Communications for Statistical Applications and Methods
2013, Vol. 20, No. 5, 367–376

DOI: http://dx.doi.org/10.5351/CSAM.2013.20.5.367

An Improved Composite Estimator for Cut-off Sampling

Hee-Jin Hwang[a], Key-Il Shin[1,b]

[a]The Bank of Korea; [b]Department of Statistics, Hankuk University of Foreign Studies

Abstract

Cut-off sampling is widely used for a highly skewed population like a business survey by discarding a part of the population (the take-nothing stratum). In this paper, we suggest a new composite estimator of the take-nothing stratum total obtained by use of the survey results of the take-nothing stratum and a take-some sub-stratum (a part of take-some stratum) for a more accurate estimate of the population total. Small simulation studies are conducted to compare the performances of known estimators and the new composite estimator suggested in this study. In addition, we use briquette consumption survey data for real data analysis.

Keywords: Best linear unbiased predictor(BLUP), Lavallee-Hidiroglou algorithm, ratio estimator, take-nothing stratum.

1. Introduction

Cut-off sampling is a well-known sampling design commonly used for a highly skewed population like a business survey. The cut-off sampling (which divides the population into three sub-populations) is a special case of stratified sampling. Sub-populations are known as the take-all, take-some and take-nothing stratums. Then the estimated population total is frequently obtained by the summation of the estimated totals of three strata.

In a business survey, the precision of the estimated total might be improved by the conduction of a census for the take-all stratum composed of large size companies. Also in some business surveys the precision might be improved by excluding the take-nothing stratum because of survey difficulties and costs. As a special case of the cut-off sampling, Hidiroglou (1986) suggested a modified cut-off sampling that divides the population into only two sub-populations (the take-all and take-some stratum). However, it is important to improve the precision for the take-nothing stratum since the precision of an estimator for the take-nothing stratum could greatly affect that of the population total.

Several methods to improve the precision of the estimate for the take-nothing stratum have been suggested with auxiliary information or administrative data (see Sarndal *et al.* (1992), Elisson and Elvers (2001) and Benedetti *et al.* (2010) for more details). Hwang and Shin (2012) also suggested a composite estimator that uses information from the take-nothing stratum and the take-some stratum; subsequently, they compared the performances of the estimators and showed the superiority of the composite estimator.

In this paper, we suggest a new composite estimator for the total of the take-nothing stratum obtained by the use of the survey results of the take-nothing stratum and the take-some sub-stratum (a part of the take-some stratum). There are several stratification methods to divide a population

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology (2012R1A1A2003919).
[1] Corresponding author: Professor, Department of Statistics, Hankuk University of Foreign Studies, Yongin, Kyonggi 449-791, Korea. E-mail: keyshin@hufs.ac.kr

into sub-populations; subsequently, the well-known L-H (Lavallee-Hidiroglou) algorithm is used for stratification in this paper. Using L-H algorithm, we divide take-some stratum into H sub-strata. After that we choose one take-some sub-stratum which is the most correlated with the take-nothing stratum. Then we obtain a composite estimator for the total of the take-nothing stratum that combines the information of the chosen sub-stratum and the take-nothing stratum. In addition, it is confirmed that the suggested composite estimator improves the precision of the estimated population total.

Section 2 explains some notations, composite estimators developed recently and the L-H algorithm. In addition, the composite estimator suggested in this study is illustrated. In Section 3, small simulation studies are conducted to compare performances of several estimators illustrated by Hwang and Shin (2012) and the composite estimator suggested in this study. In Section 4, we confirm the efficiency of the suggested estimator in use of real data and the briquette consumption survey data. Section 5 provides the conclusions.

2. Estimators of the Population Total

We use the general structure and notations used in Benedetti *et al.* (2010). Let U and N be the population and the number of the population respectively. Then U can be divided into three sub-populations or strata, $U = U_C \cup U_S \cup U_{SE}$, and denote $U_I = U_C \cup U_S$. Here U_C is the take-all stratum, U_S, the take-some stratum, U_{SE}, the take-nothing stratum, but few samples surveyed, and U_I is the inclusion stratum. Of course U_S can be divided into the H sub-strata, U_{S_h}, and $U_S = \bigcup_{h=1}^{H} U_{S_h}$. Then the estimate of the population total t_y could be calculated by the summation of the totals of the divided strata defined by

$$t_y = t_{yU_C} + t_{yU_S} + t_{yU_{SE}}, \qquad t_{yU_I} = t_{yU_C} + t_{yU_S},$$

where t_{yU_C}, t_{yU_S}, $t_{yU_{SE}}$ and t_{yU_I} are the totals of each stratum respectively. Also, given the auxiliary variable, population totals and each stratum for x are denoted by t_x, t_{xU_C}, t_{xU_S}, $t_{xU_{SE}}$ and t_{xU_I}. In addition, let I, S and SE be the indicator sets of samples corresponding to U_I, U_S and U_{SE}. Also $S = \bigcup_{h=1}^{H} S_h$, where S_h is the indicator set of h sub-stratum samples.

2.1. Estimators of the total

Some known estimators of the population total explained briefly in this section are the same as those explained in Hwang and Shin (2012).

2.1.1. Sarndal-Swansson-Wretman(SSW) estimator

Sarndal *et al.* (1992) suggested a ratio estimator by use of the ratio of two variables, an auxiliary variable x to an interesting variable y in the inclusion stratum. The Sarndal-Swansson-Wretman estimator(SSW), $\hat{t}_y^{SSW}$, is defined by

$$\hat{t}_y^{SSW} = \hat{R}_{yxU_I} t_x, \tag{2.1}$$

where $\hat{R}_{yxU_I} = \hat{t}_{yU_I}/\hat{t}_{xU_I}$, $\hat{t}_{yU_I} = t_{yU_C} + \hat{t}_{yU_S}$, $\hat{t}_{xI} = t_{xU_C} + \hat{t}_{xU_S}$, $\hat{t}_{yU_S} = \sum_{k \in S} w_k y_k$, $\hat{t}_{xU_S} = \sum_{k \in S} w_k x_k$ and w_k is a weight.

2.1.2. Composite estimators

Kim and Shin (2011) suggested a composite estimator for the total of the take-nothing stratum defined by

$$\hat{t}_{yU_{SE}}^{MODI-SSW} = \left(\alpha^{[1]}\frac{\hat{t}_{yU_{SE}}}{\hat{t}_{xU_{SE}}} + \left(1-\alpha^{[1]}\right)\frac{\hat{t}_{yU_I}}{\hat{t}_{xU_I}}\right) t_{xU_{SE}}. \tag{2.2}$$

This estimator is obtained by combining the estimator based on SSW, $\hat{t}_{yU_{SE}}^{SSW}$, with the ratio estimator $\hat{t}_{yU_{SE}}^{Ratio} = (\hat{t}_{yU_{SE}}/\hat{t}_{xU_{SE}})t_{xU_{SE}}$ is obtained by using a few samples in the take-nothing stratum. Here $\hat{t}_{yU_{SE}} = \sum_{k\in SE} w_k y_k$ and $\hat{t}_{xU_{SE}} = \sum_{k\in SE} w_k x_k$. Hwang and Shin (2012) suggested composite estimators using the best linear unbiased predictor (BLUP) for the total of the stratum U_{SE}, $\hat{t}_{yU_{SE}}^{BLUP}$. In that paper, for the total of the stratum U_{SE}, two composite estimators are suggested as in (2.3) and (2.4).

$$\hat{t}_{yU_{SE}}^{MODI-BLUP} = \hat{R}_{U_{SE}}^{MODI-BLUP} t_{xU_{SE}} = \left(\alpha^{[2]}\frac{\hat{t}_{yU_{SE}}}{\hat{t}_{xU_{SE}}} + \left(1-\alpha^{[2]}\right)\frac{\hat{T}_{yU_I}}{\hat{T}_{xU_I}}\right) t_{xU_{SE}}, \tag{2.3}$$

$$\hat{t}_{yU_{SE}}^{MODI-BLUPA} = \hat{R}_{U_{SE}}^{MODI-BLUPA} t_{xU_{SE}} = \left(\alpha^{[3]}\frac{\hat{t}_{yU_{SE}}}{\hat{t}_{xU_{SE}}} + \left(1-\alpha^{[3]}\right)\frac{\hat{T}_{yU_S}}{\hat{T}_{xU_S}}\right) t_{xU_{SE}}. \tag{2.4}$$

Here $\hat{t}_{yU_{SE}}$, $\hat{t}_{xU_{SE}}$ are defined in (2.2) and $\hat{T}_{yU_I} = \sum_{k\in I} y_k$, $\hat{T}_{yU_S} = \sum_{k\in S} y_k$, $\hat{T}_{xU_I} = \sum_{k\in I} x_k$ and $\hat{T}_{xU_S} = \sum_{k\in S} x_k$.

Also, the weight α in (2.2), (2.3) and (2.4) can be calculated using MSE or a variance of each estimator. For example, $\alpha^{[1]}$ can be calculated using (2.5).

$$\hat{\alpha}^{[1]} = \frac{\mathrm{MSE}\left(\hat{R}_{U_{SE}}^{SSW}\right)}{\mathrm{MSE}\left(\hat{R}_{U_{SE}}^{MODI}\right) + \mathrm{MSE}\left(\hat{R}_{U_{SE}}^{SSW}\right)} \approx \frac{\mathrm{Var}\left(\hat{R}_{U_{SE}}^{SSW}\right)}{\mathrm{Var}\left(\hat{R}_{U_{SE}}^{MODI}\right) + \mathrm{Var}\left(\hat{R}_{U_{SE}}^{SSW}\right)}. \tag{2.5}$$

Here $\hat{R}_{U_{SE}}^{MODI} = \hat{t}_{yU_{SE}}/\hat{t}_{xU_{SE}}$ and $\hat{R}_{U_{SE}}^{SSW} = \hat{t}_{yU_I}/\hat{t}_{xU_I}$ (see Rao (2003) for more details). Finally Kim and Shin (2011) and Hwang and Shin (2012) used the same $\hat{t}_{yU_I}$ defined in SSW for the estimate of t_{yU_I}. Therefore we obtain three composite estimators.

$$\hat{t}_{y}^{MODI-SSW} = \hat{t}_{yU_I} + \hat{t}_{yU_{SE}}^{MODI-SSW}, \tag{2.6}$$

$$\hat{t}_{y}^{MODI-BLUP} = \hat{t}_{yU_I} + \hat{t}_{yU_{SE}}^{MODI-BLUP}, \tag{2.7}$$

$$\hat{t}_{y}^{MODI-BLUPA} = \hat{t}_{yU_I} + \hat{t}_{yU_{SE}}^{MODI-BLUPA}. \tag{2.8}$$

2.2. Algorithm for stratification

For the heavily skewed population, there are several algorithms such as L-H(Lavallee-Hidiroglou) algorithm, geometric stratification algorithm, and random search algorithm to divide the population into the take-all stratum and the H sub-strata. These methods are used to calculate the optimal boundaries between each stratum to minimize the total sample size. In this paper, we use the L-H Algorithm for stratification. The brief explanation of the L-H algorithm is as follows.

The algorithm suggested by Lavallee and Hidiroglou (1988) is a method to find the best boundaries between each stratum and the least sample size by iterative calculation, given the target CV, c, and the total number of the stratum, H, (including the take-all stratum). Here the sample size n is a function of W_h, S_h, and a_h and defined by

$$n = N_H + \frac{\sum_{h=1}^{H} \frac{W_h^2 S_h^2}{a_h}}{c^2 \bar{X}^2 + \sum_{h=1}^{H} \frac{W^h S_h^2}{N}}, \tag{2.9}$$

where

n : total sample size

N : total population size, $N = \sum_{h=1}^{H} N_h$ and N_h, total population size of stratum h

S_h^2 : population variance of stratum h, $S_h^2 = \frac{1}{N_h - 1} \sum_{i=1}^{N_h} (x_{hi} - \bar{X}_h)^2$, $\bar{X}_h = N_h^{-1} \sum_{i=1}^{N_h} x_{hi}$

W_h : weight of stratum h, $W_h = N_h/N$

a_h : sample allocation rate of stratum h, $a_h = \frac{N_h S_h}{\sum_{h=1}^{H-1} N_h S_h}$

$\bar{X}$: total mean, $\bar{X} = \sum_{h=1}^{H} W_h \bar{X}_h$

c : Coefficient of Variation, CV

The sample size n is larger if the target CV, c, is smaller. Also, n becomes smaller if H would be larger. When each stratum is defined as (2.10) and n is expressed as a function of boundaries between each stratum for variable X, then optimized k (a vector of boundaries) could be obtained by the solution of (2.11). That is, if

$$U_h = \{i : k_{h-1} < x_i \le k_h\} \tag{2.10}$$

and $k_1 < \cdots < k_h < \cdots < k_{H-1}$, $k_0 = -\infty$, $k_H = \infty$, then the solution satisfied (2.8) can be obtained.

$$\frac{\partial n(k)}{\partial k_1} = \cdots = \frac{\partial n(k)}{\partial k_h} = \cdots = \frac{\partial n(k)}{\partial k_{H-1}} = 0. \tag{2.11}$$

In addition, (2.11) is expressed as a quadratic equation of k_h defined by (2.11).

$$\alpha_h k_h^2 + \beta_h k_h + \gamma_h = 0. \tag{2.12}$$

Given initial values of $k^{(0)} = (k_1^{(0)}, \ldots, k_h^{(0)}, \ldots, k_{H-1}^{(0)})'$, the solution of (2.12) is iteratively calculated and final boundaries are obtained by the converged value of $k^{(r)}$, $r = 1, 2, \ldots$. Details are found in Lavallee and Hidiroglou (1988).

2.3. Suggested estimator

The composite estimators of the total of the stratum U_{SE}, $\hat{t}_{yU_{SE}}$ (suggested in Session 2.1) are the linear combined estimators that use the estimated total of the stratum U_{SE} and that of the stratum U_I (or the take-some stratum U_S). When we estimate the total of the stratum U_{SE}, we may obtain better results by selectively using the part of the stratum U_I instead of the whole stratum U_I. The

Table 1: Parameters used in simulation study

Population types	a	b	c	d	g
Ratio	0	1.50	0.00	5.13	0.50
Regression	20	1.50	0.00	13.79	0.25
Convex	0	0.25	0.01	4.91	0.50
Concave	0	3.00	−0.01	5.60	0.50

information for better results can be obtained from the closer part of the take-some stratum to the take-nothing stratum.

For this reason we divide the take-some stratum into H sub-strata. The L-H stratification algorithm is used to divide the take-some stratum. Among the divided H sub-strata, the closer sub-stratum to the take-nothing stratum is considered to have more similar characteristics. Therefore, to estimate the total of the stratum U_{SE}, a method using only the closest sub-stratum to the take-nothing stratum is suggested. For $U_S = \bigcup_{h=1}^{H} U_{S_h}$, let U_{S_1} be the nearest sub-stratum to the take-nothing stratum. Then the newly suggested estimator of the total of stratum U_{SE} using the information of only U_{S_1} is defined by

$$\hat{t}_{yU_{SE}}^{MODI-BLUPN} = \left(\alpha^{[4]} \frac{\hat{t}_{yU_{SE}}}{\hat{t}_{xU_{SE}}} + \left(1 - \alpha^{[4]}\right) \frac{\hat{T}_{yU_{S_1}}}{\hat{T}_{xU_{S_1}}} \right) t_{xU_{SE}}, \tag{2.13}$$

where $\hat{T}_{yU_{S_1}} = \sum_{k \in S_1} y_k$ and $\hat{T}_{xU_{S_1}} = \sum_{k \in S_1} x_k$. Here $\hat{\alpha}^{[4]}$ can be similarly obtained by using (2.5). Also, like the other composite estimators, we use the same $\hat{t}_{yU_I}$ defined in the SSW to estimate t_{yU_I}. Therefore, we have the following proposed composite estimator.

$$\hat{t}_{y}^{MODI-BLUPN} = \hat{t}_{yU_I} + \hat{t}_{yU_{SE}}^{MODI-BLUPN}. \tag{2.14}$$

3. Simulation Study

We conduct a small simulation study to compare the efficiency of the newly suggested estimator and the other composite estimators. The simulation methods used in this paper are the same as those in Lee *et al.* (1995) and Hwang and Shin (2012).

First, after generating the auxiliary variable x_k from the gamma distribution with the mean 48 and variance 768, we also generate four types of populations of interesting variable, y_k. Here y_k is assumed to follow gamma distribution with mean $\mu(x) = a + bx + cx^2$ and variance $\sigma^2(x) = d^2 x^{2g}$. Using the population size $N = 10{,}000$, we pre-determine the same cut-off point for each case to compare the previous results obtained by Hwang and Shin (2012).

Values of parameters a, b, c, d and g used for four types of the generated population are shown in Table 1. The first data set is a ratio type that is a linear function of an auxiliary variable x_k and an interesting variable y_k passing through the origin. The second data set is a regression type with a positive intercept, the third data set stands for a convex function type and the fourth data set stands for a concave function type.

We use L-H algorithm to divide the take-some stratum into a sub-strata. The L-H algorithm needs the number of strata and the target CV value, c. Here we reversely calculate CV and the number of strata to meet the sample size n. We use $n = 500$ for the total sample size and the sampling fraction $f = 0.05$. Also, we consider two values, $n_{SE} = 5$ and $n_{SE} = 10$, for the sample size of the stratum U_{SE}. Table 2 summarizes the population size, N, sample size, n, and sampling fraction, f. Table 3 presents the design weights for the take-some sub-strata.

Table 2: Population size, N, sample size, n, and sampling fraction, f

Cut-off point	N_C	N_S	N_{SE}	n_S	n_{SE}	n_S/N_S	n_{SE}/N_{SE}	Target CV value (%)
80%	76	5896	4028	424	0	0.0719	0.0000	2 strata : 0.82
				419	5	0.0711	0.0012	3 strata : 0.58
				414	10	0.0702	0.0025	4 strata : 0.50
90%	76	7374	2550	424	0	0.0575	0.0000	2 strata : 1.02
				419	5	0.0568	0.0020	3 strata : 0.69
				414	10	0.0561	0.0039	4 strata : 0.52
95%	76	8332	1592	424	0	0.0509	0.0000	2 strata : 1.16
				419	5	0.0503	0.0031	3 strata : 0.78
				414	10	0.0497	0.0063	4 strata : 0.60

Table 3: Design weights used for take-some sub-strata in a simulation study

Cut-off point		2 strata		3 strata			4 strata			
		S_2	S_1	S_3	S_2	S_1	S_4	S_3	S_2	S_1
80%	population size	3798	2098	2515	2108	1273	2001	1785	1347	763
	sample size	194	220	114	109	191	31	65	142	176
	weight	19.58	9.54	22.06	19.34	6.66	64.55	27.46	9.49	4.34
90%	population size	4700	2674	3200	2657	1517	2455	2174	1727	1018
	sample size	194	220	120	125	169	30	92	123	169
	weight	24.23	12.15	26.67	21.26	8.98	81.83	23.63	14.04	6.02
95%	population size	5176	3156	3625	3000	1707	2506	2463	2110	1253
	sample size	182	232	125	127	162	78	72	120	144
	weight	28.44	13.6	29.00	23.62	10.54	32.13	34.21	17.58	8.70

We use three comparison statistics, bias, relative bias(rbias) and root mean square error(RMSE) defined by

$$\text{bias} = \bar{\hat{t}}_y - t_y,$$

$$\text{rbias}(\%) = \frac{100\left(\bar{\hat{t}}_y - t_y\right)}{t_y},$$

$$\text{rmse} = \sqrt{\frac{1}{R}\sum_{r=1}^{R}\left[\hat{t}_y(r) - t_y\right]^2},$$

where $\bar{\hat{t}}_y = \sum_{r=1}^{R} \hat{t}_y(r)/R$ and $R = 1{,}000$.

Tables 4–9 show the results of four population types. Here SSW means the Sarndal-Swansson-Wretman estimator, M-S, M-B, M-BA are the composite estimators defined by (2.6), (2.7) and (2.8) respectively. Also, M-BN_∗ stands for the suggested composite estimator defined by (2.14). For example M-BN_2 is the composite estimator obtained using the closest take-some sub-stratum among the two take-some sub-strata to the take-nothing stratum.

Table 4 shows that the RMSE criterion, M-BN_3 composite estimator provides the best results; however, the M-B estimator provides the best result for the ratio-type population. Also, M-BN_4 composite estimator is the best for the bias results.

Table 5 shows very similar results to Table 4. Using RMSE criterion, the M-BN_3 composite estimator and M-BN_2 composite estimator provide the best results. However, the M-B estimator gives the best result for the ratio-type population (see Table 4). The M-BN_4 composite estimator is the best for the bias results; however, the M-BN_2 composite estimator shows the best result for the ratio-type population. The data set in Table 5 has more information since value of the RMSE (or rbias) is smaller than those of the data set in Table 4. Table 6 also shows similar results.

Table 4: Simulation results with n_{SE} = 10, cut-off point = top 80%

Types		Estimation methods						
		SSW	M-S	M-B	M-BA	M-BN_2	M-BN_3	M-BN_4
Ratio	bias	−4119	−3641	−4095	−3546	−3678	−3361	−2042
	rbias(%)	−0.57	−0.51	−0.57	−0.49	−0.51	−0.47	−0.28
	rmes	14811	13412	13199	13455	14277	15886	21047
Linear	bias	−48259	−22258	−20488	−22353	−21282	−20035	−15405
	rbias(%)	−5.22	−2.41	−2.22	−2.42	−2.30	−2.17	−1.67
	rmes	50066	28754	28804	28653	26917	26123	27027
Convex	bias	46941	10389	6861	10725	8155	5931	4447
	rbias(%)	10.90	2.41	1.59	2.49	1.89	1.38	1.03
	rmes	49379	25679	29093	24918	18473	17294	22866
Concave	bias	−43339	−18658	−14159	−18764	−13115	−10608	−7375
	rbias(%)	−3.82	−1.64	−1.25	−1.65	−1.15	−0.93	−0.65
	rmes	46573	27320	29543	26767	20398	19667	25084

Table 5: Simulation results with n_{SE} = 10, cut-off point = top 90%

Types		Estimation methods						
		SSW	M-S	M-B	M-BA	M-BN_2	M-BN_3	M-BN_4
Ratio	bias	−579	−456	−722	−414	−256	−448	354
	rbias(%)	−0.08	−0.06	−0.10	−0.06	−0.04	−0.06	0.05
	rmes	15138	14238	14042	14267	14407	14886	20450
Linear	bias	−33817	−14793	−13979	−14813	−13999	−13762	−11021
	rbias(%)	−3.66	−1.60	−1.51	−1.60	−1.51	−1.49	−1.19
	rmes	37056	20816	20832	20780	20371	19907	22086
Convex	bias	23063	6300	5036	6204	3365	2439	638
	rbias(%)	5.36	1.46	1.17	1.44	0.78	0.57	0.15
	rmes	28021	16637	18552	16395	14525	14575	18610
Concave	bias	−21412	−9517	−8659	−9279	−6233	−3488	−2822
	rbias(%)	−1.89	−0.84	−0.76	−0.82	−0.55	−0.31	−0.25
	rmes	27923	18663	19560	18514	17494	17165	22293

Table 6: Simulation results with n_{SE} = 10, cut-off point = top 95%

Types		Estimation methods						
		SSW	M-S	M-B	M-BA	M-BN_2	M-BN_3	M-BN_4
Ratio	bias	−1501	−765	−907	−746	−1172	−206	−558
	rbias(%)	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	rmes	16085	15648	15543	15662	14912	15519	15627
Linear	bias	−22646	−9738	−9592	−9720	−9080	−8248	−7935
	rbias(%)	−0.02	−0.01	−0.01	−0.01	−0.01	−0.01	−0.01
	rmes	27666	16961	16911	16958	17142	16127	17767
Convex	bias	11917	4544	4996	4349	2007	956	115
	rbias(%)	0.03	0.01	0.01	0.01	0.00	0.00	0.00
	rmes	21030	16561	16523	16584	14764	15475	15566
Concave	bias	−12091	−5567	−6283	−5349	−3065	−2214	−1902
	rbias(%)	−0.01	0.00	−0.01	0.00	0.00	0.00	0.00
	rmes	22262	18291	18187	18299	17296	16591	17529

The performance of the suggested composite estimator will improve if the additional information of the proper size of U_{SE} stratum could be used. Also M-BN type composite estimator shows the best performance except for the ratio-type population.

Now, we investigate the results of the sample size, $n_{SE} = 5$, of the take-nothing stratum. Like the previous results (using RMSE criterion), Table 7–Table 9 show that M-BN type composite estimator provides better results; however, the M-B estimator provides the best result for the ratio-type

Table 7: Simulation results with n_{SE} = 5, cut-off point = top 80%

Types		Estimation methods						
		SSW	M-S	M-B	M-BA	M-BN_2	M-BN_3	M-BN_4
Ratio	bias	−3962	−3779	−4230	−3681	−4357	−3304	−3414
	rbias(%)	−0.55	−0.53	−0.59	−0.51	−0.61	−0.46	−0.47
	rmes	14544	13349	13201	13383	14359	15668	22819
Linear	bias	−47725	−29538	−28654	−29467	−26436	−24957	−20302
	rbias(%)	−5.16	−3.20	−3.10	−3.19	−2.86	−2.70	−2.20
	rmes	49582	35034	35710	34801	31370	30630	31101
Convex	bias	46285	8329	3665	8831	6625	4517	2058
	rbias(%)	10.75	1.93	0.85	2.05	1.54	1.05	0.48
	rmes	48553	25828	31353	24792	18006	16991	22135
Concave	bias	−43312	−22339	−18552	−22144	−15056	−12250	−8369
	rbias(%)	−3.81	−1.97	−1.63	−1.95	−1.33	−1.08	−0.74
	rmes	46472	30429	33758	29582	22632	20899	25721

Table 8: Simulation results with n_{SE} = 5, cut-off point = top 90%

Types		Estimation methods						
		SSW	M-S	M-B	M-BA	M-BN_2	M-BN_3	M-BN_4
Ratio	bias	−662	−1257	−1504	−1215	157	−746	14
	rbias(%)	−0.09	−0.17	−0.21	−0.17	0.02	−0.10	0.00
	rmes	15343	14331	14146	14359	14449	15550	23887
Linear	bias	−34072	−19341	−18917	−19305	−13972	−16676	−13200
	rbias(%)	−3.69	−2.09	−2.05	−2.09	−1.51	−1.80	−1.43
	rmes	37411	25221	25510	25139	20001	23349	26784
Convex	bias	22147	5128	3848	4996	2862	786	−337
	rbias(%)	5.14	1.19	0.89	1.16	0.66	0.18	−0.08
	rmes	27298	16488	19352	16151	14754	14596	22665
Concave	bias	−21335	−10861	−10426	−10515	−5984	−3565	−2611
	rbias(%)	−1.88	−0.96	−0.92	−0.93	−0.53	−0.31	−0.23
	rmes	27809	19806	21502	19533	16921	16802	28379

Table 9: Simulation results with n_{SE} = 5, cut-off point = top 95%

Types		Estimation methods						
		SSW	M-S	M-B	M-BA	M-BN_2	M-BN_3	M-BN_4
Ratio	bias	−1122	−1409	−1528	−1392	−1902	−1210	−978
	rbias(%)	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	rmes	15842	14927	14842	14938	15179	15270	15454
Linear	bias	−22539	−11675	−11655	−11640	−10992	−9972	−9221
	rbias(%)	−0.02	−0.01	−0.01	−0.01	−0.01	−0.01	−0.01
	rmes	27582	18413	18466	18395	18133	17865	18249
Convex	bias	11143	3152	3720	2942	1353	122	−237
	rbias(%)	0.03	0.01	0.01	0.01	0.00	0.00	0.00
	rmes	19839	15385	15607	15402	15690	15190	16335
Concave	bias	−11669	−5801	−6750	−5554	−3823	−3455	−1923
	rbias(%)	−0.01	−0.01	−0.01	0.00	0.00	0.00	0.00
	rmes	22222	18595	18738	18573	18142	17256	17480

population. The M-BN_4 composite estimator is the best for the bias results.

The results of a comparison of the two cases, n_{SE} = 10 and n_{SE} = 5 does not show any difference. That means the additional information is very helpful to improve the precision of estimation even though only small samples are surveyed in the take-nothing stratum.

The proposed composite estimator improves the precision of estimated total in most cases by

Table 10: Summary of the sample design

cut-off point	N_C	N_S	N_{SE}	n_S	n_{SE}	n_S/N_S	n_{SE}/N_{SE}	Target CV value (%)
80%	80	572	960	110	0	0.1146	0.0000	2 strata : 1.10
				100	10	0.1042	0.0104	3 strata : 0.70
				90	20	0.0938	0.0208	4 strata : 0.55
90%	80	784	748	110	0	0.1471	0.0000	2 strata : 1.50
				100	10	0.1337	0.0134	3 strata : 1.00
				90	20	0.1203	0.0267	4 strata : 0.77
95%	80	948	584	110	0	0.1160	0.0000	2 strata : 2.50
				100	10	0.1055	0.0171	3 strata : 1.70
				90	20	0.0949	0.3425	4 strata : 1.23

Table 11: Design weight for take-some sub-strata in real data analysis

Cut-off point		2 strata		3 strata			4 strata			
		S_2	S_1	S_3	S_2	S_1	S_4	S_3	S_2	S_1
80%	population size	331	241	196	221	155	162	136	149	125
	sample size	45	55	24	34	42	20	16	29	35
	weight	7.36	4.38	8.17	6.50	3.69	8.10	8.50	5.14	3.57
90%	population size	486	298	350	265	169	227	190	215	152
	sample size	50	50	35	32	33	18	18	30	34
	weight	9.72	5.96	10.00	8.28	5.12	12.61	10.56	7.17	4.47
95%	population size	557	391	423	327	198	288	259	233	168
	sample size	44	56	25	40	35	22	21	26	31
	weight	12.66	6.98	16.92	8.18	5.66	13.09	12.33	8.96	5.42

using the additional information of the only small samples of the take-nothing stratum and the proper take-some sub-stratum.

4. Real Data Analysis

For real data analysis, the total sale amount and number of sales of about 1,600 delivery companies in a 2012 Briquette Consumption survey are used. Even though the purpose of this survey is to estimate the population total by use, in this analysis we compare the precision of the estimates of population total of sale amount obtained by each estimators as explained in Section 2. To divide population into strata, we use the L-H algorithm and the take-some stratum is divided into the H sub-strata. For the cut-off point, we use the 80%, 90% and 95% point of the population total sale amount to separate take-some stratum and the stratum U_{SE}. In addition, we use $N_C = 80$ for the take-all stratum and the sample size in the stratum U_{SE}, $n_{SE} = 10$ and $n_{SE} = 20$, respectively. We replicate 1,000 times to calculate the comparison statistics. Table 10 summarizes the sampling design in this section. Also, we Table 11 presents the design weights for the take-some sub-strata. Finally the simulation results are tabulated in Table 12.

Subsequently, SSW shows the worst results in all comparison statistics (Table 12). For all cases, M_BN_$*$ is superior to M-S, M-B and M-BA. Especially, M-BN_4 almost provides the best results using RMSE and bias criterion. However, M-BN_3 shows the best results in RMSE for the case of $n_{SE} = 10$ and the top 80% cut-off point. Therefore, we can conclude that the composite estimator developed in this paper provides better results than others. Also, the results of the three cases do not show any difference in a comparison of the 80%, 90% and 95% cases regardless of the sample sizes of the stratum U_{SE}.

Therefore, we can conclude that the proposed composite estimator is very useful to estimate the population total for this data.

Table 12: Briquette consumption survey results

			Estimation methods						
			SSW	M-S	M-B	M-BA	M-BN_2	M-BN_3	M-BN_4
n_{se} = 10	80%	bias	330E5	84E5	78E5	76E5	43E5	31E5	30E5
		rbias(%)	0.0654	0.0167	0.0155	0.0152	0.0086	0.0063	0.0061
		rmes	360E5	204E5	221E5	181E5	154E5	144E5	153E5
	90%	bias	165E5	58E5	72E5	48E5	29E5	22E5	12E5
		rbias(%)	0.0327	0.0116	0.0143	0.0096	0.0058	0.0045	0.0024
		rmes	224E5	167E5	184E5	160E5	149E5	148E5	138E5
	95%	bias	67E5	22E5	39E5	14E5	13E5	12E5	4E5
		rbias(%)	0.0133	0.0044	0.0079	0.0028	0.0027	0.0023	0.0009
		rmes	180E5	164E5	165E5	164E5	148E5	156E5	148E5
n_{se} = 20	80%	bias	331E5	68E5	59E5	66E5	36E5	24E5	31E5
		rbias(%)	0.0656	0.0135	0.0117	0.0132	0.0072	0.0049	0.0062
		rmes	365E5	178E5	186E5	166E5	148E5	149E5	147E5
	90%	bias	172E5	59E5	69E5	51E5	28E5	27E5	13E5
		rbias(%)	0.0341	0.0118	0.0138	0.0101	0.0057	0.0055	0.0027
		rmes	240E5	168E5	178E5	165E5	158E5	159E5	144E5
	95%	bias	79E5	34E5	51E5	26E5	17E5	11E5	10E5
		rbias(%)	0.0157	0.0067	0.0101	0.0051	0.0023	0.0022	0.0020
		rmes	192E5	173E5	175E5	173E5	163E5	164E5	163E5

5. Conclusion

In this study we suggest a new composite estimator obtained by combining the information of a take-nothing stratum and a selected take-some sub-stratum instead of a whole take-some stratum. We surveyed a few samples from a take-nothing stratum in order to get the desired information. A small simulation study shows that the composite estimator suggested in this study is very promising to improve the precision of the estimated population total; in addition, the real data analysis confirms the results.

References

Benedetti, R., Bee, M. and Espa, G. (2010). A framework for cut-off sampling in business survey design, *Journal of Official Statistics*, **26**, 651–671.

Elisson, H. and Elvers, E. (2001). Cut-off sampling and estimation, *Proceeding of Statistics Canada Symposium*.

Hidiroglou, M. A. (1986). The construction of a self-representing stratum of large units in survey design, *The American Statistician*, **4**, 27–31.

Hwang, J. M. and Shin, K.-I. (2012). An alternative composite estimator for the take-nothing stratum of the cut-off sampling, *Communications for Statistical Applications and Methods*, **19**, 13–22.

Kim, J.-H. and Shin, K.-I. (2011). A composite estimator for the take-nothing stratum of cut-off sampling, *The Korean Journal of Applied Statistics*, **24**, 1115–1128.

Lavallee, P. and Hidiroglou, M. (1988). On the stratification of skewed populations, *Survey Methodology*, **14**, 33–43.

Lee, H., Rancourt, E. and Sarndal, C.-E. (1995). Experiment with variance estimation from survey data with imputed value, *Journal of Official Statistics*, **10**, 231–243.

Rao, J. N. K. (2003). *Small Area Estimation*, Wiley Interscience, John Wiley and Sons, New York.

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

Received May 27, 2013; Revised July 18, 2013; Accepted August 8, 2013