Journal of Statistics Applications & Probability
An International Journal

http://dx.doi.org/10.18576/jsap/090302

Implementing Logarithmic Type Estimation Procedure For Population Mean in Successive Sampling

P. Mukhopadhyay[1,*], *A. Bandyopadhyay*[1] *and G. N. Singh*[2]

[1]Department of Mathematics, Asansol Engineering College, Asansol-713305, India
[2]Department of Applied Mathematics, Indian School of Mines, Dhanbad-826004, India.

Received: 25 Sep. 2019, Revised: 11 Oct. 2020, Accepted: 17 Oct. 2020
Published online: 1 Nov. 2020

Abstract: Bahl, S. and Tuteja R.K. (1991) [1] introduced exponential structure which earned considerable success among the survey statisticians. The present paper, getting inspiration from that, attempts to introduce a new chain type ratio estimator using logarithmic function structure. The estimator is applied in successive sampling situation where complete response is presumed at both occasions. We have studied the properties of the suggested estimator and also derived its optimum condition. The strength of the proposed estimator over the conventional ones has been discussed, through numerical illustrations, taking into account four standard data sets from various situations. Encouraged with the outcomes of the proposed estimator and strategy, it has been suggested to the survey statisticians for future use.

Keywords: logarithmic estimator, successive sampling, linear convex combination

1 Introduction

Sampling theory has grown steadily in tandem with various sampling methods considered suitable in diverse practical situations. The commonest method of drawing a sample from a population is the Simple Random Sampling (SRS) method from which many other methods have been emanated. Keeping SRS method in mind, numerous new estimators have been designed in the literature on estimators, namely ratio, product, dual, regression and their variants. The exponential estimators are the newer addition to this list. In this paper, we have tried even another new estimator in the logarithmic form and investigated its dexterity over others. Our estimator makes use of auxiliary variables which was first utilized in the ratio estimator proposed [2]. Then [1] proposed new ratio and product type exponential estimators in SRS for the estimation of the mean of the study variable Y, depending on an X auxiliary variable whose population mean is known.

Following [1], [3] proposed new ratio and product type exponential estimators and family of estimators for proportion estimation in SRS inspired by ratio and product estimators proposed in SRS.

The problem of sampling on two successive occasions with a partial replacement of sampling units was introduced by [4]. He used the entire information collected during past surveys to make the current estimation more accurate. After Jessen, the theory and methodologies on successive sampling (synonymously known as matched sampling or rotation sampling) were supplemented by [5], [6], [7], [8].

Here we list a few of those various areas where successive sampling is particularly in order :

(a) **Socio-Economic and Agricultural Census:** e.g., assessing the unemployment status at different points of time along with the pattern of change in unemployment status over a period of time.

* Corresponding author e-mail: parthamukhopadhyay1967@gmail.com

(c) **Measurement of Social Hazards Due to Technological Advents:** e.g., to know the pollution level due to polluting industries and vehicles at different points of time and to know the pattern of change in pollution level over period of time.

NB: Successive sampling is particularly suitable in environmental studies e.g estimating the number of members in species on the verge of extinction (E. Artes, M. Rueda et al, Transactions on Ecology and the Environment, vol 23, 1998)

(d) **Marketing and Business Forecast:** e.g, to know the demand of a product at different points of time and to know the changing pattern in demand over a period of time.

Enthused by the above researches all the way, we now put forward the logarithmic ratio type estimator structure. We have examined the performance of the proposed strategy through data set from natural population. The findings are recommended to the survey statisticians for their application in real life.

2 The Class of Estimators and Its Formulation

To begin with, we offer another modified logarithmic ratio structure as follows:

$$\alpha\left[\bar{y}+\log\left(1+\beta\frac{\bar{X}-\bar{x}}{(\bar{X}+\bar{x})^{\beta}}\right)\right]. \tag{1}$$

The structure is consistent for if we put $\bar{x}=\bar{X}$ we have $T=\bar{Y}$.

Since we must always have positive quantity within log() function, we further require

$$1+\beta_i\frac{\bar{X}-\bar{x}}{(\bar{X}+\bar{x})^{\beta_i}}>0. \tag{2}$$

3.1 Basic Sampling Structure and Notations Followed:

When a population is changing with time, a survey carried out on a single occasion (i.e., point of time) will provide information about the characteristic of the surveyed population for the given occasion only, and it is definitely inadequate to measure the rate of change of the characteristics over time as well as the average value of the parameter for the most recent occasion (current occasion). To meet these requirements, successive sampling is ideal. Let the population of N units be $U=(U_1,U_2,\cdots,U_N)$. In this scheme, a simple random sample S of n units is selected on the first occasion without replacement, and then on the second occasion retain a SRS S_1 of $m(<n)$ units, and a fresh SRS S_2 of $(u=n-m)$ units is selected independently from the population without replacement and added to S_1. Usually $u<m<n$ is taken to minimize cost.

We define $x(y)$ as 1st (2nd or current) occasion study variables. Where $\bar{X}$ is known, and we are to estimate $\bar{Y}$ of population.

We use the following notations in the subsequent work:

– $\bar{X},\bar{Y}$: population mean of x, y, respectively.
– $\bar{x}_n(\bar{z}_n);\bar{y}_n(\bar{x}_m);\bar{y}_u(\bar{x}_u)$: sample means of the respective subscripted variables with sample size mentioned.
– ρ_{yx}: correlation coefficient between the subscripted variables.
– C_x,C_y: coefficient of variation of the subscripted variables.

Hence, in view of the above discussion and remembering the structure defined in equation (1), we suitably choose three pairs $\alpha_i,\beta_i(i=1,2,3)$, where we take $\beta_i=1,\beta_i=0,\beta_i=-1$ to construct the following linear combination of three different estimators to get the final estimator as:

$$T_p=\sum_{i=1}^{3}T_{p_i}$$

$= \alpha_1 \left[\bar{y}_m + \log\left(1 + \frac{\bar{x}_n - \bar{x}_m}{\bar{x}_n + \bar{x}_m}\right)\right] + \alpha_2 \bar{y}_m + \alpha_3 \left[\bar{y}_m + \log\left(1 - \frac{\bar{x}_n - \bar{x}_m}{(\bar{x}_n + \bar{x}_m)^{-1}}\right)\right]$

$= \bar{y}_m(\alpha_1 + \alpha_2 + \alpha_3) + \alpha_1 \frac{e_2 - e_1}{2\left(1 + \frac{e_2 + e_1}{2}\right)} - \alpha_3 \frac{\bar{X}(e_2 - e_1)}{\bar{X}^{-1}(2 + e_2 + e_1)^{-1}}.$

$$= \bar{Y}(1 + e_0)(\alpha_1 + \alpha_2 + \alpha_3) + \left[\frac{\alpha_1}{2} - 2\alpha_3 \bar{X}^2\right](e_2^2 - e_1^2). \tag{3}$$

For consistency of T_p, we must have

$$\alpha_1 + \alpha_2 + \alpha_3 = 1. \tag{4}$$

3.2 The Final Expression of Proposed Estimators

The error equations in accordance with sampling structure described above (section 2.1) are

$\bar{y}_m = \bar{Y}(1 + e_0), \bar{x}_m = \bar{X}(1 + e_1), \bar{x}_n = \bar{X}(1 + e_2)$, and the fpc's are

$f_1 = \frac{1}{m} - \frac{1}{N}; f_2 = \frac{1}{n} - \frac{1}{N}; f_1 - f_2 = \frac{1}{m} - \frac{1}{n} > 0$, and

$$E(e_0^2) = f_1 C_Y^2, E(e_1^2) = f_1 C_X^2, E(e_2^2) = f_2 C_X^2, E(e_0 e_1) = f_1 \rho_{xy} C_Y C_X, E(e_1 e_2) = f_2 C_X^2. \tag{5}$$

When second occasion sample,(of size m) is retained (called matched portion) from the first occasion sample(of size n) is already taken, and there are some correlation between errors of estimation of means for the two samples: second and first. Hence, we have

$$E(e_0 e_2) = f_2 \rho_{xy} C_Y C_X \tag{6}$$

Replacing $\bar{y}_m, \bar{x}_m, \bar{x}_n$ in (3) with the corresponding error terms in (5) and (6), we get the required expression of the estimator T_p as follows:

$$= \bar{Y}(1 + e_0)(\alpha_1 + \alpha_2 + \alpha_3) + \left[\frac{\alpha_1}{2} - 2\alpha_3 \bar{X}^2\right](e_2^2 - e_1^2). \tag{7}$$

4 Obtaining Bias and Mean Square Error for the Proposed Estimator

The structure of the estimator T_p derived in section 2.1 (3) is used next to calculate the mean square error expression. We then have

$E(T_p - \bar{Y})^2 = E\left[\bar{Y}^2 e_0^2 + 2\bar{Y}(e_0 e_2 - e_0 e_1)\left(\frac{\alpha_1}{2} - 2\alpha_3 \bar{X}^2\right) + (e_2^2 + e_1^2 - 2e_1 e_2)\left(\frac{\alpha_1}{2} - 2\alpha_3 \bar{X}^2\right)^2\right].$

(after ignoring error terms of order higher than square)

Now, requiring the estimator to be unbiased, we have

$$\left[\frac{\alpha_1}{4} + \alpha_3 \bar{X}^2\right](e_2^2 - e_1^2) = 0 \implies \frac{\alpha_1}{4} + \alpha_3 \bar{X}^2 = 0. \tag{8}$$

On letting $\frac{\alpha_1}{2} - 2\alpha_3 \bar{X}^2 = \tau$, we have

$$M(T_p) = \bar{Y}^2 f_1 C_Y^2 + 2\bar{Y}(f_1 - f_2)\tau C_Y C_X \rho_{yx} + C_X^2 (f_1 - f_2)\tau^2. \tag{9}$$

4.1 Obtaining the Minimum MSE of the Estimator

From section 3 equation (9), we have the expression of $M(T_p)$ as

$$M(T_p) = \bar{Y}^2 f_1 C_Y^2 + 2\bar{Y}(f_1 - f_2)\tau C_Y C_X \rho_{yx} + C_X^2 (f_1 - f_2)\tau^2$$

, which is a function of parameter τ. Hence, we minimize it w. r. t τ and get

$$\frac{\delta M}{\delta \tau} = 0 \implies \tau = \frac{\alpha_1}{2} - 2\alpha_3 \bar{X}^2 = \frac{\bar{Y} C_Y \rho_{yx}}{C_X}. \tag{10}$$

Solving (8), (10) and using relation (4), we get

$$\alpha_1 = \frac{\bar{Y} C_Y \rho_{yx}}{C_X}, \alpha_3 = -\frac{\bar{Y} C_Y \rho_{yx}}{4\bar{X}^2 C_X}, \alpha_2 = 1 - \frac{\bar{Y} C_Y \rho_{yx}}{C_X} + \frac{\bar{Y} C_Y \rho_{yx}}{4\bar{X}^2 C_X}.$$

Putting back the values of α_1, α_2 and α_3 in $M(T_p)$, we finally obtain optimized expression of $M(T_p)$ as

$$\begin{aligned} M(T_p)_{opt} &= \bar{Y}^2 f_1 C_Y^2 - 2\bar{Y}(f_1 - f_2) C_Y C_X \rho_{yx} \tfrac{\bar{Y} C_Y \rho_{yx}}{C_X} + \tfrac{C_X^2 (f_1 - f_2)}{4} \tfrac{4\bar{Y}^2 C_Y^2 \rho_{yx}^2}{C_X^2} \\ &= \bar{Y}^2 f_1 C_Y^2 - \bar{Y}^2 C_Y^2 \rho_{yx}^2 (f_1 - f_2) < M(\bar{y}_m) \end{aligned}. \tag{11}$$

Thus, the new estimator is always better than the sample mean.

Further, we compare the estimator obtained with the regression estimator.

$$M(\bar{y}_{reg}) - M(T_p)_{opt} = \bar{Y}^2 f_1 (1 - \rho_{yx}^2) C_Y^2 - \bar{Y}^2 f_1 C_Y^2 + \bar{Y}^2 C_Y^2 \rho_{yx}^2 (f_1 - f_2) = -\bar{Y}^2 C_Y^2 \rho_{yx}^2 f_2 < 0.$$

Thus, the estimator proposed is not more efficient than the regression estimator.

5 Combining $(T_p)_{opt}$ With T_u (Fresh Sample From Population):

Next, we construct, combining the above estimator T_p with T_u, to get our final estimator as

$$T_{pm} = \phi T_u + (1 - \phi) T_p (0 < \phi < 1), \text{where} T_u = \bar{y}_u + \log\left(1 + \frac{\bar{X} - \bar{x}_u}{\bar{X} + \bar{x}_u}\right). \tag{12}$$

Here, we take $\alpha = 1, \beta = 1$ for keeping calculation simple.

Suppose in initial sample, n units are chosen, and in the main sample m out of n are chosen. Finally, u amount of fresh members are sampled from population. Also $n = m + u$. We have two new error equations here as:

$$\bar{y}_u = \bar{Y}(1 + e_4), \bar{x}_u = \bar{X}(1 + e_5), f_3 = \tfrac{1}{u} - \tfrac{1}{N} \text{ and } T_u = \bar{y}_u + \log\left(1 + \tfrac{\bar{X} - \bar{x}_u}{\bar{X} + \bar{x}_u}\right) = \bar{Y}(1 + e_4) + \tfrac{1}{2}(-\tfrac{e_5}{2} + \tfrac{e_5^2}{4}).$$

5.1 Minimum MSE Criterion for T_u

$$M(T_u) = E(T_u - \bar{Y})^2 = \bar{Y}^2 f_3 C_Y^2 - \tfrac{1}{2}\bar{Y} E(e_4 e_5) + \tfrac{E(e_5^2)}{16}$$

$$= \bar{Y}^2 f_3 C_Y^2 - \tfrac{1}{2}\bar{Y} f_3 C_Y C_X \rho_{yx} + \tfrac{1}{16} f_3 C_X^2.$$

5.2 Optimizing the Combined Estimator T_{pm} (linear convex combination of $T_{p(opt)}$ and $T_{u(opt)}$)

Now we proceed to construct T_{pm} to estimate $\bar{y}$ in the successive occasion, according to Hansen Hurwitz strategy, and express the optimized T_{pm} as under:

$$T_{pm} = \phi T_{m_{opt}} + (1-\phi)T_{u_{opt}}$$

$$M(T_{pm}) = E(T_{pm} - \bar{Y})^2 = E[\phi T_{p_{opt}} + (1-\phi)T_{u_{opt}} - \bar{Y}]^2$$

$$= E[\phi T_{p_{opt}} + T_{u_{opt}} - \phi T_{u_{opt}} - \phi\bar{Y} - (1-\phi)T_{u_{opt}} - barY]^2$$

$$= E[\phi(T_{p_{opt}} - \bar{Y}) + (1-\phi)(T_{u_{opt}} - \bar{Y})]^2$$

$$= \phi^2 M(T_{p_{opt}}) + (1-\phi)^2 M(T_{u_{opt}}) \tag{13}$$

$$\frac{\delta M(T_{pm})}{\delta\phi} = 0 \implies \frac{M(T_{p_{opt}})}{M(T_{u_{opt}})} = \frac{1}{\phi} - 1 \implies \phi = \frac{M(T_{u_{opt}})}{M(T_{p_{opt}}) + M(T_{u_{opt}})}$$

Putting back the value of ϕ in (13), we get:

$$M(T_{pm_{opt}}) = \frac{M(T_{p_{opt}})M(T_{u_{opt}})}{M(T_{p_{opt}}) + M(T_{u_{opt}})} = \frac{[\bar{Y}^2 f_1 C_y^2 - \bar{Y}^2 C_Y^2 \rho_{yx}^2 (f_1 - f_2)][\bar{Y}^2 f_3 C_Y^2 - \frac{1}{2}\bar{Y} f_3 C_Y C_X \rho_{yx} + \frac{1}{16} f_3 C_X^2]}{\bar{Y}^2 f_1 C_y^2 - \bar{Y}^2 C_Y^2 \rho_{yx}^2 (f_1 - f_2) + \bar{Y}^2 f_3 C_Y^2 - \frac{1}{2}\bar{Y} f_3 C_Y C_X \rho_{yx} + \frac{1}{16} f_3 C_X^2} \tag{14}$$

Please note that $T_{p_{opt}}$ and $T_{u_{opt}}$ are already optimized w.r.t α and β, and $T_{pm_{opt}}$ is optimized w.r.t ϕ.

6 Verification of the Efficacy of Proposed Estimator:

The performance of new logarithmic estimator has been compared against some of the existing ones mentioned in the introduction, and the results obtained are furnished. Below we list the MSE expressions of a number of existing estimators:

$V(\bar{y}_m) = f_1 \bar{Y}^2 C_Y^2$: sample mean which is shown to be $> T$ always (vide section 4.1 eqn. (12))

$MSE(T_J) = \frac{V(T_u)V(T_n)}{V(T_u)+V(T_n)}$, where $V(T_u) = f_u \bar{Y}^2 C_Y^2$ and $V(T_n) = f_{mn} C_Y^2 (1-\rho_{yx}^2)$ (Jessen 1942 [4])

$MSE(T_{SV}) = \bar{Y}^2 C_Y^2 \left[\left(\frac{1}{m} - \frac{1}{n}\right)(1-\rho_{yx}^2) + \frac{1}{n}(1-\rho_{yz}^2)\right]$: Singh Vishwakarma, 2007 [[9], [10]]

The PRE of any estimator T with respect to sample mean estimator is defined as:

$$PRE = \frac{V(\bar{y})}{M(T)} \times 100$$

The PRE of the optimized successive estimator is computed and compared against the existing estimators for a number of natural population data sets described below:

Population I-Source: Sukhatme and Sukhatme (1970)

y: Area (acres) under wheat in 1937.
x: Area (acres) under wheat in 1936.
z: Total cultivated area (acres) in 1931.

Population II-Source: Murthy (1967)

y: Area under wheat in 1964 ,
x: Area under wheat in 1963 , and

Table 1: Parametric Values of Populations Furnished Above

Data sets	Y	X	Z	N	n	m	u	ρ_{yx}	ρ_{yz}	ρ_{xz}	C_y	C_x	C_z
Sukhatme and Sukhatme (1970)	201.41	218.41	765.35	34	15	8	7	0.93	0.83	0.9	0.74	0.76	0.61
Murthy(1967)	5182.6	5182.6	1126.5	80	30	15	15	0.91	0.99	0.94	0.35	0.94	0.75
Literacy Rate Census (2011)	76.20	68.04	68.59	34	15	8	7	0.9791	0.989	0.988	0.609	0.618	0.178
Abortion Rate in USA	19.93	19.92	20.55	51	25	13	12	0.970	0.600	0.565	0.482	0.256	0.298

Table 2: Efficiency Comparison Between T_{sv} and $T_{pm(opt)}$ against sample mean

Data Set	T_{sv}	$T_{pm(opt)}$
Murthy (1967)	130.1671	207.8189028
Sukhatme and Sukhatme (1970)	325.8603	2011.169538
Literacy Rate: Census (2011)	989.26	2014.23
Abortion Rate in USA (2011)	73.63	146.22

Table 3: Efficiency Comparison Between T_J and $T_{pm(opt)}$ against sample mean

Data Set	T_J (Jessen 1942)	$T_{pm(opt)}$
Sukhatme and Sukhatme (1970)	402.04	2011.17
Literacy Rate: Census (2011)	1576.86	2014.22

Table 4: Efficiency Comparison Between T_{sv} and $T_{pm(opt)}$ against T_{reg}

Data Set	T_{sv}	$T_{pm(opt)}$	T_{reg}
Sukhatme and Sukhatme (1970)	173.25	1069.31	188.08
Literacy Rate: Census (2011)	284.15	578.55	348.15

z: Cultivated area in 1961.

Population III-Source: Census (2011) Source: Office of Registrar General, India, Ministry of Home Affairs

y, x, z : literacy rate of India (persons) during the years 2011, 2001 and the female literacy rate (2011), respectively.

Population IV-Source: Abortion Rate in USA: Statistical Abstract of the United States: 2011

y, x, z : number of abortions rates reported in the states of US during the years 2008, 2007 and 2005, respectively.

7 Conclusion

Tables 1,2,3 & 4 suggest that the proposed log estimator $T_{pm(opt)}$ is clearly more efficient than the estimators T_{sv} for all data sets w.r.t sample mean and for data sets I and III w. r. t. regression estimator. $T_{pm(opt)}$ is more efficient than the estimators T_J for data sets I and III w. r .t. regression estimator. The efficiency steadily increases as n increases, which is natural. The reason of high efficiency should be that the estimator is optimized at different levels (w.r.t optimizing parameters are α and β, and then w. r. t ψ is the linear combination parameter of $T_{m(opt)}$ and $T_{u(opt)}$). Moreover, the estimator of fresh sampling is also optimized separately.Further, it may be observed that the efficiency falls slowly with the reduction of ρ_{yz}, and falls rapidly with the reduction of ρ_{yx} and this feature is more or less the same for other estimators also.

Thus the estimator may be recommended for application in suitable situations.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this article

References

[1] S. Bahl and R. K Tuteja. Ratio and Product Type Exponential Estimator. *Information and Optimization Sciences*, 12(1), 159-163.
[2] C. Cochran. *Sampling Techniques*. John Wiley & Sons, New York, (1977).
[3] R. Singh and P. Chauhan and N. Sawan and F. Smarandache. Ratio-product Type Exponential Estimator for Estimating Finite Population Mean Using Information on Auxiliary Attribute.
[4] R. J. Jessen. Statistical investigation of a sample survey for obtaining form facts. *Iowa Agricultural Experiment*, Bulletin no. 304, Station Road, Ames, USA, (1942).
[5] F. Yates. *Sampling methods for censuses and surveys*. Charles Grin & Co. Ltd., London, (1949).
[6] H. D. Patterson. Sampling on successive occasions with partial replacement units. *J. Roy. Statistical Society*,12(B), 241-255, (1950).
[7] B. D. Tikkiwal. *Theory of successive sampling*, Thesis for Diploma, ICAR, New Delhi, (1951).
[8] P. Mukhopadhyay and A. Bandyopadhyay and G. N. Singh. Logarithmic Estimation Procedures for Population Mean in Two-Phase Sampling, *International Journal of Applied Mathematics & Statistics*, Nov. Issue, (2019).
[9] H. P. Singh and G. K. Vishwakarma. A general class of estimators in successive sampling. *Metron* , 65(2), 201?227, (2007).
[10] H. P. Singh and G. K. Vishwakarma. A general procedure for estimating population mean in successive sampling. *Communication in Statistics-Theory and Method*, 38, 293-308, (2009).

Partha Mukhopadhyay He is an Asst. Professor in the Department of Mathematics, Asansol Engineering College, Asansol. A postgraduate in Applied Mathematics from Jadavpur University, Kolkata. He has more than 10 years of teaching experience in Asansol Engineering College at B.Tech/M.Tech level. His research interest is in the field of inferential statistics, sampling theory.

Arnab Bandyopadhyay is an Assistant Professor of Mathematics in Asansol Engineering College, India. He has more than 10 years of teaching experience and more than 5 years of research experience. His field of research interest is statistical sample surveys. He has published several international journals of repute and currents dealing with two research projects funded by Govt. of India. He is guiding several researchers for pursuing their career for Ph. D. degree in mathematics/ statistics.

Garib Nath Singh is a Professor of Statistics in the Department of Mathematics & Computing at Indian Institute of Technology (Indian School of Mines), Dhanbad, India. His research areas are Sample Surveys and Data Analysis. He has 28 years of teaching experience and produced 12 Ph.Ds and more than 150 research papers which are published in various journals of repute. He has also visited USA, Sri Lanka and Australia and delivered lectures in various international conferences.