Jour. Ind. Soc. Ag. Statistics
50(2), 1997 : 144–149

A Class of Estimators in Stratified Sampling with Two Auxiliary Variables

M. Dalabehara and L.N. Sahoo[1]
Orissa University of Agriculture and Technology, Bhubaneswar - 751003
(Received : January, 1996)

SUMMARY

Following Srivastava [2], a general class of estimators for the finite population total utilizing the available knowledge on two auxiliary variables under a stratified sampling have been given.

Key words: Asymptotic variance, Auxiliary variable, Combined estimate, Separate estimate, Stratified sampling.

1. *Introduction*

Consider a finite population divided into L disjoint strata $S_1, S_2, \ldots, S_L$. Let Y_h and X_h be the totals of S_h in respect of the study variable y and an auxiliary variable x and the overall total $Y = \sum_h Y_h$ of y-values is to be estimated. Sampling within each stratum is done independently according to any probability sampling design. Given the sample s_h in S_h, let t_{hy} and t_{hx} be unbiased estimates of Y_h and X_h respectively, such that $V(t_{hy}) = \sigma^2_{hy}$, $V(t_{hx}) = \sigma^2_{hx}$ and Cov $(t_{hy}, t_{hx}) = \sigma_{hyx}$.

As is well-known, in a stratified sampling the auxiliary variable x can be incorporated in two different ways. If the overall total $X = \sum_h X_h$ of x-values is known, a combined estimate (e.g. ratio, product or regression estimate) for Y is made. In addition to the value of X, if the values of X_h $(h = 1, 2, \ldots, L)$ are known, a separate estimator is built up from the stratum level estimators. But, in practice, the latter procedure yields better estimators than the former. In the present context, we use a second auxiliary variable to improve efficiency in the estimation of Y.

---

1 Utkal University, Bhubaneswar - 751004

2. *Use of Second Auxiliary Variable*

Sometimes even if X_h's are known, information on a cheaply ascertainable variable z whose correlation with y may be less than that of x with y (i.e. $\rho_{yz} < \rho_{yx}$) is readily available. This type of situation may also be realizable with z is used as the stratification variable. For instance, in a crop survey if y, x, and z are respectively the yield of jute, area under jute and area under cultivation, information on the total cultivated area of each village can be obtained at a low cost. If the villages in a district are stratified on the basis of their total cultivated area, then information on z can be easily known from the district records.

Let t_{hz} be an unbiased estimate of Z_h (total of z-values in S_h) so that $t_z = \sum_h t_{hz}$ will be an unbiased estimate of Z (the overall total of z-values) with $V(t_z) = \sum_h \sigma^2_{hz}$ where $\sigma^2_{hz} = V(t_{hz})$.

In this paper, following Srivastava [2], we develop a general class of estimators for Y using the covariates x and z simultaneously. We assume that, all the stratum level totals of x (i.e. X_h's) are known but for the second auxiliary variable z only the overall total (i.e. Z) is known. It is also observed that many well-known, and some less-known but potentially interesting estimators belong to this class.

3. *The Proposed Class of Estimators*

For given $s_h \subset S_h$, (h = 1, 2, ..., L) let (t_{hy}, t_{hx}) assume values in a closed convex subspace, R_2, of the two-dimensional real space containing the point (Y_h, X_h). Then following Srivastava [2] a class of estimators for Y_h is defined by

$$\hat{Y}_h = g_h(t_{hy}, t_{hx})$$

where $g_h(t_{hy}, t_{hx})$ is a known function of t_{hy} and t_{hx}, independent of Y_h, such that $g_h(Y_h, X_h) = Y_h$ and satisfying the following regularity conditions :

(i) The function $g_h(t_{hy}, t_{hx})$ is continuous in R_2

(ii) The first and second order partial derivatives of $g_h(t_{hy}, t_{hx})$ exist and are also continuous in R_2.

These conditions are assumed by Srivastava [2] for justifying Y_h to be a class of estimators for Y_h. On expanding $g_h(t_{hy}, t_{hx})$ about (Y_h, X_h) in a Taylor's series, we have to a first order of approximation

$$\hat{Y}_h \approx g_h(Y_h, X_h) + g_{h1}(t_{hy} - Y_h) + g_{h2}(t_{hx} - X_h) \tag{3.1}$$

where g_{h1} and g_{h2} denote the first order partial derivatives of $g_h(t_{hy}, t_{hx})$ w.r.t. t_{hy} and t_{hx} respectively at (Y_h, X_h). Noting that $g_h(Y_h, X_h) = Y_h$ and $g_{h1} = 1$, (3.1) can be rewritten as

$$\hat{Y}_h - Y_h = (t_{hy} - Y_h) + g_{h2}(t_{hx} - X) \tag{3.2}$$

Thus, to a first order of approximation $E(\hat{Y}_h) = Y_h$ with asymptotic variance

$$V(\hat{Y}_h) = \sigma_{hy}^2 + 2g_{h2}\,\sigma_{hyx} + g_{h2}^2\,\sigma_{hx}^2 \tag{3.3}$$

Based on the above results, a class of separate estimators for Y may be defined by $t_s = \sum_h \hat{Y}_h = \Sigma_h\, g_h(t_{hy}, t_{hx})$ with asymptotic variance

$$V(t_s) = \sum_h (\sigma_{hy}^2 + 2g_{h2}\,\sigma_{hyx} + g_{h2}^2\,\sigma_{hx}^2) \tag{3.4}$$

Whatever be the samples s_h (h = 1, 2, ..., L) and consequently an overall sample $s\ (= \bigcup_{h=1}^{L} s_h)$ chosen, let (t_s, t_z) assume values in closed convex subspace, R_2' (say), of two-dimensional real space containing the point (Y, Z). Let $f(t_s, t_z)$ be a known function of t_s and t_z which may contain Z but independent of Y such that f(Y, Z) = Y, and also admitting the regularity conditions in R_2'.

It may be noted here that, R_2 and R_2' can be regarded as the yx and yz planes respectively of a three dimensional real subspace R_3 containing the points (Y_h, X_h, Z) and (Y, X, Z). Thus the points (Y_h, X_h) and (Y, Z) can also be represented by $(Y_h, X_h, 0)$ and (Y, 0, Z).

The proposed class of estimators of Y may be defined by

$$t_G = f(t_s, t_z) \tag{3.5}$$

Since there are only a finite number of possible samples, the expectation and variance of t_G exist under the condition (i). Expanding $f(t_s, t_z)$ about (Y, Z) in a second order Taylor's series and taking expectation, we obtain the asymptotic variance of t_G

$$V(t_G) = V(t_s) + 2f_2 \ Cov(t_s, t_z) + f_2^2 \ V(t_z) \tag{3.6}$$

where f_2 denotes the first order partial derivative of $f(t_s, t_z)$ w.r.t. t_z at (Y, Z).

Writing $t_{hz} = Z_h + (t_{hz} - Z_h)$ and using (3.2)

$$Cov(\hat{Y}_h, t_{hz}) \sim \sigma_{hyz} + g_{h2}\,\sigma_{hxz}$$

so that $Cov(t_s, t_z) = \sum_h Cov(\hat{Y}_h, t_{hz}) \sim \sum_h(\sigma_{hyz} + g_{h2}\,\sigma_{hxz})$

where $\sigma_{hyz} = Cov(t_{hy}, t_{hz})$ and $\sigma_{hxz} = Cov(t_{hx}, t_{hz})$

Finally, the formula for the asymptotic variance of t_G is obtained as

$$V(t_G) = \sum_h(\sigma_{hy}^2 + 2g_{h2}\,\sigma_{hyx} + g_{h2}^2\,\sigma_{hx}^2) + f_2^2 \sum_h \sigma_{hz}^2 + 2f_2 \sum_h (\sigma_{hyz} + g_{h2}\sigma_{hxz}) \tag{3.7}$$

4. *Some Observations and Remarks*

4.1 If the second auxiliary variable z is not used, t_G reduces to t_S, i.e. a class generating a family of separate variety estimators. From (3.4) and (3.7) it follows that $V(t_G) \leq V(t_S)$ if

$$f_2^2 \sum_h \sigma_{hz}^2 + 2f_2 \sum_h (\sigma_{hyz} + g_{h2}\,\sigma_{hxz}) \leq 0$$

Thus, an estimator of t_G is more efficient than an estimator of t_S if

$$\beta_{yz} + \frac{\sum_h g_{h2}\,\sigma_{hxz}}{\sum_h \sigma_{hz}^2} \leq -\frac{f_2}{2} \tag{4.1}$$

where $\beta_{yz} = \sum_h \sigma_{hyz} \Big/ \sum_h \sigma_{hz}^2$

This condition shows that there is a scope for improving upon the estimators based on one auxiliary variable x by using the second auxiliary variable z in stratified sampling.

4.2 If x is not used i.e. x-values are treated to be a non-zero constant, t_G reduces to a class of combined variety estimators represented by

$$t_c = f(t_y, t_z) \text{ with } t_y = \sum_h t_{hy}$$

It may be noted here that, a different function satisfying the earlier regularity conditions can also be used to define t_c.

The asymptotic variance of t_c is

$$V(t_c) = \sum_h (\sigma_{hy}^2 + 2f_2\sigma_{hyz} + f_2^2\,\sigma_{hz}^2) \tag{4.2}$$

Thus, an estimator of t_G would be more efficient than that of t_c if

$$\beta_{hyx} + f_2\beta_{hzx} \le -\frac{g_{h2}}{2} \qquad (h = 1, 2, ..., L) \tag{4.3}$$

where $\beta_{hyx} = \sigma_{hyx}/\sigma_{hx}^2$, $\beta_{hzx} = \sigma_{hxz}/\sigma_{hx}^2$

4.3 The variance of t_G given in (3.7) is sought to be minimized subject to

$$g_{h2} = -(\beta_{hyx} + f_2\beta_{hzx}) = \hat{g}_{h2} \text{ (say)}$$

and

$$f_2 = -\frac{\sum_h \sigma_{hz}^2 (\beta_{hyz} - \beta_{hyx}\,\beta_{hxz})}{\sum_h \sigma_{hz}^2 (1 - \rho_{hxz}^2)} = \hat{f}_2 \text{ (say)} \tag{4.4}$$

where $\beta_{hyz} = \sigma_{hyz}/\sigma_{hz}^2$ and $\rho_{hxz} = \sigma_{hxz} / \sigma_{hx}\,\sigma_{hz}$

Thus, from (4.4), it is clear that, optimum values of g_{h2} and f_2 can not be determined uniquely. However, after obtaining an optimum value of f_2, we can use this value to calculate the optimum values of g_{h2}. Utilizing these optimum values, the minimum asymptotic variance of t_G is given by

$$V_{min}(t_G) = \sum_h \sigma_{hy}^2 (1 - \rho_{hyx}^2 - B^2) \tag{4.5}$$

where $$B = \frac{\sum_h \sigma_{hy}\,\sigma_{hz}\,(\rho_{hyz} - \rho_{hyx}\,\rho_{hyz})}{\sqrt{\sum_h \sigma_{hy}^2}\,\sqrt{\sum_h \sigma_{hz}^2\,(1-\rho_{hxz}^2)}}$$ such that

$$\rho_{hyx} = \sigma_{hyx}/\sigma_{hy}\,\sigma_{hx},\quad \rho_{hyz} = \sigma_{hyz}/\sigma_{hy}\,\sigma_{hz}$$

The estimator attaining this minimum variance is a regression-type estimator of the form

$$t_{RG} = \sum_h [t_{hy} - \hat{g}_{h2}\,(t_{hx} - X_h)] - \hat{f}_2\,(t_z - Z)$$

studied earlier by Dalabehera and Sahoo [1]. This leads to an interesting result that, one cannot improve upon t_{RG} by using x and z for the situation under consideration.

4.4 The suggested class provides us with an infinite number of estimators depending on proper choices of the functions g_h and f, and their asymptotic variances or mean square errors can be obtained from (3.7) by the substitution of corresponding values of g_{h2} and f_2. For example, simple expansion estimator (without using x or z); separate variety ratio, product, difference and regression estimators using x ; combined variety ratio, product, difference and regression estimators using z are particular members of the class. It is also interesting to note that the classes of estimators represented by t_s and t_c may be identified as subclasses of the class of estimators represented by t_G.

REFERENCES

[1] Dalebehera, M. and Sahoo, L.N., 1994. A new estimator with two auxiliary variables for stratified sampling (*Communicated for publication*).

[2] Srivastava, S.K., 1980. A class of estimators using auxiliary information in sample surveys. *Canadian Jour. Statist.*, 8, 253-254.