

**A Comparison of Variance Estimators of the Horvitz-Thompson Estimator
in Random-Order, Variable Probability, Systematic Sampling***

Stephen V. Stehman and W. Scott Overton

BU-935-MA

Revised July 1988

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 the variance estimator, v_{HT} , proposed by Horvitz and Thompson (1952). Variance estimation is complicated by the need for computing 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. An approximation formula for the pairwise inclusion probabilities, which avoids some of the practical disadvantages of the Hartley-Rao and exact formulas is assessed.

The properties of the variance estimators are shown to be associated with the 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. The new approximation to the pairwise inclusion probabilities 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.

* 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. Part of this material appeared in Stehman and Overton (1987), Estimating the Variance of the Horvitz-Thompson Estimator in Variable Probability, Systematic Samples, *Proceedings of the Section on Survey Research Methods of the American Statistical Association*, pp. 743-748.

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, u_i , 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 the sample space \mathcal{Y} , the set of all possible samples under R , and $p_R(s)$, the probability that a particular sample s will be selected. The probability that the i^{th} element will be selected in the sample is given by the inclusion probability $\pi_i = \sum_{\{s: i \in s\}} p_R(s)$. For our purposes, samples will be selected such that π_i is proportional to x_i . 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 . Combination of the πpx design with the Horvitz-Thompson estimator defines the sampling strategy addressed here. We restrict attention to the case in which $x_i \leq T_x/n$ for all i .

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

$$\hat{T}_y = \sum_{i \in s} \frac{y_i}{\pi_i}, \quad (1.1)$$

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

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

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

where $\pi_{ij} = \sum_{\{s: (i,j) \in s\}} p_R(s)$, the pairwise inclusion probability for elements i and j . Equation (1.2) holds in general, while (1.3) holds only if the sample size is fixed.

If $\pi_{ij} > 0$ for all pairs i and j in the population, two unbiased estimators of $V(\hat{T}_y)$ have been proposed, based on the formulas (1.2) and (1.3). Both variance estimators are identifiable as Horvitz-Thompson estimators of the form given in equation (1.1). The estimators are:

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

(Horvitz and Thompson (1952)), and

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

(Yates and Grundy (1953), and Sen (1953)), where the summations in (1.4) and (1.5) represent summations indexed on the sample units.

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, $V(\hat{T}_y) \equiv 0$. 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 π px 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 π px method. They found v_{HT} frequently resulted in negative estimates, and that the sampling variance of v_{YG} was much smaller than the variance of v_{HT} 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 $n=32$.

Godambe and Joshi (1965) found v_{HT} is admissible, but they were unable to establish the admissibility of v_{YG} . Sankaranarayanan (1980) claimed v_{YG} is admissible in the class of all non-negative, unbiased quadratic estimators, but Biyani (1980) contradicted this result and showed that there exist fixed sample size designs, with $n>2$, for which v_{YG} is inadmissible.

Several authors have addressed the issue of hyperadmissibility in the context of

variance estimators of \hat{T}_y . Rao and Singh (1973) claimed that v_{HT} is the unique hyperadmissible estimator of $V(\hat{T}_y)$. Both are hyperadmissible, in the appropriate circumstances, because both v_{HT} and v_{YG} are identifiable as Horvitz-Thompson estimators (1.1) when all $\pi_{ij} > 0$, and because the Horvitz-Thompson estimator is hyperadmissible for fixed sample size designs (cf., Cassel *et al* (1977), Theorem 3.4).

Cassel *et al* (1977) discounted the meaningfulness of hyperadmissibility on the grounds that survey objectives are not likely to involve many subpopulations. As we indicate in the next section, the National Stream Survey (NSS) required investigation of many subpopulations, so that hyperadmissibility becomes a dominant criterion. We now are aware (Overton and Stehman, 1987) that v_{YG} does not apply to some subpopulation identities, so that hyperadmissibility is a real issue in the NSS.

Variance estimation for variable probability sampling is complicated by the difficulty in computing the π_{ij} 's. Different π px designs can have quite different π_{ij} 's. A convenient and widely used fixed sample size, π px 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 Stream Survey (NSS)

Estimation and design issues encountered in 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 NSS design and analysis, and suppress some details of the survey to simplify discussion. The *vps* design has widespread utility, so our comments about the

specific application to the NSS are pertinent to many other situations.

The Phase I NSS design was a variable probability, systematic sample. Sampling units were selected using a point/area sampling frame imposed on topographic maps of the target area. Each point in a 64 sq. mi. 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 NSS design was 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. That is, 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. The appropriateness of this approach in the NSS has been examined in Overton and Stehman (1987) and Stehman and Overton (1987a), while Osborne (1942), Milne (1959), Payendeh (1970), and Wolter (1985) have examined this approach in other circumstances. The present paper deals with behavior of variance estimators for random-order, *vps* sampling only.

The NSS demonstrated several concerns common to surveys using this sampling design. The multiple-objective nature of the survey called for a good, general strategy of estimation. It is important to note that the sampling design of the NSS was chosen for ease of implementation and other operational advantages of the design. Efficiency of the π_{px} strategy was a secondary consideration. Further, it would be unrealistic to expect the π_{px} design to be efficient for all of the many chemical and physical attributes of interest. Some of the attributes measured may be highly correlated with the π_i 's, but many are surely not. 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 π_{px} strategy 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.

Computational convenience, in the form of a simple recursive formula, was required of the variance estimator because of the large scale of the survey and the nature of the statistics generated. Estimation of frequency distributions, the number of streams having an attribute equal to or less than a particular prescribed value, was a key activity of the stream survey analysis. Frequency distributions were estimated for many physical and chemical stream attributes in many identified subpopulations, and required estimates of variance at all observed values in the sample. v_{HT} is readily adaptable to a recursive form, so there was particular interest in verifying its utility in the context of the stream survey.

Further background on the details of the National Stream Survey can be obtained from Kaufmann *et al* (1988) and Sale *et al* (1988).

3.0 Theoretical Results

Notation:

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

π_{ij}^o = approximate formula for π_{ij} described in detail below

v_{HT}^o = Horvitz-Thompson variance estimator calculated using π_{ij}^o

v_{YG}^o = Yates-Grundy variance estimator calculated using π_{ij}^o

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

v_{HT}^{hr} = Horvitz-Thompson variance estimator calculated using π_{ij}^{hr}

v_{YG}^{hr} = Yates-Grundy variance estimator calculated using π_{ij}^{hr}

\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):

$$\pi_{ij}^o = \frac{n(n-1)x_i x_j}{T_x \left(T_x - \frac{x_i + x_j}{2} \right)} \quad (3.1)$$

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

$$= \frac{2(n-1)\pi_i \pi_j}{2n - \pi_i - \pi_j} \quad (3.3)$$

Note that in (3.2) and (3.3) the population total, T_x , does not appear, so that this form is appropriate for the Stream Survey, where T_x is unknown. Since $\pi_{ij}^o = \pi_{ji}^o$, this approximation satisfies the important symmetry property of pairwise inclusion probabilities. Further, if $x_i=1$ for all $i=1, \dots, N$, then $\pi_{ij}^o = \frac{n(n-1)}{N(N-1)}$, 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. A truncated form is usually used as a simplification to derive theoretical results (see equation 5.20 of Hartley and Rao (1962), p. 289 of Wolter (1985), and Isaki and Pinciaro (1977) for examples):

$$\pi_{ij}^{hr} = \frac{n(n-1)x_i x_j}{T_x \left[T_x - x_i - x_j + \sum_{k=1}^N x_k^2 / T_x \right]} \quad (3.4)$$

$$= \frac{(n-1)\pi_i \pi_j}{\left[n - \pi_i - \pi_j + \sum_{k=1}^N \pi_k^2 / n \right]} \quad (3.5)$$

In the simulation studies described in Section 4.0, equation (5.15) of Hartley and Rao (1962) was used instead of the truncated form (3.4) above. Note the similarities between (3.4) and (3.1), and between (3.5) and (3.2).

A more detailed analysis of the approximation formulas π_{ij}^o and π_{ij}^{hr} and a description of some additional pairwise inclusion probability formulas are in preparation.

3.2 Properties of the Variance Estimators

We begin by reporting some results on positivity of the variance estimators using the approximate π_{ij} formulas.

Theorem 3.1. $v_{YG}^o \geq 0$.

Proof: Define $a_{ij} = (\pi_i \pi_j - \pi_{ij}) / \pi_{ij}$.

$$\text{Substituting } \pi_{ij}^o \text{ for } \pi_{ij}, a_{ij} \text{ becomes } a_{ij}^o = \frac{1}{(n-1)} \left[1 - \frac{(\pi_i + \pi_j)}{2} \right]. \quad (3.6)$$

$$\text{Since } \pi_i \leq 1 \text{ and } a_{ij}^o \geq 0, v_{YG}^o = \sum_{i=1}^{n-1} \sum_{j=i+1}^n a_{ij}^o \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 \geq 0.$$

Theorem 3.2. $v_{YG}^{hr} \geq 0$ when $(\pi_i + \pi_j) \leq 1 + E(\bar{\pi}_s)$ for all $i, j \in s$, where $\bar{\pi}_s = \sum_{i \in s} \pi_i / n$.

Proof: Substituting π_{ij}^{hr} for π_{ij} in a_{ij} , a_{ij} becomes

$$a_{ij}^{hr} = \frac{1}{(n-1)} \left[1 - (\pi_i + \pi_j) + \sum_{k=1}^N \pi_k^2 / n \right] = \frac{1}{(n-1)} \left[1 - (\pi_i + \pi_j) + E(\bar{\pi}_s) \right], \quad (3.7)$$

$$\text{where } E(\bar{\pi}_s) = \sum_{k=1}^N \pi_k^2 / n.$$

$a_{ij}^{hr} \geq 0$ when $\pi_i + \pi_j \leq 1 + E(\bar{\pi}_s)$, and the theorem follows.

Conditions for positivity of v_{HT}^o and v_{HT}^{hr} are not so transparent. Both can be negative, but based on the empirical results in Section 5, v_{HT}^o has a greatly lower incidence of negativity. The negativity issue is clarified further following Theorem 3.5.

Both v_{HT} and v_{YG} are unbiased when all population π_{ij} 's are non-zero. Bias is introduced into the variance estimators when an approximation to the pairwise inclusion probabilities is used. The following theorem is presented without proof:

Theorem 3.3. Let $\hat{\pi}_{ij}$ be an approximation to π_{ij} .

$$i) \quad \text{Bias}(\hat{v}_{\text{HT}}) = \sum_{i=1}^N \sum_{j \neq i}^N \left(\frac{\pi_{ij}}{\hat{\pi}_{ij}} - 1 \right) y_i y_j \quad (3.8)$$

$$ii) \quad \text{Bias}(\hat{v}_{\text{YG}}) = -\frac{1}{2} \sum_{i=1}^N \sum_{j \neq i}^N \left(\frac{\pi_{ij}}{\hat{\pi}_{ij}} - 1 \right) \pi_i \pi_j \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2. \quad (3.9)$$

Clearly this bias depends on how closely $\hat{\pi}_{ij}$ approximates π_{ij} , but useful general conclusions regarding bias are not readily available from these formulas.

In the absence of a feasible unbiased estimator, we directed our assessment of variance estimators toward MSE and good confidence interval coverage. We believe coverage is the key criterion, while sampling variability, as measured by MSE of the variance estimators, is important because v_{YG} has been claimed superior to v_{HT} on this criterion. In the following, i and j index the sampling units within a sample. After some algebra, v_{YG} can be rewritten as:

$$v_{\text{YG}} = \sum_{i=1}^n \left(\frac{y_i}{\pi_i} \right)^2 \sum_{j \neq i}^n \left(\frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \right) + \sum_{i=1}^n \sum_{j \neq i}^n \left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \right) \frac{y_i}{\pi_i} \frac{y_j}{\pi_j}. \quad (3.10)$$

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 $(1 - \pi_i)$ that appears 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 $(1 - \pi_i)$, with the expectation taken over the sample space conditioned on $i \in s$.

Theorem 3.4 $E \left[\sum_{j \neq i}^n \left(\frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \right) \mid i \in s \right] = (1 - \pi_i).$

Proof: By the Horvitz-Thompson Theorem,

$$\begin{aligned}
 E \left[\sum_{j \neq i}^n \left(\frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \right) \mid i \in s \right] &= \sum_{j \neq i}^N \left(\frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \right) \pi_{j \cdot i} \\
 &= \sum_{j \neq i}^N \left(\frac{\pi_i \pi_j}{\pi_{j \cdot i} \pi_i} - 1 \right) \pi_{j \cdot i} \\
 &= \sum_{j \neq i}^N \pi_j - \sum_{j \neq i}^N \pi_{j \cdot i} = n - \pi_i - (n-1) = (1 - \pi_i).
 \end{aligned}$$

Note: $\sum_{j \neq i}^N \pi_{j \cdot i} = \sum_{j \neq i}^N \pi_{ij} / \pi_i = (n-1)\pi_i / \pi_i = (n-1).$

Thus from Theorem 3.4, 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* $(1 - \pi_i)$. This induces a favorable “cancellation” in v_{YG} , under certain circumstances, as follows. Rewrite (3.10) as

$$v_{YG} = \sum_{i=1}^n \left(\frac{y_i}{\pi_i} \right) \sum_{j \neq i}^n \left(\frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \right) \left(\frac{y_i}{\pi_i} \right) - \sum_{i=1}^n \left(\frac{y_i}{\pi_i} \right) \sum_{j \neq i}^n \left(\frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \right) \left(\frac{y_j}{\pi_j} \right). \quad (3.11)$$

Then, when y_i/x_i (and hence y_i/π_i) is nearly constant for all i , the case where v_{YG} is known to work well, the terms in the two summations over j will nearly cancel each other, so 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_i/x_i increases.

That $v_{YG} \equiv 0$ when $y_i/x_i = \beta$ for all i (when y is exactly proportional to x) is further evidence that v_{YG} should have small variance when the ratios y_i/x_i are nearly constant. Use of the approximation π_{ij}^o in v_{HT} results in this same property.

Theorem 3.5 If $y_i/x_i = \beta$ for all $i=1,\dots,N$, $v_{HT}^o = 0$ for all samples in the sample space.

Proof: (see Result B3, Appendix B)

From Theorem 3.5, we expect that v_{HT}^o should perform more adequately in this situation previously clearly favorable to v_{YG} . Further, v_{HT}^o would appear highly favorable relative to v_{HT}^{hr} in this circumstance. When $y_i/x_i = \beta$, Cumberland and Royall (1981, p. 356) show:

$$v_{HT}^{hr} = \beta^2 \left[T_x \bar{x}_s - \sum_{k=1}^N x_k^2 \right], \quad (3.12)$$

where \bar{x}_s = sample mean of the x 's. Under this condition, v_{HT}^{hr} will behave poorly, frequently taking on negative estimates. Since $\sum_{k=1}^N x_k^2 = T_x E(\bar{x}_s)$, the probability of a negative estimate is given by:

$$\Pr\{v_{HT}^{hr} < 0\} = \Pr\{\bar{x}_s < E(\bar{x}_s)\}. \quad (3.13)$$

By this relation, it is seen that this problem of negative estimates exists regardless of population size or sample size.

These theoretical results show the importance of the ratios y/x in determining the properties of the variance estimators. The variability of these ratios in the population is represented by the coefficient of variation, $cv(y/x)$. The theoretical results indicate a possible strong association between $cv(y/x)$ and the behavior of the variance estimators. Thus $cv(y/x)$ is an important population descriptor in the simulation studies that follow.

4.0 Design of Simulation Studies

4.1 Group I Populations

We used two simulation studies to explore the properties of v_{HT} and v_{YG} . For the first set of simulations, designated Group I, we examined two NSS pilot study data sets (Stream1 and Stream2) and several populations from the statistical literature. Six of these populations were used by Cumberland and Royall (1981) to demonstrate the superiority of v_{YG} . The Group I populations were:

- 1) Sales x =gross sales of corporations, 1974
 y =gross sales of same corporations, 1975
- 2) Cancer x =adult white female population in 1960 for counties in NC, SC, and GA
 y =breast cancer mortality, 1950-1969 (white females)
- 3) Cities x =population of US cities with population between 100,000 and 1,000,000 in 1960
 y =population of same US cities in 1970
- 4) Counties60 x =number of households of certain counties in NC, SC, and GA
 y =population, excluding residents of group quarters, 1960
- 5) Counties70 x =number of households of certain counties in NC, SC, and GA
 y =population, excluding residents of group quarters, 1970
- 6) Hospitals x =number of beds
 y =number of patients discharged

(Note: these first 6 populations are described in more detail in Royall and Cumberland, 1981.)

- 7) Paddy x =geographical area
 y =area under winter paddy (from Murthy, 1967)
- 8) Stream1 and 9) Stream2
 x =direct watershed area of stream reach
 y =length of stream reach

Scatter plots of these populations are shown in Figure I. In Figure I, the units of x and y are scaled by their respective population standard deviations, and the dashed reference line is the diagonal through the origin. The data for populations Paddy, Stream1, and Stream2 are available from the authors.

4.2 Group II Populations

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 identifiable 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 base population of $N=72$ observations was

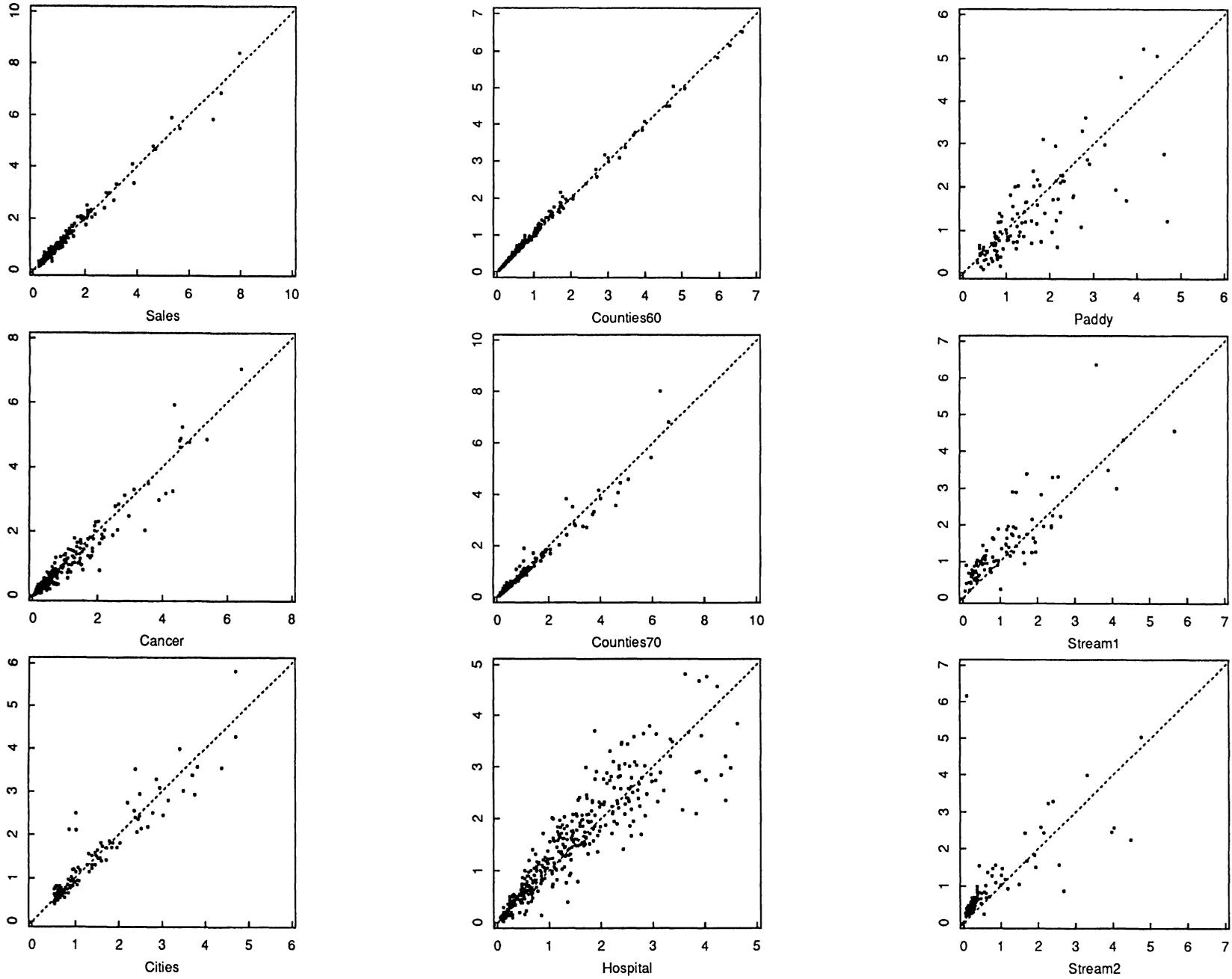


Figure I. Group I Populations (standardized units)

purposefully selected from a data set obtained from the Pilot Study of the National Stream Survey (Messer *et al*, 1986). The population variables were w =watershed area, and y =reach length. A standardized auxiliary variable, x , was derived from the original auxiliary variable, w , via the transformation $x = w\sqrt{V_y/V_w}$, where V_w and V_y are the population variances of w and y respectively. This re-scaling of the auxiliary variable does not affect the sampling scheme since multiplicative shifts in the auxiliary variable do not change the inclusion probabilities. A second standardization was achieved by representing both variables in standard units of y , so that the representation is invariant to the measurement scale of the y 's as well as the x 's.

New populations were created at various locations in the population space by adding or subtracting scalars to x and/or y of the base population. The new populations all have the same correlation between x and y , and the population ellipses have a major axis with slope 1. However, these shifts change $cv(y/x)$, and additive shifts in x change the inclusion probabilities.

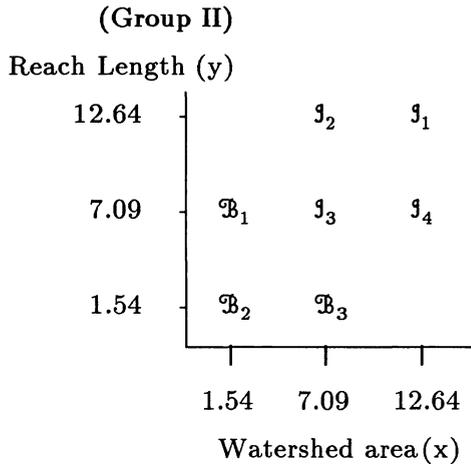
In order to assess the effect of different correlations, two other base populations, with $\rho(x,y)=0.53$ and $\rho(x,y)=0.99$, were created from the original base population. The new base populations were also moved through the population space via additive shifts in x and/or y . In the analysis reported here, we examined 7 populations at each of the three correlations.

Based on the location of the population centroids, the Group II populations were classified as \mathfrak{B} =boundary population or \mathfrak{I} =interior population. The boundary populations have high $cv(y/x)$, while the interior populations have low $cv(y/x)$ (Table 1). For a given location in the population space, $cv(y/x)$ decreases as $\rho(x,y)$ increases. The notation used to identify populations is that subscripts indicate the location of the population centroid within \mathfrak{B} or \mathfrak{I} , while superscripts denote the correlation between x and y : $lo=.53$, $m=.82$, $hi=.99$. Figure II. shows the location of the population centroids for the Group II populations. For comparison, the population centroids for the Group I populations are also plotted (the numbers represent the order of the populations in Group I as listed in Sec. 4.1).

The sampling design used in the simulations was random-order, *ups* sampling.

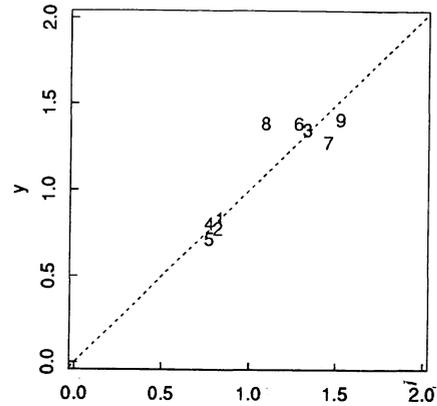
Detailed descriptions of this sampling scheme appear in Hartley and Rao (1962) and Cumberland and Royall (1981). All populations were sufficiently large that exact π_{ij} 's were not computationally feasible, so comparisons were made among v_{HT}^o , v_{HT}^{hr} , v_{YG}^o , and v_{YG}^{hr} . Version 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 computers.

Figure II. Population Space Centroids



(Units of x and y are standardized.)

Group I Population Centroids



4.3 Output of the Simulation Studies

The criteria for comparing the variance estimators are:

- 1) estimated MSE
- 2) confidence interval coverage in percent, with intervals calculated as $\hat{T}_y \pm 1.96\sqrt{\hat{v}}$
- 3) relative bias, estimated by

$$\text{rel bias} = \frac{\hat{E}(\hat{v}) - \hat{V}(\hat{T}_y)}{\hat{V}(\hat{T}_y)},$$

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} .

5.0 Results

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 , as measured by $cv(y/x)$, is small; as $cv(y/x)$ increases, no advantage is expected for v_{YG} . Further, when $cv(y/x)$ is low, v_{HT}^o should have much smaller MSE and fewer negative estimates, compared to v_{HT}^{hr} . These predictions were confirmed by the Group I simulations (Table 2).

The properties of v_{YG}^o and v_{YG}^{hr} were very similar in the Group I populations. Confidence interval coverage was identical, but v_{YG}^o uniformly outperformed v_{YG}^{hr} in terms of MSE. v_{YG}^{hr} was clearly superior to v_{HT}^{hr} only in populations Sales, Counties60, and Counties70, the populations with smallest $cv(y/x)$. The remaining populations provided examples in which v_{HT}^{hr} and v_{YG}^{hr} had very similar properties, although v_{YG}^{hr} tended to show slightly better coverage and MSE than v_{HT}^{hr} .

v_{HT}^o had much better properties than v_{HT}^{hr} in populations Sales, Counties60, and Counties70. MSE and confidence interval coverage of v_{HT}^o were dramatically better than those of v_{HT}^{hr} . Again, note that these are the three populations with smallest $cv(y/x)$. In the remaining populations, v_{HT}^{hr} had slightly smaller MSE while v_{HT}^o had slightly better coverage. Finally, comparing v_{HT}^o and v_{YG}^o , v_{YG}^o had uniformly better MSE but slightly poorer coverage than v_{HT}^o .

Generalizations from the Group I simulations are:

- a) The Horvitz-Thompson variance formula has behavior comparable to the Yates-Grundy formula in all populations except those having very small $cv(y/x)$.
- b) v_{HT}^{hr} has very poor properties when $cv(y/x)$ is small.
- c) The best estimator in terms of MSE is v_{YG}^o .
- d) The best estimator in terms of confidence interval coverage is v_{HT}^o .

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_{YG}^{hr} was far superior to v_{HT}^{hr} in the interior populations, but v_{HT}^{hr} was slightly better in the boundary populations. v_{YG}^o had smaller MSE than v_{HT}^o 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_{HT}^o was much smaller than the MSE of v_{HT}^{hr} in the interior populations, while v_{HT}^{hr} was slightly better than v_{HT}^o in the boundary populations. v_{YG}^o and v_{YG}^{hr} were virtually identical in the interior populations, but v_{YG}^o had slightly smaller MSE than v_{YG}^{hr} 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 .

Patterns in MSE were also associated with sample size (Appendix Tables A1, A2).

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

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 (Appendix Table A3). 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.

6.0 Conclusions

Our results show that the superiority of v_{YG} over v_{HT} previously reported in the statistical literature is restricted to an identifiable set of populations, and does not hold over the entire population space. 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 set of populations studied, and by identifying an association between $cv(y/x)$ and superiority of v_{YG}^{hr} to v_{HT}^{hr} . When $cv(y/x)$ is small, a condition in which π_{px} sampling is most efficient, v_{YG}^{hr} is superior to v_{HT}^{hr} . 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 comparing the results from the Group I and II studies, all the Group I populations fall into the corner of the population space near the origin, corresponding to populations \mathfrak{B}_2^m or \mathfrak{B}_2^{hi} . Thus the Group II populations span a far greater range of populations than

represented by the Group I populations. A feature of the population space approach is that it should be useful for predicting the properties of the variance estimators in real world populations. Populations Sales, Cancer, Counties60, and Counties70 should be similar to \mathfrak{B}_2^{hi} . Population Cities appears to be intermediate between \mathfrak{B}_2^m and \mathfrak{B}_2^{hi} , while the remaining populations should be similar to \mathfrak{B}_2^m . The behavior of the variance estimators in the Group I populations generally follows the results predicted by analysis of \mathfrak{B}_2^m and \mathfrak{B}_2^{hi} .

Populations Counties60 and Counties70 present an interesting comparison. Both populations have the same x 's, and the descriptive statistics and scatter plots (Figure I) are very similar. Yet the properties of v_{HT}^o and v_{HT}^{hr} are very different in the two populations. Counties70 behaves similar to population \mathfrak{J}_1^{hi} , while Counties60 is more similar in behavior to \mathfrak{J}_3^{hi} or \mathfrak{B}_2^{hi} . Both populations are located close to the origin in the population space near the diagonal through the origin. Stehman and Overton (1987b) present simulation results showing the properties of the variance estimators, particularly v_{HT}^o and v_{HT}^{hr} , change rapidly over this region of the population space for high correlation populations. The comparison between Counties60 and Counties70 is qualitatively the same as comparing \mathfrak{B}_1^{hi} and \mathfrak{B}_2^{hi} , where \mathfrak{B}_1^{hi} indicates the results for Counties60, and \mathfrak{B}_2^{hi} indicates the results for Counties70. Populations Cancer and Sales present a similar situation. Both are located near the diagonal through the origin, but Sales is slightly above, and Cancer is slightly below. The properties of v_{HT}^o and v_{HT}^{hr} are much different in population Sales as compared to population Cancer. These results illustrate the need for detailed exploration of the population space in the region near the origin.

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, nor was the variance estimator v_{YG} appropriate. MSE and confidence interval coverage of v_{HT}^o was comparable to or better than the other variance estimators studied in populations of the nature of the NSS populations, thus providing additional justification for the use of v_{HT}^o in the NSS. Extensive results on the use of v_{HT}^o in the NSS are reported in Overton and Stehman (1987) and Stehman and Overton (1987a).

The success of the Group II simulations in clarifying the pattern of behavior of the various estimators of variance has led us to a more general analysis of the population space. An expanded simulation study, to include populations from known probability distributions, is currently in progress and the results will be reported in Stehman and Overton (1987b).

Acknowledgements

We thank William Cumberland and Richard Royall for providing the data for the 6 populations from their 1981 paper.

TABLE 1. Description of Populations

Population	N	cv(X)	cv(Y)	$\rho(X,Y)$	cv(Y/X)
<u>Group I</u>					
Sales	327	1.20	1.19	.99	.14
Cancer	301	1.22	1.28	.97	.32
Cities	125	.75	.74	.95	.25
Counties 60	304	1.30	1.24	.99	.07
Counties 70	304	1.30	1.38	.98	.16
Hospitals	393	.78	.72	.91	.29
Paddy	108	.69	.78	.79	.39
Stream1	100	.92	.72	.86	.71
Stream2	100	.66	.71	.81	.41
<u>Group II</u>					
<u>High Correlation (0.99)</u>					
\mathfrak{B}_1	72	.54	.12	.99	.49
\mathfrak{B}_2	72	.54	.54	.99	.12
\mathfrak{B}_3	72	.12	.54	.99	.44
\mathfrak{J}_1	72	.07	.07	.99	.01
\mathfrak{J}_2	72	.12	.07	.99	.05
\mathfrak{J}_3	72	.12	.12	.99	.02
\mathfrak{J}_4	72	.07	.12	.99	.05
<u>Medium Correlation (0.82)</u>					
\mathfrak{B}_1	72	.65	.14	.82	.80
\mathfrak{B}_2	72	.65	.65	.82	.59
\mathfrak{B}_3	72	.14	.65	.82	.56
\mathfrak{J}_1	72	.08	.08	.82	.05
\mathfrak{J}_2	72	.14	.08	.82	.08
\mathfrak{J}_3	72	.14	.14	.82	.08
\mathfrak{J}_4	72	.08	.14	.82	.09
<u>Low Correlation (0.53)</u>					
\mathfrak{B}_1	72	.65	.14	.53	.88
\mathfrak{B}_2	72	.65	.65	.53	1.11
\mathfrak{B}_3	72	.14	.65	.53	.61
\mathfrak{J}_1	72	.08	.08	.53	.07
\mathfrak{J}_2	72	.14	.08	.53	.11
\mathfrak{J}_3	72	.14	.14	.53	.13
\mathfrak{J}_4	72	.08	.14	.53	.12

TABLE 2a. Results of Group I Simulations

Ratios of Mean Square Errors (n=16)

Population	<i>a</i>	<i>b</i>	<i>c</i>	<i>d</i>
Sales	13.28	0.09	0.95	1.29
Cancer	1.37	0.83	0.90	1.27
Cities	1.16	1.03	0.88	1.35
Counties 60	181.01	0.04	0.93	8.09
Counties 70	3.57	0.35	0.93	1.34
Hospitals	1.02	1.06	0.97	1.12
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 $MSE(v_{HT}^{hr}) / MSE(v_{YG}^{hr})$

b $MSE(v_{HT}^o) / MSE(v_{HT}^{hr})$

c $MSE(v_{YG}^o) / MSE(v_{YG}^{hr})$

d $MSE(v_{HT}^o) / MSE(v_{YG}^o)$

Table 2b. Results of Group I Simulations

Confidence Interval Coverage (nominal 95%)

Population	v_{HT}^{hr}	v_{YG}^{hr}	v_{HT}^o	v_{YG}^o
Sales	63	94	95	94
Cancer	90	93	93	93
Cities	75	87	88	87
Counties 60	56	92	98	92
Counties 70	68	88	78	88
Hospitals	94	94	94	94
Paddy	92	93	94	93
Stream1	87	88	89	88
Stream2	87	87	89	87

TABLE 3. Results of Group II Simulations
Ratios of Mean Square Errors (n=16)

Ratio of $MSE(v_{HT}^{hr})$ to $MSE(v_{YG}^{hr})$				Ratio of $MSE(v_{HT}^o)$ to $MSE(v_{YG}^o)$			
	ρ				ρ		
	<u>.53</u>	<u>.82</u>	<u>.99</u>		<u>.53</u>	<u>.82</u>	<u>.99</u>
\mathfrak{B}_1	.96	0.96	0.86	1.53	1.57	1.87	
\mathfrak{B}_2	.99	0.86	56.05	1.46	1.43	1.01	
\mathfrak{B}_3	.98	0.99	1.09	1.00	.97	.93	
\mathfrak{J}_1	9.46	85.55	16,281.00	1.43	1.79	3.25	
\mathfrak{J}_2	8.53	38.58	208.44	2.93	5.99	13.31	
\mathfrak{J}_3	3.62	31.10	5,538.02	1.41	1.65	3.33	
\mathfrak{J}_4	1.77	6.92	44.15	1.00	1.01	2.29	

Ratio of $MSE(v_{HT}^o)$ to $MSE(v_{HT}^{hr})$				Ratio of $MSE(v_{YG}^o)$ to $MSE(v_{YG}^{hr})$			
	ρ				ρ		
	<u>.53</u>	<u>.82</u>	<u>.99</u>		<u>.53</u>	<u>.82</u>	<u>.99</u>
\mathfrak{B}_1	1.14	0.97	1.80	0.72	0.75	0.82	
\mathfrak{B}_2	1.00	1.38	0.02	0.68	0.83	0.76	
\mathfrak{B}_3	0.99	1.23	0.85	0.98	0.99	1.00	
\mathfrak{J}_1	0.16	0.02	0.0002	1.05	1.02	1.02	
\mathfrak{J}_2	0.33	0.17	0.65	0.95	1.08	1.03	
\mathfrak{J}_3	0.40	0.05	0.0006	1.03	1.01	1.00	
\mathfrak{J}_4	0.55	0.17	0.05	0.99	0.98	1.01	

TABLE 4. Results of Group II Simulations
Confidence Interval Coverage (%)

Population	$\rho=.53$		$\rho=.82$		$\rho=.99$	
	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
J_1	76	93	62	94	49	93
J_2	84	94	75	94	63	93
J_3	86	93	69	93	52	93
J_4	88	93	82	93	70	93

Population	$\rho=.53$		$\rho=.82$		$\rho=.99$	
	V_{HT}°	V_{YG}°	V_{HT}°	V_{YG}°	V_{HT}°	V_{YG}°
\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
J_1	95	93	95	94	93	93
J_2	96	93	97	94	98	93
J_3	95	93	95	93	92	93
J_4	93	93	91	93	88	93

References

- Biyani, S. H. (1980). On inadmissibility of the Yates-Grundy variance estimator in unequal probability sampling. *J. Amer. Statist. Soc.* 75, 709-712.
- Brewer, K. R. W. and Hanif, M. (1983). *Sampling with Unequal Probabilities*. Springer-Verlag: New York.
- Cassel, C.-M., Sarndal, C.-E., and Wretman, J. H. (1977). *Foundations of Inference in Survey Sampling*. Wiley: New York.
- Cochran, W. G. (1977). *Sampling Methods* (3rd Edition). Wiley: New York.
- Cumberland, W. G. and Royall, R. M. (1981). Prediction models and unequal probability sampling. *J. Roy. Statist. Soc. Ser. B* 43, 353-367.
- Godambe, V. P. and Joshi, V. M. (1965). Admissibility and Bayes estimation in sampling finite populations. I. *Ann. Math. Statist.* 36, 1707-1722.
- 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, D. J. (1952). A generalization of sampling without replacement from a finite universe. *J. Amer. Statist. Assoc.* 47, 663-685.
- Isaki, C. T. and Pinciario, S. J. (1977). Numerical comparison of some estimators of variance under PPS systematic sampling. *Proceedings of the Survey Research Section, American Statistical Association*, pp. 308-313.
- Kaufmann, P. R., A. T. Herlihy, J. W. Elwood, M. E. Mitch, W. S. Overton, M. J. Sale, J. J. Messer, K. A. Cougan, D. V. Peck, K. H. Reckhow, A. J. Kinney, S. J. Christie, D. D. Brown, C. A. Hagley, and H. I. Jager. (1988). *Chemical Characteristics of Streams in the Mid-Atlantic and Southeastern United States. Volume I: Population Descriptions and Physico-Chemical Relationships*. EPA/600/3-88/021a. United States Environmental Protection Agency, Washington, D. C.
- Messer, J. J., C. W. Ariss, J. R. Baker, S. K. Droué, 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.
- Milne, A. (1959). The centric systematic area-sample treated as a random sample. *Biometrics* 15, 270-297.
- Murthy, M. N. (1967). *Sampling Theory and Methods*. Statistical Publishing Society: Calcutta.
- Osborne, J. G. (1942). Sampling errors of systematic and random surveys of cover-type areas. *J. Amer. Statist. Assoc.* 37, 256-264.
- 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.

Overton, W. S. and Stehman, S. V. (1987). *An Empirical Investigation of Sampling and Other Errors in the National Stream Survey; Analysis of a Replicated Sample of Streams*. Technical Report 119, Department of Statistics, Oregon State University, Corvallis, Oregon, 97331.

Payandeh, B. (1970). Relative efficiency of two-dimensional systematic sampling. *Forest Science* 16, 271-276.

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

Royall, R. M. and Cumberland, W. G. (1981). An empirical study of the ratio estimator and estimators of its variance. *J. Amer. Statist. Assoc.* 76, 66-77.

Sale, M. J., P. R. Kaufmann, H. I. Jager, J. M. Coe, K. A. Cougan, A. J. Kinney, M. E. Mitch, and W. S. Overton. (1988). *Chemical Characteristics of Streams in the Mid-Atlantic and Southeastern United States. Volume II: Streams Sampled, Descriptive Statistics, and Compendium of Physical and Chemical Data*. EPA/600/3-88/021b. U. S. Environmental Protection Agency, Washington, D. C.

Sankaranarayanan, K. (1980). A note on the admissibility of some non-negative quadratic estimators. *J. Roy. Statist. Soc. Ser. B* 42, 387-389.

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. (1987a). *An Empirical Investigation of the Variance Estimation Methodology Prescribed for the National Stream Survey: Simulated Sampling from Stream Data Sets*. Technical Report 118, Department of Statistics, Oregon State University, Corvallis, Oregon, 97331.

Stehman, S. V. and Overton, W. S. (1987b). *An Empirical, General Population Assessment of the Properties of Variance Estimators of the Horvitz-Thompson Estimator Under Variable Probability Systematic Sampling*. Biometrics Unit Manuscript Bu-M 936, Cornell University, 337 Warren Hall, Ithaca, New York, 14853.

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

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.

APPENDIX A: Properties of Variance Estimators from Simulations using Different Sample Sizes

TABLE A1. Ratios of Mean Square Errors

Part II: Comparison of Same Variance Estimators with Different π_{ij} Approximation Formulas

	<u>MSE(v_{HT}^o) / MSE(v_{HT}^{hr})</u>				<u>MSE(v_{YG}^o) / MSE(v_{YG}^{hr})</u>			
	<u>Sample Size</u>				<u>Sample Size</u>			
<u>Population</u>	<u>4</u>	<u>8</u>	<u>16</u>	<u>24</u>	<u>4</u>	<u>8</u>	<u>16</u>	<u>24</u>
Sales	-	.26	.09	.06	-	.95	.95	.85
Paddy	1.02	1.02	1.02	.97	.99	.98	.89	.80
Stream1	1.02	1.04	1.12	-	.93	.87	.74	-
Stream2	1.00	1.07	1.21	-	.95	.96	.93	-
<u>Medium Correlation (0.82)</u>								
\mathfrak{B}_1	1.04	1.10	1.23	1.42	.94	.87	.75	.68
\mathfrak{B}_2	1.03	1.18	1.38	1.50	.90	.91	.83	.70
\mathfrak{B}_3	1.00	1.02	.97	1.28	1.00	1.05	.99	1.00
\mathfrak{J}_1	.26	.07	.02	.01	.97	1.01	1.02	1.06
\mathfrak{J}_2	.48	.21	.17	.15	.96	1.01	1.08	.99
\mathfrak{J}_3	.49	.18	.05	.03	.97	1.01	1.01	1.00
\mathfrak{J}_4	.81	.43	.14	.08	1.04	.97	.98	1.00
<u>Low Correlation (0.53)</u>								
\mathfrak{B}_1	.89	1.03	1.15	1.27	.81	.83	.72	.64
\mathfrak{B}_2	.68	1.40	1.00	1.29	.68	1.13	.68	.68
\mathfrak{B}_3	1.01	.96	.99	1.00	1.01	.96	.98	.97
\mathfrak{J}_1	.74	.43	.16	.07	.99	1.01	1.05	.98
\mathfrak{J}_2	.79	.51	.33	.28	1.05	1.08	.95	.92
\mathfrak{J}_3	.89	.72	.40	.26	.97	.99	1.03	1.00
\mathfrak{J}_4	.95	.89	.55	.26	.98	1.03	.99	1.05

TABLE A1. Ratios of Mean Square Errors

Part I: Comparison of Different Variance Estimators with Same π_{ij} Approximation

Population	<u>MSE(v_{HT}^{hr}) / MSE(v_{YG}^{hr})</u>				<u>MSE(v_{HT}^o) / MSE(v_{YG}^o)</u>			
	<u>Sample Size</u>				<u>Sample Size</u>			
	4	8	16	24	4	8	16	24
Sales	-	3.89	13.28	23.48	-	1.07	1.29	1.68
Paddy	1.04	1.10	1.28	1.58	1.07	1.14	1.46	1.91
Stream1	1.00	1.03	.99	-	1.10	1.23	1.50	-
Stream2	.99	.98	.97	-	1.04	1.09	1.26	-
<u>Medium Correlation (0.82)</u>								
\mathfrak{B}_1	1.00	.98	.96	.93	1.11	1.23	1.57	1.94
\mathfrak{B}_2	.95	.92	.86	.86	1.09	1.20	1.43	1.83
\mathfrak{B}_3	.99	.99	.99	1.00	.98	.96	.97	1.28
\mathfrak{J}_1	3.96	17.20	85.55	308.33	1.06	1.16	1.79	3.05
\mathfrak{J}_2	2.38	7.60	38.58	98.30	1.19	1.56	5.99	15.04
\mathfrak{J}_3	2.09	6.36	31.10	106.97	1.06	1.16	1.65	3.74
\mathfrak{J}_4	1.22	2.09	6.92	22.18	.96	.91	1.01	1.76
<u>Low Correlation (0.53)</u>								
\mathfrak{B}_1	1.00	.98	.96	.96	1.10	1.22	1.53	1.91
\mathfrak{B}_2	.99	.97	.99	.97	1.00	1.20	1.46	1.82
\mathfrak{B}_3	1.00	.99	.99	.98	.99	.99	1.00	.99
\mathfrak{J}_1	1.39	2.60	9.46	29.14	1.04	1.10	1.43	2.18
\mathfrak{J}_2	1.49	2.72	8.53	23.88	1.11	1.30	2.93	7.44
\mathfrak{J}_3	1.14	1.50	3.62	9.17	1.04	1.09	1.41	2.37
\mathfrak{J}_4	1.01	1.07	1.33	2.00	1.00	1.00	1.00	1.00

TABLE A2. Confidence Interval Coverage (%)

(Nominal coverage is 95%)

Population	n = 8				n = 16				n = 24			
	v_{HT}^{hr}	v_{HT}^o	v_{YG}^{hr}	v_{YG}^o	v_{HT}^{hr}	v_{HT}^o	v_{YG}^{hr}	v_{YG}^o	v_{HT}^{hr}	v_{HT}^o	v_{YG}^{hr}	v_{YG}^o
Sales	66	91	91	90	64	94	93	93	63	95	94	94
Paddy	89	90	91	91	92	94	93	93	92	95	94	94
Stream1	87	89	88	88	90	93	90	90	-	-	-	-
Stream2	87	89	87	87	90	92	89	89	-	-	-	-
<u>Medium Correlation (0.82)</u>												
\mathcal{B}_1	81	82	80	80	87	89	85	84	88	90	86	84
\mathcal{B}_2	91	91	90	90	92	93	93	92	92	94	93	92
\mathcal{B}_3	91	91	91	91	93	92	93	93	94	92	94	94
\mathcal{J}_1	66	92	91	91	62	95	94	94	57	96	94	93
\mathcal{J}_2	79	93	91	91	75	97	94	94	69	98	93	93
\mathcal{J}_3	75	91	91	91	69	95	93	93	64	96	94	94
\mathcal{J}_4	84	91	90	91	82	91	93	93	78	91	94	94
<u>Low Correlation (0.53)</u>												
\mathcal{B}_1	80	82	79	80	87	88	85	84	82	91	85	85
\mathcal{B}_2	87	87	86	86	90	91	90	89	91	92	90	89
\mathcal{B}_3	92	91	92	91	93	93	93	93	93	94	93	94
\mathcal{J}_1	80	92	90	91	76	95	93	93	70	96	93	93
\mathcal{J}_2	86	92	91	90	84	96	94	93	80	98	93	94
\mathcal{J}_3	87	92	91	91	86	95	93	93	82	96	93	93
\mathcal{J}_4	88	90	91	90	88	93	93	93	87	93	94	93

TABLE A3. Proportion of Samples with Negative v_{HT}^{hr} *

Population	4	8	16	24
Sales	-	.28	.32	.33
Paddy	.02	.00	.00	.00
Stream1	.03	.01	.01	-
Stream2	.02	.01	.01	-
<u>Medium Correlation (0.82)</u>				
\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
<u>Low Correlation (0.53)</u>				
\mathfrak{B}_1	.01	.00	.00	.00
\mathfrak{B}_2	.00	.00	.00	.00
\mathfrak{B}_3	.00	.00	.00	.00
\mathfrak{J}_1	.12	.12	.17	.24
\mathfrak{J}_2	.12	.07	.10	.15
\mathfrak{J}_3	.06	.04	.06	.10
\mathfrak{J}_4	.04	.02	.02	.04

* Proportion of negative v_{HT}^o was less than .005 for all populations. v_{YG}^o and v_{YG}^{hr} were always positive.

APPENDIX B: Notes on v_{HT}°

Result B1. An alternative formula for v_{HT}° is the following:

$$v_{HT}^{\circ} = T_x \left[\frac{T_x}{n} s_r^2 - s_{ry} \right], \quad (B1.1)$$

where $r_i = y_i/x_i$,

$$s_r^2 = \sum_{i=1}^n (r_i - \bar{r})^2 / (n-1) \quad (\text{sample variance of } r),$$

and $s_{ry} = \sum_{i=1}^n (r_i - \bar{r}) y_i / (n-1) \quad (\text{sample covariance of } r \text{ and } y).$

Proof: In the following, all summations are over the elements in the sample.

$$s_r^2 = \sum_{i=1}^n (r_i - \bar{r})^2 / (n-1) = \sum_{i=1}^n r_i^2 / n - \sum_{i=1}^n \sum_{j \neq i}^n r_i r_j / n(n-1) \quad (B1.2)$$

$$\begin{aligned} v_{HT}^{\circ} &= \sum_{i=1}^n (y_i / \pi_i)^2 (1 - \pi_i) - \sum_{i=1}^n \sum_{j \neq i}^n [(\pi_i \pi_j - \pi_{ij}^{\circ}) / \pi_{ij}^{\circ}] (y_i / \pi_i) (y_j / \pi_j) \\ &= \left(\frac{T_x}{n} \right)^2 \sum_{i=1}^n r_i^2 \left(1 - \frac{nx_i}{T_x} \right) - \left(\frac{T_x}{n} \right)^2 \sum_{i=1}^n \sum_{j \neq i}^n a_{ij}^{\circ} r_i r_j \end{aligned}$$

(substituting $\pi_i = nx_i / T_x$ and a_{ij}° from (3.6))

$$\begin{aligned} &= \left(\frac{T_x}{n} \right)^2 \left[\sum_{i=1}^n r_i^2 - \frac{n}{T_x} \sum_{i=1}^n r_i^2 x_i \right] - \left(\frac{T_x}{n} \right)^2 \sum_{i=1}^n \sum_{j \neq i}^n \frac{1}{(n-1)} \left[1 - \frac{1}{2} \left(\frac{nx_i}{T_x} + \frac{nx_j}{T_x} \right) \right] r_i r_j \\ &= \left(\frac{T_x}{n} \right)^2 \left[\sum_{i=1}^n r_i^2 - \frac{n}{T_x} \sum_{i=1}^n r_i y_i \right] - \left(\frac{T_x}{n} \right)^2 \frac{1}{(n-1)} \sum_{i=1}^n \sum_{j \neq i}^n \left[r_i r_j - \frac{1}{2} \frac{n}{T_x} (x_i + x_j) r_i r_j \right] \\ &= \left(\frac{T_x}{n} \right)^2 \left[\sum_{i=1}^n r_i^2 - \frac{1}{(n-1)} \sum_{i=1}^n \sum_{j \neq i}^n r_i r_j - \frac{n}{T_x} \left\{ \sum_{i=1}^n r_i y_i - \frac{1}{2(n-1)} \sum_{i=1}^n \sum_{j \neq i}^n (x_i + x_j) r_i r_j \right\} \right] \\ &= \left(\frac{T_x}{n} \right)^2 \left[\sum_{i=1}^n r_i^2 - \frac{1}{(n-1)} \sum_{i=1}^n \sum_{j \neq i}^n r_i r_j \right] - \\ &\quad \frac{T_x}{n} \left[\sum_{i=1}^n r_i y_i - \frac{1}{2(n-1)} \sum_{i=1}^n \sum_{j \neq i}^n x_i r_i r_j - \frac{1}{2(n-1)} \sum_{i=1}^n \sum_{j \neq i}^n x_j r_i r_j \right] \quad (B1.3) \end{aligned}$$

Note:

1) Using equation (B1.2), the term in the first set of brackets in (B1.3) is ns_r^2 .

$$\begin{aligned} 2) \sum_{i=1}^n \sum_{j \neq i}^n x_i r_i r_j &= \sum_{i=1}^n \sum_{j \neq i}^n y_i r_j = \sum_{i=1}^n y_i \left(\sum_{j=1}^n r_j - r_i \right) = \sum_{i=1}^n y_i (n\bar{r} - r_i) \\ &= n\bar{r} \sum_{i=1}^n y_i - \sum_{i=1}^n r_i y_i = n^2 \bar{r} \bar{y} - \sum_{i=1}^n r_i y_i \end{aligned}$$

$$3) \sum_{i=1}^n \sum_{j \neq i}^n x_j r_i r_j = \sum_{i=1}^n r_i \sum_{j \neq i}^n y_j = \sum_{i=1}^n r_i (n\bar{y} - y_i) = n^2 \bar{r} \bar{y} - \sum_{i=1}^n r_i y_i.$$

Using 1) through 3), equation (B1.3) becomes

$$\begin{aligned} &\left(\frac{T_x}{n}\right)^2 ns_r^2 - \frac{T_x}{n} \left[\sum_{i=1}^n r_i y_i - \frac{1}{(n-1)} n^2 \bar{r} \bar{y} + \frac{1}{(n-1)} \sum_{i=1}^n r_i y_i \right] \\ &= \frac{T_x^2}{n} s_r^2 - \frac{T_x}{n} \left[\left(n \sum_{i=1}^n r_i y_i - n^2 \bar{r} \bar{y} \right) / (n-1) \right] \\ &= \frac{T_x^2}{n} s_r^2 - T_x \left[\frac{1}{(n-1)} \left(\sum_{i=1}^n r_i y_i - n \bar{r} \bar{y} \right) \right] \\ &= \frac{T_x^2}{n} s_r^2 - T_x s_{ry} = T_x \left[\frac{T_x}{n} s_r^2 - s_{ry} \right] \end{aligned} \tag{B1.4}$$

Result B2. $v_{HT}^0 > 0$ if $\frac{T_x}{n} s_r^2 - s_{ry} > 0$ or if $\frac{s_r^2}{s_{ry}} > \frac{n}{T_x}$.

Proof: Obvious from (B1.4).

Result B3. If $r_i = y_i/x_i = \beta$ for $i=1, \dots, N$, then $v_{HT}^0 \equiv 0$.

Proof: If $r_i = \beta$, $s_r^2 = s_{ry} = 0$ for any sample, and the result follows.