Indian Journal of Economics and Business
Vol. 21 No. 3(July, 2022)
Copyright@ Ashwin Anokha Publications & Distributions
http://www.ashwinanokha.com/IJEB.php

On the Estimation of Population Proportion in Systematic Sampling Schemes

1stZaheen Khan, 2nd Sher Akbar & 3rd Muhammad Irfan

[1] *Department of Statistics, Federal Urdu University of Arts, Sciences and Technology, Islamabad, Pakistan*
[2] *Department of Management Sciences, COMSATS University, Islamabad Pakistan.*
[3] *Department of Economics, COMSATS University, Islamabad Pakistan.*

* *Corresponding author:* zkurdu@gmail.com

Received: 15st May 2022
Revised: 20th June 2022
Accepted: 30th June 2022

Abstract: In systematic sampling, estimation of population mean or total widely discussed by the researchers and practitioners. However, estimation of population proportion has not been a matter of interest in systematic sampling. Recently, some researchers considered the estimation of population proportion in case of systematic sampling schemes. However, the results carried out by the researches in their study looks ambiguous. Therefore, in this study, we focus the estimation of population proportion in case of systematic sampling schemes. After derivation of variance expressions of sample proportion, numerical comparisons of these variances have been carried out. These comparisons are made among simple random sampling, linear systematic sampling and diagonal systematic sampling for the case where population size N is multiple of sample size n and among simple random sampling, circular systematic sampling and remainder linear systematic sampling for the case where population size N is not multiple of sample size n. The study revealed that estimation of population proportion in systematic sampling schemes provides almost similar results as we have in simple random sampling.

Keywords: Proportion, Variance, Efficiency, Comparisons.

1. Introduction

Systematic sampling is widely used design to estimate the population mean and total of quantitative variable. If it is properly organized, it shows higher efficiency over simple random sampling (SRS). However, in the literature, qualitative variable (binomial variable) like male or female, alive or die, true or false, purchased or not purchased, corona positive or negative etc. has not been discussed in systematic sampling. One can verify it from the numerous famous books of survey sampling written by (Deming, 1950), (Murthy,1967), (Cochran, 1977), (Singh and Chaudhary, 1986), (Sӓrndal, Swenssonand Wretman , 1992), (Singh and Mangat, 1986) and (Mukhopadhyay, 2009) with the names; Some Theory of Sampling, Sampling Theory and Methods, Sampling Techniques, Theory and Analysis of Sample Survey Designs, Model Assisted Survey

Sampling, Elements of Survey Sampling and Theory and Methods of Survey Sampling respectively. Furthermore, (Madow, 1953), (Sethi, 1965), (Singh and Singh, 1977), (Subramani, 2000), (Chang and Huang,2000), (Subramani,2009),(Khan ,Shabbirand Gupta , 2013), (Khan and Shabbir,2015), (Naidoo, North and Zewotir, 2016) and (Khan, Shabbir and Gupta, 2020) who proposed, respectively, Centered Systematic Sampling (CESS), Balanced Systematic Sampling (BSS), New Systematic Sampling (NSS), Diagonal Systematic Sampling (DSS),remainder linear systematic sampling (RLSS), Generalized Diagonal Systematic Sampling (GDSS), Modified Systematic Sampling(MSS), Generalized Systematic Sampling (GSS) , Multiple-Start Balanced Modified Systematic Sampling in the presence of linear trend (MBMSS) and An optimal systematic sampling scheme(OSS) also limit their discussion only to quantitative variable.

Systematic sampling is beneficial only if the variable used for ordering is as close to monotonically (increasing or decreasing) related to the survey variable y as possible. However, such relationship may not be achieved for qualitative variable in case of systematic sampling. The efficiency of the estimator for proportion will not show such improvement as we have in case of estimating the mean in systematic sampling compared to the simple random sampling and the results will be almost alike.

Recently, a manuscript published by Azeem (2021) on estimation of proportion in systematic sampling. The author adopted diagonal systematic sampling (DSS) proposed by Subramani (2000) and found that DSS more efficient than SRS and linear systematic sampling (LSS) while comparing estimator of proportion. Similarly, the author in another paper Azeemet.al(2022) adopted remainder linear systematic sampling (RLSS) proposed by Chang and Huang,(2000), and compare the efficiency of estimator of proportion with SRS and linear systematic sampling (LSS). Author, concluded that RLSS is a better choice for estimating proportion. However, this is not the case and the detail reasoning can be seen in the following sections:

2. Method

Suppose a finite population $\mathrm{U} = \{1, 2, 3\,, \dots, \mathrm{N}\}$ consist of N units numbered from 1 to N and the corresponding values of units for a given characteristic are denoted by y_1,y_2,y_3. . . y_N. Each value of $y_i(1, 2, 3, \dots, N)$ may be classify into two mutually exclusive classes. i.e. C and C', where C refers to the class of interest. Mathematically

$$y_i = \begin{cases} 1 & \text{if } \; y_i \in C \\ 0 & \text{if } \; y_i \in C' \end{cases} \quad (1)$$

The purpose is to estimate the population proportion $P = \frac{A}{N}$ (where $A = \sum_{i=1}^{N} y_i$ is the total number of units that belongs to class C from a sample of size n.

In LSS the number of population units N is divided in to k samples such that $N = nk$, and a random number r is selected from the first k units and then every k^{th} units will be selected to get a sample of n units, mathematically LSS sample (i.e. S_r can be written as:

$$S_r = r, r + k, r + 2k, \dots, r + (n-1)k \qquad (2)$$

However, in DSS proposed by [11], the sample S_r can be written as:

$$S_r = \begin{cases} r, r + (k+1), \dots, r + (n-1)(k+1) if \quad r \le k - n + 1 \\ r, r + (k+1), r + 2(k+1), \dots, r + t(k+1) = (t+1)k.\, (t+1)k + 1, (t+2)k + 2, \dots, \\ (n-1)k + (n-t-1) \qquad\qquad if \quad r > k - n + 1 \end{cases} \quad (3)$$

If population size *N* is not a multiple of the sample size *n* i.e. $N \neq nk$, we may not use LSS or DSS to get a sample of size *n*; instead, we use circular systematic sampling (CSS) or RLSS. In CSS all *N* units of population will be arranged around a circle and a unit *r* is selected at random from all *N* units and then every k^{th} unit is selected to get a sample of size *n*, mathematically;

$$S_r = \begin{cases} r + (j-1)k & r + (j+1)k \le N \\ r + (j-1)k - N & r + (j+1)k > N \end{cases} \qquad where\ j = 1, 2, 3, \dots, n. \quad (4)$$

In RLSS, the N population units be written as $N = nk + d = (n-d)k + d(k+1)$, where d is the remainder. In this scheme, we split total number of population units into two groups. The first group consist of first $(n-d)k$units and the second group consist of remaining $d(k+1)$units. A random number $r_1$is selected from the first k units of the first group and then every k^{th} unit thereafter is selected to get a sample $S_{r_1}$of $(n-d)$ units. Similarly, a random number r_2 is selected from the first (*k*+1) units of the second group and then every $(k+1)^{th}$ unit thereafter is selected to get a sample $S_{r_2}$of *d* units. A sample of size *n* can be obtain by combining the $(n-d)$selected units of first group and *d* selected units of second group i.e. $S_{r_{12}}{=}S_{r_1}, S_{r_2}$; mathematically:

2.1. Estimation

The sample proportion is given by:

$$p_t = \frac{a_t}{n} \quad (5)$$

where$a_t = \sum_{i \epsilon s} y_i$ and ▯▯= SRS, LSS, DSS, CSS and RLSS.

Note that, in case of RLSS

$$p_{RLSS} = \frac{(n-d)k}{N} p_{r_1} + \frac{d(k+1)}{N} p_{r_2} \quad (6)$$

wherep_{r_1} and p_{r_2} is the proportion of units that belong to class C of the sub-samples S_{r_1} and S_{r_2} respectively.

$$E(p_i) = P_i \qquad i = 1, 2 \quad (7)$$

Where $P_1$and P_1 are the proportions of units that belong to class C in the first and second group respectively. Now taking expectation of Equation (5), we have

$$E(p_t) = E(p_{RLSS}) = P \quad (8)$$

p_t and p_{RLSS} are unbiased estimator of P.

The variance expressions of SRS, LSS, DSS, CSS and RLSS are given by:

$$V(p_{SRS}) = \frac{(N-n)}{(N-1)n} P(1-P) \quad (9)$$

$$V(p_{LSS}) = \frac{1}{k}\sum_{i=1}^{k}(p_i - P)^2 \quad (10)$$

and

$$V(p_{DSS}) = \frac{1}{k}\sum_{i=1}^{k}(p_i - P)^2 \quad (11)$$

The variance expression for LSS and DSS looks similar but arrangement of units are different, so the results will be different. The variance of CSS is given by:

$$V(p_{CSS}) = \frac{1}{N}\sum_{i=1}^{N}(p_i - P)^2 \quad (12)$$

wherep_i is the proportion of ith sample from N possible samples.

The variance of RLSS is given by:

$$V(p_{RLSS}) = \frac{1}{N^2}\left[((n-d)k)^2 V(p_{r_{1)}} + (d(k+1))^2 V(p_{r_{2)}}\right] \quad (13)$$

where $V(p_{r_1}) = \frac{1}{k}\sum_{i=1}^{k}(p_i - P_1)^2$

and $V(p_{r_1}) = \frac{1}{k+1}\sum_{i=1}^{k+1}(p_i - P_2).^2$

3. Findings and Discussions

In this section, initially, the comparisons have been made for the variances of SRS, LSS and DSS for the case where population size is multiple of sample size .i.e.$N = nk$. Here, we considered the same data which is already used by Azeem (2021). The data is collected from 250 students of social sciences at the University of Malakand. The authors in their paper mentioned that "the respondents were asked to report a yes answer if he/she is aware of the adverse effects of carbonated drinks otherwise report 'no'. In the data the value '1' is assigned to the response 'yes' and '0' otherwise. The authors compared the variances of SRS, LSS and DSS and concluded that DSS is the most suitable sampling design for estimation of proportion. However, it is not true; for verification, one can see the following detail:

In this study, the total numbers of respondents who respond 'yes' are 194 .i.e. $A = \sum_{i=1}^{N} y_i = 194$, thus$P = \frac{194}{250} = 0.776$. If we take a sample of size n = 5 and using Equation (9) the

$V(p_{SRS}) = \frac{(250-5)}{(250-1)5}0.776(1-0.776) = 0.0342$, Similarly, if n = 10,$V(p_{SRS}) = \frac{(250-10)}{(250-1)10}0.776(1-0.776) = 0.0168$

However, the results of variances mentioned in Azeem (2021) are 0.0392 and 0.0224 for the sample sizes 5 and 10 respectively. Clearly, the results given by Azeem (2021) look erroneous. Similarly, the whole comparison presented in Table.3 by Azeem (2021) is not correct. Thus, the findings based on this research cannot be generalized.

In this study, we find out the variances of proportion under SRS, LSS and DSS by taking the similar sizes of population and sample that are considered by Azeem (2021) and the results are presented in Table 1. Here, one can see that we cannot prefer any one of the sampling design over the other.

Another study proposed by Azeemetal (2022) based on the comparison of variances of SRS, LSS and RLSS. This study is conducted for the case where population size is not multiple of sample size i.e. $N \neq nk$. The authors numerically compared the variances for said schemes and the results are presented in Tables 3; 4 and 5. In this comparison LSS is compared with SRS and RLSS which is not possible; because, LSS is applicable only for those cases where . $N = nk$. However, we can use CSS instead of LSS for the case where $N \neq nk$. Moreover, the numerical results for SRS and RLSS presented in these tables are also looks not correct. Consequently, the conclusion based on this study is also worthless.

Here, we have made a simulation study in which the data is generated from Bernoulli distribution by fixing $P = 0.3$ for various population sizes. Variance of proportion is calculated for SRS, CSS and RLSS. This process is replicated 100 times and finally the average variance for each design is calculated and the results are presented in Table 2.

One can see in Table 2, that we cannot prefer any one scheme over the other by comparing the variances of sample proportion. Therefore, we can say that the systematic schemes behave like simple random sampling while dealing with the sample proportion as an estimator of the population proportion.

Table-1 Comparison of Variances of SRS, LSS and DSS

Sr. no	N	n	k	SRS	VARLSS	VARDSS
1	250	5	50	0.03421	0.03142	0.01862
2	250	10	25	0.01675	0.01862	0.01542
3	249	3	83	0.05764	0.06187	0.07258
4	248	4	62	0.04262	0.04457	0.04457
5	248	8	31	0.02096	0.01180	0.01785
6	246	3	82	0.05739	0.05164	0.06248
7	246	6	41	0.02834	0.03335	0.04419
8	245	5	49	0.03425	0.02879	0.03042
9	245	7	35	0.02426	0.02716	0.02249
10	243	9	27	0.01857	0.01463	0.01646
11	242	11	22	0.01511	0.02459	0.01632
12	240	3	80	0.05688	0.05540	0.05818
13	240	4	60	0.04248	0.03561	0.04186
14	240	5	48	0.03384	0.03873	0.03040
15	240	6	40	0.02808	0.02693	0.03526
16	240	8	30	0.02088	0.01530	0.01217
17	240	10	24	0.01656	0.01915	0.01582
18	240	12	20	0.01368	0.01269	0.02658
19	240	15	16	0.01080	0.00984	0.01262
20	238	7	34	0.02410	0.02424	0.02904
21	238	14	17	0.01169	0.00713	0.01434

Table-1 Comparison of Variances of SRS, LSS and DSS

Sr. No	N	n	k	d	SRS	CSS	RLSS
1	52	5	10	2	0.03857	0.03900	0.03769
2	53	5	10	3	0.03798	0.03647	0.04134
3	54	5	10	4	0.03726	0.03766	0.03850
4	102	10	10	2	0.01881	0.01824	0.01922
5	105	10	10	5	0.01906	0.01776	0.01883
6	108	10	10	8	0.01933	0.01979	0.01960
7	154	15	10	4	0.01270	0.01226	0.01223
8	158	15	10	8	0.01256	0.01270	0.01235
9	162	15	10	12	0.01273	0.01202	0.01331
10	5010	50	100	10	0.00417	0.00424	0.00429
11	5020	50	100	20	0.00416	0.00412	0.00411
12	5030	50	100	30	0.00416	0.00416	0.00418
13	10030	100	100	30	0.00208	0.00209	0.00213
14	10050	100	100	50	0.00207	0.00207	0.00207
15	10070	100	100	70	0.00207	0.00212	0.00211
16	15040	150	100	40	0.00139	0.00135	0.00141
17	15080	150	100	80	0.00139	0.00141	0.00140
18	15120	150	100	120	0.00139	0.00137	0.00138
19	100050	200	500	50	0.00105	0.00103	0.00105
20	100100	200	500	100	0.00104	0.00107	0.00105
21	100150	200	500	150	0.00105	0.00098	0.00096

4. Conclusion

In the proposed study, we considered the sample proportion as an estimator of the population proportion. The variances of sample proportion under different sampling scheme by taking various population and sample sizes have been analyzed. The results have been presented in Tables 1 and 2. From this study it is revealed that the suggestions to the researchers given by Azeem (2021) and Azeem et al (2022) on the basis of their study are not practicable. Moreover, we can say that the DSS in the case where $N = nk$ and RLSS in the case where $N \neq nk$ will not show any significant results over SRS or other systematic schemes as claimed by Azeem (2021) and Azeem et al(2022). The main reason of not getting improved results is that, the systematic sampling is beneficial only if the survey variable y is closely related to the variable used for ordering. However, such relationship may not be achieved for qualitative variable in case of systematic sampling. Therefore, estimation of proportion in systematic sampling will not show any improvement (high efficiency) as compared to the simple random sampling and the results will be almost alike. Finally, it is suggested that estimation of proportion using SRS will be the better option than the systematic schemes. As it is more flexible and almost give the similar results as we have in systematic schemes in case of estimating proportion.

References

Deming WE.(1950). Some Theory of Sampling. John Wiley and Sons.

Murthy MN. (1967). Sampling Theory and Methods. Statistical Publishing Society Calcutta, India.

Cochran WG. (1977). Sampling Techniques. 3rd ed. New Delhi, Wiley Eastern Limited.

Singh D & Chaudhary F. (1986).Theory and Analysis of Sample Survey Designs. John Wiley and Sons.

Sa¨rndal CE, Swensson B & Wretman J. (1992).Model Assisted Survey Sampling. Springer Verlag.

Singh S & Mangat NS.(1986) Elements of Survey Sampling. Kluwer Academic.

Mukhopadhyay P. (2009). Theory and Methods of Survey Sampling. PHI Learning Private Limited, New Delhi, India.

Madow WG. (1953) On the Theory of Systematic Sampling. The Annals of Mathematical Statistics, 24, 101-106.

Sethi VK. (1965). On Optimum Pairing of Units. Sankhya. 27 (B), 315-320.

Singh D& Singh P. (1977). New Systematic Sampling. Journal of Statistical Planning and Inference. 1, 163-177.

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

Chang H.J& Huang K. C. (2000). Remainder Linear Systematic Sampling. Sankhya. 62(B), 249-256.

Subramani J. (2009). Further Results on Diagonal Systematic Sampling Scheme for Finite Populations. Journal of the Indian Society of Agriculture Statistics. 63(3), 277-282.

Khan Z, Shabbir J& Gupta S N. (2013). A New Sampling Design for Systematic Sampling. Communications in Statistics - Theory and Methods. 42, 2659-2670.

Khan Z&Shabbir J. (2015). Modified Systematic Sampling in the Presence of Linear Trend. Hacettepe Journal of Mathematics and Statistics. 44(2), 417-442.

Naidoo LR, North D & Zewotir T (2016). Multiple-Start Balanced Modified Systematic Sampling in the Presence of Linear Trend. Communications in Statistics - Theory and Methods. 45(14), 4307-4324.

Khan Z, Shabbir J, Gupta S N (2020). An Optimal Systematic Sampling Scheme. Journal of Statistical Computation and Simulation. 90(11), 2023-2036.

Azeem M. (2021). On Estimation of Population Proportion in Diagonal Systematic Sampling. Life Cycle Reliability and Safety Engineering. 10, 249-254.

Azeem M, Khan Z&Khattak SW. (2022).Estimation of Population Proportion in Remainder Systematic Sampling. Indian Journal of Economics and Business. 21(2), 217-227.