JACKKNIFE VARIANCE ESTIMATION FOR TWO-PHASE SAMPLES WITH HIGH SAMPLING FRACTIONS

Hyunshik Lee, Westat; and Jae-Kwang Kim, Hankuk University of Foreign Studies
Hyunshik Lee, Westat, 1650 Research Boulevard, Rockville, Maryland 20850

Key Words: Cluster sampling, Stratification, Simple random sampling, Finite population correction (fpc), Replicate weighting

1. Introduction

The two-phase sample design is often employed in sample surveys for various reasons. It has a long history, first introduced by Neyman (1938). Traditionally, the technique is used to collect some auxiliary data that are not available for the sampling frame from a large first-phase sample to use the data at the second-phase sampling. The technique, sometimes called "double sampling" has many applications in different forms (e.g., Rao, 1973; Cochran, 1977; Breidt and Fuller, 1993; Rao and Sitter, 1995; Hidiroglou and Särndal, 1998; Fuller, 1998).

In this paper, we focus on variance estimation for two-phase sampling with stratified sampling at both phases and with high sampling fractions. Particularly, we are interested in the jackknife technique for the variance estimation. Rao and Shao (1992) proposed a consistent jackknife variance estimator for the reweighted expansion estimator (REE) in the context of hot deck imputation treating the respondents as the second-phase sample. Kott and Stukel (1997) considered the same problem and concluded that the jackknife variance estimator works well for the reweighted expansion estimator (REE) if the first-phase sampling is with replacement. Rao and Sitter (1997) considered a case of two-phase sampling with nontrival sampling rate, where the first-phase strata are the same as the second-phase strata. Binder et al. (2000) studied the variance estimation problem for a similar two-phase sample design but without the restriction of the with-replacement sampling assumption. However, they used the Taylor linearization method. Also, Kim, Navarro, and Fuller (2000) tackled the same problem using the jackknife variance estimator but with with-replacement sampling assumption at the first-phase.

In this paper, our main concern is how to properly incorporate the finite population correction (fpc) in the jackknife variance estimator for a two-phase sample, where the first-phase sampling is a stratified cluster sampling and the second-phase is simple random sampling of the elements from the strata of elements defined within the first-phase strata.

In the next section, our proposed jackknife variance estimator is presented. In Section 3, an application of the proposed variance estimator is discussed. In the final section, some concluding remarks are given.

2. Proposed Jackknife Variance Estimator

The first-phase sample is selected by stratified cluster sampling. Let there be H first-phase strata and let stratum h have N_h clusters from which n_h clusters are sampled by simple random sampling with a sampling fraction of f_{h1} (i.e., $f_{h1} = n_h / N_h$).

Elements in the sampled clusters are stratified using the information collected from the clusters and a simple random sample of elements is selected from each second-phase stratum.

Let w_{hij} be the first-phase weight of element j of cluster i in stratum h $\left(w_{hij} = n_h^{-1} N_h\right)$ and y_{hij} denote the value of survey variable y for element j in cluster (hi). The cluster total of y for cluster (hi) is given by $y_{hi} = \sum_j y_{hij}$ and the stratum mean of cluster total y_{hi} in stratum h by $\bar{y}_h = n_h^{-1} \sum_{i=1}^{n_h} y_{hi}$.

Suppose for a moment that there is no second-phase sampling, then the total Y of the y-variable would be estimated by

$$\hat{Y}_1 = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_j w_{hij} y_{hij} = \sum_{h=1}^{H} \left(\sum_{(ij) \in A_h} w_{hij} y_{hij} \right), \qquad (1)$$

where A_h is the set of indices of the elements selected into the first phase sample in stratum h. Without loss of generality, we assume that the first n_h clusters are selected in stratum h. A correct variance estimator would then be

$$V\left(\hat{Y}_1\right) = \sum_{h=1}^{H} \frac{1 - f_{h1}}{n_h \left(n_h - 1\right)} \sum_{i=1}^{n_h} \left(y_{hi} - \bar{y}_h\right)^2 . \qquad (2)$$

The jackknife estimator for (2) is given by

$$v_j\left(\hat{Y}_1\right) = \sum_{h=1}^{H} \sum_{j=1}^{n_h} \left(\hat{Y}_1^{(hj)} - \hat{Y}_1 \right)^2 , \qquad (3)$$

where

$$\hat{Y}_1^{(st)} = \sum_{h=1}^{H}\left(\sum_{(ij)\in U_h} w_{hij}^{(st)} y_{hij}\right), \qquad (4)$$

is a replicate analogue of (1) based on replicate weights defined by

$$w_{hij}^{(st)} = \begin{cases} \delta_h w_{hij} & \text{if } h = s \text{ and } i = t, \\ \dfrac{\left(n_h - \delta_h\right) w_{hij}}{n_h - 1} & \text{if } h = s \text{ and } i \neq t, \\ w_{hij} & \text{if } h \neq s, \end{cases} \qquad (5)$$

with $\delta_h = 1 - \sqrt{\left(1 - f_{h1}\right)\dfrac{n_h - 1}{n_h}}$ for $i = 1, 2, ..., n_h$, and $h = 1, 2, ..., H$.

Note that the jackknife factors and the fpc factors are already incorporated in the above replicate weights. This incorporation of the factors in the replicate weights has an advantage over the traditional way of defining the replicate weights since it abolishes the need to specify such factors separately in the replication variance estimation software such as WesVar (2000).

Now, consider the second-phase sampling. Let there be G_h second-phase strata in first-phase stratum *h* and let A_{hg} be the set of elements in second-phase stratum *g* in stratum *h*, from which a second-phase simple random sample a_{hg} is selected. Let also M_{hg} denote the size of A_{hg} and m_{hg} denote the size of a_{hg}.

To estimate *Y*, two estimators are available. One is the double expansion estimator (DEE) and the other is the reweighted expansion estimator (REE) (Kott and Stukel, 1997; Kim, Navarro, and Fuller, 2000). For the design described above, the two estimators are identical. The REE is given by

$$\begin{aligned} \hat{Y}_2 &= \sum_{h=1}^{H}\sum_{g=1}^{G_h}\left(\sum_{(ij)\in A_{hg}} w_{hij} \frac{\sum_{(ij)\in a_{hg}} w_{hij} y_{hij}}{\sum_{(ij)\in a_{hg}} w_{hij}}\right), \\ &=: \sum_{h=1}^{H}\sum_{g=1}^{G_h}\left(\sum_{(ij)\in a_{hg}} \alpha_{hij} y_{hij}\right), \end{aligned} \qquad (6)$$

where for $(ij) \in a_{hg}$

$$\alpha_{hij} = \frac{\sum_{(ij)\in A_{hg}} w_{hij}}{\sum_{(ij)\in a_{hg}} w_{hij}} w_{hij} = \frac{M_{hg}}{m_{hg}}\frac{N_h}{n_h}, \qquad (7)$$

is the two-phase sampling base weight. This is also the DEE weight because the first-phase sampling weights are the same within each stratum. The numerator $\sum_{(ij)\in A_{hg}} w_{hij}$ part is subject to the first-phase sampling variability only, but the denominator $\sum_{(ij)\in a_{hg}} w_{hij}$ is also subject to the second-phase sampling variability. The first-phase sampling variability comes from estimating the size of second-phase stratum. Within a particular second-phase stratum *hg*, the overall two-phase sampling design for the stratum is equivalent to the stratified random sample of size m_{hg} from the finite population of size $M_{hg} f_{h1}^{-1}$, where f_{h1} is the sampling rate for the first-phase sampling. The overall sampling rate of the two-phase sample is $f_{h1} f_{hg2}$ with $f_{hg2} = M_{hg}^{-1} m_{hg}$. Thus, the proposed replication method applies different replication weighting to the two weights resulting from the two phases of sampling in order to incorporate differential sampling rates.

The proposed jackknife variance estimator of the estimator (6) is then given by

$$v_J\left(\hat{Y}_2\right) = \sum_{s=1}^{H}\sum_{t=1}^{n_s}\left(\hat{Y}_2^{(st)} - \hat{Y}_2\right)^2, \qquad (8)$$

where $\hat{Y}_2^{(st)}$ is defined as

$$\begin{aligned} \hat{Y}_2^{(st)} &= \sum_{h=1}^{H}\sum_{g=1}^{G_h}\left(\sum_{(ij)\in A_{hg}} w_{hij}^{(st)} \frac{\sum_{(ij)\in a_{hg}} w_{hij,g}^{*(st)} y_{hij}}{\sum_{(ij)\in a_{hg}} w_{hij,g}^{*(st)}}\right), \\ &= \sum_{h=1}^{H}\sum_{g=1}^{G_h}\left(\sum_{(ij)\in a_{hg}} \alpha_{hij}^{(st)} y_{hij}\right), \end{aligned} \qquad (9)$$

where $w_{hij}^{(st)}$ is defined in (5) and $w_{hij,g}^{*(st)}$ is defined by

$$w_{hij,g}^{*(st)} = \begin{cases} \delta_{hi,g} & \text{if } h = s \text{ and } i = t, \\ \dfrac{m_{sg} - \delta_{hi,g} m_{st,g}}{m_{sg} - m_{st,g}} & \text{if } h = s \text{ and } i \neq t, \\ 1 & \text{if } h \neq s, \end{cases} \qquad (10)$$

with $m_{st,g}$ being the number of the second-phase sample units in (*st*)-th cluster and the *g*-th second-phase stratum, and

$$\delta_{hi,g} = 1 - \left[\frac{1 - f_{h1} f_{hg2}}{1 + \sum_{(hi') \neq (hi)} \left(\frac{m_{hi',g}}{m_{hg} - m_{hi',g}} \right)^2} \right]^{1/2} .$$

Note that (10) is much more complex than (5) because multiple elements are deleted when a replicate is formed.

Further weight adjustment for nonresponse adjustment and/or poststratification can be performed in the usual way using the replicate weights $\alpha_{hij}^{(st)}$.

The idea of applying different replication weights for the weight components corresponding to the different phases of sampling was first proposed by Fuller in the doctoral dissertation of Kim (2000), where the first-phase is simple random sampling and the second-phase is stratified random sampling.

3. An Application

The SPeNSE sample design is a two-phase design where the first-phase sample was selected by the stratified simple random sampling of educational agencies (clusters) and the second-phase sample was selected from the service provider pool assembled from selected agencies. The participating agencies from the first-phase sample provided rosters of service providers by personnel type. The roster pool is stratified by personnel type within each first-phase stratum and a simple random sample of providers was selected from each second-phase stratum.

The sampling rates for some first-phase and second-phase strata were very high and, thus, there was a concern on over estimation of the variance if the fpc's were not incorporated in the variance estimation.

On the other hand, if the fpc is applied naively to the first-phase sample, the variance is underestimated since the first phase fpc is also applied to the second phase sampling variance unnecessarily. This was the motivation of this research and the correct jackknife variance estimator proposed here comes in between the two.

In SPeNSE the former was implemented to be conservative rather than liberal. It was our desire to make sure that the conservatism is not excessive by comparing the actually used variance estimates with the correct ones. To this end, the magnitude of the positive bias of the conservative variance estimator was investigated by comparing the two variance estimators for ten key variables.

The replicate weights calculated as described in the previous section are further adjusted for second-phase nonresponse and poststratification in the same way as done in the SPeNSE final weights. The correct variance estimates are then calculated for the same ten key variables and compared. The comparison shows that the correct variance estimates are fairly close to the conservative ones that were actually used in SPeNSE. The relative difference (RELDIF) between the conservative and the correct is defined by

$$\text{RELDIF} = 100 \times \frac{\text{DIF}}{\text{Current STD estimate}},$$

where

$$\text{DIF} = \text{Current STD estimate} - \text{Correct STD estimate}.$$

The absolute value of RELDIF is mostly less than 10 percent. However, the correct variance estimates can even be larger than the conservative due to randomness. It should be noted that the correctness of the proposed variance estimator is in expectation. This means that, although the conservative variance estimator overestimates the true variance in expectation, a particular estimate can be smaller than the correct one, which is correct in expectation.

There are only three cases with RELDIF over 20 percent, two of which have the cell size of 1 each. The other case has the cell size of 163. Considering the cell sample size, this is the only case with unusually large RELDIF. Since the results of the comparisons for 10 variables are too voluminous to present here, only one sample table is presented in Table 1.

This study confirmed that the conservatism of the variance estimator used for SPeNSE is minor.

4. Concluding Remarks

Two-phase sampling or multi-phase sampling method is often used in practice for various reasons. Jackknife variance estimation is very flexible in incorporation of the design weights, nonresponse adjustments, and complex form of statistics of interest. However, when it is used for a multi-phase sample, special care is needed because replicate weighting should be done in phases incorporating the multi-phase sampling features in the variance estimation. In this paper, we propose a jackknife variance estimator that is particularly suitable when sampling at both phases is done by stratified sampling with high sampling fractions. The proposed method is also applicable for variance estimation of imputed data in the similar context as for Rao and Shao (1992) but when fpc is nonnegligible.

Table 1. Comparison of standard errors estimated by conservative and correct variance estimators for the variable SOVERALL (overall performance)

Personnel	Level* of	Estimate	Estimated standard error			Cell	Marginal
type	SOVERALL	(percent)	Conservative	Correct	RELDIF	sample size	sample size
1	1	.	.	.	.	0	870
1	2	0.58	0.28	0.26	7.86	6	870
1	3	12.26	2.96	2.49	15.92	97	870
1	4	66.90	3.58	3.27	8.60	573	870
1	5	20.25	2.50	2.37	5.17	194	870
2	1	0.04	0.05	0.04	26.92	1	1040
2	2	0.58	0.36	0.36	0.00	5	1040
2	3	15.88	3.45	3.25	5.83	136	1040
2	4	64.58	3.82	3.52	7.89	657	1040
2	5	18.92	2.27	1.95	14.43	241	1040
3	1	0.06	0.07	0.04	43.66	1	854
3	2	2.73	0.91	0.92	-0.66	15	854
3	3	19.90	1.97	1.93	1.68	173	854
3	4	58.12	2.18	2.03	6.57	496	854
3	5	19.19	1.92	1.80	6.35	169	854
4	1	0.02	0.02	0.03	-13.04	1	2575
4	2	1.13	0.26	0.25	6.06	30	2575
4	3	16.30	1.02	0.96	6.35	421	2575
4	4	62.17	1.56	1.54	1.03	1620	2575
4	5	20.37	1.52	1.45	4.41	503	2575
5	1	0.13	0.16	0.14	7.10	1	869
5	2	0.70	0.31	0.29	6.41	7	869
5	3	14.41	1.54	1.61	-4.94	118	869
5	4	69.38	1.83	1.88	-2.78	598	869
5	5	15.38	1.46	1.36	6.45	145	869
6	1	.	.	.	.	0	887
6	2	1.19	0.43	0.40	7.89	10	887
6	3	11.20	1.22	1.22	0.00	99	887
6	4	53.60	1.85	1.86	-0.54	484	887
6	5	34.01	1.86	1.90	-1.72	294	887
7	1	0.16	0.20	0.17	11.73	1	847
7	2	0.58	0.27	0.27	0.74	5	847
7	3	15.86	1.67	1.56	6.81	135	847
7	4	61.76	2.09	1.98	5.21	528	847
7	5	21.64	2.17	2.09	3.37	178	847
MARGINAL	1	0.13	0.15	0.13	11.84	5	7942
MARGINAL	2	0.72	0.22	0.21	1.38	78	7942
MARGINAL	3	15.50	1.32	1.22	7.06	1179	7942
MARGINAL	4	61.23	1.64	1.56	4.94	4956	7942
MARGINAL	5	22.42	1.74	1.67	3.57	1724	7942
Average			1.45	1.37	6.08		

*Note: 1: poor; 2: fair; 3: good; 4: very good; 5: exceptional.

5. References

Binder, D.A., Babyak, C., Brodeur, M., Hidiroglou, M.A., and Jocelyn, W. (2000). Variance estimation for two-phase stratified sampling. *The Canadian Journal of Statistics*, 28, pp. 751-764.

Breidt, K., and Fuller, W.A. (1993). Regression weighting for multi-phase samples. *Sankhya*, 55, pp. 297-309.

Cochran, W.G. (1977). *Sampling techniques*, 3rd edition. New York: Wiley.

Fuller, W.A. (1998). Replication variance estimation for two-phase samples. *Statistica Sinica*, 8, pp. 117-132.

Hidiroglou, M.A., and Särndal, C.E. (1998). Use of auxiliary information for two-phase sampling. *Survey Methodology*, 24, pp. 11-20.

Kim, J.K. (2000). *Variance estimation after imputation*. Ph.D. dissertation, Iowa State University.

Kim, J.K., Navarro, A., and Fuller, W. (2000). Variance Estimation for 2000 Census Coverage Estimates. In *Proceedings of the Section on Survey Research Methods of the American Statistical Association*, pp. 515-520.

Kott, P.S., and Stukel, D.M. (1997). Can the Jackknife Be Used with a To-Phase Sample? *Survey Methodology*, 23, pp. 81-89.

Neyman, J. (1938). Contribution to the theory of sampling human populations. *Journal of the American Statistical Association*, 33, pp. 101-116.

Rao, J.N.K. (1973). On double sampling for stratification and analytic surveys. *Biometrika*, 60, pp. 125-133.

Rao, J.N.K., and Shao, J. (1992). Jackknife variance estimation with survey data under hot desk imputation. *Biometrika*, 79, pp. 811-822.

Rao, J.N.K., and Sitter, R. (1995). Variance estimation under two-phase sampling with application to imputation for missing data. *Biometrika*, 82, pp. 453-460.

Rao, J.N.K., and Sitter, R. (1997). Variance estimation under stratified two-phase sampling with applications to measurement bias. *Survey Measurement and Process Quality*, pp. 753-768.

Westat (2000). *WesVar 4.0: User's Guide*. Rockville: Westat, Inc.