Jour. Ind. Soc. Ag. Statistics
56(3), 2003 : 222-229

Post Stratified Estimator in Two Stage Design using Auxiliary Information

M.S. Narang, P.C. Mehrotra[1] and A.K. Bhatia[2]
Indian Agricultural Statistics Research Institute, New Delhi-110012
(Received : December, 1996)

SUMMARY

Post stratification in two stage sampling utilizing auxiliary information at the second stage has been attempted. Units at both the stages are selected with simple random sampling without replacement. It is empirically demonstrated that the suggested procedure not only provides estimate of the character under study with higher precision but also improves the precision of the estimate (pooled over the strata) vis-a-vis the common post stratified two stage sampling procedure not utilizing auxiliary information.

Key words : Post-stratification, Two-stage sampling, Auxiliary information.

1. Introduction

The problem of post stratification in two-stage sampling with post-stratification on the basis of ultimate-stage units was studied by Mehrotra [2] and Kumar [3]. In many practical situations auxiliary information is available which under certain conditions could be used with advantage.

It is common knowledge that estimators which utilize auxiliary information could be highly efficient when such data are available. In this paper an attempt has been made to develop estimator of population parameter in a stratified two stage design with post-stratification at the second stage using available auxiliary information. The estimator used was the conventional ratio estimator.

2. Preliminaries

2.1 Let the population consists of N distinct and identifiable primary stage units (psu) with the i-th psu having M_i second-stage units (ssu). Let a simple

---

1 427, Deepali, Pitampura, Delhi 110034
2 NBAGR, Karnal

random sample (SRS) of n psu's be drawn from the population and in turn a SRS of m_i ssu's be drawn from the i-th selected psu.

Stratify the selected ssu's in each of the sampled psu into k strata on the basis of some characteristic of the ssu's (stratifying variable). At this stage i.e. after the ssu's have been stratified it would be interesting to examine the nature and number of post-strata formed within each of the selected n psu's. It is possible that a selected psu may contain more than one ssu belonging to the same stratum. Again, a selected psu may not contain any ssu belonging to one or more of the k strata and as a consequence that psu will not find a place in those strata. Further, a selected psu may possess at least one ssu pertaining to two or more of the k strata or to each of the k strata and accordingly will find a place in each of the two or more of the k strata or to each of the k strata, as the case may be.

Let the number of sample psu's containing at least one ssu belonging to the h-th stratum (h = 1, 2,------k) be denoted by n_h ($0 < n_h \leq n$) and let $m_{i(h)}$ denote the number of ssu's from the i-th psu falling in the h-th stratum ($0 < m_{i(h)} \; \overline{X} \; m_i$). Also, Let N_h and $M_{i(h)}$ denote the corresponding numbers in the population. Further, let the number of psu's having a ssu belonging to both the h-th and h'-th strata be denoted by $n_{hh'}$ and that in the population by $N_{hh'}$. Likewise let $n_{h(hh')}$ and $n_{h'(hh')}$ denote the number of psu's having a ssu belonging only to the h-th or h'-th stratum respectively for the strata pair (hh') with $N_{h(hh')}$ and $N_{h'(hh')}$ as the corresponding numbers in the population. Therefore, it is obvious to see that

$$N_h = N_{hh'} + N_{h(h\,h')}, \quad N_{h'} = N_{hh'} + N_{h'(hh')}$$

and $$n_h = n_{hh'} + n_{h(h\,h')}, \quad n_{h'} = n_{hh'} + n_{h'(hh')} \tag{1}$$

Let y be the study character for which the population total Y is to be estimated. An auxiliary character x, highly correlated with y, is assumed to be available for the population i.e. ($\overline{X}_{i(h)}$ and $M_{i(h)}$), the population mean for the auxiliary character and the size for the h-th stratum in the i-th psu. Ratio estimator, which utilizes auxiliary information at the estimation stage may then be usefully applied for estimating Y.

Now, we proceed to work out the post-stratified ratio estimator with post-stratification at the second stage on the basis of second-stage units under the assumption that the population mean of auxiliary character ($\overline{X}_{i(h)}$) as well as strata sizes ($M_{i(h)}$) are known.

2.2 The expectation and variance of the proposed estimator would be obtained as follows

Let E_1 and E_2 denote the expectations at first and second stage of selection respectively, whereas $E_2(\cdot \mid m_{i(h)})$ denote the conditional expectation at the second stage for a given value of $m_{i(h)}$. And let V_1 and V_2 denote the variances at

the first and second stage of selection respectively. Also, let $V_2(\cdot \mid m_{i(h)})$ denote the conditional variance at the second stage for a given value of $m_{i(h)}$.

3. Proposed Estimator

Let $x_{ij(h)}$ and $y_{ij(h)}$ denote the value of auxillary character and character under study respectively for the j-th ssu of the i-th psu in the h-th stratum.

An estimator of the population total for the h-th stratum namely $Y_{(h)}$ for the study character y will be given by

$$\hat{Y}_{RS(h)} = \frac{N_h}{n_h} \sum_{i=1}^{n_h} M_{i(h)} \bar{y}_{Ri(h)} \tag{2}$$

the suffix RS implies that the estimator is a ratio type and with stratification.

where $\bar{y}_{Ri(h)} = \frac{\bar{y}_{i(h)}}{\bar{x}_{i(h)}} \bar{X}_{i(h)}$

and $\bar{y}_{i(h)} = \frac{1}{m_{i(h)}} \sum_{j=1}^{m_{i(h)}} y_{ij(h)}, \bar{x}_{i(h)} = \frac{1}{m_{i(h)}} \sum_{j=1}^{m_i(h)} x_{ij(h)}$

And an estimator of the population total Y will be given by

$$\hat{Y}_{RS} = \sum_{h=1}^{k} \frac{N_h}{n_h} \sum_{i=1}^{n_h} M_{i(h)} \bar{y}_{R_{i(h)}} \tag{3}$$

4. Bias of the Proposed Estimator

The expectation of (3) would be

$$E(\hat{Y}_{RS}) = E_1E_2E_2 \sum_{h=1}^{k} \frac{N_h}{n_h} \sum_{i=1}^{n_h} (M_{i(h)} \bar{y}_{Ri(h)} \mid m_{i(h)})$$

Following Stephen [4] and Sukhatme and Sukhatme [5] we get

$$\text{Bias } (\hat{Y}_{RS}) = \sum_{h=1}^{k} \sum_{i=1}^{N_h} \frac{M_i f_i}{m_i} w_{i(h)} \sigma_i(h) + \sum_{h=1}^{k} \sum_{i=1}^{N_h} \frac{M_i f_i}{m_i^2} (1 - w_{i(h)}) \sigma_{i(h)} \tag{4}$$

where $w_{i(h)} = M_{i(h)} / M_i$, $\sigma_{i(h)} = \bar{Y}_{i(h)} \delta_{i(h)} / w_{i(h)}$ and $f_i = 1 - m_i / M_i$

$$\delta_{i(h)} = C^2_{i(h)x} - \rho_{i(h)xy} C_{i(h)x} C_{i(h)y}$$

$$C_{i(h)x} = S_{i(h)x} / \bar{X}_{i(h)}, C_{i(h)y} = S_{i(h)y} / \bar{Y}_{i(h)}$$

$$\rho_{i(h)xy} = \frac{E\left(y_{ij(h)} - \bar{y}_{i(h)}\right)\left(x_{ij(h)} - \bar{x}_{i(h)}\right)}{\sqrt{E\left(y_{ij(h)} - \bar{y}_{i(h)}\right)^2 E\left(x_{ij(h)} - \bar{x}_{i(h)}\right)^2}}$$

The first term, in equation (4) for bias in $\hat{Y}_{RS}$, is the bias as in the usual separate ratio estimator under proportional allocation and the second term is the additional component in bias arising from randomness in the $m_{i(h)}$'s on account of post-stratification.

5. Estimator of Bias

$$\hat{B}(\hat{Y}_{RS}) = \sum_{h=1}^{k} \frac{N_h}{n_h} \sum_{i=1}^{n_h} \frac{M_i f_i}{m_i} \hat{T}_{i(h)} + \sum_{h=1}^{k} \frac{N_n}{n_h} \sum_{i=1}^{n_h} \frac{M_i f_i}{m_i^2} \frac{(1 - w_{i(h)})}{w_{i(h)}} \hat{T}_{i(h)} \quad (5)$$

where $$\hat{T}_{i(h)} = \text{Est.}\left\{\bar{Y}_{i(h)}\ \delta_{i(h)}\right\} = \frac{1}{\bar{X}_{i(h)}}\left\{\hat{R}_{i(h)}\, s^2_{i(h)x} - s_{i(h)xy}\right\}$$

$$\hat{R}_{i(h)} = \frac{\bar{y}_{i(h)}}{\bar{x}_{i(h)}}\ ,\ s^2_{i(h)x} = \frac{1}{m_{i(h)} - 1} \sum_{j=1}^{m_{i(h)}} (x_{ij(h)} - \bar{X}_{i(h)})^2$$

$$s_{i(h)xy} = \frac{1}{m_{i(h)} - 1} \sum_{j=1}^{m_{i(h)}} (x_{ij(h)} - \bar{X}_{i(h)})(y_{ij(h)} - \bar{Y}_{i(h)})$$

$$= \frac{1}{m_{i(h)} - 1} \sum_{j=1}^{m_{i(h)}} (x_{ij(h)} - \bar{X}_{i(h)})\,(y_{ij(h)} - \bar{Y}_{i(h)})$$

6. Variance of the Estimator

The variance of the proposed estimator would be

$$V\left(\hat{Y}_{RS}\right) = \sum_{h=1}^{k} N_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right) S'^2_{b(h)} + \sum_{h \neq h'}^{k} N_h\, N_{h'} \left(\frac{1}{n_{hh'}} - \frac{1}{N_{hh'}}\right) S'_{b(hh')} \frac{n^2_{hh'}}{n_h n_{h'}}$$

$$+ \sum_{h=1}^{k} \frac{N_h}{n_h} \sum_{i=1}^{N_h} M^2_{i(h)} \left\{\frac{1}{m_i w_{i(h)}} + \frac{f_i(1 - w_{i(h)})}{m_i^2 w^2_{i(h)}} - \frac{1}{M_i w_{i(h)}}\right\} \phi_{i(h)}$$

$$+ \sum_{h=1}^{k} \frac{N_h}{n_h} \sum_{i=1}^{N_h} \Psi^2_{i(h)} \left\{\frac{f_i\left(1 - w_{i(h)}\right)}{m_i^3 w^3_{i(h)}}\right\}$$

$$-\sum_{h\neq h'}^{k}\frac{N_h N_{h'}}{n_h n_{h'}}\frac{n_{hh'}}{N_{hh'}}\sum_{i=1}^{N_{hh'}}\psi_{i(h)}\psi_{i(h')}\left(\frac{f_i}{m_i^3 w_{i(h)} w_{i(h')}}\right) \qquad (6)$$

where $S'^2_{b(h)} = \frac{1}{N_h - 1}\sum_{i=1}^{N_h}\left[Y'_{i(h)} - \frac{1}{N_h}\sum_{i=1}^{N_h}Y'_{i(h)}\right]^2$

$$S'_{b(hh')} = \frac{1}{N_{hh'}-1}\sum_{i=1}^{N_{hh'}}\left[Y'_{i(h')} - \frac{1}{N_{hh'}}\sum_{i=1}^{N_{hh'}}Y'_{i(h')}\right]\left[Y'_{i(h')} - \frac{1}{N_{hh'}}\sum_{i=1}^{N_{hh'}}Y'_{i(h')}\right]$$

$$\phi_{i(h)} = S^2_{i(h)y} - 2R_{i(h)}\, S_{i(h)yx} + R^2_{i(h)}\, S^2_{i(h)x}$$

$$S^2_{i(h)y} = \frac{1}{M_{i(h)} - 1}\sum_{j=1}^{M_{i(h)}}\left(y_{ij(h)} - \overline{Y}_{i(h)}\right)^2$$

$$\psi_{i(h)} = M_{i(h)}\,\overline{Y}_{i(h)}\,\delta_{i(h)}\ ;\ \psi_{i(h')} = M_{i(h')}\,\overline{Y}_{i(h')}\,\delta_{i(h')} \text{ and } w_{i(h')} = \frac{M_{i(h')}}{M_i}$$

The average value of $V\left(\hat{Y}_{RS}\right)$ in repeated sample of size n (unconditional variance) will be

$$V\left(\hat{Y}_{RS}\right) = EV\left(\hat{Y}_{RS}\middle|n_1, n_2 \ldots n_k\right) + VE\left(\hat{Y}_{RS}\middle|n_1, n_2 \ldots n_k\right)$$

Since, $E\left(\hat{Y}_{RS}\middle|n_1, n_2 \ldots n_k\right)$ is a constant independent of n_h, we see that

$$V\left(\hat{Y}_{RS}\right) = EV\left(\hat{Y}_{RS}\middle|n_1, n_2 \ldots n_k\right)$$

$$= N^2\sum_{h=1}^{k}\left(\frac{1}{n} - \frac{1}{N}\right)w_h S'^2_{b(h)} + \frac{N^2 f}{n^2}\sum_{h=1}^{k}\left(1 - w_h\right)S'^2_{b(h)}$$

$$+ N^2\sum_{h\neq h'}^{k}\left[w_{hh'}\left\{1 + \frac{f\left(1-w_h\right)}{nw_h} + \frac{f\left(1-w_{h'}\right)}{nw_{h'}} + \frac{f}{n}\right\}\left(\frac{1}{n} - \frac{1}{N}\right)\right.$$

$$\left. - \frac{f}{nN}\left(1 + w_{hh'}\right)\right]S'_{b(hh')}$$

$$+ N^2\sum_{h=1}^{k}\frac{1}{nN}\left(1 + \frac{f\left(1-w_h\right)}{nw_h}\right)\sum_{i=1}^{N_h}\frac{f_i M_i^2}{m_i}w_{i(h)}\,\phi_{i(h)}$$

$$+ N^2\sum_{h=1}^{k}\frac{1}{nN}\left(1 + \frac{f\left(1-w_h\right)}{nw_h}\right)\sum_{i=1}^{N_h}\frac{f_i M_i^2}{m_i^2}\left(1 - w_{i(h)}\right)\phi_{i(h)}$$

$$+ N^2 \sum_{h=1}^{k} \frac{1}{nN}\left(1+\frac{f(1-w_h)}{nw_h}\right) \sum_{i=1}^{N_h} \psi_{i(h)}^2 \frac{f_i(1-w_{i(h)})}{m_i^3 w_{i(h)}^3}$$

$$- N^2 \sum_{h=1}^{k} \frac{1}{nN}\left\{1+\frac{f(1-w_h)}{nw_h}+\frac{f(1-w_{h'})}{nw_{h'}}+\frac{f}{n}\right\}$$

$$\times \sum_{i=1}^{N_{hh'}} \psi_{i(h)}\psi_{i(h')} \frac{f_i}{m_i^3 w_{i(h)} w_{i(h')}} \qquad (7)$$

The first term in (7) is the variance of the total of a stratified two-stage sample taken with proportional allocation. The second term represents the adjustment due to post-stratification at the first-stage. The third factor has resulted as the psu's are cutting across the boundaries of the strata. The fourth term is the usual expression for the variance of the separate ratio estimator under proportional allocation. The later three terms are the additional components of variance due to the involvement of post-stratification at the ultimate stage. However, the last two terms may be ignored being of $0(1/m_i^3)$.

7. Estimate of the Variance

An estimator of the variance of the estimate following Des Raj [1] will be

$$\hat{V}(\hat{Y}_{RS}) = \sum_{h=1}^{k} N_h^2 \left(\frac{1}{n_h}-\frac{1}{N_h}\right)\frac{1}{n_h-1}\sum_{i}^{n_h}\left(y_{i(h)}-\bar{y}_{(h)}\right)^2$$

$$\times \sum_{h\neq h}^{k} N_h N_{h'} \frac{n_{hh'}^2}{n_h n_{h'}}\left(\frac{1}{n_{hh'}}-\frac{1}{N_{hh'}}\right)\frac{1}{n_{hh'}-1}\sum_{i}^{n_{hh'}}\left(y_{i(h)}-\bar{y}_{(h)}\right)\left(y_{i(h')}-\bar{y}_{(h')}\right)$$

$$+ \sum_{h=1}^{k} \frac{N_h}{n_h}\sum_{i}^{n_h} M_{i(h)}^2 \left\{\left(\frac{1}{m_{i(h)}}-\frac{1}{M_{i(h)}}\right)\left(s_{i(h)y}^2 - 2\hat{R}_{i(h)}s_{i(h)xy} + \hat{R}_{i(h)}^2 s_{i(h)x}^2\right)\right\}$$

where

$$\hat{R}_{i(h)} = \left[\frac{1}{m_{i(h)}}\sum_{j}^{m_{i(h)}} y_{ij(h)} \Big/ \frac{1}{m_{i(h)}}\sum_{j}^{m_{i(h)}} x_{ij(h)}\right]$$

$$y_{i(h)} = M_{i(h)}\,\hat{R}_{i(h)}\,\bar{x}_{i(h)} \;;\; \bar{y}_{(h)} = \frac{1}{n_h}\sum_{i}^{n_h} y_{i(h)}$$

$$s_{i(h)x}^2 = \frac{1}{m_{i(h)}-1}\sum_{j}^{m_{i(h)}}\left(x_{ij(h)}-\bar{x}_{i(h)}\right)^2$$

$$s^2_{i(h)y} = \frac{1}{m_{i(h)} - 1} \sum_{j=1}^{m_{i(h)}} \left(y_{ij(h)} - \bar{y}_{i(h)}\right)^2$$

$$s_{i(h)xy} = \frac{1}{m_{i(h)} - 1} \sum_{j}^{m_{i(h)}} \left(y_{ij(h)} - \bar{y}_{i(h)}\right)\left(x_{ij(h)} - \bar{x}_{i(h)}\right); \quad \bar{y}_{i(h)} = \frac{1}{m_{i(h)}} \sum_{j}^{m_{i(h)}} y_{ij(h)}$$

$$\bar{x}_{i(h)} = \frac{1}{m_{i(h)}} \sum_{j}^{m_{i(h)}} x_{ij(h)}$$

8. *Empirical Illustration*

In order to demonstrate numerically the relative performance of the proposed estimator of the population total utilizing auxiliary information we have considered a hypothetical population conducting a simulation study. For this purpose we have generated a population of 50 psu's each with varying number of ssu's aggregating to a total of 5000 ssu's. The hypothetical population was generated by varying the various parameters obtained from data for Kapurthala Tehsil of District Kapurthala in Punjab. The variable under study was irrigated area and the auxiliary variable was total area in a village. The correlation between these two variables was 0.87. We drew a sample of 20 psu's and from each sampled psu's a 10 percent sample of ssu's on an average. Thus, 20 psu's and 200 ssu's were selected with an overall sampling fraction of 4 percent. The estimate of population total, its bias, variance and mean square error as also the relative efficiency of the proposed estimator including the percent standard error were worked out from the sample. The results are presented below in Table 8.1.

Table 8.1. Relative performance of two-stage ratio estimator with post stratification at the second stage using auxiliary information

Estimator	Estimate 10^4	Bias 10^4	Variance 10^7	Mean square error 10^7	Relative efficiency w.r.t.1	SE %	Relative bias %
1. Without use of auxiliary information	115.29	–	454.81	–	100	5.8	–
2. Utilizing auxiliary information	115.28	0.56	426.24	429.35	106	5.7	0.5

From the above table it is seen that the proposed estimator is more precise compared to the one without the use of auxiliary information and the S.E is quite low being 5.7 percent. Also, the magnitude of the bias is quite small and relative bias (%) is negligible.

9. Conclusion

The methodology developed and presented in this paper on post stratified two stage sampling with stratification at the second stage using auxiliary information will be of immense use in many situations often encountered in agriculture, education and social studies. Attempts in the development of such methodologies were very much restricted in the past due to complex calculations of the variances expressions etc. of the estimators. Now with availability of fast electronic computers such constraints are no more there.

REFERENCES

[1] Des Raj (1968). *Sampling Theory.* Tata McGraw Hill Publishing Co., New Delhi.

[2] Mehrotra, P.C. (1993). A scheme for post-stratification in two stage sampling. *Jour. Ind. Soc. Agril. Stat.*, **45 (1).**

[3] Kumar, K. (1989). On post stratification for improvement of estimation procedure in a multi-stage sampling design. Unpublished Ph.D. Thesis submitted to P.G. School, IARI, New Delhi.

[4] Stephan, F.F. (1945). The expected value and variance of the reciprocal and other negative powers of a positive Bernoullian. *Ann. Math. Statist.*, **16.**

[5] Sukhatme P.V. and Sukhatme B.V. (1970). *Sampling Theory of Surveys with Applications.* (second revised edition), Iowa State University Press, Ames, Iowa, U.S.A and Indian Society of Agricultural Statistics, New Delhi, India.