Remainder Linear Systematic Sampling

Author(s): Horng-Jinh Chang and Kuo-Chung Huang

REMAINDER LINEAR SYSTEMATIC SAMPLING

By HORNG-JINH CHANG

and

KUO-CHUNG HUANG

Tamkang University, Taipei

SUMMARY. In this paper, a slight modification of linear systematic sampling procedure that maintains the simplicity of selection is proposed. The procedure depends on the remainder only and can be used for population size being not a multiple of sample size. When the remainder is zero, the suggested method reduces to the usual linear systematic sampling procedure. Moreover, we also compare the efficiency of the proposed sampling design with simple random sampling, circular systematic sampling and new partially systematic sampling for various types of populations.

1. Introduction

Linear systematic sampling design is very simple and operationally convenient in practice. but, when the population size is not a multiple of the sample size, results in a variable sample size and in this case sample mean as an estimator of the population mean is not unbiased. To overcome these difficulties and in some cases to increase the efficiency. a number of modifications of linear systematic sampling design have been proposed, such as centered systematic sampling, balanced systematic sampling and modified systematic sampling etc.. In particular, Mahalanobis (1946), Deming (1950), Gautschi (1957) and Shiue (1960) considered a multiple-start systematic sampling scheme; Murthy and Rao (1988) discussed the circular systematic sampling method. Singh and Padam Singh (1977) proposed a method termed new systematic sampling; Padam Singh and Garg (1979) suggested a balanced random sampling design; Leu and Tsui (1996) considered a new partially systematic sampling. Note that all of the modified designs above provided a fixed sample size and an unbiased estimator of the population mean.

In the present study, a new sampling scheme called remainder linear systematic sampling is established to extend the linear systematic sampling for the case in which population size is not a multiple of sample size. The efficiency of the proposed procedure has been compared with simple random sampling and in some cases, circular systematic sampling and new partially systematic sampling for various types of populations.

---

2.   Remainder Linear Systematic Sampling

Consider a finite population $U = (U_1, U_2, \ldots, U_N)$ consisting of N units from which a sample of size n is to be drawn. The corresponding values obtained for any specific characteristic in the N units that comprise the population are denoted by $y_1, y_2, \ldots, y_N$, where y_i is associated with U_i. Here, the interest is to estimate the population mean $\overline{Y} = \sum_{i=1}^{N} y_i / N$.

The sample interval is given by $k = N/n$, if N is an integer multiple of the sample size n with $N = nk$. In case population size is not a multiple of the sample size, the sample interval $k = INT(N/n)$ is the truncated integer. The general form of the population size can be represented as

$$N = nk + r \tag{2.1}$$

where $0 \leq r < n$ and N, n, k, r are integers.

Divide the population into two strata such that the first stratum contains the front $(n - r)k$ units and the second stratum contains the hind $r(k + 1)$ units. For the first stratum, choose every k continuous units as a group. We therefore have $(n - r)$ groups with k units each. For the second stratum, choose every $(k + 1)$ continuous units as a group, the we have r groups with $(k + 1)$ units each. The proposed sampling procedure is as follows.

(a) Select a random start k_1 from 1 to k, and every k-th number thereafter from the first $(n - r)$ groups as the first stratum sample. That is, the k_1-th unit of each group of the first stratum is in the sample. This sample set is denoted by s' with indices $kl + k_1, l = 0, 1, \ldots, (n - r) - 1$

(b) Select a random start k_2 from 1 to $(k + 1)$, starting with the $[(n - r)k + k_2]$-th unit, and every $(k + 1)$-th number thereafter from the r groups as the second stratum sample. That is, the k_2-th unit of each group of the second stratum is in the sample. This sample set is denoted by s'' with indices

$$(k + 1)l' + (n - r)k + k_2, \quad l' = 0, 1, \ldots, r - 1.$$

(c) Our resulting sample s of size n is the combination of s' and s''.

Note that, if the population size is an integer multiple of the sample size, the sampling procedure reduces to the usual linear systematic sampling.

Using the remainder linear systematic sampling, the Horvitz-Thompson estimator can be used for the estimation of population mean. Under the suggested sampling procedure, the first order inclusion probability for the unit U_i is given by

$$
\begin{aligned}
\pi_i, \quad &= \quad \frac{1}{k}, & \text{if } 1 \leq i \leq (n - r)k. \\
&= \quad \frac{1}{k + 1}, & \text{if } (n - r)k + 1 \leq i \leq N.
\end{aligned}
$$

Moreover, the second order inclusion probabilities for a pair of units U_i and U_j, where $i \neq j$, are

$$
\begin{aligned}
\pi_{ij} &= \frac{1}{k}, & &\text{if } U_i \in s', U_j \in s'. \\
&= \frac{1}{k+1}, & &\text{if } U_i \in s'', U_j \in s''. \\
&= \frac{1}{k(k+1)} & &\text{if } U_i \in s', U_j \in s''. \\
&= 0, & &\text{otherwise.}
\end{aligned}
$$

It is easy to verify that the estimator of the population mean is unbiased under the suggested sampling design since the Horvitz-Thompson estimator is unbiased. For obtaining the variance of remainder linear systematic sampling, we use the following notations:

$\overline{y}_{1i}$: is the i-th sample mean of the first stratum,
$\overline{y}_{2i}$: is the i-th sample mean of the second stratum,
$\overline{Y}_1$: is the population mean of the first stratum,
$\overline{Y}_2$: is the population mean of the second stratum.

We then have

THEOREM 1. *The variance of the Horvitz-Thompson estimator under the remainder linear systematic sampling design is*

$$
V_{RLSS}(\overline{y}_{HT}) = \frac{1}{N^2} \left\{ (n-r)^2 k^2 \left[\frac{1}{k} \sum_{i=1}^{k} (\overline{y}_{1i} - \overline{Y}_1)^2 \right] \right.
$$

$$
\left. + \quad r^2 (k+1)^2 \left[\frac{1}{k+1} \sum_{i=1}^{k+1} (\overline{y}_{2i} - \overline{Y}_2)^2 \right] \right\}. \tag{2.2}
$$

On the basis of the single inclusion probabilities of remainder systematic sampling, the estimator of the finite population mean can be obtained as

$$
\overline{y}_{HT} = \frac{(n-r)k\overline{y}_{1i} + r(k+1)\overline{y}_{2i}}{N}. \tag{2.3}
$$

In addition, since the random starts k_1 and k_2 are chosen independently and the sample is drawn in a systematic manner, Theorem 1 is trivial.

Theorem 1 means that the variance of remainder linear systematic sampling procedure is the sum of the weighted variance of each stratum where the weight is the square of the population proportion of each stratum.

When s is the sample drawn, the corresponding variance estimator is

$$
\widehat{V}_{RLSS}(\overline{y}_{HT}) = \frac{1}{n^2} \left[\sum_{U_i \in s} \left(\frac{1 - \pi_i}{\pi_i^2} \right) y_i^2 + \sum_{\substack{U_i, U_j \in s \\ i \neq j}} \left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j \pi_{ij}} \right) y_i y_j \right]. \tag{2.4}
$$

Notice that the estimator $\widehat{V}_{RLSS}(\overline{y}_{HT})$ is not unbiased because some joint inclusion probabilities π_{ij} will be zero.

3.   Efficiency of Remainder Linear Systematic Sampling Procedure

The efficiency of remainder linear systematic sampling in relation to that of simple random sampling, circular systematic sampling or new partially systematic sampling depends on the characters of the populations. Thus, the efficiencies of these sampling procedures are compared for various types of populations.

3.1 *Population in random order.* Regard the finite population as drawn from an infinite super-population in which the expectation is denoted by symbol $\mathcal{E}$. If the variates $y_i (i = 1, 2, \ldots, N)$ are drawn at random from a super-population in which

$$\mathcal{E}\,(y_i) = \mu, \quad \mathcal{E}\,(y_i - \mu)^2 = \sigma_i^2$$

$$\mathcal{E}\,(y_i - \mu)(y_j - \mu) = 0, (i \neq j),$$

it is known as a population in random order (Cochran (1977)). The variance of the Horvitz-Thompson estimator is given in (2.2). By taking expectation of this quantity over the super-population model, we get the expected variance

$$\mathcal{E}\,V_{RLSS}(\overline{y}_{HT}) = \frac{1}{N^2}\left[(k-1)\sum_{i=1}^{(n-r)k}\sigma_i^2 + k\sum_{i=(n-r)k+1}^{N}\sigma_i^2\right]. \qquad (3.1)$$

Comparing to the expected variance for the simple random sample mean

$$\mathcal{E}\,V_{SRS}(\overline{y}) = \frac{N-n}{N^2 n}\sum_{i=1}^{N}\sigma_i^2, \qquad (3.2)$$

we have

$$\mathcal{E}\,V_{SRS}(\overline{y}) - \mathcal{E}\,V_{RLSS}(\overline{y}_{HT}) = \frac{1}{N^2 n}\left[r\sum_{i=1}^{N}\sigma_i^2 - n\sum_{i=(n-r)k+1}^{N}\sigma_i^2\right].$$

Therefore, remainder linear systematic sampling procedure is more efficient if and only if

$$\frac{r}{n} > \frac{\displaystyle\sum_{i=(n-r)k+1}^{N}\sigma_i^2}{\displaystyle\sum_{i=1}^{N}\sigma_1^2}. \qquad (3.3)$$

This means that the suggested method is better than simple random sampling if and only if the second stratum sample proportion is more than the second stratum variance proportion.

In the special case of equal model variances, it is obvious to see that

$$\mathcal{E} \, V_{SRS}(\overline{y}) - \mathcal{E} \, V_{RLSS}(\overline{y}_{HT}) = \frac{r(r-n)}{N^2 n}\sigma^2.$$

Equality occurs only when $r = 0$. In other words, remainder linear systematic sampling has the same sampling variance averaged over the model as simple random sampling without replacement only when $N = nk$ and $\sigma_i^2 = \sigma^2$ for all $i = 1, 2, \ldots, N$.

3.2 *Population with linear trend.* For the population with linear trend, we assume that $y_i = i$. It is obvious to see that

$$\frac{1}{k}\sum_{i=1}^{k}(\overline{y}_{1i} - \overline{Y}_1)^2 = \frac{k^2 - 1}{12} \tag{3.4}$$

and

$$\frac{1}{k+1}\sum_{i=1}^{k+1}(\overline{y}_{2i} - \overline{Y}_2)^2 = \frac{k(k+2)}{12}. \tag{3.5}$$

Substituting the results of (3.4) and (3.5) into equation (2.2), the variance of remainder linear systematic sampling procedure, we get

$$V_{RLSS}(\overline{y}_{HT}) = \frac{k}{12N^2}\left[(n-r)^2 k(k^2-1) + r^2(k+1)^2(k+2)\right]. \tag{3.6}$$

For the same population size N, the variance of simple random sampling is

$$V_{SRS}(\overline{y}) = \frac{(N-n)(N+1)}{12n}. \tag{3.7}$$

From the formulas of (3.6) and (3.7), it can be shown that $V_{RLSS}(\overline{y}_{HT}) \leq V_{SRS}(\overline{y})$. Equality occurs only when $r = 0$ and $n = 1$. That is, for population size being not a multiple of sample size, the variance of remainder linear systematic sampling is lower than that of simple random sampling.

Better results can be obtained with remainder linear systematic sampling by applying end corrections which removes the linear trend as a component of variance. Now let

$$
\begin{aligned}
a &= k_1, \\
b &= k_1 + (n - r - 1)k, \\
w &= \frac{2[(n-r)kk_1 + r(k+1)k_2] - (N+r)(k+1)}{2(n-r-1)Nk},
\end{aligned}
$$

then the estimator $\overline{y}'_{HT} = [(n-r)k\overline{y}_{1i} + r(k+1)\overline{y}_{2i}]/N + w(y_a - y_b)$ is equal to the finite population mean under a perfect linear trend. Hence, in a population having a linear trend, the estimator $\overline{y}'_{HT}$ with end corrections provides the exact population mean and thus $V(\overline{y}'_{HT}) = 0$. Note that, in the special case of $r = 0$, i.e. $N = nk$, the weight w reduces to $\frac{2k_1 - k - 1}{2(n-1)k}$, and thus, $\overline{y}'_{HT}$ is the same as the Yates's estimator.

3.3 *Auto-correlated population*. In this section, we consider the case of population size which is not a multiple of the sample size. Suppose that the corresponding values $y_i(i = 1, 2, \ldots, N)$ are drawn at random from a super-population in which

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

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

The expected variance for simple random sampling, circular systematic sampling and new partially systematic sampling are, respectively,

$$\sigma_{SRS}^2 = \left(\frac{1}{n} - \frac{1}{N}\right)\sigma^2 \left[1 - \frac{2}{N(N-1)} \sum_{d=1}^{N-1}(N-d)\rho_d\right], \tag{3.8}$$

$$\sigma_{CSS}^2 = \left(\frac{1}{n} - \frac{1}{N}\right)\sigma^2 + \frac{2}{Nn^2}\sigma^2 \sum_{t=1}^{N}\left[\sum_{i=0}^{n-1}\sum_{j>i}^{n-1}\rho_{|(ik+t)-(jk+t)|}\right] - \frac{2}{N^2}\sigma^2 \sum_{d=1}^{N-1}(N-d)\rho_d, \tag{3.9}$$

and

$$
\begin{aligned}
\sigma_{NPS}^2 &= \left(\frac{1}{n} - \frac{1}{N}\right)\sigma^2 - \frac{2}{N^2}\sigma^2 \sum_{d=1}^{N-1}(N-d)\rho_d \\
&\quad + \frac{2}{Nn^2}\sigma^2 \left\{\frac{a(a-1)}{u(u-1)} \sum_{t=1}^{N}\left[\sum_{i=0}^{u-1}\sum_{j>i}^{u-1}\rho_{|(t+j)-(t+i)|}\right]\right. \\
&\quad + \frac{a}{u} \sum_{t=1}^{N}\left[\sum_{i=0}^{u-1}\sum_{j=1}^{n-a}\rho_{|(t+i)-(jk+t+u-1)|}\right] \\
&\quad \left. + \sum_{t=1}^{N}\left[\sum_{i=0}^{n-a}\sum_{j>i}^{n-a}\rho_{|(ik+t+u-1)-(jk+t+u-1)|}\right]\right\}. \tag{3.10}
\end{aligned}
$$

When the remainder linear systematic sampling design is used, the expected variance is shown to be

$$
\begin{aligned}
\sigma_{RLSS}^2 &= \frac{\sigma^2}{N^2}\left\{k(N - n + r)\right. \\
&\quad - 2\sum_{d=1}^{(n-r)k-1}[(n-r)k - d]\rho_d - k^2 \sum_{d=1}^{n-r-1}(n-r-d)\rho_{dk} \\
&\quad \left. + 2\sum_{d=1}^{r(k+1)-1}[r(k+1) - d]\rho_d - (k+1)^2\sum_{d=1}^{r-1}(r-d)\rho_{d(k+1)}\right\}. \tag{3.11}
\end{aligned}
$$

From expressions (3.8) through (3.11), it is not possible to obtain a general result about the relative efficiency of the sampling procedures under consideration.

However, comparisons can be made for three 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 numerical studies, as the illustrative examples provided by Leu and Tsui (1996), are presented in Table 1, Table 2 and Table 3.

Table 1. VARIANCES OF FOUR SAMPLING PROCEDURES FOR LINEAR CORRELOGRAM

			Linear correlogram			
N	n	u, a	σ^2_{SRS}	σ^2_{CSS}	σ^2_{NPS}	σ^2_{RLSS}
10	4	4,2	0.0550	0.0300	0.0392	0.0200
15	7	9,4	0.0271	0.0068	0.0177	0.0059
25	8	10,3	0.0295	0.0052	0.0123	0.0049
25	12	11,5	0.0150	0.0023	0.0063	0.0019
35	8	5,2	0.0331	0.0073	0.0120	0.0051
35	12	5,2	0.0187	0.0023	0.0032	0.0021
35	17	15,7	0.0104	0.0012	0.0039	0.0009

Table 2. VARIANCES OF FOUR SAMPLING PROCEDURES FOR EXPONENTIAL CORRELOGRAM

			Exponential correlogram			
N	n	u, a	σ^2_{SRS}	σ^2_{CSS}	σ^2_{NPS}	σ^2_{RLSS}
10	4	4,2	0.1337	0.1096	0.1169	0.0998
15	7	9,4	0.0705	0.0406	0.0577	0.0453
25	8	10,3	0.0811	0.0518	0.0626	0.0547
25	12	11,5	0.0413	0.0219	0.0296	0.0234
35	8	5,2	0.0933	0.0683	0.0786	0.0702
35	12	5,2	0.0530	0.0322	0.0352	0.0327
35	17	15,7	0.0293	0.0150	0.0206	0.0159

Table 3. VARIANCES OF FOUR SAMPLING PROCEDURES FOR HYPERBOLIC CORRELOGRAM

			Hyperbolic correlogram			
N	n	u, a	σ^2_{SRS}	σ^2_{CSS}	σ^2_{NPS}	σ^2_{RLSS}
10	4	4,2	0.0763	0.0496	0.0586	0.0389
15	7	9,4	0.0440	0.0141	0.0307	0.0161
25	8	10, 3	0.0559	0.0192	0.0314	0.0210
25	12	11,5	0.0285	0.0071	0.0148	0.0078
35	8	5,2	0.0679	0.0298	0.0407	0.0296
35	12	5,2	0.0386	0.0114	0.0142	0.0114
35	17	15,7	0.0213	0.0046	0.0106	0.0049

From Table 1, Table 2 and Table 3, it is found that remainder linear systematic sampling design is more efficient than simple random sampling scheme and new partially systematic sampling scheme for these three special types of populations.

Compared with circular systematic sampling, the suggested sampling procedure is better than circular systematic sampling procedure under linear correlogram. For the cases of exponential and hyperbolic correlogram, the proposed method is as good as the circular systematic sampling method.

4.   Concluding Remarks

Remainder linear systematic sampling is an extension of the usual linear systematic sampling procedure that can be used when the population size is not a multiple of the sample size. It is more efficient than simple random sampling for populations in random order if and only if the second stratum sample proportion is more than the second stratum variance proportion. Also this scheme is better than simple random sampling for populations with linear trend, and more efficient than other given methods for populations exhibiting linear, exponential and hyperbolic correlogram. Further, it is superior to circular systematic sampling while compared for populations exhibiting linear correlogram and is as good as circular systematic sampling for populations exhibiting exponential and hyperbolic correlogram.

References

COCHRAN, W.G. (1977). *Sampling Techniques*, Third edition, 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.

LEU, C.H. AND TSUI, K.W. (1996). New partially systematic sampling. *Statistica Sinica*, **6**, 617-630.

MURTHY, M.N. AND RAO, T.J. (1988). Systematic sampling with illustrative examples. *Handbook of Statistics*, Vol. **6**, 147-185.

PADAM SINGH AND GARG, J.N. (1979). On balanced random sampling. *Sankhyā* C, **41**, 60-68.

SINGH D. AND PADAM SINGH (1977). New systematic sampling. *J. Statist. Plann. Inference*, **1**, 163-178.

HORNG-JINH CHANG AND KUO-CHUNG HUANG
GRADUATE INSTITUTE OF MANAGEMENT SCIENCES
TAMKANG UNIVERSITY
TAMSUI, TAIPEI, TAIWAN 25137
R.O.C.
E-mail: chj@mail.tku.edu.tw