Variance Estimation for a Small Number of PSUs

Hyunshik Lee
Westat, 1600 Research Blvd., Rockville, MD 20850

Abstract

In some surveys, the number of primary sampling unit (PSU) selected is small but the number of ultimate sampling units is large. The usual consistent variance estimator based solely on the between-PSU variance is not stable because the number of degrees of freedom is small. One alternative is to use a variance estimator that estimates the within-PSU variance by treating the PSUs as strata. It has a larger number of degrees of freedom but underestimates the variance. Combining these two estimators, we can produce a variance estimator that is more stable than the usual consistent variance estimator and less biased than the within-PSU variance estimator. The performance of such a hybrid estimator is demonstrated by simulation using a population similar to the population for an actual survey where this hybrid estimator has been used.

Key Words: Between- and within-PSU variances, variance ratio, bias and variance of a variance estimator

1. Introduction

In surveys, cluster sampling is often used to select a sample of study units, where study units or ultimate sampling units (USUs) are grouped into clusters and a sample of clusters are first selected to reach USUs. Sometimes, such clusters naturally occur. For example, for a survey of students, schools are first selected, and then students are selected from the selected schools. The first stage sampling units of clusters are customary called the primary sampling units (PSUs).

For such a design, the variance is usually estimated assuming that PSUs are selected with replacement and using the PSU level variance of the PSU level survey estimates. This variance estimate includes both within-PSU variance and between-PSU variance, and is consistent if the sampling fraction is small so that the replacement sampling assumption is reasonably true. The number of degrees of freedom (DF) of the variance estimator is largely determined by the size of the PSU sample. Therefore, if the PSU sample size is small, DF is small, and the variance estimate is very unreliable. A small number of PSUs is selected sometimes to reduce the survey cost, especially when data collection has to be done at the physical locations of PSUs and the data collection cost per PSU is high.

To illustrate the issue of a small DF, assume that the population mean is to be estimated and PSU sample means are normally distributed with a common variance. Then DF is given by DF = $m - H$, where m is the PSU sample size and H is the number of strata in a stratified sample design. The rule of thumb for the minimum desirable DF is 30 (also refer to Korn and Graubard, 1999, pp. 194-195).

Korn and Graubard (1999, pp. 193-202) discuss 3 methods to deal with the small DF issue:

1) Collapse strata, which is helpful if H is large;
2) Ignore PSU-clustering and estimate the variance at the secondary sampling unit (SSU) level or USU level;
3) Estimate the variance at the USU level assuming the sample is a simple random sample (SRS), and then inflate the SRS variance by the average design effect.

Other methods are also available in the literature:

4) Successive difference model (Kish, 1965, p.119) for systematic sampling of PSUs;
5) Using generalized variance functions (Wolter, 2007, Chapter 7).

Methods 1 and 2 gives biased variance estimates in most situations; Method 1 potentially overestimates the true variance by including between-strata variance, which does not exist under the stratified design, whereas Method 2 underestimates by ignoring clustering effect. Method 1 can be useful if the overestimation is not excessive. Method 2 may be used if the underestimation is not severe and the usual consistent variance estimator is too unreliable. Method 3 needs to pool across similar surveys over time or across different variables within the same survey to calculate the average design effect. If the average design effect is very different from the true design effect for a given variable, the variance estimate for the variable will be biased. However, it is a viable option for a survey, which is conducted repeatedly or for variables with similar design effects within a survey.

Method 4 was studied by DuMouchel, Govindarajulu, and Rothman (1973) in comparison with the collapsing strata strategy when one unit is selected from each stratum. Viewing systematic sampling as a means of creating implicit strata with one selected unit per stratum, their study is applicable to our situation. They compared pairing of adjacent strata to create a collapsed stratum of two units each and the successive pairing of Kish and concluded that the latter is more efficient in most situations. Moreover, it has a larger number of degrees of freedom.

Method 5 is a general variance estimation method applicable when there is a functional relationship between the expected value of the point estimate and the variance of the point estimate. Since this function has to be estimated, data are pooled over variables with similar functional relationship. This pooling could make the method biased for estimating the variance of some estimates. Kalton (1995) also discusses Methods 1, 2, 4, and 5.

In this paper, we want to address the small DF problem for the situation that a small number of PSUs is selected by systematic sampling from a sorted list with a hope of improving the sampling efficiency through implicit stratification rather than explicit stratification. It is assumed that many secondary sampling units are selected so that we can use the Method 2 approach. However, we want to correct the bias of the Method 2 variance estimator by combining the Method 2 variance estimator, which is stable but biased, with the usual unstable but consistent variance estimator. This new variance estimator is called the Hybrid variance estimator in this paper.

2. Hybrid Variance Estimator

Let V_1 be the variance estimator that estimates the within-PSU variance. This variance is estimated using the SSUs as PSUs and treating the PSUs as strata. It is in fact the Method 2 variance estimator. Let V_2 be the usual design consistent variance estimator. V_1 is stable with a large DF but biased, whereas V_2 is design consistent but unstable with a small DF. The Hybrid variance estimator, H, is then defined by combining V_1 and V_2, as follows:

$$H = BV_1, \tag{1}$$

where

$$B = E(V_2)/E(V_1). \tag{2}$$

The expectation of H is equal to that of V_2, which implies that the Hybrid variance estimator is consistent as is V_2. However, the problem is B is an unknown population quantity. So we need to estimate it. One simple estimator would be:

$$b = \hat{B} = V_2/V_1. \tag{3}$$

This estimator is inherently unstable because V_2 is used in the numerator. So we use an average of many estimates, b's, for variables that are believed to have similar B's. With this average for K such variables, where K is sufficiently large, the Hybrid estimator is defined by:

$$H = \bar{b}V_1, \tag{4}$$

where

$$\bar{b} = \frac{1}{K}\sum_{i=1}^{K} b_i. \tag{5}$$

The National Survey of the Use of Booster Seats (NSUBS) sponsored by the National Highway Traffic Safety Administration (NHTSA) used the Hybrid variance estimator. But its performance was not studied, and this is the motivation of this study.

3. Performance of the Hybrid Variance Estimator

The performance of the proposed Hybrid variance estimator was studied via simulation using a population data generated based on the 2009 NSUBS data. The sample design of the survey and the Hybrid variance estimator used are described below.

- 16 PSU's (a PSU is a county or a group of counties) are systematically selected from 50 NOPUS PSUs; two of 16 are certainties. The old National Occupant Protection Use Survey (NOPUS) was another restraint use survey, which used 50 PSUs, from which the PSU sample for NSUBS was selected.
- Observation sites are stratified within PSUs by site types of the following:
    - Daycare centers;

- Recreation centers;
- Fast food restaurants;
- Gas stations.

- In 2009, 674 sites were selected independently from each site type stratum within PSU, of which 166 were found to be ineligible due to frame problem. Out of 508 eligible sites, 433 sites responded.
- Vehicles with children of age under 13 are observed for two hours. In 2009, there were 6,033 observed vehicles with 9,471 observed children, of which 7,284 children were interviewed.
- Using 57 most important estimates of children's restraint use, *B* is estimated by the mean of the middle 50% of *b*'s for the 57 estimates (50 % trimmed mean was used to avoid undue influence of extreme values). Children's restraint use is estimated by various children's characteristics such as age, gender, weight, height, geographic region, race/ethnicity, weather condition, restraint type, and also by driver's characteristics.

The Simulation study was conducted as described below:

- The population data were generated for the 50 NOPUS PSUs using PSU level characteristics such as geographic region and metropolitan status.
- 4,800 samples were selected following the NSUBS design closely but conditionally from the 50 NOPUS PSUs.
- From each sample, 57 point estimates and Hybrid variance estimates were calculated.
- The variance of a point estimate over 4800 samples is considered the "true" conditional variance for the point estimate.
- The average of 4,800 Hybrid variance estimates is treated as the conditional expectation of the Hybrid variance estimator (similarly for V_1 and V_2).
- This is compared with the "true" conditional variance for each of 57 variables to examine the size of the bias.
- Relative difference (RD) is examined, where RD of X and Y is defined by $\mathrm{RD}(X,Y) = 100(X-Y)/Y$.

Table 1 shows RD's of the simulated expectation of the Hybrid, V_1, and V_2 with respect to the "true" (simulated) variance for the 57 children's restraint use estimates. It also compares the Hybrid and V_1 with V_2, which is supposed to be consistent.

Table 1: Simulation Results - Relative Differences

Variance Estimators	*N*	*Mean*	*Std Dev*
Hybrid vs. Sim Var	57	37.6	41.8
V_1 vs. Sim Var	57	-8.3	27.7
V_2 vs. Sim Var	57	36.2	45.7
Hybrid vs. V_2	57	7.8	41.9
V_1 vs. V_2	57	-28.1	28.5

Table 1 shows that the average RD of the Hybrid and true (simulated) variance for 57 estimation cases is about 38%, which means that the Hybrid overestimates the variance about 38% in average. It also shows that the Hybrid and V_2 perform similarly, which is

also demonstrated by the comparison of the Hybrid with V_2. V_1 performs surprisingly well as its average RD with respect to the simulated variance is -8%, which although negative as expected is fairly close to zero. The comparison between V_1 and V_2 demonstrates that the between PSU variance is not negligible.

The simulation result about the correction factor (B) is shown in Table 2. The average B over 57 estimation cases is 1.59, which is tracked quite closely by the 50% trimmed mean that was used to define the Hybrid variance estimator. The average of (simulated) expectation $E(b)$ of b is 1.88, which is 18% larger than the average B.

Table 2: Simulation Results about the Correction Factor (B)

Factor	*N*	*Mean*	*STD*	*Min*	*Med*	*Max*
$E(V_2)/E(V_1)$	57	1.59	0.64	0.48	1.41	3.79
$E(b)$	57	1.88	0.77	0.95	1.62	4.54
50% Trim b	57	1.52	N/A	N/A	N/A	N/A

Figure 1 shows the scatter plot of 57 points of ($E(V_1)$, $E(V_2)$). Strong linear relation that passes through the origin is exhibited, which provides an empirical justification of the ratio model (2) used by the Hybrid variance estimator.

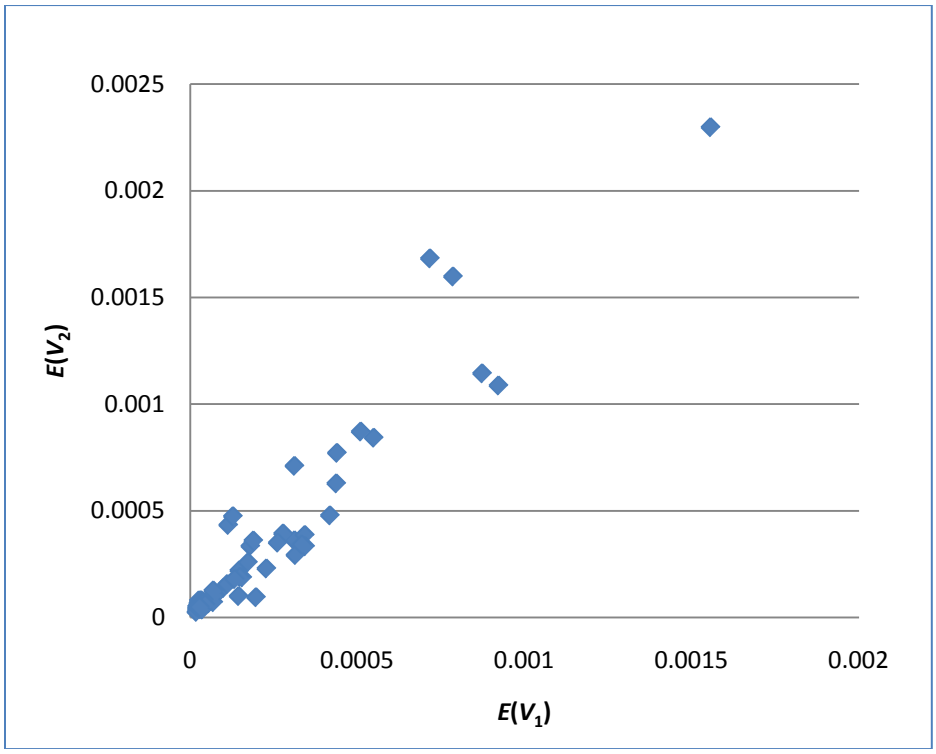

Figure 1: Scatter plot of $E(V_1)$ vs. $E(V_2)$

We also examined the coverage properties of the normal theory-based 95% confidence intervals formed by various variance estimators. Table 3 presents the results, which are somewhat surprising because we expected conservative coverages by the Hybrid and V_2. Contrary to our expectation, they are lower than the nominal value of 95% yet acceptable as is usual to see similar results from survey estimates with well accepted variance

estimates. It would have been better if the t-distribution had been used. However, it is not trivial to determine the true degrees of freedom. Preliminary result, when the t-distribution is used with rough DFs, shows coverage improvements of 1-2% but the intervals are still conservative. Skewness of the intervals can be seen from the asymmetry of the lower and upper coverages. V_1, although its bias is the smallest, produces too liberal confidence intervals.

Table 3: Simulation Results – Coverage of 95% CI

Var Est	*N*	*Lower Cov.*	*Upper Cov.*	*Total Cov*	*Half Length*
Hybrid	57	39.7%	51.8%	91.5%	2.96%
V_1	57	37.5%	48.0%	85.5%	2.43%
V_2	57	39.2%	51.0%	90.2%	2.92%

4. Summary and Some Concluding Remarks

The simulation study results presented in the previous section are summarized in the following:

- The Hybrid variance estimator is 38% positively biased in average compared to the simulated variance.
- The Hybrid and V_2 estimators perform similarly in terms of bias.
- V_2 is 43% more variable than the Hybrid, whereas the Hybrid is 60% more variable than V_1. This is part of the simulation results but not shown in Section 3.
- The 50% trimmed *b*-factor closely tracks the average *B*-factor.
- Although the Hybrid and V_2 are substantially positively biased, the 95% confidence intervals are still liberal yet acceptable.

The Hybrid variance estimator and V_2 perform similarly in terms of the bias but the Hybrid cuts down over 40% of the instability of V_2. So it meets our expectation. However, it is puzzling that V_2 is not consistent as expected. Possible explanations are:

- Partially due to non-replacement sampling, while the variance estimator, V_2 was formulated based on the replacement sampling assumption;
- Probably something to do with systematic sampling – the usual way of reflecting implicit stratification by constructing variance strata from the sorted list of sampling frame introduces a positive bias (Chromy, 2010).

It is a well accepted practice to provide slightly conservative variance estimates. The reader may think that 38% overestimation is too excessive. However, liberal yet acceptable coverage property provides a justification for the use of the Hybrid variance estimator for NSUBS. The application of the method for other situations may be different, so careful examination may be necessary.

Obviously, there is plenty of room for further study. Some possible directions for future study are given below:

- Use the successive difference model (Method 4) for V_2 to define the Hybrid variance estimator;
- Try the GVF method (Method 5) to see if it works for NSUBS;

- Try design effect adjustment method (Method 3) to see if it works for NSUBS.
- Study the confidence interval coverage using the t-distribution.

Acknowledgements

I would like to thank Drs. Bob Fay and Graham Kalton at Westat for their helpful comments on the draft of this paper.

References

Chromy, J.R. (2010). Horvitz-Thompson Variance Weights: Exact vs. Approximate. Presented at the 2010 Joint Statistical Meetings, Vancouver, Canada.

DuMouchel, W.H., Govindarajulu, Z., and Rothman, E. (1973). A Note on Estimating the Variance of the Sample Mean in Stratified Sampling. *Canadian Journal of Statistics*, 1, 267-274.

Kalton, G. (1995). Variance Estimation with Few Degrees of Freedom. *Bulletin of the International Statistical Institute*, 56(4), 1642-1645.

Kish, L. (1965). *Survey Sampling*. New York: John Wiley and Sons, Inc.

Korn, E.L., and Graubard, B.I. (1999). *Analysis of Health Surveys*. New York: John Wiley and Sons, Inc.

Wolter, K.M. (2007). *Introduction to Variance Estimation*, 2nd edition. New York: Springer.