ON POST-STRATIFICATION FOR CLUSTER SAMPLING

P.C. MEHROTRA, A.K. SRIVASTAVA AND K.K. TYAGI

Indian Agricultural Statistics Research Institute, New Delhi

(Received : September, 1982)

SUMMARY

Post-stratification in uni-stage unequal cluster sampling on the basis of the elements of the selected clusters has been discussed. It has been empirically demonstrated that the suggested procedure not only provides estimates of the character under study according to the strata variable, but also improves the precision of the overall estimate compared to the usual cluster sampling procedure.

INTRODUCTION

In cluster sampling, clusters are taken as sampling units and all the elements of selected clusters are observed. In some situations, stratification may be needed or it may be available for elements within the clusters. The stratum to which an elements belongs may not be known until the data have been collected in case of variables suitable for stratification. Variables like age, sex, race, educational level, size of holding etc. are common examples. The strata sizes may be obtained fairly accurately from official statistics, but the units can be classified into strata only after the sample is selected. Discussion on post-stratification are broadly available in text books such as [1] and [2]. In this paper, we discuss post-stratification in cluster sampling on the basis of elements of the selected clusters.

2. SUGGESTED PROCEDURE

Let the population consist of N clusters where the ith cluster contains M_i elements ($i=1, 2,\ldots, N$). Let a sample of n clusters be drawn with simple random sampling without replacement.

Classify the elements of each of the sampled n clusters into k strata with respect to some characteristic of the elements of the

sample clusters. It is possible that some of the selected clusters may not have any elements belonging to some of the strata. Let the number of sample clusters, containing at least one element belonging to the h th stratum $(h=1, 2,\ldots\ldots k)$ be denoted by n_h, $(0 \leqslant n_h \leqslant n)$ and let N_h denote the corresponding number in the population. Further, let the number of elements of the i-th cluster that fall in the h-th stratum out of the total number of elements M_i of that cluster, be denoted by $M_{i(h)}$, $(0 \leqslant M_{i(h)} \leqslant M_i)$

3. Proposed Estimator

When we do not take into account the stratification, an unbiased estimator of the population total is

$$\hat{\bar{Y}}_c = \frac{N}{n} \sum_{i}^{n} M_i \bar{Y}_i \tag{1}$$

$$\text{where } \bar{Y}_i = \frac{1}{M_i} \sum_{j=1}^{M_i} y_{ij}$$

and y_{ij} is the value of the jth element $(j=1, 2,\ldots M_i)$ in the ith cluster $(i=1,2,\ldots,N)$ and the subscript 'c' indicates that the estimator relates to cluster sampling. Its variance is

$$V(\hat{\bar{Y}}_c) = N^2\left(\frac{1}{n} - \frac{1}{N}\right) S^2 \tag{2}$$

$$\text{where } S^2 = \frac{1}{N-1} \sum_{i=1}^{N} \left(M_i \bar{Y}_i - \sum_{i=1}^{N} \frac{M_i \bar{Y}_i}{N} \right)^2$$

However, when we consider the post-stratification within clusters on the basis of elements, an estimator of the population total $Y_{(h)}$ for the h th stratum in respect of the character under study y may be considered as

$$\frac{Nh}{nh} \sum_{i}^{nh} \bar{Y}_{i(h)} M_{i(h)}$$

where $\bar{Y}_{i(h)}$ is the mean value of the study character for the $M_{i(h)}$ elements of the ith cluster falling in the hth stratum and is given by

$$\bar{Y}_{i(h)}=\frac{1}{M_{i(h)}}\sum_{j=1}^{M_i(_h)} y_{ij(h)}$$

and $y_{ij(h)}$ is the value of the jth element of the ith cluster falling in the hth stratum.

An unbiased estimator of the population total Y is given by

$$\hat{Y}_{cs}=\sum_{h=1}^{k}\frac{N_h}{n_h}\sum_{i}^{n_h}\bar{Y}_{i(h)}\,M_{i(h)}=\sum_{h=1}^{k}\frac{N_h}{n_h}\sum_{i}^{n_h}Y_{i(h)} \qquad (3)$$

where $Y_{i(h)}$ is the total for the study character in the ith cluster falling in the hth stratum and the subscript 'cs' indicates that the estimator relates to post-stratification in cluster sampling on the basis of elements.

4. Variance of the Estimator

For fixed $n_1, n_2 \ldots, n_k$ the variance of the proposed estimator will be given by

$$V(\hat{Y}_{cs}/n_h's)=\sum_{h=i}^{k}N_h^2\left(\frac{1}{n_h}\frac{1}{N_h}\right)S_h^2+$$

$$\sum_{h\neq h'}^{k}\mathrm{Cov}\left\{\left(\frac{N_h}{n_h}\sum_{i}^{n_h}Y_{i(h)},\frac{N_h'}{n_h'}\sum_{i}^{n_h'}Y_i(_{h'})\right)\Big/r_{h'}s\right.$$

where $$S_h^2=\frac{1}{N_h-1}\sum_{i=1}^{N_h}\left(Y_{i(h)}-\frac{1}{N_h}\sum_{i=1}^{N_h}Y_{i(h)}\right)^2$$

A cluster may contain elements belonging to one or more of the k strata. Let the number of clusters having an element belonging to both the hth and h'th strata in the sample be denoted by $n_{hh'}$ and that in the population by $N_{hh'}$. Let $n_h(_{hh'})$ and $n_h'(_{hh'})$ denote the number of clusters having an element belonging only to the hth and h'th strata respectively for the strata pair (hh'). Define $N_h\,(_{hh'})$ and

$N_h'(_{hh'})$ as the corresponding numbers in the population. Obviously

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

$$\text{and } n_h = n_{hh'} + n_{h(hh')}$$

The conditional covariance for given n_h's at (4) above may be written as

$$\frac{N_h N_{h'}}{n_h n_{h'}} \operatorname{Cov}\left\{ \sum_{i}^{n_{h(hh')}} Y_{i(h)} + \sum_{i}^{n_{hh'}} Y_{i(h)},\ \sum_{i}^{n_{h'(hh')}} Y_{i(h')} + \sum_{i}^{n_{hh'}} Y_{i(h')} \right\}$$

$$= n_{hh'}{}^2 \frac{N_h N_{h'}}{n_h n_{h'}} \operatorname{Cov}\left\{ \frac{1}{n_{hh'}} \sum_{i}^{n_{hh'}} Y_{i(h)},\ \frac{1}{n_{hh'}} \sum_{i}^{n_{hh'}} Y_{i(h')} \right\}$$

$$= \frac{N_h N_{h'}}{n_h n_{h'}} nhh'^2 \left(\frac{1}{n_{hh'}} - \frac{1}{N_{hh'}} \right) S_{hh'}$$

where,

$$S_{hh'} = \frac{1}{N_{hh'} + 1} \sum_{i=1}^{n_{hh'}} \left((Y_{i(h)} - \frac{1}{N_{hh'}} \sum_{i=1}^{N_{hh'}} Y_{i(h)}) \right)$$

$$\left(Y_{i(h')} - \frac{1}{N_{hh'}} \sum_{i=1}^{N_{hh'}} Y_{i(h')} \right)$$

It is, of course assumed that N_h and $N_{(hh')}$, are known.

$V(\hat{Y}_{cs})$ is thus given by

$$V(\hat{Y}_{cs}) = \sum_{h=1}^{k} N_h^2 \left\{ E\left(\frac{1}{n_h}\right) - \frac{1}{N_h} \right\} S_h^2 +$$

$$\sum_{h \neq h'}^{k} N_h N_{h'} \left[E\left(\frac{n_{hh'}}{n_h n_{h'}} \right) - \frac{1}{n_{hh'}} E\left(\frac{n_{hh'}{}^2}{n_h n_{h'}} \right) \right] S_{hh'} \qquad (5)$$

The expectations required in the above equation may be obtained to the first order of appropriation utilizing the usual technique in ratio method of estimation as follows :

$$E\left(\frac{1}{n_h}\right)=\frac{1}{nw_h}\left(1+\frac{1-w_h}{nw_h}\right) \tag{6}$$

$$E\left\{\frac{n_{hh'}}{n_h n_{h'}}\right\}=\frac{w_{hh'}}{nw_h w_{h'}}\left(1+\frac{1-w_{h'}}{nwh'}+\frac{1-wh}{nwh}+\frac{1}{n}\right) \tag{7}$$

$$E\left\{\frac{n_{hh'}^2}{n_h n_{h'}}\right)=\frac{w_{hh'}^2}{w_h w_{h'}}\left(1+\frac{1-w_{h'}}{nw_h}+\frac{1-w_h}{nw_h}+\frac{1-w_{hh'}}{nw_{hh'}}+\frac{3}{n}\right) \tag{8}$$

where

$$w_h=\frac{N_h}{N}, \quad wh'=\frac{N_{h'}}{N}$$

and $w_{hh'}=\frac{N_{hh'}}{N}$

Substituting from (6), (7) and (8) in the variance expression (5), we get

$$V(\hat{Y}_{cs})=\sum_{h=1}^{k} N_h^2\left\{\frac{1}{nw_h}\left(1+\frac{1-w_h}{nw_h}\right)-\frac{1}{N}\right\}S_h^2+$$

$$\sum_{h\neq h'}^{k} N_h N_{h'}\left(\frac{w_{hh'}}{nw_h w_{h'}}\left\{1+\frac{1-w_{h'}}{nw_{h'}}+\frac{1-w_h}{nw_h}+\frac{1}{n}\right\}-\right.$$

$$\left.\frac{1}{N_{hh'}}\left\{\frac{w_{hh'}^2}{w_h w_{h'}}\left(1+\frac{1-w_{h'}}{nw_{h'}}+\frac{1-w_h}{nw_h}+\frac{1-w_{hh'}}{nw_{hh'}}+\frac{3}{n}\right)\right\}\right)S_{hh'}$$

$$=N^2\sum_{h=1}^{k}\left(\frac{1}{n}-\frac{1}{N}\right)w_h S_h^2+\frac{N^2}{n^2}\sum_{h=1}^{k}(1-wh)S_h^2+$$

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

$$-N^2\sum_{h\neq h'}^{k}\frac{1}{nN}(1+w_{hh'})S_{hh'} \tag{9}$$

The variance expression given at (9) is seen to be made up of four components. The first component is the variance of a stratified sample taken with proportional allocation, the second represents the adjustment due to post-stratification of the sample clusters and the last two terms the combined effect of splitting the clusters and of post-stratification of element.

5. Empirical Illustration

To show the usefulness of the suggested procedure, we present in the following paragraphs a numerical illustration. For this purpose, we have considered a population of 120 clusters in a tehsil. The elements of the clusters are holding of varying sizes and the character under study being the area under wheat crop in units of 10 hectares during the rabi season. Consider a sample of 15 clusters with simple random sampling without replacement. On the basis of the character-size of holding-of the elements of the sample clusters the elements are classified into three strata viz; first stratum with holdings of small size, second stratum with holding of medium size and the third stratum with holdings of large size. The value of n_h, $n_{h(hh')}$ etc. in the sample and the corresponding values in the population were as given in the following table.

	N	N_1	N_2	N_3	N_{12}	N_{13}	N_{23}
Population	120	103	95	97	86	84	81
	n	n_1	n_2	n_3	n_{12}	n_{13}	n_{23}
Sample	15	13	11	12	11	10	10

The estimates of stratum-wise and overall totals and their variances are given in the following table.

Suggested procedure	*Stratum number*	Y	$V(\hat{Y})$
	First	2163	177452
	Second	4232	397374
	Third	11292	2536491
	Overall	17687	3111317
Usual procedure	Overall	17280	5731167

Thus, it is seen that the suggested procedure has not only provided estimates of the total according to the strata variable but has

also improved the precision of the estimate of the overall total. The efficiency of the suggested procedure compared to that of the usual procedure is 185 per cent.

ACKNOWLEDGEMENT

The authors are thankful to the referees for their useful suggestions.

REFERENCES

[1] Cochran, W.G. (1977) : Sampling techniques, (3rd Ed.) *John Wiley and Sons. New York.*

[2] Sukhatme, P.V. and Sukhatme, B.V. (1970) : Sampling theory of surveys with applications, (2nd Ed.) *Iowa State University Press, Ames, Iowa, U.S.A.*