Jour. Ind. Soc. Ag. Statistics
55(2), 2002 : 158-173

Estimation in Post-Stratification Using Prior Information and Grouping Strategy

D. Shukla, Ajay Bankey and Manish Trivedi
Dr. H.S. Gour University, Sagar (M.P.) 470003
(Received : December, 1998)

SUMMARY

This paper presents an estimator in a post stratified set-up of sampling assuming prior knowledge of Population Proportion of Mean Matrix (PPM) and coefficients of variation of strata. A concept of PPM matrix is introduced and properties are derived. The method of choice of weights for combining post-stratified sample means, is proposed along with their optimum selection. A general strategy of grouping strata is introduced with the application of two plans.

Key words : Post-stratification, SRSWOR, PPM, Estimator, Optimum.

1. Introduction

The stratification requires information on strata sizes and availability of frames for each stratum. The former is easier to manage but the latter is often hard to get and therefore reduces the effective application of stratification. As a solution, the post-stratification technique is used and according to Sukhatme *et al.* [12] with a large sample size, the post-stratification is always as precise as the stratified sampling with proportional allocation. Some useful research contributions in the area of post-stratification are by Holt and Smith [3], Jagers *et al.* [4], Jagers [5], Agrawal and Panda [1], Little [7], Gelman and Little [2], Lazzerni and Little [6] and Shukla and Trivedi ([10], [11]).

2. Notations and Assumptions

Let N be the size of a population $U = [U_{11}, U_{12}, \ldots U_{iN_i} \ldots U_{KN_K}]$ consisting of K strata, with N_i units belonging to the i^{th} strata such that $\sum_{i=1}^{K} N_i = N$. Let variable under study be Y with values $Y_{ij}, i = 1, 2, \ldots K$, $j = 1, 2, \ldots N_i$ on U_{ij} along with stratum means $\overline{Y}_i$ of i^{th} stratum and grand mean

$\overline{Y}$. A sample of size n is drawn from N by SRSWOR and post-stratified into k strata with n_i units falling in the i^{th} stratum such that $\sum_{i=1}^{k} n_i = n$. The sample mean is $\overline{y}$ based on n and $\overline{y}_i$ is based on n_i units. The components of variability are

$$S_i^2 = \left(\frac{1}{N_i - 1}\right)\left[\sum_{j=1}^{N_i} \left(Y_{ij} - \overline{Y}_i\right)^2\right]$$

$$S^2 = \left(\frac{1}{N-1}\right)\left[\sum_{i=1}^{K} (N_i - 1)S_i^2 + \sum_{i=1}^{K} N_i(\overline{Y}_i - \overline{Y})^2\right]$$

$$C_{Y_i} = \frac{S_i}{\overline{Y}_i}, C_Y = \frac{S}{\overline{Y}}$$

2.1 Population Proportion of Mean Matrix (PPM)

Define a matrix of order $(k+1)\times(k+1)$ as $P = (p_{ij}); i, j = i, 2, \ldots k$ where

$$p_{ij} = \overline{Y}_i : \overline{Y}_j = \left(\frac{\overline{Y}_i}{\overline{Y}_j}\right), \quad p_{i(k+1)} = \overline{Y}_i : \overline{Y} = \left(\frac{\overline{Y}_i}{\overline{Y}}\right)$$

$$p_{(k+1)j} = \overline{Y} : \overline{Y}_j = \left(\frac{\overline{Y}}{\overline{Y}_j}\right), \quad p_{(k+1)(k+1)} = \overline{Y} : \overline{Y} = 1$$

Some important properties of matrix P

(i) It is a square matrix

(ii) Diagonal elements of P are unity *i.e.* $p_{ij} = 1$ for $i = j$ and $p_{(k+1)(k+1)} = 1$

(iii) Non diagonal elements possess a relation

$$p_{ji} = \frac{1}{p_{ij}}, p_{i(k+1)} = \frac{1}{p_{(k+1)j}} \text{ for } i \neq j$$

(iv) The knowledge of only lower (or upper) diagonal elements is enough to determine P completely.

Some important assumptions are: (i) while post stratifying n, it is presumed that probability of n_i being zero is very small, (ii) prior information on lower (or upper) diagonal elements of P is available, (iii) prior information on coefficients of variation C_{Y_i}, of each strata, is available.

The assumption (i) is obvious with moderate k for a large n. Moreover, (ii) and (iii) are easily possible through expert guess, past experience, successive surveys or pilot surveys. As an example, an agricultural survey of the rural area of a district, village may classified as "Big Size" and "Small Size" according to the area under cultivation, and "Crop Production" is a variable of main interest. The possible guesses are

(i) average crop production by small-villages is nearly one-third to big-villages and approximately half to the average of entire rural area

(ii) average production of big size group is nearly $2\frac{1}{2}$ times of entire rural area. This provides a PPM matrix of order 3×3 as

$$\begin{array}{l} \\ \text{Small} \\ \text{Big} \\ \text{All} \end{array} \begin{array}{c} \begin{array}{ccc} \text{Small} & \text{Big} & \text{All} \end{array} \\ \begin{bmatrix} 1:1 & 1:3 & 1:2 \\ 3:1 & 1:1 & 2.5:1 \\ 2:1 & 1:2.5 & 1:1 \end{bmatrix} \end{array} = \begin{bmatrix} 1 & 1/3 & 1/2 \\ 3 & 1 & 5/2 \\ 2 & 2/5 & 1 \end{bmatrix}$$

Remark. When reliable information on P is available through an expert guess, past experience or a pilot survey, it needs to be effectively utilized in estimation problems. Searls [9] has utilized the prior information on the coefficient of variation C_Y for constructing an efficient estimator.

3. Proposed Estimator

With $W_i = \left(\frac{N_i}{N}\right)$, a class of post-stratified estimator for $\overline{Y}$, is

$$\left[\left(\overline{y}_{ps}\right)_{c_i}\right] = \sum_{i=1}^{k} \left(C_i\ W_i\right) \overline{y}_i \tag{3.1}$$

where C_i is an unknown constant of the i^{th} stratum and the quantity $(C_i W_i)$ constitutes a new weight structure for combining strata means in the sample.

Remark 3.1. As special case when $C_i = C, \forall\ i$ then

$$\left[\left(\overline{y}_{ps}\right)_{c}\right] = C \sum_{i=1}^{k} W_i \overline{y}_i \tag{3.2}$$

and C = 1 provides usual post-stratified estimator $\overline{y}_{ps} = \sum_{i=1}^{k} W_i \overline{y}_i$ (3.3)

Remark 3.2. The proposed (3.1) is a general class of estimators having (3.2) and (3.3) as members. Moreover, (3.3) is unbiased for $\overline{Y}$. The weight structure $C_i W_i$ is to be chosen subject to the level of minimum mean square error, in the class (3.1).

3.1 Motivation and Justification

(i) In the set up (N, n) of SRSWOR, for sample mean $\bar{y}$ with $V(\bar{y}) = [(n^{-1}) - (N^{-1})]S^2$; Searls [9] has proposed estimator $\bar{y}_s = C\bar{y}$ with optimal choice

$$C = C^* = \left[1 + \left\{\left(n^{-1}\right) - \left(N^{-1}\right)\right\}C_Y^2\right]$$

(ii) The $\bar{y}_s$ observed efficient over $\bar{y}$ at $C = C^*$ assuming known coefficient of variation (C_Y) of the population

(iii) A motivation is derived from (i) and (ii) for a post-stratified set-up of sampling $\left(N, n = \sum_{i=1}^{k} n_i\right)$ in the form of proposed class (3.1) assuming

Case I: when constant C is same for all k strata

Case II: when it is different for all k strata

Case III: when it is same for a group of strata

3.2 Bias and Mean Square Error

$$E\left[\left(\bar{y}_{ps}\right)_{c_i}\right] = \sum_{i=1}^{k} C_i W_i \bar{Y}_i \text{ and } Bias\left[\left(\bar{y}_{ps}\right)_{c_i}\right] = \left[\sum_{i=1}^{k} C_i W_i \bar{Y}_i - \bar{Y}\right] \quad (3.2.1)$$

Note 3.1. Wherever follows, we denote $E[\{.\}/n_i]$ and $V[\{.\}/n_i]$ as conditional expectation and conditional variance under given n_i. A standard result is

$$E\left[\frac{1}{n_i}\right] = \left[\frac{1}{nW_i} + \frac{(N-n)}{(N-1)}\frac{(1-W_i)}{n^2W_i^2}\right]$$

Remark 3.3. The mean square error of the class (3.1) is

$$MSE\left[\left(\bar{y}_{ps}\right)_{c_i}\right] = V\left[\left(\bar{y}_{ps}\right)_{c_i}\right] + \left[Bias\left\{\left(\bar{y}_{ps}\right)_{c_i}\right\}\right]^2$$

$$= E\left[V\left[\left\{\left(\bar{y}_{ps}\right)_{c_i}\right\}/n_i\right]\right] + V\left[E\left[\left\{\left(\bar{y}_{ps}\right)_{c_i}\right\}/n_i\right]\right] + E\left[E\left[\left\{\left(\bar{y}_{ps}\right)_{c_i} - \bar{Y}\right\}^2/n_i\right]\right]$$

$$= E\left[\sum_{i=1}^{k} C_i^2 W_i^2 \left(\frac{1}{n_i} - \frac{1}{N_i}\right)S_i^2\right] + \left[\sum_{i=1}^{k} C_i W_i \bar{Y}_i - \bar{Y}\right]^2$$

$$= \sum_{i=1}^{k} C_i^2 W_i^2 \left[E\left\{ \frac{1}{n_i} - \frac{1}{N_i} \right\} \right] S_i^2 + \left[\sum_{i=1}^{k} C_i W_i \overline{Y}_i - \overline{Y} \right]^2$$

$$= \left[\frac{1}{n} - \frac{1}{N} \right] \sum_{i=1}^{k} C_i^2 W_i S_i^2 + \frac{(N-n)}{(N-1)(n^2)} \left[\sum_{i=1}^{k} (1 - W_i) C_i^2 S_i^2 \right]$$

$$+ \left[\sum_{i=1}^{k} C_i W_i \overline{Y}_i - \overline{Y} \right]^2$$

which is obtained using note 3.1. On substitution of $C_i^2 = [1 + (C_i^2 - 1)]$

$$MSE\left[(\overline{y}_{ps})_{c_i} \right] = V[(\overline{y}_{ps})] + \sum_{i=1}^{k} (C_i^2 - 1) A_i S_i^2 + [B]^2 \tag{3.2.2}$$

$$V[(\overline{y}_{ps})] = \left[\frac{1}{n} - \frac{1}{N} \right] \sum_{i=1}^{k} W_i S_i^2 + \frac{(N-n)}{(N-1)n^2} \left[\sum_{i=1}^{k} (1 - W_i) S_i^2 \right] \tag{3.2.3}$$

$$\text{At } C_i = C, \forall i, \; MSE[(\overline{y}_{ps})_c] = V[(\overline{y}_{ps})] + (C^2 - 1)[D] + [(C-1)^2 \overline{Y}^2] \tag{3.2.4}$$

$$A_i = \left[\left(\frac{1}{n} - \frac{1}{N} \right) W_i + \frac{(N-n)}{(N-1)n^2} (1 - W_i) \right]$$

$$B = \left[\sum_{i=1}^{k} C_i W_i \overline{Y}_i - \overline{Y} \right]; \; D = \left[\sum_{i=1}^{k} A_i S_i^2 \right] = V(\overline{y}_{ps})$$

4. Choice of C_i

The proposed estimator (3.1) is efficient over $\overline{y}_{ps}$, when C_i satisfies condition

$$\left[\sum_{i=1}^{k} (C_i^2 - 1) A_i S_i^2 \right] + [B]^2 < 0 \tag{4.1}$$

Moreover, from (3.2.4), the estimator (3.2) is efficient over $\overline{y}_{ps}$, when the selection of C, fulfils condition $(C-1)\left[(C+1)D + (C-1)\overline{Y}^2\right] < 0$ (4.2)

Remark 4.1. In (4.1), if choices $C_i > 1$ for all i then $\left[(\bar{y}_{ps})_{C_i}\right]$ can never be efficient over $\bar{y}_{ps}$. If at least one or some of them are less than unity, there is a high chance of getting gain over usual estimator. In (4.2), the choice C < 1 supports this fact.

Differentiate (3.2.2) with C_i and equate to zero, we have

$$2\left[A_i S_i^2 + W_i^2 \overline{Y}_i^2\right] C_i + \left[\sum^{k} \sum^{k}_{i \neq j} C_j W_i W_j \overline{Y}_i \overline{Y}_j\right] - 2\overline{Y}\left[W_i \overline{Y}_i\right] = 0 \quad (4.3)$$

Divide by $2\overline{Y}_i^2$, we have the systems of k equations in C_i as

$$\left.\begin{aligned}
&\left(A_1 C_{Y_1}^2 + W_1^2\right) C_1 + \frac{1}{2}\left(W_2 W_1 \frac{\overline{Y}_2}{\overline{Y}_1}\right) C_2 + \frac{1}{2}\left(W_3 W_1 \frac{\overline{Y}_3}{\overline{Y}_1}\right) C_3 + \ldots \\
&\qquad \ldots + \frac{1}{2}\left(W_k W_1 \frac{\overline{Y}_k}{\overline{Y}_1}\right) C_k = \frac{\overline{Y}}{\overline{Y}_1} W_1 \\
&\frac{1}{2}\left(W_1 W_2 \frac{\overline{Y}_1}{\overline{Y}_2}\right) C_1 + \left(A_2 C_{Y_2}^2 + W_2^2\right) C_2 + \frac{1}{2}\left(W_3 W_2 \frac{\overline{Y}_3}{\overline{Y}_2}\right) C_3 + \ldots \\
&\qquad \ldots + \frac{1}{2}\left(W_k W_2 \frac{\overline{Y}_k}{\overline{Y}_2}\right) C_k = \frac{\overline{Y}}{\overline{Y}_2} W_2 \\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\
&\frac{1}{2}\left(W_1 W_k \frac{\overline{Y}_1}{\overline{Y}_k}\right) C_1 + \frac{1}{2}\left(W_1 W_k \frac{\overline{Y}_2}{\overline{Y}_k}\right) C_2 + \frac{1}{2}\left(W_3 W_k \frac{\overline{Y}_3}{\overline{Y}_k}\right) C_3 + \\
&\qquad \ldots + \left(A_k C_{Y_k}^2 + W_k^2\right) C_k = \frac{\overline{Y}}{\overline{Y}_k} W_k
\end{aligned}\right] \quad (4.4)$$

The (4.4) has k equations for k unknown C_i. The other elements W_i and elements of P are known, therefore the system could be easily solved for C_i using any standard technique of solution of equations.

4.1 Criteria for Optimum Choice

The necessary condition for the proposed estimator (3.1) to be more efficient than $\bar{y}_{ps}$ is that C_i values $(i = 1, 2, \ldots k)$ obtained as a solution of system of equations (4.4) must satisfy (4.1).

Remark 4.1.1. In matrix notation, (4.4) could be like **AC** = **B** where

$$\mathbf{A} = \left[a_{ij}\right]_{k\times k} \text{ and } a_{ij} = \begin{bmatrix} A_i C_{Y_i}^2 + W_i^2 & \text{if } i = j = 1, 2, 3 \ldots k \\ \frac{1}{2} W_i W_j \frac{\overline{Y}_j}{\overline{Y}_i} = \frac{1}{2} W_i W_j p_{ij} & \text{if } i \neq j \end{bmatrix}$$

$$\mathbf{B} = \left[b_j\right]_{k\times k} \text{ and } b_j = \frac{\overline{Y}}{\overline{Y}_j} W_j,\ \mathbf{C}' = \left[C_1, C_2, C_3, \ldots C_k\right]_{1\times k}$$

Remark 4.1.2. The optimum MSE of $\left[(\overline{y}_{ps})_c\right]$ at the value

$$C_{opt} = \left[1 + \frac{D}{\overline{Y}^2}\right]^{-1}$$

$$\text{MSE}\left[(\overline{y}_{ps})_c\right]_{opt} = C_{opt}\, V(\overline{y}_{ps})$$

5. *Empirical Study*

In order to examine the performance of the proposed estimator some empirical illustrations are given in Tables 5.1 and 5.2 for various types of data sets. The efficiency comparisons of these data sets are given in Table 5.3.

Table 5.1. Data set I (From Sarndal *et al.* [8])

Stratum No.	N_i	$\sum_{j=1}^{N_i} Y_{ij}$	$\sum_{j=1}^{N_i} Y_{ij}^2$	$\overline{Y}_i$	S_i^2	W_i	Sample Size
1	105	1098.9	21855.0	$\overline{Y}_1 = 10.4657$	$S_1^2 = 99.560?$	0.8467	30
2	19	3445.9	1822736.8	$\overline{Y}_2 = 181.3631$	$S_2^2 = 66543.195$	0.1532	
Total	124	4544.8	1844591.8	$\overline{Y} = 36.65$	$S^2 = 12213.202$	-	

Matrices **P** and **A** and vector **B** are

$$\mathbf{P} = \begin{bmatrix} 1 & 0.06 & 0.28 \\ 17.33 & 1 & 4.95 \\ 3.50 & 0.20 & 1 \end{bmatrix}_{3\times 3} \quad \mathbf{A} = \begin{bmatrix} 0.7365 & 1.1242 \\ 0.0037 & 0.0327 \end{bmatrix}_{2\times 2} \quad \mathbf{B} = \begin{bmatrix} 2.9654 \\ 0.0309 \end{bmatrix}_{2\times 1}$$

Table 5.2: Other Data Set

Data Set No.	Stratum No.	N_i	$\sum_{j=1}^{N_i} Y_{ij}$	$\sum_{j=1}^{N_i} Y_{ij}^2$	$\overline{Y}_i$	S_i^2	W_i	Sample Size
II	1	120	1843.92	33232.5	$\overline{Y}_1 = 15.366$	$S_1^2 = 41.1667$	0.8467	30
	2	32	6108.00	3341961.8	$\overline{Y}_2 = 190.875$	$S_2^2 = 70196.693$	0.1532	
	Total	152	7951.92	3375194.3	$\overline{Y} = 52.3152$	$S^2 = 19745.073$		
III	1	105	573.89	13491.0	$\overline{Y}_1 = 5.4657$	$S_1^2 = 99.5603$	0.8467	30
	2	32	3350.89	3058164.1	$\overline{Y}_2 = 176.3631$	$S_2^2 = 66543.1979$	0.1532	
	Total	124	3924.79	3071655.1	$\overline{Y} = 31.6516$	$S^2 = 13642.412$		
IV	1	180	1831.89	26951.0	$\overline{Y}_1 = 10.1772$	$S_1^2 = 46.41066$	0.8181	40
	2	40	7212.00	4496134.2	$\overline{Y}_2 = 180.3$	$S_2^2 = 81943.864$	0.1818	
	Total	220	9043.89	4523085.2	$\overline{Y} = 41.1086$	$S^2 = 16635.48$		
V	1	44	345.29	135704.7	$\overline{Y}_1 = 7.8477$	$S_1^2 = 3092.9064$	0.468	25
	2	50	5759.90	4141861.3	$\overline{Y}_2 = 115.198$	$S_2^2 = 70986.377$	0.532	
	Total	94	6105.19	4277566.0	$\overline{Y} = 64.9489$	$S^2 = 19745.073$		

Data Set No.	Stratum No.	N_i	$\sum_{j=1}^{N_i} Y_{ij}$	$\sum_{j=1}^{N_i} Y_{ij}^2$	$\overline{Y}_i$	S_i^2	W_i	Sample Size
VI	1	105	573.89	13491.0	$\overline{Y}_1 = 5.4657$	$S_1^2 = 99.5603$	0.7342	25
	2	38	5903.78	3313737.1	$\overline{Y}_2 = 155.367$	$S_2^2 = 64770.4$	0.2657	
	Total	143	6477.69	3327228.1	$\overline{Y} = 45.2985$	$S^2 = 21364.77$		
VII	1	44	345.29	135704.7	$\overline{Y}_1 = 7.8477$	$S_1^2 = 3092.9064$	0.4681	25
	2	40	5541.00	4060226.9	$\overline{Y}_2 = 138.525$	$S_2^2 = 84427.1789$	0.4255	
	3	10	218.90	81634.6	$\overline{Y}_3 = 21.89$	$S_3^2 = 8538.1$	0.1064	
	Total	94	6105.19	4277566.2	$\overline{Y} = 64.9489$	$S^2 = 41731.61$		
VIII	1	105	573.89	13491.9	$\overline{Y}_1 = 5.4657$	$S_1^2 = 99.5603$	0.7342	30
	2	19	3046.89	1686387.6	$\overline{Y}_2 = 160.3631$	$S_2^2 = 66543.19$	0.3286	
	3	19	2856.89	1627349.6	$\overline{Y}_3 = 150.36$	$S_3^2 = 66543.19$	0.1328	
	Total	143	6477.69	3327228.2	$\overline{Y} = 45.29857$	$S^2 = 21364.7$		

5.1 Calculation of Variance, MSE and Bias

Table 5.3

Date Set No.	$V(\bar{y})$	$V(\bar{y}_{ps})$	Estimator $(\bar{y}_{ps})_c$				Estimator $(\bar{y}_{ps})_{ci}$			
			Optimum C–Value	Optimum M.S.E.	Bias	Estimator $(\bar{y}_{ps})_c$	Optimum C-Value	Optimum M.S.E.	Bias	Estimator $(\bar{y}_{ps})_{ci}$
I	358.418	307.634	0.8136	250.3110	6.8296	$E_1 = 30.16\%$ $E_2 = 18.66\%$	$C_1 = 3.1291$ $C_2 = 0.5875$	181.3048	7.4069	$E_1 = 49.41\%$ $E_2 = 41.06\%$
II	487.130	421.259	0.8666	364.9950	6.9788	$E_1 = 25.0\%$ $E_2 = 13.3\%$	$C_1 = 3.24$ $C_2 = 0.6358$	319.2280	12.5418	$E_1 = 34.46\%$ $E_2 = 24.22\%$
III	344.727	307.634	0.7650	235.3608	7.4362	$E_1 = 31.72\%$ $E_2 = 23.49\%$	$C_1 = 4.7969$ $C_2 = 0.5365$	162.5331	15.8894	$E_1 = 52.85\%$ $E_2 = 47.16\%$
IV	379.302	348.270	0.8324	283.2392	6.8898	$E_1 = 25.32\%$ $E_2 = 18.77\%$	$C_1 = 3.7389$ $C_2 = 0.5871$	222.7312	9.2737	$E_1 = 41.27\%$ $E_2 = 36.12\%$
V	1225.390	1192.540	0.8014	1074.2314	12.9000	$E_1 = 12.30\%$ $E_2 = 09.86\%$	$C_1 = 2.6597$ $C_2 = 0.7507$	1045.4735	9.1753	$E_1 = 14.68\%$ $E_2 = 12.30\%$
VI	562.754	569.758	0.8342	452.9073	7.5080	$E_1 = 19.51\%$ $E_2 = 20.50\%$	$C_1 = 7.4562$ $C_2 = 0.5694$	407.7135	8.1330	$E_1 = 27.55\%$ $E_2 = 28.44\%$
VII	1225.399	1192.548	0.7797	1129.3808	14.3090	$E_1 = 07.80\%$ $E_2 = 05.29\%$	$C_1 = 2.3368$ $C_2 = 0.7393$ $C_3 = 2.5643$	1125.4070	6.5112	$E_1 = 08.13\%$ $E_2 = 05.63\%$
VIII	562.754	569.758	0.7826	545.9367	9.8447	$E_1 = 02.99\%$ $E_2 = 04.18\%$	$C_1 = 0.7032$ $C_2 = 0.7086$ $C_3 = 6.8313$	500.8108	11.2640	$E_1 = 11.00\%$ $E_2 = 12.10\%$

where $E_1 = \frac{V(\bar{y}) - MSE(.)}{V(\bar{y})} \times 100$, $E_2 = \frac{V(\bar{y}_{ps}) - MSE(.)}{V(\bar{y}_{ps})} \times 100$

6. Counter Examples

Two populations containing three strata where the selection of the constants (C_1, C_2, C_3) from (4.4) produces results inferior to the usual estimator.

Set IX

Stratum No.	N_i	$\sum_{j=1}^{N_i} y_{ij}$	$\sum_{j=1}^{N_i} y_{ij}^2$	Mean	S_i^2	Sample Size
1	44	3453	40977	$\bar{Y}_1 = 78.4772$	$S_1^2 = 3092.9064$	
2	40	5541	4060227	$\bar{Y}_2 = 138.5250$	$S_2^2 = 84427.1789$	25
3	10	2189	556015	$\bar{Y}_3 = 218.9000$	$S_3^2 = 8538.1000$	
Total	N = 94	6105.19	4277566	$\bar{Y} = 118.9690$	$S^2 = 19745.07$	

$V(\bar{y}_{ps}) = 1191.3347$; $V(\bar{y}) = 579.7488$; $MSE\left[(\bar{y}_{ps})_{c_i}\right] = 3417.1887$ at the values $C_1 = 1.7717$, $C_2 = 1.9173$ and $C_3 = 2.6281$

Set X

Stratum No.	N_i	$\sum_{j=1}^{N_i} y_{ij}$	$\sum_{j=1}^{N_i} y_{ij}^2$	Mean	S_i^2	Sample Size
1	105	573.89	297038.25	$\bar{Y}_1 = 5.4657$	$S_1^2 = 99.5603$	
2	19	3350.89	1464905.10	$\bar{Y}_2 = 176.3631$	$S_2^2 = 66543.19$	40
3	40	5541.00	3553848.10	$\bar{Y}_3 = 138.525$	$S_3^2 = 84427.19$	
Total	N = 164	9465.79	5313791.40	$\bar{Y} = 57.7182$	$S^2 = 32614.23$	

$V(\bar{y}_{ps}) = 594.505$; $V(\bar{y}) = 616.4886$; $MSE\left[(\bar{y}_{ps})_{c_i}\right] = 2979.954$ at the values $C_1 = 5.2506$, $C_2 = 3.055$ and $C_3 = 0.3705$

6.1 Reason for Counter Examples

When C_i values obtained as a solution from (4.4) fail to satisfy (4.1) these values may not result in providing more efficient estimator. In such situation it is desireable to re-design the estimation strategy through the grouping strategy discussed in the Section 7.

7. General Grouping Strategy

Choose two positive integers r_1 and r_2 such that $r_1 + r_2 = k$ and define two groups G_1 (containing any r_1 strata) and G_2 (containing any r_2 strata). The constant C_1 is to be used for G_1 and C_2 for G_2, and then consider a modified form of estimator

$$\left[\left(\bar{y}'_{ps}\right)_{c_i}\right] = C_1\left[\sum_{i=1}^{r_1} W_i\bar{y}_i\right] + C_2\left[\sum_{i=r_1+1}^{r_1+r_2} W_i\bar{y}_i\right] \tag{7.1}$$

Remark 7.1. The problem at this juncture is that some strata may be large in terms of size (like middle income group) and some may be bigger in terms of means (like mean expenditure of high income group). Therefore, grouping of any r_1 strata among k in G_1 need not be a fruitful strategy.

7.1 Grouping Plan (1, k – 1)

Step I : Choose a row i (i = 1, 2, 3, ... k) of the PPM matrix having $p_{ij} \le 1$ for all j = 1, 2, 3, ... k + 1. Assume only one such row exists definitely.

Step II : Put corresponding i[th] stratum in the group G_1 and change its notations by $W_{(1)}$, $\bar{Y}_{(1)}$, $S^2_{(1)}$, and $\bar{y}_{(1)}$.

Step III : Put rest all the (k – 1) strata into group G_2 changing their notations $W_{(m)}, \bar{Y}_{(m)}, S^2_{(m)}, \bar{y}_{(m)}, (m = 2, 3, 4, \dots k)$.

Step IV : Use the estimator $\left[\left(\bar{y}^{(1)}_{ps}\right)_{c_i}\right] = \left[C_1 W_{(1)}\,\bar{y}_{(1)} + C_2\left\{\sum_{m=2}^{k} W_{(m)}\,\bar{y}_{(m)}\right\}\right]$

(7.2)

7.2 Grouping Plan (k – 1, 1)

This is opposite to the former

Step I : Choose a row i (i = 1, 2, 3, ... k) of the PPM matrix having $p_{ij} \ge 1$ for all $j = 1, 2, 3, \dots k + 1$. Assume a definite existence of only one such row.

Step II : Put all the (k – 1) strata into group G_1 (not including i[th] strata) with changing notations $W_{(m)}, \bar{Y}_{(m)}, S^2{(m)}, \bar{y}_{(m)}$ $(m = 1, 2, 3, \dots k - 1)$.

Step III : Put the i^{th} strata into the group G_2 with notations $W_{(k)}, \overline{Y}_{(k)}, S^2_{(k)}, \overline{y}_{(k)}$.

Step IV : Use the estimator

$$\left[\left(\overline{y}^{(2)}_{ps}\right)_{c_i}\right] = C_1\left[\sum_{m=1}^{k-1} W_{(m)}\overline{y}_{(m)}\right] + C_2\, W_{(k)}\overline{y}_{(k)} \qquad (7.3)$$

7.3 Optimum Equations

The (4.4) reduces into only two equations with two unknowns containing known elements of the PPM matrix under these plans. A solution of these provides the optimum C_1 and C_2.

Under Plan (1, k – 1)

$$C_1\left[A_{(1)}C^2_{Y_{(1)}} + W^2_{(1)}\right] + C_2\, W_{(1)}\left[\sum_{m=2}^{k} W_{(m)}\left\{\frac{\overline{Y}_{(m)}}{\overline{Y}_{(1)}}\right\}\right] = W_{(1)}\frac{\overline{Y}}{\overline{Y}_{(1)}} \qquad (7.3.1)$$

$$C_1 W_{(1)}\left\{\frac{\overline{Y}_{(1)}}{\overline{Y}}\right\}\left[1 - W_{(1)}\left\{\frac{\overline{Y}_{(1)}}{\overline{Y}}\right\}\right]$$

$$+ C_2\left[\sum_{m=2}^{k} A_{(m)}C^2_{Y_{(m)}}\left\{\frac{\overline{Y}_{(m)}}{\overline{Y}}\right\}^2 + \left(1 - W_{(1)}\frac{\overline{Y}_{(1)}}{\overline{Y}}\right)^2\right]$$

$$= \left(1 - W_{(1)}\frac{\overline{Y}_{(1)}}{\overline{Y}}\right) \qquad (7.3.2)$$

where

$$C_{Y_{(m)}} = \frac{S_{(m)}}{\overline{Y}_{(m)}}, \quad A_{(m)} = \left[\frac{1}{n} - \frac{1}{N}\right]W_{(m)} + \frac{(N-n)}{(N-1)n^2}\left(1 - W_{(m)}\right)$$

Under Plan (k – 1, 1)

$$C_1\left[\sum_{m=1}^{k-1} A_{(m)}\, C^2_{Y_{(m)}}\left\{\frac{\overline{Y}_{(m)}}{\overline{Y}}\right\}^2 + \left(1 - W_{(k)}\frac{\overline{Y}_{(k)}}{\overline{Y}}\right)^2\right]$$

$$+ C_2 W_{(k)}\left\{\frac{\overline{Y}_{(k)}}{\overline{Y}}\right\}\left[1 - W_{(k)}\left\{\frac{\overline{Y}_{(k)}}{\overline{Y}}\right\}\right] \quad \left[\left(\overline{y}_{ps}\right)_c\right] = c\sum_{i=1}^{k} W_i\overline{y}_i \qquad (7.3.3)$$

$$C_1 W_{(k)}\left[\sum_{m=1}^{k-1} W_{(m)}\left\{\frac{\overline{Y}_{(m)}}{\overline{Y}_{(k)}}\right\}\right] + C_2\left[A_{(k)}C^2_{Y_{(k)}} + W^2_{(k)}\right] = W_{(k)}\frac{\overline{Y}}{\overline{Y}_{(k)}} \qquad (7.3.4)$$

Table 7.1. Comparison of MSE and bias under grouping plans

	For Estimator $(\overline{y}_{ps})_{c_i}$								
	Without Grouping			With Grouping Plan (1, k – 1)			With Grouping Plan (k – 1, 1)		
Date Set No	Optimum C-Values	Optimum MSE & (Bias)	Gain in efficiency	Optimum C-Values	Optimum MSE & (Bias)	Gain in efficiency	Optimum C-Values	Optimum MSE & (Bias)	Gain in efficiency
VII	$C_1 = 2.564$ $C_2 = 0.739$ $C_3 = 2.336$	1125.407 (6.511)	5.63%	$C_1 = 1.171$ $C_2 = 0.758$	923.416 (14.177)	22.6%	$C_1 = 1.098$ $C_2 = 0.750$	922.366 (14.145)	22.6%
VIII	$C_1 = 6.831$ $C_2 = 0.708$ $C_3 = 0.703$	500.810 (11.264)	12.10%	$C_1 = 7.608$ $C_2 = 0.268$	167.345 (3.688)	70.6%	$C_1 = 0.824$ $C_2 = 0.757$	444.876 (9.819)	21.9%
IX	$C_1 = 1.771$ $C_2 = 0.717$ $C_3 = 2.628$	3417.190 (49.605)	–186%	$C_1 = 2.638$ $C_2 = 0.229$	382.046 (3.190)	69.9%	$C_1 = 0.387$ $C_2 = 3.315$	559.462 (4.681)	53%
X	$C_1 = 5.250$ $C_2 = 3.055$ $C_3 = 0.370$	2979.950 (35.599)	–401%	$C_1 = 2.636$ $C_2 = 0.872$	463.587 (0.920)	22%	$C_1 = 0.751$ $C_2 = 1.026$	499.607 (8.728)	15%

Remark 7.2. As special case with two strata (k = 2) and two groups, the strategy reduces into plan (1, 1) which will provide improved estimator subject to condition (4.1).

7.4 Comparison of (1, k – 1) and (k – 1, 1)

Both plans are based on different criteria of selecting the i^{th} strata, for G_1 and G_2. The plan (1, k – 1) is focused on lowest mean, biggest grouping idea of strata while plan (k – 1, 1) has a basis of biggest grouping, highest mean of the strata.

8. Conclusions

Proposed estimator (3.1) is a general class having estimators (3.2) and (3.3) as members. When information about elements of PPM matrix and coefficients of variation are known, it could be utilised in the efficient estimation by using the proposed estimator. The weight (C_i W_i) could be made optimal by solving the system of equations satisfying (4.1). Among several unknown constants C_i, if at least one or some of them are less than unity, there is a high chance of getting gain over usual post-stratified estimator. Under laid down assumptions, the optimum selection of constant C_i is easy to compute improving the efficiency. Among all data sets I to VIII, there is considerable gain in efficiency over the usual estimator when (3.1) and (3.2) are used. In spite of that, lack of gain in efficiency, using (3.1), is shown in two counter examples. To cope with this, a general strategy of grouping strata is proposed which is found effective and easy in application. The strategy has grouping plans $(1, k-1)$ and $(k-1, 1)$ and both generate efficient estimators on those data sets where the usual (3.1) proved less efficient. While comparing two plans over same data sets, it is found that plan (1, k – 1) is better than plan (k – 1, 1), but all together both are recommendable over the situation of not using the grouping strategy. Also, both plans are effective in reducing the bias component of the estimator (3.1).

ACKNOWLEDGEMENT

Authors are thankful to referee for critical comments and suggestions which have improved the contents of the manuscript.

REFERENCES

[1] Agrawal, M.C. and Panda, K.B. (1993). An efficient estimator in post-stratification. *Metron*, **51**, 3-4, 179-187.

[2] Gelman, A. and Little, T.C. (1997). Post-stratification into many categories using hierarchical logistic regression survey method. *Survey Methodology*.

[3] Holt, D. and Smith, T.M.F. (1979). Post-stratification. *J. Roy. Statist. Soc.*, A **142**, 33-36.

[4] Jagers, P., Oden A. and Trulsson, L. (1985). Post-stratification and ratio estimator. *Int. Stat. Rev.*, **53**, 221-238.

[5] Jagers, P. (1986). Post-stratification against bias in sampling. *Int. Stat. Rev.*, **54**, 159-167.

[6] Lazzerni, I.C. and Little, R.J.A. (1997). Random effect models for smoothing post-stratification weights. *Journal of Official Statistics.*

[7] Little, R.J.A. (1993). Post-stratification : A modeler's perspective. *Jour. Amer. Statist. Assoc.*, **88**, 1001-1012.

[8] Sarndal, C.E., Swensson, B. and Wertman, J. (1992). *Model Assisted Survey Sampling.* Springer-Verlag, New York.

[9] Searls (1964). The utilisation of a known coefficient of variation in the estimation procedure. *Jour. Amer. Statist. Assoc.*, **59**, 1225-1226.

[10] Shukla, D. and Trivedi, Manish (1999). A new estimator in post-stratification sampling scheme, Baysian Analysis. *Proceedings of NSBA-TA*, 16-18 Jan. 1999.

[11] Shukla, D. and Trivedi, Manish (2001). Mean estimation in deeply stratified population under post-stratification. *Jour. Ind. Soc. Agril. Stat.*, **54(2)**, 221-235.

[12] Sukhatme, P.V, Sukhatme, B.V., Sukhatme, S. and Asok, C. (1984). *Sampling Theory of Survey with Applications.* Iowa State University Press, Indian Society of Agricultural Statistics, New Delhi.