A Composite Estimator for Cut-Off Sampling

Hee-Jin Hwang*
The Bank of Korea, Seoul, Republic of Korea, hjhwang@bok.or.kr

Key-Il Shin
Hankuk University of Foreign Studies, Seoul, Republic of Korea,
keyshin@hufs.ac.kr

Abstract

The cut-off sampling has been widely used for a highly skewed population like a business survey by discarding a part of population, so called take-nothing stratum and taking all samples from take-all stratum.

In this paper for the estimation of the population total, we suggest a composite estimator which is obtained by use of the survey results of take-nothing stratum and a sample sub-stratum. Small simulation studies are conducted to compare the performances of some estimators including the composite estimator suggested in this study. Also for real data analysis, we use the Korea briquette consumption survey data.

Key words : best linear unbiased predictor(BLUP), Lavallee-Hidiroglou algorithm, ratio estimator, take-nothing stratum

1. Introduction

The cut-off sampling has been widely used for a highly skewed population like a business survey by discarding a part of population, so called take-nothing stratum and taking all samples from take-all stratum. Then the estimated population total is obtained by the summation of the estimated totals of the take-all stratum, sample stratum and take-nothing stratum.

In a business survey, the precision of the estimated total might be improved by the conduction of census for the take-all stratum composed of the large size companies. While in some business surveys the precision might be improved by excluding take-nothing stratum because of difficulty in survey and the cost. However in some cases since the precision of an estimator for the take-nothing stratum could greatly affect that of the population total, it is necessary to increase the precision for the take-nothing stratum.

Several methods in order to increase the precision of the estimation for the take-nothing stratum have been suggested with auxiliary information or administrative data. Recently Hwang and Shin (2012) suggested a composite estimator which uses information of the take-nothing stratum and the sample stratum.

In this paper, we suggest a composite estimator for the total of take-nothing stratum which is obtained by use of the survey results of the take-nothing stratum and the sample sub-stratum, a part of the sample stratum. For this, we divide sample stratum into H sub-strata. There are some stratifying methods to divide a population into H sub-populations and in this paper, the well-known L-H(Lavallee-Hidiroglou) algorithm is used. Then we choose one sample sub-stratum, for example i-th sub-stratum, which is the most correlated with the take-nothing stratum. Then we obtain a composite estimator for the total of take-nothing stratum combining information of the only i-th sub-stratum and the take-nothing stratum.

Section 2 explains some notations and some composite estimators developed recently and the L-H algorithm. Also the composite estimator suggested in this study is briefly illustrated. In section 3, small simulation studies are conducted to compare

performances of the several estimators illustrated by Hwang and Shin (2012) and the estimator suggested in this study. In section 4, we confirm the efficiency of the suggested estimator in use of a real data, the Korea briquette consumption survey data. And a conclusion is in section 5.

2. The estimator of the population total in the take-nothing stratum

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 $U_I = U_c \cup U_s$. Here U_c is take-all stratum, U_s, sample stratum, U_{SE}, take-nothing stratum, but few samples surveyed, and U_I is the inclusion stratum. Of course U_s can be divided into H sub-strata, U_h, and $U_s = \bigcup_{h=1}^{H} U_h$. Then the estimate of the population total t_y is calculated by the summation of each estimated total of divided stratum defined by

$$t_y = t_{yC} + t_{yS} + t_{ySE},\ t_{yI} = t_{yC} + t_{yS}$$

where t_{yC}, t_{yS}, t_{ySE} and t_{yI} are the totals of each stratum respectively. Also, given auxiliary variable x, the total of the population and each stratum for x is expressed as t_x, t_{xC}, t_{xS}, t_{xSE} and t_{xI}.

2.1 Previous estimator of the total

In this paper, the estimators of the population total are the same as those suggested in Hwang and Shin (2011). We briefly explain them in this section.

1) Sarndal-Swansson-Wretman(SSW) Estimator

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

$$\hat{t}_y^{SSW} = \hat{R}_{yxI} t_x \quad \text{here, } \hat{R}_{yxI} = \hat{t}_{yI}/\hat{t}_{xI} \tag{2.1}$$

2) Composite estimators

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

$$\hat{t}_{ySE}^{MODI-SSW} = \left(\alpha^{[1]} \frac{\hat{t}_{ySE}}{\hat{t}_{xSE}} + (1-\alpha^{[1]}) \frac{\hat{t}_{yI}}{\hat{t}_{xI}}\right) t_{xSE} \tag{2.2}$$

This estimator is obtained by combining the estimator of based on SSW, $\hat{t}_{ySE}^{SSW} = (\hat{t}_{yI}/\hat{t}_{xI}) t_{xSE}$ with the ratio estimator $\hat{t}_{ySE}^{Ratio} = \frac{\hat{t}_{ySE}}{\hat{t}_{xSE}} t_{xSE}$ using few samples in the take-nothing stratum.

Also, Hwang and Shin(2012) suggested composite estimators using the BLUP(best linear unbiased predictor) for the total of the stratum SE, $\hat{t}_{ySE}^{BLUP}$. In that paper, for the total of the stratum SE, two composite estimators are suggested as following in (2.3) and (2.4).

$$\hat{t}_{ySE}^{MODI-BLUP} = \hat{R}_{SE}^{MODI-BLUP} t_{xSE} = \left(\alpha^{[2]} \frac{\hat{t}_{ySE}}{\hat{t}_{xSE}} + (1-\alpha^{[2]}) \frac{T_{yS_I}}{T_{xS_I}}\right) t_{xSE} \tag{2.3}$$

$$\hat{t}_{ySE}^{MODI-BLUPA} = \hat{R}_{SE}^{MODI-BLUPA} t_{xSE} = \left(\alpha^{[3]} \frac{\hat{t}_{ySE}}{\hat{t}_{xSE}} + (1-\alpha^{[3]}) \frac{\hat{t}_{yS}}{\hat{t}_{xS}}\right) t_{xSE} \tag{2.4}$$

where the definitions of T_{yS_I}, T_{xS_I}, $\hat{t}_{ySE}$, $\hat{t}_{xSE}$, $\hat{t}_{ySI}$, $\hat{t}_{xSI}$ are the same as those in Hwang and Shin(2012). Also the weight α in (2.2), (2.3) and (2.4) can be calculated by using MSE or variance of each estimator. For example $\alpha^{[1]}$ can be calculated using the following.

$$\hat{\alpha}^{[1]} = \frac{MSE(R_{SE}^{SSW})}{MSE(R_{SE}^{MODI}) + MSE(R_{SE}^{SSW})} \approx \frac{Var(R_{SE}^{SSW})}{Var(R_{SE}^{MODI}) + Var(R_{SE}^{SSW})} \qquad (2.5)$$

Here $R_{SE}^{MODI} = \hat{t}_{ySE}/\hat{t}_{xSE}$ and $R_{SE}^{SSW} = \hat{t}_{yI}/\hat{t}_{xI}$. See more details in Rao (2003).

2.2 Algorithm for stratification

For the heavily skewed population, there are some algorithms to divide the population into the take-all stratum and H sub-strata to survey. In this study we use the L-H algorithm suggested by Lavallee and Hidiroglou (1988). See more details in Lavallee and Hidiroglou (1988).

2.3 Suggested estimators of the total sum

The composite estimators of the total of the stratum SE, $\hat{t}_{ySE}$, suggested in session 2.1, are the linear combined estimators with the estimated total of the stratum SE and that of the stratum I (or the stratum S). When we estimate the total of the stratum SE, $\hat{t}_{ySE}$, we could obtain better results by selectively using the partial information of the stratum I instead of using the whole information of the stratum I. The partial information could be obtained from the closer part of sample stratum to take-nothing stratum.

For this reason we divide the sample stratum into H sub-strata. To divide the sample stratum, L-H stratification algorithm is used. Among the divided H strata, the closer sub-stratum to the take-nothing stratum is assumed to have more similar characteristics. Therefore, to estimate the total of the stratum SE, the method using only the closest sub-stratum to the take-nothing stratum is suggested. That is, we assume that for $S = \cup_{h=1}^{H} S_i$, S_1 is the nearest sample stratum to the take-nothing stratum and the newly suggested estimator of the total of stratum SE using the information of only S_1, is following.

$$\hat{t}_{ySE}^{MODI-BLUPN} = \left(\alpha^{[4]}\frac{\hat{t}_{ySE}}{\hat{t}_{xSE}} + (1-\alpha^{[4]})\frac{\hat{t}_{yS_1}}{\hat{t}_{xS_1}}\right) t_{xSE} \qquad (2.6)$$

3. Simulation study

To compare the efficiency of newly suggested estimators in Session 2.3, we conducted a small simulation study. The simulation method used in this paper is the same as that in Lee et al. (1995). Here we consider four types of underlying population structure. The first data set is a ratio type that a linear function of an auxiliary variable x_k and an interesting variable y_k passes through the origin. The second data set is a regression type with positive intercept, the third data set stands for a convex function, and the fourth data set stands for a concave function. For dividing sample stratum into sub-strata, we use L-H algorithm. Also, the boundary of take-nothing stratum is determined by top 80%. Also we use $N = 5{,}000$ and $n = 500$ for the population and sample size respectively. Also the sample size in the stratum SE is $n_{SE} = 10$.

We use three comparison statistics, bias, relative bias(rbias) and root mean square

error(RMSE). Following Table 3.1 shows the results. Here SSW means Sarndal-Swansson-Wretman estimator, M-S, M-B, M-BA are the estimators of (2.2), (2.3) and (2.4) respectively. Also M-BN_* stands for the composite estimator suggested in this paper. For example M-BN_2 is the composite estimator obtained by using the closest sub-stratum from two sample sub-strata to the take-nothing.

Table 3.1 Simulation results with $n_{SE} = 10$

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
	rmse	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
	rmse	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
	rmse	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
	rmse	46573	27320	29543	26767	20398	19667	25084

Table 3.1 shows that using MSE criterion, M-BN_3 composite estimator gives the best results except the ratio-type population. For the ratio-type population, M-B estimator gives the best result. Also in bias results, M-BN_4 is the best.

4. Real data analysis

For real data analysis, total sale amount and number of sales of about 1,600 delivery companies in 2012 Korea briquette consumption survey are used for analysis. 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 explained in section 2. To divide population into strata, we use L-H algorithm and the sample stratum is divided into H sub-strata. The cur-off point to separate sample stratum and SE stratum, we use the 80% point in population total sale amount. The results are tabulated in Table 4.1.

As you can see Table 4.1, SSW shows the worst results in all comparison statistics. For the case of $n_{SE} = 10$, M-BN_3 gives the best results using MSE criterion. On the other hand, for the case of $n_{SE} = 20$, M-BN_4 shows the best results in MSE. Also in bias case, M-BN_4 is the best for $n_{SE} = 10$ and M-BN_3 is the best for $n_{SE} = 20$. Therefore, we can conclue that the composite estimator developed in this paper gives better results than the others.

Table 4.1 Korea briquette consumption survey results

		Estimation methods						
		SSW	M-S	M-B	M-BA	M-BN-2	M-BN-3	M-BN-4
$n_{se} = 10$	bias	330E5	84E5	78E5	76E5	43E5	31E5	30E5
	rbias(%)	0.0654	0.0167	0.0155	0.0152	0.0086	0.0063	0.0061
	rmse	360E5	204E5	221E5	181E5	154E5	144E5	153E5
$n_{se} = 20$	bias	331E5	68E5	59E5	66E5	36E5	24E5	31E5
	rbias(%)	0.0656	0.0135	0.0117	0.0132	0.0072	0.0049	0.0062
	rmse	365E5	178E5	186E5	166E5	148E5	149E5	147E5

5. Conclusion

In this study we suggest a composite estimator obtained by combining the information of a selected sample sub-stratum and take-nothing stratum. In order to get information of take-nothing stratum, we survey few samples from that stratum. Small simulation study shows that the composite estimator suggested in this study is very promising to improve the precision of the estimated total. Also the real data analysis confirms that result.

Acknowledgement

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)

References

[1] 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, 6, 1115-1128.

[2] 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, 1, 13-22,

[3] 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.

[4] Lavallee, P. and Hidiroglou, M.(1988) "On the stratification of skewed populations", *Survey Methodology*, **14**, 33-43.

[5] 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.

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

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