# Comparison of Variance Estimation Techniques in a Simulation Study

Pedro J. Saavedra, Tigran Markaryan and Wendy A. Wyatt
Pedro J. Saavedra, ORC Macro, 11785 Beltsville Dr., Calverton, MD 20705

**Keywords: Bootstrap, jackknife, cluster sampling, Taylor Series, Monte Carlo simulations**

## Introduction

A simulation study was conducted using a frame with 200,000 records, 1,000 PSUs, one categorical variable designed to represent an adjustment variable and twenty variables with different distributions, intra-class correlations and relationships to the adjustment variable. Inaccuracies in the original PSU size estimates and differential nonresponse were built into the simulations. The study used two different simulated samples that varied considerably the distributions of the variables in each from which variances were estimated. The choice of two samples and many estimands and methods rather than a more intensive selection of multiple samples examining one or two methods and estimands was made for its heuristic value. Each of the two samples of 100 PSUs and 25 units per PSU were drawn from the frame, and point estimates were obtained by adjusting to a simulated race/ethnicity variable. Variance estimates were then obtained by drawing multiple samples and obtaining the standard deviation of the estimates across samples. These estimates were compared with those obtained from each of the two samples using several jackknife, bootstrap, and Taylor Series linearization techniques. The results are discussed in terms of the distribution of the variables and the techniques used.

Most variance estimation studies compare variance estimates from a sample by examiningh each method to a preferred method. In a few studieses a real sample has been used to create a population and then the different variance estimates have been compared. This paper compares different estimates obtained from a sample with the variances obtained by drawing multiple samples from the frame. Since frame data is seldom available and must be simulated, it is often not possible to determine the degree to which estimates obtained from a sample are biased, and whether the bias is a function of the sample or the estimation method.

## General Considerations

Consider an extreme situation. A ten percent sample with one extreme outlier is drawn from a population. In this extreme case the population has 100 units and a sample of 10 is drawn. One unit has a value of 1,100 and the other 99 units have a value of 100. The result is that ten percent of the samples will yield an estimate of 200 and the remaining ninety percent will yield an estimate of 100. Variance estimates will be 10,000 for ten percent of the samples and 0 for the remaining 90 percent. The true variance of the mean is 1,000.

The above example would be true with any variance estimate derived from the samples. No estimation approach could solve the problem that an estimate from the sample will remain only an estimate and would be sensitive to the peculiarities of the particular sample selected. However, different variance estimators from a sample may be more or less sensitive to the peculiarities of the sample, the particular characteristics of the sampling design or the weighting adjustments.

This motivation of this paper proceeds from frequent disagreements over the preferred method of choice under different circumstances. Most often one can compare different variance estimates obtained from a sample, but does not know what the true variance is. Frequently there is no good procedure through which variances can be estimated, and in order to obtain estimates one has to assume that one aspect of the design is unimportant.

Such assumptions are common in situations where a cluster sample is selected without replacement with Probabilities Proportional to Size (PPS), and where adjustments are made to the weights in order to match some external population count and to adjust for possible underrepresentation of certain groups. Most of the time when the sampling fraction is small, the sample is treated as if it were selected without replacement. If one uses SUDAAN or some other software which relies on Taylor Series, one often ends up ignoring some aspects of weight adjustments. The fact that the variance estimates themselves have an error of estimate -- and that such an error depends on the methodology -- is often not taken into account by some practitioners. But ultimately one does not usually have frame data to establish comparisons.

The purpose of this study is to examine several variance estimation methods using a simulated frame and sample. The simulated frame and design were kept very simple, with the sole intent of comparing estimation methods for means and ratios. This is a preliminary report on some simulations based on one sample and one frame, with estimates calculated for twenty means and twenty ratios.

## The Frame and Sampling Design

The simulated frame consisted of 200,000 units divided into 1,000 PSUs. The units were randomly assigned to the 1,000 PSUs by first randomly assigning each a value from 1 to 1,200 and then subtracting 1,000 from those numbers greater than 1,000. Thus the average size of twenty percent of the PSUs was twice as large as the average size of the remaining 800, but the actual population of the PSUs varied.

In cluster sampling one seldom has an exact measure of size for the PSUs, though once a PSU is selected one can make a more precise determination of the number of units. We simulated this process by assigning each unit a random number with a mean of 1.0 and a standard deviation of 1.0. Then the sum of these size measures within each PSU became the reported size of the PSU. We also defined five "races" with proportions .35, .25, .25, .10 and .05 respectively. These were spread randomly across PSUs. Variables were later assigned a certain variation due to race and it was assumed that the population count by race was known. This variable was designed to simulate the weight adjustments for race or ethnicity that are common in many surveys.

The sampling design involved selection of 100 PSUs with PPS using the Goodman-Kish approach with no stratification. The PSUs were sorted randomly and a sampling interval equal to one tenth of the sum of the size measures was chosen using a random starting point. It was assumed that the real size of a PSU could be ascertained once the PSU was selected, and 25 units were selected from each PSU.

Using Poisson sampling we designated 5% of the sample as non-respondents. Adjustment for non-response was implemented at the PSU level. Finally an adjustment for race was implemented in order to make the weights in the race category add up to the population. Hence, the preliminary weights were as follows:

1) A PSU weight equal to .01 times the sum of the sizes of the PSUs divided by the PSU size.

2) A unit weight equal to the number of units in the PSU divided by 25.

3) An adjustment for non-response equal to 25 divided by the number of respondents in the PSU.

4) An adjustment for "race" equal to the population total for the unit's "race" category divided by the sum of the products of the first three weights across all units in the sample.

Two different samples with two sets of variables were drawn.

## The Variables

For Study 1 twenty variables were simulated in the following ways.

1) Each variable started as a normal variate with mean of 0 and standard deviation of 1.

2) The variables were transformed in groups of four as follows:

a) Variables 1 through 4 and 17 through 20 were left intact.

b) Variables 5 through 8 were transformed by preserving the sign and taking the square root of the absolute value.

c) Variables 9 through 12 were transformed by taking the exponential minus 1.

d) Variables 13 through 16 were transformed by taking the square, but preserving the sign.

3) The variables thus transformed were normed to a mean of 1,000 and a standard deviation of 100.

4) Twenty additional variables were defined at the PSU level with a mean of zero and a standard deviation of ten. These were added respectively to each of the twenty original variables.

5) Twenty variables were defined for each race so so as to have a standard deviation between races equal to k where k is the index of the variable in question. The value of each of these for a unit;s race was added to the corresponding variable. Thus each variable has a progressively higher variance between races.

6) The variables were once again normed to a

mean of 1000 and a standard deviation of 100.

7) Means were estimated for each of the twenty variables. In addition, estimates were calculated for the ratios of $V_j/V_{j+5}$ for j=1 to 15, and $V_j/V_{j-15}$ for j=16 to 20. These ratios were multiplied by 1,000 to make examination of results easier.

For Study 2 the variables were defined as before with the following exceptions.

1) The between-PSUs standard deviation was 50 instead of 10.

2) The race differences were four times as large as in Study 1.

3) One percent of the cases for each variable had the original value (normalized with a mean of zero) multiplied by five.

Everything else was the same for Study 2, except that the a new sample was drawn.

**The Variance Estimators.**

The standard against which all estimators were measured was obtained by drawing 1,000 samples and obtaining estimates for each mean and each ratio. Since this estimator is not an exact estimate of the variance, we included a second set of 1,000 samples from the frame as one of the estimators used. Thus one could measure the agreement of every other method to the criterion with the agreement of the same method using different random numbers.

The methods of variance estimation drawn from the sample included:

1) *Delete-a-group jackknife* . The 100 sampled PSUs were ordered in the initial order (which means that the PSUs in the top 20% with respect to size appeared first) and a group variable was created with twenty values, where the first, twenty-first, forty-first, sixty-first and eighty-first PSUs were assigned a value of one and so forth. Units in PSUs with a value of 1 were dropped and the fourth adjustment factor was recalculated. The standard jackknife formula with twenty groups was then applied (Kott, 1998).

2) *Ordinary jackknife* . The same procedure as above, but each PSU constituted a group unto itself.

3) *Bootstrap -- 200 replicates.* Two hundred replicate samples of 100 PSUs were selected with replacement from the main sample. Each unit retained the first three components of the weight, but the fourth component (the adjustment to population totals by the "race" category) was recalculated. The standard deviation of the estimates was used.

4) *Bootstrap -- 1,000 replicates.* Same as above, but with 1,000 replicates.

5) *Half Sample Replication.* Here 1,000 half samples of 50 PSUs were drawn randomly from the main sample. As before, the first three components of the weight were preserved, and these weights were adjusted by categorical totals.

6) *Taylor Series -- With Replacement.* SUDAAN was used to set a design for sampling equiprobably with replacement at the first stage. The final weights were used for the individual units.

7) *Taylor Series - Without Replacement.* Unequal probabilities without replacement were used with SUDAAN. However, the joint probabilities were approximated with .99 times the product of the probabilities. The final weights were used as weights and no post-stratification was defined.

8) *Taylor Series -- With Replacement and Post-Stratification.* SUDAAN was used to set a design for sampling equiprobably with replacement at the first stage. Weights prior to final adjustment were used and the "race" variable totals were provided.

9) *Taylor Series - Without Replacement and with Post-Stratification.* Unequal probabilities without replacement were used with SUDAAN. However, the joint probabilities were approximated with .99 times the product of the probabilities. Weights prior to final adjustment were used and the "race" variable totals were provided.

**Results**

The above designs are presented as 1 to 9 in the table and graphs. The second set of estimates from the frame is presented as 10. The estimates from the frame can be trusted to be closest to the true variance in that they are not dependent of the specific sample selected. Thus the first one was set up as the standard and three measures were obtained to measure the similarity of the nine methods and the second frame estimate to the first frame estimate. These were the mean difference, the root mean square difference and the mean absolute difference of each estimate with the first estimate. A summary of these results are presented in Table 1. The root mean square difference are presented in the four charts that follow.

For Study 1 the Jackknife with 20 groups had both the mean variance (across all the variables) closest to the mean of the twenty variables and the worst variable by variable correspondence of any of the methods. In other words, if one applied this method to many variables and took the average standard error, one is likely to obtain a result that comes close to what one would derive from the frame, but the variable by variable estimates would be off. Indeed it also has the largest variation in the estimates of any method. The close mean held up only for the Study 2 ratios, but for all four sets the root mean squares and absolute deviations were the worst.

All of the methods tend to overestimate the variance, and the Taylor Series Without Replacement and using the adjusted weights instead of specifying the post-stratification consistently overestimated the most. It should be noted that the joint probabilities were not correctly specified, since the calculation in this case would be complex, but instead a number slightly lower than the product of the probabilities was used.

Several tentative conclusions can be derived from the simulations. The use of two samples was not intended to yield definitive conclusions, but rather to provide heuristic insights into the potential hazards involved in the selection of method. The first insight is that the delete-a-group jackknife has a relatively poor differential performance. In other words, while it neither over-estimates nor underestimates systematically, its estimates can be off the mark in either direction. This is to be expected due to the limited number of degrees of freedom. On the other hand, this procedure may be quite useful when trying to obtain a design effect across variables. The second is that Taylor Series requires that post-stratification be specified when it is in fact used in the weighting. One surprise was that that the Monte Carlo half-sample seems to do quite well and is easy to program. While we did not use Fay's method (Judkins, 1990) it seems clear that it could be easily applied to the Monte-Carlo half-sample if one were interested in small subdomain estimates, and this could be a procedure of choice because of its ease in programming.

## Bibliography

Effron, B. (1982). *The Jackknife, the Bootstrap, and Other Resampling Plans*, SIAM Monographs No. 38.

Judkins, D. (1990) Fay's method for variance estimation. *Journal of Official Statistics*, 6 223-240.

Kott, Phillip S. (1998). *Using the Delete-A-Group Jackknife Variance Estimator in NASS Surveys*, RD Research Report No. RD-98-01, USDA, NASS: Washington, DC. ,

Särndal, C., Swensson, B. and Wretman, J. (1997) *Model Assisted Survey Sampling,* Springer Series in Statistics.

Shah, B.V. , Barnwell, B.G. and Bieler, G.S. (1996). *SUDAAN User's Manual: Release 7.0.* Research Triangle Institute , Research Triangle Park, NC

## Table 1 Summary Statistics for Various Methods

**Study 1 Means**

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Diff. | 0.024 | 0.080 | 0.065 | 0.061 | 0.057 | 0.101 | 0.192 | 0.077 | 0.167 | -0.001 |
| RMSQ | 0.454 | 0.177 | 0.198 | 0.174 | 0.185 | 0.189 | 0.248 | 0.175 | 0.228 | 0.060 |
| A.D. | 0.378 | 0.151 | 0.149 | 0.149 | 0.159 | 0.163 | 0.219 | 0.150 | 0.194 | 0.051 |

**Study 1 Ratios**

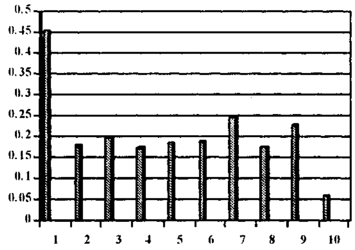
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Diff. | 0.191 | 0.142 | 0.105 | 0.122 | 0.095 | 0.169 | 0.295 | 0.137 | 0.263 | -0.027 |
| RMSQ | 0.608 | 0.304 | 0.351 | 0.284 | 0.244 | 0.331 | 0.406 | 0.301 | 0.372 | 0.122 |
| A.D. | 0.479 | 0.248 | 0.282 | 0.238 | 0.205 | 0.261 | 0.358 | 0.244 | 0.331 | 0.104 |

**Study 2 Means**

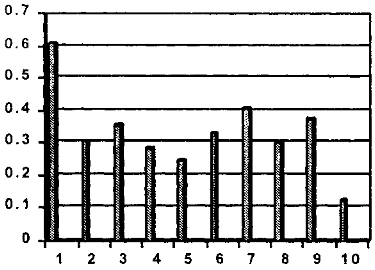
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Diff. | 0.315 | 0.295 | 0.272 | 0.272 | 0.251 | 0.330 | 0.369 | 0.288 | 0.297 | -0.045 |
| RMSQ | 0.729 | 0.470 | 0.485 | 0.485 | 0.452 | 0.498 | 0.524 | 0.466 | 0.452 | 0.120 |
| A.D | 0.539 | 0.345 | 0.377 | 0.359 | 0.338 | 0.379 | 0.405 | 0.341 | 0.336 | 0.092 |

**Study 2 Ratios**

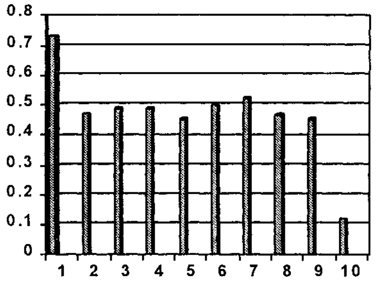
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Diff. | 0.261 | 0.407 | 0.373 | 0.367 | 0.337 | 0.421 | 0.477 | 0.398 | 0.447 | -0.073 |
| RMSQ | 1.117 | 0.637 | 0.904 | 0.888 | 0.805 | 0.642 | 0.677 | 0.630 | 0.659 | 0.222 |
| A.D. | 0.886 | 0.540 | 0.720 | 0.703 | 0.666 | 0.557 | 0.586 | 0.534 | 0.558 | 0.156 |



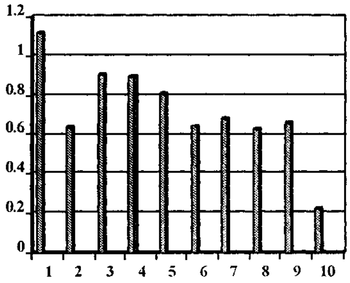
Study 1 Means



Study 1 Ratios



Study 2 Means



Study 2 Ratios