

Estimation of Population Proportion in Remainder Systematic Sampling

Dr. Muhammad Azeem¹, Dr. Zahid Khan² & Dr. Sanam Wagma Khattak³

¹ Assistant Professor, Department of Statistics, University of Malakand, Chakdara, Dir (Lower), Pakistan.

² Lecturer, Department of Statistics, University of Malakand, Chakdara, Dir (Lower), Pakistan.

³ Lecturer, Department of Economics, University of Peshawar.

Corresponding author Email: azeemstats@uom.edu.pk

Received date: 19th January 2022

Revised date: 05th March 2022

Accepted date: 11th April 2022

Abstract: In survey sampling, the linear systematic sampling is a popular method of drawing a sample from a finite population. A limitation of linear systematic sampling is that it requires the population size be a constant multiple of the required sample size. This restriction puts a limit on its usefulness as the size of population is not usually a constant multiple of the required sample size. An alternative sampling scheme is the remainder systematic sampling which can be used for any sample size and population size. In literature, remainder systematic sampling and its modified forms have been studied by many researchers but all of these studies are limited to quantitative characteristics only. In many practical situations, the researchers face situations where the variable under study is a binary qualitative variable. In this paper, the efficiency of the estimates of population proportion under remainder systematic sampling scheme is studied. It is found that the remainder systematic sampling scheme provides more efficient estimates of population proportion than simple random sampling and linear systematic sampling.

Keywords: Simple Random Sampling, Linear Systematic Sampling, Sampling Variance, Remainder Systematic Sampling, Population Proportion.

Introduction

In the field of survey sampling, the linear systematic sampling method, simply called systematic sampling, dates back to Madow and Madow (1944). In this method of sample selection, a sample of size n units from a finite population of size N units is obtained in such a way that the first unit is drawn from the first k ($=N/n$) units. After selecting the first unit, every k th unit of the population is then selected in the sample. Thus, if the population units are arranged into a table having n rows and k columns, the linear systematic sampling method actually selects one column from a total of k columns. Lahiri (1951) suggested what is called circular systematic sampling scheme. Chang and Huang (2000) developed a modified version of the systematic sampling called remainder systematic sampling in an attempt to make systematic sampling

widely applicable for any sample and population size. The diagonal systematic sampling scheme was introduced by Subramani (2000). Sampath and Varalakshmi (2008) introduced a modified systematic sampling scheme which they called diagonal circular systematic sampling. Subramani (2009) introduced a generalized version of diagonal systematic sampling method. Khan et al. (2014) suggested the conditions under which the Sampath and Varalakshmi (2008) sampling scheme is applicable. Khan et al. (2015) proposed a generalized version of systematic sampling and it was shown that diagonal systematic sampling is a special case of the new generalized sampling scheme. Recently, Naidoo et al. (2018) suggested a new modified version of remainder systematic sampling design. In addition to the above studies, various aspects of systematic sampling has been studied by many researchers including Yates (1948), Madow (1953), Bellhouse and Rao (1975), Cochran (1977), Fountain and Pathak (1989), Subramani (2012, 2013), Subramani and Gupta (2014), Khan et al. (2013), Naidoo et al. (2015) and Subramani (2018) etc. Recently, Azeem et al. (2021) presented a new approach to using systematic sampling, and findings of the study revealed that the new sampling design achieved more gain in efficiency than the existing sampling designs.

In many practical situations, the researcher is interested to estimate the proportion or total number of population units possessing some binary qualitative characteristics like employment status (employed or unemployed), gender (male or female), literacy status (literate or illiterate), area type (rural or urban) etc. In this paper, the problem estimation of population proportion under remainder systematic sampling is studied and a comparative analysis of the efficiency of estimates of population proportion under various sampling designs is carried out through a simulation study.

1. Notations

Let the population of finite size consists of $N = nk + r = (n - r)k + r(k + 1)$ units and it is required to estimate the proportion of population units possessing a particular characteristic of interest based on a sample of size n so that $n = (n - r) + r$, where k is a positive integer. Let every population unit y_i ($i = 1, 2, \dots, N$) belongs to one of the two mutually exclusive classes C and C' where C is the class of units having the characteristic of interest. That is, let

$$y_i = \begin{cases} 1, & \text{if } i\text{th unit of population belongs to class } C, \\ 0, & \text{otherwise.} \end{cases} \quad (2.1)$$

Let the number of population units belonging to class C are denoted by A and the number of sample units belonging to class C are denoted by a . Moreover, let the population proportion be denoted by:

$$P = \frac{A}{N}, \quad (2.2)$$

where $A = \sum_{i=1}^N y_i$.

In remainder systematic sampling scheme, the steps involved are as follows:

1. Divide the population of size N into two sets: Set-1 and Set-2, in such a manner that Set-1 receives the first $(n-r)k$ units y_i ($i=1, 2, \dots, (n-r)k$) and Set-2 receives the remaining $r(k+1)$ units y_i ($i=(n-r)k+1, (n-r)k+2, \dots, (n-r)k+r(k+1)$).
2. In Set-1, arrange the $(n-r)k$ units in a table having $n-r$ rows and k columns such that $n-r \leq k$. In Set-2, arrange the $r(k+1)$ units in a table having r rows and $k+1$ columns (see Table 1-2).
3. Select a pair of random numbers r_1 and r_2 where $1 \leq r_1 \leq k$ and $1 \leq r_2 \leq k+1$. In Set-1, the units are drawn in such a way that the selected $n-r$ units are the entries in the r_1 th column of Table 1. In Set-2, the selected r units are the elements in the r_2 th column of the Table 2. Finally, the selected units from both sets are combined to get the sample of size n .

Table 1: Arrangement of the population units in Set-1

S.No.	1	2	...	k
1	y_1	y_2	...	y_k
2	y_{k+1}	y_{k+2}	...	y_{2k}
3	y_{2k+1}	y_{2k+2}	...	y_{3k}
...
$n-r$	$y_{(n-r-1)k+1}$	$y_{(n-r-1)k+2}$...	$y_{(n-r)k}$

Table 2: Arrangement of the population units in Set-2

S.No.	1	2	...	$k+1$
1	$y_{(n-r)k+1}$	$y_{(n-r)k+2}$...	$y_{(n-r)k+(k+1)}$
2	$y_{(n-r)k+(k+1)+1}$	$y_{(n-r)k+(k+1)+2}$...	$y_{(n-r)k+2(k+1)}$
3	$y_{(n-r)k+2(k+1)+1}$	$y_{(n-r)k+2(k+1)+2}$...	$y_{(n-r)k+3(k+1)}$
...
r	$y_{(n-r)k+(r-1)(k+1)+1}$	$y_{(n-r)k+(r-1)(k+1)+2}$...	$y_{(n-r)k+r(k+1)}$

It is clear that the remainder sampling procedure has $k(k+1)$ possible samples each of size n . The probabilities of inclusion are given by:

$$\pi_i = \begin{cases} \frac{1}{k} & \text{if } i\text{th unit belongs to Set-1,} \\ \frac{1}{k+1} & \text{if } i\text{th unit belongs to Set-2.} \end{cases} \quad (2.3)$$

$$\pi_{ij} = \begin{cases} \frac{1}{k} & \text{if } i\text{th and } j\text{th units are from the same column of Set-1,} \\ \frac{1}{k+1} & \text{if } i\text{th and } j\text{th units are from the same column of Set-2,} \\ \frac{1}{k(k+1)} & \text{if } i\text{th and } j\text{th units are from Set-1 and Set-2 respectively,} \\ 0 & \text{otherwise.} \end{cases} \quad (2.4)$$

The selected sampling units are:

$$S_{r_1 r_2} = \left\{ y_{r_1}, y_{k+r_1}, \dots, y_{(n-r-1)k+r_1}, y_{(n-r)k+r_2}, y_{(n-r)k+(k+1)+r_2}, \dots, y_{(n-r)k+(r-1)(k+1)+r_2} \right\},$$

where $r_1 = 1, 2, \dots, k$; $r_2 = 1, 2, \dots, k+1$.

3. Estimator of Population Proportion and its Properties

The sample proportion based on simple random sampling is given by:

$$p_{SRS} = \frac{a_{SRS}}{n}, \quad (3.1)$$

where

$$a_{SRS} = \sum_{i=1}^n y_i.$$

The variance of p_{SRS} under simple random sampling without replacement is given by:

$$Var(p_{SRS}) = \frac{N-n}{N-1} \frac{PQ}{n}, \quad \text{where } Q = 1 - P. \quad (3.2)$$

The sample proportion based on linear systematic sampling is given by:

$$p_{sy} = \frac{a_{sy}}{n}, \quad (3.3)$$

where

$$a_{sy} = \sum_{i=0}^{n-1} y_{ik+r} .$$

The variance of p_{sy} is given by:

$$Var(p_{sy}) = \frac{1}{k} \sum_{i=1}^k (p_{sy} - P)^2 . \quad (3.4)$$

The sample proportion based on remainder systematic sampling scheme is given by:

$$p_{rsy} = \frac{(n-r)kp_{1rsy} + r(k+1)p_{2rsy}}{N} , \quad (3.5)$$

where

$$p_{1rsy} = \frac{a_{1rsy}}{n-r} , \quad (3.6)$$

$$p_{2rsy} = \frac{a_{2rsy}}{r} , \quad (3.7)$$

$$a_{1rsy} = \sum_{i=0}^{n-r-1} y_{ik+r_1} , \quad (3.8)$$

and

$$a_{2rsy} = \sum_{i=0}^{r-1} y_{i(k+1)+r_2} . \quad (3.9)$$

Theorem 1: Under remainder systematic sampling scheme, the sample proportion can be written in the form of Horvitz-Thompson estimator p_{HT} introduced by Horvitz and Thompson (1952) and is unbiased for population proportion P .

Proof: By definition

$$\begin{aligned} p_{rsy} &= \frac{(n-r)kp_{1rsy} + r(k+1)p_{2rsy}}{N} , \\ &= \frac{1}{N} [ka_{1rsy} + (k+1)a_{2rsy}] , \\ &= \frac{1}{N} \left[k \sum_{i \in s_1} y_i + (k+1) \sum_{i \in s_2} y_i \right] , \end{aligned}$$

where s_1 and s_2 denote the samples drawn from Set-1 and Set-2 respectively.

$$\begin{aligned}
 p_{rsy} &= \frac{1}{N} \left(\sum_{i \in S_1} \frac{y_{1i}}{1/k} + \sum_{i \in S_2} \frac{y_{2i}}{1/(k+1)} \right), \\
 &= \frac{1}{N} \sum_{i \in S} \frac{y_i}{\pi_i} = p_{HT}.
 \end{aligned} \tag{3.10}$$

Taking expectation on both sides of (3.5) yields:

$$E(p_{rsy}) = \frac{(n-r)k}{N} E(p_{1rsy}) + \frac{r(k+1)}{N} E(p_{2rsy}). \tag{3.11}$$

Now

$$\begin{aligned}
 E(p_{1rsy}) &= E\left(\frac{1}{n-r} \sum_{i=1}^{n-r} y_{1i}\right) = \frac{1}{n-r} \sum_{i=1}^{n-r} E(y_{1i}), \\
 &= \frac{1}{n-r} \sum_{i=1}^{n-r} \left[\sum_{i \in S_1} y_{1i} \frac{1}{(n-r)k} \right], \\
 &= \frac{1}{(n-r)k} \sum_{i \in S_1} y_{1i} = P_1.
 \end{aligned} \tag{3.12}$$

Similarly,

$$E(p_{2rsy}) = \frac{1}{r(k+1)} \sum_{i \in S_2} y_{2i} = P_2, \tag{3.13}$$

where S_1 and S_2 denotes all units in Set-1 and Set-2 respectively.

Substituting (3.12) and (3.13) in (3.11) and simplification yields:

$$E(p_{rsy}) = P.$$

Remark 1: Although Horvitz-Thompson estimator is usually used to estimate the population mean or total, it can also be used to estimate population proportion by simply treating the variable of interest as binary variable with possible values 0 and 1.

Remark 2: Using Sen-Yates-Grundy approach suggested by Sen (1953) and Yates and Grundy (1953), the variance of p_{rsy} can be written as:

$$Var(p_{rsy}) = Var_{SYG}(p_{HT}) = \frac{1}{N^2} \left\{ \frac{1}{2} \sum_{i=1}^N \sum_{\substack{j=1 \\ j \neq i}}^N (\pi_i \pi_j - \pi_{ij}) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 \right\}. \tag{3.14}$$

Remark 3: The Sen-Yates-Grundy estimator for (3.14) is given by:

$$\text{var}(p_{rsy}) = \text{var}_{SYG}(p_{HT}) = \frac{1}{2N^2} \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n \left(\frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \right) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2. \quad (3.15)$$

The values of π_i and π_{ij} can be used from (2.3) and (2.4) in expression (3.14) and (3.15) to obtain the sampling variance of the sample proportion and its estimator under the remainder systematic sampling scheme.

4. Simulation Study and Conclusion

For the purpose of efficiency comparison, population data was generated from Bernoulli distribution for different choices of proportion ($P = 0.3, 0.5, 0.7$) using R language command. From these generated populations, random samples were drawn repeatedly by three different sampling schemes: simple random sampling, linear and remainder systematic sampling. The results of the variances of sample proportion for different choices of n, k and r have been presented in Table 3-5 (see Appendix). It is clear that if sample size is small, the sample proportion based on remainder systematic sampling scheme is more efficient than both simple random sampling and linear systematic sampling. Moreover, it is also clear that as the sample size increases, the sampling variances of the sample proportion based on all three sampling designs under consideration approach to zero. Therefore, it is recommended to researchers to use remainder systematic sampling scheme for getting precise estimates of population proportion of units which possess the characteristic of interest.

References

1. Azeem, M., Asif, M., Ilyas, M., Rafiq, M., and Ahmad, S. (2021). An efficient modification to diagonal systematic sampling for finite populations. *AIMS Mathematics*, 6(5), 5193-5204.
2. Bellhouse, D.R. and Rao, J.N.K. (1975). Systematic sampling in the presence of linear trends. *Biometrika*, 62, 694-697.
3. Chang, H.J. and Huang, K.C. (2000). Remainder linear systematic sampling. *Sankhya*, B, 62, 376-384.
4. Cochran, W.G. (1977). *Sampling Techniques*, 3rd Edition, John Wiley & Sons.
5. Fountain, R.L. and Pathak, P.L. (1989). Systematic and non-random sampling in the presence of linear trends. *Communications in Statistics – Theory and Methods*, 18, 2511-2526.
6. Horvitz, D.G. and Thompson, D.J. (1952). A generalization of sampling without replacement from finite universe. *Journal of the American Statistical Association*, 47, 663-685.

Estimation of Population Proportion in Remainder Systematic Sampling

7. Khan, Z., Gupta, S. and Shabbir, J. (2014). A note on diagonal circular systematic sampling. *Journal of Statistical Theory and Practice*, 8, 439-443.
8. Khan, Z., Shabbir, J. and Gupta, S. (2013). A new sampling design for systematic sampling. *Communications in Statistics – Theory and Methods*, 42(18), 3359-3370.
9. Khan, Z., Shabbir, J. and Gupta, S. (2015). Generalized systematic sampling. *Communications in Statistics – Simulation and Computation*, 44, 2240-2250.
10. Lahiri, D.B. (1951). A method for selection providing unbiased estimates. *International Statistical Association Bulletin*, 33, 133-140.
11. Madow, W.G. (1953). On the theory of systematic sampling III-comparison of centered and random start systematic sampling. *The Annals of Mathematical Statistics*, 24, 101-106.
12. Madow, W.G. and Madow, L.H. (1944). On the theory of systematic sampling. I. *Annals of Mathematical Statistics*, 25, 1-24.
13. Naidoo, L.R., North, D., Zewotir, T. and Arnab, R. (2015). Balanced modified systematic sampling in the presence of linear trend. *South African Statistical Journal*, 49, 187-203.
14. Naidoo, L.R., North, D., Zewotir, T. and Arnab, R. (2018). Remainder modified systematic sampling in the presence of linear trend. *Communications in Statistics – Theory and Methods*, 47(10), 2469-2481. DOI: 10.1080/03610926.2017.1295076.
15. Sampath, S. and Varalakshmi, V. (2008). Diagonal circular systematic sampling. *Model Assisted Statistics and Applications*, 3(4), 345-352.
16. Sen, A.R. (1953). On the estimate of variance in sampling with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, 5, 119-127.
17. Subramani, J. (2000). Diagonal systematic sampling scheme for finite populations. *Journal of Indian Society of Agricultural Statistics*, 53(2), 187-195.
18. Subramani, J. (2009). Further results on diagonal systematic sampling for finite populations. *Journal of the Indian Society of Agricultural Statistics*, 63(3), 277-282.
19. Subramani, J. (2012). A modification on linear systematic sampling for odd sample size. *Bonfring International Journal of Data Mining*, 2(2), 32-36.
20. Subramani, J. (2013). A modification on linear systematic sampling. *Model Assisted Statistics and Applications*, 8(3), 215-227.
21. Subramani, J. (2018). On circular systematic sampling in the presence of linear trend. *Biometrics and Biostatistics International Journal*, 7(4), 2018.

22. Subramani, J. and Gupta, S.N. (2014). Generalized modified linear systematic sampling scheme for finite populations. *Hacettepe Journal of Mathematics and Statistics*, 43(3), 529-542.
23. Yates, F. (1948). Systematic sampling. *Philosophical Transactions of the Royal Society*, A 241, 345-377.
24. 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, 153-161

APPENDIX

Table 3: Variances of the sample proportion under different sampling schemes for $P = 0.30$

k	n	r	$Var(p_{SRS})$	$Var(p_{sy})$	$Var(p_{rsy})$
10	5	2	0.0409	0.0551	0.0369
		3	0.0448	0.1049	0.0381
		4	0.0317	0.0307	0.0270
	10	2	0.0162	0.0093	0.0083
		5	0.0196	0.0338	0.0166
		8	0.0163	0.0312	0.0104
	15	4	0.0141	0.0109	0.0106
		8	0.0131	0.0148	0.0077
		12	0.0145	0.0190	0.0072
100	50	10	0.0047	0.0043	0.0042
		20	0.0046	0.0042	0.0039
		30	0.0047	0.0052	0.0041
	100	30	0.0023	0.0020	0.0016
		50	0.0023	0.0019	0.0018
		70	0.0023	0.0022	0.0021
	150	40	0.0014	0.0014	0.0013
		80	0.0014	0.0014	0.0013
		120	0.0014	0.0012	0.0011
500	200	50	0.0010	0.0011	0.0010
		100	0.0011	0.0011	0.0010
		150	0.0010	0.0010	0.0009
	500	200	0.0005	0.0004	0.0004
		300	0.0004	0.0004	0.0004
		400	0.0004	0.0004	0.0004
	800	200	0.0003	0.0003	0.0003
		400	0.0003	0.0003	0.0003
		600	0.0003	0.0003	0.0003

Estimation of Population Proportion in Remainder Systematic Sampling

Table 4: Variances of the sample proportion under different sampling schemes for $P = 0.50$

k	n	r	$Var(p_{SRS})$	$Var(p_{sy})$	$Var(p_{rsy})$
10	5	2	0.0461	0.0284	0.0265
		3	0.0425	0.0662	0.0374
		4	0.0451	0.0427	0.0242
	10	2	0.0227	0.0227	0.0218
		5	0.0256	0.0204	0.0194
		8	0.0240	0.0418	0.0142
	15	4	0.0166	0.0172	0.0143
		8	0.0134	0.0099	0.0063
		12	0.0165	0.0117	0.0105
100	50	10	0.0049	0.0049	0.0044
		20	0.0053	0.0059	0.0043
		30	0.0052	0.0050	0.0049
	100	30	0.0027	0.0026	0.0024
		50	0.0023	0.0028	0.0023
		70	0.0024	0.0023	0.0022
	150	40	0.0015	0.0016	0.0015
		80	0.0015	0.0013	0.0013
		120	0.0016	0.0017	0.0014
500	200	50	0.0012	0.0011	0.0011
		100	0.0013	0.0014	0.0013
		150	0.0013	0.0013	0.0012
	500	200	0.0005	0.0005	0.0005
		300	0.0005	0.0005	0.0005
		400	0.0005	0.0005	0.0005
	800	200	0.0003	0.0003	0.0003
		400	0.0003	0.0003	0.0003
		600	0.0003	0.0003	0.0003

Table 5: Variances of the sample proportion under different sampling schemes for $P = 0.70$

k	n	r	$Var(p_{SRS})$	$Var(p_{sy})$	$Var(p_{rsy})$
10	5	2	0.0337	0.0338	0.0272
		3	0.0399	0.0373	0.0327
		4	0.0410	0.0467	0.0363
	10	2	0.0168	0.0173	0.0148
		5	0.0163	0.0160	0.0072
		8	0.0182	0.0232	0.0160
	15	4	0.0119	0.0166	0.0099
		8	0.0140	0.0198	0.0106
		12	0.0126	0.0143	0.0082
100	50	10	0.0040	0.0039	0.0036
		20	0.0042	0.0039	0.0037
		30	0.0040	0.0040	0.0036
	100	30	0.0021	0.0018	0.0018
		50	0.0021	0.0023	0.0021
		70	0.0020	0.0021	0.0020
	150	40	0.0015	0.0012	0.0012
		80	0.0015	0.0015	0.0014
		120	0.0014	0.0014	0.0012
500	200	50	0.0011	0.0011	0.0010
		100	0.0010	0.0011	0.0010
		150	0.0011	0.0011	0.0011
	500	200	0.0004	0.0005	0.0004
		300	0.0004	0.0004	0.0004
		400	0.0004	0.0004	0.0004
	800	200	0.0003	0.0003	0.0003
		400	0.0003	0.0003	0.0003
		600	0.0003	0.0003	0.0002