DOI: 10.52131/pjhss.2024.v12i1.2053

Pakistan Journal of Humanities and Social Sciences

Volume 12, Number 01, 2024, Pages 637–644
Journal Homepage:
https://journals.internationalrasd.org/index.php/pjhss

A Novel Mixed Sampling Design Under the Existence of A Linear Trend

Muhammad Azeem [1], and Zahid Khan [2]

[1] Department of Statistics, University of Malakand, KP, Pakistan. Email: azeemstats@uom.edu.pk
[2] Department of Statistics, University of Malakand, KP, Pakistan. Email: zahid.uom1@gmail.com

ARTICLE INFO

Article History:

Received:	December 10, 2023
Revised:	March 17, 2024
Accepted:	March 18, 2024
Available Online:	March 20, 2024

Keywords:
Linear Trend
Stratified Sampling
Mixed-Mode Sampling
Milk Yield Data

Funding:
This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

ABSTRACT

In almost every field of social sciences, data is obtained from a small part drawn from a large population. Two common procedures to draw a sample from a population of interest are stratified random sampling and systematic random sampling. Each sampling procedure has been designed to be used in particular sample surveys. In real-world surveys, researchers may face a situation in which the simultaneous use of two or more sampling designs may be feasible to precisely estimate the population parameter rather than simply relying on a single sampling design. We propose a novel mixed-mode sampling scheme for real-world sample surveys in which the population units exhibit some sort of linear trend. Our proposed mixed sampling design combines the diagonal and linear systematic sampling with stratified random sampling. In real-world circumstances in which some sort of linear trend exists in the population units, we prove that our proposed mixed sampling design performs more efficiently than the available sampling designs. Additionally, unlike the systematic sampling design, the suggested sampling scheme can be applied practically for any sample and population size.

Corresponding Author's Email: azeemstats@uom.edu.pk

1. Introduction

In real-world sample surveys, the most widely applied random sampling schemes for drawing a small sample from a large population include simple random, stratified, and cluster random sampling schemes. Every sampling design features some advantages but also suffers from some disadvantages. The simplest method of drawing a probability sample is the simple random sampling scheme. Simple random sampling may be a useful method in situations in which there is homogeneity among the population units. On the other hand, simple random sampling generally provides less efficient estimates than other sampling designs especially when compared to stratified random sampling. A stratified sample divides the population under consideration into mutually exclusive subpopulations generally called strata. After stratification, simple random samples are chosen from each stratum / subpopulation. Besides stratified sampling, systematic sampling has also gotten popularity among survey researchers due to its simplicity. The systematic random sampling method was originally developed in the study of Madow and Madow (1944). Systematic sampling draws a sample of n units so that the first unit for the sample is chosen from the first few units of the population. After selection of the first unit, a particular order of unit-selection is followed to select the remaining units. One serious drawback of the systematic sampling procedure is the requirement that the sizes of the population and sample must reconcile, which happens very rarely in real-world situations. To address this issue, Lahiri (1951) presented a novel systematic sampling design terming it the circular systematic sampling scheme. The remainder systematic sampling, another variant of systematic random sampling, was suggested by Chang and Huang (2000). Remainder systematic sampling is practically useful for almost every population and sample size. Subramani (2000) suggested an improved variant of the systematic sampling scheme in which the units for the sample are obtained diagonally from the population matrix.

Huang (2004) presented a novel mixed-mode sampling method which unified simple random sampling with systematic random sampling to unbiasedly estimate the sampling variance. Later on, a precise variant of the systematic random sampling was developed in Subramani (2012) research study. The Subramani (2012) sampling design was further enhanced by a general systematic random sampling scheme by Subramani and Gupta (2014). Recently, Azeem (2021), Azeem (2022) and Khan, Azeem, Aftab, and Ismail (2022) studied various aspects of the estimation of parameters under different sample selection procedures.

In real-world sample surveys, there are situations where the population units can follow a decreasing or an increasing form of a linear trend. For example, consider the milk yield data of cows recorded on daily basis over a particular time period. Generally, the data will follow a decreasing linear trend as the daily milk produced by cows, from the time of delivery, will decrease over time.

In real-world situations in which a linear trend is present, one may use linear systematic sampling design. However, linear systematic sampling has a limitation that the population size must reconcile with the sample size which happens very rarely. In practice, this condition rarely meets, which makes it impossible to use systematic sampling. The diagonal systematic sampling also suffers from this drawback, which limits its application to real-world surveys. Likewise, simple random sampling is not feasible when the population units are heterogeneous. In such situations, one may use a stratified sampling design. However, if there is a linear trend in the data, we can achieve further improvement in efficiency over stratified sampling if we use a mixed-mode sampling scheme.

This study suggests a novel mixed sampling scheme which combines the diagonal systematic random sampling, and stratified sampling as well as the linear systematic random sampling. The algebraic expressions for the efficiency conditions have also been obtained. We prove that our new suggested mixed sampling method performs more efficiently than the available sampling methods under linear trend.

2. Proposed Mixed Sampling Design

Taking motivation from the study of Huang (2004), we introduce a new mixed sampling scheme. Consider a population containing *N* elements numbered as 1, 2, 3, …, *N*. Suppose we need to choose a random sample of *n* units, when the population size can be expressed as $N = kk + k(n-k-r) + (k+1)r = nk + r$, with $k < n$ where r denotes some positive number. Our proposed mixed-mode sampling design contain the following sample-selection steps:

1. Start with partitioning the population under study into three mutually exclusive groups of units: Group-1, Group-2 and Group-3, with such a strategy that Group-1 receives the first $kk = k^2$ population units. Likewise, Group-2 receives the $k(n-k-r)$ population units next to the first group, whereas Group-3 receives the last $(k+1)r$ population units.
2. In Group-1, place the k^2 units into a $k \times k$ diagonal matrix. In Group-2, place the $k(n-k)$ units into another matrix having order $(n-k)\times k$, whereas in Group-3, place the $r(k+1)$ units into a third matrix having order $r\times(k+1)$. The units assigned to different groups have been given in Table 1, Table 2, and Table 3.
3. In Group-1, utilize diagonal systematic random sampling to collect a random sample containing *k* units. To draw the sample, choose a random number r_1 so that $1 \le r_1 \le k$. Obtain the units so that the k units chosen in the sample are the elements of the diagonal / broken diagonal in the Group-1 matrix. Likewise, from Group-2, obtain a systematic random sample consisting of $n-k$ units. To obtain the sample, choose a random value r_2 such that $1 \le r_2 \le k$. Select the sample from Group-2 so that the units chosen in the sample are the entries of the r_2th column in the Set-2 matrix. Moreover, in Group-3, consider the r rows and the $k+1$ columns as r strata with each stratum having size

$k+1$. To obtain the sample, choose a single unit randomly from each stratum. Finally, gather the units chosen from all of the three sets to obtain the required sample.

4.

Table 1: Unit Placement in the First Group

Rows	Columns			
	1st	2nd	…	k th
1st	x_1	x_2	…	x_k
2nd	x_{k+1}	x_{k+2}	…	x_{2k}
…	…	…	…	…
k th	$x_{(k-1)k+1}$	$x_{(k-1)k+2}$	…	x_{kk}

Table 2: Unit Placement in the Second Group

Rows	Columns			
	1st	2nd	…	kth
(k+1) th	x_{kk+1}	x_{kk+2}	…	$x_{kk+k=(k+1)k}$
(k+2) th	$x_{(k+1)k+1}$	$x_{(k+1)k+2}$	…	$x_{(k+2)k}$
…	…	…	…	…
$(n-r)$th	$x_{(n-r-1)k+1}$	$x_{(n-r-1)k+2}$	…	$x_{(n-r)k}$

Table 3: Unit Placement in the Third Group

Rows	Columns			
	1st	2nd	…	$(k+1)$th
$(n-r+1)$th	$x_{(n-r)k+1}$	$x_{(n-r)k+2}$	…	$x_{(n-r+1)k}$
$(n-r+2)$th	$x_{(n-r+1)k+1}$	$x_{(n-r+1)k+2}$	…	$x_{(n-r+2)k}$
…	…	…	…	…
n th	$x_{(n-1)k+1}$	$x_{(n-1)k+2}$	…	x_{nk}

We observe that one can draw k probable samples from each of the first two sets. Likewise, a total of $(k+1)^r$ possible samples can be drawn from Group-3. In this manner, our suggested mixed sampling scheme contains $k^2(k+1)^r$ total possible samples, with each sample containing n units. The probabilities of inclusion of units under the mixed sampling design may be expressed as:

$$\pi_i = \begin{cases} \dfrac{1}{k}, & \text{if the } i\text{th population unit is in the first or second group,} \\ \dfrac{1}{k+1}, & \text{if the } i\text{th population unit is in the third group.} \end{cases} \qquad (1)$$

Also,

$$\pi_{ij} = \begin{cases} 1/k, & \text{if both } i\text{th and } j\text{th units are in the same diagonal} \\ & \text{in the first group or to the same column in the second group,} \\ 1/(k+1)^2, & \text{if both } i\text{th and } j\text{th unit are in different rows of the third group,} \\ 1/k^2, & \text{if the } i\text{th and } j\text{th unit are in the first and second group, respectively,} \\ 1/(k+1)k, & \text{if } \textit{the } i\text{th and } j\text{th unit are in the first and third group, respectively,} \\ & \text{or if } i\text{th and } j\text{th unit belong to the second group and third group, respectively,} \\ 0, & \text{otherwise.} \end{cases}$$

(2)

Let us denote the mean of the units belonging to the first, second, and third group using the notations $\bar{x}_1$, $\bar{x}_2$, and $\bar{x}_3$, respectively. The sample mean under our suggested mixed-mode sampling scheme can be expressed as:

$$\bar{x}_p = w_1\bar{x}_1 + w_2\bar{x}_2 + w_3\bar{x}_3 . \tag{3}$$

The notations in Eq. (3) are explained as:

$$\bar{x}_1 = \frac{\sum_{i=0}^{k-1} x_{(k+1)i+r_1}}{k}, \qquad \text{if } r_1 = 1, \tag{4}$$

or

$$\bar{x}_1 = \frac{\sum_{i=0}^{t} x_{i(k+1)+r_1} + \sum_{i=1}^{k-t-1} x_{(t+i)k+i}}{k}, \qquad \text{if } r_1 > 1, \tag{5}$$

$$\bar{x}_2 = \frac{\sum_{i=0}^{n-k-r-1} x_{ik+r_2}}{n-k-r}, \tag{6}$$

and

$$\bar{x}_3 = \frac{\sum_{i=1}^{r} x_i}{r}, \tag{7}$$

where

$w_1 = \frac{kk}{N}$, $w_2 = \frac{k(n-k-r)}{N}$, $w_3 = \frac{(k+1)r}{N}$, where $\sum w_i = 1$.

3. Practical Existence of Linear Tendency Among Units

The term linear trend may be regarded as some sort of an arithmetic progression which can exist among the population units. There are many real-world situations in which a moderate or a severe level of a linear trend can exist. For instance, consider an educational institute where roll numbers are allotted to the students based on their grades in the previous academic program. In such situations, intelligent students generally get the top enrollment numbers. Thus, if their examination scores are placed in order of their roll numbers, we will observe the existence of a severe linear trend in the data.

To give another example of the practical existence of linear trend, let us consider the milk yield of cows observed daily, starting from the day of delivery. One may anticipate that the quantity of the milk yield will decrease over weeks, thus leading to a decreasing trend.

Now let us analyze the performance of different sampling designs in real-world scenarios in some decreasing or increasing type linear trend is present. Let the *N* population units exhibit linear trend. Then:

$$x_i = c + id, \qquad \text{where } i=1, 2, 3, \ldots, N. \tag{8}$$

The sampling variance using simple random sampling can be obtained as:

$$Var(\bar{x}_r) = (N+1)(k-1)\frac{d^2}{12}. \tag{9}$$

The sampling variance using the systematic random sampling design can be obtained as:

$$Var(\bar{x}_{sy}) = (k+1)(k-1)\frac{d^2}{12}. \tag{10}$$

The sampling variance using the diagonal systematic random sampling can be obtained as:

$$Var\left(\bar{x}_{dsy}\right)=\left(k-n\right)\left[2+\left(k-n\right)n\right]\frac{d^2}{12n}. \quad (11)$$

The sampling variance based on remainder systematic random sampling can be expressed as follows:

$$Var\left(\bar{x}_{rsy}\right)=\left[\left(n-r\right)^2 k\left(k^2-1\right)+\left(k+2\right)r^2\left(k+1\right)^2\right]kd^2 \big/ 12N^2. \quad (12)$$

Likewise, the sampling variance using the Subramani (2012) procedure can be obtained as:

$$Var\left(\bar{x}_{msy}\right)=\left[1+\left(n-1\right)^2\right]\left(k-1\right)\left(k+1\right)\frac{d^2}{12n^2}. \quad (13)$$

The sampling variance based on our proposed sampling design is derived as:

$$Var\left(\bar{x}_p\right)=w_1^2 Var\left(\bar{x}_1\right)+w_2^2 Var\left(\bar{x}_2\right)+w_3^2 Var\left(\bar{x}_3\right). \quad (14)$$

As the first group contains a total of $k\times k=k^2$ units, putting $k=n$ in Eq. (11) gives:

$$Var\left(\bar{x}_1\right)=0. \quad (15)$$

Group-2 has $\left(n-k-r\right)k$ units and since the variance of $\bar{x}_{sy}$ does not depend on n, therefore, it follows that:

$$Var\left(\bar{x}_2\right)=Var\left(\bar{x}_{sy}\right)=\left(k+1\right)\left(k-1\right)\frac{d^2}{12}. \quad (16)$$

Group-3 consists of a total of $r\left(k+1\right)$ units where stratified random sampling is used. This leads to:

$$Var\left(\bar{x}_3\right)=\frac{\left(k+1\right)\left(k-1\right)}{12}\frac{d^2}{r}. \quad (17)$$

Using Eq. (15), (16) and (17) in Eq. (14), the sampling variance of $\bar{y}_p$ can be derived as follows:

$$Var\left(\bar{x}_p\right)=\left[k^2\left(n-k-r\right)^2+r\left(k+1\right)^2\right]\frac{\left(k+1\right)\left(k-1\right)d^2}{12N^2}. \quad (18)$$

Under linear trend, all of the variances discussed above have been obtained by using the sum of the arithmetic progression given by:

$$\sum_{i=1}^{N} i = N\left(N+1\right)/2.$$

Also,

$$1^2+2^2+....+N^2=\sum_{i=1}^{N} i^2=\frac{\left(2N+1\right)N\left(N+1\right)}{6}.$$

In the absence of linear trend, we cannot use the sum of the series to obtain the variances. As a result, the mathematical expressions of the variances under each sampling scheme will be different from those presented in Eq. (9) to Eq. (13) and Eq. (18). Consequently, this will affect the efficiency of the mean under each sampling scheme.

4. Efficiency Comparison with Existing Sampling Methods

Under the existence of linear trend among units, the mixed sampling design performs more efficiently as compared to simple random sampling design if we have:

$$Var\left(\bar{x}_p\right)<Var\left(\bar{x}_r\right). \quad (19)$$

Putting Eq. (9) and Eq. (18) in Eq. (19) and simplifying, we get:

$$\left[\left(n-k-r\right)^2 k^2+\left(k+1\right)^2 r\right]<\frac{N^2\left(N+1\right)}{k+1}. \tag{20}$$

One can note that condition (20) always holds, to conclude that the new mixed sampling design performs more efficiently as compared to simple random sampling in all practical scenarios.

Our suggested mixed sampling design performs more efficiently than the systematic sampling if we have:

$$Var\left(\overline{x}_p\right)<Var\left(\overline{x}_{sy}\right). \tag{21}$$

Putting Eq. (10) and Eq. (18) in Eq. (21) and simplifying, we get:

$$\frac{\left(n-k-r\right)^2 k^2+\left(k+1\right)^2 r}{N^2}<1. \tag{22}$$

Condition (22) is always true if $k<n$.

Our suggested mixed-mode sampling is more efficient as compared to the diagonal systematic random sampling design if we have:

$$Var\left(\overline{x}_p\right)<Var\left(\overline{x}_{dsy}\right). \tag{23}$$

Using Eq. (11) and Eq. (18) in Eq. (23) and some algebraic simplification yields:

$$\frac{n}{N^2}\left[k^2\left(n-k-r\right)^2+r\left(k+1\right)^2\right]\left(k+1\right)\left(k-1\right)<\left[\left(k-n\right)n+2\right]\left(k-n\right). \tag{24}$$

Our suggested mixed-mode sampling design performs more efficiently as compared to the Subramani (2012) method if:

$$Var\left(\overline{x}_p\right)<Var\left(\overline{x}_{msy}\right). \tag{25}$$

Using Eq. (13) and Eq. (18) in Eq. (25) leads to the condition:

$$k^2\left[\left(n-1\right)^2+\left(n-k-r\right)^2+1\right]+\left(k+1\right)^2 r<N^2. \tag{26}$$

Our suggested mixed-mode sampling performs more efficiently than the remainder systematic random sampling design if we have:

$$Var\left(\overline{x}_p\right)<Var\left(\overline{x}_{rsy}\right). \tag{27}$$

Putting Eq. (12) and Eq. (18) in Eq. (27) gives:

$$\left[k^2-2k\left(n-r\right)\right]k^2\left(k-1\right)<\left[\left(2rk+1\right)+k^2\left(r-1\right)\right]\left(k+1\right)r. \tag{28}$$

5. Concluding Remarks

In this study, we presented a novel mixed sampling scheme which improves the efficiency of the already available sampling schemes. The proposed sampling design is useful in real-world situations where the population units exhibit some sort of a decreasing or an increasing linear tendency. The sampling variance under various sampling schemes, using different values of *k*, *N*, and *n*, have been provided in Table 4. The improvement over the available sampling schemes can also be observed graphically from the box-plot given in Figure 1. The values for *k*, *N*, and *n* have been selected so that $N=nk+r$ and $n>k$. One may note that the constant d^2 works as a multiplicative constant in the variance expression of each sampling scheme, therefore we have used d = 1 for the sake of simplicity. Table 4 indicates that our proposed mixed sampling design outperforms the available sampling designs in terms of efficiency.

Figure 1: Sampling Variance Under Different Sampling Schemes

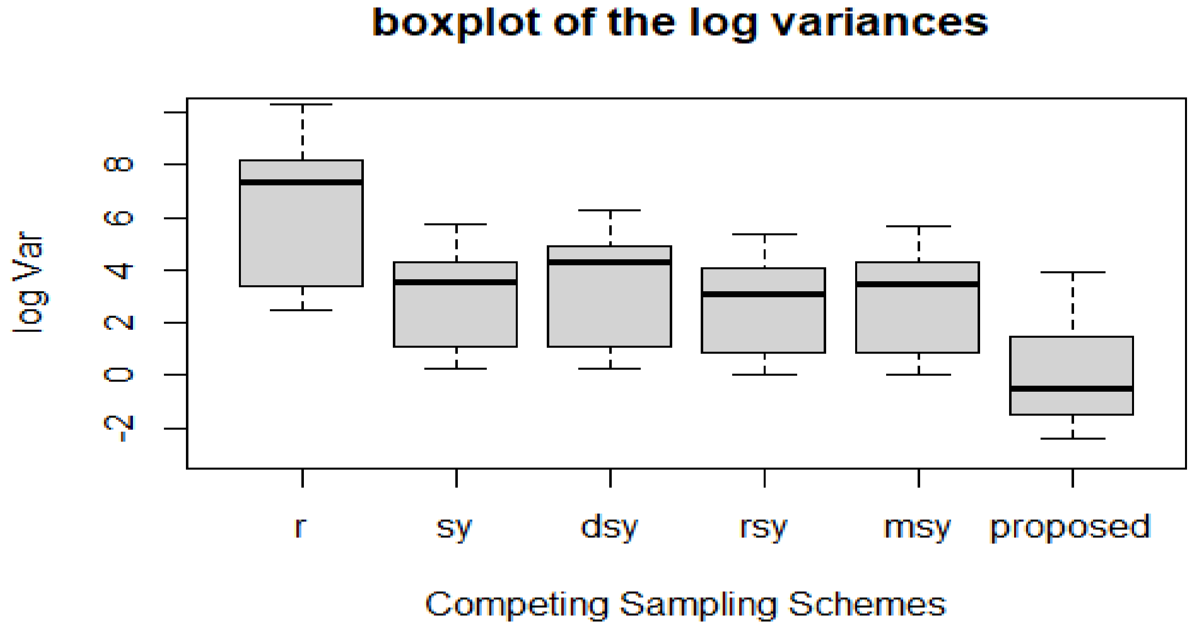

Table 4: Sampling Variance Under Different Sampling Procedures

n	*k*	*r*	$Var(\bar{x}_r)$	$Var(\bar{x}_{sy})$	$Var(\bar{x}_{dsy})$	$Var(\bar{x}_{rsy})$	$Var(\bar{x}_{msy})$	$Var(\bar{x}_p)$
10	4	6	11.750	1.250	2.901	1.001	1.030	0.090
		7	12.001	1.250	2.901	1.190	1.030	0.110
		8	12.250	1.250	2.901	1.420	1.030	0.140
		9	12.501	1.250	2.901	1.701	1.030	0.190
	6	4	27.080	2.920	1.270	1.690	2.391	0.140
		5	27.500	2.920	1.270	1.780	2.391	0.190
		6	27.920	2.920	1.270	2.011	2.391	0.291
		7	28.330	2.920	1.270	2.350	2.391	0.430
		8	28.750	2.920	1.270	2.801	2.391	0.611
50	10	40	375.750	8.250	133.201	8.071	7.930	0.161
		44	375.750	8.250	133.201	9.490	7.930	0.230
		48	375.750	8.250	133.201	11.160	7.930	0.401
	20	30	1584.921	33.250	74.900	19.870	31.951	0.440
		35	1584.921	33.250	74.900	22.801	31.951	0.850
		40	1584.921	33.250	74.900	27.201	31.951	1.920
	30	20	3627.420	74.920	33.271	40.640	71.980	0.640
		30	3627.420	74.920	33.271	42.740	71.980	3.960
		40	3627.420	74.920	33.271	57.670	71.980	13.271
100	20	20	3199.921	33.250	533.201	22.450	32.590	11.810
		50	3247.421	33.250	533.201	17.530	32.590	3.021
		80	3294.921	33.250	533.201	25.151	32.590	0.270
	40	20	13068.250	133.250	299.901	90.260	130.610	21.390
		50	13165.751	133.250	299.901	68.360	130.610	1.981
		80	13263.250	133.250	299.901	95.600	130.610	6.201
	60	20	29603.250	299.921	133.270	203.411	293.980	12.530
		50	29750.751	299.921	133.270	152.531	293.980	4.470
		80	29898.250	299.921	133.270	211.390	293.980	49.151

Like any other sampling scheme, the proposed mixed-mode sampling method also suffers from some drawbacks. In real-world scenarios where a linear tendency among population units does not exist, the proposed sampling scheme may be less efficient than other sampling designs like simple random sampling or stratified sampling. It is therefore advised for the survey practitioners that they need to be aware about the behavior of the population units before applying the proposed mixed-mode sampling design.

We have evaluated the efficiency of the sample mean under the new mixed-mode sampling scheme. It is recommended for future researchers to evaluate the efficiency of the sample variance and sample proportion using the suggested mixed sampling design.

Under the existence of a linear trend, the increase in efficiency compared to the existing sampling designs makes our suggested mixed sampling design preferable for implementation in real-world surveys.

References

Azeem, M. (2021). On Estimation of Population Proportion in Diagonal Systematic Sampling. *Life Cycle Reliability and Safety Engineering, 10*(3), 249-254. doi:https://doi.org/10.1007/s41872-021-00165-0

Azeem, M. (2022). A Modified Version of Diagonal Systematic Sampling in the Presence of Linear Trend. *Plos one, 17*(3), e0265179.

Chang, H.-J., & Huang, K.-C. (2000). Remainder Linear Systematic Sampling. *Sankhyā: The Indian Journal of Statistics, Series B*, 249-256.

Huang, K.-C. (2004). Mixed Random Systematic Sampling Designs. *Metrika, 59*(1), 1-11. doi:https://doi.org/10.1007/s001840300264

Khan, Z., Azeem, M., Aftab, M. H., & Ismail, M. (2022). Modified Median Ranked Set Sampling Scheme for Estimating Population Mean.

Lahiri, D. B. (1951). A Method of Sample Selection Providing Unbiased Ratio Estimates. *Bulletin Institut International de Statistique, 33*, 133.

Madow, W. G., & Madow, L. H. (1944). On the Theory of Systematic Sampling, I. *The Annals of Mathematical Statistics, 15*(1), 1-24.

Subramani, J. (2000). Diagonal Systematic Sampling Scheme for Finite Populations. *Journal of the Indian Society of Agricultural Statistics, 53*(2), 187-195.

Subramani, J. (2012). A Modification on Linear Systematic Sampling for Odd Sample Size. *Bonfring International Journal of Data Mining, 2*(2), 32-36.

Subramani, J., & Gupta, S. N. (2014). Generalized Modified Linear Systematic Sampling Scheme for Finite Populations. *Hacettepe Journal of Mathematics and Statistics, 43*(3), 529-542.