Horvitz-Thompson Variance Weights: Exact vs. Approximate

James R. Chromy
RTI International, 3040 Cornwallis Road, P.O. Box 12194, Research Triangle Park, NC 27709

Abstract

Sequential probability minimum replacement sample designs provide a practical methodology for selecting PPS samples that satisfy the requirement of positive pairwise probabilities and nonnegative variance weights. The exact solutions for the variance weights can lead to some unacceptable variance estimates such as zero estimates regardless of the observed values. This paper explores some alternative approximate variance estimators that avoid this problem. Although not strictly unbiased, the variance estimates from alternate estimators can be shown to be nearly unbiased and to have less variability than the unbiased variance estimators based on the exact variance weights. Some comparisons to PPS systematic designs are also addressed with alternate variance weights.

Key Words: PPS sampling, PPS sequential sampling, PPS systematic sampling, variance approximations, probability minimum replacement

The discussion in this paper follows from the discussion in a paper presented in the D.G. Horvitz Memorial session at the 2009 Joint Statistical Meetings (Chromy 2009).

1. Overview

PPS sampling is widely used to provide an opportunity to achieve a zero or near zero variance when the variable being observed is proportional to or nearly proportional to the size measure used in selecting the sample. It is also commonly used in selecting initial stages of the sample in multi-stage designs. Hansen and Hurwitz (1943) provide unbiased estimates for totals and their variances when utilizing with replacement PPS sampling. Horvitz and Thompson (1952) developed a general theory for unbiased estimation of a population total and the variance of the estimate when selecting PPS samples without replacement. They noted that unbiased estimation of the variance was only possible if the pairwise probabilities of selection were positive. Even when the pairwise probabilities are positive, their computation is sufficiently complex that many practitioners use simpler approximate variance formulas which either ignore the gains from utilizing without replacement sampling or approximate the gain with an approximate finite population factor incorporated into Hansen-Hurwitz' with replacement variance estimation formula.

This paper examines the bias for some approximate variance estimation formulas when applied to variables in a simulated population when using two PPS without replacement designs: PPS-sequential (Chromy 1979, 1981) and PPS systematic (Madow 1949). Both methods are available in SAS Proc SurveySelect (SAS Institute Inc., 2004) and both procedures are designed to take advantage of implicit stratification achieved through sorting the sampling frame by a frame variable related to the variables of interest. Williams and Chromy (1980) proposed a serpentine ordering method based on one or

more categorical variables and a final continuous variable. The option for serpentine sorting is also available in SAS Proc SurveySelect.

Note that both PPS sequential and PPS systematic designs can be used to select probability minimum replacement (PMR) designs. PMR designs allow use of size measures greater than 1 over the sample size when size measures are highly unequal and the sample size is fairly large. To accommodate PMR designs it is useful to think in terms of expected sample size at the unit level (number of times a unit is selected). In PMR designs, a unit will be selected either $\lfloor E(n_i) \rfloor$ or $\lfloor E(n_i) \rfloor + 1$ times. When all expected sample sizes are less than 1, the special case of PPS without replacement results. Notation for PMR designs is employed in this paper, but the simulated population used for the empirical study reduced to the special PPS without replacement case. To compare the usual without replacement notation to the PMR notation, it useful to define the probabilities $\pi_i = P\{n_i = \lfloor E(n_i) \rfloor + 1\}$ and $\pi_{ij} = P\{n_i = \lfloor E(n_i) \rfloor + 1, n_j = \lfloor E(n_j) \rfloor + 1\}$.

3. PPS Sample Designs

3.1 PPS Sequential Designs: When applied in the PPS without replacement sampling context with samples of size at least two, PPS sequential designs produce positive pairwise probabilities for all pairs of units with positive probability, $\pi_{ij} > 0$ .for all $i \neq j$. When applied in a PMR case, in addition to requiring a sample of at least size 2, it is also necessary for $\sum_{i=1}^{N} \pi_i > 2$; then $E(n_i n_j) > 0$ and $\pi_{ij} > 0$ .for all $i \neq j$.

The population total, $T = \sum_{i=1}^{N} X_i$, then has an unbiased estimator, $\hat{T} = \sum_{i=1}^{n} x_i / E(n_i)$.The Yates-Grundy (1953) form of the variance estimator is then

$$V(\hat{T}) = \frac{1}{2} \sum_{i=1}^{N} \sum_{j \neq i}^{N} \{E(n_i)E(n_j) - E(n_i n_j)\} \left[\frac{X_i}{E(n_i)} - \frac{X_j}{E(n_j)} \right]^2$$

with unbiased variance estimator

$$\hat{V}(\hat{T}) = \frac{1}{2} \sum_{i=1}^{n} \sum_{j \neq i}^{n} \left(\frac{E(n_i)E(n_j) - E(n_i n_j)}{E(n_i n_j)} \right) \left[\frac{x_i}{E(n_i)} - \frac{x_j}{E(n_j)} \right]^2 .$$

Note than the value $\{E(n_i)E(n_j) - E(n_i n_j)\}$ in the variance expression is the negative covariance of the achieved sample sizes for units *i* and *j*. For PPS sequential designs, the covariance of any two sample sizes is always less than or equal to zero. It tends toward zero as units in the frame get farther apart. As a result, all terms in the variance and in the variance estimator are positive or zero.

3.2 PPS Systematic Designs: While the primary focus of this paper is on approximate variance estimators for the PPS sequential sampling, the same types of approximations can be applied to PPS systematic sampling designs. Systematic sampling also has the PMR property and has the same form of the formula for estimating the population total. The Yates-Grundy variance expression can also be applied to PPS systematic, but no unbiased variance estimator exists since $E(n_i n_j) = 0$ and $\pi_{ij} = 0$ for some $i \neq j$. Systematic sampling and PPS systematic sampling greatly reduce the number of possible samples. The unit sample sizes for elements that can appear in the same sample together have a positive covariance and therefore the associated coefficients in the Yates-Grundy

variance expression are negative. Positive coefficients in the variance formula are associated with sample elements that cannot occur together in the same sample.

5. Approximate Successive Difference Variance Estimators

The approximate variance estimators considered in this research are successive difference estimators. Pairs of elements included in the estimator are limited to those that appear adjacent to each other in the ordered sample. Note that successive difference estimators can be viewed as the average of two pseudo stratum estimators with each pseudo strata containing two sample elements. For even n, one set of pseudo strata includes the pairs (1,2), (3,4),…,(n-3,n-2), (n-1, n) and the other contains the pairs (2,3), (4,5),….,(n-2,n-1), (n,1). For odd n, the pseudo stratum approach requires the formation of one pseudo stratum with 3 elements. The successive difference estimator avoids problems with odd sample sizes. Because of the relationship of the pseudo stratum estimator to the successive difference estimator, the results reported in this paper should also apply to approximate variance estimators based on the pseudo strata of size 2.

The approximate successive difference variance estimators studied were based on the with replacement (WR) form of the variance applied to pseudo stratum samples of size 2. A finite population correction factor is added to the formula to adjust for sampling without replacement (or with minimum replacement). The general form is

$$\hat{V}(\hat{T}_t) = \frac{1}{2}\sum_{i=1}^{n-1} fpc_{t(i,i+1)}\left[\frac{x_i}{E(n_i)} - \frac{x_{i+1}}{E(n_{i+1})}\right]^2 + \frac{1}{2} fpc_{t(n,1)}\left[\frac{X_n}{E(n_n)} - \frac{X_1}{E(n_1)}\right]^2.$$

The subscript t denotes the type of approximate finite population correction factor applied. All four alternative variance estimators studied can be represented in this general form; the approximate finite population correction factors defining each approximation are shown in Table 1.

Table 1. Approximate Estimators Defined by Assumed Finite Population Correction

t	Description	Computation
1	With replacement (WR)	$fpc_{1(ij))} = 1$
2	Without replacement approximation 1 (WOR1)	$fpc_{2(ij))} = \frac{N-2}{N}$
3	Without replacement approximation 2 (WOR2), a particular solution from Kott (1998)	$fpc_{3(ij))} = \frac{\widehat{N}_{ij} - 2}{\widehat{N}_{ij}}$ $\widehat{N}_{ij} = \frac{1}{E(n_i)} + \frac{1}{E(n_j)}$
4	Without replacement approximation 3 (WOR3)	$fpc_{4(ij))} = \frac{2 - \pi_i - \pi_j}{2}$

This paper is limited to examining the expected values of these approximate variance estimators in a simulated population with all variables known. The expected value of the each approximate estimator can be evaluated in the simulated population as:

$$E[\hat{V}(\hat{T}_t)] = \frac{1}{2}\sum_{i=1}^{N}\sum_{j\neq i}^{N} fpc_{t(ij)} E(n_i n_j \text{ and } i \text{ adjacent to } j)\left[\frac{X_i}{E(n_i)} - \frac{X_j}{E(n_j)}\right]^2$$

The conditional expectation with adjacency was computed by decomposing the overall expectation of the product of the sample sizes into four components based on the

outcomes of the PPS sequential selection procedure: the achieved sample sizes for units *i* and *j* and the cumulative sample sizes achieved through units *i* and *j*. Only those components with the cumulative sample counts differing by exactly one were included in computing the adjusted expectation.

For PPS systematic sampling the probability of being selected and being adjacent in the sample was more straightforward and involved determining the range of uniform (0,1) random numbers that would result in the selection of both units and then checking for their cumulative expected sample size separating the two was less than 1.[1]

6. Generation of a Hypothetical Population

The sampling schemes used in this empirical study were based on selecting PPS samples of size 10 from a population of size 50 with unequal sizes. It was contrived to permit serpentine ordering with over a two-level categorical variable and a continuous sorting variable that was ascending for the first 25 units belonging to category 1 and descending for the last 25 units belonging to category 2. This is the type of situation that samplers assume exists and makes sampling from an ordered list particularly effective when using either PPS sequential or PPS systematic sampling.

The steps in generating the hypothetical population were:

1. Generate an approximate size measure, Z, from the lognormal distribution with normal mean of 8.8 for category 1 units and 9.2 for category 2 units. Both categories were simulated with a normal variance of 0.5.
2. Generate a true size from the lognormal with parameters, log Z and 0.05.
3. Generate a sorting variable, V, from the normal(0,1) distribution.
4. Generate a proportion variable, P, from the beta distribution to achieve intraclass correlations, ρ=0.01, 0.02, 0.04, 0.06, 0.10 by setting $\alpha=0.1(1-\rho)/\rho$ and $\beta=0.9(1-\rho)/\rho$.
5. Setting the magnitude of the ordered disturbance variable, γ, to 0.01, 0.02, 0.04, 0.06, and 0.10.
6. Generating 80 variables, Y, from the model, $Y=PXe^{\gamma V}$ based on the four intraclass correlation variables and five ordered disturbance variables all replicated four times.

It was then possible compute the expected value of the unbiased variance estimator and for each of the four approximate estimators for the PPS sequential design. Note that the formula for the expected value matches the formula for the true variance when using the PPS sequential design.

While no unbiased variance estimator exists for the variance from a PPS systematic sample, it is possible to compute the true variance using the same formula and to compute the expectation of the approximate variance estimators.

7. Empirical Comparisons

7.1 PPS Sequential Sampling: Figure 1 shows variance weights for the simulated population when using PPS sequential sampling. Note in the left panel, that these weights become small (<0.001) as units get farther apart. The positive coefficients in the corners reflect that the method allows for a random start and the ordered frame can be

[1] Additional detail can be provided by the author, but a full description would dominate the remainder of the discussion in this paper.

viewed as being arrayed around a circle. Picking a random start on the circle then guarantees that the products of all pairs of sample sizes have positive expectation. The WR approximation weights are shown in the right panel actually become zero when units are farther apart and no longer can be adjacent in any selected sample. The WR weights are an upper bound for the WOR approximate weights.

Figure 1. Left Panel: Positive Pair Weights (>0.001) for PPS Sequential Sampling for a PPS Sequential Sample of 10 out of 50: True Variance (Expectation of Unbiased Estimator)
Right Panel: Positive Pair Weights for the Expected Value of the Approximate Variance Estimators for PPS Sequential Sampling (>0.001 for the WR approximation)

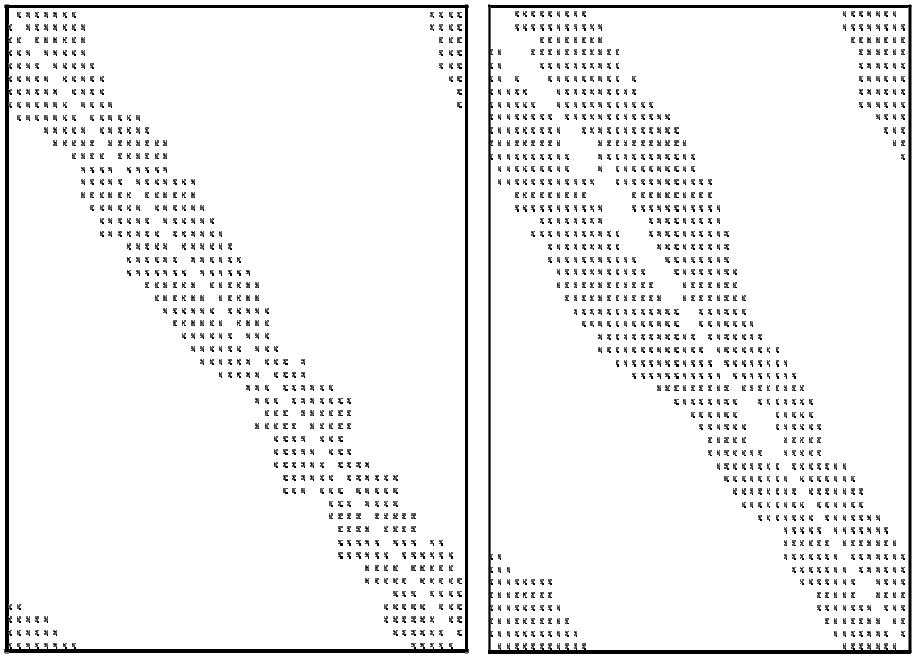


Table 2 shows relative standard errors for the simulated variable outcomes. The entries are the average over four replications with five levels of the perturbation factor and four levels of the intraclass correlation. The approximation based on the WR weights is consistently high and is generally considered to be conservative by practitioners. The other three approximations tend to over-correct and have expected values below the true value but only moderately so. Their stability relative to the unbiased estimates remains to be investigated. The relationship of the expected values of the approximate estimates appears to remain fairly close and consistent across different levels of implicit stratification effects and intraclass correlations used to generate the observed variables. Further investigation of these relationships with some variation in the sampling rate appears warranted before making a general recommendation. Based on the expected values alone and on the current availability of computing power, the use of the unbiased estimator for single stage designs should be strongly considered as more software products support this approach.

Table 2. Relative Standard Errors (Percent) of Estimates and Four Approximations for PPS Sequential Samples of Size10 from a Simulated Population of Size 50: Averages Over 4 Replications for Five Levels of a Perturbation Factor, V, Simulating Increased Ordering Effects and Four Levels of an Intraclass Correlation, RHO, Simulating Increased Variability at the PSU Level

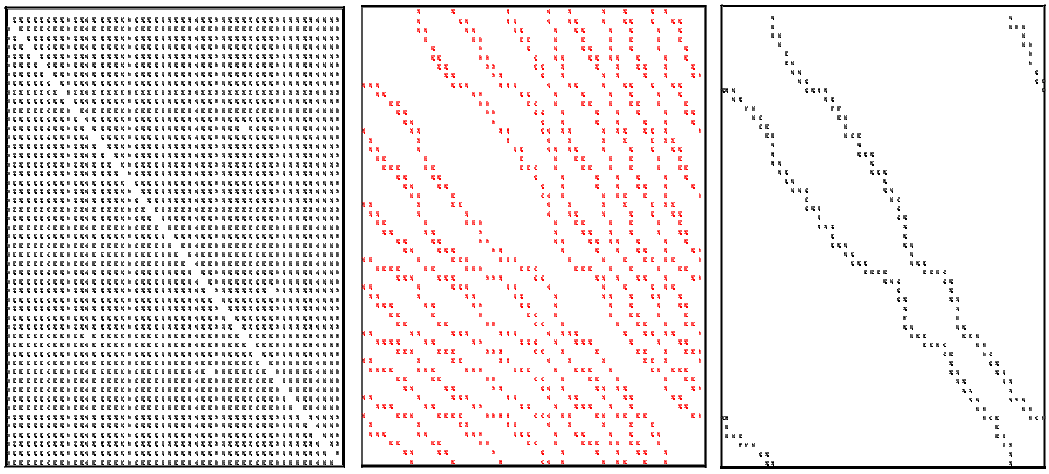
V	RHO	TRUE	WR	WOR1	WOR2	WOR3
0.01	0.01	7.9	8.3	7.5	7.4	7.2
0.01	0.05	17.9	19.7	17.6	17.3	16.9
0.01	0.10	26.3	28.0	25.1	24.6	24.2
0.01	0.20	34.7	36.5	32.7	32.0	31.4
0.02	0.01	8.8	9.2	8.2	8.1	7.9
0.02	0.05	18.9	20.5	18.3	17.7	17.4
0.02	0.10	24.9	27.5	24.6	24.2	23.8
0.02	0.20	41.1	43.7	39.1	38.9	38.1
0.04	0.01	8.3	8.9	7.9	7.8	7.7
0.04	0.05	16.0	17.4	15.5	15.3	15.0
0.04	0.10	26.7	29.7	26.5	26.0	25.4
0.04	0.20	39.1	41.6	37.2	36.4	35.8
0.06	0.01	7.5	8.1	7.2	7.1	6.9
0.06	0.05	17.8	19.9	17.8	17.5	17.1
0.06	0.10	26.9	29.6	26.5	25.9	25.4
0.06	0.20	41.7	44.7	40.0	40.0	39.3
0.10	0.01	8.0	8.7	7.7	7.7	7.5
0.10	0.05	19.3	21.7	19.4	18.9	18.5
0.10	0.10	26.4	27.5	24.6	24.2	23.8
0.10	0.20	35.2	38.5	34.5	33.7	33.0
Average		22.7	24.5	21.9	21.5	21.1

7.2 PPS Systematic Sampling: Figure 2 shows the pattern of variance weights for the simulated population when using PPS systematic sampling. The left panel illustrates that contributions to the true variance come from essentially all pairs. The expected sample sizes for units that can be in the same systematic sample are positively correlated yielding a positive covariance. Since the weights in the Yates-Grundy form of the variance estimate are negative covariances of the achieved sample sizes, this means that units that can appear in the same sample have negative contributions to the true variance. When using an approximation based on successive differences, a positive applied to a nearest neighbour subset of pairs that have negative weights in the true variance expression as shown in the right panel of Figure 2.

Table 3 shows relative standard errors for the simulated variables when using PPS systematic sampling. True values and four approximations are shown. The WR approximation is conservatively high in every case studied. The WOR approximations are also generally high with several exceptions where the opposite is true. One might expect that as the implicit stratification effect due to ordering on a perturbation factor increases that the true variance would increase more rapidly than approximations since

the true variance weights pairs that are not necessarily near each other. This occurs for some cases, but not consistently.

Figure 2. Results for PPS Systematic Sampling of 10 out of 50
Left Panel: Non-zero Variance Weights (Absolute Value >0.001)
Center Panel: Negative Variance Weights (<0.001)
Right Panel: Positive Variance Weights for WR Approximation



8. General Conclusions

In comparing the suitability of successive difference approximations for PPS sequential and PPS systematic designs, approximations for the PPS sequential design will be close because at least one of two conditions is likely to hold: (1) adjacent units capture the benefits of implicit stratification, and (2) the weights in the unbiased variance estimator tend toward zero for elements that are not likely to be adjacent in the sample. Approximations for the PPS systematic depend solely on the first condition. This may explain the less consistent of behaviour of the expected values of approximations for PPS systematic sampling. These conclusions should also apply to approximations based on pseudo strata formed by portioning the ordered list of selected units.

The WR approximation to the variance for PPS sequential and PPS systematic appears to be the safe road to follow for those who do not wish to overstate the statistical significance of their results. This appears to hold up well without excessively overestimating standard errors with a sampling rate as high as 1 in 5 used in the simulations. Using the approximate WOR formulas runs some risk of underestimating the true variance, but this did not appear to be excessive. At higher sampling rates, use of the WR formula may clearly lead to denying the precision of the data and some recognition of the finite population correction may be essential. In multistage sampling, the concern about overestimating the variance is diminished if the first stage component of variance is moderately small. Further modeling and simulation of multi-stage designs is needed to enlighten decisions about the use of finite population correction factors for multi-stage designs.

Table 3. Relative Standard Errors (Percent) of Estimates and Four Approximations for PPS Systematic Samples of Size10 from a Simulated Population of Size 50: Averages Over 4 Replications for Five Levels of a Perturbation Factor, V, Simulating Increasing Ordering Effects and Four Levels of an Intraclass Correlation, RHO, Simulating Increased Variability at the PSU Level

V	RHO	TRUE	WR	WOR1	WOR2	WOR3
0.01	0.01	6.1	9.3	8.3	8.2	8.1
0.01	0.05	17.3	20.7	18.5	18.1	17.7
0.01	0.10	25.3	29.3	26.2	25.7	25.1
0.01	0.20	28.7	38.4	34.4	33.5	32.8
0.02	0.01	8.8	9.2	8.3	8.1	8.0
0.02	0.05	14.9	21.6	19.4	18.6	18.2
0.02	0.10	23.4	29.0	25.9	25.3	24.9
0.02	0.20	44.9	46.6	41.7	41.6	40.6
0.04	0.01	8.6	9.3	8.4	8.2	8.1
0.04	0.05	12.7	18.8	16.8	16.5	16.1
0.04	0.10	26.9	30.8	27.5	26.9	26.2
0.04	0.20	36.8	45.2	40.4	39.5	38.7
0.06	0.01	6.8	8.4	7.5	7.3	7.2
0.06	0.05	18.7	21.0	18.8	18.4	18.1
0.06	0.10	23.4	31.1	27.8	27.2	26.7
0.06	0.20	39.9	47.9	42.8	42.8	41.8
0.10	0.01	7.0	9.2	8.2	8.1	8.0
0.10	0.05	19.4	23.3	20.8	20.3	19.8
0.10	0.10	22.8	29.5	26.4	25.8	25.4
0.10	0.20	36.6	40.5	36.3	35.3	34.5
Average		21.4	26.0	23.2	22.8	22.3

References

Chromy, J. R. (1979). Sequential Sample Selection Methods. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 401-406.

Chromy, J. R. (1981). Variance Estimators for a Sequential Selection Procedure. Current Topics in Survey Sampling. D. Krewski, R. Platek and J. N. K. Rao. New York, Academic Press**:** 329-347.

Chromy, J. R. (2009), "Some Generalizations of the Horvitz-Thompson Estimator," *Proceedings of the Survey Research Methods Section, American Statistical Association*, 217-227.

Horvitz, D. G. and D. J. Thompson (1952). "A Generalization of Sampling Without Replacement from a Finite Universe." The Journal of the American Statistical Association **47**: 663-685.

Kott, P. S. (1988), "Model-Based Finite Population Correction for the Horvitz-Thompson Estimator," *Biometrika*, 75, 797-799.

SAS Institute Inc. (2004). SAS/STAT 9.1 User's Guide. Cary, NC, SAS Institute, Inc.

Williams, R. L. and J. R. Chromy (1980). SAS Sample Selection MACROS. Proceedings of the Fifth Annual SAS Users Group International Conference, SAS Institute, pp.392-6.

Yates, F. and P. M. Grundy (1953). "Selection without Replacement from within Strata and with Probability Proportional to Size." Journal of the Royal Statistical Society **B15**: 253-261.