Varaince Estimation Under Balanced Sampling Plans Excluding Adjacent Units

James H. Wright [1]
Department of Mathematics, Bucknell University, Lewisburg, PA 17837
[1] Research supported by NSA Grant No. H98230-05-1-0022

ABSTRACT

Sampling plans that exclude the selection of adjacent units within a given sample, while maintaining a constant second-order inclusion probability for non-adjacent units, have been proposed as a means of collecting information from populations where neighboring units provide similar responses. Although significant advancements have been achieved concerning the generalization and existence of such sampling plans for finite, one-dimensional populations, many other aspects of these plans warrant further investigation. Results of an investigation of three biased variance estimators of the Horvitz-Thompson estimator of the population mean under such plans are presented

Key Words: variance estimation; balanced sampling plans excluding adjacent units; polygonal designs; circularly and linearly ordered populations; finite population sampling

1. Introduction

Hedayat, Rao, and Stufken (1988a, 1988b) first introduced balanced sampling plans that exclude contiguous units, i.e., plans that prevent the selection of contiguous units within a given sample while maintaining a constant second-order inclusion probability for non-contiguous units. Major advancements have been made in the identification and generalizations of such plans, see Stufken (1993), Stufken and Wright (2001), and Wright and Stufken (2005); however, primarily due to a lack of a sufficient number of identifiable plans, little has been done to investigate corresponding analytical issues. Due to their nature, an unbiased estimator of the variance of the Horvitz-Thompson estimator of the population mean, μ, cannot be obtained from a single application of these plans. In this paper, the relative efficiency of balanced sampling excluding adjacent units with respect to simple random sampling under various population structures, as well as three biased approximation techniques for the variance of the Horvitz-Thompson estimator of μ will be investigated.

A sampling plan is defined as $d = \{(s_k, p_k), k = 1, \ldots, b\}$, where the s_k's are subsets of units and $p_k > 0$ is the probability of selection of the subset s_k such that $\sum_{k=1}^{b} p_k = 1$. The corresponding set $S_d = \{s_k, k = 1, \ldots, b\}$ is called the support of the sampling plan and b is called the support size. The first-order inclusion probability for unit i is defined as $\pi_i = \sum_{\substack{s_k \in S_d \\ i \in s_k}} p_k$ is the probability that unit i is in the selected sample. The second-order inclusion probability for units i and j defined as $\pi_{ij} = \sum_{\substack{s_k \in S_d \\ i,j \in s_k}} p_k$ is the probability that both units i and j are in the selected sample.

For an observed sample, s, the Horvitz-Thompson estimator of μ is defined as

$$\hat{\mu}_{HT} = \frac{\sum_{i \in s} \breve{y}_i}{N}, \qquad (1.1)$$

where $\breve{y}_i = \frac{y_i}{\pi_i}$. Provided that all first-order inclusion probabilities of units in the population are positive, (1.1) is an unbiased estimator of μ.

The variance of the Horvitz-Thompson estimator of μ is equal to

$$V(\hat{\mu}_{HT}) = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \Delta_{ij} \breve{y}_i \breve{y}_j, \qquad (1.2)$$

where $\Delta_{ij} = \pi_{ij} - \pi_i \pi_j$. Provided that the second-order inclusion probabilities of all pairs of units in the population are positive, (1.2) may be estimated unbiasedly by

$$\hat{V}(\hat{\mu}_{HT}) = \frac{1}{N^2}\sum_{i,j\in s}\sum breve{\Delta}_{ij}\breve{y}_i\breve{y}_j \,, \qquad (1.3)$$

where

$$\breve{\Delta}_{ij} = \frac{\Delta_{ij}}{\pi_{ij}} = \begin{cases} 1-\dfrac{\pi_i\pi_j}{\pi_{ij}}, & \text{for } i \neq j \\ 1-\pi_i, & \text{for } i = j. \end{cases}$$

2. Results for Balanced Sampling Plans Excluding Adjacent Units

Commonly, such as in environmental and ecological populations, neighboring units within a finite population, spatially or sequentially ordered, may provide similar information. When sampling from such populations, one may desire a sample that avoids the selection of adjacent units. This paper will focus on one-dimensional orderings, of which two situations can occur. The population may follow a circular ordering, in which the first unit of the population is contiguous with the last unit, or a linear ordering, in which the first unit is not contiguous with the last unit.

Hedayat, Rao, and Stufken (1988a, 1988b) proposed a sampling plan for a given circular population of size N, for which a sample of size n is obtained without replacement such that the second-order inclusion probabilities are 0 for contiguous units and constant for non-contiguous units. Stufken (1993) introduced an extension to balanced sampling avoiding adjacent units. Under such a plan, called a circular BSA(N, n, α), units that are adjacent, i.e., within a distance of $\pm\alpha$ (*mod N*), do not appear within the same sample, while the second-order inclusion probabilities of non-adjacent units are constant. For given N, n, and α, the first- and second-order inclusion probabilities under a circular BSA(N, n, α) are

$$\pi_i = \frac{n}{N} \quad \text{for } i = 1,\ldots,N,$$

and

$$\pi_{ij} = \begin{cases} 0 & \text{if } i-j \equiv \pm 1,\ldots,\pm\alpha \pmod N \\ \dfrac{n(n-1)}{N(N-2\alpha-1)} & \text{otherwise.} \end{cases}$$

While the Horvitz-Thompson estimator of μ, which reduces to the observed sample mean, $\bar{y}_s$, is an unbiased estimator, an unbiased estimator of the variance of $\hat{\mu}_{HT}$ cannot be obtained. Note that under a circular BSA(N, n, α)

$$\Delta_{ii} = \frac{n}{N}\left(\frac{N-n}{N}\right),$$

and

$$\Delta_{ij} = \begin{cases} -\left(\dfrac{n}{N}\right)^2 & \text{if } i-j = \pm 1,\ldots,\pm\alpha \pmod N \\ \dfrac{n}{N}\left[\dfrac{n(2\alpha+1)-N}{N(N-(2\alpha+1))}\right] & \text{otherwise.} \end{cases}$$

Making the appropriate substitutions and simplifying (1.2), one obtains

$$V(\hat{\mu}_{HT}) = \frac{1}{N^2}\left[\left(\frac{N-n}{n}\right)\sum_{i=1}^{N} y_i^2 - \sum\sum_{\substack{i,j \\ i-j=\pm1,\ldots,\pm\alpha \pmod N}} y_i y_j + \left(\frac{n(2\alpha+1)-N}{n(N-(2\alpha+1))}\right)\sum\sum_{\substack{i,j \\ i-j\neq\pm1,\ldots,\pm\alpha \pmod N}} y_i y_j\right]. \qquad (2.1)$$

Clearly the second summation in (2.1) is the cause of concern since there is no way of estimating cross-products of adjacent terms unbiasedly under a circular BSA(N, n, α). Further note that the other two summations in (2.1) can be estimated unbiasedly.

Theorem 2.1: A necessary condition of existence of a circular BSA(N, n, α) is

$$N \geq (2\alpha+1)n$$

for $n \geq 3$ and $\alpha \geq 1$, and

$$N \geq (2\alpha+1)n + 1,$$

for the following combinations of (n, α): $\{(n, 1); n \geq 5\}$, $\{(n, 2); 6 \leq n \leq 12\}$, $\{(n, 3); 5 \leq n \leq 9\}$, $\{(n, 4); n = 6, 7, 8\}$, and $\{(n, 5); n = 6, 7\}$.

Furthermore, the necessary conditions stated in Theorem 2.1 have been shown to be sufficient for certain combinations of n and α. For proofs of the above results, see Stufken (1993) and Wright and Stufken (2005), respectively.

A number of design generation techniques have been developed and are summarized in the following theorem.

Theorem 2.2: The existence of a circular BSA(N, n, α) implies the existence of:

1) a circular BSA($N + 2\alpha + 1$, n, α),

2) a linear BSA($N - \alpha$, $n - 1$, α) and a linear BSA($N - (\alpha + 1)$, $n - 1$, α), and

3) $2\xi + 1$ circular BSA(N', n', α)'s, where $N' = N - \xi, \ldots, N + \xi$ for $\xi = n - n'$ where $n' < n$.

The existence of a linear BSA(N, n, α) implies the existence of:

4) a circular BSA($N + \alpha$, n, α) and a circular BSA($N + \alpha + 1$, n, α).

The proof of 1) is a direct extension of a proof from Hedayat, Rao, and Stufken (1988a and 1988b), a proof of 2) can be found in Stufken (1993), and the proofs of 3) and 4) are detailed in Wright and Stufken (2005).

A cyclic construction method for the development of circular BSA(N, n, α) sampling plans was presented by Stufken (1993). The vast majority of directly identified circular BSA(N, n, α)'s possess a cyclical structure, and sets of generators for all such plans may be obtained at:

http://www.facstaff.bucknell.edu/jwright/Research/One-Dim_Designs/circular/.

3. Relative Efficiencies of BSA Sampling

Since BSA sampling yields an unbiased estimator of μ, relative efficiencies for estimating μ under simulated populations with various structures were obtained as a means of comparing BSA sampling with respect to simple random sampling. Circularly and linearly ordered populations of 60, 120, and 180 units were simulated using positive correlation structures with three strengths of correlation – strong, moderate, and weak. All simulated populations have a mean of 0 and a variance of 1.

BSA sampling strategies with n of 5 and 10 and α of 1 and 2 were investigated under all populations. For all combinations of sampling strategies and populations, the true variance of the Horvitz-Thompson estimator of μ was obtained. The relative efficiency of a sampling strategy (SS) for a given population with respect to simple random sampling is defined as

$$RE_{SRS}(SS) = \frac{V_{SRS}(\hat{\mu}_{HT})}{V_{SS}(\hat{\mu}_{HT})}. \qquad (3.1)$$

If (3.1) yields a value greater than 1, then one may conclude that the given sampling strategy is more efficient than simple random sampling in terms of estimation of μ. The resulting relative efficiencies are graphically depicted in Figure 3.1.

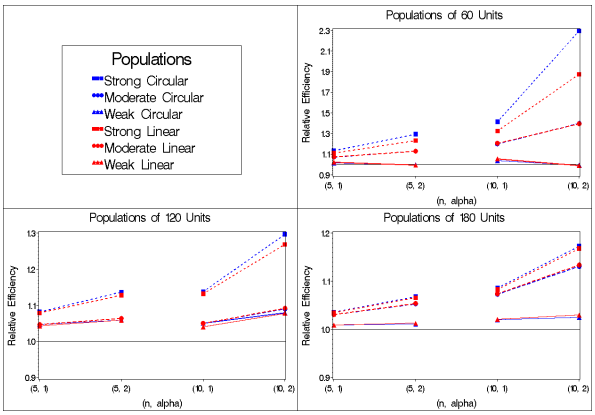

Figure 3.1 – Relative Efficiencies of BSA Sampling With Respect to SRS

As expected, BSA sampling is generally more efficient than simple random sampling for populations with positive correlation structure. While BSA sampling plans with $\alpha = 1$ are more efficient than the corresponding plans with $\alpha = 2$ for populations with the weak positive correlation structure, the opposite is true for populations with moderate and strong positive correlation structures. Furthermore, across all population structures, the difference in the relative efficiencies between circularly and linearly ordered

populations becomes negligible as population size increases.

4. Estimation Using Approximation Techniques

As mentioned previously, an unbiased estimate of the variance of the Horvitz-Thompson estimator for μ cannot be obtained from a single application of BSA sampling plans, and, specifically, the second summation in (2.1) is the cause of concern. Various approximations have been suggested and will be detailed in this section.

One common suggestion for approximating the variance of the Horvitz-Thompson estimator of μ is to replace the second summation in (2.5) by $\sum_{i,j \atop i-j=\pm 1,\ldots,\pm\alpha \pmod N} \sum \mu^2 = 2\alpha N\mu^2$. Under this approximation technique, the variance estimator of the Horvitz-Thompson estimator of μ is

$$\hat{V}_1(\hat{\mu}_{HT}) = \left(\frac{1}{Nn^2}\right)\left(\frac{N-(2\alpha+1)n}{2(n-1)}\right)\sum_{\substack{i,j\in s\\ i<j}}\sum (y_i - y_j)^2 + \left(\frac{2\alpha}{Nn}\right)\sum_{i\in s}(y_i - \bar{y}_s)^2, \quad (4.1)$$

which is clearly nonnegative provided that $N \geq (2\alpha + 1)n$. However, if responses for adjacent units are similar, one may achieve a better approximation through other techniques that exploit the relationship.

Hedayat, Rao, and Stufken (1988a, 1988b) proposed two approximations that attempt to exploit the ordering of units within the population. The first suggestion is to replace y_j in the second summation of (2.5) with y_i. Hence, $y_{i_1}y_{i_2} + y_{i_2}y_{i_1}$ is replaced by $y_{i_1}^2 + y_{i_2}^2$, where $i_1, i_2 \in s$. Since $y_{i_1}^2 + y_{i_2}^2 - 2y_{i_1}y_{i_2} = (y_{i_1} - y_{i_2})^2 > 0$, this approximation leads to a smaller quantity than the actual variance. As the number of pairs of adjacent units increases, which corresponds to an increase in α for fixed n, the difference between this approximation and the true variance will steadily increase. Under this approximation technique, the variance estimator of the Horvitz-Thompson estimator of μ is

$$\hat{V}_2(\hat{\mu}_{HT}) = \left(\frac{1}{Nn^2}\right)\left(\frac{N-(2\alpha+1)n}{2(n-1)}\right)\sum_{\substack{i,j\in s\\ i<j}}\sum (y_i - y_j)^2, \quad (4.2)$$

which is clearly nonnegative provided that $N \geq (2\alpha + 1)n$, and yields a smaller estimate than (4.1).

The second approximation proposed by Hedayat, Rao, and Stufken, which has been modified here for general α, is to replace y_j in the second summation of (2.5) with a weighted average of y_i and $y_{i+\alpha+1}$, for $j > i$. Specifically, y_j is replaced by

$$\frac{(\alpha+1-(j-i))y_i + (j-i)y_{i+\alpha+1}}{\alpha+1}. \quad (4.3)$$

Under this approximation technique, the variance estimator of the Horvitz-Thompson estimator of μ has the form

$$\hat{V}_3(\hat{\mu}_{HT}) = \left(\frac{1}{Nn^2}\right)\left(\frac{N-(2\alpha+1)n}{2(n-1)}\right)\sum_{\substack{i,j\in s\\ i<j}}\sum (y_i - y_j)^2 + \left(\frac{\alpha}{Nn}\right)\sum_{i\in s}(y_i - \bar{y}_s)^2 - \left(\frac{\alpha(N-(2\alpha+1))}{Nn(n-1)}\right)\sum_{i,i+\alpha+1\in s}(y_i - \bar{y}_s)(y_{i+\alpha+1} - \bar{y}_s). \quad (4.4)$$

As an observation, (4.4) may yield a negative variance estimate.

While the proposed estimators (4.1), (4.2), and (4.4) all attempt to estimate the variance of the Horvitz-Thompson estimator of μ, each variance estimator has some positive and negative qualities. For instance, if one wants a conservative estimate of the variance of the Horvitz-Thompson estimator of μ, then (4.1) may be used since the resulting variance estimator tends to typically be positively biased. In addition, note that one must obtain a sample that contains the i^{th} and the $(i + \alpha + 1)^{st}$ units of the population for the estimator (4.4) to possibly exploit the circular ordering of responses of units within the population to achieve a potentially "refined" estimate. As a means of comparing the performance of the three variance estimators, a simulation study was performed – the results of which will be reported in Section 5.

5. Simulation Study

Since unbiased estimates of the true variances under the various sampling strategies cannot be obtained, relative efficiencies of the sampling strategies are not sufficient measures of performance. For example, an approximation technique may consistently underestimate the true variance by a considerable amount even though the corresponding relative efficiency of the sampling strategy may be greater than 1. As a result, even though the corresponding relative efficiency may be greater than 1, simple random sampling may actually provide a more accurate estimate of the variance of the estimator of μ.

Utilizing the populations detailed in Section 3, circular BSA sampling plans for sample sizes of 5 and 10 using α's of 1 and 2 were developed. A total of 100,000 samples were simulated for each sampling plan, and estimates of the population mean, as well as the variance of the estimator, were obtained under the various population structures. For each of the 100,000 variance estimates obtained under a given population structure and sampling strategy, a relative measure of estimation was computed as

$$RM\left(\hat{V}_i\left(\hat{\mu}_{HT}\right)\right)=\frac{\hat{V}_i\left(\hat{\mu}_{HT}\right)-V\left(\hat{\mu}_{HT}\right)}{V\left(\hat{\mu}_{HT}\right)}. \qquad (5.1)$$

(5.1) is typically bounded below by –1; however, note that a lower value is obtained if the estimated variance is negative. As a means of summarizing the distribution of the relative measures for a given sampling strategy and population structure, "modified" 5-number summaries – 10th percentile, 25th percentile, median, 75th percentile, 90th percentile – were obtained and used to gauge the accuracy and precision of the corresponding estimator.

An estimator was classified as yielding generally "accurate" estimates when the median relative measure had a magnitude less than 0.15 and 25th and 75th percentiles of the distribution differed in sign with magnitudes of at least 0.10. If the 25th percentile of the distribution of relative measures was negative with a magnitude greater than 0.50, then it was determined that the corresponding variance estimator generally underestimated the true variance. Similarly, if the 75th percentile of the distribution of relative measures was positive with a magnitude greater than 0.50, then it was determined that the corresponding variance estimator generally overestimated the true variance. The corresponding inter-quartile range (IQR) of the distribution of relative measures was used as a measure of precision for the corresponding variance estimator.

As expected, each approximation technique yields more precise and accurate variance estimates corresponding to an increasing sampling fraction under large populations, and there is little difference between the corresponding distributions of relative measures for a given approximation technique across circularly and linearly ordered populations. While the IQR's of the distribution of relative measures become more consistent across the approximation techniques as the population size increases, under all scenarios, Approximation 2 provides the smallest IQR, followed by Approximation 3 and Approximation 1, respectively. However, when the differences are especially large, Approximation 2 tends to severely underestimate the true variance. Specifically, while Approximation 2 performs well under large populations with strong correlation structure with respect to the other approximation techniques, in general, it performs poorly under all other population structures.

Even though Approximation 1 grossly overestimates the true variance under small populations with strong correlation structure utilizing a large sampling fraction, the technique appears to provide acceptable estimates under the other population structures. Approximation 3 performs fairly well under populations with some form of correlation structure – the stronger the better. Furthermore, for scenarios when both Approximation 1 and Approximation 3 provide adequate estimates, Approximation 3 tends to not overestimate the true variance as consistently as Approximation 1.

As mentioned earlier, Approximation 3 may in fact yield a negative estimate. Table 5.1 details the percentage of negative estimates obtained by using Approximation 3 under various sampling plans. Note that no negative estimates were obtained under sampling plans with $\alpha = 1$ or when $n = 5$.

While there appears to be an increasing trend in the percentage of negative estimates obtained for a given population as $(2\alpha + 1)n$ approaches N, which should be expected given the inherent structure between the potential units of a sample created by excluding adjacent units, the observed percentages are mostly negligible for large N. Nonetheless, if

Table 5.1 – Percentage of Negative Estimates Obtained Using Approximation 3 Under BSA(N, 10, 2) Plans

Population	N = 60	120	180
Circular, Strong	11.37%	0.24%	0.05%
Circular, Moderate	2.46%	0.04%	< 0.01%
Circular, Weak	1.02%	0.06%	0.01%
Linear , Strong	7.82%	0.17%	0.03%
Linear, Moderate	3.01%	0.01%	0.02%
Linear, Weak	1.52%	0.06%	< 0.01%

$(2\alpha + 1)n \approx N$, the potential of obtaining a negative variance estimate under Approximation 3 may serve as a deterrent for using the approximation.

Additionally, coverage probabilities of desired 95% confidence intervals for the population mean were empirically estimated. For all three variance estimation techniques, desired 95% confidence intervals of the form

$$\hat{\mu}_{HT} \pm t_{(n-1);0.975} \sqrt{\hat{V}_i(\hat{\mu}_{HT})} \qquad (5.2)$$

were obtained for each of the 100,000 repetitions under all combinations of population structures and sampling plans. A sampling strategy and/or proposed interval estimator were determined to be inadequate if the corresponding empirical coverage reported greatly differed from 95%. The resulting empirical coverage probabilities are graphically depicted in Figure 5.1. Note that all simulations that yielded a negative estimate of variance using Approximation 3 were omitted from consideration.

In general, (5.2) appears to be a fairly conservative interval estimator for μ. While some scenarios yielded empirical coverage probabilities much lower than desired, they were limited to when a BSA(60, 10, 2) sampling plan was used. For all populations of 120 and 180 units, the empirical coverages ranged from 94.5% to 98.5%, regardless of sampling plan and utilized approximation technique. In addition, there appears to be little differences in the empirical coverages across sampling fractions and population orderings for the larger populations.

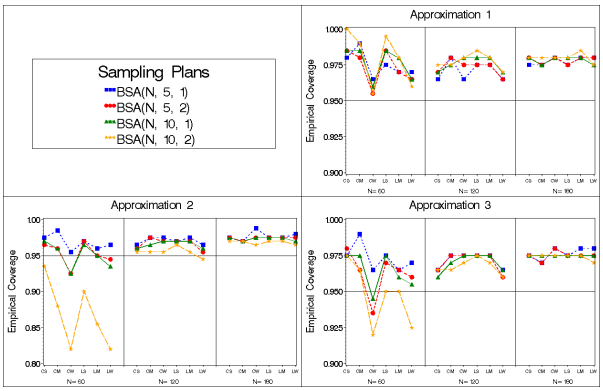

Figure 5.1 – Desired 95% Confidence Interval Coverage Probabilities

6. Closing Remarks

In general, BSA sampling plans are more efficient than simple random sampling under populations exhibiting a positive correlation structure. Specifically, significant gains in efficiency were obtained under BSA sampling plans with an $\alpha = 2$ applied to linearly or circularly ordered populations with moderate to strong positive correlation structure. While each of the three variance approximation techniques experience problems under small populations – potential for overestimation using Approximation 1, underestimation using Approximation 2, and negative estimates using Approximation 3 – the effects of the corresponding tendency diminishes as the population size increases. While Approximation 3 appears to be viable estimate under large populations with strong correlation structure, Approximation 1 would be preferred under large populations with weak to moderate correlation structure.

References

1. Hedayat, A. S., Rao, C. R., Stufken, J., 1988a. Sampling plans excluding contiguous units. J. Statist. Plann. Inference 19, 159 – 170.

2. Hedayat, A. S., Rao, C. R., Stufken, J., 1988b. Designs in survey sampling avoiding contiguous units. In: Krishnaiah, , P. R, Rao, C. R. (Eds.), Handbook of Statistics, vol. 6. Elsevier, Amsterdam, pp. 575 – 583.

3. Särndal, C.-E., Swensson, B., Wretman, J., 1992. Model Assisted Survey Sampling. Springer-Verlag, New York.
4. Stufken, J., 1993. Combinatorial and statistical aspects of sampling plans to avoid the selection of adjacent units. J. Combin. Inform. System Sci. 18, 81 – 92.

5. Stufken, J., Wright, J. H., 2001. Polygonal designs with blocks of size $k \leq 10$. Metrika 54, 179 – 184.

6. Wright, J. H., Stufken, J., 2005. New balanced sampling plans excluding adjacent units. J. Statist. Plann. Inference (submitted).