J. Ind. Soc. Agril. Statist.
58(3), 2004 : 300-306

Double Sampling for Ratio Estimation with Non-response

Rifat Tabasum and I.A. Khan
Aligarh Muslim University, Aligarh – 202 002

(Received : June, 2003)

SUMMARY

In this paper we have considered ratio estimator under double sampling in presence of non-response when the population mean of the auxiliary variable is unknown and obtained the first phase sample size, second phase sample size and sub-sampling fraction for the proposed estimator which minimize the survey cost for a specified precision. The cost obtained for the proposed estimator is compared theoretically and numerically with that of the cost obtained by Hansen and Hurwitz estimator and found that survey cost for our proposed estimator is less than the cost obtained by Hansen and Hurwitz estimator.

Key words : Hansen-Hurwitz estimator, Optimum allocation, Double sampling, Non-response.

1. Introduction

In general during surveys, it is observed that information in most cases are not obtained at the first attempt even after some call-backs. An estimate obtained from such incomplete data may be misleading because of the biased estimator. This is the case of non-response and the usual approach to face the non-response is to recontact the non-respondents and obtained the information as much as possible. The work of Hansen and Hurwitz (1946) pioneering the treatment of non-response, assumes that a sub-sample of initial non-respondents is recontacted with a more expensive method, they suggested the first attempt by mail questionnaire and the second attempt by a personal interview. Survey based on Hansen-Hurwitz technique costs more because of extra work of personal interviews. Using Hansen and Hurwitz (1946) procedure, Cochran (1977) proposed the ratio and regression estimators of the population mean of the study variable in which information on the auxiliary variable is obtained from all the sample units, and the population mean of the auxiliary variable is known, while some sample units failed to supply information on the study variable. Further improvement in the estimation procedure for population mean in presence of non-response using auxiliary character was suggested by Rao ((1986), (1987)) and Khare and Srivastava ((1993), (1997)).

In this paper we have considered ratio estimator for population mean under double sampling in presence of non-response where population mean of auxiliary variable is not known and obtained the optimum values of the first and second phase sample and sub-sampling fraction which minimize the survey cost for specified precision. The cost obtained for our proposed estimator is compared theoretically and numerically through an artificially generated population with that of the cost obtained by Hansen and Hurwitz estimator and found that the cost of our proposed estimator is less than the cost obtained by Hansen-Hurwitz estimator under certain given conditions.

2. Sampling Scheme

For the estimate of population mean $\overline{X}$ of the auxiliary variable x, a large first phase sample of size n' is selected from a population of N units by simple random sampling without replacement (SRSWOR). A smaller second phase sample of size n is selected from n' by simple random sampling without replacement (SRSWOR) and the character y under study is measured on it. The ratio estimator of the mean of y is $\overline{y}'_r = (\overline{y}/\overline{x})\overline{x}'$ where, $\overline{x}'$ is the sample mean from n' units. $\overline{y}$ and $\overline{x}$ are obtained from the second phase sample when there is no non-response in the second phase sample. If however, there is non-response in the second phase sample, take a sub-sample of the non-respondents and re-contact them.

Let us assume that at the first phase, all the n' units supplied information on the auxiliary variable x. At the second phase from sample n, let n_1 units supply information on y and n_2 refuse to respond. Using Hansen and Hurwitz (1946) approach to sub-sampling, from the n_2 non-respondents a sub-sample, of size m units is selected at random and is enumerated by direct interview, such that $m = \frac{n_2}{k}, k > 1$.

Here we assume that response is obtained for all the m units. This method of double sampling can be applied in a household survey where the household size is used as an auxiliary variable for the estimation of family expenditure. Information can be obtained completely on the family size, while there may be some non-response on the household expenditure. The whole population is divided into two classes, one consists of N_1 units, which would respond on the first attempt at the second phase and the other consists of N_2 units, which would not respond on the first attempt at the second phase of sampling but will respond on the second attempt.

3. The Double Sampling Ratio Estimator

We define the double sampling ratio estimator as

$$\bar{y}_R^* = \frac{\bar{y}^*}{\bar{x}^*}\bar{x}' = r^*\bar{x}' \tag{3.1}$$

where, $\bar{x}^*$ and $\bar{y}^*$ are the Hansen-Hurwitz estimators for $\bar{X}$ and $\bar{Y}$, respectively and are given by

$$\bar{x}^* = w_1\bar{x}_1 + w_2\bar{x}_{2m}; \ \bar{y}^* = w_1\bar{y}_1 + w_2\bar{y}_{2m} \tag{3.2}$$

where, $w_i = \frac{n_i}{n}$ and $W_i = \frac{N_i}{N}$; $i = 1, 2$

The ratio estimators are generally biased, but the bias is negligible if the sample size is large enough. The approximate variance of $\bar{y}_R^*$ (for large sample size) is given by

$$V(\bar{y}_R^*) \cong \left(\frac{1}{n'} - \frac{1}{N}\right)S_y^2 + \left(\frac{1}{n} - \frac{1}{n'}\right)S_r^2 + \frac{W_2(k-1)}{n}S_{2r}^2 \tag{3.3}$$

where

$$S_r^2 = S_y^2 + R^2S_x^2 - 2RS_{xy}$$

$$S_{2r}^2 = S_{2y}^2 + R^2S_{2x}^2 - 2RS_{2xy} \tag{3.4}$$

'R' is the population ratio of $\bar{Y}$ to $\bar{X}$. S_x^2, S_y^2 are the variance for the whole population and S_{2x}^2, S_{2y}^2 are the population variance for the stratum of non-respondents for the variable x and y respectively. S_{xy} and S_{2xy} are the covariances for the whole population and the population of non-respondents respectively.

Let us consider a cost function for $\bar{y}_R^*$ as

$$C = c'n' + cn + c_1n_1 + c_2m \tag{3.5}$$

where

$c' =$ The unit cost associated with first phase sample, n'

$c =$ The cost of the first attempt on y with the second phase sample, n

$c_1 =$ The unit cost for processing the respondent data on y at the first attempt in n_1

$c_2 =$ The unit cost associated with the sub-sample, m of n_2

Since the value of n_1 and m is not known until the first attempt is made, so the expected cost will be used in planning the survey. The expected values of n_1 and m are $W_1 n$ and $\frac{W_2 n}{k}$. Thus the expected cost is given by

$$E(C) = C^* = c'n' + \left(c + c_1 W_1 + \frac{c_2 W_2}{k}\right)n \tag{3.6}$$

To determine the optimum values of k, n and n' that minimize the cost for a fixed variance V_0 we consider the function

$$\phi = C^* + \lambda\left\{V(\bar{y}_R^*) - V_0\right\}$$

$$\phi = c'n' + \left(c + c_1 W_1 + \frac{c_2 W_2}{k}\right)n$$

$$+\lambda\left\{\left(\frac{1}{n'} - \frac{1}{N}\right)S_y^2 + \left(\frac{1}{n} - \frac{1}{n'}\right)S_r^2 + \left(\frac{W_2(k-1)}{n}\right)S_{2r}^2 - V_0\right\} \tag{3.7}$$

where λ is Lagrange's multiplier.

Using Lagrange's multiplier technique the optimum values of k, n, and n' thus obtained are

$$k_0 = \sqrt{\frac{c_2(S_r^2 - W_2 S_{2r}^2)}{S_{2r}^2(c + c_1 W_1)}}$$

$$n_0 = \frac{\sqrt{S_r^2 + W_2(k_0 - 1)S_{2r}^2}\left\{\sqrt{(S_y^2 - S_r^2)c'} + \sqrt{S_r^2 + W_2(k_0 - 1)S_{2r}^2}\sqrt{c + c_1 W_1 + \frac{c_2 W_2}{k_0}}\right\}}{\left(V_0 + \frac{S_y^2}{N}\right)\sqrt{c + c_1 W_1 + \frac{c_2 W_2}{k_0}}}$$

and $$n_0 = \frac{\left\{\sqrt{(S_y^2 - S_r^2)c'} + \sqrt{S_r^2 + W_2(k_0 - 1)S_{2r}^2}\sqrt{c + c_1 W_1 + \frac{c_2 W_2}{k_0}}\right\}\sqrt{S_y^2 - S_r^2}}{\left(V_0 + \frac{S_y^2}{N}\right)\sqrt{c'}}$$

4. Hansen – Hurwitz Estimator

The variance of the Hansen-Hurwitz estimator $\bar{y}^*$ is

$$V(\bar{y}^*) = \left(\frac{1}{n} - \frac{1}{N}\right)S_y^2 + \frac{W_2(k-1)}{n}S_{2y}^2 \tag{4.1}$$

The expected cost function is given by

$$C_1^* = \left(c + c_1 W_1 + \frac{c_2 W_2}{k}\right)n \tag{4.2}$$

To determine the optimum values of k and n that minimize the cost for fixed variance we consider the function as

$$\phi' = C_1^* + \left\{V(\bar{y}^*) - V_0\right\}$$

$$\phi' = \left(c + c_1 w_1 + \frac{c_2 w_2}{k}\right)n + \lambda\left\{\left(\frac{1}{n} - \frac{1}{N}\right)S_y^2 + \frac{W_2(k-1)S_{2y}^2}{n} - V_0\right\} \tag{4.3}$$

Using Lagrange multiplier technique the optimum values of k and n thus obtained are as follows

$$k_{0HH} = \sqrt{\frac{c_2(S_y^2 - W_2 S_{2y}^2)}{S_{2y}^2(c + c_1 W_1)}} \text{ and } n_{0HH} = \frac{S_y^2 + W_2(k-1)S_{2y}^2}{\left(V_0 + \frac{S_y^2}{N}\right)} \tag{4.4}$$

5. *Theoretical Comparison of the Estimators*

It is well known in literature that the double sampling ratio estimator will be more efficient than the simple random sampling estimator if $R < 2\beta$ or $\rho > \frac{1}{2}\frac{RS_x}{S_y}$. Also the double sampling ratio estimator in presence of non-response will be more efficient than the Hansen-Hurwitz estimator if $R < 2\beta_2$ or $\rho_2 > \frac{1}{2}\frac{RS_{2x}}{S_{2y}}$ where $\beta = \frac{S_{xy}}{S_x^2}$, $\beta_2 = \frac{S_{2xy}}{S_{2x}^2}$, $\rho = \frac{S_{xy}}{S_x S_y}$ and $\rho_2 = \frac{S_{2xy}}{S_{2x}S_{2y}}$

For our proposed estimator the cost of the survey for given precision, under optimum allocation will be less than the cost of the survey for Hansen-Hurwitz estimator if $C_1^* - C^* > 0$.

$$\text{i.e., } \left(c + c_1 W_1 + \frac{c_2 W_2}{k_{0HH}}\right)n_{0HH} - c'n_0' - \left(c + c_1 W_1 + \frac{c_2 W_2}{k_0}\right)n_0 > 0$$

So the condition that cost for our proposed estimator will be less than that of Hansen-Hurwitz estimator is given by

$$c' < \frac{c_1'}{\theta_2}(\theta_1 - 1) + \frac{c_2'}{\theta_2}\left(\frac{\theta_1}{k_{0HH}} - \frac{1}{k_0}\right)$$

Where, $\theta_1 = \frac{n_{0HH}}{n_0}$, $\theta_2 = \frac{n_0'}{n_0}$, $c_1' = c + c_1W_1$ and $c_2' = c_2W_2$

6. *Numerical Example for the Comparison of the Proposed Estimators*

The expected cost C^* for our proposed estimator $\bar{y}_R^*$ and expected cost C_1^* considered by Hansen-Hurwitz estimator $\bar{y}^*$ are compared (using artificially generated population). The parameters of the population are

$N = 500$, $N_2 = 150$, $R = 1.48$, $\rho = 0.81$, $S_x^2 = 350.54$, $S_y^2 = 1213.82$, $S_{xy} = 530.07$, $S_{2x}^2 = 150.04$, $S_{2y}^2 = 610.67$, $S_{2xy} = 253.68$, $S_r^2 = 412.49$, $S_{2r}^2 = 188.35$, $\beta = 1.69$, $\beta_2 = 1.69$, $\rho_2 = 0.83$

Table 6.1. Values of k_0, n_0', n_0' and expected C^* for double sampling ratio estimator $\bar{y}_R^*$ and values of C_1^* for Hansen-Hurwitz estimator $\bar{y}^*$

						For fixed variance $V_o = 5.41$						
W_1	W_2	c'	c	c_1	c_2	k_{0HH}	n_{0HH}	k_0	n_0	n_0'	Expected Cost C^*	Expected Cost C_1^*
0.7	0.3	0.1	0.5	1	2	1.67	170	1.77	78	411	158	256
		0.2	0.6	1.4	3	1.78	173	1.89	83	357	240	353
		0.3	0.8	1.6	4	1.87	175	1.98	86	333	314	435
		0.4	0.9	1.9	5	1.94	177	2.05	89	317	389	522

From Table 6.1, it is observed that the expected cost C^* for double sampling ratio estimator is less than the expected cost C_1^* of Hansen-Hurwitz estimator.

Conclusion

The ratio estimator based on the double sampling procedure has been proposed when there is non-response on the main character and the population mean of the auxiliary variable is not known. The potentially serious non-response bias is eliminated by sub-sampling the non-respondents as in the Hansen and Hurwitz procedure (1946). From the above table, we conclude that the total expected cost C^* of the double sampling ratio estimator is less than that of cost C_1^* obtained using Hansen-Hurwitz estimator.

REFERENCES

Cochran, W.G. (1977). *Sampling Techniques.* 3rd edition. John Wiley & Sons, New York.

Hansen, M.H., and Hurwitz, W.N. (1946). The problem of non-response in sample surveys. *J. Amer. Statist. Assoc.*, **41**, 517-529.

Khare, B. B. and Srivastava, S. (1993). Estimation of population mean using auxiliary character in presence of non-response. *Natl. Aca. Sci. Lett.*, India, **16(3)**, 111-114.

Khare, B. B. and Srivastava, S. (1997). Transformed ratio type estimators for the population mean in the presence of non-response. *Comm. Stat. – Theory Methods*, **26(7)**, 1779-1791.

Rao, P.S.R.S. (1986). Ratio estimation with sub-sampling the non-respondents. *Survey Methodology*, **12(2)**, 217-230.

Rao, P.S.R.S. (1987). Ratio and regression estimates with sub-sampling the non-respondent. *Paper presented at a special contributed session of the International Statistical Association Meeting, Sept.*, 2-16, 1987, Tokyo, Japan.