UNEQUAL PROBABILITY SAMPLING WITH RANDOM STRATIFICATION

C. Asok and B. V. Sukhatme
Iowa State University, Ames, Iowa

Summary: For estimating the population total of the characteristic under study in the case of unequal probabilities and without replacement procedure it is often advantageous to select the units such that the inclusion probability of each unit is proportional to its size. Such sampling schemes are called Inclusion Probability Proportional to size (I.P.P.S.) schemes. However there is a scarcity of such schemes in the literature which are practically useful for sample size $n > 2$. In this article use of random stratification has been examined in devising I.P.P.S. procedures for sample size $n > 2$ by using the procedures for sample size 2. In particular, an alternate generalization of Durbin's procedure for $n > 2$ has been suggested which is adoptable with much ease in practice and which yields a more efficient estimator than the Sampford's procedure.

1. Introduction: For estimating the population total Y of a characteristic y defined over a population of size N it is often advantageous to select the sample with unequal probability and without replacement. In such situations a more often considered estimator is the Horvitz-Thompson (H.T.) estimator, viz.,

$$\hat{Y}_{H.T.} = \sum_{1}^{n} y_i/\pi_i \qquad (1)$$

with variance

$$V(\hat{Y}_{H.T.}) = \sum_{1}^{N} Y_i^2/\pi_i + \sum_{i}^{N} \sum_{j(\neq i)}^{N} \frac{\pi_{ij}}{\pi_i\pi_j} Y_iY_j - Y^2 \qquad (2)$$

where n is the sample size, π_i is the probability of including the ith unit in the sample and π_{ij} is the probability for the ith and jth units to be both included in the sample. When information on an auxiliary characteristic x assuming the value X_i on the ith unit is available for all the units where Y_i is approximately proportional to X_i, considerable reduction in the variance can be achieved by making $\pi_i \alpha X_i$.

Such a scheme must obviously satisfy the condition

$$\pi_i = np_i , \qquad (3)$$

where $p_i = X_i/X$, X being the sum of all the X_i's.

Even though several authors have proposed schemes for sample size two that satisfy the condition $\pi_i = 2p_i \leq 1$, not many of these are useful for generalizing to samples of size $n > 2$. As such there is a scarcity of sampling schemes satisfying the condition $\pi_i = np_i$ in the literature which are practically useful for samples of arbitrary size. Moreover the strict applicability of the existing methods of unequal probability sampling without replacement including the calculation of unbiased estimates of sampling error is out of question in certain kinds of large scale survey work on grounds of practicability. Thus there is a need for evolving methods which retain the advantages of unequal probability sampling without replacement but are rather easier to apply in practice and only involve a slight loss of exactness. In this article we will investigate the role of random stratification in developing schemes that are practically useful and are applicable in large scale surveys, by making use of the schemes that are useful for smaller sample sizes. In evaluating the expressions for the inclusion probabilities π_{ij} and hence for the variance of the Horvitz-Thompson estimator we adopt an asymptotic theory considered first by Hartley and Rao (4) and subsequently used by Rao (5). The authors have considered the same approach also in their earlier papers (1,2).

Before dealing with the specific schemes to be proposed we first give a brief outline of some of the concepts and results presented in (2).

Let N be a multiple of K and let N/K = M, say. Let G denote a typical group of M units out of N units of the population $\mathcal{U}$. There are infact $\binom{N}{M}$ such groups. Let $\mathcal{G} = \{G_1, G_2, \ldots, G_{\binom{N}{M}}\}$ be the set of all such groups.

Definition 1: An ordered K-tuple $\theta_i = (G_{i_1}, G_{i_2}, \ldots, G_{i_K})$ is said to be a partition of the population $\mathcal{U}$ if

$$G_{i_j} \in \mathcal{G} , \qquad j = 1,2,\ldots,K$$

$$G_{i_j} \cap G_{i_{j'}} = \emptyset, \quad j \neq j'$$

and $\mathcal{U} = \bigcup_{j=1}^{K} G_{i_j}$

Definition 2: Two partitions $\theta_i = (G_{i_1}, G_{i_2}, \ldots, G_{i_K})$ and $\theta_{i'} = (G_{i'_1}, G_{i'_2}, \ldots, G_{i'_K})$ of the population $\mathcal{U}$ are said to be equivalent if one is just a rearrangement of the other, i.e., if each G_{i_j} is some $G_{i'_{j'}}$ and viceversa.

Definition 3: Two partitions θ_i and $\theta_{i'}$ are said to be distinct if they are not equivalent.

Theorem 1: The total number, A, of distinct partitions of the population $\mathcal{U}$ with groups of size M each is given by

$$A = \frac{N!}{K!\,(M!)^K} \qquad (4)$$

$\mathcal{G} = \{\theta_1, \theta_2, \ldots, \theta_A\}$ denotes the set of all distinct partitions.

Theorem 2: The total number, A_1, of distinct partitions of the population $\mathcal{U}$ with groups of size M each such that a particular pair of units (U_i, U_j) falls in the same group is given by

$$A_1 = \frac{(N-2)!}{(K-1)!\ (M-2)!\ (M!)^{K-1}} \tag{5}$$

Theorem 3: The total number, A_2, of distinct partitions of the population $\mathcal{U}$ with groups of size M each such that a particular pair of units (U_i, U_j) fall in different groups is given by

$$A_2 = \frac{(N-2)!}{(K-2)!\ \{(M-1)!\}^2\ (M!)^{K-2}} \tag{6}$$

$\mathcal{G}_1(i,j) = \{\theta_1, \theta_2, \ldots, \theta_{A_1}\}$ denotes the set of all distinct partitions such that (U_i, U_j) is in the same group.

$\mathcal{G}_2(i,j) = \{\theta_1, \theta_2, \ldots, \theta_{A_2}\}$ denotes the set of all distinct partitions such that (U_i, U_j) are in different groups.

The following relations among $\mathcal{G}_1(i,j)$, $\mathcal{G}_2(i,j)$ and $\mathcal{G}$ are immediate

$$\mathcal{G}_1(i,j) \cup \mathcal{G}_2(i,j) = \mathcal{G}$$

and

$$\mathcal{G}_1(i,j) \cap \mathcal{G}_2(i,j) = \emptyset$$

An obvious check on (4), (5) and (6) is provided by the relation

$$A_1 + A_2 = A \tag{7}$$

2. A Simple Scheme for Sample Size Two Utilizing Random Stratification: We will consider the scheme of selecting a sample of size two by adopting the Durbin's procedure with random stratification. The scheme is as follows:

(i) Split the population at random into three groups of equal sizes and select one group from among the three groups with probability proportional to the sum of the p_t's of the units belonging to that group.

(ii) Select two units utilizing the Durbin's procedure from the group that has been selected in step (i) utilizing the p_ts.

For this scheme of sampling the probability of including the ith unit in the sample is

$$\pi_i = \frac{1}{A} \cdot \sum_{\mathcal{G}} \pi^{(g)} \cdot \frac{2p_i}{S_g} \tag{8}$$

where S_g is the sum of the p_t's of the units belonging to the gth primary stage unit, and $\pi^{(g)}$, the probability of selecting the gth primary stage unit is

$$\pi^{(g)} = S_g$$

Hence we get from (8)

$$\pi_i = 2p_i \tag{9}$$

Expression for the pairwise inclusion probability π_{ij} for this scheme is

$$\pi_{ij} = \frac{1}{A} \cdot \sum_{\mathcal{G}_1(i,j)} S_q D_{i_j}\ , \tag{10}$$

where S_q is the sum of the p_t's of the units belonging to the qth primary stage unit that contains the pair (U_i, U_j) of a given partition of $\mathcal{G}_1(i,j)$; and D_{i_j} is the probability of including the pair (U_i, U_j) together under the Durbin's procedure, given that the qth primary stage unit has been selected in step (i). We have

$$D_{i_j} = \frac{\frac{2p_i}{S_q} \cdot \frac{p_j}{S_q}}{1 + \Sigma' \frac{p_t/S_q}{1 - 2p_t/S_q}} \cdot \left[\frac{1}{1-2p_i/S_q} + \frac{1}{1-2p_j/S_q}\right] \tag{11}$$

where Σ' denotes the summation taken over all the units that belong to the qth primary stage unit. Expression (10) for π_{ij} can alternately be written as

$$\pi_{ij} = \frac{A_1}{A} \cdot \frac{1}{A_1} \sum_{\mathcal{G}_1(i,j)} S_q \cdot D_{i_j}$$

$$= \frac{A_1}{A} \cdot E[S_q \cdot D_{i_j}] \tag{12}$$

where E denotes the expectation over the scheme of selecting $(\frac{N}{3} - 2)$ units from among the (N-2) population units excluding U_i and U_j.

Assuming that N is moderately large and p_i is of $O(N^{-1})$ we will get by using the asymptotic theory of Hartley and Rao (4) that

$$\pi_{ij} = 2p_ip_j[1 + \{(p_i + p_j) - \Sigma p_t^2\} + \{2(p_i^2 + p_j^2) - 2\Sigma p_t^3 - 16p_ip_j + 15\ (p_i + p_j)\ \Sigma p_t^2 - 15(\Sigma p_t^2)^2\}] \tag{13}$$

correct to $O(N^{-4})$.

From (9) and (13) it follows that the scheme under consideration satisfies the conditions of Theorem 2 in (1) with $a_2 = -16$ and so we have by applying the theorem,

$$V(\hat{Y}_{H.T.}) = \frac{1}{2}[\Sigma p_i z_i^2 - \Sigma p_i^2 z_i^2] - \frac{1}{2}[2\Sigma p_i^3 z_i^2 - \Sigma p_i^2 \cdot \Sigma p_i^2 z_i^2 + 16(\Sigma p_i^2 z_i)^2]\ , \tag{14}$$

correct to $O(N^0)$,

where $$z_i = (\frac{Y_i}{p_i} - Y) \tag{15}$$

To the same order of approximation, the variance of the H.T. estimator under the Durbin's procedure is given by (Equation (31) with n = 2 in (1))

$$V(\hat{Y}_{H.T.})_D = \frac{1}{2}[\Sigma p_i z_i^2 - \Sigma p_i^2 z_i^2] - \frac{1}{2}[2\Sigma p_i^3 z_i^2 - \Sigma p_i^2 \cdot \Sigma p_i^2 z_i^2] \quad (16)$$

From (14) and (16) we have

$$V(\hat{Y}_{H.T.})_D - V(\hat{Y}_{H.T.}) = 8(\Sigma p_i^2 z_i)^2 \geq 0,$$

which shows that the H.T. estimator for the scheme under consideration has a uniformly smaller variance than the H.T. estimator for the Durbin's procedure.

3. Use of Random Stratification in Getting Improved Estimates: In Section 2 we have presented a scheme for sample size 2 that utilizes the idea of random stratification and provides a better estimate than most of the existing schemes. In this section we will discuss the role of random stratification in getting an improved estimate for any given scheme that satisfy the conditions of Theorem 2 of [1].

The procedure for selecting a sample of size 2n is as follows:

(i) Split the population at random into three groups of equal sizes and select two groups from among these three such that the inclusion probability of any group is proportional to the sum of the p_t's of the units belonging to that group.

(ii) Select n units each from the two selected groups independently by adopting any I.P.P.S. scheme that satisfies the conditions of Theorem 2 of [1].

With the same notations used in section 2 we have for this scheme of sampling the inclusion probability

$$\pi_i = \frac{1}{A} \cdot \sum_{G} 2S_g \cdot \frac{np_i}{S_g} \quad (17)$$

$$= 2np_i$$

and

$$\pi_{ij} = \frac{1}{A} \sum_{G_1(i,j)} 2S_q \cdot \pi^{(q)}_{ij} + \frac{1}{A} \sum_{G_2(i,j)} \pi^{(r,s)} \cdot \frac{np_i}{S_r} \cdot \frac{np_j}{S_s} \quad (18)$$

where $\pi^{(q)}_{ij}$ is the probability of including the pair of units (U_i, U_j) when step (ii) is adopted in the qth group that contains U_i and U_j in a given partition of $G_1(i,j)$; and $\pi^{(r,s)}$ is the probability of including the rth and sth groups together when step (i) is adopted where the rth group contains U_i and sth group contains U_j in a given partition of $G_2(i,j)$.

The expression for $\pi^{(q)}_{ij}$ from theorem 2 of [1] is

$$\pi^{(q)}_{ij} = \frac{n(n-1)p_ip_j}{S_q^2}\Big[1 + \Big\{\frac{p_i + p_j}{S_q} - \frac{\Sigma' p_t^2}{S_q^2}\Big\} + \Big\{\frac{2(p_i^2 + p_j^2)}{S_q^2} - \frac{2\Sigma' p_t^3}{S_q^3} + \frac{a_n p_i p_j}{S_q^2} - (a_n + 1)\frac{(p_i + p_j)\Sigma' p_t^2}{S_q^3} + (a_n + 1)\frac{(\Sigma' p_t^2)^2}{S_q^4}\Big]\quad (19)$$

correct to $O(N^{-4})$, where Σ' denotes the summation over all the units belonging to the qth group and a_n is a constant that may depend on n.

Using a similar procedure as in the previous section we will get

$$\frac{1}{A} \sum_{G_1(i,j)} 2S_q \cdot \pi^{(q)}_{ij} = 2n(n-1)p_ip_j[1 + \{(p_i + p_j) - \Sigma p_t^2\} + \{2(p_i^2 + p_j^2) - 2\Sigma p_t^3 + (9a_n - 16)p_ip_j - (9a_n - 15)(p_i + p_j)\Sigma p_t^2 + (9a_n - 15)(\Sigma p_t^2)^2\}] \quad (20)$$

correct to $O(N^{-4})$.

The second component in (18) is

$$\frac{1}{A} \sum_{G_2(i,j)} \pi^{(r,s)} \frac{np_i}{S_r} \cdot \frac{np_j}{S_s} = n^2 \cdot \frac{1}{A} \sum_{G_2(i,j)} \pi^{(r,s)} \frac{p_i}{S_r} \cdot \frac{p_j}{S_s}$$

where the expression for $\pi^{(r,s)}$ is known to be given by

$$\pi^{(r,s)} = \pi^{(r)} + \pi^{(s)} - 1$$

$$= 2S_r + 2S_s - 1,$$

substituting in the above we get after considerable algebra

$$\frac{1}{A} \sum_{G_2(i,j)} \pi^{(r,s)} \frac{np_i}{S_r} \cdot \frac{np_j}{S_s} = 2n^2 p_ip_j[1 + \{(p_i + p_j) - \Sigma p_t^2\} + \{2(p_i^2 + p_j^2) - 2\Sigma p_t^3 - 7p_ip_j + 6(p_i + p_j)\Sigma p_t^2 - 6(\Sigma p_t^2)^2\}] \quad (21)$$

correct to $O(N^{-4})$.

Substituting from (20) and (21) into (18) we get

$$\pi_{ij} = 2n(2n-1)p_ip_j[1 + \{(p_i + p_j) - \Sigma p_t^2\} + \{2(p_i^2 + p_j^2) - 2\Sigma p_t^3 + b_n p_i p_j - (b_n + 1)(p_i + p_j)\Sigma p_t^2 + (b_n+1)(\Sigma p_t^2)^2\}], \quad (22)$$

correct to $O(N^{-4})$, where

$$b_n = \frac{(n-1)(9a_n - 16) - 7n}{(2n - 1)} \qquad (23)$$

Equations (17) and (22) show that this scheme again satisfies the conditions of Theorem 2 of [1].

Hence it follows that instead of using any given I.P.P.S. scheme for sample size 2n, we will get a better estimate by adopting the procedure described in this section, if the condition

$$a_{2n} - b_n > 0 \qquad (24)$$

is satisfied.

As is shown in the following the condition $a_{2n} - b_n > 0$ is satisfied for the Goodman and Kish procedure as well as the Sampford's procedure.

(i) Goodman and Kish procedure: For the Goodman and Kish procedure we have

$$a_n = 2 \text{ for all } n$$

Substituting this in (23) we get

$$b_n = \frac{2(n-1)-7n}{(2n-1)} = \frac{-(5n+2)}{(2n-1)} < 2 = a_{2n}$$

(ii) Sampford's procedure: For the Sampford's procedure

$$a_n = -(n-2),$$

from which we get

$$a_{2n} - b_n = \frac{n(5n+2)}{(2n-1)} > 0$$

The unequal probability schemes that are easily applicable for general sample sizes are rather scarce in the literature owing to the complications involved. Thus the procedure considered in this section would be advantageous to adopt for getting a sample of four units by applying it to any given simple procedure presented for sample size 2.
From (23) we have

$$a_4 - b_2 = a_4 - 3a_2 + 10 \qquad (25)$$

Thus for all those schemes useful for sample size 2, one can adopt the procedure under consideration advantageously if the condition $a_4 - 3a_2 + 10 > 0$ is satisfied. For example, for the procedure of Yates and Grundy (1953) and the procedure of Durbin (1953), this condition is satisfied.

4. Randomized m-Stage Procedure with Durbin's Scheme: When the Durbin's scheme (1967) is adopted in step (ii) of the procedure described in Section 3 for getting a sample of size 4, it gives a more efficient estimator than the Sampford's procedure for sample size 4 which is a generalization of the Durbin's scheme. It can be easily seen through a conditional argument that the procedure of section 3 can be adopted in successive stages, for any sample size of the form $n = 2^m$, where m is any positive integer, which provides a more efficient estimator. In this section we will consider the scheme utilizing the Durbin's procedure by considering a randomized m-stage design for selecting a sample of size 2^m.

The procedure is as follows:

(i) Split the population of N units at random into three equal groups and select two groups from among the three such that the inclusion probability of any group is proportional to the sum of the p_t's of the units belonging to that particular group.

Within each of the above selected groups, which could be denoted as primary stage units, perform the following procedure independently.

(ii) Split the units belonging to this group at random into three equal groups and select two groups from among the three such that the inclusion probability of any group is proportional to the sum of the p_t's of the units belonging to that group.

Repeat the procedure described in step (ii) independently within each of the selected units at each stage until we select 2^{m-1} units of the (m-1)th stage.

(iii) Within each of the (m-1)th stage units that are selected in step (ii), apply the Durbin's procedure independently for selecting a sample of size 2.

The above procedure would yield a sample of size 2^m.
In what follows we assume for mathematical convenience that N is a multiple of 3^{m-1}.

The notations we use here are similar to those adopted in section 2.

We denote the total number of distinct arrangements that can be made of the population of N units into three equal groups by $R_N(2)$; the total number of distinct arrangements that can be made of the population of N units into three equal groups such that a given pair of units (U_i, U_j) belong to two different groups by $R_N(2,1)$; and the total number of distinct arrangements that can be made of the population of N units into three equal groups such that a given pair of units (U_i,U_j) belong to the same group by $R_N(2,2)$.

It follows from Theorems 1, 2 and 3 that

$$R_N(2) = N!/6(\tfrac{N}{3}!)^3 \qquad (26)$$

$$R_N(2,1) = (N-2)!/\{(\tfrac{N}{3} - 1)!\}^2(\tfrac{N}{3}!) \qquad (27)$$

and

$$R_N(2,2) = (N-2)!/2\{(\tfrac{N}{3} - 1)!\}(\tfrac{N}{3}!)^2 \qquad (28)$$

Let $\mathcal{R}_N(m)$ denote the collection of all arrangements such that within each stage the arrangements are distinct, and $R_N(m)$ be the cardinality of the set $\mathcal{R}_N(m)$.

By an inductive argument it can be seen that

$$R_N(m) = R_N(2) \cdot \{R_{N/3}(m-1)\}^3 \qquad (29)$$

With respect to any particular pair (U_i,U_j) of the population units, the collection $\mathcal{R}_N(m)$ of all arrangements is the union of mutually disjoint sets $\mathcal{R}_N(m,t)$ $(t = 1,2,\ldots,m)$ where $\mathcal{R}_N(m,1)$ denotes the collection of all arrangements wherein the pair (U_i,U_j) belong to different primary stage units, $\mathcal{R}_N(m,t)$, $2 \le t \le m-1$, denotes the collection of all arrangements wherein the pair (U_i,U_j) belong to the same primary stage unit, same secondary stage unit,..., same (t-1)th stage unit, but different tth stage units; $\mathcal{R}_N(m,m)$ denotes the collection of all arrangements wherein the pair (U_i,U_j) belong to the same (m-1)th stage unit.

Let $R_N(m,t)$ be the cardinality of the set $\mathcal{R}_N(m,t)$ $1 \le t \le m$.

Theorem 4: For the $R_N(m,t)$, $1 \le t \le m$ and $R_N(m)$, the following relations hold.

$$\frac{R_N(m,t)}{R_N(m,t-1)} = \frac{1}{3}, \text{ for } 2 \le t \le m-1 \qquad (30)$$

$$\frac{R_N(m,m)}{R_N(m,m-1)} = \frac{(N-3^{m-1})}{2N} \qquad (31)$$

$$\frac{R_N(m,t)}{R_N(m)} = \frac{2N}{3^t(N-1)}, \ 1 \le t \le m-1 \qquad (32)$$

and

$$\frac{R_N(m,m)}{R_N(m)} = \frac{(N-3^{m-1})}{3^{m-1}(N-1)} \qquad (33)$$

Now, for the randomized m-stage procedure with the Durbin's scheme the inclusion probability for the ith population unit is,

$$\pi_i = \frac{1}{R_N(m)} \cdot \sum_{\mathcal{R}_N(m)} \left[\frac{2p_i}{S_{g_1g_2\cdots g_{m-1}}} \cdot \frac{2S_{g_1g_2\cdots g_{m-1}}}{S_{g_1g_2\cdots g_{m-2}}} \cdot \frac{2S_{g_1g_2\cdots g_{m-2}}}{S_{g_1g_2\cdots g_{m-3}}} \cdots \frac{2S_{g_1g_2}}{S_{g_1}} \cdot 2S_{g_1}\right]$$

$$= \frac{1}{R_N(m)} \cdot \sum_{\mathcal{R}_N(m)} 2^m p_i$$

$$= 2^m \cdot p_i$$

$$= np_i , \qquad (34)$$

where $S_{g_1g_2\cdots g_\ell}$ $(1 \le \ell \le m-1)$ denotes the sum of the p_t's of the units belonging to the g_ℓth ℓth stage unit of the $g_{\ell-1}$th $(\ell-1)$th stage unit of the ... g_2th second stage unit of the g_1th primary stage unit.

Probability of including the pair of units (U_i,U_j) together in the sample is

$$\pi_{ij} = \frac{1}{R_N(m)} \cdot \left[\sum_{\mathcal{R}_N(m,1)} C_1 + \sum_{\mathcal{R}_N(m,2)} C_2 + \cdots + \sum_{\mathcal{R}_N(m,t)} C_t + \cdots + \sum_{\mathcal{R}_N(m,m)} C_m\right] \qquad (35)$$

where $C_t (1 \le t \le m)$ is the conditional probability of selecting the pair $(U_i,U_j$ given the arrangement belonging to the tth category.

Evaluation of $\frac{1}{R_N(m)} \cdot \sum_{\mathcal{R}_N(m,1)} C_1$: In a given arrangement of the first category let U_i belong to the r_{m-1}th (m-1)th stage unit of the r_{m-2}th (m-2)th stage unit of the ... r_2th second stage unit of the r_1th primary stage unit and let U_j belong to the s_{m-1}th (m-1)th stage unit of the s_{m-2}th (m-2)th stage unit of the ... s_2th second stage unit of the s_1th primary stage unit.

The conditional probability C_1 is given by

$$C_1 = \left\{\frac{2p_i}{S_{r_1r_2\cdots r_{m-1}}} \cdot \frac{2S_{r_1r_2\cdots r_{m-1}}}{S_{r_1r_2\cdots r_{m-2}}} \cdot \frac{2S_{r_1r_2\cdots r_{m-2}}}{S_{r_1r_2\cdots r_{m-3}}} \cdots \frac{2S_{r_1r_2}}{S_{r_1}}\right\} \cdot \left\{\frac{2p_j}{S_{s_1s_2\cdots s_{m-1}}} \cdot \frac{2S_{s_1s_2\cdots s_{m-1}}}{S_{s_1s_2\cdots s_{m-2}}} \cdot \frac{2S_{s_1s_2\cdots s_{m-2}}}{S_{s_1s_2\cdots s_{m-3}}} \cdots \frac{2S_{s_1s_2}}{S_{s_1}}\right\} (2S_{r_1} + 2S_{s_1} - 1)$$

$$= 2^{2m-1} p_i p_j \left(\frac{1}{S_{r_1}} + \frac{1}{S_{s_1}} - \frac{1}{2S_{r_1}S_{s_1}}\right)$$

Thus we have

$$\frac{1}{R_N(m)} \sum_{\mathcal{R}_N(m,1)} C_1 = 2^{2m-1} p_i p_j \cdot \frac{R_N(m,1)}{R_N(m)} \cdot E\left[\frac{1}{S_{r_1}} + \frac{1}{S_{s_1}} - \frac{1}{2S_{r_1}S_{s_1}}\right]$$

where E denotes the average taken over all the arrangements belonging to the first category.

Proceeding in a similar way as in section (2), we get

$$\frac{1}{R_N(m)} \sum_{\mathcal{R}_N(m,1)} C_1 = 2^{2m-1} p_i p_j [1 + \{(p_i+p_j) - \Sigma p_t^2\} + \{2(p_i^2+p_j^2) - 2\Sigma p_t^3 - 7p_ip_j + 6(p_i+p_j)\Sigma p_t^2 - 6(\Sigma p_t^2)^2\}] \qquad (36)$$

correct to $O(N^{-4})$.

In a similar way we get for $2 \le t \le m-1$,

$$\frac{1}{R_N(m)} \sum_{\mathcal{R}_N(m,t)} C_t = 2^{2m-t} p_i p_j [1 + \{(p_i+p_j) - \Sigma p_t^2\} + \{2(p_i^2 + p_j^2) - 2\Sigma p_t^3 - (3^{2t} - 2)p_ip_j + (3^{2t}-3)(p_i+p_j)\Sigma p_t^2 - (3^{2t}-3)(\Sigma p_t^2)^2\}] \qquad (37)$$

correct to $O(N^{-4})$,

and

$$\frac{1}{R_N(m)} \sum_{R_N(m,m)} C_m = 2^m p_i p_j [1 + \{(p_i + p_j) - \Sigma p_t^2\} + \{2(p_i^2 + p_j^2) - 2\Sigma p_t^3 - (2 \cdot 3^{2m-2} - 2)p_i p_j + (2 \cdot 3^{2m-2} - 3)(p_i + p_j)\Sigma p_t^2 - (2 \cdot 3^{2m-2} - 3)(\Sigma p_t^2)^2\}] \quad (38)$$

correct to $O(N^{-4})$.
Substituting from (36), (37) and (38) into (35) we get

$$\pi_{ij} = 2^m(2^m - 1)p_i p_j [1 + \{(p_i + p_j) - \Sigma p_t^2\} + \{2(p_i^2 + p_j^2) - 2\Sigma p_t^3 + B_m \cdot p_i p_j - (B_m + 1)(p_i + p_j)\Sigma p_t^2 + (B_m + 1)(\Sigma p_t^2)^2\}] \quad (39)$$

correct to $O(N^{-4})$, where

$$B_m = \frac{1}{7(2^m - 1)} [23 \cdot 2^m - 32 \cdot 9^{m-1} - 14] \quad (40)$$

Thus for the randomized m-stage procedure with the Durbin's scheme, the expression for π_i is given by (34) and the expression for π_{ij} correct to $O(N^{-4})$ is given by (39).

Since the conditions of theorem 2 of [1] are satisfied, the variance of the H.T. estimator correct to $O(N^0)$ for randomized m-stage procedure with the Durbin's scheme is

$$V(\hat{Y}_{H.T.})_{RD} = \frac{1}{2^m} \cdot [\Sigma p_i z_i^2 - (2^m - 1)\Sigma p_i^2 z_i^2] \frac{-(2^m - 1)}{2^m} \cdot [2\Sigma p_i^3 z_i^2 - \Sigma p_i^2 \cdot \Sigma p_i^2 z_i^2 - B_m \cdot (\Sigma p_i^2 z_i)^2] \quad (41)$$

Randomized m-stage procedure with the Durbin's scheme is an alternative to the Sampford's procedure as a generalization of the Durbin's scheme for samples of size $n > 2$.

Since the simplicity of this randomized procedure to be adopted in large scale surveys is evident relative to the procedure of Sampford, it will be interesting to study the relative performance of the two methods.

Theorem 5: Variance corresponding to the randomized m-stage procedure with the Durbin's scheme for sample size 2^m is uniformly smaller than the variance corresponding to the Sampford's procedure for sample size 2^m and the efficiency of the randomized m-stage procedure relative to the Sampford's procedure increases as the sample size increases.

Proof: Variance of the Horvitz-Thompson estimator correct to $O(N^0)$ for the Sampford's procedure as given in [1] is

$$V(\hat{Y}_{H.T.})_{Samp} = \frac{1}{2^m} \cdot]\Sigma p_i z_i^2 - (2^m - 1)\Sigma p_i^2 z_i^2] \frac{-(2^m - 1)}{2^m} \cdot [2\Sigma p_i^3 z_i^2 - \Sigma p_i^2 \Sigma p_i^2 z_i^2 + (2^m - 2)(\Sigma p_i^2 z_i)^2] \quad (42)$$

From (41) and (42) we get

$$V(\hat{Y}_{H.T.})_{Samp} - V(\hat{Y}_{H.T.})_{RD} = \frac{1}{7} \cdot D_m \cdot (\Sigma p_i^2 z_i)^2 \quad (43)$$

where $D_m = \frac{1}{2^m} \cdot (32 \cdot 9^{m-1} - 7 \cdot 4^m - 2^{m+1})$

We have

$$D_2 = 42 > 0 \quad (44)$$

and

$$D_{m+1} = \frac{1}{2^{m+1}} (32 \cdot 9^m - 7 \cdot 4^{m+1} - 2^{m+2}) > \frac{9}{2^{m+1}} (32 \cdot 9^{m-1} - 7 \cdot 4^m - 2^{m+1}) = \frac{9}{2} D_m, \text{ for all m} \quad (45)$$

(44) and (45) together imply that D_m is nonnegative and monotome increasing.

Hence it follows from (43) that $V(\hat{Y}_{H.T.})_{Samp} - V(\hat{Y}_{H.T.})_{RD}$ is nonnegative and is larger for larger values of m.

Instead of the Durbin's scheme one can use any efficient scheme at the (m-1)th stage of the randomized m-stage procedure wherein the gains are expected to be substantial. The formulae for π_{ij} and hence the variance of the corresponding Horvitz-Thompson estimator could be derived using exactly the same technique. Applicability of these randomized varying probability schemes in large scale surveys is quite evident compared to the complicated procedures that are existent in the literature whose applicability is doubtful in large scale surveys.

Acknowledgement: This work was supported in part under OEC contract No. 0-73-6640 by the U.S. Office of Education, Department of Health, Education and Welfare.

References:

1. Asok, C. and Sukhatme, B. V. (1974). On Sampford's procedure of unequal probability sampling without replacement. Proceedings of the Social Statistics Section. American Statistical Association: 220-225.
2. _____ (1975). On Rao, Hartley and Cochran's procedure of unequal probability sampling without replacement. March 1975 meetings of the Institute of Mathematical Statistics, St. Paul, Minnesota.
3. Durbin, J. (1967). Design of multi-stage surveys for the estimation of sampling errors. Applied Statistics 16: 152-164.
4. Hartley, H. O. and Rao, J. N. K. (1962). Sampling with unequal probabilities and without replacement. Annals of Mathematical Statistics 33: 350-374.
5. Rao, J. N. K. (1965). On two simple schemes of unequal probability sampling without replacement. Journal of the Indian Statistical Association 3: 173-180.
6. Sampford, M. R. (1967). On sampling without replacement with unequal probabilities of selection. Biometrika 54: 499-513.