Statistica Sinica **6**(1996), 617-630

NEW PARTIALLY SYSTEMATIC SAMPLING

Ching-Ho Leu and Kam-Wah Tsui

National Cheng Kung University and University of Wisconsin-Madison

Abstract: When the sample mean, $\overline{y}$, is used to estimate the mean of a finite population, the usual systematic sampling procedure cannot provide an unbiased estimator of $V(\overline{y})$, the variance of $\overline{y}$. In this paper, we propose a new partially systematic sampling procedure which provides an unbiased estimator of $V(\overline{y})$. Moreover, the population size N is not required to be a multiple of the sample size used. We also compare the efficiency of this new sampling procedure with simple random sampling, the usual or circular systematic sampling procedures, and systematic sampling with multiple starts for populations exhibiting certain characteristics, such as "random", "linear", or "autocorrelated" trends, or periodic variation.

Key words and phrases: Partially systematic sampling, inclusion probability, Horvitz-Thompson estimator.

1. Introduction

The usual systematic sampling design, in its simplest and most commonly used form, selects every kth unit from a finite population of N units, assuming that the sample size n is N/k, for some integer k. This sampling design is attractive because of its simplicity and operational convenience. Moreover, when the sample mean is used to estimate the population mean of a characteristic of interest, using a systematic sampling design yields higher efficiency than using a simple random sampling design or a stratified random sampling design in many situations. The main disadvantage with the (single-start) systematic sampling design is that it cannot provide an unbiased estimator of variance of the sample mean. Some modifications of the usual systematic sampling design have therefore been proposed to overcome this problem. Gautschi (1957) considered a multiple-start systematic sampling design; Singh and Padam Singh (1977) suggested another modified systematic sampling design; Padam Singh and Garg (1979) proposed a balanced random sampling design; Zinger (1980) and Wu (1984) considered a partically systematic sampling design which selects a systematic sample of size n and then selects a supplementary sample using either a simple random sampling design or the usual systematic sampling design from the remaining $N-n$ units. Agrawel, Singh and Singh (1984) provided a modified systematic sampling scheme combining the concept of random interval with the use of unequal selection probabilities.

In this paper, we propose a new partially systematic sampling procedure. Our procedure maintains simplicity of selection and enables one to obtain an unbiased variance estimator of the sample mean. We also compare the efficiency of our proposed procedure to the simple random sampling procedure, the usual systematic sampling procedure, and in some cases, the systematic sampling procedure with multiple starts for certain types of finite populations.

2. New Partially Systematic Sampling

Consider a finite population $U = (U_1, \ldots, U_N)$ of size N from which a sample of size n is to be drawn. For each $i = 1, \ldots, N$, a real-valued y_i is associated with U_i. The problem of interest is to estimate the population mean $\bar{Y} = \sum_{i=1}^{N} y_i/N$.

Our sampling procedure depends on two parameters, k and a. Define $u = N - (n-a)k$. If N is a multiple of n, let $k = N/n$, $a = 2$; otherwise, let k be the integer closest to $N/(n-1)$, and let a be an integer satisfying $2 \leq a \leq [n/2] + 1$, where $[w]$ denotes the integer part of w. To select a sample of size n,
(a) first, select at random an index t from $1, \ldots, N$;
(b) using the convention that for any $i = 1, \ldots, N$, the unit with index $i + N$ stands for the unit with index i , select a simple random sample of a indices from the set $\{t, t+1, \ldots, t+u-1\}$, and denote by s'_t the set of a units of the finite population with the selected indices;
(c) let s''_t be the systematic sample of the finite population, with indices $t + (u-1) + lk$, $l = 1, \ldots, n-a$;
(d) our resulting sample s_t of size n is the union of s'_t and s''_t.

2.1. The choice of size a and interval k

It is desired to choose the parameters a, u and k in our sample selection procedure in such a way that
(R1) every sample contains n distinct units,
(R2) there is an even spread of sample units over the entire population,
(R3) the inclusion probability for every pair of units is non-zero.

Note that $N = u + (n-a)k$, there is only one traversal of the points on the circle, and hence the sample will always contain n distinct units. Theorem 1 below provides conditions ensuring that (R3) is satisfied.

Theorem 1. *Under our proposed sampling procedure, the inclusion probability for every pair of units is non-zero if*

$$\text{(a)} \quad a \geq 2 \qquad \textit{and} \qquad \text{(b)} \quad u \geq k. \tag{2.1}$$

Proof. Without loss of generality, we can consider the pair $\{U_1, U_j\}$ only, where $j > 1$ is an arbitrary index. Suppose $a \geq 2$. When the starting index $t = 1$

and $U_j \in s'_t$, the possible differences of $j-1$ are covered in $\{1, \ldots, u-1\}$. Recall that a unit with index $N+v$ for some v, $1 \leq v \leq N$, is identified as the unit with index v. When $U_j \in s''_t$ and the starting index t is such that $t+u-1=i$ is in $\{1, \ldots, u-1\}$, the possible differences of $j-1$ are then covered in $\{k+(i-1), 2k+(i-1), \ldots, (n-a)k+(i-1)\}$ where $i=1, \ldots, u-1$. Now as $u \geq k$, all possible differences from 1 up to $(n-a)k+(u-1)=N-1$ occur at least once. Consequently, the inclusion probability for every pair of units is non-zero.

For circular systematic sampling, Sengupta and Chattopadhyay (1987) and Bellhouse (1984) addressed the problem of the choosing k so that the sample satisfies the requirements (R1) and (R2). For a given size n, they suggested setting the sampling interval k equal to the integer nearest to N/n when $N \neq jk$ for some integer $j \leq (n-1)$. Otherwise, set k equal to the greatest integer $[N/n]$ in N/n. Combining these suggestions and the conditions in Theorem 1, and letting $k_1 = [N/(n-1)]$ and $k_2 = [N/n]+1$, we arrive at the following recommendations for the choice of parameters a and k (a detailed discussion is given in the Appendix):
(i) If $k_1 = 1$, choose $a = [n/2]$ and $k = 1$.
(ii) If $N = nk$, choose $k = N/n$ and $a = 2$. Otherwise, choose k as follows:
(iii) When $k_1 \geq 2$ and $k_1 \geq k_2$, choose $k = k_1$ and $a = 2$.
(iv) When $k_1 \geq 2$ and $k_1 < k_2$, let a be the smallest integer which satisfies $(a-1)k_2 \geq n$; if $a \geq k_2$, choose $k = k_1$; if $a < k_2$, then choose $k = k_1$ or $k = k_2$, depending on which one is close to u/a.

Note that our sampling procedure is the same as the modified systematic sampling procedure considered by Singh and Padam Singh (1977) when $a = u$. However, under their sampling design, the sample size n must satisfy $n \geq \sqrt{(2N+4)} - 1$. In contrast, with our sampling scheme and appropriate choices of a and k, this restriction becomes unnecessary.

2.2. Calculation of inclusion probabilities and estimation procedure

In this section, we derive the inclusion probabilities for individual as well as for pairwise units. For a given starting index t, and for $i = 1, \ldots, N$, the indicator variables c_{ti} are defined as

$$c_{ti} = \begin{cases} 1, & \text{if } U_i \in s_t; \\ 0, & \text{otherwise.} \end{cases}$$

Note that $\sum_{i=t}^{t+u-1} c_{ti} = a$ and $\sum_{i=1}^{N} c_{ti} = n$ by the construction of the sample s_t. The inclusion probabilities for individual units and pairwise units are given by

$$\pi_i = \frac{1}{N}\sum_{t=1}^{N} E[c_{ti}], \qquad \pi_{ij} = \frac{1}{N}\sum_{t=1}^{N} E[c_{ti}c_{tj}].$$

Theorem 2. *Under our proposed sampling procedure, $\pi_i = n/N$, $i = 1, \ldots, N$.*

Proof. For every sample s_t, there are a units in s'_t with indices in the set $I_t = \{t, t+1, \ldots, t+u-1\}$ and $(n-a)$ units in s''_t. For a fixed i, there are u possible values of t such that i is in I_t and $(n-a)$ values of t such that $U_i \in s''_t$. It follows that

$$\pi_i = \frac{1}{N}\Big[u \cdot \frac{a}{u} + (n-a) \cdot 1\Big] = n/N, \quad \text{for all } i = 1, \ldots, N. \tag{2.2}$$

To determine the inclusion probabilities for a pair of units (U_i, U_j), $i \neq j$, under our sampling procedure, we note that there are four possible cases:
(A) $(U_i, U_j) \in s'_t$,
(B) $(U_i, U_j) \in s''_t$,
(C) $U_i \in s'_t$ and $U_j \in s''_t$,
(D) $U_j \in s'_t$ and $U_i \in s''_t$.

Now for k and a chosen according to the recommendations in Section 2.1 above, there exists one positive integer m satisfying

$$(m-1)k + 1 \leq u \leq mk.$$

The units U_i and U_j are said to be at distance v if $|j - i| = v$. Lemma 1 below counts the number of pairs of units at various distances.

Lemma 1.
(i) *In case* (C), *for each fixed $l = 1, \ldots, (m-1)$, there are l pairs of units at distance $|j-i| = v$, where $v = lk, \ldots, (l+1)k-1$ or $v = N-lk, \ldots, N-(l-1)k-1$. Furthermore, for each fixed $l' = 0, 1, \ldots, (n-a-m)$, there are m pairs of units at distance $|j-i| = v'$, where $v' = (m+l')k, \ldots, (l'+1)k+u-1$, and there are $m-1$ pairs of units at distance $|j-i| = v''$, where $v'' = (l'+1)k+u, \ldots, (m+l'+1)k-1$.*
(ii) *In case* (D), *for each fixed $l = 1, \ldots, (m-1)$, there are l pairs of units at distance $|j-i| = v$, where $v = (l-1)k+1, \ldots, lk$ or $v = N-(l+1)k+1, \ldots, N-lk$. Moreover, for each fixed $l' = 0, 1, \ldots, (n-a-m)$, there are m pairs of units at distance $|j-i| = v'$, where $v' = (m+l'-1)k+1, \ldots, l'k+u$; and there are $m-1$ pairs of units at distance $|j-i| = v''$, where $v'' = l'k+u+1, \ldots, (m+l')k$.*

Proof. Without loss of generality, let j be greater than i. The possible differences of $j - i$ in case (C) are

$$\begin{array}{ccccc}
k & 2k & 3k & \cdots & (n-a)k \\
k+1 & 2k+1 & 3k+1 & \cdots & (n-a)k+1 \\
\vdots & \vdots & \vdots & & \vdots \\
k+u-1 & 2k+u-1 & 3k+u-1 & \cdots & (n-a)k+u-1.
\end{array}$$

Since $(m-1)k+1 \leq u \leq mk$, the number of pairs of units at distance $|j-i| = v$, for $v = k, \ldots, N-1$, reduces to the result (i). Note that $(n-a-l)k+u = N-lk$. On the other hand, the possible differences of $j-i$ in case (D) are

$$\begin{array}{cccccc}
1 & k+1 & 2k+1 & 3k+1 & \cdots & (n-a-1)k+1 \\
2 & k+2 & 2k+2 & 3k+2 & \cdots & (n-a-1)k+2 \\
\vdots & \vdots & \vdots & \vdots & & \vdots \\
u & k+u & 2k+u & 3k+u & \cdots & (n-a-1)k+u.
\end{array}$$

Since $(m-1)k+1 \leq u \leq mk$, the number of pairs of units at distance $|j-i|$ runs from $1, \ldots,$ to $N-k$, which is result (ii).

Theorem 3. *The inclusion probabilities for a pair of units* (U_i, U_j), $i \neq j$, *under our sampling procedure is given by*

(i) $N\pi_{ij} = \frac{a(a-1)}{u(u-1)} \max(0, (u-dk)) + [(n-a)-d] + \frac{a}{u}[\min(2d, 2m-1)]$, *if* $|i-j| = dk$ *or* $N-dk$, $d = 1, \ldots, (n-a)$,

(ii) $N\pi_{ij} = \frac{a(a-1)}{u(u-1)}(u-|i-j|) + \begin{cases} \frac{a}{u}(2l-1), & \textit{for } (l-1)k+1 \leq j-i \leq lk-1; \\ \frac{a}{u}(2m-1), & \textit{for } (m-1)k+1 \leq j-i \leq u-1, \end{cases}$

where $l = 1, \ldots, m-1$, *if* $|i-j| \leq u-1$, *excluding the cases already covered in* (i).

(iii) $$N\pi_{ij} = \frac{a(a-1)}{u(u-1)}[|i-j| - (N-u)] + \begin{cases} \frac{a}{u}(2l-1), & \textit{for } N-(l-1)k-1 \geq j-i \geq N-lk+1; \\ \frac{a}{u}(2m-1), & \textit{for } N-(m-1)k-1 \geq j-i \geq N-u+1, \end{cases}$$

where $l = 1, \ldots, m-1$, *if* $|i-j| \geq N-u+1$, *excluding the cases already covered in* (i).

(iv) $N\pi_{ij} = 2(m-1)\frac{a}{u}$, *if* $l'k+u+1 \leq |i-j| \leq (l'+m)k-1$, *where* $l' = 1, \ldots, (n-a-m)$, *excluding the cases already covered in* (i).

(v) $N\pi_{ij} = 2m\frac{a}{u}$, *if* $(l'+m)k+1 \leq |i-j| \leq (l'+1)k+u-1$, *where* $l' = 1, \ldots, (n-a-m-1)$, *excluding the cases already covered in* (i).

Proof. Without loss of generality, let j be greater than i. For a given starting index t, and given that i, j are in $\{t, t+1, \ldots, t+u-1\}$, the conditional probability that $(U_i, U_j) \in s'_t$ is $a(a-1)/[u(u-1)]$ and the conditional probability that $U_i \in s'_t$ is a/u. The conditional probability that $U_j \in s''_t$ is 0 or 1 depends on whether $j = lk+t+u-1$ or not, where $l = 1, 2, \ldots,$ or $(n-a)$. When $j-i = dk$, there are four possible cases (A),(B),(C),(D) as described preceding Lemma 1. In case (A), if $u \geq dk$, there will be $u-dk$ units at distance dk. Also, there are $(n-a)-1$ units at distance k, $(n-a)-2$ units at distance $2k, \ldots,$ $(n-a)-d$ units at distance $dk, \ldots,$ and 1 unit at distance $(n-a-1)k$ in case (B). Using Lemma 1 for the cases (C) and (D) and the fact that the selection is circular yields the result (i).

When $j - i \le u - 1$, excluding the case $j - i = dk$, there are three cases (A) (C) and (D) only. In case (A), there are $u - 1$ units at distance 1, $u - 2$ units at distance 2, $\ldots$, and 1 units at distance $u - 1$. Together with Lemma 1, result (ii) follows. The proof for result (iii) follows from result (ii) directly, because of the circular nature of the selection.

In the remaining cases, we need only consider (C) and (D). Using Lemma 1 again completes the proof of the theorem.

To estimate the population mean $\bar{Y} = \sum_{i=1}^{N} y_i/N$, we use the Horvitz-Thompson estimator, which is the usual sample mean by Theorem 2:

$$\hat{\bar{Y}} = \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y}. \tag{2.3}$$

The variance of the Horvitz-Thompson estimator, and the corresponding variance estimator are given by Yates-Grundy (1953), respectively:

$$V(\bar{y}_n) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j>i}^{N}\Big(1 - \frac{N^2}{n^2}\pi_{ij}\Big)(y_i - y_j)^2 \tag{2.4}$$

$$\hat{V}(\bar{y}_n) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j>i}^{N} c_{ti}c_{tj}\Big(\frac{1}{\pi_{ij}} - \frac{N^2}{n^2}\Big)(y_i - y_j)^2, \tag{2.5}$$

where the c_{ti}'s are indicator variables defined in the beginning of Section 2.2.

3. Efficiency of New Partially Systematic Sampling Procedure

Given starting index t, the following notation is used:

$\bar{y}_{at} =$ the mean of the y_i's associated with the simple random sample of size a from the subpopulation of units with indices $\{t, t+1, \ldots, t+u-1\}$ of size u,

$\bar{y}_{st} =$ the sample mean of the y_i's associated with the units in s''_t,

$\bar{y}_{\cdot t} = \sum_{i=0}^{u-1} y_{it}/u$, where $y_{it} = y_{t+i}$,

$\bar{y}_{nps} = [a\bar{y}_{at} + (n-a)\bar{y}_{st}]/n =$ new partially systematic (nps) sample mean.

The variance of the new partially systematic sample mean can be written as

$$V(\bar{y}_{nps}) = \frac{1}{N}\frac{a}{n^2}\Big(1 - \frac{a}{u}\Big)\sum_{t=1}^{N}\frac{\sum_{i=0}^{u-1}(y_{it} - \bar{y}_{\cdot t})^2}{u-1} + \frac{1}{N}\sum_{t=1}^{N}\Big(\frac{a\bar{y}_{\cdot t} + (n-a)\bar{y}_{st}}{n} - \bar{Y}\Big)^2. \tag{3.1}$$

The performance of the new partially systematic sampling in relation to that of systematic or simple random sampling depends on the properties of the finite population. Thus the relative efficiencies of these sampling procedures are compared in this section for various types of superpopulations. We regard the values of superpopulation y_i's as drawn from an infinite superpopulation in which the expectation is denoted by symbol $\mathcal{E}$.

3.1. Population in random order

If the variates y_i $(i = 1, \ldots, N)$ are drawn at random from a superpopulation in which

$$\mathcal{E}(y_i) = \mu, \qquad \mathcal{E}(y_i - \mu)^2 = \sigma_i^2$$
$$\mathcal{E}(y_i - \mu)(y_j - \mu) = 0 \quad (i \neq j),$$

it is known as a population in random order (Cochran (1977)). Since

$$\begin{aligned} V(\bar{y}_{nps}) = \frac{1}{N}\Big(\frac{1}{n} - \frac{1}{N}\Big)\sum_{i=1}^{N}(y_i - \mu)^2 + \frac{2}{Nn^2}\bigg\{ & \frac{a(a-1)}{u(u-1)}\sum_{t=1}^{N}[\sum_{i=0}^{u-1}\sum_{j>i}^{u-1}(y_{it} - \mu)(y_{jt} - \mu)] \\ & + \frac{a}{u}\sum_{t=1}^{N}[\sum_{i=0}^{u-1}\sum_{j=1}^{n-a}(y_{it} - \mu)(y_{jk+t+u-1} - \mu)] \\ & + \sum_{t=1}^{N}[\sum_{i=1}^{n-a}\sum_{j>i}^{n-a}(y_{ik+t+u-1} - \mu)(y_{jk+t+u-1} - \mu)]\bigg\} \\ & - \frac{2}{N^2}\sum_{i=1}^{N}\sum_{j>i}^{N}(y_i - \mu)(y_j - \mu). \end{aligned} \tag{3.2}$$

Taking expectation over the superpopulation model, we get the expected variance

$$\sigma_{nps}^2 = \frac{1}{N}\Big(\frac{1}{n} - \frac{1}{N}\Big)\sum_{i=1}^{N}\sigma_i^2. \tag{3.3}$$

The expected variance for the usual or circular systematic sample mean and simple random sample mean are also known to be equal to (3.3). Hence, the three sampling procedures under consideration are equally efficient for populations with y_i's in random order. In general, under the superpopulation model considered in this section, Equation (3.3) holds for any sampling design with fixed sample size n and inclusion probability n/N for all units.

Rao (1975) and Rao and Bellhouse (1978) gave a random permutation model to represent populations in random order. They assume that the finite population consists of N fixed numbers $z_1, \ldots, z_N$ and that the measurements $y_1, \ldots, y_N$ are obtained as a random permutation of $z_1, \ldots, z_N$. Hence, $P(y_u = z_1) = 1/N$ and $P(y_u = z_i, y_v = z_j) = 1/[N(N-1)]$. The random permutation model is formalized as the linear model in which

$$y_i = \bar{Y} + e_i \tag{3.4}$$

$$E_m(e_i) = 0, \quad E_m(e_i^2) = \sigma^2, \quad E_m(e_ie_j) = -\sigma^2/(N-1), \quad (i \neq j),$$

where the operator E_m denotes expectation with respect to the model,

$$\bar{Y} = \sum_{i=1}^{N} y_i/N = \sum_{i=1}^{N} z_i/N,$$

and

$$\sigma^2 = \sum_{i=1}^{N}(y_i - \bar{Y})^2/N = \sum_{i=1}^{N}(z_i - \bar{Z})^2/N.$$

Then it is shown that the variances for the usual systematic sample mean and simple random sample mean averaged over model (3.4) are equal to $(k-1)\sigma^2/(N-1) = (k-1)S^2/(nk)$, where $(N-1)S^2 = \sum_{i=1}^{N}(y_i - \bar{Y})^2$ and $N = nk$. This result also follows under our proposed sampling procedure. Note that $V(\bar{y}_{nps})$ can be written as in (3.2) except that $\bar{Y}$ replaces μ there. Straightforward algebra then shows that $V(\bar{y}_{nps})$ averaged over (3.4) is $(k-1)\sigma^2/(N-1)$. We conclude that the three sampling procedures under consideration are also equally efficient under the random permutation model. More generally, with this random permutation model, it may be shown that the variance of the sample mean, averaged over the model, is the same for all fixed-size designs with inclusion probability n/N for all units.

3.2. Autocorrelated population

We assume that the observations y_i $(i = 1, \ldots, N)$ are drawn from a superpopulation in which

$$\mathcal{E}(y_i) = \mu, \qquad \mathcal{E}(y_i - \mu)^2 = \sigma^2$$

and

$$\mathcal{E}(y_i - \mu)(y_j - \mu) = \rho_{|j-i|}\sigma^2 \quad (i \neq j).$$

For the case $N = nk$, Cochran (1946) obtained the expected variance for the usual systematic sample mean and the simple random sample mean as, respectively,

$$\begin{aligned}\sigma^2_{sys} &= \frac{N-1}{N}\sigma^2\Big[1 - \frac{2}{N(N-1)}\sum_{d=1}^{N-1}(N-d)\rho_d\Big] \\ &\quad -\frac{n-1}{n}\sigma^2\Big[1 - \frac{2}{n(n-1)}\sum_{d=1}^{n-1}(n-d)\rho_{dk}\Big], \qquad (3.5)\end{aligned}$$

and

$$\sigma^2_{srs} = \Big(1 - \frac{n}{N}\Big)\frac{\sigma^2}{n}\Big[1 - \frac{2}{N(N-1)}\sum_{d=1}^{N-1}(N-d)\rho_d\Big].$$

When circular systematic sampling is used, the expected variance of the sample mean is shown to be

$$\begin{aligned}\sigma^2_{css} \;=\; & \Big(\frac{1}{n}-\frac{1}{N}\Big)\sigma^2+\frac{2}{Nn^2}\sigma^2+\sum_{t=1}^{N}\Big[\sum_{i=0}^{n-1}\sum_{j>i}^{n-1}\rho_{|(ik+t)-(jk+t)|}\Big] \\ & -\frac{2}{N^2}\sigma^2\sum_{d=1}^{N-1}(N-d)\rho_d. \qquad (3.6)\end{aligned}$$

The expected variance of the sample mean using the new partially systematic sampling procedure in this case is

$$\begin{aligned}\sigma^2_{nps} \;=\; & \Big(\frac{1}{n}-\frac{1}{N}\Big)\sigma^2+\frac{2}{Nn^2}\sigma^2\Big\{\frac{a(a-1)}{u(u-1)}\sum_{t=1}^{N}[\sum_{i=0}^{u-1}\sum_{j>i}^{u-1}\rho_{|(t+j)-(t+i)|}] \\ & +\frac{a}{u}\sum_{t=1}^{N}[\sum_{i=0}^{u-1}\sum_{j=1}^{n-a}\rho_{|(t+i)-(jk+t+u-1)|}]+\sum_{t=1}^{N}[\sum_{i=0}^{n-a}\sum_{j>i}^{n-a}\rho_{|(ik+t+u-1)-(jk+t+u-1)|}]\Big\} \\ & -\frac{2}{N^2}\sigma^2\sum_{d=1}^{N-1}(N-d)\rho_d. \qquad (3.7)\end{aligned}$$

When $N = nk$ and $n = ml$, instead of choosing only one random start, we can select a simple random sample of size l (without replacement) from the first kl elements and then every klth element following those selected. We call the resulting sample a systematic sample with multiple (l) random starts (msy). Gautschi (1957) showed that the expected variance of the sample mean for such a sample is

$$\sigma^2_{msy}=\frac{k-1}{N}\sigma^2\Big[1-\frac{2}{N(kl-1)}\sum_{d=1}^{N-1}(N-d)\rho_d-\frac{2kl}{m(kl-1)}\sum_{d=1}^{m-1}(m-d)\rho_{kld}\Big]. \qquad (3.8)$$

From expressions (3.5) through (3.8), it is difficult to provide general result about the relative efficiency of the sampling procedures under consideration. However, comparisons can be made for some types of correlograms considered by Cochran (1946), such as
(i) linear correlogram : $\rho_d = 1 - d/L$, $L \geq N - 1$,
(ii) exponential correlogram : $\rho_d = e^{-\lambda d}$,
(iii) hyperbolic correlogram : $\rho_d = tanh(d^{-3/5})$.

The results are presented in Table 1 and Table 2.

From Tables 1 and 2, we see that our partially systematic sampling procedure is more efficient than the simple random sampling procedure for these special types of finite population but is less efficient than the usual or circular systematic sampling procedures. This can be viewed as a tradeoff for being able to obtain an unbiased estimator of the variance of the sample mean in using our partially systematic sampling procedure instead of using the usual or circular systematic sampling procedure. Compared with the multiple-starts systematic sampling, we see that $\sigma^2_{msy} > \sigma^2_{nps}$ under exponential correlogram. For the case of hyperbolic correlogram, $\sigma^2_{msy} > \sigma^2_{nps}$ except when $n = 4$, and for linear correlogram $\sigma^2_{msy} > \sigma^2_{nps}$ except when $l = 2$ and $n = 4$ or 6. Hence the performance of our proposed sampling procedure is better than the systematic sampling procedure with multiple starts in most of the cases.

When $N = nk$ and $\rho_d = \rho^d$, let $a = 2$ and $u = 2k$, as suggested in the Section 2.1. Then Equation (3.5) reduces to

$$\sigma^2_{sys} = \Big(\frac{1}{n} - \frac{1}{N}\Big)\sigma^2 - \frac{2\sigma^2}{n^2}\Big[\frac{n\rho^k}{1-\rho^k} - \frac{\rho^k(1-\rho^{nk})}{(1-\rho^k)^2}\Big] - \frac{2\sigma^2}{N^2}\Big[\frac{N\rho}{1-\rho} - \frac{\rho(1-\rho^N)}{(1-\rho)^2}\Big]$$

and (3.7) reduces to

$$\sigma^2_{nps} = \sigma^2_{sys} + O(n^{-2}).$$

That is, the difference of the expected variances corresponding to the usual systematic sampling and our proposed sampling procedure is of small order $O(n^{-2})$ in this case.

Table 1. Variances of the sample mean corresponding to three sampling procedures for various correlograms where *css*, *srs*, and *nps* denote respectively, circular systematic sampling, simple random sampling and our new partially systematic sampling.

				Linear			Exponential			Hyperbolic	
N	n	u, a	σ^2_{css}	σ^2_{srs}	σ^2_{nps}	σ^2_{css}	σ^2_{srs}	σ^2_{nps}	σ^2_{css}	σ^2_{srs}	σ^2_{nps}
10	4	4, 2	0.0300	0.0550	0.0392	0.1096	0.0337	0.1169	0.0496	0.0763	0.0586
10	5	4, 2	0.0100	0.0367	0.0191	0.0501	0.0891	0.0655	0.0158	0.0508	0.0288
15	5	6, 2	0.0119	0.0474	0.0247	0.0805	0.1234	0.0990	0.0295	0.0770	0.0483
15	7	9, 4	0.0068	0.0271	0.0177	0.0406	0.0705	0.0577	0.0141	0.0440	0.0307
25	5	10, 2	0.0128	0.0555	0.0288	0.1186	0.1527	0.1344	0.0506	0.1052	0.0734
25	8	10, 3	0.0052	0.0295	0.0123	0.0518	0.0811	0.0626	0.0192	0.0559	0.0314
25	12	11, 5	0.0023	0.0150	0.0063	0.0219	0.0413	0.0296	0.0071	0.0285	0.0148
35	5	14, 2	0.0131	0.0588	0.0306	0.1400	0.1658	0.1522	0.0658	0.1207	0.0893
35	8	5, 2	0.0073	0.0331	0.0120	0.0683	0.0933	0.0786	0.0298	0.0679	0.0407
35	12	5, 2	0.0023	0.0187	0.0032	0.0322	0.0530	0.0352	0.0114	0.0386	0.0142
35	17	15, 7	0.0012	0.0104	0.0039	0.0150	0.0293	0.0206	0.0046	0.0213	0.0106

Table 2. Variances of the sample mean corresponding to three sampling procedures for various correlograms with $a = 2$, where msy, srs, and nps denote respectively, systematic sampling with multiple (l) random starts, simple random sampling and our new partially systematic sampling.

			Linear			Exponential			Hyperbolic		
N	n	u, l	σ^2_{msy}	σ^2_{srs}	σ^2_{nps}	σ^2_{msy}	σ^2_{srs}	σ^2_{nps}	σ^2_{msy}	σ^2_{srs}	σ^2_{nps}
16	4	8, 2	0.0352	0.0664	0.0449	0.1595	0.1744	0.1561	0.0818	0.1103	0.0852
16	8	4, 2	0.0065	0.0221	0.0065	0.0430	0.0581	0.0365	0.0175	0.0368	0.0142
24	4	12, 2	0.0376	0.0723	0.0492	0.1877	0.1985	0.1854	0.1068	0.1357	0.1103
24	6	8, 2	0.0156	0.0434	0.0179	0.1056	0.1191	0.0972	0.0529	0.0814	0.0486
24	6	8, 3	0.0226	0.0434	0.0179	0.1126	0.1191	0.0972	0.0641	0.0814	0.0486
24	8	6, 2	0.0081	0.0289	0.0083	0.0656	0.0794	0.0568	0.0298	0.0543	0.0249
24	8	6, 4	0.0151	0.0289	0.0083	0.0751	0.0794	0.0568	0.0427	0.0543	0.0249
24	12	4, 2	0.0029	0.0145	0.0026	0.0283	0.0397	0.0227	0.0112	0.0271	0.0080
36	4	18, 2	0.0391	0.0761	0.0519	0.2077	0.2152	0.2065	0.1299	0.1574	0.1332
36	6	12, 2	0.0167	0.0496	0.0196	0.1284	0.1345	0.1184	0.0696	0.0984	0.0648
36	6	12, 3	0.0244	0.0496	0.0196	0.1298	0.1345	0.1184	0.0812	0.0984	0.0648
36	9	8, 3	0.0100	0.0285	0.0069	0.0749	0.0807	0.0617	0.0418	0.0590	0.0284
36	12	6, 2	0.0036	0.0190	0.0032	0.0435	0.0538	0.0359	0.0194	0.0393	0.0145
36	12	6, 3	0.0051	0.0190	0.0032	0.0478	0.0538	0.0359	0.0244	0.0393	0.0145
36	12	6, 4	0.0067	0.0190	0.0032	0.0499	0.0538	0.0359	0.0278	0.0393	0.0145
36	18	4, 2	0.0013	0.0095	0.0010	0.0187	0.0269	0.0144	0.0072	0.0197	0.0046

3.3. Population in linear trend

Suppose the finite population is such that the y_i's satisfy the relationship $y_i = \alpha + \beta i$ for some constant α and β. The variance of the sample mean under the usual systematic sampling and simple random sampling procedure in this case can be reduced to a simple form (Cochran (1977)). When $N = nk$,

$$\begin{aligned} V(\bar{y}_{srs}) &= \beta^2 \frac{(k-1)(N+1)}{12}, \\ V(\bar{y}_{sys}) &= \beta^2 \frac{(k^2-1)}{12}, \end{aligned}$$

and the variance of our partially systematic sampling procedure is

$$V(\bar{y}_{nps}) = \beta^2 \frac{(k^2-1)}{12} + \frac{\beta^2}{3n^2}(k-1)[(n-2)(3k-1)+k]. \tag{3.9}$$

Thus the variance under the systematic sampling is smaller than that under our proposed sampling procedure and the difference of the two variances is $O(n^{-1})$. Note that $V(\bar{y}_{srs}) = V(\bar{y}_{nps})$ when $n = 2$, and if $n > 2$ then $V(\bar{y}_{srs}) > V(\bar{y}_{nps})$ in this case.

When $N = nk$ and $n = ml$, Gautschi (1957) showed that the variance of the systematic sampling with multiple (l) random starts under this population is

$$V(\bar{y}_{msy}) = \beta^2 \frac{(k-1)(kl-1)}{12}. \tag{3.10}$$

From (3.10) and (3.9), replacing n by ml, we have $V(\bar{y}_{msy}) = V(\bar{y}_{nps})$ if $l = 2$ and $m = 1$. Moreover, $V(\bar{y}_{msy}) > V(\bar{y}_{nps})$ if (1) $l = 2$ and $m \geq 5$ or (2) $l \geq 3$ and $m \geq 2$. Since $l \geq 2$, our proposed sampling procedure is better than the systematic sampling with multiple random starts for most of the cases in the population with linear trend.

3.4. Population with periodic variation

It is well known that the effectiveness of systematic sampling depends on the value of k when the population is periodic. The worst case is when k is equal to the period of the population, or is an integral multiple of the period; the systematic sample is no more precise than a single observation taken from the population. The best case occurs when k is an odd multiple of the half-period resulting in a zero sampling variance of the sample mean.

In our new sampling procedure, we select a simple random sample of size $a \geq 2$ from the set $\{t, t+1, \ldots, t+u-1\}$, which is more effective than the worst case of systematic sampling. If k is an odd multiple of the half-period and the observations of the simple random sample of size a are the same as that of the systematic sample then the sampling variance of the mean is zero also. Thus, for periodic populations the range of the variance of the sample mean under our partially systematic sampling is smaller than that under usual systematic sampling, while the lower limit of the variance is the same in the two sampling procedures. Therefore, our proposed sampling procedure can be considered as more efficient than systematic sampling for populations with periodic variation.

Acknowledgements

The authors wish to thank two referees for their helpful comments and suggestions.

Appendix

If $k_1 = [N/(n-1)]$, there exists an integer b such that

$$(n-1)k_1 \leq N = k_1(n-1) + b < (n-1)(k_1+1), \tag{A.1}$$

where $0 \le b < n-1$. This implies

$$(a-1)k_1 \le N-(n-a)k_1 < (a-1)k_1+n-1. \tag{A.2}$$

If $k_2 = [N/n]+1$, there exists an integer c such that

$$n(k_2-1) \le N = n(k_2-1)+c < nk_2, \tag{A.3}$$

where $0 \le c < n$. This implies

$$ak_2-n \le N-(n-a)k_2 < ak_2. \tag{A.4}$$

Case I. If $k_1 \ge k_2$, let $k = k_1 \ge 2$. Then, from (A.2) and (A.3),

$$(a-1)k \le N-(n-a)k = u < nk_2-(n-a)k \le nk_2-(n-a)k_2 = ak_2 \le ak. \tag{A.5}$$

We, therefore, choose a sample of size a from u units to satisfy the requirement (R2). As $a \ge 2$, we have

$$(a-1)k \ge 2(a-1) = a+(a-2) \ge a$$

$$(a-1)k \ge k. \tag{A.6}$$

This implies $u \ge a$ and $u \ge k$ which satisfies condition (2.1) in Theorem 1. For simplicity, we recommend to choose $a = 2$.

Case II. If $k_2 > k_1$, there exists a positive integer d such that $k_2 = k_1+d$. Thus, from (A.1) and (A.3), $N = k_1(n-1)+b = n(k_1+d-1)+c$, which implies $-k_1+b = n(d-1)+c \ge 0$. Hence $n-1 > b \ge k_1 > 0$. Moreover, $nd = n-k_1+b-c \le n+(n-1) < 2n$. Therefore, $d = 1$. If we choose $k = k_2$, then $ak-n \le u = N-(n-a)k < (a-1)k$. The length of this interval is $n-k_2$. From Theorem 1, we need $ak-n \ge k$, i.e. $(a-1)k \ge n$ and $a < k_2$, as these conditions imply $u \ge k$ and $u > a$ which satisfy Equation (2.1). Hence, we may choose a_1 to be the smallest integer so that $(a_1-1)k_2 \ge n$. If we choose $k = k_1$, then $ak \le u < (a-1)k+(n-1)$. The length of this interval is, again, $n-k-1 = n-k_2$. Since $[N/(n-1)] = [N/n]$ implies that n is large and that k is small relative to N, the range of the value of u varies as N or n vary. In this case, we offer the following guidance for the choice of a and k. When $k = k_1$, let the upper bound of u be not greater than $2(a-1)k$, and choose a_2 as the smallest integer which satisfies $(a_2-1)k_2 \ge n-1$. However, a_2 may be greater than a_1, and we desire the value of a to be as small as possible. Hence, $a = a_1$. If $a \ge k_2$, we choose $k = k_1$ because condition (A.6) and condition (2.1) still hold in this situation. If $a < k_2$, we choose $k = k_1$ or $k = k_2$, depending on which one is closer to u/a in order to satisfy requirement (R2). Note that, when $k_2 > k_1 \ge 2$, the upper bound of a is $[n/2]+1$ because $([n/2]+1-1)\cdot k_2 \ge n$ for all $n \ge 2$.

References

Agrawel, R., Singh, D. and Singh, P. (1984). Systematic sampling using varying probabilities. *J. Indian Soc. Agricultural Statist.* **36**, 99-109.

Bellhouse, D. R. (1984). On the choice of the sampling interval in circular systematic sampling. *Sankhyā Ser.B* **46**, 247-248.

Cochran, W. G. (1946). Relative accuracy of systematic and stratified random samples for a certain class of populations. *Ann. Math. Statist.* **17**, 164-177.

Cochran, W.G. (1977). *Sampling Techniques*, 3rd edition. John Wiley, New York.

Gautschi, W. (1957). Some remarks on systematic sampling. *Ann. Math. Statist.* **28**, 385-394.

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.

Murthy, M. N. and Rao, T. J. (1988). Systematic sampling with illustrative examples. *Handbook of Statistics* **6**, 147-185.

Padam Singh and Garg, J. N. (1979). On balanced random sampling. *Sankhyā Ser.C* **41**, 60-68.

Rao, J. N. K. (1975). On the foundations of survey sampling. In: *A Survey of Statistical Design and Linear Models* (Edited by J. N. Srivastava), 489-506, North-Holland, Amsterda.

Rao, J. N. K. and Bellhouse, D. R. (1978). Optimal estimation of a finite population mean under generalized random permutation models. *J. Statist. Plann. Inference* **2**, 125-141.

Sengupta, S. and Chattopadhyay, S. (1987). A note on circular systematic sampling. *Sankhyā Ser.B* **49**, 186-187.

Singh D. and Padam Singh (1977). New systematic sampling. *J. Statist. Plann. Inference* **1**, 163-178.

Wu, C. F. J. (1984). Estimation in systematic sampling with supplementary observations. *Sankhyā Ser.B* **46**, 306-315.

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**, 253-261.

Zinger, A. (1980). Variance estimation in partially systematic sampling. *J. Amer. Statist. Assoc.* **75**, 206-211.

Department of Statistics, National Cheng-Kung University, Tainan 70101, Taiwan.

Department of Statistics, University of Wisconsin, Madison, WI 53706, U.S.A.

(Received April 1991; accepted September 1995)