Special Proceedings: ISBN #: 978-81-950383-1-2
24th Annual Conference of SSCA, 23-27 February 2022; pp 107-113

An Application of Agrawal-Panda Type Post Stratified Estimates in Cluster Sampling with Application in Estimating Average Student Enrolment in a Class

Manish Trivedi[1] **and Purnendukisor Bandyopadhyay**[2]
[1]*School of Sciences, Indira Gandhi National Open University*
[2] *Department of Higher Education, Ministry of Education*

Received: 20 February 2022; Revised: 23 July 2022; Accepted: 23 July 2022

Abstract

The school education system in India cater to the students. Estimating the number of students in a typical class is essential for all types of policy planning, which target the students as recipient of a service. Due to a variety of reasons, it might not be possible to undertake a census of all the schools to have complete information about number of students. Therefore, resorting to sample surveys of schools in target areas can be a viable solution before launching of any programme. Due to paucity of data or incomplete frame information, it may not always be possible to stratify the schools in target areas before the survey. However, once the sample is selected, this information can be collected from the selected schools and post-stratified estimates can be developed from them. In this paper, an attempt has been made to develop an estimate using the Agrawal-Panda type methods of post stratified estimates of population mean on cluster sample using data from the student enrolments in 3 districts of one State for which recent data is available and it helps to derive the population parameters, including the variance and MSE of the proposed estimators.

Key words: Cluster sampling; Post stratification; Mean square error.

1. Introduction

Using stratified sampling for improving estimates of population parameters, where heterogeneity of population parameters is already known or anticipated, is a time-tested method. However, many-a-times, at the time of sample selection, the frame does not have enough data to divide the frame in appropriate strata. Sometimes, due to the nature of the variables under study, it becomes very difficult to use stratification before drawing of the samples, particularly, when the variables under study might be dependent on gender differentials, where the study variables might require stratification within nearly each household to be covered in the survey.

Post stratified sampling strategies are used in sample surveys in similar cases. Methodological developments on post-stratification in case of incomplete frame information has been studied since long (Holt and Smith,1979). This was followed by examination of the feasibility of ratio estimators in a post stratified setup (Jager *et. al*., 1985). Critical thoughts on application of post stratification was presented subsequently (Smith, 1991). In a seminal paper on post stratification, weight structure which has made use of both the population information and the sampling fractions from different strata to arrive at the final estimate was examined

Corresponding Author: Purnendukisor Bandyopadhyay
Email: purnendukb@yahoo.com

(Agrawal and Panda, 1993, 1995). This has resulted in further study in the area of post stratification. Estimation methodologies for population mean under stratified population using prior information with grouping strategy was developed (Shukla *et. al.* 2001, 2002). This paper has essentially been motivated by the weight structure given in the Agrawal-Panda post stratified estimate. The structure has been first applied in a stratified sampling set up. A few alternate estimates using this structure has been proposed. Expectation and variance/Mean Square Error (MSE) of these alternate estimates have then been derived. After deriving these estimates, the variance / MSE of the estimates have been compared with that of the usual post stratified estimates using real life data.

Concluding remarks have been provided on performance of the proposed estimators and scopes for further work in this area.

2. Methodology

Let U be a finite population having N clusters. Let the clusters be of unequal sizes. The population can be divided into k strata such that i^{th} stratum contains N_i clusters, $\sum_{i=1}^{k} N_i = N$. A random sample of n clusters ($n < N$) is drawn from the N clusters by simple random sampling without replacement (SRSWOR). The sample is post stratified such that n_i clusters are from the i^{th} stratum $\sum_{i=1}^{k} n_i = n$.

2.1. Notations used for the population parameters

The notations used throughout are

Y Variable under study

Y_{ijl} l^{th} value of the variable Y in i^{th} stratum and j^{th} cluster

$\bar{Y}_{ij}$ Mean value of Y in i^{th} stratum and j^{th} cluster

W_i Population proportion of clusters in i^{th} stratum = $\frac{N_i}{N}$

M_{ij} Size of j^{th} cluster in i^{th} stratum

M_i Total elements in i^{th} stratum = $\sum_{j=1}^{N_i} M_{ij}$

$\bar{M}_i$ Average size of clusters in i^{th} stratum = $\frac{M_i}{N_i}$

u_{ij} $M_{ij} \Big/ M_i$

$\bar{Y}_i$ Population mean of i^{th} stratum = $\sum_{i-1}^{k}\sum_{j=1}^{N_i} Y_{ijl} \Big/ M_i = \bar{Y}_{N_i} = \sum_{j=1}^{N_i} M_{ij}\,\bar{Y}_{ij} \Big/ M_i = \frac{1}{N_i}\sum_{j=1}^{N_i} u_{ij}\bar{Y}_{ij}$

$\bar{Y}_{..}$ $= \bar{Y} = \sum_{i=1}^{k} W_i \bar{Y}_i$

$\bar{\bar{Y}}_N$ = mean of N cluster means $= \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{N_{ij}} \bar{Y}_{ij}$

$\bar{\bar{Y}}_{N_i}$ Mean of $N_i clusters\ of\ i^{th} stratum = \frac{1}{N_i}\sum_{j=1}^{N_i} \bar{Y}_{ij}$

$S_{b_i}^2 = \frac{1}{(N_i - 1)} \sum_{J=1}^{N_i} \left(\bar{Y}_{ij} - \bar{Y}_{N_i}\right)^2$

$S_{b_i}'^2 = \frac{1}{(N_i - 1)} \sum_{J=1}^{N_i} \left(u_{ij}\,\bar{Y}_{ij} - \bar{Y}_{N_i}\right)^2$

$$S'_{bu_i} = \sqrt{\sum_{J=1}^{N_I} \frac{\left(u_{ij} - 1\right)^2}{\left(N_i - 1\right)}}$$

$$S'_{byu_i} = \sum_{J=1}^{N_I} \frac{\left(u_{ij}\,\bar{Y}_{ij} - \bar{Y}_i\right)\left(u_{ij} - 1\right)}{\left(N_i - 1\right)}$$

$$S_{iM\bar{Y}} = \frac{1}{(N_i - 1)} \sum_{j=1}^{N_i} \left(M_{ij} - \bar{M}_i\right)\left(\bar{Y}_{ij} - \bar{\bar{Y}}_{N_i}\right)$$

$$S''^{\,2}_{b_i} = \frac{1}{(N_i - 1)} \sum_{J=1}^{N_i} u^2{}_{ij}\left(\bar{Y}_{ij} - \bar{Y}_{N_i}\right)^2$$

$$S'''_{b_i} = \frac{1}{(N_i - 1)} \sum_{J=1}^{N_i} \left[u_{ij}\left(\bar{Y}_{ij} - \bar{Y}_i\right) - \bar{\bar{Y}}_N\left(u_{ij} - 1\right)\right]$$

2.2. Notations used for the sample estimates

The notations used for the sample estimates are

$\bar{y}_{ij}$ Mean of j^{th} cluster in i^{th} stratum included in the sample, $i = 1, 2, \ldots, k$; $j = 1, 2, \ldots, n_i$

$\bar{y}_i = \frac{\sum_{j=1}^{n_i} M_{ij}\bar{y}_{ij}}{\sum_{j=1}^{n_i} M_{ij}}$, sample mean of i^{th} stratum

$P_i = \frac{n_i}{n}$: sample proportion of clusters from i^{th} stratum

$f = \frac{n}{N} : sampling\ fraction$

$$\bar{u}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} u_{ij}$$

2.3. Usual estimators

Some of the usual estimators in cluster sampling scheme are:

$$\bar{\bar{y}}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \bar{y}_{ij}$$

$$\bar{y}'_i = \frac{1}{n_i} \sum_{j=1}^{n_i} u_{ij}\,\bar{y}_{ij}$$

$$\bar{y}''_i = \sum_{j=1}^{n_i} u_{ij}\,\bar{y}_{ij} \Big/ \sum_{j=1}^{n_i} u_{ij}$$

$\bar{y}_{ps} = \frac{1}{N}\sum_{i=1}^{k} N_i\bar{y}_i$ is the usual post stratified estimator

Variance in stratified sampling is $Var\left(\bar{y}_{stratified}\right) = \sum_{i=1}^{k}\left(\frac{N_i}{N}\right)^2 \left(\frac{N_i - n_i}{N_i}\right) \frac{S_i^2}{n_i}$

The variance of usual post stratified estimator is

$$Var\left(\bar{y}_{ps}\right) = Var\left(\bar{y}_{stratified}\right) + \frac{N - n}{n^2\,N}\sum_{i=1}^{k}(1 - W_i)\,S_i^2$$

3. Proposed Estimators

For developing the proposed estimates, first let us recall the weight structure used in the seminal paper of Agrawal-Panda (1993), which is $W_{i\,\alpha}^{*} = \left[\alpha\ \frac{n_i}{n} + (1 - \alpha)\ \frac{N_i}{N}\right]$, α being a suitably chosen constant.

Taking inspiration from this weight structure, some proposed post stratified estimates, which we have examined in this paper are

$$\bar{y}_a^{*} = \sum_{i=1}^{k} W_{i\,\alpha}^{*}\,\bar{\bar{y}}_i$$

$$\bar{y}_b^{*} = \sum_{i=1}^{k} W_{i\,\alpha}^{*}\,\bar{y}_i'$$

$$\bar{y}_c^{*} = \sum_{i=1}^{k} W_{i\,\alpha}^{*}\,\bar{y}_i''$$

$$\bar{y}_d^{*} = \sum_{i=1}^{k} W_{i\,\alpha}^{*}\,\bar{u}_i\,\bar{\bar{y}}_i$$

3.1. Deriving properties of the proposed estimators

Theorem 1: The estimator $\bar{y}_a^{*}$ is biased for $\bar{Y}$ and the MSE is

$$MSE\,(\bar{y}_a^{*}) = \left(\frac{1}{n} - \frac{1}{N}\right)\sum_{i=1}^{k} W_i S_{b_i}^2 + \frac{(N-n)\,(1-\alpha)^2}{(N-1)\,n^2}\sum_{i=1}^{k}(1 - W_i)\,S_{b_i}^2 + \frac{(N-n)}{Nn}\alpha^2\left[S_a^2 - \sum_{i=1}^{k} W_i S_{b_i}^2\right] + \left[\sum_{i=1}^{k}\frac{(N_i-1)\,W_i}{N_i\bar{M}_i}\ S_{iM\bar{Y}}\right]$$

where $S_a^2 = \frac{1}{(N-1)}\ \sum_{i=1}^{k}\sum_{j-1}^{N_i}\left[\bar{Y}_{ij} - \bar{\bar{Y}}_N\right]^2$

Theorem 2: The estimator $\bar{y}_b^{*}$ is unbiased for $\bar{Y}$ and the variance is

$$Var\,(\bar{y}_b^{*}) = \left(\frac{1}{n} - \frac{1}{N}\right)\sum_{i=1}^{k} W_i S_{b_i}'^{\,2} + \frac{(N-n)\,(1-\alpha)^2}{(N-1)\,n^2}\sum_{i=1}^{k}(1 - W_i)\,S_{b_i}'^{\,2} + \frac{(N-n)}{Nn}\alpha^2\left[S_b^2 - \sum_{i=1}^{k} W_i S_{b_i}'^{2}\right]$$

where $S_{b_i}'^{\,2} = \frac{1}{(N_i-1)}\ \sum_{J=1}^{N_i}\left(u_{ij}\,\bar{Y}_{ij} - \bar{Y}_{N_i}\right)^2$ and $S_b^2 = \sum_{i=1}^{k}\frac{1}{(N_i-1)}\sum_{j-1}^{N_i}\left[u_{ij}\bar{Y}_i - \bar{Y}\right]^2$

Theorem 3: The estimator $\bar{y}_c^*$ is unbiased for $\bar{Y}$ and the variance is

$$Var\,(\bar{y}_c^*) = \left(\frac{1}{n} - \frac{1}{N}\right)\sum_{i=1}^{k} W_i S_{b_i}^{\prime\,2} + \frac{(N-n)\,(1-\alpha)^2}{(N-1)\,n^2}\sum_{i=1}^{k}(1-W_i)\,S_{b_i}^{\prime\,2} + \frac{(N-n)}{Nn}\alpha^2\left[S_c^2 - \sum_{i=1}^{k} W_i S_{b_i}^{\prime\,2}\right]$$

where $S_{b_i}^{\prime\,2} = \frac{1}{(N_i-1)}\sum_{J=1}^{N_i} u^2{}_{ij}\left(\bar{Y}_{ij} - \bar{Y}_{N_i}\right)^2$ and $S_c^2 = \frac{1}{(N-1)}\sum_{i=1}^{k}\sum_{j-1}^{N_i} u_{ij}^2\left[\bar{Y}_{ij} - \bar{Y}\right]^2$

Theorem 4: The estimator $\bar{y}_d^*$ is biased for $\bar{Y}$ and the MSE is

$$MSE\,(\bar{y}_d^*) = \left(\frac{1}{n} - \frac{1}{N}\right)\sum_{i=1}^{k} W_i S_{b_i}^2 + \frac{(N-n)\,(1-\alpha)^2}{(N-1)\,n^2}\sum_{i=1}^{k}(1-W_i)\,S_{b_i}^2 + \frac{(N-n)}{Nn}\alpha^2\left[S_a^2 - \sum_{i=1}^{k} W_i S_{b_i}^2\right] + \left[\sum_{i=1}^{k}\left\{\left(W_i - \frac{1}{n}\right) - \frac{(1-\alpha)\,(N-n)\,(1-W_i)}{(N-1)\,n^2\,W_i}\right\}\frac{S_{iM\bar{Y}}}{\bar{M}_i}\right]^2$$

4. Numerical Illustrations

The above proposed estimators were compared with usual post stratified estimators using 3 different values of α, so that it can show the performance of the proposed estimators for varying levels of weights distributed between the sample and population proportion of clusters.

The Department of School Education and Literacy (DoSEL), Government of India conducts the annual survey of schools. This data collection and dissemination system is called the Unified District Information System for Education Plus (UDISE+), the details of which can be accessed at https://udiseplus.gov.in/#/page/about. This is an annual census conducted in more than 15 lakh schools of India. The DoSEL disseminates anonymised unit level data. The numerical illustration has been done using class-wise enrolment of students in different schools at 3 districts of one State. Here, each district was considered as a separate stratum, each school as a cluster and variable under study was number of students in a class. The data set provides information on all the schools of a district, i.e., the entire frame was available for drawal of samples and computation of variance, MSE, etc. of the proposed estimators. In the present paper, samples were drawn by SRSWOR from the frame of schools of a specific State and then post-stratified among the 3 strata (i.e., 3 districts of this State). The results are given below:

Variable	Stratum			overall
	I	II	III	
N_i	176	184	58	418
n_i	23	20	7	50
$\bar{Y}$				24.2
$\bar{\bar{Y}}_N$				20.7
$\bar{y}_i$	29.47	14.51	21.14	
$\sum_{j=1}^{n_i} u_{ij}$	22.79	19.97	6.06	

Variable	Stratum			overall
	I	II	III	
$\bar{y}_{ps}$				21.73
$\bar{\bar{y}}_i$	22.78	13.30	17.61	
$\bar{y}_i'$	29.20	14.49	18.30	
$\bar{y}_i''$	29.47	14.51	21.14	
$\bar{y}_a^*$, $\alpha = 0.1$				17.93
$\bar{y}_a^*$, $\alpha = 0.5$				18.08
$\bar{y}_a^*$, $\alpha = 0.9$				18.23
$\bar{y}_b^*$, $\alpha = 0.1$				21.27
$\bar{y}_b^*$, $\alpha = 0.5$				21.50
$\bar{y}_b^*$, $\alpha = 0.9$				21.74
$\bar{y}_c^*$, $\alpha = 0.1$				21.79
$\bar{y}_c^*$, $\alpha = 0.5$				22.03
$\bar{y}_c^*$, $\alpha = 0.9$				22.26
$\bar{y}_d^*$, $\alpha = 0.1$				17.50
$\bar{y}_d^*$, $\alpha = 0.5$				17.65
$\bar{y}_d^*$, $\alpha = 0.9$				17.79
$Var\left(\bar{y}_{stratified}\right)$				16.88
$Var\left(\bar{y}_{ps}\right)$				**19.47**
$MSE\ (\bar{y}_a^*)$, $\alpha = 0.1$				**16.346**
$MSE\ (\bar{y}_a^*)$, $\alpha = 0.5$				**16.361**
$MSE\ (\bar{y}_a^*)$, $\alpha = 0.9$				**16.814**
$Var\ (\bar{y}_b^*)$, $\alpha = 0.1$				23.229
$Var\ (\bar{y}_b^*)$, $\alpha = 0.5$				19.710
$Var\ (\bar{y}_b^*)$, $\alpha = 0.9$				**12.234**
$Var\ (\bar{y}_c^*)$, $\alpha = 0.1$				**16.942**
$Var\ (\bar{y}_c^*)$, $\alpha = 0.5$				**17.325**
$Var\ (\bar{y}_c^*)$, $\alpha = 0.9$				**18.756**
$MSE\ (\bar{y}_d^*)$, $\alpha = 0.1$				**13.74**
$MSE\ (\bar{y}_d^*)$, $\alpha = 0.5$				**13.76**
$MSE\ (\bar{y}_d^*)$, $\alpha = 0.9$				**14.21**

As can be seen from the above results, the variance / MSE of the proposed estimators were less than the usual post stratified estimator $\bar{y}_{ps}$. Moreover, for a few specific choices of α, the variance/ MSE of proposed estimators were less than the variance of usual stratified SRSWOR estimator. In the given data set, the proposed estimator $\bar{y}_b^*$ has shown least variance with a choice of $\alpha = 0.9$.

5. Concluding Remarks

The four estimates namely $\bar{y}_a^*, \bar{y}_b^*, \bar{y}_c^* \; and \; \bar{y}_d^*$ have been developed using the weight structure of Agrawal and Panda (1993) after applying this on a post stratified cluster sampling scheme. The properties of these estimators have been derived and comparison among these estimators' variances/ Mean Square Errors have been made to find out whether the proposed estimator is behaving well or not. The methodology has been applied on a data set which is newly available, and which contain population level information. This has enabled to compute the theoretical variances and MSEs from real life recent data. The variances/ MSEs of the proposed post stratified estimators are lower than the traditional post stratified estimators for some of the values of α. This shows that the proposed post stratified estimator can be more efficient over the usual PS estimator for these values of α.

Acknowledgement

The authors are grateful to the editors for their useful suggestions that led to considerable improvement in the presentation of the contents. The authors are grateful, particularly to Prof. V. K. Gupta, who gave numerous suggestions for improvement of the paper and Dr. D. Roy Choudhury who was a constant source of inspiration for the authors. The authors humbly and respectfully acknowledge the constant guidance of Prof. V. K. Singh and the encouragement and support from the faculty of the School of Sciences, IGNOU. Last, but not the least, the authors are grateful to the Ministry of Education for their encouragement in taking up research work in addition to usual duties.

References

Agarwal, M. C. and Panda, K. B. (1993). An efficient estimator in poststratification. *Metron*, **5**(**3-4**), 179-187.

Bryant, E. C., Hartley, H. O. and Jessen, R. J. (1960). Design and estimation in two-way stratification. *Journal of the American Statistical Association*, **55**, 105-124.

Holt, D. and Smith, T. M. F. (1979). Post-stratification. *Journal of the Royal Statistical* Society, **A, 142**, 33-36.

Shukla, D. and Trivedi, M. (2001). Mean estimation in deeply stratified population under post-stratification. *Journal of* the *Indian Society of Agricultural Statistics*, **54(2)**, 221-235.

Smith, T. M. F. (1991). Post-stratification. *The Statistician,* **40**, 315-323.

Unified District Information System for Education Plus – data sharing portal https://src.udiseplus.gov.in/udise-share/ accessed various times during 2021.