ESTIMATING THE VARIANCE OF THE HORVITZ-THOMPSON ESTIMATOR IN VARIABLE PROBABILITY, SYSTEMATIC SAMPLES

Stephen V. Stehman and W. Scott Overton
Cornell University and Oregon State University
Stephen V. Stehman, Biometrics Unit, 337 Warren Hall, Ithaca, NY 14853

Abstract

We examined two common estimators of variance of the Horvitz-Thompson estimator when the sampling design was random-order, systematic, with unequal probabilities, and fixed sample size. The variance estimator, v_{YG}, due to Yates and Grundy (1953) and Sen (1953) has gained favor in the statistical literature, based on certain theoretical and empirical results, over an estimator, v_{HT}, proposed by Horvitz and Thompson (1952). Both variance estimators require calculating pairwise inclusion probabilities. An approximate formula (Hartley and Rao, 1962) frequently has been used, but computing this approximation or the true pairwise inclusion probabilities is often impractical.

The properties of the variance estimators are shown to be associated with the population coefficient of variation of the ratios y/x, where y is the response variable of interest, and x is an auxiliary variable used to select the sample. The superiority of v_{YG} is most pronounced when cv(y/x) is very small. v_{HT} computed using the Hartley-Rao approximation formula has particularly poor properties in this circumstance. For larger cv(y/x), v_{YG} and v_{HT} have more similar behavior, and v_{HT} is sometimes better. A new approximation formula for the pairwise inclusion probabilities is given which has practical advantages over the Hartley-Rao formula. This new approximation improves the properties of v_{HT} especially when cv(y/x) is small.

The stream survey component of the National Surface Water Survey, conducted by the Environmental Protection Agency, is used as an example to illustrate some practical and theoretical concerns to be addressed when examining the variance estimation problem.

1.0 Estimators of Variance of the Horvitz-Thompson Estimator

We consider a finite population of size N. A response variable of interest, y_i, and an auxiliary variable, $x_i>0$, are defined for each element, $i=1,\ldots,N$, of the population. A sample of fixed size, n, will be selected without replacement from this population. Define a sampling rule, R, to be the protocol or scheme for selecting samples. Then R determines $\mathcal{S}$, the set of all possible samples (the sample space) under R, and $p_R(s)$, the probability that a particular sample s will be selected. The probability that unit i will be selected in the sample, the inclusion probability, is given by $\pi_i = \sum_{\{s:i\in s\}} p_R(s)$.

For our purposes, samples will be selected such that π_i is proportional to x_i; i.e., in sampling from a list, this results in $\pi_i = nx_i/T_x$, where T_x is the population total of the x's. This design will be denoted πpx. We restrict attention to the case in which $x_i \leq T_x/n$.

If $\pi_i>0$ $\forall$ i, the Horvitz-Thompson estimator,

$$\hat{T}_y = \sum_{i=1}^{n} \frac{y_i}{\pi_i}, \tag{1.1}$$

is unbiased for the population total, $T_y = \sum_{i=1}^{N} y_i$, and has variance

$$\begin{aligned} V(\hat{T}_y) &= \sum_{i=1}^{N} \left(\frac{y_i}{\pi_i}\right)^2 (1-\pi_i)\,\pi_i \\ &\quad + \sum_{i=1}^{N}\sum_{j\neq i}^{N} \left(\pi_{ij} - \pi_i\pi_j\right) \frac{y_i}{\pi_i}\frac{y_j}{\pi_j} \end{aligned} \tag{1.2}$$

$$= \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \left(\pi_i\pi_j - \pi_{ij}\right) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2, \tag{1.3}$$

where $\pi_{ij} = \sum_{\{s:(i,j)\in s\}} p_R(s)$

is the pairwise inclusion probability. Equation (1.2) holds in general, while (1.3) holds only if the sample size is fixed.

Two estimators of $V(\hat{T}_y)$ have been proposed, based on the formulas (1.2) and (1.3). Both estimators are unbiased if $\pi_{ij}>0$ for all pairs i and j in the population. The estimators are:

$$v_{HT} = \sum_{i=1}^{n}\left(\frac{y_i}{\pi_i}\right)^2(1-\pi_i) + \sum_{i=1}^{n}\sum_{j\neq i}^{n}\left(\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}}\right)\frac{y_i y_j}{\pi_i\pi_j} \tag{1.4}$$

(Horvitz and Thompson (1952)), and

$$v_{YG} = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \left(\frac{\pi_i\pi_j - \pi_{ij}}{\pi_{ij}}\right) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2 \tag{1.5}$$

(Yates and Grundy (1953), and Sen (1953)).

v_{YG} frequently has been claimed superior to v_{HT} on the basis of fewer negative estimates and smaller sampling variance. Theoretical comparison of the two variance estimators has yielded only limited insight. It is known that when the ratio $r_i = y_i/x_i$ is constant for all $i=1,\ldots,N$, $V(\hat{T}_y) \equiv 0$. In this situation, $v_{YG} \equiv 0$, but v_{HT} does not identically equal 0; being unbiased, v_{HT} therefore must be capable of negative values. Thus, at least for populations in which y_i is nearly proportional to x_i, v_{YG} would appear to have smaller sampling variance. This is the important case in which πpx sampling is very efficient.

Several empirical studies have shown advantages for v_{YG}. Rao and Singh (1973) studied 34 natural populations, selecting samples of size n=2, using Brewer's πpx method. They found v_{HT} frequently resulted in negative estimates, and that the sampling variance of v_{HT} was much larger for many of their populations. Similar results were obtained by Cumberland and Royall (1981). They examined 6 populations using random-order, variable probability, systematic sampling to select samples of size n=32.

Variance estimation for variable probability sampling is complicated by the difficulty in computing the π_{ij}'s. Different πpx designs can have quite different π_{ij}'s. A convenient and widely used fixed sample size, πpx design is designated *variable* *probability systematic* (*vps*), and this design will be the focus of our attention. Hidiriglou and Gray (1980) provided a FORTRAN program for computing the exact (or true) π_{ij}'s for random-order, *vps* sampling. Computing times for these exact π_{ij}'s were excessively high for our purposes. The approximate formula for the π_{ij}'s under random-order, *vps* sampling due to Hartley and Rao (1962) has commonly been used in this circumstance (for example, Cumberland and Royall, 1981). A disadvantage of the exact formula and the Hartley-Rao formula is that x_i must be known for all population elements, not just the sample elements.

2.0 An Example: The National Surface Water Surveys

Estimation and design issues encountered in the National Surface Water Surveys (NSWS), and particularly the National Stream Survey (Overton, 1985, 1987, Messer et al, 1986) illustrate some of the practical and theoretical issues concerning variance estimators of the Horvitz-Thompson estimator. We consider a small part of the actual stream survey design and analysis, and suppress some details of the survey to simplify discussion.

The Phase I Stream Survey design was a *vps* sample. Sampling units were selected using a point/area sampling frame imposed on topographic maps of the target area. Each point in the square dot grid was associated with a target reach or "no reach", where a reach was a well-defined stream segment. This protocol resulted in reaches being sampled with probability proportional to direct watershed area.

The stream survey design is a *fixed configuration*, *vps* sample, not a random-order, *vps* sample. However, the approach used to estimate variances in the stream survey was to treat the observed configuration as random. The variance estimators employed result from use of π_{ij}'s appropriate to a random-order, *vps* design. This approach is based on the perception that, for many natural populations, the systematic patterns generated by the dot-grid sampling procedure do not preclude treating the sample as though it were taken from a randomized list. A study of the appropriateness of this approach in the stream survey is currently underway. Preliminary indications are favorable, and the report of those studies will appear elsewhere (Stehman and Overton, 1987). The present paper deals only with behavior of variance estimators under random-order, *vps* sampling.

The stream survey had several concerns common to surveys using this sampling design. The multiple-objective nature of the survey called for a good, general strategy of estimation. Requiring different variance estimators for different response variables was not practical.

It is important to note that the sampling design of the stream survey was chosen for ease of implementation and other operational advantages of the design. Efficiency of the πpx design was a secondary consideration. Further, it would be unrealistic to expect the πpx design to be efficient for all of the many chemical and physical attributes of interest. Thus we are interested in properties of the variance estimators, v_{HT} and v_{YG}, under a broad range of conditions, not restricted solely to circumstances in which the πpx design is known to be efficient.

Another practical concern in the stream survey was that the auxiliary variable, direct watershed area, was measured only on the sample units. The exact pairwise inclusion formula and the Hartley-Rao approximate formula were therefore not available for use. A formula for the pairwise inclusion probabilities was needed that was computationally feasible and did not require knowledge of all x_i's in the population.

3.0 Results

Notation:

v_{HT} (or v_{YG}) = Horvitz-Thompson (or Yates-Grundy) variance estimator calculated using (exact) π_{ij}

π^o_{ij} = approximate formula for π_{ij} described below

v^o_{HT} = Horvitz-Thompson variance estimator calculated using π^o_{ij}

v^o_{YG} = Yates-Grundy variance estimator calculated using π^o_{ij}

π^{hr}_{ij} = approximate formula for π_{ij} derived in Hartley and Rao (1962)

v^{hr}_{HT} = Horvitz-Thompson variance estimator calculated using π^{hr}_{ij}

v^{hr}_{YG} = Yates-Grundy variance estimator calculated using π^{hr}_{ij}

$\hat{v}$ = generic designation for any of the above variance estimators

3.1 Pairwise Inclusion Probability Formulas

The formula for approximating the pairwise inclusion probabilities is derived in terms of random-order, *vps* sampling from a list frame (Overton, 1985):

$$\pi^o_{ij} = \frac{(n-1)\pi_i\pi_j}{n-\frac{1}{2}(\pi_i+\pi_j)} \tag{3.1}$$

$$= \frac{2(n-1)\pi_i\pi_j}{2n-\pi_i-\pi_j} \tag{3.2}$$

Note that in (3.1) and (3.2) the population total, T_x, does not appear, so that this form is appropriate for the stream survey, where T_x is unknown. When $x_i=1$ for all $i=1,\dots,N$, then $\pi^o_{ij} = n(n-1)/N(N-1)$, the pairwise inclusion probability appropriate for a simple random sample. Thus the approximation gives the correct result in this simple case.

The Hartley-Rao formula is much more complicated. The truncated form usually used to derive theoretical results (see equation (5.20) of Hartley and Rao (1962), and Cumberland and Royall (1981) for examples) is:

$$\pi^{hr}_{ij} = \frac{(n-1)\pi_i\pi_j}{\left[n-\pi_i-\pi_j+\sum_{k=1}^{N}\pi_k^2/n\right]} \tag{3.3}$$

In the simulation studies described in Section 4.0, equation (5.15) of Hartley and Rao (1962) was used instead of the truncated form (3.3) above. Note the similarities between (3.1) and (3.3).

3.2 Properties of the Variance Estimators

The issue of sampling variability is particularly critical since v_{YG} has been claimed superior to v_{HT} on this criterion. Rewriting v_{YG} as follows,

$$v_{YG} = \sum_{i=1}^{n} \left(\frac{y_i}{\pi_i}\right)^2 \sum_{j\neq i}^{n} \left[\frac{\pi_i\pi_j - \pi_{ij}}{\pi_{ij}}\right] + \sum_{i=1}^{n}\sum_{j\neq i}^{n} \left[\frac{\pi_{ij} - \pi_i\pi_j}{\pi_{ij}}\right] \frac{y_i}{\pi_i}\frac{y_j}{\pi_j}, \tag{3.4}$$

it is seen that v_{YG} and v_{HT} (equation 1.4) have very similar forms, the difference being that v_{YG} uses the term $\sum_{j\neq i}^{n} \left[\frac{\pi_i\pi_j - \pi_{ij}}{\pi_{ij}}\right]$ in the first summation in place of the term $(1-\pi_i)$ in v_{HT}.

The quantity $\sum_{j\neq i}^{n} \left[\frac{\pi_i\pi_j - \pi_{ij}}{\pi_{ij}}\right]$ is an unbiased estimator of $(1-\pi_i)$, the expectation taken over the sample space conditioned on $i\varepsilon s$. Thus the essential difference between v_{YG} and v_{HT} is that v_{YG} replaces the term $(1-\pi_i)$ in v_{HT} with a random variable having *expectation* $(1-\pi_i)$. Replacing the known quantity $(1-\pi_i)$ with this random variable induces a favorable "cancellation" in v_{YG}, under certain circumstances, as follows. Rewriting (3.4),

$$v_{YG} = \sum_{i=1}^{n} \left(\frac{y_i}{\pi_i}\right) \sum_{j\neq i}^{n} \left[\frac{\pi_i\pi_j - \pi_{ij}}{\pi_{ij}}\right] \left(\frac{y_i}{\pi_i}\right) - \sum_{i=1}^{n} \left(\frac{y_i}{\pi_i}\right) \sum_{j\neq i}^{n} \left[\frac{\pi_i\pi_j - \pi_{ij}}{\pi_{ij}}\right] \left(\frac{y_j}{\pi_j}\right), \tag{3.5}$$

when y/x (and hence y/π) is nearly constant for all units in the population, the terms in the two summations over j will essentially cancel each other. v_{YG} will be nearly zero with very little sampling variability. The sampling variability of v_{YG} should increase as the variability in the ratios y/x increases.

The case of zero variability in the ratios ($y_i/x_i=\beta$ for $i=1,\ldots,N$) is of special interest. Under this circumstance $v_{YG} \equiv 0$ (for any representation of π_{ij}). But $v^{o}_{HT} \equiv 0$ (proof omitted), while v^{hr}_{HT} and v_{HT} are not identically 0. Thus we expect that v^{o}_{HT} would perform similarly to v_{YG}, and better than v_{HT} or v^{hr}_{HT}, in populations having small variation in the y/x ratios.

4.0 Design of Simulation Studies

We used two simulation studies to explore the properties of the variance estimators. For the first set of simulations, designated Group I, we examined two stream survey data sets and two populations from the statistical literature (Table 1). One of these populations, Sales, was used by Cumberland and Royall (1981) to demonstrate the superiority of v_{YG}.

Table 1. Group I Populations

Population	N	cv(x)	cv(y)	ρ(x,y)	cv(y/x)
Sales[1]	327	1.20	1.19	.99	.14
Paddy[2]	108	0.69	0.78	.79	.39
Stream1[3]	100	0.92	0.72	.86	.71
Stream2[3]	100	0.66	0.52	.81	.41

[1] Cumberland and Royall (1981), x = gross sales of corporation in 1974, y = sales in 1975

[2] Murthy (1967), x = geographical area, y = area under winter paddy

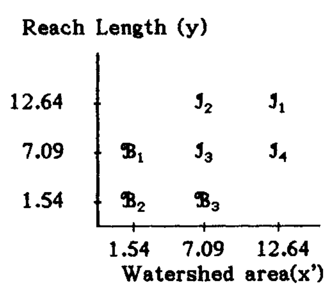
[3] x = direct watershed area, y = length of reach

We undertook the Group II simulations as a systematic exploration of a structured set of populations. By standardizing some population parameters, we hoped to associate properties of the variance estimators with key attributes of the populations. This approach also permitted expanding the scope of populations previously studied in the statistical literature.

For the Group II simulations, a baseline population was purposefully selected from a stream survey data set (x=direct watershed area, y=reach length, N=72). A modified auxiliary variable, x', was derived from the original auxiliary variable via the transformation $x' = \sqrt{(V_y/V_x)}\, x$, where V_x and V_y were the population variances of x and y respectively. This modification of the auxiliary variable equalized the variances of x' and y, created a population with major axis of slope 1, and maintained the same probability structure on the sample space achieved by the original x. By adding (or subtracting) increments of 15 to x' and/or y, we shifted the baseline population through the "population space". Shifting the population in this way maintains the same correlation of x' and y and the slope of the major axis remains 1. However, these shifts change cv(y/x), and additive shifts in x' change the inclusion probabilities.

Populations with ρ(x,y) values of 0.53 and 0.99 were created from the original baseline population, and these populations were also shifted through the population space. Based on the location of their population centroids, the Group II populations were classified as $\mathfrak{B}$=boundary populations or $\mathfrak{I}$=interior populations (see Figure I).

Figure I. Population Space Centroids (ρ=.82)



The boundary populations have high cv(y/x'), while the interior populations have low cv(y/x'). For a given location in the population space, cv(y/x') decreases with increasing ρ(x',y). (Notation identifying populations: subscripts denote the particular population within $\mathfrak{B}$ or $\mathfrak{I}$, superscripts denote ρ(x',y): lo=.53, m=.82, hi=.99.)

Table 2. Group II Populations: cv(y/x')

Population	ρ=.53	ρ=.82	ρ=.99
$\mathfrak{B}_1$	.88	.80	.49
$\mathfrak{B}_2$	1.11	.59	.12
$\mathfrak{B}_3$	.61	.56	.44
$\mathfrak{I}_1$	.07	.05	.01
$\mathfrak{I}_2$	.11	.08	.05
$\mathfrak{I}_3$	.13	.08	.02
$\mathfrak{I}_4$	.12	.09	.05

The sampling design used in the simulations was random-order, *vps* sampling. Detailed descriptions of this sampling scheme appear in Hartley and Rao (1962) and Cumberland and Royall (1981). All populations were sufficiently large that exact π_{ij}'s were not computationally feasible, so the comparisons were among v^{o}_{HT}, v^{hr}_{HT}, v^{o}_{YG}, and v^{hr}_{YG}. Version 1.49 of the GAUSS Mathematical and Statistical System (Aptech Systems, Inc., Kent, WA) was used to run the simulations on IBM XT or AT machines.

5.0 Results of the Simulation Studies

The criteria for comparing the variance estimators are:

1) estimated MSE
2) confidence interval coverage achieved using the variance estimators, with intervals calculated as $\hat{T}_y \pm 1.96\sqrt{\hat{v}}$
3) relative bias, estimated by:
   $$\text{rel bias} = [\hat{E}(\hat{v}) - \hat{V}(\hat{T}_y)]/\hat{V}(\hat{T}_y),$$
   where $\hat{E}(\hat{v})$ was the simulated expected value of $\hat{v}$, and $\hat{V}(\hat{T}_y)$ was an unbiased estimate of $V(\hat{T}_y)$ obtained from the simulations
4) proportion of samples resulting in negative $\hat{v}$.

The results of the simulations are based on 5,000 replications of the sampling procedure. (Note: Tables have been condensed showing results only for some sample sizes and, in the Group II simulations, some correlations. Please contact the authors for copies of complete tables.)

5.1 Group I Simulations

The results of Section 3.2 predict that v_{YG} should outperform v_{HT} when the variability of the ratios y/x is small. As the variability in the y/x ratios increases, no apparent advantage is expected for v_{YG}. Further, when cv(y/x) is low, v^{o}_{HT} should have much smaller MSE and fewer negative estimates, compared to v^{hr}_{HT}. The predictions were confirmed by the Group I simulations. The relevant MSE comparisons and confidence interval coverages are presented in Table 3.

TABLE 3. Results of Group I Simulations

Ratios of Mean Square Errors (n=16)

Population	*a*	*b*	*c*	*d*
Sales	13.28	0.09	0.95	1.29
Paddy	1.28	1.01	0.89	1.46
Stream1	0.99	1.12	0.74	1.50
Stream2	0.97	1.21	0.93	1.26

a $\text{MSE}(v^{hr}_{HT})$ / $\text{MSE}(v^{hr}_{YG})$

b $\text{MSE}(v^{o}_{HT})$ / $\text{MSE}(v^{hr}_{HT})$

c $\text{MSE}(v^{o}_{YG})$ / $\text{MSE}(v^{hr}_{YG})$

d $\text{MSE}(v^{o}_{HT})$ / $\text{MSE}(v^{o}_{YG})$

Confidence Interval Coverage (nominal 95%)

Population	v^{hr}_{HT}	v^{hr}_{YG}	v^{o}_{HT}	v^{o}_{YG}
Sales	63	94	95	94
Paddy	92	93	94	93
Stream1	87	88	89	88
Stream2	87	87	89	87

The properties of v^{o}_{YG} and v^{hr}_{YG} were very similar in the Group I populations. Confidence interval coverage was identical, but v^{o}_{YG} uniformly outperformed v^{hr}_{YG} in terms of MSE. Comparing v^{hr}_{YG} to v^{hr}_{HT}, only in population Sales, where cv(y/x) is very small, is v^{hr}_{YG} clearly superior. The two stream populations provide examples of populations in which v^{hr}_{HT} and v^{hr}_{YG} have very similar properties.

v^{o}_{HT} had much better properties than v^{hr}_{HT} in population Sales. MSE and confidence interval coverage of v^{o}_{HT} were dramatically better than those of v^{hr}_{HT}, and the proportion of negative estimates dropped from .32 (n=16) for v^{hr}_{HT} to 0 for v^{o}_{HT}. In the other three populations, v^{hr}_{HT} had slightly smaller MSE while v^{o}_{HT} had slightly better coverage. Finally, comparing v^{o}_{HT} and v^{o}_{YG}, v^{o}_{YG} had uniformly better MSE but slightly poorer coverage than v^{o}_{HT}.

Generalizations from the Group I simulations are:

a) The Horvitz-Thompson variance formula is much better behaved, relative to the Yates-Grundy formula, in populations Stream1 and Stream2 than in populations Sales and Paddy; population Sales demonstrates the worst in v^{hr}_{HT}.
b) The best estimator in terms of MSE is v^{o}_{YG}.
c) The best estimator in terms of confidence interval coverage is v^{o}_{HT}.

5.2 Group II Simulations

Differences in behavior of the variance estimators were identifiable with the two population classes, $\mathfrak{B}$ and $\mathfrak{J}$. Considering MSE, v^{hr}_{YG} was far superior to v^{hr}_{HT} in the interior populations, but v^{hr}_{HT} was slightly better in the boundary populations. v^{o}_{YG} had smaller MSE than v^{o}_{HT} in all populations except $\mathfrak{B}_3^{m}$, but only in population $\mathfrak{J}_2$ was the difference very dramatic. Comparing the same variance estimator with different π_{ij} formulas, MSE of v^{o}_{HT} was much smaller than the MSE of v^{hr}_{HT} in the interior populations, while v^{hr}_{HT} was slightly better than v^{o}_{HT} in the boundary populations. v^{o}_{YG} and v^{hr}_{YG} were virtually identical in the interior populations, but v^{o}_{YG} had slightly smaller MSE than v^{hr}_{YG} in the boundary region, particularly in populations $\mathfrak{B}_1^{lo}$ and $\mathfrak{B}_1^{m}$, and $\mathfrak{B}_2^{lo}$ and $\mathfrak{B}_2^{m}$.

TABLE 4. Results of Group II Simulations

Ratios of Mean Square Errors (n=16, ρ=.82 only)

Population	*a*	*b*	*c*	*d*
$\mathfrak{B}_1$	0.96	0.97	0.75	1.57
$\mathfrak{B}_2$	0.86	1.38	0.83	1.43
$\mathfrak{B}_3$	0.99	1.23	0.99	0.97
$\mathfrak{J}_1$	85.55	0.02	1.02	1.79
$\mathfrak{J}_2$	38.58	0.17	1.08	5.99
$\mathfrak{J}_3$	31.10	0.05	1.01	1.65
$\mathfrak{J}_4$	6.92	0.17	0.98	1.01

columns *a,b,c,d* as in Table 3

Patterns in MSE were also associated with sample size. MSE of v^{hr}_{HT} relative to the other variance estimators became increasingly worse with increasing sample size in the interior populations. Similarly, the MSE of v^{o}_{HT}, relative to v^{o}_{YG} and v^{hr}_{YG}, generally increased with sample size, though this pattern was not evident in $\mathfrak{B}_2^{lo}$, $\mathfrak{B}_3^{lo}$, or $\mathfrak{J}_4^{lo}$. No association was evident between sample size and the ratio of MSE's of v^{o}_{YG} and v^{hr}_{YG} in the interior region, but for populations $\mathfrak{B}_1$ and $\mathfrak{B}_2$, the MSE advantage of v^{o}_{YG} over v^{hr}_{YG} increased with sample size.

Confidence interval coverage was dependent on the choice of π_{ij} approximation, but the results followed a pattern similar to that observed for MSE. The major difference in coverage was observed in the interior populations, where v_{HT}^{hr} had substantially poorer coverage than any of the other three variance estimators. For the boundary populations, all 4 variance estimators provided similar coverage.

Table 5. Results of Group II Simulations

Confidence Interval Coverage (%) (n=16)

Results using π_{ij}^{hr}

	$\rho=.53$		$\rho=.82$		$\rho=.99$	
Popn	v_{HT}^{hr}	v_{YG}^{hr}	v_{HT}^{hr}	v_{YG}^{hr}	v_{HT}^{hr}	v_{YG}^{hr}
$\mathfrak{B}_1$	87	85	87	85	90	89
$\mathfrak{B}_2$	90	90	92	93	59	93
$\mathfrak{B}_3$	93	93	93	93	93	93
$\mathfrak{J}_1$	76	93	62	94	49	93
$\mathfrak{J}_2$	84	94	75	94	63	93
$\mathfrak{J}_3$	86	93	69	93	52	93
$\mathfrak{J}_4$	88	93	82	93	70	93

Results using π_{ij}^{o}

	$\rho=.53$		$\rho=.82$		$\rho=.99$	
Popn	v_{HT}^{o}	v_{YG}^{o}	v_{HT}^{o}	v_{YG}^{o}	v_{HT}^{o}	v_{YG}^{o}
$\mathfrak{B}_1$	88	84	89	84	92	89
$\mathfrak{B}_2$	91	89	93	92	92	93
$\mathfrak{B}_3$	93	93	92	93	93	93
$\mathfrak{J}_1$	95	93	95	94	93	93
$\mathfrak{J}_2$	96	93	97	94	98	93
$\mathfrak{J}_3$	95	93	95	93	92	93
$\mathfrak{J}_4$	93	93	91	93	88	93

None of the simulations resulted in a sample for which v_{YG}^{o} or v_{YG}^{hr} was negative. The proportion of negative v_{HT}^{hr} was greater for the interior populations than for the boundary populations. Further, the proportion of negative estimates increased with $\rho(x,y)$. The proportion of negative v_{HT}^{o} was less than .005 for all populations and sample sizes.

Table 6. Proportion of Samples with Negative v_{HT}^{hr} ($\rho=.82$)

	n			
Population	4	8	16	24
$\mathfrak{B}_1$	.00	.00	.00	.00
$\mathfrak{B}_2$	.01	.00	.00	.00
$\mathfrak{B}_3$	.00	.00	.00	.00
$\mathfrak{J}_1$	.26	.30	.34	.39
$\mathfrak{J}_2$	.15	.16	.22	.30
$\mathfrak{J}_3$	.15	.17	.24	.31
$\mathfrak{J}_4$	.07	.07	.10	.15

6.0 Conclusions

Our results show that the superiority of v_{YG} over v_{HT} previously reported in the statistical literature is attributable partly to the restricted range of populations studied, and partly to the poor behavior of the Hartley-Rao approximation in the Horvitz-Thompson variance estimator. Cumberland and Royall (1981) identified the superiority of v_{YG}^{hr} over v_{HT}^{hr} in populations appropriately modelled by regression through the origin. Our results clarify the picture by generalizing the population space, and by identifying an association between cv(y/x) and superiority of v_{YG}^{hr}. When cv(y/x) is small, a condition in which πpx sampling is most efficient, v_{YG}^{hr} is superior. When cv(y/x) is larger, the behavior of v_{HT}^{hr} is comparable to, and in some cases better than v_{YG}^{hr}.

Introduction of the new approximation, π_{ij}^{o}, provides a different assessment. The properties of v_{HT}^{o} were much better than the properties of v_{HT}^{hr} when cv(y/x) was small, and v_{YG}^{o} had smaller MSE than v_{YG}^{hr} when cv(y/x) was large. Thus π_{ij}^{o} improved both variance estimators in those circumstances in which the estimator performed relatively poorly using π_{ij}^{hr}. Bias of the variance estimators was usually larger using π_{ij}^{o} than using π_{ij}^{hr}, but we consider confidence interval coverage and MSE more meaningful criteria for assessing these variance estimators. In no circumstance did π_{ij}^{o} lead to substantially poorer MSE or confidence interval coverage for either variance estimator.

In the National Stream Survey, v_{HT}^{o} provided a convenient and computationally efficient variance estimator. Variance formulas using either π_{ij}^{hr} or the exact π_{ij}'s were not possible in this survey. Establishing that v_{HT}^{o} had MSE and confidence interval coverage comparable to, or better than the other variance estimators studied, in populations of the nature of the stream populations, provided additional justification for the use of v_{HT}^{o} in the stream survey.

Acknowledgments

This paper is a contribution of the Aquatic Effects Research Program, funded by the U. S. Environmental Protection Agency, through the National Acid Precipitation Assessment Program. This paper has not been subjected to EPA's peer and policy review, and therefore does not necessarily reflect the views of the Agency. The authors thank Charles E. McCulloch, Cornell University, for helpful comments on this manuscript.

References

Brewer, K. R. W., and Hanif, M. (1983). *Sampling with Unequal Probabilities,* New York: Springer-Verlag.

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

Cumberland, W. G., and Royall, R. M. (1981). Prediction models and unequal probability sampling. *J. Roy. Statist. Soc. Ser. B* **43**, 353-367.

Hartley, H. O., and Rao, J. N. K. (1962). Sampling with unequal probability and without replacement. *Ann. Math. Statist.* **33**, 350-374.

Hidiriglou, M. A., and Gray, G. B. (1980). Construction of joint probability of selection for systematic p.p.s. sampling. *Applied Statistics* **29**, 107-112.

Horvitz, D. G., and Thompson, M. E. (1952). A generalization of sampling without replacement from a finite universe. *J. Amer. Statist. Assoc.* 47, 663-685.

Messer, J.J., C.W. Ariss, J.R. Baker, S.K. Drousé, K.N. Eshleman, P.N. Kaufmann, R.A. Linthurst, J.M. Omernik, W.S. Overton, M.J. Sale, R.D. Shonbrod, S.M. Stanbaugh, and J.R. Tutshall, Jr. (1986). *National Surface Water Survey: National Stream Survey, Phase I — Pilot Survey.* EPA—600/4—86—026, U.S. Environmental Protection Agency, Washington, D.C.

Murthy, M. N. (1967). *Sampling Theory and Methods,* Calcutta: Statistical Publishing Society.

Overton, W. S. (1985). *A Sampling Plan for Streams in the National Stream Survey.* Technical Report 114, Department of Statistics, Oregon State University, Corvallis, Oregon, 97331.

Overton, W. S. (1987). *Phase II Analysis Plan, National Lake Survey — Working Draft, April 15, 1987.* Technical Report 115, Department of Statistics, Oregon State University, Corvallis, Oregon, 97331.

Rao, J. N. K., and Singh, M. P. (1973). On the choice of estimator in survey sampling. *Austral. J. Statist.* 15, 95-104.

Sen, A. R. (1953). On the estimate of the variance in sampling with varying probabilities. *J. Indian Soc. Agric. Statist.* 7, 119-127.

Stehman, S. V., and Overton, W. S. (1987). *Results of simulation studies of the estimation methodology prescribed for the National Stream Survey.* Technical Report 118, Department of Statistics, Oregon State University, Corvallis, Oregon, 97331.

Wolter, K. M. (1985). *Introduction to Variance Estimation,* New York: Springer-Verlag.

Yates, F., and Grundy, P. M. (1953). Selection without replacement from within strata with probability proportional to size. *J. Roy. Statist. Soc. Ser. B* 15, 235-261.