Compare the evaluating methods of variance of estimate on complex two-phase sampling with different sampling units on each phase

Young Jin*
Statistical research institute, Daejeon, Rep. of Korea jinyknk@gmail.com

Sang Eun Lee
University of Kyunggi, Suwon, Rep. of Korea sanglee62@kgu.ac.kr

Key il Shin
Hankuk University of Foreign Studies, Seoul, Rep. of Korea
keyshin@hufs.ac.kr

Abstract

The known estimations of complex two-phase sampling is commonly used for the sampling units in the first phase and the units of the second phase are the same. In this study, we considered the sampling scheme where the sampling units in the first phase are clusters and the units of the second phase are not cluster anymore which are the list samples. For example, in first phase, cluster's are selected as sampling units and before selecting sample in second phase frame, make the listing sampling frame using the elements in clusters from the first phase, after that, in second phase, elements are the sampling units. In this paper, estimate on complex two-phase sampling with different sampling units on each phase are presented. Also, the calculating methods of the variance which are exact and jackknife methods are compared. For the simulations, Farm household economy survey (FHES) in 2010 and Agriculture survey (AS) in 2011 are used.

Key Words: Jackknife, agriculture sampling, master sample, two-stage cluster sampling

1. Background

In every 5 years, we have sampled farm households using stratified-cluster sampling for Farm household economy survey (FHES) with 2010 Census of Agriculture (CA), in Rep. of Korea.

Although the period's differences between survey period for CA and sampling design for FHES are about 2-3 years, the properties (e.g. non-farm household, farming type) of the sampling units have been changed. There are caused for rapidly changing of rural environment and multi-agriculture patterns.

Therefore, we propose complex two-phase sampling for FHES using the most recent information of sampling units.

We have sampled the farm households from the master sample (the first phase sample) with Agriculture survey (AS) in 2011. In first phase, the sampling units are clusters and the units of the second phase are farm households from 2011 AS.

On the study, the analytical variance using the classic method tends to underestimate the variance of the property, since it ignores the clusters underlying in the sampling scheme. (Robothama, Young and Saavedra-Nievas, 2008).

In this study, it is the purpose to derive the estimator on complex two-phase sampling and compare the evaluating methods of variance of estimator between complex two-phase sampling and jackknife method.

2. Estimate of Complex two-phase sampling

In first phase consisting of at least two-stages; a first-stage for obtaining a stratified regional sample and a second-stage to obtain a cluster sample, which is constructed from farm-households. In the second phase these farm households form clustered sampling units are restructured by list sampling frame and stratified by farming type and a final sample of farm households is selected for income, assets and liabilities using stratified systematic sampling method.

An approach is presented in this paper that in the first phase. Overall, the complex two-phase design is similar to two-stage cluster sampling. So, the variance estimator for complex two-phase sampling is derived by the application of two-stage cluster sampling.

2.1 Estimates

A mean estimator of the national farm income given by

$$\hat{\bar{Y}}_M = \frac{1}{M}\frac{N}{n}\sum_{h=1}^{L}\sum_{j=1}^{n_h^{(2)}}\frac{M_j^*}{m_j^*}\frac{m_h^{(1)}}{m_h^{(2)}}\sum_{k=1}^{m_{hj}^{(2)}}Y_{hjk}^{(2)} = \frac{1}{M}\frac{N}{n}\sum_{h=1}^{L}\sum_{j=1}^{n_h^{(2)}}\frac{M_{hj}^{*(1)}}{m_{hj}^{*(2)}}\sum_{k=1}^{m_{hj}^{(2)}}Y_{hjk}^{(2)} \quad (1)$$

where N: no. of cluster, M: no. of farm households
h: strata, j:cluster, k:individual farm household
(1): the first phase sample, (2): the second phase sample

An unbiased variance estimator of the national farm income given by

$$\hat{Var}(\hat{\bar{Y}}_M) = v(\hat{\bar{Y}}_M^{(2)}) = \frac{N^2}{M^2}\frac{1-f_1}{n}\frac{\sum_{j=1}^{n}\left(\hat{\bar{Y}}_j^* - \hat{\bar{Y}}_N\right)^2}{n-1} + \frac{1}{M^2}\frac{N}{n}\sum_{j=1}^{n}M_j^{*2}\frac{s_{2j}^2}{m_j^{*(1)}}(1-f_{2j})$$

$$+\frac{1}{M^2}\frac{N^2}{n^2}\sum_{h=1}^{L}\left\{\frac{1-f_1^{(2)}}{n_h^{(2)}}\frac{\sum_{j=1}^{n_h^{(1)}}\left(\hat{M}_{hj}^{*(2)}\right)^2\left(\hat{\bar{Y}}_{hj}^{*(2)} - \hat{\bar{Y}}\right)^2}{n_h^{(1)}-1} + \frac{1}{n_h^{(1)}}\frac{1}{n_h^{(2)}}\sum_{j=1}^{n_h^{(1)}}\left(\hat{M}_{hj}^{*(2)}\right)^2\frac{s_{2hj}^{2(2)}}{m_{hj}^{(2)}}(1-f_{2hj}^{(2)})\right\} \quad (2)$$

where

$$f_1 = \frac{n}{N},\quad \hat{\bar{Y}}_j^* = \hat{M}_j^*\bar{y}_j^*,\quad \hat{\bar{Y}}_N = \frac{1}{n}\sum_{j=1}^{n}\hat{\bar{Y}}_N^*,\quad s_{2j}^2 = \frac{1}{m_j^{*(1)}-1}\sum_{k=1}^{m_j^{*(1)}}(y_{jk}^{(1)} - \hat{\bar{Y}}_j^{*(1)})^2,\quad f_{2j} = \frac{m_j^{*(1)}}{M_j^*},$$

$$f_1^{(2)} = \frac{n_h^{(2)}}{n_h^{(1)}},\quad \hat{M}_j^{*(2)} = \frac{M_j^*}{m_j^*}\frac{m_h^{(1)}}{m_h^{(2)}}m_{hj}^{(2)} = \frac{m_h^{(1)}}{m_h^{(2)}}m_{hj}^{(2)},\quad \hat{\bar{Y}}_{hj}^{*(2)} = \frac{1}{m_{hj}^{(2)}}\sum_{k=1}^{m_{hj}^{(2)}}y_{hjk}^{(2)},\quad \hat{\bar{Y}}_h = \frac{1}{n_h^{(2)}}\sum_{k=1}^{m_h^{(2)}}y_{hjk}^{(2)},$$

$$s_{2hj}^{2(2)} = \frac{1}{m_j^{*(2)}-1}\sum_{k=1}^{m_j^{*(2)}}(y_{jk}^{(2)} - \hat{\bar{Y}}_j^{*(2)})^2,\quad f_{2hj}^2 = \frac{m_{hj}^{(2)}}{m_{hj}^{*(1)}}$$

2.2 Methods of evaluating the variance

We present the evaluating about the variance using the suggested variance estimator and jackknife methods.

2.2.1 Exact estimate

Variance estimator for complex two-phase sampling is calculated using equation (1) and (2).

2.2.2 Jackknife estimate

Jackknife variance estimator is as follows,

$$v_j = \frac{R-1}{R}\sum_{r=1}^{R}(t_r - t)^2 \qquad (3)$$

3. Simulation

A simulation study is used to evaluate between the suggested variance estimator and jackknife. About 74,300 simulated data for sampling units in the first phase are generated with 2,646 farm household data (income, asset and liability) of 2010 FHES and 75,000 surveyed farm data of 2011 AS.

Farm households from the generated pseudo first phase clustered samples are stratified by farming type and 2,646 farm households are sampled by the units of the second phase using stratified systematic sampling. These procedures repeat 500 times without replacement. After that, we calculate the mean and relative root mean square error (RRMSE).

RRMSE is as follows,

$$RRMSE = RMSE/\theta \times 100 \qquad (3)$$

where $RMSE = \sqrt{\frac{1}{S}\sum_{s=1}^{S}(\hat{\theta}_s - \theta)^2}$

(S : *No. of the sample selection*), $\theta = \sum_{s=1}^{S}\hat{\theta}_s \Big/ S$)

Also, we estimate the variance and relative standard error (RSE) for complex two-phase sampling and jackknife method.

It is observed that RSEs of complex two-phase sampling for national farm income, assets and liabilities are lower in most of the variables when compared with RRMSEs and RSEs of jackknife (Table 1). However, RSEs of jackknife are similar to RRMSE in all the variables and RSEs of F.H. income and assets are lowest (2.4 and 2.7).

When compared to the regional estimators using complex two-phase sampling and Jackknife method, some RRMSEs using 500 times sampling data and some RSEs of jackknife method (RSE_jkw) show similar trends and are

regionally stable. It is observed that some RSEs of complex two-phase sampling (RSE_cd) show the lowest trends among them but are regionally unstable for very large values (Fig. 1).

Table 1

Estimated relative root mean square error (RRMSE), standard error (S.E.) and relative standard error (RSE) using complex two-phase sampling and jackknife method

	Estimators				
		Sampling Design		Jackknife	
Variables	RRMSE	S.E.	RSE	S.E.	RSE
1.Farm Household Income	2.7	952	2.9	783	2.4
2.Agriculture Income	5.2	471	4.4	545	5.1
3.Gross Farm Income	4.6	1143	3.8	1194	4.0
4.Non-farm business Income	9.2	543	6.4	779	9.2
5.Assets	2.8	11371	2.9	10410	2.7
6.Liabilities	6.2	1289	4.2	2008	6.5

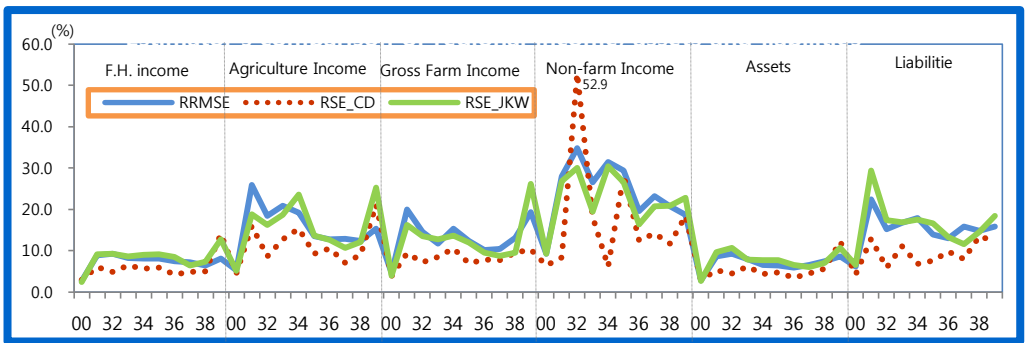

Fig.1. Regional RRMSE and RSE using complex two-phase sampling and jackknife method

4. Conclusion

It is in the empirical study to estimate the variance of the farm household's mean when a complex two-phase sampling scheme is applied. It derives the variance estimator for complex tow-phase sampling where the sampling units in the first phase are clusters and the units of the second phase are the list samples.

Some RSEs of complex two-phase sampling are lowest value but regionally unstable for existing large. Most RSEs of Jackknife method have been similar values with RRMSE and regionally stable.

In complex two-phase sampling, the jackknife method will continue being a good approximation for estimating the variance of the mean.

References

Cochran,W.G., 1977. Sampling Techniques, 3rd. John Wiley, NY.

Robothama, H., Young, Z.I., 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, Fisheries Research 93, 135-139.