Jour. Ind. Soc. Ag. Statistics
Vol. XXXIX, No. 2 (1987), pp. 191-201

ON DOUBLE SAMPLING FOR STRATIFICATION AND USE OF AUXILIARY INFORMATION

ABEL F. IGE and T. P. TRIPATHI*
University of Ilorin, Nigeria

(Received : March, 1986)

SUMMARY

In the usual procedure of double sampling for Stratification (DSS), the auxiliary information collected on the first phase sample is used only at the designing stage for stratifying the sampled units and for estimating the strata weights W_h. The usual unbiased estimator for the population mean $\bar{Y}$ based on DSS does not utilize the entire information collected on the first phase sample and the stratified subsamples. Similarly the unstratified double sampling (USDS) procedure utilizes the auxiliary information collected on the first phase sample, only at the estimation stage for defining the usual ratio, difference and regression estimators in USDS. This motivates us to propose alternative sampling strategies, based on DSS, utilizing rhe auxiliary information obtained on the first phase sample both at the designing as well as at estimation stages. The general properties of the proposed strategies are studied and conditions for optimality obtained. The situations in which the proposed estimators are better than the usual unbiased estimator in DSS are identified and some of the proposed estimators are compared with corresponding estimators based on USDS with and without cost considerations.

Keywords : Double sampling; stratification; sampling strategy; combined difference and ratio estimators.

Introduction

Suppose we want to estimate the population mean $\bar{Y}$ of a variate y and consider it desirable to stratify the population, consisting of N units, on the basis of the values of an auxiliary character x but the frequency

*On leave from Indian Statistical Institute, Calcutta, India.

distribution of x is unknown. The sampling frame for various strata and the strata weights $W_h = N_h/N$, $h = 1, \ldots, L$, are not known although the strata may be fixed in advance. In such a situation we use the technique of double sampling for stratification (DSS) which consists of the following steps (Rao [3]).

(i) We select a preliminary large sample $S_{(1)}$ of size n' rather inexpensively using simple random sampling without replacement (SRSWOR) and observe the auxiliary character x alone.

(ii) The sample $S_{(1)}$ is stratified into L strata on the basis of the observed x. Let n_h' denote the number of units in $S_{(1)}$ falling into stratum $h (h = 1, \ldots, L, \Sigma n_h' = n')$ and $n' = \{n_1', \ldots n_L'\}$ denote the resulting configuration of $S_{(1)}$.

(iii) Subsamples of sizes $n_h = \nu_h n_h'$, $0 < \nu_h < 1$ $(h = 1, \ldots, L)$ ν_h being predetermined for each h, are selected from strata, independently from each other, using SRSWOR and the character of main interest y is observed.

Let $n = \Sigma n_h$; $n = \{n_1, \ldots, n_L\}$ and $\bar{y}_h = \sum_{i=1}^{n_h} y_{hi}/n_h$ Note that $w_h = \dfrac{n_h'}{n'}$ is an unbiased estimate of strata weights $W_h = N_h/N$. Throughout we assume that n' is large enough so that $Pr(n_h' = 0) = 0$ for all h.

The usual unbiased estimator for $\bar{Y}$ in DSS, defined by

$$\bar{y}_{ds} = \Sigma\, w_h\, \bar{y}_h. \tag{1.1}$$

with

$$V(\bar{y}_{ds}) = \left(\frac{1-f}{n'}\right) S_y^2 + \frac{1}{n'} \sum_h \left(\frac{1}{\nu_h} - 1\right) W_h S_{hy}^2 \tag{1.2}$$

is well known [Rao (1973); Cochran (1977)], where

$$f = \frac{n'}{N};\ S_y^2 = \frac{\sum_{i=1}^{N} (y_i - \bar{Y})^2}{N-1}$$

$$S_{hy}^2 = \sum_{i=1}^{n_h} (y_{hi} - \bar{Y}_h)^2/(N_h - 1) \quad \text{and} \quad \bar{Y}_h = \sum_{i=1}^{n_h} y_{hi}/N_h.$$

It is to be noted that in the discussion on DSS by various survey statisticians, the auxiliary information collected on $S_{(1)}$ is used only for stratification. However in defining the estimator $\bar{y}_{ds}$, the entire information already at hand through $S_{(1)}$ and the Stratified subsamples is not fully utilized. We are of the opinion that Survey statisticians should

endeavour to devise methods of utilizing all the available information (auxiliary or otherwise) at their disposal to obtain better sampling strategies. This motivates us to propose alternative sampling strategies which utilize the auxiliary information obtained on $S_{(1)}$ not only for stratification but at the estimation stage as well.

2. Combined Estimators in Double Sampling for Stratification

We note that

$$\bar{x}' = \sum_h w_h \bar{x}'_h \quad \text{and} \quad \bar{x}_{st} = \sum_h w_h \bar{x}_h$$

where

$$\bar{x}'_h = \sum_{i=1}^{n'_h} x_{hi}/n'_h \quad \text{and} \quad \bar{x}_h = \sum_{i=1}^{n_h} x_{hi}/n_h$$

are unbiased estimators of the population mean $\bar{X} = \sum_h W_h \bar{X}_h$ of x.

Utilizing the information collected on x-variate we define the combined difference and ratio estimators in DSS by

$$e_{DC} = \bar{y}_{ds} - \lambda (\bar{x}_{ds} - \bar{x}') \tag{2.1}$$

and

$$e_{RC} = \bar{y}_{ds} (\bar{x}'/\bar{x}_{ds}) \tag{2.2}$$

respectively, where λ is a suitably chosen constant. We also consider the weighted estimator

$$e_1^* = (1 - w)\, \bar{y}_{ds} + w e_{RC} \tag{2.3}$$

where w is a suitably chosen weight.

We note that e_{DC} is unbiased for $\bar{Y}$ and exact expression for its variance is given by

$$V(e_{DC}) = \frac{(1-f)}{n'} S_y^2 + \frac{1}{n'} \sum_h \left(\frac{1}{\nu_h} - 1 \right) W_h (S_{hy}^2 - 2\lambda S_{hyx} + \lambda^2 S_{hx}^2) \tag{2.4}$$

Furthermore, for large samples $S_{(1)}$, the approximate expressions for the biases and mean square errors (MSEs) of the estimators e_{RC} and e_1^*

are given by

$$B(e_{RC}) = \frac{1}{\bar{x}_n'} \sum_h \left(\frac{1}{\nu_h} - 1 \right) W_h (R S_{hx}^2 - S_{hyx}); \; B(e_1^*) = wB(e_{RC})$$

$$M(e_{RC}) = [V(e_{DC})] \text{ with } \lambda = R; \; M(e_1^*) = [V(e_{DC})] \text{ with } \lambda = wR \tag{2.5}$$

where $R = \bar{Y}/\bar{X}$.

Let $\beta = \sum_h a_h \beta_h / \sum_h a_h$ with $a_h = \left(\frac{1}{\nu_h} - 1 \right) W_h S_{hx}^2$

be the weighted average of the strata population regression coefficient

$$\beta_h = S_{hyx}/S_{hx}^2 \text{ of } y \text{ on } x$$

and (2.6)

$$\rho = \frac{\sum_h \left(\frac{1}{\nu_h} - 1 \right) W_h \rho_{hyx} S_{hy} S_{hx}}{\left[\sum_h \left(\frac{1}{\nu_h} - 1 \right) W_h S_{hy}^2 \sum_h \left(\frac{1}{\nu_h} - 1 \right) W_h S_{hx}^2 \right]^{\frac{1}{2}}}$$

where

$\rho_{hyx} = S_{hyx}/S_{hy} S_{hx}$ is the correlation coefficient between y and x in stratum h.

The optimum λ and w and the resulting MSEs are given by

$$\lambda_0 = \beta, \quad w_0 = \beta/R$$

$$V_0(e_{DC}) = \frac{(1-f)}{n'} S_y^2 + \frac{1}{n'} (1 - \rho^2) \sum_h \left(\frac{1}{\nu_h} - 1 \right) W_h S_{hy}^2 = M_0(e_1^*) \tag{2.7}$$

The estimated values of λ_0 and w_0 may be given by

$$\hat{\lambda}_0 = \hat{\beta} = \sum_h \left(\frac{1}{\nu_h} - 1 \right) W_h s_{hyx} \Big/ \sum \left(\frac{1}{\nu_h} - 1 \right) w_h s_{hx}^2$$

$$\hat{w}_0 = \hat{\lambda}_0 \bar{x}_{ds}/\bar{y}_{ds} \tag{2.8}$$

where

$$s_{hyx} = \sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)(x_{hi} - \bar{x}_h)/(n_h - 1)$$

and $s_{hx}^2 = s_{hxx}$ are unbiased for S_{hyx} and S_{hx}^2 respectively. The estimator e_{DC} with $\lambda = \hat{\lambda}_0$, viz $e_{Drg} = \bar{y}_{ds} - \hat{\beta}\,(\bar{x}_{ds} - \bar{x}')$ may be called the regression estimator for $\bar{Y}$ in DSS. For large samples, $M(e_{Drg})$ and $M(\hat{e}_1^*)$ would again be given by (2.7) where $\hat{e}_1^*$ is e_1^* with $w = \hat{w}_0$, It is noted that

$$\sum_h w_h^2 \left(\frac{1}{n_h} - \frac{1}{n_h'}\right) s_{hyx} = \frac{1}{n'} \sum_h \left(\frac{1}{\nu_h} - 1\right) w_h\, S_{hyx}$$

is an unbiased estimator of $\frac{1}{n'} \sum_h \left(\frac{1}{\nu_h} - 1\right) W_h\, S_{hyx}$. Rao (1973) showed that

$$s_{yds} = \frac{N}{N-1}\left[\sum_h \frac{w_h}{n_h}(n_h - 1)\, s_{hy}^2 + \sum_h w_h (\bar{y}_h - \bar{y}_{ds})^2 + \hat{V}(\bar{y}_{ds})\right]$$

is a non-negative unbiased estimator of S_y^2 where

$$\hat{V}(\bar{y}_{ds}) = \frac{(N-1)}{N} \sum_h \left(\frac{n_h' - 1}{n' - 1} - \frac{n_h - 1}{N - 1}\right) \frac{w_h\, s_{hy}^2}{n_h}$$

$$+ \frac{(1-f)}{n' - 1} \sum_h w_h (\bar{y}_h - \bar{y}_{ds})^2 \qquad (2.9)$$

with

$$s_{hy}^2 = \sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)^2/(n_h - 1)$$

is a non-negative unbiased estimator of $V(\bar{y}_{ds})$.

Using the above results, a non-negative unbiased estimator of $V(e_{DC})$ would be given by

$$\hat{V}(e_{DC}) = \frac{(1-f)}{n'} s_{yds}^2 + \sum_h w_h \frac{(n_h' - n_h)}{n_h\, n_h'} (s_{hy}^2 - 2\lambda s_{hyx} + \lambda^2 s_{hx}^2)$$

$$= \frac{(1-f)}{n'} s_{yds} + \frac{1}{n'} \sum_h \left(\frac{1}{\nu_h} - 1\right) w_h (s_{hy}^2 - 2\lambda s_{hyx} + \lambda^2 s_{hx}^2).$$

Non-negative but biased estimators for the MSEs of e_{RC} add e_1^* may be given by the above expression in (2.10) again with $\lambda = \hat{R} = \bar{y}_{ds}/\bar{x}_{ds}$ and

$\lambda = w\hat{R}$ respectively. Furthermore, a non-negative but biased estimator of the minimum variance in (2.7) is given by (2.10) with λ replaced by $\hat{\lambda}_0$ in (2.8).

3. Relative Performance of Estimators

From (1.2) and (2.7) we observed that the estimators e_{DC} and e_1^* with optimum choices of λ and w respectively are uniformly better than the usual unbiased estimator $\bar{y}_{ds}$. Noting that $V_0(e_{DC})$ is a monotonically decreasing function of $|\rho|$ and that $|\rho|$ is a monotonically increasing function of $|\rho_{hyx}|$ it may be stated that if all the strata correlation coefficients $\rho_{hyx} (h = 1, \ldots, L)$ have high values or have the same sign, $V_0(e_{DC})$ would decrease considerably resulting into appreciable gain in efficiency of optimum estimators over $\bar{y}_{ds}$. However if values of ρ_{hyx} are low or positive in some strata while negative in others, the values of $|\rho|$ may not be high and the gains may not be appreciable.

In general, e_{DC} would be better than $\bar{y}_{ds}$ if λ lies between 0 and 2β and e_1^* would be better than $\bar{y}_{ds}$ is w lies between 0 and $2\beta \mid R$. Furthermore, for $R > 0$, e_{RC} would be better than $\bar{y}_{ds}$ if $\beta \mid R > \frac{1}{2}$. It is interesting to note that the above results are quite similar to those in ordinary stratified random sampling.

If $\rho_{hyx} < 0$ for all $h = 1, \ldots, L$ and $R > 0$, e_{RC} should not be used as it would be worse than $\bar{y}_{ds}$. Even for the populations where β is expected to be quite small relative to R, e_{RC} should not be used. However there is no such restriction in the use of e^* which is more general than e_{RC}. Thus in practice use of e_1^* may be preferred over e_{RC}. However if we insist on unbiasedness e_{DC} is to be preferred over e_1^*.

Using the same amount of information needed for defining e_{DC} one may define a separate difference estimator in DSS by

$$e_{Ds} = \sum_h W_h \{\bar{y}_h - \lambda_h (\bar{x}_h - x'_h)\} \tag{3.1}$$

which is unbiased for $\bar{Y}$ with exact variance expression

$$V(e_{Ds}) = \frac{(1-f)}{n'} S_y^2 + \frac{1}{n'} \sum_h W_h \left(\frac{1}{\nu_h} - 1\right) (S_{hy}^2 - 2\lambda_h S_{hyx} + \lambda^2 S_{hx}^2) \tag{3.2}$$

The optimum value of λ_h would be β_h and

$$V_0(e_{DS}) = \frac{(1-f)}{n'} S_y^2 + \frac{1}{n'} \sum_h W_h \left(\frac{1}{\nu_h} - 1\right) (1 - \rho_{hyx}^2) S_{hy}^2.$$

Noting that

$$V_0(e_{DC}) - V_0(e_{DS}) = \frac{1}{n'}\sum_h \left(\frac{1}{\nu_h} - 1\right) W_h S_{hx}^2 (\beta - \beta^h)^2$$

the separate estimator e_{DS} would be preferrable over the combined estimator e_{DC} provided good guessed values β_h^* of β_h are available for each h so that e_{DS} may be made better than $\bar{y}_{as}$ and e_{DC} by taking $\lambda_h = \beta_h^*$.

4. Comparison with Estimators Based on Unstratified Double Sampling

Let the first sample $S_{(1)}$ of size n' be drawn as before and the second sample $S_{(2)}$ of size n be a sub-sample of $S_{(1)}$ selected according to SRSWOR instead of being selected in the form of stratified subsamples. We shall call this procedure unstratified double sampling (USDS). Based on the first phase sample we define $\bar{x}' = \sum_{i=1}^{n'} x_i/n'$ and based on the second phase sample we define $\bar{y} = \sum_{i=1}^{n} y_i/n$, $\bar{x} = \sum_{i=1}^{n} x_i/n$. The difference and ratio estimators in USDS are defined by

$$\bar{y}'_D = \bar{y} - \lambda^* (\bar{x} - \bar{x}')$$

and (4.1)

$$\bar{y}'_R = \frac{\bar{y}}{\bar{x}} \bar{x}'.$$

(Raj [2], Cochran [1]). The estimator $\bar{y}'_D$ is unbiased with exact expression for its variance as

$$V(\bar{y}'_D) = \frac{(1-f)}{n'} S_y^2 + \left(\frac{1}{n} - \frac{1}{n'}\right)(S_y^2 - 2\lambda^* S_{yx} + \lambda^2 S_x^2) \qquad (4.2)$$

For large n, the approximate expression for the MSE of $\bar{y}'_R$ is

$$M(\bar{y}'_R) = [V(\bar{y}'_D)]_{\lambda^* = R} \qquad (4.3)$$

We note that in case of USDS we utilize information on $S_{(1)}$ only for defining the estimators while in case of DSS we have used it both for stratification as well as for defining the estimators.

For comparisons we assume in case of DSS estimators, that allocation of second sample to different strata is proportional to the random strata

size n_h', that is

$$n_h \propto n_h'$$

so that for each $h = 1, \ldots, L$ (4.4)

$$\nu_h = n/n'$$

We note that

$$S_{yx} = \sum_h \frac{(W_h - N^{-1})}{(1 - N^{-1})} S_{hyx} + \sum_h \frac{W_h}{(1 - N^{-1})} (\bar{Y}_h - \bar{Y})(\bar{x}_h - \bar{x})$$

$$= \sum_h W_h S_{hyx} + \sum_h W_h (\bar{Y}_h - \bar{Y})(\bar{x}_h - \bar{x}) \quad (4.5)$$

neglecting the terms $1/N$. From (2.4), (2.5) and (4.2) to (4.5)

$$V(\bar{y}_D') - V(e_{DC}) = \left(\frac{1}{n} - \frac{1}{n'}\right) \sum_h W_h [\bar{Y}_h - \bar{Y}) - \lambda(\bar{x}_h - \bar{x})]^2$$

$$M(\bar{y}_R') - M(e_{RC}) = \left(\frac{1}{n} - \frac{1}{n'}\right) \sum_h W_h [\bar{Y}_h - R\bar{x}_h]^2$$

where the same λ is used in both of $\bar{y}_D'$ and e_{DC}. Thus the sampling strategies (DSS, e_{DC}) and (DSS, e_{RC}) are better than the strategies (USDS, $\bar{y}_D'$) and (USDS), $\bar{y}_R'$) respectively. It may be remarked that the above results are similar to those obtained by comparing difference and ratio estimators in SRSWOR with the combined difference and ratio estimators in ordinary stratified sampling with proportional allocation.

Let $V\min(\bar{y}_D')$ denote the variance of $\bar{y}_D'$ at the optimum choice $B = \rho_{yx} S_y/S_x$ of λ^* where ρ_{yx} is the population correlation coefficient between y and x. From (3.2) and (4.2) to (4.5).

$$V_{\min}(\bar{y}_D') - V(e_{DS}) = \left(\frac{1}{n} - \frac{1}{n'}\right) \left[A + \sum_h W_h S_{hx}^2 \{(B - \beta_h)^2 - (\lambda_h - \beta_h)^2\}\right] \quad (4.6)$$

where

$$A = \Sigma W_h \{\bar{Y}_h - \bar{Y}) - B(\bar{x}_h - \bar{x})\}.$$

We note that if λ_h does not depart substantially from β_h, the strategy (DSS, e_{DS}) would be better than the strategy (USDS, $\bar{y}_D'$).

Following Cochran ([1] p. 341 and 331), using (4.3) and assuming that cost per unit is the same for all strata, the cost function

$$C = c'n' + cn, \; c' < c \quad (4.7)$$

may be used for both the strategies (USDS, $\bar{y}_D$) and (DSS, e_{DC}).

Let $V_{\text{opt}}(\bar{y}'_D)$ denote $V(\bar{y}'_D)$ with those values of n' and n which minimize $V(\bar{y}'_D)$ for given cost C. For simplicity, let us assume that $\lambda = \lambda^* = B$ which is the best choice of λ^*. Now using (4.5), we get, after algebraic simplifications,

$$C[V_{\text{opt}}(\bar{y}'_D) - V_{\text{opt}}(e_{DC})] = A(c - c') + 2(cc')^{1/2}[\{V_1V_2 + A(2\rho_{yx}^2 - 1)\, S_y^2 + A^2\}^{1/2} - \{V_1V_2\}^{1/2}] \quad (4.8)$$

where

$$V_1 = \rho_{yx}^2\, S_y^2 + A$$

and

$$V_2 = \sum_h W_h S_{hy}^2 - 2B\Sigma W_h S_{hyx} + B^2 \Sigma W_h S_{hx}^2.$$

We note that even under the above situations which are most suitable to $\bar{y}'_D$ the estimator e_{DC} may be better than $\bar{y}'_D$. A sufficient condition for the sampling strategy (DSS, e_{DC}) to be better than the strategy (USDS, $\bar{y}'_D$) is obviously given by

$$A + (2\rho_{yx}^2 - 1)\, S_y^2 \geqslant 0 \quad (4.9)$$

which always hold if

$$|\rho_{yx}| > \frac{1}{\sqrt{2}}.$$

5. A Numerical Investigation of the Estimators

To investigate the relative efficiency of the proposed estimators over $\bar{y}_{as}$, we consider the population data in Cochran [1] [p. 168, Table 6.3] about Jefferson county Iowa. We note as in Cochran [1] [p. 332, Example] that the Jefferson data did not generate from a double sampling procedure, but could be used to illustrate it.

Table 1 gives the percent relative efficiency of the proposed estimators over $\bar{y}_{as}$ for some sample sizes (n', n).

Using the Jefferson data, we observed that the percent relative efficiency of the separate difference estimator e_{DS} over the combined difference estimator e_{DC} at optimum values of λ and λ_h is approximately 103 for the sample sizes in Table 1. Furthermore, the percent relative efficiency of the combined ratio estimator e_{RC} over the unstratified ratio estimator $\bar{y}'_R$ is approximately 108. We also note that the percent relative

TABLE 1—PERCENT RELATIVE EFFICIENCY OF THE PROPOSED ESTIMATORS OVER $\bar{y}_{ds}$

Estimators \ Sample Sizes (n', n)	(1000, 100)	(1000, 150)	(1000, 200)	(700, 70)	(700, 100)	(700, 140)	(500, 50)	(500, 70)	(500, 100)
e_{RC}	125.23	123.95	122.66	124.20	123.24	121.49	124.15	122.74	120.77
0.1	120.90	119.88	118.84	120.02	119.3	117.89	120.03	118.89	117.31
$e_{DC}\lambda = 0.1793^{**}$	127.38	125.97	124.54	126.28	125.18	123.25	126.18	124.63	122.47
0.3	113.33	112.72	112.09	112.68	112.37	111.51	112.81	112.11	111.16
0.6	125.24	123.96	122.66	124.21	123.24	121.49	124.15	122.74	120.78
$e_1^*w = 0.7997^{**}$	127.38	125.97	124.54	126.28	125.18	123.25	126.18	124.63	122.47
0.9	126.84	125.46	124.06	125.75	124.69	122.81	125.67	124.15	122.04
0.1, 0.05	119.25	118.31	117.37	118.41	109.20	116.50	118.46	117.42	115.98
$e_{DC}(\lambda_1, \lambda_2) = (0.2404, 0.1160^{**})$	131.67	129.99	128.30	130.4	129.08	126.77	130.24	128.40	125.84
0.4, 0.2	119.22	118.29	117.34	118.38	109.19	116.05	118.43	117.39	115.94

**denote the optimum values of λ, w, λ_1 and λ_2 respectively.

efficiency of the combined difference estimator e_{DC} over the unstratified difference estimator $\bar{y}'_D$ for the same values of $\lambda = \lambda^*$ 0.1 or 0.3 lie between 110 and 112 for the sample sizes in Table 1.

REFERENCES

[1] Cochran, W. G. (1977) : *Sampling Techniques*, 3rd Edition, New York : Wiley.

[2] Raj, Des (1968) : *Sampling Theory*, New York : McGraw-Hill.

[3] Rao, J. N. K. (1973) : On double sampling for stratification and analytical surveys, *Biometrika*, **60** : 125-133.