $\bar{y}_{hi} = m_{hi}^{-1}\sum_{j=1}^{m_{hi}} y_{hij}$, $\bar{\hat{Y}}_h^C = {n_h}^{-1}\sum_{i=1}^{n_h} M_{hi}\bar{y_{hi}}$, $f_{1h} = n_h/N_h$, $f_{2hi} = m_{hi}/M_{hi}$ and ${s_{2hi}}^2 = \sum_{j=1}^{m_{hi}}(y_{hij}-\bar{y}_{hi})^2/(m_{hi}-1)$, the variance estimator for i^{th} cluster in h^{th} stratum.

2.1.2. Ratio estimator of stratified two-stage cluster sampling

The ratio estimator is known as one of the calibration estimates in stratified two stage cluster sampling with L strata. We assume that the number of elements in each population stratum, M_h is known. Then the estimator of total, $\hat{Y}^R$ is:

$$\hat{Y}^R = \sum_{h=1}^{L}\frac{M_h}{\sum_{i=1}^{n_h} M_{hi}}\sum_{i=1}^{n_h}\frac{M_{hi}}{m_{hi}}\sum_{j=1}^{m_{hi}} y_{hij}. \tag{2.3}$$

We use the additional information of the number of elements in stratum h, M_h and may have more accurate estimate of total than that of (2.1). The usual variance estimator of $\hat{Y}^R$ is:

$$\hat{V}\left(\hat{Y}^R\right) = \sum_{h=1}^{L}\left[\frac{N_h^2}{n_h}\frac{\sum_{i=1}^{n_h} M_{hi}^2\left(\bar{y_{hi}}-\bar{\hat{y}}_h\right)^2}{n_{h-1}}(1-f_{1h}) + \frac{N_h}{n_h}\sum_{i=1}^{n_h} M_{hi}^2(1-f_{2hi})\frac{{s_{2hi}}^2}{m_{hi}}\right]. \tag{2.4}$$

Here $\bar{\hat{y}}_h = {n_h}^{-1}\sum_{i=1}^{n_h}\bar{y_{hi}}$ and all other definitions are the same as in (2.2).

2.2. Comparison of the estimators

Now the sum of the design weight $\hat{w}$ can be obtained by plugging 1 into Equation (2.1) instead of y_{hij}. That is $\hat{w} = \sum_{h=1}^{L} N_h/n_h \sum_{i=1}^{n_h} M_{hi}$.

If the size of population in stratum h, M_h is known, then the ratio estimator can be used. If not, or inaccurate information, the value of M_h must be estimated. The estimate of the size of population in stratum h is as follows:

$$\hat{M}_h = \frac{N_h}{n_h}\sum_{i=1}^{n_h} M_{hi} = N_h\hat{\bar{M}}_h. \tag{2.5}$$

Therefore if there are some problems of estimating the mean of size of each cluster, obviously the estimate of total should be worse. That is, if N_h/n_h and $M_h/\sum_{i=1}^{n_h} M_{hi} = \sum_{i=1}^{N_n} M_{hi}/\sum_{i=1}^{n_n} M_{hi}$ are identical or at least similar, then the estimated value using (2.1) can be used.

Now assume that M_{hi} is random and $M_{hi} \sim (\bar{M}_h, \sigma_h^2)$ with $\bar{M}_h$, mean of the number of elements of cluster and σ_h^2, the variance of M_{hi}. Then $\hat{\bar{M}}_h = {n_h}^{-1}\sum_{i=1}^{n_h} M_{hi} \sim (\bar{M}_h, \sigma_h^2/n_h)$. Therefore if small number of sample clusters is used or σ_h^2 is large, then the accuracy of estimate of total obtained by the two stage cluster estimator becomes worse.

3. Suggested composite estimator

3.1. The composite estimator

Practically the number of elements in each cluster is hardly the same. Of course, EDs in census have about 60 households and therefore $\sigma_h^2 \approx 0$ and $\hat{M}_h$ can be estimated close to the true value.

However, for instance, EDs in Agriculture census have the different cluster sizes. In addition, the number of sample clusters is rather small in most of cases. Census cannot be used as sampling frame if the survey is not for official statistics. In those cases, the accuracy of the two stage cluster estimator can be declined.

To overcome this situation, we suggest a new composite estimator, $\hat{Y}_{CP}$, which combines two estimators $\hat{Y}_C$ and $\hat{Y}_R$.

The suggested linear composite estimator is defined by

$$\hat{Y}^{CP} = \alpha\hat{Y}^C + (1 - \alpha)\hat{Y}^R, \tag{3.1}$$

where α is the weight of the suggested composite estimator.

Now the estimated value of α can be obtained by

$$\hat{\alpha} = \frac{\widehat{\text{MSE}}\left(\hat{Y}^R\right)}{\widehat{\text{MSE}}\left(\hat{Y}^R\right) + \widehat{\text{MSE}}\left(\hat{Y}^C\right)} \approx \frac{\hat{V}\left(\hat{Y}^R\right)}{\hat{V}\left(\hat{Y}^R\right) + \hat{V}\left(\hat{Y}^C\right)}. \tag{3.2}$$

For more details of the estimated value of the weight α, see Rao (2003) or Hwang and Shin (2013). Here $\hat{V}(\hat{Y}^C)$ and $\hat{V}(\hat{Y}^R)$ can be calculated using (2.2) and (2.4). However, we use the replication variance estimation, especially the Jackknife variance estimator since the variance estimator (2.4) is not unbiased.

3.2. Jackknife variance estimator

From Cochran (1977), we have that the variance estimate of the ratio estimator in stratified two stage cluster sampling is obtained approximately. So that in this study we use the delete one cluster Jackknife variance estimation method. This method is also used in Lee *et al.* (2015). The Jackknife variance method is a commonly used non-parametric methods that can reduce bias. Quenouille (1949) suggested at first and Tukey (1958) used this for variance estimation. More details are found in Wolter (1985).

The general set up for the Jackknife variance method is as follows. First using the size of n samples, $Y_1, Y_2, \ldots, Y_n$, we can calculate $\hat{\theta} = f(Y_1, Y_2, \ldots, Y_n)$, an estimator of parameter θ. Then similarly after deleting the k^{th} element, the estimator deleted one element can be obtained as below.

$$\hat{\theta}_{n(k)} = f(Y_1, Y_2, \ldots, Y_{k-1}, Y_{k+1}, \ldots, Y_n). \tag{3.3}$$

The Jackknife variance estimator is:

$$V\left(\hat{\theta}_{JK}\right) = \frac{(n-1)}{n}\sum_{i=1}^{n}\left(\hat{\theta}_{n(k)} - \bar{\hat{\theta}}_{n(k)}\right)^2, \tag{3.4}$$

where $\bar{\hat{\theta}}_{n(k)}$ is the mean of $\hat{\theta}_{n(k)}$.

Now the delete one cluster Jackknife method which is used in this study is that deleting k^{th} cluster is used instead of deleting one element. Using delete one cluster Jackknife variance method, the variance estimator of the ratio estimate is as follows.

First from (2.3), the estimator of total of stratum h is $\hat{Y}_h^R = (M_h/\sum_{i=1}^{n_h} M_{hi}) \sum_{i=1}^{n_h} M_{hi}/m_{hi} \sum_{j=1}^{m_{hi}} y_{hij}$. After deleting k^{th} cluster, the estimator of total of stratum h is

$$\hat{Y}_{h(k)}^{JK} = \frac{M_h}{\sum_{i=1}^{n_{h(k)}} M_{hi}} \sum_{i=1}^{n_{h(k)}} \frac{M_{hi}}{m_{hi}} \sum_{j=1}^{m_{hi}} y_{hij}, \tag{3.5}$$

where

$$n_{h(k)} = \begin{cases} n_{h(k)} = n_h - 1, & \text{if } k \in \text{clusters in stratum } h, \\ n_{h(k)} = n_h, & \text{otherwise.} \end{cases}$$

Now the estimator of total is $\hat{Y}_{(k)}^{JK} = \sum_{h=1}^{L} \hat{Y}_h(k)^{JK}$ and from (3.5) the Jackknife variance estimator is:

$$V\left(\hat{Y}^{JK}\right) = \frac{(n-1)}{n} \sum_{k=1}^{n} \left(\hat{Y}_{(k)}^{JK} - \bar{\hat{Y}}_{(k)}^{JK}\right)^2, \tag{3.6}$$

where $n = \sum_{h=1}^{L} n_h$ is the total number of sample clusters and $\bar{\hat{Y}}_{(k)}^{JK}$ is the mean of $\hat{Y}_{(k)}^{JK}$.

4. A simulation study

In this section, the composite estimator which is the linear combination of the two stage cluster and the ratio estimators is compared with the other two estimators to improve the estimate accuracy for the stratified two stage cluster simple random sampling with different cluster sizes. For simplicity, we assume the number of strata is $L = 1$.

For the simulation study, the number of clusters in population is $N = N_1 = 500$, the number of sample clusters is $n = n_1 = 20, 40$ and the number of sample elements in cluster is $m_{hi} = m_{1i} = 5, 10$. First the normal data are generated with difference size of cluster unit, $M_{hi} = M_{1i}$. Also for existing variation of means between clusters, we consider different mean m_{Y_i} in normal distribution. Lastly, for considering changes of population size M_1 at the time of survey period, we generate values of M_1 from normal distribution with mean $M_{true} = \sum_{i=1}^{N} M_{1i}$ and variance $\sigma_{M_1}^2$. Following are the simulation set up and the values used for simulation.

(1) $M_{1i} \overset{iid}{\sim} N\left(\bar{M}_1, \sigma_{\bar{M}_1}^2\right), \quad \left(\bar{M}_1, \sigma_{\bar{M}_1}\right) = (30, 4), (60, 4), (60, 8).$

(2) $M_1 \overset{iid}{\sim} N\left(M_{true},\ \sigma_{M_1}^2\right), \quad \text{CV}_1 = \sigma_{M_1}/M_{true} = 0, 0.01, 0.02, 0.03.$

(3) $m_{Y_i} \overset{iid}{\sim} N\left(\mu_Y, \sigma_{mY}^2\right), \quad (\mu_Y,\ \sigma_{mY}) = (200, 24).$

(4) $y_{1ij} \overset{iid}{\sim} N\left(m_{Y_i}, \sigma_Y^2\right), \quad \sigma_Y = 10.$

We also use root mean squared error (RMSE), Bias, absolute bias (ABias) for the comparison statistics defined by

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

Table 1: Comparison results with $(\bar{M}_1, \sigma_{\bar{M}_1}) = (30, 4)$

n	m_{h_i}	CV_1	Bias			Abias			RMSE		
		(%)	ratio	cluster	com	ratio	cluster	com	ratio	cluster	com
20	5	0	−5653	−7574	−6897	61365	101170	76930	77050	125924	96350
		1	−312	−6488	−2963	67699	102045	79506	86071	127995	99599
		2	2441	1268	1653	77656	91127	77482	98211	112349	96215
		3	−5318	−305	−1894	95540	84619	78215	120844	106776	98430
	10	0	435	3219	1734	63175	86358	71569	79513	109888	91192
		1	−4181	−484	−2323	68463	89846	74357	85076	111740	93145
		2	3943	1785	2909	76286	86361	74568	95774	110127	93214
		3	3917	2067	3918	90996	87383	76466	114142	108929	96018
40	5	0	−3387	−2708	−3163	42639	55372	46015	53180	69039	56768
		1	−1303	−1938	−1692	50192	58119	49758	63363	73246	62801
		2	−1873	−2795	−1878	63968	59279	54812	80239	74044	69342
		3	1307	3439	3257	84915	63190	64886	106087	78810	81292
	10	0	136	3384	1507	43763	59919	47779	55228	74335	59948
		1	1983	261	1074	50405	56557	48926	63690	71156	61293
		2	−32	3881	2199	64503	58688	54718	80708	73364	68497
		3	578	504	1549	80686	60774	62346	101530	76981	78455

Abias = absolute bias, RMSE = root mean squared error.

Table 2: Comparison results with $(\bar{M}_1, \sigma_{\bar{M}_1}) = (60, 4)$

n	m_{h_i}	CV_1	Bias			Abias			RMSE		
		(%)	ratio	cluster	com	ratio	cluster	com	ratio	cluster	com
20	5	0	2738	−719	544	129244	149364	139756	162020	187351	175130
		1	−1885	−5480	−3629	138010	164342	147252	170317	205318	183362
		2	4821	−74	2827	161636	168591	152781	202792	209229	190531
		3	8014	552	5989	188795	146808	143700	238981	183651	180516
	10	0	−5629	−4427	−5035	125727	139736	133095	156323	173100	164868
		1	−12732	−12519	−12479	134527	138234	133149	166828	172477	165391
		2	3199	447	2194	152801	136211	133179	192956	170680	167668
		3	−1471	2703	3639	188964	138715	137314	234252	171749	171321
40	5	0	2473	3068	2624	91554	95604	91590	114135	121122	115838
		1	−1467	4726	2419	94104	98611	90347	119672	124781	114614
		2	−5675	−763	−1864	135500	100956	104694	168242	124306	129777
		3	−4411	−176	1172	167519	95438	109145	208048	120531	135661
	10	0	−473	779	−74	89211	98252	92475	111957	122360	115568
		1	−769	−884	−913	96250	95815	91510	119786	119180	113429
		2	2785	2044	3635	132502	98426	101113	164583	122046	126581
		3	4458	−264	4092	163370	92351	105927	204709	117070	133566

Abias = absolute bias, RMSE = root mean squared error.

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

$$\text{ABias} = \frac{1}{R}\sum_{r=1}^{R}\left|\hat{Y}^{(r)} - Y\right|.$$

Here we use replication number, $R = 1{,}000$ and the results are tabulated in Table 1 to Table 3. From Table 1 to Table 3, we can see the similar trend of results. First of all, based on Bias, we cannot see any pattern by changes in the numbers of sample clusters or elements of each cluster. The three estimators are all unbiased. Now when we change the number of sample clusters, 20 to 40 which means, the number of sample clusters is increasing, the results of Abias and RMSE show that all

Table 3: Comparison results with $(\bar{M}_1, \sigma_{\bar{M}_1}) = (60, 8)$

n	m_{h_i}	CV_1 (%)	Bias			Abias			RMSE		
			ratio	cluster	com	ratio	cluster	com	ratio	cluster	com
20	5	0	2091	−224	454	126682	173960	145137	160600	223836	186664
		1	−18738	−13050	−15411	137025	194506	157803	169711	239948	194287
		2	−2874	−4180	−3381	157295	174411	150204	200333	221292	191979
		3	16182	6295	11413	194807	169451	159063	243860	212264	198785
	10	0	−5714	−4375	−6311	123797	163102	139152	155073	206022	174892
		1	347	1286	859	134821	172857	147388	170415	215237	182762
		2	972	−501	1154	164665	173700	154188	205641	215881	192029
		3	−8868	−10257	−8500	191113	175508	161688	240872	220967	201210
40	5	0	3127	−493	1293	89605	121985	97760	110815	151083	120078
		1	−1255	5618	1833	98782	121432	98211	123481	152000	123374
		2	−1760	−12	12	126862	126322	109212	160940	157566	139188
		3	−2221	−575	1250	165024	127531	124456	206835	158103	157917
	10	0	2234	1055	1960	91919	130454	100770	115739	163531	127420
		1	6628	2888	5196	103656	125007	103859	129933	156408	130869
		2	−3762	−9226	−4681	131463	121479	112264	164179	152589	139289
		3	550	1255	2797	177316	122268	130812	216391	153856	159909

Abias = absolute bias, RMSE = root mean squared error.

estimators improve the accuracy of estimates. However, obviously increasing m_{h_i}, 5 to 10, does shows little improvement.

It is clear that if CV_1 gets larger, then the ratio estimator rapidly deteriorates. Obviously the results of the two stage cluster estimator do not change because of not using the information of the size of population, $M_h = M_1$. Now for the case of $CV_1 = 0$, with known and accurate population information, the ratio estimator is the best. However, if CV_1 is greater than 0.01 then the two stage cluster estimator is better than the ratio estimator.

However, the suggested composite estimator shows stability on every case. Especially, the suggested composite estimator has relative merit for the case of $CV_1 = 0.01$. That is, the suggested estimator is at least the same as the better one when based on Abias and RMSE statistics. Furthermore, when CV_1 gets large, the ratio estimator rapidly deteriorates. However, the suggested estimator still gives stable results. Also, the two stage cluster estimator has a large bias compared to the suggested estimator (relatively).

Consequently, the ratio estimator is the best if we do have known and accurate information on population. However, knowing the accurate information about the population at the survey period is not quite possible and difficult. Therefore the suggested composite estimator will be best when the population is changed but not knowing the exact information.

5. A real data analysis

In this section, we use Taxi Company data from 170 taxi companies in Korea. The data includes the variables: region, company name, travel distance/transfer distance and we are interested in the travel distance variable. For this real data analysis, we adopt the Salvati *et al.* (2010) method for generating pseudo population. With 170 surveyed taxi company data, the pseudo population is generated by re-samples of clusters, with a replacement 30 times. That means we have 5,100 clusters in pseudo population. The administrative district are made of 2 strata. Then the number of clusters in each stratum, N_j and the number of elements in each cluster, M_{hi} are obtained. Among these, we deleted M_{hi} which is greater than 150 or less than 7 because in practice the differences between M_{hi} are hardly large.

Table 4: Comparison results for real data analysis (Stratum 1)

n	m_{h_i}	CV_1 (%)	Bias ratio	Bias cluster	Bias com	Abias ratio	Abias cluster	Abias com	RMSE ratio	RMSE cluster	RMSE com
15	3	0	−250	−3390	−1671	7130	9513	8234	8877	11889	10336
		3	238	−2794	−1334	7595	8722	7601	9642	11019	9624
		5	271	−2312	−682	9024	9172	8232	11335	11510	10405
		10	806	−2624	−443	12863	8952	8989	16165	11276	11414
	5	0	−266	−3037	−1573	5745	8273	6751	7255	10456	8492
		3	136	−2224	−992	6583	7905	6816	8234	9946	8539
		5	79	−2960	−854	12692	7870	8247	15828	9954	10481
		10	−297	−2666	−1110	7943	8017	6947	9916	10085	8752
30	3	0	32	−1619	−1050	4404	4985	4665	5495	6242	5818
		3	517	−1404	−745	5380	4716	4608	6860	5997	5851
		5	325	−1644	−865	7054	4871	4953	8924	6245	6346
		10	−816	−1805	−1174	11905	4931	6024	14941	6262	7986
	5	0	56	−1635	−1149	3265	3775	3487	4094	4754	4375
		3	−37	−1499	−1044	4569	3764	3672	5775	4752	4565
		5	326	−1743	−997	6441	3925	3959	8168	4920	5017
		10	605	−1584	−505	11675	3744	4843	14606	4800	6511

Abias = absolute bias, RMSE = root mean squared error.

Table 5: Comparison results for real data analysis (Stratum 2)

n	m_{h_i}	CV_1 (%)	Bias ratio	Bias cluster	Bias com	Abias ratio	Abias cluster	Abias com	RMSE ratio	RMSE cluster	RMSE com
15	3	0	−1662	1635	−317	7353	8819	7906	9429	11240	10049
		3	−1015	2166	689	7822	8884	8155	9949	11047	10204
		5	−1258	1843	384	9360	9317	8432	11844	11535	10660
		10	−325	1849	406	12348	8360	8816	15759	10581	11607
	5	0	−1364	1951	221	6574	8249	7017	8410	10378	8851
		3	−1060	2218	64	7377	8359	7121	9339	10448	9090
		5	−935	2549	842	8126	8499	7444	10469	10615	9383
		10	−975	2051	1142	12961	8321	9004	16500	10323	11344
30	3	0	−909	−387	−559	4058	4293	4238	5174	5410	5371
		3	−732	−551	−587	5090	4326	4366	6403	5372	5504
		5	−530	−628	−384	6612	4436	4777	8374	5592	6118
		10	−1898	−963	−852	11508	4236	5549	14338	5367	7552
	5	0	−1278	−954	−1056	3304	3566	3449	4192	4452	4327
		3	−997	−678	−730	4498	3527	3561	5723	4501	4575
		5	−776	−646	−629	6350	3640	3971	8039	4621	5173
		10	−1161	−718	−464	11630	3586	5098	14489	4640	7158

Abias = absolute bias, RMSE = root mean squared error.

From each stratum, sample clusters with $n_h = 15, 30$ are selected and from each cluster, samples with $m_{hi} = 3, 5$ are selected. Three estimators stated in Section 3 are calculated and the results are compared based on the comparison statistics in Section 4. We use CV $= 0, 0.03, 0.05$ and 0.1 to consider the changes on population. The number of replications is 1,000 and the results are in Table 4 and Table 5.

Both Tables 4 and 5 show similar results. First, based on Bias, we cannot see any trend or pattern because three estimators are all unbiased. However, based on Abias and RMSE, the results vary depending on the value of CV. Especially from Table 4 in case of $n = 15$, the results of the two stage cluster and the ratio estimates are switched with respect to RMSE in between CV = 0.05 and $CV = 0.1$. However, the results of the case of $n = 30$, are switched between CV = 0.03 and CV = 0.05. The suggested composite estimator also shows stable and relatively good results. Table 5 shows similar results to Table 4.

6. Summary and conclusion

When the total number of population elements is unknown, the two stage cluster estimator should be used to estimate the total. However, the ratio estimator is recommended if the total number of population elements is known and accurate. The ratio estimator or the two stage cluster estimator can be used when the total number of population elements is known but not accurate. At this point, the choice of estimators is depending on the accuracy about information of population and on the number of cluster elements. However, those terms cannot be assured in general. Therefore, we suggest a composite estimator which is a linear combination of the two stage cluster and the ratio estimators.

There exists a time difference between sampling frame and survey period. Also, cluster sampling designs are usually used to reduce the survey costs. Therefore, the number of population elements is not known in most cases and the sizes of each cluster are not the same. For that case, the suggested composite estimator will give more stable and better estimates compared with two other estimators under those conditions.

Acknowledgement

This study was supported by Hankuk University of Foreign Studies research fund (2015).

References

Cochran WG (1977). *Sampling Technique*, John Wiley and Sons, New York.

Hwang HJ and Shin KI (2013). An improved composite estimator for cut-off sampling, *Communications for Statistical Applications and Method*, **20**, 367–376.

Lee SE, Jin Y, and Shin KI (2015). A Note on complex two-phase sampling with different sampling units of each phase, *Communications for Statistical Applications and Method*, **22**, 435–443.

Quenouille LLE MH (1949). The joint distribution of serial correlation coefficients, *The Annals of Mathematical Statistics*, **20**, 561–571.

Rao JNK (2003). *Small Area Estimation*, Wiley-Interscience.

Salvati N, Chandra H, Ranalli MG, and Chambers R (2010). Small area estimation using a non-parametric model-based direct estimator, *Computational Statistics and Data Analysis*, **54**, 2159–2171.

Tukey JW (1958). Bias and confidence in not quite large samples, *Annals of Mathematical Statistics*, **29**, 614.

Wolter KM (1985). *Introduction to Variance Estimation*, Springer, New York.

Received October 13, 2015; Revised January 11, 2016; Accepted January 13, 2016