Resampling Variance Estimation for a Two-Phase Sample

Hyunshik James Lee and David A. Marker
Westat, 1600 Research Blvd., Rockville, MD 20850

Abstract

Two-phase sampling is often used in a wide variety of surveys. Variance estimation from a two-phase sample has been a subject of active research. The re-sampling method of variance estimation has been used for this problem. However, the method confronts a challenging problem when the first phase sampling fraction is high. In the extreme (but not uncommon) case some first-phase strata are take-all but there is subsampling at the second phase. This issue is studied for a real survey.

Key Words: Double-expansion estimate, reweighted expansion estimate, jackknife variance estimator, relative bias

1. Introduction

The two-phase or double sampling technique is used for multiple reasons. As one example, nonresponse, which is a perennial problem in sample surveys, is treated as the second-phase sampling under the quasi-randomization framework of Oh and Sheuren (1983), and much research has been done in this area. The two-phase sampling technique was originally proposed by Neyman (1938) for a situation where auxiliary information is not available but can be obtained cheaply (Cochran, 1977, ch. 12, pp 327-358). So a large sample is selected at the first-phase, and auxiliary data are collected for the first-phase sample. A second-phase sample with a smaller sample size is selected from the first-phase sample, and survey data are collected for this sample. Auxiliary information from the first-phase sample is used to improve estimation through ratio or regression estimation with the survey data collected at the second-phase sample.

However, two-phase sampling is also used for other survey situations. The Stormwater Survey of the Environment Protection Agency (EPA) is such an example. This is a survey of construction establishments, which uses stratified equal probability sampling. After a sample was selected, to reduce respondent burden EPA imposed a cap that no more than two establishments should be selected from any multi-establishments company. To implement this cap, two-phase sampling was used, where the second-phase sampling occurred in multi-establishment companies with more than two establishments selected in the first-phase sample. Note that multi-establishment companies cut across the first-phase stratum boundaries.

A two-phase sample must be distinguished from a two-stage sample that selects the second-stage sample in a nested fashion within the first-stage sample units (i.e., primary sampling units). When sampling is done across the cluster, however, it is a two-phase sample, and the variance estimation gets much more complicated than for a two-stage design.

The EPA Stormwater Survey stratified the sampling frame by industry and size. For each industry stratum, three size strata were formed, one of which is certainty or take-all. The other two size strata were take-some strata, some of which were selected with a high sampling rate.

The estimator of a survey variable used in the survey is the re-weighted estimator (REE). Defining the multi-establishment companies with more than two establishments selected in the first-phase sample as the second-phase strata (indexed by h) and utilizing an auxiliary variable (denoted by x), the REE for the population total and mean for a survey variable, y, is given by:

$$\hat{Y} = \sum_{h=1}^{H} \frac{\hat{X}_{h1}}{\hat{X}_{h2}} \hat{Y}_{h2}$$

$$\hat{\bar{Y}} = \hat{Y} \big/ \hat{N}$$

$$\hat{N} = \sum_{h=1}^{H} \hat{N}_{h1}$$

Where $\hat{N}_{h1} = \sum_{i \in h1} w_i$, $\hat{X}_{h1} = \sum_{i \in h1} w_i x_i$, $\hat{X}_{h2} = \sum_{i \in h2} w_i w_{i2} x_i$, $\hat{Y}_{h2} = \sum_{i \in h2} w_i w_{i2} y_i$, and $w_i$and w_{i2} are the first- and second-phase sampling weights. Symbol $h1$ denotes the second-phase strata, where the second-phase sample $h2$ is selected, that is, $h2 \subseteq h1,$ for all h. Note that the second-phase strata include the first-phase sample, which does not require sub-sampling (i.e., all single establishment companies and multi-establishment companies with no more than two establishments selected in the first-phase sample). Note also that $\hat{X}_{h2}$ and $\hat{Y}_{h2}$ are the double expansion estimators (DEE) for the second phase stratum h.

The issue here is how to estimate the variance of the re-weighted estimators using the jackknife method. It appears easier to address the variance estimation issue for two-phase samples through the Taylor linearization method (see Binder et al., 2000; Kim and Kim, 2007). However, we prefer using the replication method for its advantages over the Taylor method since it can be applied for non-linear statistics without linearization

2. Jackknife Variance Estimators for Two-Phase Sampling

Two jackknife variance estimators, which are suitable for our situation, have been proposed in the literature. One is by Shao and Thompson (2009), who proposed a jackknife variance estimator using the model assisted approach for an industry survey with a similar design (take-all and take-some strata) as the EPA's Stormwater Survey. The second is by Kim and Yu (2011), who proposed a jackknife variance estimator for two-phase samples with high sampling rates but without certainties.

Assuming a super-population model m, Shao and Thompson expressed the variance of the estimator, $\hat{Y}$ of the population total, Y, of a survey variable, y, as follows:

$$\begin{aligned} V_{m,s}(\hat{Y} - Y) &= E_m[V_s(\hat{Y})] + V_m[E_s(\hat{Y}) - Y] \\ &= V_1 + V_2 \end{aligned}$$

where s denotes the sampling distribution. The first component is estimated by $\hat{V}_s(\hat{Y})$, which can be obtained easily with finite population correction (FPC) incorporated using the usual jackknife variance estimator. The second component is more difficult to estimate but Shao and Thompson proposed to estimate this approximately by the variance of

$$\breve{Y} = \sum_{h=1}^{H} \sqrt{1-\frac{X_{ch2}}{X_{ch1}}}\left(\frac{X_{ch1}}{X_{ch2}}\right) Y_{ch2}$$

where ch are second phase strata crossed by combined first-phase certainty strata. In doing so, they assume that contributions from take-some strata to V_2 is negligible. Combining the variance estimators for the two variance terms, the Shao-Thompson jackknife variance estimator is given by

$$\hat{V}_{ST} = \sum_{l} \frac{n_l - 1}{n_l} \sum_{j \in u_l} \left(\tilde{Y}_{(lj)} - \frac{1}{n_l} \sum_{k \in s_l} \tilde{Y}_{(lk)} \right)^2$$

where

$$\tilde{Y}_{(lj)} = \begin{cases} \breve{Y}_{(lj)} & \text{if 1st phase stratum } l \text{ is certainty} \\ \hat{Y}_{(lj)} \sqrt{1 - \dfrac{n_l}{N_l}} & \text{otherwise} \end{cases}$$

Their proposed method was meant to handle a situation where a business survey sample is selected by a take-some/take-all design but it suffers from nonresponse, which can be treated as the second-phase sampling.

The second method Kim and Yu (2011) proposed followed the approach that Kim, Navarro, and Fuller (2006) took for jackknife variance estimation for a two-phase sample – this variance estimator is referred to as the KNF estimator and given by:

$$\hat{V}_{KNF} = \sum_{k} c_k \left(\hat{Y}^{(k)} - \hat{Y} \right)^2$$

where $c_k = (1 - f_k)(n_k - 1)/n_k$ and $\hat{Y}^{(k)}$ is a replicate estimate for the population total. However, the KNF jackknife variance estimator is consistent for a two-phase sample only when the sampling rate is small, and thus, it is inconsistent when the sampling rate is high. Kim and Yu addressed this issue and provided a bias-corrected estimator for high sampling rates. Their estimator has the same form as the KNF estimator as shown below:

$$\hat{V}_{KY} = \hat{V}^{*}_{KNF} = \sum_{k=1}^{L} c_k \left(\hat{Y}^{*(k)} - \hat{Y} \right)^2$$

where the k-th replicate estimate $\hat{Y}^{*(k)}$ is given by:

$$
\begin{aligned}
\hat{Y}^{*(k)} &= \frac{\sum_i w_i^{(k)} M_{i2}^{(k)} w_{i2} y_i}{\sum_i w_i^{(k)} M_{i2}^{(k)} w_{i2}} \\
M_{i2}^{(k)} &= 1 + (\delta_{ki} - p_k) b_i \\
b_i &= \sqrt{\frac{(1 - w_{i2}^{-1}) w_i^{-1}}{\sum_{k=1}^{L} c_k p_k (1 - p_k)}}
\end{aligned}
$$

and δ_k is Bernoulli random variable with parameter p_k, that is, δ_k is 1 with a probability p_k and 0 with a probability $(1 - p_k)$. The KNF estimator does not have the M-terms (or they are all one), which are used in the Kim-Yu estimator to remove the bias.

Similarly, we can define two versions of the variance estimator for $\hat{\hat{Y}}$.

One problem with the Kim and Yu method is that it cannot handle take-all (certainty) strata at the first-phase. We took an ad-hoc measure by using 0.01 for FPC instead of 0 to avoid eliminating replicates formed for the take-all strata – this is equivalent to assume that the sampling rate is 99 percent instead of 100 percent. This will introduce a bias in the jackknife variance estimator, which depends on the magnitude of the variance coming from the take-all strata through the second-phase sampling.

Both variance estimators have some problems for our situation. The Shao-Thompson estimator ignores variance contributions from high-sampling-rate-take-some strata, and the approximation used in the second term can be non-negligible (as shown by the authors of the paper), whereas the Kim-Yu estimator uses an ad-hoc measure to address the problem with take-all strata. Because of these uncertainties, we conducted a simulation study to make sure these weaknesses do not cause a big problem.

3. Simulation Study

We used the sampling frame for the EPA Stormwater Survey as the population for the simulation study. The sample design stratifies the population into seven industry strata and three size strata, one of which is a take-all, within each industry stratum. We used only five industry strata that consist of residential developers, where the second-phase sampling mostly occurred. The frame size of these five industry strata is 98,521. The sample size of 1,681 was allocated to 15 industry-size strata to meet EPA objectives, resulting in 220 certainties and 1,461 non-certainties selected by an equal probability sampling method within strata with the sampling rates ranging between 0.3 percent and 44 percent. In the second phase, multi-establishment companies with more than two establishments selected at the first phase were subsampled by the probability proportionate to size (PPS) method to select only two establishments.

For the simulation study, we selected 1,000 samples following the sample design. From each sample, we calculated the total estimate for annual sales and three variance estimates (two-versions of the Shao-Thompson and the Kim-Yu). At the end of the simulation, we calculated the following statistics:

- The variance of 1,000 point estimates for the total (and mean) annual sales to be used as the Monte Carlo variance denoted by V

- Three variance estimators:
    - i. V_{J1} – Shao-Thompson's first term
    - ii. V_{ST} – Shao-Thompson
    - iii. V_{KY} - Kim-Yu with $p_k = 0.5$
- Monte Carlo expectations of the variance estimators (i.e., the average of the variance estimators over 1,000 samples) denoted by $\mathrm{E}(\widehat{\mathrm{V}})$
- Relative biases of the three variance estimators defined by

$$\text{Relative Bias } (\hat{V}) = \frac{E(\hat{V})}{V} - 1$$

- 95 percent Confidence Interval Coverage by each of the three variance estimators for the total (or mean) estimate

Table 1 presents the simulation results.

Table 1: Simulation Results - Relative Bias and 95% Confidence Interval (CI) Coverage

Variance Estimators	*Rel. Bias for $\hat{V}(\hat{Y})$*	*CI Coverage*	*Rel. Bias for $\hat{V}(\widehat{\bar{\bar{Y}}})$*	*CI Coverage*
V_{J1}	-0.034	93.3%	-0.031	93.1%
V_{ST}	0.119	94.7%	-0.031	93.1%
V_{KY}	-0.015	93.4%	-0.014	93.8%

The simulation study results can be summarized as follows:

- Both methods work very well for our situation.
- The Kim-Yu method is easier to implement, but using 0.01 as the take-all FPC is arbitrary, it may not be a good choice for other situations.
- The Shao-Thompson method does not include the second term for non-certainty first-phase strata and can underestimate the variance. It could also overestimate when the sample size of the first-phase certainty is small as shown in Shao and Thompson (2009). Such tendency is also shown in our study (see the value for the relative bias for $\widehat{\mathrm{V}}_{\mathrm{ST}}(\widehat{\mathrm{Y}})$).
- For the EPA Stormwater Survey, either can be used, although the Kim-Yu method performed slightly better.
- The first-term estimator of the Shao-Thompson method works quite well, which is surprising. This indicates that the contribution to the total variance due to the second-phase sampling is negligible.

3. Conclusions and Discussion

The two methods we studied work well for the problem we have for both means and totals. Would they work as well for more complex statistics? How would they perform in other situations? These questions require more study, and both methods have their own strengths and weaknesses, which should be weighed carefully to choose for each situation.

Another item for future study is to study the bootstrap method. Saigo (SMJ, 2007) studied the bootstrap method for two-phase sampling with SRS at both phases. We would want to extend his method for other two-phase sampling situations, including PPS sampling.

References

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

Cochran, W.G. (1977). *Sampling Techniques*, 3rd ed. New York: John Wiley.

Kim, J.K., and Kim, J.J. (2007). Nonresponse weighting adjustment using estimated response probability. *The Canadian Journal of Statistics*, 35, 501-514.

Kim, J.K., Navarro, A., and Fuller, W.A. (2006). Replicate variance estimation after multi-phase stratified sampling. *Journal of American Statistical Association*, 101, 312-320.

Kim, J.K., and Yu, C.L. (2011). Replication variance estimation under two-phase sampling. *Survey Methodology*, 37, 67-74.

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

Oh, H.L., Scheuren, F.J. (1983). Weighting adjustment for unit nonresponse. In W.G. Madow, I. Olkin, and D.B. Rubin, eds., *Incomplete Data in Sample Surveys, Volume 2: Theory and Bibliography*, pp. 143-184.

Saigo, H. (2007). Mean-adjusted bootstrap for two-phase sampling. *Survey Methodology*, 33, 61-68.

Shao, J., and Thompson, K.J. (2009). Variance estimation in the presence of nonrespondents and certainty strata. *Survey Methodology*, 35, 215-225.