Pairwise Inclusion Probability Formulas in Random-Order, Variable Probability, Systematic Sampling

Stephen V. Stehman
Biometrics Unit, Cornell University, Ithaca, NY 14850 USA

W. Scott Overton
Department of Statistics, Oregon State University, Corvallis, Oregon 97331 USA

BU-1008-M

Revised August 1989

Pairwise Inclusion Probability Formulas in Random-Order, Variable Probability, Systematic Sampling

Stephen V. Stehman and W. Scott Overton
Biometrics Unit, Cornell University, and
Department of Statistics, Oregon State University

Technical Report 131

August 1989

Abstract

Variable probability, systematic sampling combined with the Horvitz-Thompson estimator is a commonly used sampling strategy. Estimating the variance of the Horvitz-Thompson estimator requires calculating the pairwise inclusion probabilities for the variable probability design. For random-order, variable probability systematic sampling, these probabilities may be computationally burdensome, and require knowledge of the auxiliary variable, x, for all elements in the universe. An easy to compute approximation requiring only the sample values of x is described and compared to the exact pairwise inclusion probabilities and to a well-known approximation derived by Hartley and Rao (1962). The analytical and empirical properties of the new approximation compare favorably to the properties of exact pairwise inclusion probabilities and provide some advantages over the Hartley-Rao approximation.

DEPARTMENT OF STATISTICS
Oregon State University
Corvallis, Oregon

1. INTRODUCTION

Variable probability sampling designs permit each of the N elements in the finite universe to have a different probability of being selected in the sample of n elements. The probability that element u will be selected in the sample, or the inclusion probability of element u, is given by $\pi_u = \sum_{\{s:u \epsilon s\}} p(s)$, where p(s) is the probability of selecting sample s, and the summation is over all samples in the sample space that contain element u. The pairwise inclusion probability of elements i and j $(i \neq j)$ is given by $\pi_{ij} = \sum_{\{s:(i,j) \epsilon s\}} p(s)$, where the summation is now over all samples containing elements i and j.

Variable probability sampling designs are most effective for precise estimation when the inclusion probabilities, π_u's, are proportional to the x_u's, where x_u is the value of an auxiliary variable for element u. Of the many variable probability designs (cf. Brewer and Hanif, 1983; Chaudhuri and Vos, 1988), random-order, variable probability systematic (random-order *vps*) sampling is one of the most practical designs guaranteeing π's proportional to the x's (Sunter, 1986). The Horvitz-Thompson estimator combined with random-order, *vps* sampling is a flexible, general sampling strategy.

In practice, a complication with this strategy is calculating the pairwise inclusion probabilities. Two commonly used estimators of the variance of the Horvitz-Thompson estimator, one due to Horvitz and Thompson (1952),

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

the other due to Yates and Grundy (1953) and Sen (1953),

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

require the pairwise inclusion probabilities. For random-order, *vps* sampling, Hidiriglou and Gray (1980) have written a FORTRAN program based on an algorithm provided by Connor (1966) for computing the π_{ij}'s. For large N, calculating the true π_{ij}'s can be computationally burdensome, and in some surveys, sufficient information is unavailable to compute the exact π_{ij}'s. For example, all population x's are needed to compute the exact π_{ij}'s, but in the National Stream Survey (Overton, 1985), only the sample x's were available. Approximation of the pairwise inclusion probabilities is often necessary. An early approximation formula for the π_{ij}'s was derived by Hartley and Rao (1962). This formula has been used in practice (cf. Wolter, 1985, and Choudry *et al*, 1985), and Cumberland and Royall (1981) derived some model-based, theoretical properties of variance estimators computed with this approximation.

A new π_{ij} approximation formula is described that requires only the sample x's. The properties of the new approximation are compared to the exact π_{ij}'s and the Hartley-Rao formula, and relationships among the π_{ij} approximations are given. Several small populations are used to empirically assess the different π_{ij} formulas.

Definitions and Notation:

1. $\bar{\pi}=\sum_{i=1}^{N}\pi_i/N=n/N$ (population mean of the π's)

2. $V(\pi)=\sum_{i=1}^{N}(\pi_i-\bar{\pi})^2/(N-1)$ (population variance of the π's)

3. $\bar{X}=\sum_{i=1}^{N}x_i/N$ (population mean of the x's)

4. $\bar{\pi}_s=\sum_{i=1}^{n}{}_s\,\pi_i/n$ (mean of the π's in the sample)

5. $\bar{x}_s=\sum_{i=1}^{n}{}_s\,x_i/n$ (mean of the x's in the sample)

6. $\pi_{ii}\equiv\pi_i$ (1)

7. For simple random sampling, $\pi_{ij}=n(n-1)/N(N-1)$ (2)

8. $T_x=\sum_{i=1}^{N}x_i=$ population total of the x's

9. $\pi px=$ inclusion probability proportional to x

2. PROPERTIES AND RELATIONS

The following properties hold for general variable probability, fixed-size sampling designs (cf. Hanurav, 1962; Brewer and Hanif, 1983):

1. $\sum_{i=1}^{N}\pi_i=n$ (3)

2. $\pi_{ij}=\pi_{ji}$ (symmetry) (4)

3. $\pi_{ij}=\pi_i$ when $\pi_j=1$ (5)

4. $\sum_{j\neq i}^{N}\pi_{ij}=(n-1)\pi_i$ (6)

5. $\sum_{i=1}^{N}\sum_{j\neq i}^{N}\pi_{ij}=n(n-1)$ (7)

6. $E(\bar{\pi}_s)=\frac{1}{n}\sum_{i=1}^{N}\pi_i^2$ (8)

<u>Proof</u>: $E(\bar{\pi}_s)=E(\sum_{i=1}^{n}\pi_i/n)=\sum_{i=1}^{N}\pi_i^2/n.$ □

7. For a πpx design, $E(\bar{x}_s)=nE(\bar{\pi}_s)/T_x=\sum_{i=1}^{N}x_i^2/T_x$ (9)

8. $\sum_{i=1}^{N}\pi_i^2=(N-1)V(\pi)+n^2/N$ (10)

<u>Proof</u>: $V(\pi)\equiv\left[\sum_{i=1}^{N}\pi_i^2-Nn^2/N^2\right]/(N-1)=\left[\sum_{i=1}^{N}\pi_i^2-n^2/N\right]/(N-1),$

which implies $(N-1)V(\pi)=\sum_{i=1}^{N}\pi_i^2-n^2/N.$ □

9. $\sum_{j\neq i}^{N}\sum^{N}(\pi_i\pi_j-\pi_{ij})=n-\sum_{i=1}^{N}\pi_i^2$ (11)

<u>Proof</u>: $\sum_{j\neq i}^{N}\sum^{N}\pi_{ij}=n(n-1)$ by (7), and

$\sum_{j\neq i}^{N}\sum^{N}\pi_i\pi_j=\sum_{i=1}^{N}\pi_i\sum_{j\neq i}^{N}\pi_j=\sum_{i=1}^{N}\pi_i(n-\pi_i)=n-\sum_{i=1}^{N}\pi_i^2$ from (3). □

10. $\sum_{j\neq i}^{N}\sum^{N}\pi_j\pi_{ij}=(n-1)\sum_{i=1}^{N}\pi_i^2=n(n-1)E(\bar{\pi}_s)$ (12)

<u>Proof</u>: $\sum_{j\neq i}^{N}\sum^{N}\pi_j\pi_{ij}=\sum_{i=1}^{N}\left[(\sum_{j=1}^{N}\pi_j\pi_{ij})-\pi_i\pi_{ii}\right]$

using equation (1) and reversing the order of summation in the double sum,

$=\sum_{j=1}^{N}\sum_{i=1}^{N}\pi_j\pi_{ij}-\sum_{i=1}^{N}\pi_i^2=\sum_{j=1}^{N}\pi_j\sum_{i=1}^{N}\pi_{ij}-\sum_{j=1}^{N}\pi_j^2$

$=\sum_{j=1}^{N}\pi_j\Big((n-2)\pi_j\Big)-\sum_{j=1}^{N}\pi_j^2$ by (6)

$=(n-1)\sum_{j=1}^{N}\pi_j^2=n(n-1)E(\bar{\pi}_s)$ by (8). □

11. For any πpx design, $V(\pi)=n^2V(x)/T_x^2=n^2[cv(x)]^2/N^2$. (13)

Proof: $V(\pi)=V(nx_i/T_x)=n^2V(x)/T_x^2$. □

The following theorem shows the intuitive result that the expected average of the sample π's for a πpx design is always greater than the expected average of the sample π's for an equal probability sample.

Theorem 2.1 $E(\bar{\pi}_s)\geq\bar{\pi}=n/N$ for πpx designs.

Proof: Expanding $E(\bar{\pi}_s)$ using (10) and (8), we obtain $E(\bar{\pi}_s)=(1/n)\left[(N-1)V(\pi)+n^2/N\right]=(N-1)V(\pi)/n+n/N\geq n/N$, since $V(\pi)\geq 0$. Equality holds only if $\pi_i=n/N$ for all $i=1,\ldots,N$. □

Corollary: $E(\bar{x}_s)\geq\bar{X}$.

Proof: Since $\bar{\pi}_s=n\bar{x}_s/T_x$, $E(\bar{x}_s)\geq\frac{T_x}{N}=\bar{X}$. □

3. APPROXIMATION FORMULAS

Several formulas for approximating pairwise inclusion probabilities are described in the following subsections. All of the approximations are derived for random-order, *vps* sampling.

3.1 Hartley-Rao Approximation

An approximation formula, based on an asymptotic theory of randomized systematic sampling for n fixed and $N\rightarrow\infty$, was derived by Hartley and Rao (1962):

$$\pi_{ij}^{hr} = \frac{(n-1)}{n}\pi_i\pi_j + \frac{(n-1)}{n^2}(\pi_i^2\pi_j+\pi_i\pi_j^2) - \frac{(n-1)}{n^3}\pi_i\pi_j\sum_{j=1}^{N}\pi_j^2$$
$$+\frac{2(n-1)}{n^3}(\pi_i^3\pi_j+\pi_i\pi_j^3+\pi_i^2\pi_j^2) - \frac{3(n-1)}{n^4}(\pi_i^2\pi_j+\pi_i\pi_j^2)\sum_{j=1}^{N}\pi_j^2 \tag{14}$$
$$+\frac{3(n-1)}{n^5}\pi_i\pi_j\Big(\sum_{j=1}^{N}\pi_j^2\Big)^2 - \frac{2(n-1)}{n^4}\pi_i\pi_j\sum_{j=1}^{N}\pi_j^3 .$$

Hartley and Rao claim this approximation is correct to $O(N^{-4})$. A truncated version of (14) is often used in practice to simplify computations and derivation of analytical results (cf. Cumberland and Royall, 1981 and Wolter, 1985). This truncated form is:

$$\pi_{ij}^{hrt} = \frac{n(n-1)x_i x_j}{T_x\left[T_x - x_i - x_j + \sum_{k=1}^{N} x_k^2/T_x\right]}$$
$$= \frac{(n-1)\pi_i\pi_j}{\left[n-\pi_i-\pi_j+\sum_{k=1}^{N}\pi_k^2/n\right]} \tag{15}$$
$$= \frac{(n-1)\pi_i\pi_j}{\left[n-\pi_i-\pi_j+E(\bar{\pi}_s)\right]}. \tag{16}$$

Hartley and Rao report that v_{YG} calculated with π_{ij}^{hr} is correct to $O(N^0)$, while v_{YG} calculated with π_{ij}^{hrt} is correct to $O(N^1)$. In this paper, π_{ij}^{hrt} is used to derive analytic results, while π_{ij}^{hr} is used for empirical comparisons.

Some empirical assessment of the accuracy of the approximation π_{ij}^{hr} has been done. Choudry *et al* (1985) reported that the Hartley-Rao approximation values were close to the exact π_{ij} values when N was greater than 15 in their study of the Canadian Labor Force Survey, but recommended exact π_{ij}'s for N<16. Hartley and Chakrabarty (1967) examined π_{ij}^{hr} for small N (N≤15) and suggested some modification in the formulas for these small populations. When both π_i and π_j were large, π_{ij}^{hr} did not perform well. Hajek (1981) questioned the approximations used by Hartley and Rao in deriving their formula, and presented a counterexample to support his claim. He wrote "... there is hardly a hope for reasonable asymptotic expressions for π_{ij} for n=2 and N→∞." Joshi (1983) claimed Hajek's counterexample was invalid, while Iachan (1983) supported Hajek's general conclusion. Bellhouse (1988) further elaborated on this controversy.

A disadvantage of the Hartley-Rao approximation formula is the requirement that all population x's must be known, so alternative approximation formulas are needed. It is also of interest that other approximations perform better than the Hartley-Rao formula under some circumstances.

3.2 New Approximation Formulas

Overton (1985) described an approximation to the pairwise inclusion probabilities, based on sampling with variable probability from a randomly ordered list. This approximation is derived by the following heuristic argument. On a line segment of length T_x, fixed points are located at intervals of length $k=T_x/n$, the sampling interval. Each point represents an "indicator", the position at which a sample element will be selected. The line segment is also partitioned by subsegments of the length of the x_i, $i=1,2,\ldots,N$. The approximate π_{ij} formula is then obtained by approximating the conditional probability, $\pi_{j \cdot i}$, and using the relation $\pi_{ij}=\pi_{j \cdot i}\pi_i$.

To approximate the conditional probability, locate unit i at the left end of the line segment, so that it occupies the length x_i from the origin. The first sample point thus identifies element i, and the remaining n-1 indicator points are uniformly spaced over T_x-x_i. Then the remaining N-1 population units are randomly permuted and placed end to end on the line segment. Clearly for a finite number of x's placed along this line segment, some arrangements are not possible. We will ignore these irregularities caused by the discrete set of x's in making this approximation. The left end of unit j cannot be located within x_j of the right hand end of the line segment, since it would then overlap the end of the line. The left end of unit j can then be placed anywhere on the line segment of length $T_x-x_i-x_j$. Thus the conditional probability that the part of the line segment covered by unit j contains a particular "indicator" point is approximated by $x_j/(T_x-x_i-x_j)$. There are (n-1) remaining indicator points on the line segment, so the probabililty that the part of the line segment covered by unit j contains any indicator point is approximately $(n-1)\left[x_j/(T_x-x_i-x_j)\right]$. An approximate formula for π_{ij} is then

$$\pi^*_{ij}=\pi_{j\,.\,i}\pi_i=\frac{(n-1)x_j}{(T_x-x_i-x_j)}\pi_i=\frac{(n-1)\pi_i\pi_j}{(n-\pi_i-\pi_j)}. \tag{17}$$

A simple adjustment to π^*_{ij} is made so that the approximation gives the correct pairwise inclusion probability for a simple random sample. The adjusted approximation is:

$$\pi^{o}_{ij}=\frac{n(n-1)x_i x_j}{T_x\left(T_x-\frac{x_i+x_j}{2}\right)} \tag{18}$$

$$=\frac{(n-1)\pi_i\pi_j}{n-\frac{1}{2}(\pi_i+\pi_j)} \tag{19}$$

$$=\frac{2(n-1)\pi_i\pi_j}{2n-\pi_i-\pi_j}. \tag{20}$$

Early in the investigation of the approximations, the adjustment to formula (18) from (17) was believed unnecessary. However, simulation studies showed far better behavior of the variance estimators using (18).

Although this derivation is based on list sampling, in which all the x's are known, the resultant formula (20) requires neither all the x's nor even T_x. Additionally, the symmetry property is satisfied. Both π^{o}_{ij} and π^{hr}_{ij} can be viewed as modifications of π^{*}_{ij} to satisfy symmetry (see equations 16, 17, and 19). The modification is incorporated into the denominator of π^{*}_{ij}: π^{o}_{ij} adds $(\pi_i+\pi_j)/2$ to the denominator of π^{*}_{ij}, while π^{hr}_{ij} adds $E(\bar{\pi}_s)$.

3.3 Alternative Approximation Formulas

Other pairwise inclusion probability approximation formulas can be easily constructed. Several are listed here as a matter of record. D. S. Robson suggested the following approximation:

$$\pi^{r}_{ij} = \frac{n(n-1)x_i x_j}{2T_x}\left[\frac{1}{T_x - x_i} + \frac{1}{T_x - x_j}\right]$$

$$= \frac{(n-1)\pi_i\pi_j}{\left[(n-\pi_i)(n-\pi_j)/(n-(\pi_i+\pi_j)/2)\right]}$$

$$= \frac{(n-1)\pi_i\pi_j}{2}\left[\frac{1}{n-\pi_i} + \frac{1}{n-\pi_j}\right].$$

This approximation satisfies equation (7) (see Lemma 4.4), but does not satisfy equation (6).

Another approximation, based on a modification of π^{hr}_{ij}, is

$$\pi^{hr1}_{ij} = (n-1)\pi_i\pi_j/\left[n - \pi_i - \pi_j + \frac{N}{n}\left(\frac{\pi_i^2+\pi_j^2}{2}\right)\right].$$

This approximation estimates the term $\sum_{k=1}^{N}\pi_k^2$ in π^{hr}_{ij} from (15) by N*average(of π_i^2 and π_j^2). Still another approximation based on π^{hr}_{ij} estimates the term $E(\bar{\pi}_s)$ in (16) unbiasedly by $\bar{\pi}_s$. Then, $\pi^{hr2}_{ij} = (n-1)\pi_i\pi_j/\left[n - \pi_i - \pi_j + \bar{\pi}_s\right]$.

Each of these approximations satisfies equations (2) and (4). All of these formulas have one of the advantages of π^{o}_{ij} in that only sample x's are required. Undoubtedly, many other approximations could be derived using other *ad hoc* methods. Of the approximations listed in this section, only π^{r}_{ij} was subjected to further study.

4. PROPERTIES OF APPROXIMATIONS

In this section, some of the properties of the approximation formulas relative to those of the exact π_{ij}'s. The basis of the investigation will be the properties of the true π_{ij}'s given by equations (5), (6), and (7). Empirical assessment of the approximations is based on the populations investigated by Rao and Singh (1973) described in Table 1.

Table 1 Description of Populations.

Popn	$E(\bar{\pi}_s)$	$Var(\pi)$	$max\{\pi\}$
1	0.12	0.16	0.20
3	0.17	0.38	0.28
4	0.44	5.31	0.69
5	0.48	6.32	0.86
6	0.21	1.16	0.27
9	0.17	0.66	0.34
10	0.15	3.72	0.24
11	0.20	0.07	0.24
12	0.24	1.50	0.42
13	0.29	2.22	0.64
14	0.21	1.14	0.36
15	0.26	1.41	0.42
19	0.12	0.21	0.22
23	0.30	1.44	0.41
24	0.30	1.44	0.41
26	0.35	2.88	0.53
27	0.17	0.35	0.20
31	0.21	0.76	0.31

Key:

Popn - Population number given in Rao and Singh (1973).

$E(\bar{\pi}_s)$, $Var(\pi)$, and the maximum π are calculated for n=2.

From equation (5), if $\pi_i=1$, then $\pi_{ij}=\pi_j$. Hartley and Rao (1962) and Hartley and Chakrabarty (1967) stated that π_{ij}^{hr} would not provide an accurate approximation for large π_i. Result 4.1 shows the effect of a large π_i on the approximation π_{ij}^{hrt}.

Result 4.1 If $\pi_i=1$, then

i) $\pi_{ij}^{hrt}<\pi_{ij}$ when $\pi_j>E(\bar{\pi}_s)$, and

ii) $\pi_{ij}^{hrt}\geq\pi_{ij}$ when $\pi_j\leq E(\bar{\pi}_s)$.

Proof: If $\pi_i=1$, $\pi_{ij}^{hrt}=\frac{(n-1)\pi_j}{n-1-\pi_j+E(\bar{\pi}_s)}=\pi_j/\left[1+\frac{E(\bar{\pi}_s)-\pi_j}{n-1}\right]$. □

The following result shows that when a large π_i is present, π_{ij}^o generally underestimates the true π_{ij} unless π_j is also relatively large.

Result 4.2 If $\pi_i=1$, then $\pi_{ij}^o \le \pi_j \ \forall \ j$.

Proof: If $\pi_i=1$, $\pi_{ij}^o=\frac{(n-1)\pi_j}{\left[n-\frac{1}{2}-\frac{\pi_j}{2}\right]}=\pi_j\left[\frac{2(n-1)}{2n-1-\pi_j}\right]\le\pi_j$. Equality holds if $\pi_j=1$. □

Note that π_{ij}^o does give the correct result when $\pi_i=\pi_j=1$. This indicates that the approximation is adequate when both π_i and π_j are large.

In practice, it is a simple matter to modify the definition of either π_{ij}^o or π_{ij}^{hr} so as to conform to equation (5) when $\pi_i=1$. The issue is what happens to the approximation when π_i is close to 1. To observe the numerical effect of a large π_i on the performance of the approximation formulas, two populations with a large π_i were included in the empirical study

Hartley and Rao (1962) claim π_{ij}^{hr} satisfies equation (6) to $O(N^{-4})$. Assessing π_{ij}^o relative to the property identified by equation (6), we obtain

$$\sum_{j\neq i}^{N}\pi_{ij}^o=(n-1)\pi_i\sum_{j\neq i}^{N}\frac{2\pi_j}{2n-\pi_i-\pi_j}=(n-1)\pi_i\sum_{j\neq i}^{N}\frac{\pi_j}{n-(\pi_i+\pi_j)/2}.$$

Although a useful algebraic simplification for the above relationship has not been found, empirical results (Table 2) indicate:

1) $\sum_{j\neq i}^{N}\pi_{ij}^o>(n-1)\pi_i$ for $\pi_i<E(\bar{\pi}_s)$, and

2) $\sum_{j\neq i}^{N}\pi_{ij}^o<(n-1)\pi_i$ for $\pi_i>E(\bar{\pi}_s)$.

Table 2 Empirical Assessment of Property Defined by Equation (6): $\sum_{j\neq i}^{N} \pi_{ij} = (n-1)\pi_i$.

Note: Values reported in columns 2 through 4 are

$$\left[(n-1)\pi_i - \sum_{j\neq i}^{N} \hat{\pi}_{ij}\right] \times 10{,}000,$$

where $\hat{\pi}_{ij}$ is the appropriate approximation formula. Values in the table for the exact π_{ij}'s would be 0. Rows are ordered by increasing value of π_i.

a) Population 3

	Approximation		
π_i	π_{ij}^{hr}	π_{ij}^{o}	π_{ij}^{r}
0.07	-0.12	-18.74	-19.51
0.07	-0.14	-19.31	-20.09
0.09	-0.25	-19.74	-20.48
0.11	-0.28	-18.98	-19.69
0.11	-0.28	-18.26	-18.95
0.11	-0.28	-18.05	-18.74
0.11	-0.28	-17.83	-18.51
0.13	-0.24	-15.56	-16.19
0.14	-0.03	-10.52	-11.11
0.16	0.25	-6.05	-6.63
0.18	1.59	8.35	7.66
0.20	2.94	19.21	18.34
0.23	6.46	41.24	39.78
0.28	22.29	108.55	104.13

b) Population 5

	Approximation		
π_i	π_{ij}^{hr}	π_{ij}^{o}	π_{ij}^{r}
0.01	-0.00	-12.31	-16.11
0.02	-0.01	-24.38	-31.85
0.03	-0.10	-47.77	-62.26
0.04	-0.20	-59.07	-76.90
0.13	-4.97	-156.49	-201.85
0.17	-10.53	-193.15	-248.56
0.24	-24.60	-234.87	-303.12
0.24	-24.60	-234.87	-303.12
0.26	-27.93	-239.36	-309.70
0.86	1779.57	1899.55	1553.46

Table 2 (Continued)

c) Population 6

π_i	Approximation		
	π_{ij}^{hr}	π_{ij}^{o}	π_{ij}^{r}
0.15	-0.21	-23.44	-23.93
0.17	0.01	-20.03	-20.41
0.17	0.10	-18.73	-19.08
0.18	0.64	-12.69	-12.95
0.20	1.98	-1.08	-1.27
0.20	1.98	-1.08	-1.27
0.21	2.29	1.17	0.97
0.21	2.62	3.51	3.30
0.24	5.73	22.45	22.03
0.27	12.41	53.81	52.63

d) Population 11

π_i	Approximation		
	π_{ij}^{hr}	π_{ij}^{o}	π_{ij}^{r}
0.16	0.02	-19.67	-19.99
0.17	0.34	-15.73	-15.97
0.18	0.47	-14.26	-14.47
0.19	1.15	-7.59	-7.73
0.20	1.60	-3.77	-3.89
0.20	1.60	-3.77	-3.89
0.20	2.13	0.38	0.26
0.21	2.76	4.87	4.74
0.24	6.95	29.37	28.92
0.24	7.58	32.55	32.03

e) Population 15

π_i	Approximation		
	π_{ij}^{hr}	π_{ij}^{o}	π_{ij}^{r}
0.06	-0.10	-33.19	-35.90
0.08	-0.29	-41.17	-44.36
0.09	-0.56	-46.80	-50.28
0.12	-1.10	-52.11	-55.76
0.19	-1.72	-45.09	-48.35
0.19	-1.72	-45.09	-48.35
0.25	1.40	-17.20	-20.19
0.27	5.12	3.01	-0.13
0.33	25.34	77.11	71.85
0.42	97.77	247.10	231.47

For all approximations, the magnitude of the deviation from equation (6) is very small, yet later results show the different approximations can result in very different behaviors of the variance estimators. A pattern evident from the populations examined is that the deviation from equation (6) for all approximations is related to the size of the inclusion probability of the unit. For the three approximations examined, $\sum_{j\neq i}^{N} \hat{\pi}_{ij} > (n-1)\pi_i$ for the smaller inclusion probabilities, and $\sum_{j\neq i}^{N} \hat{\pi}_{ij} < (n-1)\pi_i$ for larger π_i's. Generally, π_{ij}^{hr} has the smallest deviation from (6), while π_{ij}^{o} and π_{ij}^{r} are similar to each other in absolute deviation from (6).

The following lemma provides a direct comparison of π_{ij}^{o} and π_{ij}^{r}:

<u>Lemma</u> <u>4.3</u> $\pi_{ij}^{o} \leq \pi_{ij}^{r}$.

<u>Proof</u>:
$$\pi_{ij}^{r} - \pi_{ij}^{o} = \frac{(n-1)\pi_i\pi_j(2n-\pi_i-\pi_j)}{2(n-\pi_i)(n-\pi_j)} - \frac{2(n-1)\pi_i\pi_j}{2n-\pi_i-\pi_j}$$
$$= (n-1)\pi_i\pi_j\left[\frac{2n-\pi_i-\pi_j}{2(n-\pi_i)(n-\pi_j)} - \frac{2}{2n-\pi_i-\pi_j}\right]$$
$$= (n-1)\pi_i\pi_j\left[\frac{(2n-\pi_i-\pi_j)^2-4(n-\pi_i)(n-\pi_j)}{2(n-\pi_i)(n-\pi_j)(2n-\pi_i-\pi_j)}\right]$$
$$= (n-1)\pi_i\pi_j\left[\frac{(\pi_i-\pi_j)^2}{2(n-\pi_i)(n-\pi_j)(2n-\pi_i-\pi_j)}\right]. \qquad (21)$$

Since $\pi_i \leq 1 \ \forall\ i$, all terms in (21) are positive and the lemma follows. □

Lemma 4.3 leads to the obvious result that $\sum_{j\neq i}^{N} \pi_{ij}^{o} \leq \sum_{j\neq i}^{N} \pi_{ij}^{r}$.

To complete this assessment, recall the property identified by equation (7), $\sum_{j\neq i}^{N}\sum^{N}\pi_{ij}=n(n-1)$. The following lemma shows that π^{r}_{ij} satisfies (7):

<u>Lemma 4.4</u> $\sum_{j\neq i}^{N}\sum^{N}\pi^{r}_{ij}=n(n-1)$.

<u>Proof</u>: $\sum_{j\neq i}^{N}\sum^{N}\pi^{r}_{ij}=\sum_{j\neq i}^{N}\sum^{N} n(n-1)x_i x_j \frac{1}{2T_x}\left[\frac{1}{T_x-x_i}+\frac{1}{T_x-x_j}\right]$

$$=\frac{n(n-1)}{2T_x}\sum_{j\neq i}^{N}\sum^{N} x_i x_j\left[\frac{T_x-x_j+T_x-x_i}{(T_x-x_i)(T_x-x_j)}\right]$$

$$=\frac{n(n-1)}{T_x}\left[\sum_{j\neq i}^{N}\sum^{N}\frac{x_i x_j}{T_x-x_i}\right]$$

$$=\frac{n(n-1)}{T_x}\left[\sum_{i=1}^{N}\frac{x_i}{T_x-x_i}\sum_{j\neq i}^{N}x_j\right]$$

$$=\frac{n(n-1)}{T_x}\left[\sum_{i=1}^{N}\frac{x_i}{T_x-x_i}(T_x-x_i)\right]$$

$$=\frac{n(n-1)}{T_x}T_x=n(n-1).$$ □

Then for π^{o}_{ij}, we obtain

<u>Theorem 4.5</u> $\sum_{i=1}^{N}\sum_{j\neq i}^{N}\pi^{o}_{ij}\leq n(n-1)$.

<u>Proof</u>: Follows from Lemmas 4.3 and 4.4. □

Equation (7) is examined empirically for several small populations in Table 3. The results confirm Theorem 4.5, and also show that, at least for these populations, $\sum_{j\neq i}^{N}\sum^{N}\pi^{hr}_{ij}<n(n-1)$. For all populations in Table 3, π^{o}_{ij} satisfies (7) more closely than does π^{hr}_{ij}, using the deviations from n(n-1) as the measure of closeness. The magnitude of the deviation from (7) tends to increase with the population variance of the π's. Although π^{hr}_{ij} satisfies condition (6) more closely than π^{o}_{ij} for most populations, the "errors" in π^{o}_{ij} tend to cancel when the

<u>Table 3</u> Empirical Assessment of Property Defined by Equation (7): $\sum_{j\neq i}^{N}\sum^{N}\pi_{ij}=n(n-1)$.

Note:

1) Values reported in columns 2 and 3 are $\left[2-\sum_{j\neq i}^{N}\sum^{N}\pi_{ij}\right]\times 1000$.

2) $\sum_{i=1}^{N}\sum_{j\neq i}^{N}\pi^{r}_{ij}=n(n-1)$, so tabled values for π^{r}_{ij} would be 0.

3) Values for exact π_{ij}'s are 0.

4) Rows are ordered by Var(π) for the population.

Popn.	(2) π^{hr}_{ij}	(3) π^{o}_{ij}
11	2.46	0.24
6	2.75	0.39
1	0.92	0.51
19	1.18	0.68
27	1.72	0.39
10	0.54	0.24
3	3.16	1.43
9	6.04	3.16
31	5.47	2.24
14	9.82	5.20
15	12.41	4.66
24	11.94	3.09
23	11.94	3.09
12	19.25	9.47
13	64.50	28.81
26	27.00	8.72
4	71.46	26.07
5	168.66	69.73

summation in (7) is taken over all pairs of elements. Note also that π^r_{ij} is virtually identical to π^o_{ij} in (6), and has perfect cancellation in (7).

4.2 Assessment of "Weights"

The assessment of the properties of the π_{ij} formulas provides some insight into the relative accuracy of the approximations. To further investigate the effects on the variance estimators v_{HT} and v_{YG} of these differences in properties of the π_{ij} approximations, it is useful to define $a_{ij}=(\pi_i\pi_j-\pi_{ij})/\pi_{ij}$ for the $(i,j)^{th}$ pair of population units. Then,

$$v_{YG}=\frac{1}{2}\sum_{j\neq i}^{n}\sum^{n} a_{ij}\left(\frac{y_i}{\pi_i}-\frac{y_j}{\pi_j}\right)^2, \quad \text{and} \tag{22}$$

$$v_{HT}=\sum_{i=1}^{n}\left(\frac{y_i}{\pi_i}\right)^2(1-\pi_i)-\sum_{j\neq i}^{n}\sum^{n} a_{ij}\frac{y_i y_j}{\pi_i\pi_j}. \tag{23}$$

By defining a_{ij} in this way, the cross-product terms in v_{YG} and v_{HT} are functions of the a_{ij}'s and ratios y_i/π_i and y_j/π_j. The a_{ij}'s will be called "weights", because they weight the contribution of the terms involving functions of the ratios y_i/π_i and y_j/π_j. Analogous weights a^o_{ij} and a^{hrt}_{ij} are defined by substituting π^o_{ij} and π^{hrt}_{ij} respectively for π_{ij} in the formula for a_{ij}. Then,

$$a^o_{ij}=\frac{1}{n-1}\left[1-\frac{\pi_i+\pi_j}{2}\right], \text{ and } a^{hrt}_{ij}=\frac{1}{n-1}\left[1-\pi_i-\pi_j+E(\bar{\pi}_s)\right].$$

<u>Theorem 4.6</u>

i) If $(\pi_i+\pi_j)/2<E(\bar{\pi}_s)$, then $a^{hrt}_{ij}>a^o_{ij}$.

ii) If $(\pi_i+\pi_j)/2>E(\bar{\pi}_s)$, then $a^{hrt}_{ij}<a^o_{ij}$.

<u>Proof</u>: From $a_{ij}^{hrt} - a_{ij}^{o} = \frac{1}{n-1}\left[E(\bar{\pi}_s) - \frac{\pi_i+\pi_j}{2}\right]$, the theorem follows directly. □

<u>Corollary 4.7</u> If $(\pi_i+\pi_j)/2 < n/N$, then $a_{ij}^{hrt} > a_{ij}^{o}$.

<u>Proof</u>: The corollary follows directly since $n/N \le E(\bar{\pi}_s)$ by Theorem 2.1. □

Theorem 4.6 and its corollary show that when π_i and π_j are small relative to $E(\bar{\pi}_s)$, $a_{ij}^{hrt} > a_{ij}^{o}$. Corollary 4.7 provides the more interpretable bound, since n/N is the inclusion probability for equal probability sampling. Corollary 4.7 states that for the (i, j) pairs in which the average of π_i and π_j is less than the inclusion probability for an equal probability sample, the Hartley-Rao approximation provides greater weight to those pairs than the weight provided by π_{ij}^{o}.

An immediate implication of Corollary 4.7 is the following. Suppose that for a particular sample, most of the sample π's are small relative to $E(\bar{\pi}_s)$. By Corollary 4.7, we know that $a_{ij}^{hrt} > a_{ij}^{o}$. The effect on the variance estimators of using either π_{ij}^{o} or π_{ij}^{hrt} can be readily seen by examining (22) and (23). For example, v_{YG} calculated with π_{ij}^{hr} will be larger than v_{YG} calculated with π_{ij}^{o} if all sample π's are less than $E(\bar{\pi}_s)$. Also, the contribution to the negative piece of v_{HT} in equation (23) will be much larger when a_{ij}^{hrt} is used than when a_{ij}^{o} is used, if the sample π's are all less than $E(\bar{\pi}_s)$.

The properties of the approximate weights can also be evaluated in a manner analogous to the procedure employed to assess the π_{ij} approximations that used equations (6) and (7). Starting first with a property analogous to equation (7), define

$$\bar{a} = \sum_{i=1}^{N} \sum_{j \neq i}^{N} a_{ij}/N(N-1) \tag{24}$$

to be the average weight over all population pairs. $\bar{a}^{hrt}$ and $\bar{a}^{o}$ are defined similarly substituting a_{ij}^{hrt} and a_{ij}^{o} for a_{ij} in equation (24). The following lemmas lead to the result that $\bar{a}^{hrt} > \bar{a}^{o}$.

<u>Lemma 4.8</u> $\bar{a}^{hrt} = \frac{1}{n-1}\left[1-n/N\right] + \frac{1}{(n-1)}\left[E(\bar{\pi}_s) - n/N\right]$.

<u>Proof</u>: First note the folowing relationships,

$$\sum_{j \neq i}^{N}\sum^{N} \pi_i = \sum_{i=1}^{N} \pi_i (N-1) = n(N-1), \quad \text{and} \tag{25}$$

$$\sum_{j \neq i}^{N}\sum^{N} \pi_j = \sum_{i=1}^{N} (n - \pi_i) = nN - n = n(N-1). \tag{26}$$

Then,

$$\bar{a}^{hrt} = \sum_{i=1}^{N} \sum_{j \neq i}^{N} a_{ij}^{hrt}/N(N-1)$$

$$= \sum_{j \neq i}^{N}\sum^{N} \left[\frac{\pi_i \pi_j}{(n-1)\pi_i\pi_j/\left[n - \pi_i - \pi_j + E(\bar{\pi}_s)\right]} - 1\right]/N(N-1)$$

$$= \sum_{j \neq i}^{N}\sum^{N} \left[n - \pi_i - \pi_j + E(\bar{\pi}_s) - n + 1\right]/(n-1)N(N-1)$$

$$= \frac{1}{(n-1)} \sum_{j \neq i}^{N}\sum^{N} \left[1 - \pi_i - \pi_j + E(\bar{\pi}_s)\right]/N(N-1)$$

using (25) and (26) and combining like terms,

$$= \frac{1}{(n-1)}\left[(1 + E(\bar{\pi}_s) - 2n/N\right].$$

<u>Lemma 4.9</u> $\bar{a}^{o} = \frac{1}{(n-1)}\left[1 - n/N\right]$.

<u>Proof</u>: $$\bar{a}^{o} = \frac{1}{N(N-1)} \sum_{j \neq i}^{N}\sum^{N} \left(\frac{\pi_i\pi_j(2n - \pi_i - \pi_j)}{2(n-1)\pi_i\pi_j} - 1\right)$$

$$=\frac{1}{N(N-1)}\sum_{j\neq i}^{N}\sum^{N}\left(\frac{2n-\pi_i-\pi_j-2n+2}{2(n-1)}\right)$$

$$=\frac{1}{(n-1)2N(N-1)}\sum_{j\neq i}^{N}\sum^{N}(2-\pi_i-\pi_j)$$

$$=\frac{2N(N-1)-2n(N-1)}{2N(n-1)(N-1)} \quad \text{using (25) and (26).}$$

The lemma follows after algebraic simplification. □

Theorem 4.10 For any πpx design, $\bar{a}^{hrt} > \bar{a}^{o}$.

Proof: From Lemmas 4.8 and 4.9, $\bar{a}^{hrt} > \bar{a}^{o}$ if $E(\bar{\pi}_{\bullet}) > n/N$. This condition is always true by Theorem 2.1. □

No closed form expression has been found for $\bar{a}$, but empirical results from the populations examined in Table 4 indicate the following relationships:

1) $\bar{a} > \bar{a}^{o}$ in all populations, and
2) $\bar{a}$ may be greater or less than $\bar{a}^{hr}$.

Theorem 4.6 established that for small π_i and π_j, $a_{ij}^{hrt} > a_{ij}^{o}$. Theorem 4.10 shows that the population average weight for a_{ij}^{hrt} is greater than the population average weight for a_{ij}^{o}. Moreover, the average weight for a_{ij}^{o} is smaller than the average weight for the exact a_{ij}'s. Once again the influence of the population average weights on the variance estimators is best seen by examining equations (22) and (23).

Since the a_{ij}'s enter the sample with unequal probability, we could also consider a weighted population average, where the a_{ij}'s are weighted by their sample inclusion probability, π_{ij}. That is, define

<u>Table 4</u>. Comparison of Average Population a_{ij}'s (n=2):

$$\bar{a} = \sum_{j \neq i}^{N} \sum^{N} a_{ij} / N(N-1).$$

Note:

1) Values reported are the average weight times 100.

2) Rows are ordered by Var(π) of the populations.

Popn.	$\bar{a}$	$\bar{a}^{hr}$	$\bar{a}^{o}$
11	80.41	80.58	80.00
6	80.62	80.82	80.00
1	91.71	91.72	90.00
19	92.28	92.27	90.00
27	87.29	86.92	84.62
10	91.60	91.61	88.24
3	88.75	88.73	85.71
9	93.65	93.47	87.50
31	88.23	88.37	83.33
14	101.11	99.08	88.24
15	88.40	87.86	80.00
24	82.51	81.26	75.00
23	82.51	81.26	75.00
12	100.81	101.68	87.50
13	123.52	111.62	87.50
26	86.35	87.68	75.00
4	135.19	110.36	80.00
5	210.65	123.14	80.00

$$\bar{a}_w = \sum_{j\neq i}^{N}\sum^{N} \pi_{ij}a_{ij} / \sum_{j\neq i}^{N}\sum^{N} \pi_{ij} = \sum_{j\neq i}^{N}\sum^{N} \pi_{ij}a_{ij}/n(n-1).$$

Weighted averages of a_{ij}^{hrt} and a_{ij}^{o} could be defined similarly, substituting a_{ij}^{hrt} or a_{ij}^{o} for a_{ij} in the above equation. These weighted averages have another interpretation. Note that

$$E\left[\sum_{j\neq i}^{n}\sum^{n} a_{ij}\right] = \sum_{j\neq i}^{N}\sum^{N} a_{ij}\pi_{ij} = n(n-1)\,\bar{a}_w.$$

Thus the comparison of weighted population average weights, for any of a_{ij}, a_{ij}^{o}, and a_{ij}^{hrt}, would be equivalent to comparison of the expected value of the sum of the sample weights. These expected values are the same for the exact π_{ij}'s and the approximations π_{ij}^{o} and π_{ij}^{hrt}.

Theorem 4.11 $E\left[\sum_{j\neq i}^{n}\sum^{n} a_{ij}\right] = E\left[\sum_{j\neq i}^{n}\sum^{n} a_{ij}^{o}\right] = E\left[\sum_{j\neq i}^{n}\sum^{n} a_{ij}^{hrt}\right] = n - \sum_{i=1}^{N} \pi_i^2.$

Proof:

i) $E\left[\sum_{j\neq i}^{n}\sum^{n} a_{ij}\right] = \sum_{j\neq i}^{N}\sum^{N} a_{ij}\pi_{ij} = \sum_{i=1}^{N}\sum_{j\neq i}^{N} (\pi_i\pi_j - \pi_{ij}) = n - \sum_{i=1}^{N} \pi_i^2$ by equation (11).

ii) $E\left[\sum_{j\neq i}^{n}\sum^{n} a_{ij}^{o}\right] = \sum_{j\neq i}^{N}\sum^{N} \pi_{ij}\left(\frac{2-\pi_i-\pi_j}{2(n-1)}\right)$

$$= \frac{1}{2(n-1)}\left[2\sum_{j\neq i}^{N}\sum^{N}\pi_{ij} - \sum_{i=1}^{N}\pi_i \sum_{j\neq i}^{N}\pi_{ij} - \sum_{j\neq i}^{N}\sum^{N}\pi_j\pi_{ij}\right]$$

using equations (7) and (3) on the first two terms inside the brackets and using equation (12) on the last term yields

$$= \frac{1}{2(n-1)}\left[2n(n-1) - \sum_{i=1}^{N}\pi_i(n-1)\pi_i - (n-1)\sum_{i=1}^{N}\pi_i^2\right]$$

$$= \frac{1}{2(n-1)}\left[2n(n-1) - 2(n-1)\sum_{i=1}^{N}\pi_i^2\right] = n - \sum_{i=1}^{N}\pi_i^2.$$

iii) $E\left[\sum_{s}^{n}\sum_{s}^{n}{}_{j\neq i} a_{ij}^{hrt}\right] = \sum\sum_{j\neq i}^{N\,N} \pi_{ij}\left(\frac{1-\pi_i-\pi_j+E(\bar{\pi}_s)}{n-1}\right)$

$$= \frac{1}{n-1}\left[1+E(\bar{\pi}_s)\right] \sum\sum_{j\neq i}^{N\,N} \pi_{ij} - \frac{1}{n-1}\sum\sum_{j\neq i}^{N\,N} (\pi_i+\pi_j)\pi_{ij}$$

using (7) to simplify the first double sum, and (6) and (12) to simplify the second double sum

$$= n\left[1+E(\bar{\pi}_s)\right] - \frac{1}{n-1}\left[2(n-1)\sum_{i=1}^{N}\pi_i^2\right] \text{ using (8) for } E(\bar{\pi}_s)$$

$$= n + \sum_{i=1}^{N}\pi_i^2 - 2\sum_{i=1}^{N}\pi_i^2 = n - \sum_{i=1}^{N}\pi_i^2.$$

[

From Theorem 4.11, the approximations a_{ij}^{o} and a_{ij}^{hrt} have the appealing property that the expected sum of their sample weights equals the expected sum of the sample weights for the exact π_{ij}'s. An empirical verification of Theorem 4.11 is given in Table 5.

Table 5 shows $E(\sum_{s}^{n}\sum_{s}^{n}{}_{j\neq i} a_{ij})$ and the corresponding expected values using a_{ij}^{o} and a_{ij}^{hr}. The expected values for a_{ij} and a_{ij}^{o} are nearly the same, but in all populations, the expected value for a_{ij}^{hr} exceeds that of a_{ij}. A possible explanation is that π_{ij}^{hrt} was used in Theorem 4.11, while the more complicated formula π_{ij}^{hr} was used in calculating the values in Table 5. Thus it appears that Theorem 4.11 does not apply to the Hartley-Rao form of equation (14).

Theorems 4.6, 4.10, and 4.11 provide the following qualitative picture:

1) when π_i and π_j are small relative to $E(\bar{\pi}_s)$, $a_{ij}^{hr} > a_{ij}^{o}$, and the inequality reverses when π_i and π_j are large;

Table 5. Comparison of Weighted Averages of a_{ij}'s (n=2):

$$\bar{a}_w = \sum_{j\neq i}^{N}\sum^{N} \pi_{ij} a_{ij} / \sum_{j\neq i}^{N}\sum^{N} \pi_{ij} = \sum_{j\neq i}^{N}\sum^{N} \pi_{ij} a_{ij} / n(n-1).$$

Note:

1) Values reported are 100x(average weight).

2) Rows ordered by Var(π) of the population.

Popn.	$\bar{a}_w$	$\bar{a}_w^{hr}$	a_w^o
11	79.67	79.89	79.67
6	79.47	79.71	79.47
1	88.48	88.57	88.49
19	87.99	88.09	87.99
27	82.53	82.67	82.53
10	85.26	85.44	85.26
3	83.22	83.48	83.22
9	82.58	83.03	82.58
31	79.12	79.56	79.12
14	79.11	79.88	79.13
15	73.67	74.57	73.67
24	69.97	70.88	69.97
23	69.97	70.88	69.97
12	76.25	77.59	76.25
13	70.83	73.63	70.83
26	64.94	66.71	64.94
4	56.09	60.16	56.09
5	51.56	56.22	51.56

2) the unweighted average over all population pairs of the a_{ij}^{hr}'s is larger than the unweighted average of the a_{ij}^{o}'s;

3) the discrepancies between both approximations and the true a_{ij}'s average out over the set of all population pairs when the unequal probability of selecting an (i, j) pair is taken into consideration; while individual weights based on the approximations are not exactly equal to the true weights, there is a cancellation of errors so that the overall average weight is correct.

5. SUMMARY

Empirical and analytical investigation of several pairwise inclusion probability approximation formulas revealed that the properties of π_{ij}^{o} compared favorably with those of π_{ij}^{hr}. That π_{ij}^{o} requires only the sample x's is a major advantage for variance estimation because obtaining all the x's in the universe is often impractical, particularly if sampling from a map or area frame. The variance estimators v_{HT} and v_{YG} calculated with π_{ij}^{o} possess good theoretical properties and have performed well in empirical studies (Stehman and Overton, 1987ab, 1989). The advantageous characteristics of π_{ij}^{o} greatly diminish the problem of estimating the variance of the Horvitz-Thompson estimator for random-order, variable probability systematic sampling.

6. REFERENCES

Bellhouse, D. R. (1988). Systematic Sampling. Handbook of Statistics, Vol. 6, P. R. Krishnaiah and C. R. Rao, eds., 125-145. Elsevier Science Publishers: Amsterdam.

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

Chaudhuri, A., and Vos, J. W. E. (1988). Unified Theory and Strategies of Survey Sampling. North-Holland: New York.

Choudhry, G. H., Lee, H., and Drew, J. D. (1985). Cost-variance optimization for the Canadian Labour Force Survey. Survey Methodology 11, 33-50.

Connor, W. S. (1966). An exact formula for the probability that two specified sampling units will occur in a sample drawn with unequal probabilities and without replacement. J. Amer. Statist. Assoc. 61, 384-390.

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

Hajek, J. (1981). Sampling from a Finite Population. Marcel Dekker: New York.

Hanurav, T. V. (1962). Some sampling schemes in probability sampling. Sankhya Ser. A 24, 421-428.

Hartley, H. O., and Chakrabarty, R. P. (1967). Evaluation of approximate variance formulas in sampling with unequal probabilities and without replacement. Sankhya Ser. B 25, 201-208.

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.

Iachan, R. (1983). Reply to a letter to the editor by V. M. Joshi. American Statistician 37, 96.

Joshi, V. M. (1983). Letter to the editor. American Statistician 37, 96.

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.

Stehman, S. V., and Overton, W. S. (1987a). A comparison of Variance Estimators of the Horvitz-Thompson Estimator in Random-Order, Variable Probability, Systematic Sampling. Biometrics Unit Manuscript Bu-935-M, Cornell University.

Stehman, S. V., and Overton, W. S. (1987b). Estimating the variance of the Horvitz-Thompson Estimator in Variable Probability, Systematic Samples. Proceedings of the Section of Survey Research Methods of the American Statistical Association, pp. 743-748.

Stehman, S. V., and Overton, W. S. (1989). An Empirical, General Population Assessment of the Properties of Variance Estimators of the Horvitz-Thompson Estimator under Random-Order, Variable Probability Systematic Sampling. Biometrics Unit Manuscript Bu-936-M, Cornell University.

Sunter, A. (1986). Solutions to the problem of unequal probability sampling without replacement. Internat. Statist. Rev. 54, 33-50.

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.

Seely, Justus. Minimal sufficient statistics and completeness for multivariate normal families. January 1975.

Birkes, David, H. D. Brunk, John W. Lee. Generalized exponential families. January 1975.

Pierce, Donald A. and Bradley R. Sands. Extra-Bernoulli variation in binary data. November 1975.

Scheurman, H. Lynn and K. Norman Johnson. Computational experience with a renewable resource (timber) harvest model. November 1975.

Lembersky, Mark R. Maximum return from a collection of standard and special appreciation assets. January 1976.

Brunk, H. D. and N. F. G. Martin. An inequality involving powers of sums of powers. January 1976.

Birkes, David and Justus Seely. Three-way classification designs of resolutions III, IV, and V. March 1976.

Brunk, H. D. Univariate density estimation by orthogonal series. March 1976.

Brunk, H. D. Stone spaces in probability. March 1976.

Jan, Shan Wu, David Birkes, and Roger Petersen. A test of hypothesis for a patterned correlation matrix: A Monte Carlo study. February 1976.

5. Pierce, Donald A. A random effects model for matched pairs of binomial data. September 1976.

6. Johansen, Søren. Homomorphisms and general exponential families. October 1976.

7. Jacroux, Michael and Justus Seely. Some sufficient conditions for establishing (M,S)-optimality. February 1977.

8. Thompson, Steve. The study of bear populations. June 1977.

9. Thomas, D. R. and D. A. Pierce. Neyman's smooth goodness-of-fit test when the hypothesis is composite. April 1977.

0. Ramsey, Fred L. and J. Michael Scott. Use of circular plot surveys in estimating the density of a population with Poisson scattering. January 1978.

1. Kopecky, Kenneth J. and Donald A. Pierce. Efficiency of smooth goodness-of-fit tests. April 1978

2. Brunk, H. D. Models for integration. July 1978.

3. Lindstrom, F. Tom. Function minimization by the Nelder-Mead simplex method. October 1978.

4. Lindstrom, F. T. and D. S. Birkes. On the estimation of parameters in a non-linear one compartmen stochastic open model by apprximate likelihood functions. January 1979.

5. Ramsey, Fred L. Parametric models for line-transect surveys. February 1979.

6. Dodge, Y. and D. R. Thomas. On the performance of non-paramtric and normal theory multiple compar son procedures. February 1979.

7. Pierce, Donald A., William A. Stewart and Kenneth J. Kopecky. Distribution-free regression analysis of grouped survival data. August 1979.

8. Brunk, H. D. Bayesian least squares estimates of univariate regression functions. August 1979.

9. Lindstrom, F. T. An objective function surface exploration algorithm. August 1979.

0. Arthur, J. A comparison of optimization methods on a non-linear regression problem. August 1979.

1. Brunk, H. D. Estimation of stimulus-response curves by Bayesian least squares. November 1979.

2. Seely, Justus. Parametrizations and correspondence in linear models. November 1979.

3. Lindstrom, F. Tom. FUNMIN: An online-interactive mixed simplex-Marquardt least squares or likel hood minimization algorithm. January 1980.

4. Kafadar, Karen. An empirical investigation of small samples from symmetric populations for constructing robust confidence intervals. May 1980.

5. Chou, Chung-Kuang and David A. Butler. Hazardous-inspection initiation policies. June 1980.

6. Kafadar, Karen. Borrowing scale information in estimating location robustly. June 1980.

7. Butler, David A. and Gerald J. Lieberman. An early-accept modifcation to the test plans of milit standard 781C. June 1980.

8. Rosenfield, Donald B., Roy D. Shapiro and David A. Butler. Optimal strategies for selling an ass December 1980.

9. Dodge, Y. and F. T. Lindstrom. An alternative approach to least squares estimation when dealing with contaminated data. April 1981.

80. Brunk, H. D. Boolean universes for stochastic models. December 1981.

81. Butler, David A. and G. J. Lieberman. Inspection policies for fault location. September 1982.

82. Thompson, Steven K. and Fred L. Ramsey. Adaptive sampling of anaimal populations. March 1983.

83. Hussein, Abou-Bakr A. and David R. Thomas. Optimal design of np-charts. March 1983.

84. Shih, John and H. D. Brunk. Bayes least squares linear estimation of densities. March 1983.

85. Ramsey, Fred L. and Christopher P. Marsh. Diet dissimilarity. May 1983.

86. Seely, Justus and David Birkes. Parametrizations and resolution IV. May 1983.

87. Schafer, Daniel W. The adequacy of least squares before and after correction for attenuation. June 1983.

88. Seely, Justus. Some continuity properties of dispersion matrices of estimators in linear models. August 1983.

89. Brunk, H. D. Bayes least squares linear regression is asymptotically full Bayes: Estimation of special densities. November 1983.

90. Lee, Youngjo. The unbiased estimation of risk for discontinuous estimation of the mean of the multivariate normal distribution. November 1983.

91. Lee, Youngjo and O. Kempthorne. An estimation procedure for K-nested linear models. November 198

92. Lee, Youngjo. Asymptotic properties of the shrinkage estimation. November 1983.

93. Stark, Steven B. A model study of the alfalfa leafcutter bee - seed production system. November 1983.

94. Schafer, Daniel W. Combining information for measurement error model regression. January 1984.

95. Gray, Robert J. and Donald A. Pierce. Goodness-of-fit tests for censored survival data. January 1984.

96. Pierce, Donald A. and Dale L. Preston. Hazard function modelling for dose-response analysis of cancer incidence in the A-bomb survivor data. March 1984.

97. Pierce, Donald A., Dale L. Preston and Toranosuke Ishimaru. A method for analysis of cancer incidence in Japanese A-bomb survivors, with application to acute leukemia. March 1984.

98. Lee, Youngjo. A short note on the positive-part shrinkage estimator for the location parameter under arbitrary convex loss function. March 1984.

99. Rossi, Richard J. and F. L. Ramsey. P-values for extreme values of the normal, student's t, chi square and F statistics. May 1984.

100. Arthur, Jeffrey L. Vector maximization by efficient directions. August 1984.

101. Thomas, Dana L. and David R. Thomas. Confidence bands for percentiles in the linear regression model. August 1984.

102. Brunk, H. D. Coherence and consistency of inferences. September 1984.

103. Carter, Nancy J. and G. David Faulkenberry. Predicting unit variate values in a finite population. December 1984.

104. Limam, Mohamed and David R. Thomas. Simultaneous tolerance intervals for the linear regression model. February 1985.

105. Birkes, David R., Cliff B. Pereira and Justus F. Seely. Admissibility among affine sets of linear estimators under quadratic loss. March 1985.

106. Wildman, Valerie Jean and Fred L. Ramsey. Estimating effective area surveyed with the cumulative distribution function. May 1985.

107. Limam, Mohamed and David R. Thomas. Simultaneous tolerance intervals in the random one-way model with covariates. August 1985.

108. Pierce, Donald A. and Dale L. Preston. Analysis of cancer mortality in the atomic bomb survivor cohort.. September 1985.

109. Brunk, H. D. The compact ramification of a set, and induced measures. November 1985.

110. Schafer, D. W. Covariate measurement error in generalized linear models. November 1985.

111. Birkes, David and Yadolah Dodge. The number of minimally connected block designs. January 1986.

112. Brunk, H. D. Fully Coherent Inference. June 1986.

113. Overton, W. Scott. Working Draft, Analysis Plan for the Eastern Lake Survey, March 8, 1985. July, 1986.

114. Overton, W. Scott. A sampling plan for streams, in the national surface water survey. July 1986.

115. Overton, W. Scott. Phase II Analysis Plan, National Lake Survey -- Working Draft, April 15, 1987. April 1987.

116. Arthur, Jeffrey L., James O. Frendewey and Robert T. Sumichrast. GENLP: A FORTRAN Routine for Generating Linear Programs with Known Optimal Solutions. May 1987