USE OF ANCILLARY INFORMATION IN COLLAPSING OF STRATA

By

S. Mohanty and Padam Singh

Institute of Agricultural Research Statistics, New Delhi

1. Introduction

If the population is highly variable, resort may be made for deep stratification so that only one unit is sampled from each stratum. In this case usual formula for estimating $V(\bar{y}_{st})$ cannot be used. Cochran (1963) suggested a technique called the method of "Collapsed strata" for estimating $V(\bar{y}_{st})$ which consists in grouping the strata in pairs such that two strata in pair should have equal mean and size. As these assumptions are too restrictive, Seth (1966) suggested a modification to Cochran's procedure.

In the present paper a general method of collapsing any number of strata and the use of an ancillary character for collapsing have been suggested. Further, the same ancillay character is used in improving the estimate of the parameter under study.

2. The Method

Let L be the number of strata and N_i, $\bar{Y}_i, S_i^2$ denote the number of units in the ith stratum, ith stratum mean and variance of the character Y under study respectively. Let $y_1, y_2, \ldots\ldots, y_L$ be the sample drawn by selecting one unit from each stratum. The stratified random sampling estimate is given by

$$\bar{y}_{st}=\sum_{i=1}^{L} \frac{N_i}{N} y_i \qquad \ldots(2.1)$$

and its variance is given by

$$V(\bar{y}_{st})=\sum_{i=1}^{L} \frac{N_i^2}{N^2} S_i^2 \qquad \ldots(2.2)$$

The estimate of (2.2) as proposed by Cochran (1963) is

$$\sum_{i=1}^{L/2} \frac{N_i^2}{N^2}(y_{i1}-y_{i2})^2 \qquad \ldots(2.3)$$

where $N_{i1}=N_{i2}=N_i$, i_1th, i_2th strata are collapsed to form the ith stratum and L is assumed to be an even integer. The above estimate is unbiased when $\bar{Y}_{i1}=\bar{Y}_{i2}$. Seth (1966) suggested another estimate of (2.2) given by

$$\frac{1}{N^2}\sum_{i=1}^{L/2}(N_{i1}{}^p y_{i1}-N_{i2}{}^p y_{i2})(N_{i1}{}^{2-p} y_{i1}-N_{i2}{}^{2-p} y_{i2}), \qquad \ldots(2.4)$$

where the strata are collapsed in such a way that the strata means are proportional to some power 'p' of the number of units in the stratum. The bias of the estimate (2.4) is small when

$$N_{i1}{}^p\bar{Y}_{i1}-N_{i2}{}^p\bar{Y}_{i2} \text{ is negligible.}$$

Now the general method of collapsing consists of collapsing k_i strata to form a new ith stratum ($i=1,2.....l$) such that the k_i strata to be collapsed must satisfy the relation

$$N_{ij}{}^{p_i}\bar{Y}_{ij}=N_{ij'}{}^{p_i}\bar{Y}_{ij'} \qquad \ldots(2.5)$$

for all $j\neq j'=1,2.....,k_i$. When $p_i=1$, the above relation means that the k_i strata to be collapsed must have same total value of the character Y under study.

The estimate of (2.1) for the above case is given by

$$\frac{1}{N^2}\sum_{i=1}^{l}\sum_{j\neq j'}^{k_i}(N_{ij}{}^{p_i}y_{ij}-N_{ij'}{}^{p_i}y_{ij'})(N_{ij}{}^{2-p_i}y_{ij}-N_{ij'}{}^{2-p_i}y_{ij'}) \qquad \ldots(2.6)$$

The expected value of (2.6) is given by

$$V(\bar{y}_{st})+\sum_{i=1}^{l}\sum_{j\neq j'}^{k_i}(N_{ij}{}^{p_i}\bar{Y}_{ij}-N_{ij'}{}^{p_i}\bar{Y}_{ij'})(N_{ij}{}^{2-p_i}\bar{Y}_{ij}-N_{ij'}{}^{2-p_i}\bar{Y}_{ij'}) \qquad \ldots(2.7)$$

The estimate (2.6) is unbiased when the conditions given in (2.5) are satisfied.

The proposed estimate (2.6) over-estimate the variance as it appears from (2.7) and is likely to be less efficient compared to the sampling designs where strata are formed in such a manner that more than one unit are selected per stratum to form the sample.

3. Use of Ancillary Information

As it is not possible to examine the conditions given in (2.5) for strata to be collapsed to form a new stratum, we may use an ancillary character X which is highly correlated with Y and which is readily available. Let $\bar{X}_i$ be the mean of the ancillary variate for the ith stratum ($i=1,2.....$, L). We collapsed those k_i strata to form a new ith stratum ($i=1,2,.....$, l) such that for the strata to be collapsed the conditions

$$N_{ij}p_i\bar{X}_{ij}=N_{ij'}p_i\bar{X}_{ij},'j\neq j'=1,2,.....k_i \text{ are satisfied.}$$

The above conditions ensure that the conditions in (2.5) are also (almost) satisfied as Y and X are highly correlated and hence the bias in the estimate (2.6) tends to zero.

After collapsing the strata by use of the ancillary variate Y, the same variate X is again used to improve the estimate of $\bar{Y}$ by using ratio method of estimation. Here we will consider both separate and combined methods of ratio estimation.

(*a*) Separate ratio estimate is defined as

$$\bar{y}_{RS}=\sum_{i=1}^{l}\frac{N_{i\cdot}}{N}\times\frac{\bar{y}_{i\cdot}}{\bar{x}_{i\cdot}}\times\bar{X}_{N_{i\cdot}}, \qquad ...(3.1)$$

where $$N_{i\cdot}=\sum_{j=1}^{k_i}N_{ij},\ \bar{y}_{i\cdot}=\sum_{j=1}^{k_i}N_{ij}\,y_{ij}/N_{i\cdot},$$

$$x_{i\cdot}=\sum_{j=1}^{k_i}N_{ij}\,x_{ij}/N_{i\cdot},$$

and $$\bar{X}_{N_{i\cdot}}=\sum_{r=1}^{k_i}\sum_{j=1}^{N_{ir}}X_{ij}/N_{i\cdot}.$$

The variance of (3.1) is given by (upto first degree of approximation)

$$V(\bar{y}_{RS})=\sum_{i=1}^{l}\frac{N_i^2}{N^2}\,[V(\bar{y}_i)+R_i^2V(\bar{x}_i)-2R_i.\ \mathrm{Cov}(\bar{x}_{i\cdot},\bar{y}_{i\cdot})] \qquad ...(3.2)$$

where $R_i=\bar{Y}_{N_i.}/\bar{X}_{N_i.}$.

Now under the assumptions $N_{ij}\bar{Y}_{ij}=N_{ij}'\bar{Y}_{ij}'$ (considering $p_i=1$ for all i) for all $j\neq j'=1, 2,\ldots\ldots, k_i$, we get

$$\left.\begin{aligned}
\hat{V}(\bar{y}_{i\cdot}) &= \frac{1}{N_i^2}\sum_{j\neq j'=1}^{k_i}(y_{ij}N_{ij}-y_{ij}'N_{ij}')^2\\
\hat{V}(\bar{x}_{i\cdot}) &= \frac{1}{N_i^2}\sum_{j\neq j'=1}^{k_i}(x_{ij}N_{ij}-x_{ij}'N_{ij}')^2\\
\text{and } \hat{\mathrm{Cov}}(\bar{x}_{i\cdot},\bar{y}_{i\cdot}) &= \frac{1}{N_i^2}\sum_{j\neq j'=1}^{k_i}(y_{ij}N_{ij}-y_{ij}'N_{ij}')(x_{ij}N_{ij}-x_{ij}'N_{ij}').
\end{aligned}\right\} \qquad ...(3.3)$$

So the estimate of (3.2) in given by

$$\begin{aligned}\hat{V}(\bar{y}_{RS}) = \frac{1}{N^2}\sum_{i=1}^{l}\sum_{j\neq j'=1}^{k_i}&[(y_{ij}N_{ij}-y_{ij}'N_{ij}')^2+r_i^2(x_{ij}N_{ij}-x_{ij}'N_{ij}')^2\\ &-2r_i.(x_{ij}N_{ij}-x_{ij}'N_{ij}')(y_{ij}N_{ij}-y_{ij}'N_{ij}')] \qquad ...(3.4)\end{aligned}$$

where $r_i.=\hat{R}_i.=\bar{y}_{i\cdot}/\bar{x}_{i\cdot}$, which on simplification becomes

$$\frac{1}{N^2}\sum_{i=1}^{l}\sum_{j\neq j'=1}^{k_i}[(y_{ij}N_{ij}-y_{ij}'N_{ij}')-r_i.(x_{ij}N_{ij}-x_{ij}'N_{ij}')]^2 \qquad ...(3.5)$$

(*b*) The combined ratio estimate, with usual notation is defined as

$$\bar{y}_{RC}=\bar{y}_{St}\ \bar{X}_N/\bar{x}_{St}. \qquad ...(3.6)$$

The variance of (3.6) upto first degree of approximation,

when $\sum_{i=1}^{l}k_i$

is sufficiently large is given by

$$V(\bar{y}_{RC})=\sum_{i=1}^{l}(N_i.^2/N^2)[V(\bar{y}_{i\cdot})+R^2V(x_{i\cdot})-2R\ \mathrm{Cov}(\bar{x}_{i\cdot},\bar{y}_{i\cdot})] \qquad ...(3.7)$$

where $R=\bar{Y}_N / \bar{X}_N$. After substituting the values of the estimate from (3.3), the estimate of (3.7) is given by

$$\hat{V}(\bar{y}_{RC})= \frac{1}{N^2}\sum_{i=1}^{l}\sum_{j\neq j'=1}^{k_i}[(N_{ij}y_{ij}-N_{ij}'y_{ij}')-r(N_{ij}x_{ij}-N_{ij}'x_{ij}')]^2 \quad ...(3.8)$$

where $r=\bar{y}_{st}/\bar{x}_{st}$.

The estimates (3.1) and (3.6) will be more efficient than the estimate given in (2.1), if

$$\rho_i>\tfrac{1}{2}R_i\cdot\sqrt{\frac{V(\bar{x}_{i\cdot})}{V(\bar{y}_{i\cdot})}},\ \rho_i>\tfrac{1}{2}R\sqrt{\frac{V(\bar{x}_{\cdot})}{V(\bar{y}_{i\cdot})}} \text{ respectively,}$$

where ρ_i is the population correlation coefficient between y and x in the ith stratum, $i=1, 2,.....l$.

The proposed estimates (3.5) and (3.8) are, of course, subject to the same objection as estimate (2.7), that, they overestimate the variance if the conditions in (2.5) are not satisfied and are likely to be less efficient compared to the sampling designs where strata are so formed that at least two samples are drawn per stratum to form the sample.

4. Double Sampling

So far we have considered the case when $\bar{X}$ is known for each stratum. But this is not the case always. So, at times it pays to devote a part of the resources in estimating the population mean of the ancillary variate for each stratum by taking a large sample. Let a sample of size n'_{ij} from (ij)th stratum is first drawn to estimate $\bar{X}_{N_{i\cdot}}$, where (ij)th $(j=1, 2,....., k_i)$ strata are collapsed to form the new ith stratum.

Let $n'_{i\cdot}=n'_{i1}+n'_{i2}+.....+n'_{ik_i}$ and

$$\bar{x}_{n'_{i\cdot}}=\frac{1}{n'_{i\cdot}}\sum_{j=1}^{k_i}n'_{ij}\bar{x}_{ij}.$$

Then from n'_{ij} units of each stratum one unit is selected for which the Y character is measured.

(*a*) The separate ratio estimate under double sampling is defined as

$$\bar{y}_{\text{DRS}} = \sum_{i=1}^{l} \frac{N_{i.}}{N} \frac{\bar{y}_{i.}}{\bar{x}_{i.}} \bar{x}_{n'_{i.}} \qquad \ldots(4.1)$$

The variance of the above estimate is given by

$$V(\bar{y}_{\text{DRS}}) = \sum_{i=1}^{l} \frac{N_i^2}{N^2} \cdot \left[V(\bar{y}_{i.}) + R_i^2 \left\{ V(\bar{x}_{i.}) - V(\bar{x}_{n'_{i.}}) \right\} \right.$$
$$\left. - 2R_i \{ \text{Cov}\,(\bar{x}_{i.}, \bar{y}_{i.}) - \text{Cov}\,(\bar{x}_{n'_{i.}}, \bar{y}_{n'_{i.}}) \} \right] \qquad \ldots(4.2)$$

The estimate of the above variance, after substituting the estimate of each term is given by

$$\hat{V}(\bar{y}_{\text{DRS}}) = \frac{1}{N^2} \sum_{i=1}^{l} \sum_{j \neq i=1}^{ki} \left[\frac{N_{i.}}{(N_{i.} - k_i)} \frac{(n'_{i.} - k_i)}{n'_{i.}} \left\{ \left(y_{ij} N_{ij} - y_{ij'} N_{ij'} \right) - \right. \right.$$
$$\left. \left. r_{i.} \left(x_{ij} N_{ij} - x_{ij'} N_{ij'} \right) \right\}^2 \right] + \frac{1}{N^2} \sum_{i=1}^{l} \sum_{j \neq j'=1}^{ki} \left(\frac{1}{n'_{i.}} - \frac{1}{N_{i.}} \right)$$
$$\frac{N_{i.}\, k_i}{(N_{i.} - k_i)} \times \left(N_{ij} y_{ij} N - {}_{ij'}\, y_{ij'} \right)^2 \qquad \ldots(4.3)$$

(*b*) The double sampling combined ratio estimate is defined as

$$\bar{y}_{\text{DRC}} = \frac{\bar{y}_{st}}{\bar{x}_{st}} \bar{x}_{n'}, \qquad \ldots(4.4)$$

where $\bar{x}_{n'} = \frac{1}{N} \sum_{i=1}^{l} N_{i.} \bar{x}_{n'_{i.}}$.

The variance of the above estimate is given by

$$V(\bar{y}_{\text{DRC}}) = V(\bar{y}_{st}) + R^2 [V(\bar{x}_{st}) - V(\bar{x}_{n'})] - 2R[\text{Cov}(\bar{x}_{st}, \bar{y}_{st}) - \text{Cov}(\bar{x}_{n'}, \bar{y}_{n'})]$$

and the estimate is given by

$$\hat{V}(\bar{y}_{\text{DRC}}) = \frac{1}{N^2} \sum_{i=1}^{l} \sum_{j \neq j'=1}^{k_i} \frac{N_{i.}}{(N_{i.} - k_i)} \frac{(n'_{i.} - k_i)}{n'_{i.}} \left\{ \left(y_{ij} N_{ij} - y_{ij'} N_{ij'} \right) \right.$$
$$\left. - r \left(x_{ij} N_{ij} - x_{ij'} N_{ij'} \right) \right\}^2 + \frac{1}{N^2} \sum_{i=1}^{l} \sum_{j \neq j'=1}^{ki} \frac{N_{i.}\, k_i}{(N_{i.} - k_i)}$$
$$\left(\frac{1}{n'_{i.}} - \frac{1}{N_i} \right) \left(y_{ij} N_{ij} - y_{ij'} N_{ij'} \right)^2 \qquad \ldots(4.5)$$

5. ILLUSTRATION

The data on the block level estimate for wheat crop in Patna district of Bihar for the year 1967-68 have been taken for illustrating the use of ancillary information in collapsing strata For this purpose two blocks consisting of 15 V.L.W. circles as strata have been taken. The character under study is the crop cutting yield per plot of size 1/80 of an acre in kg and the ancillary information is the eye estimate per acre in kg. For our purpose we have selected one village per V.L.W. circle. Our sample consists of one field for crop cutting experiment and four fields for eye estimates in each selected village. The collapsing of strata have been done on the basis of the auxiliary variate such that total of auxiliary variate remains more or less same for the strata to be collapsed. Following the double sampling procedure given in section 4, the results are as follows :

Method of estimation	*Mean in kg./plot*	*Variance*	*% gain in efficiency over stratified random sampling*
Stratified random sampling	5.797	0.8770	
Double sampling separate ratio	5.872	0.2870	23.87
Double sampling combined ratio	5.846	0.2885	23.47

with parameters $n'_{ij}=4$ for all i and j, $k_1=k_2=k_3=3$, $k_4=2$, $k_5=4$, $l=5$ and $r_{i.}=(0.0135, 0.0133, 0.0174, 0.0128, 0.0136)$, $r=0.0141$.

SUMMARY

Here an attempt has been made to utilize ancillary information in collapsing any number of strata in case of deep stratification and the same ancillary information is used in estimating the parameter under study more efficiently.

ACKNOWLEDGEMENT

We are thankful to the referee for his valuable comments to improve the paper.

REFERENCES

1. Cochran, W.G. (1963) Sampling Technique, John Wiley and Sons, Inc. New York.
2. Seth, G.R. (1966) "On collapsing of strata," Journal of Indian Society of Agricultural Statistics, Vol. XVIII, 1966.