ON THE EFFICIENCY COMPARISON OF TWO πps SAMPLING STRATEGIES

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

1. Introduction: Let $\{U_1, U_2, \ldots, U_N\}$ be a finite population of a known number N of identifiable units and consider the problem of estimating the population total Y of a characteristic y with the value Y_i on unit U_i. When information on an auxiliary characteristic x highly correlated with y is available it is often advantageous to select a sample of size n with varying probabilities and without replacement. For any sampling design an unbiased estimator of Y proposed by Harvitz and Thompson is

$$\hat{Y}_{H.T.} = \sum_{i \in s} Y_i/\pi_i, \tag{1.1}$$

where the sum is over all distinct units of the sample s and π_i is the probability of including U_i in a sample of size n. The variance of $\hat{Y}_{H.T.}$ is given by

$$V(\hat{Y}_{H.T.}) = \sum^N \frac{Y_i^2}{\pi_i} + \sum_i^N \sum_{j(\neq i)}^N \frac{\pi_{ij}}{\pi_i \pi_j} Y_i Y_j - Y^2, \tag{1.2}$$

where π_{ij} is the probability for the i-th and j-th units to be both in the sample. (1.2) evidently reduces to zero when π_i is proportional to Y_i which suggests that considerable reduction in the variance can be achieved by making $\pi_i \propto x_i$. Such a scheme must obviously satisfy the condition

$$\pi_i = np_i , \tag{1.3}$$

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

Several schemes have been proposed in the literature that satisfy condition (1.3). However not many of them are applicable for sample size greater than two. Further none of these procedures, owing to the complications involved, are strictly applicable in large scale surveys. In this connection it is worthwhile to quote Durbin (1953, p. 267). He says:

"The strict application of the usual methods of unequal probability sampling without replacement, including the calculation of unbiased estimates of sampling error is out of the question in certain kinds of large-scale survey work on grounds of practicability. There is therefore a need for 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." Also, practically nothing is known as to how the different procedures compare among themselves as measured by the variance of the corresponding estimators. In an earlier article the authors have compared the procedure of Goodman and Kish with that of Sampford and have concluded that Sampford's procedure yields a uniformly better estimator than the procedure of Goodman and Kish. In this paper we compare the procedure of Goodman and Kish with that of Hanurav. Also we confine to the simple but important case of sample size 2. We desecribe the Hanurav's procedure in the following:

Without loss of generality let

$$0 < X_1 \leq X_2 \leq \cdots \leq X_{N-1} \leq X_N \leq \frac{X}{2} . \tag{1.4}$$

we first describe sampling scheme A covering the special case

$$X_{N-1} = X_N . \tag{1.5}$$

Sampling Scheme A:
Step 1. Select two units from the population with probability p_K for U_K and with replacement. If the sample consists of distinct units accept it; otherwise reject the sample and proceed to Step 2. Select two units from the population with probabilities proportional to p_K^2 and with replacement. Again, if the sample consists of distinct units accept it; otherwise reject and proceed to further steps. In general, if the 1st, 2nd, ... (m-1)th steps result in rejections, the units are drawn in the m^{th} step with probabilities proportional to $p_1^{2^{m-1}}$, $p_2^{2^{m-1}}$, ... $p_N^{2^{m-1}}$. It has been shown by Hanurav that sampling scheme A terminates after a finite number of steps with probability 1. Also the inclusion probabilities π_i and π_{ij} are given by

$$\pi_i = 2p_i \tag{1.6}$$

and
$$\pi_{ij} = 2p_i p_j [1 + \sum_{K=1}^{\infty} W_K] , \tag{1.7}$$

$$\text{where } W_K = \frac{(p_i p_j)^{2^K - 1}}{S_{(1)} S_{(2)} \cdots S_{(K)}} \tag{1.8}$$

$$\text{with } S_{(t)} = \sum_{K=1}^{N} p_K^{2^t} . \tag{1.9}$$

Now restriction (1.5) is dropped to generalize scheme A. This generalized scheme, which is denoted as sampling scheme B, is described as follows:

Sampling Scheme B:
Step 1. Conduct a binomial trail with probability of success δ given by

$$\delta = \frac{2(1-p_N)(p_N - p_{N-1})}{(1-p_N-p_{N-1})} . \tag{1.10}$$

If the trial results in success proceed to step 2; otherwise proceed to step 3.
Step 2. Select one of the units $U_1, U_2, \ldots, U_{N-1}$ with probabilities proportional to $p_1, p_2, \ldots, p_{N-1}$. If U_j is the unit thus selected, accept U_N and U_j as the unordered sample.
Step 3. Proceed with the sampling scheme A with the probabilities p_i replaced by p_i^* given by

$$p_i^* = \frac{p_i}{1-p_N+p_{N-1}} \quad \text{for} \quad 1 \le i \le N-1$$

$$\text{and} \quad p_N^* = p_{N-1}^* = \frac{p_N - \frac{1}{2}\delta}{1-\delta} \tag{1.11}$$

For sampling scheme B, π_i and π_{ij} are given by

$$\pi_i = 2p_i \tag{1.12}$$

$$\pi_{ij} = (1-\delta)\phi_{ij} \text{ for } 1 \le i \ne j < N,$$

$$\text{and} \quad \pi_{Nj} = \delta\{p_j/(1-p_N)\} + (1-\delta)\phi_{Nj}, \tag{1.13}$$

$$\text{where} \quad \phi_{ij} = 2p_i^* p_j^* \, (1 + \sum_{K=1}^{\infty} W_K^*), \tag{1.14}$$

$$W_K^* = \frac{(p_i^* p_j^*)^{2^K - 1}}{S_{(1)}^* S_{(2)}^* \cdots S_{(K)}^*}, \tag{1.15}$$

$$S_{(t)}^* = \sum_{K=1}^{N} p_K^{*2^t}. \tag{1.16}$$

Contrary to Hanurav's claim it can be seen from (1.7) and (1.13) as to how complicated it is to calculate the pairwise probabilities. Since the expression for pairwise probability is in terms of an infinite series, its exact numerical value for given data can never be calculated and as such one must resort to some kind of approximation for getting the pairwise probabilities and hence the variance. As the condition $\pi_{ij} < \pi_i \pi_j$ is satisfied for this procedure one can conclude that it yields an estimator which has a uniformly smaller variance than the customary estimator in sampling with replacement. As the method has been subsequently extended by Hanurav to cover the case of arbitrary sample size it will be of interest to study the relative performance of this method relative to the procedure of Goodman and Kish and of Sampford. Hartley and Rao used an asymptotic approach for deriving the expression for π_{ij} of the Goodman and Kish procedure and hence the variance of the H.T. estimator. As such it would be realistic for comparison purposes to derive the approximate expressions for π_{ij} and hence the variance for the Hanurav's procedure using the asymptotic approach of Hartley and Rao. These approximations should be of value for their own sake, since the simplicity of computation is one of the factors to be considered in choosing a sampling procedure. We will first evaluate the π_{ij} and hence the variance for scheme A under the asumption of Hartley and Rao viz., N is large and p_i is of $O(N^{-1})$.

2. Evaluation of π_{ij} and $V(\hat{Y}_{H.T.})$ for scheme A:

In order to evaluate the variance correct to $O(N^{1})$ we have to evaluate π_{ij} correct to $O(N^{-3})$. Also for using in the case of smaller size populations the variance correct to $O(N^0)$ is to be evaluated by evaluating π_{ij} correct to $O(N^{-4})$.

When p_i is of $O(N^{-1})$ it can be easily seen that $S_{(t)}$ is of $O(N^{-2^t+1})$ from which it follows that W_K will be of $O(N^{-K})$.

Hence the expression for π_{ij} correct to $O(N^{-4})$ is

$$\pi_{ij} = 2p_i p_j [1 + \frac{p_i p_j}{\Sigma p_t^2} + \frac{p_i^3 p_j^3}{\Sigma p_t^2 \Sigma p_t^4}]. \tag{2.1}$$

Substituting from (1.6) and (2.1) into (1.2), simplifying and retaining terms to $O(N^0)$ only we get

$$V(\hat{Y}_{H.T.})_A = \Sigma \frac{Y_i^2}{2p_i} + \frac{1}{2}[Y^2 - \Sigma Y_t^2] + \frac{1}{2\Sigma p_t^2}[(\Sigma Y_t p_t)^2 - \Sigma Y_t^2 p_t^2] + \frac{1}{2\Sigma p_t^2 \Sigma p_t^4}(\Sigma Y_t p_t^3)^2 - Y^2$$

$$= \frac{1}{2}[\Sigma \frac{Y_t^2}{p_t} - Y^2] - \frac{1}{2}[\Sigma Y_t^2 - \frac{(\Sigma Y_t p_t)^2}{\Sigma p_t^2}] - \frac{1}{2\Sigma p_t^2}[\Sigma Y_t^2 p_t^2 - \frac{(\Sigma Y_t p_t^3)^2}{\Sigma p_t^4}]$$

$$= \frac{1}{2}\Sigma p_t z_t^2 - \frac{1}{2}[\Sigma p_t^2 z_t^2 - \frac{(\Sigma p_t^2 z_t)^2}{\Sigma p_t^2}] - \frac{1}{2\Sigma p_t^2}[\Sigma p_t^4 z_t^2 - \frac{(\Sigma p_t^4 z_t)^2}{\Sigma p_t^4}] \tag{2.2}$$

where $z_t = \frac{Y_t}{p_t} - Y$.

When the variance is considered to $O(N^1)$ only we get

$$V(\hat{Y}_{H.T.})_A = \frac{1}{2}\Sigma p_t z_t^2 - \frac{1}{2}[\Sigma p_t^2 z_t^2 - \frac{(\Sigma p_t^2 z_t)^2}{\Sigma p_t^2}] \tag{2.4}$$

The term of $O(N^2)$ in the above viz., $\frac{1}{2}\Sigma p_t z_t^2$ is the variance of the customary estimator in the case of sampling with replacement. Hence the term of $O(N^1)$, viz., $\frac{1}{2}[\Sigma p_t^2 z_t^2 - \frac{(\Sigma p_t^2 z_t)^2}{\Sigma p_t^2}]$ represents the reduction in variance achieved by adopting scheme A of Hanurav. Following the method adopted in obtaining (2.2) one can in fact obtain the exact variance for scheme A as follows:

From (1.6) and (1.4) with $W_0 = 1$ we get

$$\sum_i \sum_{j(\ne i)} \frac{\pi_{ij}}{\pi_i \pi_j} Y_i Y_j = \frac{1}{2} \sum_i \sum_{j(\ne i)} Y_i Y_j (\sum_{K=0}^{\infty} W_K),$$

$$= \frac{1}{2} \sum_{K=0}^{\infty} \sum_i^N \sum_{j(\ne i)}^N W_K Y_i Y_j$$

$$= \frac{1}{2} \sum_{K=0}^{\infty} R_K \tag{2.5}$$

where $$R_K = \sum_{i}^{N} \sum_{j(\neq i)}^{N} W_K Y_i Y_j$$

$$= \frac{1}{S_{(1)}S_{(2)}\cdots S_{(K)}}[(\Sigma Y_t p_t^{2^K-1})^2 - \Sigma Y_t^2 p_t^{2^{K+1}-2}] .$$

Substituting from (2.5) into (1.2) we get

$$V(\hat{Y}_{H.T.})_A = \Sigma \frac{Y_t^2}{2p_t} - \frac{1}{2}\Sigma Y_t^2 + \frac{1}{2\Sigma p_t^2}[(\Sigma Y_t p_t)^2 - \Sigma Y_t^2 p_t^2] + \frac{1}{2\Sigma p_t^2 \Sigma p_t^4}[(\Sigma Y_t p_t^3)^2 - \Sigma Y_t^2 p_t^6] + \ldots$$

Rearranging the terms we get

$$V(\hat{Y}_{H.T.})_A = \frac{1}{2}[\Sigma \frac{Y_t^2}{p_t} - Y^2] - \sum_{r\geq 1} \frac{1}{2\Sigma p_t^2 \Sigma p_t^4 \cdots \Sigma p_t^{2^{r-1}}}[\Sigma Y_t^2 p_t^{2^r-2} - \frac{(\Sigma Y_t p_t^{2^r-1})^2}{\Sigma p_t^{2^r}}]$$

Substituting $\Sigma z_t^2 p_t^{2^r} - \frac{(\Sigma z_t p_t^{2^r})^2}{\Sigma p_t^{2^r}}$ for $\Sigma Y_t^2 p_t^{2^r-2} - \frac{(\Sigma Y_t p_t^{2^r-1})^2}{\Sigma p_t^{2^r}}$ and $\Sigma p_t z_t^2$ for $\Sigma \frac{Y_t^2}{p_t} - Y^2$ in the above we get

$$V(\hat{Y}_{H.T.})_A = \frac{1}{2}\Sigma p_t z_t^2 - \sum_{r\geq 1} \frac{1}{2\Sigma p_t^2 \Sigma p_t^4 \cdots \Sigma p_t^{2^{r-1}}} [\Sigma p_t^{2^r} z_t^2 - \frac{(\Sigma p_t^{2^r} z_t)^2}{\Sigma p_t^{2^r}}] \tag{2.6}$$

3. Evaluation of π_{ij} and $V(\hat{Y}_{H.T.})$ for Scheme B:

For evaluating $V(\hat{Y}_{H.T.})_B$ correct to $O(N^0)$ we have to evaluate π_{ij} correct to $O(N^{-4})$.

From (1.11) we get by expanding,

$$p_i^* = \frac{p_i}{1-p_N+p_{N-1}} = p_i\{1 + (p_N-p_{N-1}) + (p_N-p_{N-1})^2 + (p_N-p_{N-1})^3 + \ldots\} \text{ for } i = 1,2,\ldots,N-1$$

and $$p_N^* = p_{N-1}\{1 + (p_N-p_{N-1}) + (p_N-p_{N-1})^2 + (p_N-p_{N-1})^3 + \ldots\} . \tag{3.1}$$

Substituting in (1.16) we get

$$S^*_{(\ell)} = \sum_{K=1}^{N} p_K^{*2^\ell} = (\Sigma p_t^{2^\ell} - p_N^{2^\ell} + p_{N-1}^{2^\ell})\{1 + (p_N-p_{N-1}) + (p_N-p_{N-1})^2 + \ldots\}^{2^\ell} \tag{3.2}$$

From (3.1) and (3.2) we get that p_i^* is of $O(N^{-1})$ and $S^*_{(\ell)}$ is of $O(N^{-2^\ell+1})$.

Hence it follows from (1.15) that W_K^* is of $O(N^{-K})$. As δ is of $O(N^{-1})$, it is evident from (1.13) that in order to evaluate π_{ij} to $O(N^{-4})$ we should first evaluate ϕ_{ij} to $O(N^{-4})$. Substituting from (3.1) and (3.2) into (1.14) we get correct to $O(N^{-4})$,

$$\phi_{ij} = 2p_i p_j[1+\{2(p_N-p_{N-1}) + \frac{p_i p_j}{\Sigma p_t^2}\} + \{\frac{2(p_N-p_{N-1})p_i p_j}{\Sigma p_t^2} + 3(p_N-p_{N-1})^2 + \frac{(p_N^2-p_{N-1}^2)p_i p_j}{(\Sigma p_t^2)^2} + \frac{p_i^3 p_j^3}{\Sigma p_t^2 \Sigma p_t^4}\}]$$

for $1 \leq i \neq j > N$

and $$\phi_{Nj} = 2p_{N-1}p_j[1+\{2(p_N-p_{N-1}) + \frac{p_{N-1}p_j}{\Sigma p_t^2}\} + \{\frac{2(p_N-p_{N-1})p_{N-1}p_j}{\Sigma p_t^2} + 3(p_N-p_{N-1})^2 + \frac{(p_N^2-p_{N-1}^2)p_{N-1}p_j}{(\Sigma p_t^2)^2} + \frac{p_{N-1}^3 p_j^3}{\Sigma p_t^2 \Sigma p_t^4}\}]. \tag{3.3}$$

Substituting from (1.10) and (3.3) into (1.13) we get correct to $O(N^{-4})$,

$$\pi_{ij} = 2p_i p_j[1 + \frac{p_i p_j}{\Sigma p_t^2} + \{\frac{p_i^3 p_j^3}{\Sigma p_t^2 \Sigma p_t^4} + \frac{(p_N^2-p_{N-1}^2)p_i p_j}{(\Sigma p_t^2)^2} - (p_N^2 - p_{N-1}^2)\}] \text{ for } i \leq i \neq j < N$$

and $$\pi_{Nj} = 2p_j(p_N-p_{N-1})\{1+(p_N+p_{N-1})+(p_N+p_{N-1})^2\} + 2p_{N-1}p_j[1 + \frac{p_{N-1}p_j}{\Sigma p_t^2} + \{\frac{p_{N-1}^3 p_j^3}{\Sigma p_t^2 \Sigma p_t^4} + \frac{(p_N^2-p_{N-1}^2)p_{N-1}p_j}{(\Sigma p_t^2)^2} - (p_N^2-p_{N-1}^2)\}] . \tag{3.4}$$

Substituting from (3.4) into (1.2) we get correct to $O(N^0)$,

$$V(\hat{Y}_{H.T.})_B = \frac{1}{2}\Sigma p_t z_t^2 - \frac{1}{2}[\Sigma p_t^2 z_t^2 - \frac{(\Sigma p_t^2 z_t)^2}{\Sigma p_t^2}] - \frac{1}{2\Sigma p_t^2}[\Sigma p_t^4 z_t^2 - \frac{(\Sigma p_t^4 z_t)^2}{\Sigma p_t^4}] + \frac{1}{2\Sigma p_t^2}(p_N^2-p_{N-1}^2)(\Sigma p_t^2 z_t)[\frac{\Sigma p_t^2 z_t}{\Sigma p_t^2} - 2z_N] \tag{3.5}$$

From (2.2) and (3.5) we get that the H.T. estimator has the same variance correct to $O(N^1)$ for either of the schemes A and B and is given by

$$V(\hat{Y}_{H.T.})_A = V(\hat{Y}_{H.T.})_B = \frac{1}{2}\Sigma p_t z_t^2 - \frac{1}{2}[\Sigma p_t^2 z_t^2 - \frac{(\Sigma p_t^2 z_t)^2}{\Sigma p_t^2}] \tag{3.6}$$

Hartley and Rao have obtained the variance of the H.T. estimator for the procedure of Goodman and Kish and the variance correct to $O(N^1)$ is given by

$$V(\hat{Y}_{H.T.})_{G.K.} = \frac{1}{2}\Sigma p_t z_t^2 - \frac{1}{2}\Sigma p_t^2 z_t^2 \tag{3.7}$$

Later it has been shown by Rao (1963, 1965) that the variance is given by (3.7) also for the procedures of Durbin (1953, 1967) and Yates and Grundy (1953). From this one might be tempted to conjecture that (3.7) infact holds for any πps procedure. (3.6) shows that this may not always be the case.

From (3.6) and (3.7) we have

$$V(\hat{Y}_{H.T.})_A - V(\hat{Y}_{H.T.})_{G.K.} = V(\hat{Y}_{H.T.})_B - V(\hat{Y}_{H.T.})_{G.K.} = \frac{1}{2}\frac{(\Sigma p_t^2 z_t)^2}{\Sigma p_t^2} \geq 0 \qquad (3.8)$$

From (3.6), (3.7) and (3.8) we conclude that

Theorem 1: When the variance of the corresponding H.T. estimator is considered to $O(N^1)$, Goodman and Kish procedure has a uniformly smaller variance than the Hanurav's procedure.

For N not sufficiently large the approximation to $O(N^1)$ of the variance may not be quite satisfactory and as such one might have to consider the approximation to $O(N^0)$. Variance of the H.T. estimator for the Goodman and Kish procedure derived by Hartley and Rao (1962) may be written as

$$V(\hat{Y}_{H.T.})_{G.K.} = \frac{1}{2}[\Sigma p_t z_t^2 - \Sigma p_t^2 z_t^2] - \frac{1}{2}[2\Sigma p_t^3 z_t^2 - \Sigma p_t^2 \cdot \Sigma p_t^2 z_t^2 - 2(\Sigma p_t^2 z_t)^2] \qquad (3.9)$$

Comparisons of (3.9) with either (2.2) or (3.5) is rather hard and may not lead to any positive conclusion. So in the next section we will try to compare both the procedures under a well known super population model.

4. Comparison of the two procedures under a super population model:

In order to study the relative performance of different I.P.P.S. (Inclusion Probability Proportional to Size) schemes as measured by the variance of the corresponding H.T. estimator, it is convenient to assume some knowledge regarding the relationship between the variate y and the auxiliary characteristic x. Since unequal probability sampling is resorted to in the situations where y is approximately proportional to x it is reasonable to assume the model

$$y_i = \alpha + \beta x_i + e_i \qquad (4.1)$$

where α and β are unknown constants and e_i is a random variable such that $E(e_i|x_i) = 0$, $E(e_i^2|x_i) = ax_i^g$, $a \geq 0$, $g \geq 0$, and $E(e_i e_j|x_i, x_j) = 0$.

Theorem 2: Average variance of the corresponding H.T. estimator for any I.P.P.S. scheme under model (4.1) is

$$V^*(\hat{Y}_{H.T.}) = \alpha^2[\Sigma_i \frac{1}{\pi_i} + \Sigma_i \Sigma_{j(\neq i)} \frac{\pi_{ij}}{\pi_i \pi_j} - N^2] + ax^g(\frac{\Sigma p_t^{g-1}}{n} - \Sigma p_t^g) \qquad (4.2)$$

Proof: Taking the expectation of $V(\hat{Y}_{H.T.})$ under model 4.1 we get

$$V^*(\hat{Y}_{H.T.}) = \alpha^2[\Sigma_i \frac{1}{\pi_i} + \Sigma_i \Sigma_{j(\neq i)} \frac{\pi_{ij}}{\pi_i \pi_j} - N^2] + \alpha\beta[2\Sigma_i \frac{x_i}{\pi_i} + \Sigma_i \Sigma_{j(\neq i)} \frac{\pi_{ij}}{\pi_i \pi_j}(x_i + x_j) - 2Nx]$$
$$+ \beta^2[\Sigma_i \frac{x_i^2}{\pi_i} + \Sigma_i \Sigma_{j(\neq i)} \frac{\pi_{ij}}{\pi_i \pi_j} x_i x_j - x^2] + a[\Sigma_i \frac{x_i^g}{\pi_i} - \Sigma_i x_i^g],$$

which, upon using the relations $\pi_i = np_i$ and $\Sigma_{j(\neq i)} \pi_{ij} = (n-1)\pi_i$, reduces to (4.2).

Thus from (4.2) it follows that when $\alpha = 0$, the average variance of the corresponding H.T. estimator will be the same for all the I.P.P.S. schemes. However, if $\alpha \neq 0$, it can be observed from (4.2) that among all the I.P.P.S. schemes, the H.T. estimator corresponding to the scheme for which the value $\Sigma_i \Sigma_{j(\neq i)} \frac{\pi_{ij}}{\pi_i \pi_j}$ is least will have the least average variance. Thus a reasonable investigation will be to rank the various I.P.P.S. schemes according to the value of $\Sigma_i \Sigma_{j(\neq i)} \frac{\pi_{ij}}{\pi_i \pi_j}$ (= C, say). For this investigation we will confine to the case n = 2 only.

For the schemes of Durbin (1967), Yates and Grundy (1953), Durbin (1953), Goodman and Kish (1950) and for scheme A of Hanurav (1967) the approximate expressions for π_{ij} correct to $O(N^{-4})$ are respectively given by

$$\pi_{ij}^{(1)} = 2p_i p_j[1 + \{(p_i + p_j) - \Sigma p_t^2\} + \{2(p_i^2 + p_j^2) - 2\Sigma p_t^3 - (p_i + p_j)\Sigma p_t^2 + (\Sigma p_t^2)^2\}] \qquad (4.3)$$

$$\pi_{ij}^{(2)} = 2p_i p_j[1 + \{(p_i + p_j) - \Sigma p_t^2\} + \{2(p_i^2 + p_j^2) - 2\Sigma p_t^3 + \tfrac{3}{4}p_i p_j - \tfrac{7}{4}(p_i + p_j)\Sigma p_t^2 + \tfrac{7}{4}(\Sigma p_t^2)^2\}] \qquad (4.4)$$

$$\pi_{ij}^{(3)} = 2p_i p_j[1 + \{(p_i + p_j) - \Sigma p_t^2\} + \{2(p_i^2 + p_j^2) - 2\Sigma p_t^3 + p_i p_j - 2(p_i + p_j)\Sigma p_t^2 + 2(\Sigma p_t^2)^2\}] \qquad (4.5)$$

$$\pi_{ij}^{(4)} = 2p_i p_j[1 + \{(p_i + p_j - \Sigma p_t^2\} + \{2(p_i^2 + p_j^2) - 2\Sigma p_t^3 + 2p_i p_j - 3(p_i + p_j)\Sigma p_t^2 + 3(\Sigma p_t^2)^2\}] \qquad (4.6)$$

and

$$\pi_{ij}^{(5)} = 2p_i p_j[1 + \frac{p_i p_j}{\Sigma p_t^2} + \frac{p_i^3 p_j^3}{\Sigma p_t^2 \Sigma p_t^4}] \qquad (4.7)$$

Expressions (4.3) and (4.6) are from Asok and Sukhatme (1974) and expressions (4.4) and (4.5) are from Rao (1963). Using equations (4.3) to (4.7) and the relation $\pi_i = 2p_i$, the values of $\Sigma_i \Sigma_j \frac{\pi_{ij}}{\pi_i \pi_j}$ correct to $O(N^0)$ for the five schemes are respectively given by

$$C_1 = \frac{1}{2}[N(1 - N\Sigma p_t^2) + \{3N\Sigma p_t^2 + N^2(\Sigma p_t^2)^2 - 2N^2\Sigma p_t^3 - 2\}] \qquad (4.8)$$

$$C_2 = \frac{1}{2}[N^2 + N(1 - N\Sigma p_t^2) + \{\tfrac{3}{2}N\Sigma p_t^2 + \tfrac{7}{4}N^2(\Sigma p_t^2)^2 - 2N^2\Sigma p_t^3 - \tfrac{5}{4}\}] \qquad (4.9)$$

$$C_3 = \frac{1}{2}[N^2 + N(1 - N\Sigma p_t^2) + \{N\Sigma p_t^2 + 2N^2(\Sigma p_t^2)^2 - 2N^2\Sigma p_t^3 - 1\}] \qquad (4.10)$$

$$C_4 = \frac{1}{2}[N^2 + N(1 - N\Sigma p_t^2) - \{N\Sigma p_t^2 - 3N\Sigma(\Sigma p_t^2)^2 + 2N^2\Sigma p_t^3\}] \qquad (4.11)$$

and

$$C_5 = \frac{1}{2}[N^2 + \frac{1}{\Sigma p_t^2}(1-N\Sigma p_t^2)+\{\frac{(\Sigma p_t^3)^2}{\Sigma p_t^2 \cdot \Sigma p_t^4} - 1\}] \qquad (4.12)$$

It can be easily verified from (4.8) thru (4.11) that

$$C_1 \le C_2 \le C_3 \le C_4 \qquad (4.13)$$

which is also a direct consequence of the comparisons made by Rao (1963, 1965) of the above four schemes without any model assumptions.
From (4.11) and (4.12) we get

$$C_5-C_4 = \frac{1}{2}[\frac{1}{\Sigma p_t^2}(1-N\Sigma p_t^2)^2+\{\frac{(\Sigma p_t^3)^2}{\Sigma p_t^2 \cdot \Sigma p_t^4} - 3N^2(\Sigma p_t^2)^2 + N\Sigma p_t^2+2N^2\Sigma p_t^3-1\}] \qquad (4.14)$$

Now, assuming $p_1, p_2, \ldots, p_N$ to be having a specific distribution Δ with moments μ_r' we can replace Σp_t^r in (4.14) by $N\mu_r'$ because we have from Khintchine's law of large numbers

$$\operatorname*{plim}_{N\to\infty} m_r' = \operatorname*{plim}_{N\to\infty} \frac{1}{N}\Sigma p_t^r = \mu_r' . \qquad (4.15)$$

In view of the relation $\Sigma p_t = 1$, we however should have

$$\mu_1' = \frac{1}{N} \qquad (4.16)$$

In the following we will investigate the relative efficiency of Hanurav's scheme A in relation to the other procedures mentioned here under various distributions of p_t.

Case (i) - χ^2 distribution: When the p_ts are distributed as $\frac{1}{\nu N}\chi^2_{(\nu)}$ where $\chi^2_{(\nu)}$ is the chi-square variate with ν degrees of freedom, from the relation

$$\Sigma p_t^r = N\mu_r' \qquad (4.17)$$

we get

$$\Sigma p_t^2 = \frac{\nu+2}{\nu N} \qquad (4.18)$$

$$\Sigma p_t^3 = \frac{(\nu+2)(\nu+4)}{\nu^2 N^2} \qquad (4.19)$$

and

$$\Sigma p_t^4 = \frac{(\nu+2)(\nu+4)(\nu+6)}{\nu^3 N^3} . \qquad (4.20)$$

Substituting these values in (4.14) we get

$$C_5-C_4 = \frac{1}{2}[\frac{\nu N}{\nu+2}\{1 - \frac{\nu+2}{\nu}\}^2+\{\frac{\nu+4}{\nu+6} - \frac{3(\nu+2)^2}{\nu^2} + \frac{\nu+2}{\nu} + \frac{2(\nu+2)(\nu+4)}{\nu^2} - 1\}]$$

which after simplification reduces to

$$C_5-C_4 = \frac{2N}{\nu(\nu+2)} + \frac{4(2\nu+3)}{\nu^2(\nu+6)} \ge 0 \qquad (4.21)$$

Case (ii) - β distribution: When the p_t's follow a beta distribution of the first kind with parameters (α_1-1, α_2) where α_1 and α_2 are related by the equation

$$\mu_1' = \frac{\alpha_1}{\alpha_1+\alpha_2+1} = \frac{1}{N}$$

or

$$\alpha_2 = (N-1)\alpha_1 - 1 \qquad (4.22)$$

we get after substituting $N\mu_r'$ for Σp_t^r,

$$\Sigma p_t^2 = \frac{\alpha_1+1}{N\alpha_1+1} \qquad (4.23)$$

$$\Sigma p_t^3 = \frac{(\alpha_1+1)(\alpha_1+2)}{(N\alpha_1+1)(N\alpha_1+2)} \qquad (4.24)$$

and

$$\Sigma p_t^4 = \frac{(\alpha_1+1)(\alpha_1+2)(\alpha_1+3)}{(N\alpha_1+1)(N\alpha_1+2)(N\alpha_1+3)} . \qquad (4.25)$$

Substituting from (4.23) thru (4.25) in (4.14) we get

$$C_5-C_4 = \frac{1}{2(\alpha_1+1)(\alpha_1+3)(N\alpha_1+1)^2(N\alpha_1+2)}[N\alpha_1^3(N^3 + 2N^2 - 7N + 4) + \alpha_1^2(3N^4 + 4N^3 - 22N^2 + 14N + 1) + \alpha_1(12N^3 - 33N^2 + 18N + 3) - 6(N-1)] \qquad (4.26)$$

Now,

$$N^3+2N^2-7N+4=N(N^2-1)+2N(N-3)+4 > 0 \text{ for } N \ge 3 \qquad (4.27)$$

$$3N^4+4N^3-22N^2+14N+1=N^3(3N-7)+11N^2(N-2)+14N+1 > 0 \text{ for } N \ge 3 \qquad (4.28)$$

and

$$\alpha_1(12N^3 - 33N^2 + 18N + 3) - 6(N-1) = (\alpha_1-1)[3N^2(4N-11) + 3(6N+1)] + 3N^2(4N-11) + 3(4N+3) > 0 \quad \text{for } N \ge 3, \qquad (4.29)$$

because $\alpha_1 > 1$.
(4.27) thru (4.29) implies that

$$C_5 - C_4 > 0. \qquad (4.30)$$

Case (iii) - Uniform distribution: When the p_t's follow a uniform distribution over the interval $(0, \frac{2}{N})$, we get from $\Sigma p_t^r = N\mu_r'$,

$$\Sigma p_t^2 = \frac{4}{3N}, \ \Sigma p_t^3 = \frac{2}{N^2} \text{ and } \Sigma p_t^4 = \frac{16}{5N^3} .$$

Substitution of these values in (4.14) gives,

$$C_5 - C_4 = \frac{(4N - 3)}{96} > 0 \qquad (4.31)$$

In view of equations (4.13), (4.21), (4.30) and (4.31) it follows that when the variance is considered to $O(N^0)$, Hanurav's strategy would be inferior to those of Durbin (1967), Yates and Grundy (1953), Durbin (1953), and Goodman and Kish (1950) when the p_t's follow chi-square, beta or uniform distributions.

5. Numerical Illustration: The data presented in Table 1 is for the 20 districts in the Andhra Pradesh State of India. The x variable gives the populations (rounded off in thousands) in these districts as per the 1951 census and the y variable gives the exact population as per the 1961 census

Table 1

Population Figures for the 20 districts in the Andhra Pradesh State of India

i	X_i	Y_i
1	2123	2342291
2	2072	2288976
3	2301	2609311
4	1697	1978434
5	1736	2076103
6	2560	3009997
7	1794	2033963
8	1666	1913169
9	1628	1342140
10	1483	1764223
11	1617	1909644
12	1447	1590689
13	1821	2063601
14	1109	1226465
15	835	1021503
16	831	1009301
17	1428	1620417
18	1329	1545750
19	808	1057225
20	1287	1574797

In Table 2 are presented the variances correct to $O(N^2)$, $O(N^1)$ and $O(N^0)$ for the procedure of Goodman and Kish as well as Hanurav when samples of size 2 are considered. The variance correct to $O(N^2)$ for either of the procedures represents the true variance for the customary estimator in the case of probability proportional to size sampling with replacement. Values of the successive approximations indicate that the convergence is quite satisfactory inspite of the fact that the population size is much smaller than one usually encounters in practice. The relative difference between the two variances for larger sample sizes is however expected to be much higher than it is in this case.

Table 2

Approximations to $V(\hat{Y}_{H.T.})$

Order of approximation	Goodman and Kish procedure	Hanurav's procedure
$O(N^2)$	364525×10^7	364525×10^7
$O(N^1)$	347021×10^7	347068×10^7
$O(N^0)$	346217×10^7	346249×10^7

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. Durbin, J. (1953). Some results in sampling theory when the units are selected with unequal probabilities. Journal of the Royal Statistical Society, Series B, 15: 262-269.

3. Durbin, J. (1967). Design of multi-stage surveys for the estimation of sampling errors. Applied Statistics 16: 152-164.

4. Goodman, R. and Kish, L. (1950). Controlled selection - A technique in probability sampling. Journal of the American Statistical Association 45: 350-372.

5. Hanurav, T. V. (1966). An optimum πPS sampling. Annals of Mathematical Statistics 37: 1859.

6. Hanurav, T. V. (1967). Optimum utilization of auxiliary information: πPS sampling of two units from a stratum. Journal of the Royal Statistical Society, Series B, 29: 374-391.

7. Hartley, H. O. and Rao, J. N. K. (1962). Sampling with unequal probabilities and without replacement. Annals of Mathematical Statistics 33: 350-374.

8. Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. Journal of the American Statistical Association 47: 663-685.

9. Rao, J. N. K. (1963). On three procedures of unequal probability sampling without replacement. Journal of the American Statistical Association 58: 202-215.

10. Rao, J. N. K. (1965). On two simple schemes of unequal probability sampling without replacement. Journal of the Indian Statistical Association 3: 173-180.

11. Sampford, M. R. (1967). On sampling without replacement with unequal probabilities of selection. Biometrika 54:499-513.

12. Yates, F. and Grundy, P. M. (1953). Selection without replacement from within strata with probability proportional to size. Journal of the Royal Statistical Society, Series B, 15:253-261.