Jour. Ind. Soc. Ag. Statistics
48(1), 1996 : 105–113

On Double Sampling for Stratification with Sub-Sampling the Non- Respondents

F.C. Okafor
Department of Statistics, University of Nigeria, Nsukka, Nigeria
(Received : June, 1990)

SUMMARY

In the usual procedure of double sampling for stratification, DSS, it is generally assumed that there is total response on both the auxiliary variable used in estimating the stratum weights W_h and on the main character of interest. It may happen in practice that there is total response on the auxiliary variable and incomplete response on the main character. For example in household survey information on household size is readily available; while during the actual survey some households may withhold information on their family expenditure. Motivated by this we derive DSS estimators in the presence of non-response based on the subsampling of the non-respondents. The condition under which the proposed estimators are better than the usual DSS estimators, $\bar{y}_{ds}$, of the population mean $\bar{Y}$ is given.

Keywords : Double sampling; Non-response; Stratification; Subsampling.

Introduction

Stratification is one way of utilizing the auxiliary information to improve the precision of an estimate. Sometimes the information on the auxiliary character needed for stratification of units, e.g. age, sex, household size etc. is not available. In this situation we resort to double sampling or two phase sampling in which the information needed for stratification is collected at the first phase of sampling. In other words the first sample is used to distinguish the strata and obtain estimates of the stratum weights. While a smaller second phase sample is used to collect information on the main character of interest. This type of sampling is called double sampling for stratification (DSS).

Rao [5] proposed a DSS strategy for the estimation of the population mean $\bar{Y}$ of the variate, y, using the values of the auxiliary variate collected at the first phase for stratification only. Ige and Tripathi [3] went a step further and used the information collected at the first phase for stratification as well as in constructing ratio and difference estimators of the population mean $\bar{Y}$. So far all the authors who have dealt on DSS have assumed that all the units

selected, responded favourably to the enquiry. This may not be true in practice especially in mail interview and even in personal interview where some units may fail to supply information required. Hansen and Hurwitz [2] discussed a method of tackling total non-response in mail interview. This involves taking a simple random subsample of the non-respondents and interviewing them personally. It is assumed that at this second call all respond. The two estimates of the population mean obtained from the respondents at the first mail interview and the non-respondents at the second personal interview is suitably combined to yield the desired estimate, the population mean. Rao [6] applied this method of subsampling the non-respondents for the ratio estimation of the mean when the population mean of the auxiliary character is known.

In this paper an attempt is made to present a DSS strategy when there is total non response on the main character and total response on the auxiliary based on Rao [5] and Ige and Tripathi [3] DSS strategies.

2. *Rao [5] DSS Strategy in the Presence of Non Response*

An initial large sample of size n′ is selected from the population of N units by simple random sampling without replacement (SRSWOR). Information on the auxiliary variable x is collected with which an unbiased estimate $w_h = n'_h/n'$ of the true stratum weight, $W_h = N_h/N$, is calculated, n'_h is the number of units in the initial sample that falls in stratum h (h = 1, 2, ... L; $\sum_{h=1}^{L} n'_h = n'$). In each stratum a subsample of size $n_h = v_h n'_h$ ($0 < v_h < 1$, v_h is predetermined) is selected from n'_h by SRSWOR ($\sum_{h=1}^{L} n_h = n$, the second phase sample size). It is assumed that n′ is large so that $\Pr(n'_h = 0) = 0$ for all strata. The main character,y, is then observed on the n_h units. The DSS estimate of the population mean is given as

$$\bar{y}_{ds} = \sum_{h=1}^{L} w_h \bar{y}_h, \quad \bar{y}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} y_{hi} \tag{2.1}$$

The variance of $\bar{y}_{ds}$ as given by Rao [5] is

$$V(\bar{y}_{ds}) = \left(\frac{1}{n'} - \frac{1}{N}\right) S_y^2 + \frac{1}{n'} \sum_{h=1}^{L} W_h \left(\frac{1}{v_h} - 1\right) S_{yh}^2 \tag{2.2}$$

$$S_y^2 = \frac{1}{N-1} \sum_{i=1}^{N} (y_i - \bar{Y})^2; \quad S_{yh}^2 = \frac{1}{N_h - 1} \sum_{i=1}^{N_h} (y_{hi} - \bar{Y}_h)^2$$

The result in (2.1) above assumes total response. Let n_{1h} units respond at the first call from the n_h units selected in stratum h and n_{2h} not respond.

Following Hansen and Hurwitz [2], select a subsample of $m_{2h} = n_{2h}/k_h$ units ($k_h > 1$, a known constant) from the non respondents.

Interview these units with improved method. The estimator for $\bar{Y}$ becomes

$$\bar{y}^*_{ds} = \sum_{h=1}^{L} w_h \bar{y}^*_h \tag{2.3}$$

$$\bar{y}^*_h = \frac{n_{1h}\bar{y}_{1h} + n_{2h}\bar{y}_{m_{2h}}}{n_h}$$

$\bar{y}_{1h}$ = sample mean for the respondents based on n_{1h} units.

$\bar{y}_{m_{2h}}$ = sample mean for the non respondents based on m_{2h} units.

Clearly $\bar{y}^*_{ds}$ is an unbiased estimator of $\bar{Y}$ since

$$E(\bar{y}^*_{ds}) = E_d E_r (\bar{y}^*_{ds} \mid n'_h, n_{2h}) = \bar{Y}$$

E_d = expectation for DSS

E_r = expectation for subsampling the non respondents

and

$$V(\bar{y}^*_{ds}) = E_d V_r (\bar{y}^*_{ds} \mid n'_h, n_{2h}) + V_d E_r (\bar{y}^*_{ds} \mid n'_h, n_{2h})$$

$V_d E_r (\bar{y}^*_{ds} \mid n'_h, n_{2h}) = V(\bar{y}_{ds})$ given in (2.2)

$$E_d V_r (\bar{y}^*_{ds} \mid n'_h, n_{2h}) = \frac{1}{n'} \sum_{h=1}^{L} W_{2h} \frac{k_h - 1}{v_h} S^2_{2yh} \tag{2.4}$$

Combining (2.2) and (2.4)

$$V(\bar{y}^*_{ds}) = V(\bar{y}_{ds}) + \frac{1}{n'} \sum_{h=1}^{L} W_{2h} \frac{(k_h - 1)}{v_h} S^2_{2yh} \tag{2.5}$$

$W_{2h} = \frac{N_{2h}}{N}$, population proportion of the non-respondents in stratum h.

S^2_{2yh} is the population variance of the non respondent group in stratum h.

3. *Ige and Tripathi [3] DSS Strategies in the Presence of Non-response*

Ige and Tripathi [3] gave the following ratio and difference estimators of the mean when there is total response as

$$e_{RC} = \bar{y}_{ds}\,\bar{x}'/\bar{x}_{ds}, \quad \bar{x}' = \sum_{h=1}^{L} w_h \bar{x}'_h \tag{3.1}$$

$$e_{DC} = \bar{y}_{ds} - \lambda\,(\bar{x}_{ds} - \bar{x}') \tag{3.2}$$

$$e_{DS} = \sum_{h=1}^{L} w_h \{\bar{y}_h - \lambda_h\,(\bar{x}_h - \bar{x}'_h)\} \tag{3.3}$$

with variances

$$V(e_{RC}) = \left(\frac{1}{n'} - \frac{1}{N}\right) S_y^2 + \frac{1}{n'} \sum_{h=1}^{L} W_h \left(\frac{1}{\nu_h} - 1\right)(S_{yh}^2 + R^2 S_{xh}^2 - 2RS_{xyh}) \tag{3.4}$$

$$R = \bar{Y}/\bar{X}\,; S_{xyh} = \frac{1}{N_h - 1} \sum_{i=1}^{N_h} (x_{hi} - \bar{X}_{h.})(y_{hi} - \bar{Y}_{h.})$$

$$V(e_{DC}) = V(e_{RC}) \text{ with } R = \lambda$$

$$V(e_{DS}) = \left(\frac{1}{n'} - \frac{1}{N}\right) S_y^2 + \frac{1}{n'} \sum_{h=1} W_h \left(\frac{1}{\nu_h} - 1\right)(S_{yh}^2 + \lambda_h^2 S_{xh}^2 - 2\lambda_h S_{xyh}) \tag{3.5}$$

Again the above estimators assume total response. In the case of some refusals, the subsampling procedure used in section 2 will be used and the estimators become

$$e_{RC}^* = \frac{\bar{y}_{ds}^*}{\bar{x}_{ds}}\,\bar{x}' \tag{3.6}$$

$$e_{DC}^* = \bar{y}_{ds}^* - \lambda\,(\bar{x}_{ds} - \bar{x}') \tag{3.7}$$

$$e_{DS}^* = \sum_{h=1}^{L} W_h\,(\bar{y}_h^* - \lambda_h\,(\bar{x}_h - \bar{x}')) \tag{3.8}$$

Their variances are given using the same procedure above as

$$V(e_{RC}^*) = V(e_{RC}) + \frac{1}{n'} \sum_{h=1}^{L} W_{2h} \frac{k_h - 1}{\nu_h} S_{2yh}^2 \tag{3.9}$$

$$V(e_{DC}^*) = V(e_{DC}) + \frac{1}{n'} \sum_{h=1} W_{2h} \frac{k_h - 1}{\nu_h} S_{2yh}^2 \tag{3.10}$$

The optimum value of λ used in (3.7) is given by

$$\lambda_0 = \beta_{ds} = \frac{\sum_h W_h \left(\frac{1}{\nu_h} - 1 \right) S_{xyh}}{\sum_h W_h \left(\frac{1}{\nu_h} - 1 \right) S_{xh}^2}$$

Substituting λ_0 in (3.10) the optimum variance of e_{DC}^* becomes

$$V_0(e_{DC}^*) = \left(\frac{1}{n'} - \frac{1}{N} \right) S_y^2 + \frac{1 - \rho_{ds}^2}{n'} \sum W_h \left(\frac{1}{\nu_h} - 1 \right) S_{yh}^2 + \frac{1}{n'} \sum_h W_{2h} \frac{k_h - 1}{\nu_h} S_{2yh}^2 \tag{3.11}$$

where

$$\rho_{ds} = \left[\sum_h W_h \left(\frac{1}{\nu_h} - 1 \right) S_{xyh} \right] \left[\sum_h W_h \left(\frac{1}{\nu_h} - 1 \right) S_{xh}^2 \cdot \sum W_h \left(\frac{1}{\nu_h} - 1 \right) S_{yh}^2 \right]^{\frac{1}{2}}$$

While the variance of e_{DS}^* is

$$V(e_{DS}^*) = V(e_{DS}) + \frac{1}{n'} \sum_h W_{2h} \frac{k_h - 1}{\nu_h} S_{2yh}^2 \tag{3.12}$$

Optimum λ_h used in (3.8) is

$$\lambda_{0h} = \beta_h = S_{xyh} / S_{xh}^2$$

Hence, the optimum variance of e_{DS}^* is

$$V_0(e_{DS}^*) = \left(\frac{1}{n'} - \frac{1}{N} \right) S_y^2 + \frac{1}{n'} \sum_h \left(\frac{1}{\nu_h} - 1 \right) (1 - \rho_h^2) S_{yh}^2 + \frac{1}{n'} \sum_h W_{2h} \frac{k_h - 1}{\nu_h} S_{2yh}^2 \tag{3.13}$$

Remark: It may happen that not all strata experience non response. In other words in some strata all the units may respond while in others some units may fail to respond.

In this case we set in (2.3), (3.6), (3.7) and (3.8)

$\bar{y}_h^* = \bar{y}_h$ if complete response occurs in stratum h

$= \bar{y}_h^*$ if there is subsampling of non-respondents in stratum h

For the variance set in (2.5), (3.9), (3.10) and (3.12)

$S_{2yh}^2 = 0$ if complete response in stratum h

$= S_{2yh}^2$ if there is subsampling of non respondents in stratum h

4. *Comparison of the Proposed Estimators*

4.1 Theoretical Comparison

We shall now compare the proposed estimators with the DSS estimator, $\bar{y}_{ds}$.

From (2.2) and (2.5) we find that $\bar{y}_{ds}^*$ has a higher variance than $\bar{y}_{ds}$ due to the subsampling of the non-respondents.

Comparing $\bar{y}_{ds}$ and e_{RC}^* we deduce that e_{RC}^* will be better than $\bar{y}_{ds}$ in spite of subsampling the non-respondents if

$$\sum_h W_h \left(\frac{1}{\nu_h} - 1 \right) (2RS_{xyh} - R^2 S_{xh}^2) > \sum_h W_{2h} \frac{k_h - 1}{\nu_h} S_{2yh}^2$$

While the condition under which e_{DC}^* is to have a smaller variance than $\bar{y}_{ds}$ is obtained from (2.2) and (3.10) as

$$\rho_{ds}^2 \sum_h W_h \left(\frac{1}{\nu_h} - 1 \right) S_{yh}^2 > \sum_h W_{2h} \frac{k_h - 1}{\nu_h} S_{2yh}^2$$

Finally e_{DS}^* has a smaller variance than $\bar{y}_{ds}$ if

$$\sum W_h \left(\frac{1}{\nu_h} - 1 \right) \rho_h^2 S_{yh}^2 > \sum W_{2h} \frac{k_h - 1}{\nu_h} S_{2yh}^2$$

4.2 Empirical Comparison

To investigate the relative efficiency of the proposed estimators with respect to $\bar{y}_{ds}$ we make use of census data in Murthy [4] [p.127, Table 4.12]. For the purpose of the analysis both the area of each village and the area cultivated in the village are converted to hectares and grouped into three strata

with area of the village as the stratifying variable, x. The idea is to use DSS to estimate the mean area under cultivation. Within each strata, the population was subdivided into respondent and non-respondent groups. Villages with larger area are considered to belong to the non respondent group.

Table 4.1 shows the parameters obtained from the census data after stratification, and used for the calculation of the relative efficiency for sample sizes (n, n′).

Table 4.1. Population Parameter

Stratum	W_h	W_{2h}	S^2_{yh}	S^2_{xh}	S^2_{2yh}	S_{xyh}
0–930	0.336	0.148	39974.81	54624.49	14549.99	35507.36
931–1700	0.325	0.133	61455.48	54862.44	17386.54	17473.07
1701–4300	0.313	0.125	172425.05	428164.23	71175.11	137254.78

$R = 0.54299$

For the calculation of the efficiencies shown in Tables 4.2 and 4.3 we assumed that N = 10,000 and that $k_h = k = 2$ in all strata.

Table 4.2. Relative Efficiency of the Proposed Estimators over $\bar{y}_{ds}$

Sample Sizes (n′, n) Estimators	(5000, 2500)	(5000, 1000	(5000, 500	(2000, 1000)	(2000, 400)	(2000, 200)	(1000, 500)	(1000, 200)	(1000, 100)
$\bar{y}^*_{ds}$	0.8767	0.8724	0.8709	0.9037	0.8856	0.8780	0.9102	0.8895	0.8802
e^*_{RC}	0.9449	0.9909	1.0081	0.9577	0.9920	1.0076	0.9607	0.9923	1.0075
e^*_{DC} $(\lambda = \lambda_0)$	0.9741	1.0455	1.0732	0.9802	1.0400	1.0882	0.9817	1.0384	1.0667
e^*_{DS} $(\lambda_h = \lambda_{0h})$	0.9831	1.0628	1.0940	0.9871	1.0551	1.0876	0.9881	1.0529	1.0856

From Table 4.2 we notice that the combined ratio estimator e^*_{RC} when there is non response has no much improvement over $\bar{y}_{ds}$ in the estimation of the population mean, $\bar{Y}$.

While the combined difference, e^*_{DC} and separate difference estimator, e^*_{DS} has a slight improvement over $\bar{y}_{ds}$, the gain in efficiency range between 4% and 9%. We also observe from the same table that when the second phase

sampling fraction is $\frac{1}{2}$, all the estimators showed a loss in efficiency. But when it is less than $\frac{1}{2}$, e^*_{DC} and e^*_{DS} exhibit a gain in efficiency.

Table 4.3. Relative Efficiency of e^*_{DC} and e^*_{DS} over $\bar{y}_{ds}$ for n′ = 5000

Sample Sizes n / Estimators	2500	1250	1000	625	500	250	200	100
e^*_{DC} ($\lambda = \lambda_0$)	0.9741	1.0325	1.0455	1.0660	1.0732	1.0878	1.0908	1.0969
e^*_{DS} ($\lambda_h = \lambda_{0h}$)	0.9831	1.0482	1.0628	1.0860	1.0940	1.1107	1.1141	1.1210

Table 4.3 shows us the relative efficiency of e^*_{DC} and e^*_{DS} for varying second phase sampling fraction. We note that as the second phase sampling fraction decreases the efficiency of e^*_{DC} and e^*_{DS} increases, from a loss of 4% when the sampling fraction is $\frac{1}{2}$ to a gain of 12% when the sampling fraction falls to 1/50.

5. *Optimum Allocation*

Consider the cost function

$$C = C_1 n' + \sum_h C_{2h} n_h + \sum_h C_{21h} n_{1h} + \sum_h C_{22h} n_{2h}/k_h \qquad (5.1)$$

C's are the cost per unit.

C_1 is cost of getting information on the first phase sample.

C_{2h} is cost of first attempt on the main character in stratum h.

C_{21h} is cost of processing the results on the main character from the respondents at the first attempt at the second phase sample in stratum h.

C_{22h} is cost of getting and processing results on the main character from the subsample of the non respondents at the second phase sample in stratum h.

Since the value of n_{1h} is not known until the first attempt is made, the expected cost will be used. Using double expectation the expected cost is

$$E(C) = C^* = C_1 n' + n' \Sigma C_{2h} v_h W_h + n' \Sigma C_{21h} v_h W_{1h} + n' \Sigma C_{22h} v_h W_{2h}/k_h \qquad (5.2)$$

$$W_{1h} = 1 - W_{2h}$$

To obtain the optimum values of n', v_h and k_h we adopt a stepwise minimization technique. First using Lagrange's multiplier we minimize the variance of $\bar{y}^*_{ds}$ (see (2.5)) subject to the fixed expected cost C* given in (5.2). This results in the optimum value of k_h given by

$$k_{0h} = \frac{1}{2} \{ (C_{22h} (1 - S^2_{2yh}))^2 + 4C_{22h} S^2_{2yh} C_h \Delta_h \}^{\frac{1}{2}} / S^2_{2yh} C_h \qquad (5.3)$$

where $C_h = C_{2h} W_h + C_{21h} W_{1h}$

$$\Delta_h = W_h S^2_{yh} - W_{2h} S^2_{2yh}$$

By plugging k_{0h} in (5.2) and (2.5) and following Cochran [1] the optimum value of v_h is

$$v_{0h} = \{ C_1 (\Delta_h + W_{2h} k_{0h} S^2_{2yh}) \}^{\frac{1}{2}} \div \{ (S^2_y - \Sigma W_h S^2_{yh}) (C_h + C_{22h} W_{2h} / k_{0h}) \}^{\frac{1}{2}} \qquad (5.4)$$

The optimum n' is hence obtained for either fixed cost or fixed variance using (5.2) or (2.5). For e^*_{DC}, the optimum values of k_h and v_h are obtained by replacing S^2_{yh} in (5.3) and (5.4) with $S^2_{yh} + \lambda^2 S^2_{xh} - 2\lambda S_{xyh}$. While for e^*_{RC} and e^*_{DS}, S^2_{yh} is replaced with $S^2_{yh} + R^2 S^2_{xh} - 2RS_{xyh}$ and $S^2_{yh} + \lambda^2_h S^2_{xh} - 2\lambda_h S_{xyh}$ respectively.

REFERENCES

[1] Cochran, W.G., 1977. Sampling Techniques, 3rd Edition, N.Y. Wiley P.

[2] Hansen, M.H. and Hurwitz, W.N., 1946. The Problem of Non response in Sample Surveys. *Journ. of the American Statistical Association*, **41**, 517-529.

[3] Ige, A.F. and Tripathi, T.P., 1987. On Double Sampling for Stratification and Use of Auxiliary information. *Jour. Ind. Soc. Agric Stat*, **39**. No 2, 191-201.

[4] Murthy, M.N., 1967. Sampling Theory and Methods, 1st ed. 127-130.

[5] Rao, J.N.K., 1973. On Double Sampling for Stratification and Analytical Surveys. *Biometrika*, **60**, 125-133.

[6] Rao, P.S.R.S., 1986. Ratio Estimation with Subsampling the non-respondents. Survey Methodology, 12, 2, 217-230.