Jurnal Teknologi

Full paper

A New Cost Effective Estimator in the Presence of Non-response for Two-phase Sampling

Muhammad Ismail[a*], Muhammad Hanif[b], Muhammad Qaiser Shahbaz[a]

[a]*Department of Mathematics, COMSATS Institute of IT, Lahore Pakistan*
[b]*School National College of Business Administration & Economics, Lahore Pakistan*

*Corresponding author: drismail39@gmail.com

Article history

Received :21 January 2013
Received in revised form :
7 May 2013
Accepted :25 June 2013

Graphical abstract

$$\bar{y}^* = (r_1/n)\bar{y}_{r_1 1} + (r_2/n)\bar{y}_{k2}$$

Abstract

In In this paper we have proposed a new estimator of population mean in the presence of non–response using information of a single auxiliary variable. We have obtained survey cost for the fixed variance of the proposed estimator and compared it with the cost obtained by Tabasum and Khan (2004) and Hansen Hurwitz (1946). After the comparison we saw that the cost of our proposed estimator is lesser than Tabasum and Khan (2004) and Hansen Hurwitz (1946) estimators.

Keywords: Non-response; Hansen Hurwitz estimator; Tabasum and Khan estimator; auxiliary variable; two-phase sampling

Abstrak

Dalam makalah ini, kami telah mencadangkan satu penganggar baru bagi min populasi dengan kehadiran tak-sambut penggunaan informasi bagi satu pemboleh ubah sokongan. Kami telah mendapatkan kos kaji selidik bagi varian tetap penganggar yang dicadangkan dan dibandingkan dengan kos yang diperoleh oleh penganggar Tabasum dan Khan dan penganggar Hansen Hurwitz. Selepas perbandingan, didapati bahawa kos penganggar yang dicadang adalah yang paling kurang berbanding dengan kos penganggar Tabasum and Khan dan kos penganggar Hansen Hurwitz.

Kata kunci: Tak-sambut; penganggar Hansen Hurwitz; penganggar Tabasum dan Khan; pemboleh ubah sokongan

1.0 INTRODUCTION

Non–response has been a major problem of almost every sample surveys. The incomplete data create many problems for researcher and this problem cannot be eliminated even by increasing the sample size. The non–response always exists when surveying human populations as people hesitate to respond in surveys. In sensitive issues the non–response rate increases. The pioneer researchers in this area were Hansen and Hurwitz (1946). After that many survey statisticians have suggested methods of estimating population characteristics in the presence of non–response. The sub-sampling method has been a popular method in case of non–response. Due to sub sampling the cost survey is increased.

In this paper we have proposed a new estimator for population mean under Two-phase sampling in the presence of non-response. We have also derived mean square error for that proposed estimator and obtain the optimum values of the sample sizes at first phase, second phase and sampling fraction which minimize the survey cost.

We have compared empirically the survey cost of new proposed estimators with the Hansen and Hurwitz (1946) estimator and Tabasum and Khan (2004) estimator. We found that the cost of our proposed estimator is less than the cost obtained by Hansen and Hurwitz (1946) estimator and Tabasum and Khan (2004) estimator.

2.0 MATERIALS AND METHODS

2.1 Two-phase Sampling Scheme

Suppose a simple random sample without replacement (SRSWOR) of size n is drawn from a population of size N. From the available sample, r_1 units respond to survey variable Y and r_2 units do not respond. Corresponding to sample respondents and non–respondents, the population is also divided in same sort of groups containing N_1 and N_2 units. Out of r_2 non–respondents, a sub-sample of k ($k=r_2/h$, $h>1$) units is drawn and information is obtained from these k units. Hansen and Hurwitz (1946) suggested following estimator of population mean when sub-sampling is used to overcome non–response:

$$\bar{y}^* = (r_1/n)\bar{y}_{r_1 1} + (r_2/n)\bar{y}_{k2}, \qquad (2.1)$$

where $\bar{y}_1 = r_1^{-1}\sum_{i=1}^{r_1} y_i$ and $\bar{y}_{k2} = k^{-1}\sum_{i=1}^{k} y_i$ are means of variable of interest. The estimator (2.1) is unbiased with variance:

$$Var(\bar{y}^*) = \lambda_2 S_y^2 + \theta S_{y_2}^2, \qquad (2.2)$$

with $S_y^2 = \sum_{i=1}^{N}(y_i - \bar{Y})^2/(N-1)$, $S_{y_2}^2 = \sum_{i=1}^{N_2}(y_i - \bar{Y}_2)^2/(N_2-1)$,

$\lambda = (1-f)/n$, $f = n/N$,

$\theta = W_2(h-1)/n$, $W_2 = N_2/N$, $\lambda_1 = n_1^{-1} - N^{-1}$,

$\lambda_2 = n_2^{-1} - N^{-1}$, $\lambda_3 = n_2^{-1} - n_1^{-1}$

$\bar{Y} = N^{-1}\sum_{i=1}^{N} y_i$ and $\bar{Y}_2 = N_2^{-1}\sum_{i=1}^{N_2} y_i$.

2.2 Non-response in Two Phase Sampling

The two phase sampling procedure has been effectively used in the presence of non–response to increase the precision of estimates. The two phase sampling procedure in case of non–response is described as:

i) Select a first phase sample of size n_1 using SRSWOR and record information on auxiliary variable X.

ii) Select a second phase sample of size n_2 using SRSWOR from first phase sample of size n_1. The r_1 units respond and r_2 units do not respond. Collect information on study variable Y from responding units.

iii) Select a subsample of size k ($k=r_2/h$, $h>1$) and record information on study variable from these selected units.

Using above two phase sampling procedure, various authors have proposed different estimators of population mean in the presence of non–response. Some notable references are of Cochran (1977), Rao (1986), Naik and Gupta (1991), Tripathi and Khare (1997), Tabasum and Khan (2004, 2006) and Khare and Srivastava (1993, 1995, 2010), Singh and Kumar (2008a, 2008b, 2008c, 2009, 2011).

2.3 New Proposed Estimator with Cost Function and Optimum Values Modeling Approach

The proposed estimator for the situation, when non-response occurs in study variable y and auxiliary variable x is

$$t_d = \bar{y}^* - \left(\sqrt{\bar{x}^*} - \sqrt{\bar{x}_1}\right)$$

We know that

$$\bar{e}_y^* = \bar{y}^* - \bar{Y} \quad \Rightarrow \quad \bar{y}^* = \bar{Y} + \bar{e}_y^*$$

$$\bar{e}_x^* = \bar{x}^* - \bar{X} \quad \Rightarrow \quad \bar{x}^* = \bar{X} + \bar{e}_x^*$$

$$\bar{e}_{x_1} = \bar{x}_1 - \bar{X} \quad \Rightarrow \quad \bar{x}_1 = \bar{X} + \bar{e}_{x_1}$$

Putting the values of $\bar{e}_y^*$, $\bar{e}_x^*$, and $\bar{e}_{x_1}$ in (1) we get

$$t_d = \left(\bar{Y} + \bar{e}_y^*\right) - \left(\sqrt{\bar{X} + \bar{e}_x^*} - \sqrt{\bar{X} + \bar{e}_{x_1}}\right)$$

or

$$t_d - \bar{Y} = \bar{e}_y^* - \frac{1}{2\sqrt{\bar{X}}}\left(\bar{e}_x^* - \bar{e}_{x_1}\right)$$

Taking Square and apply Expectation on both sides, we have

$$MSE(t_d) \approx E\left(\bar{e}_y^{*2}\right) + \frac{1}{4\bar{X}}\left[E\left(\bar{e}_x^{*2}\right) + E\left(\bar{e}_{x_1}^2\right) - 2E\left(\bar{e}_x^*\bar{e}_{x_1}\right)\right] - \frac{1}{\sqrt{\bar{X}}}\left[E\left(\bar{e}_y^*\bar{e}_x^*\right) - E\left(\bar{e}_y^*\bar{e}_{x_1}\right)\right]$$

We know that

$E\left(\bar{e}_y^{*2}\right) = \lambda_2 S_y^2 + \theta S_{y_2}^2$, $E\left(\bar{e}_y^*\bar{e}_{x_1}\right) = \lambda_1 S_{xy}$

$E\left(\bar{e}_x^{*2}\right) = \lambda_2 S_x^2 + \theta S_{x_2}^2$, $E\left(\bar{e}_x^*\bar{e}_y^*\right) = \lambda_2 S_{xy} + \theta S_{xy2}$

$E\left(\bar{e}_{x_1}^2\right) = \lambda_1 S_x^2$, $E\left(\bar{e}_{x_1}\bar{e}_x^*\right) = \lambda_1 S_x^2$ and $R = \frac{\bar{Y}}{\bar{X}}$

We can write

$$Var(T_{R1d}) = \lambda_2 S_y^2 + \theta S_{y_2}^2 + \frac{1}{4\bar{X}}\left(\lambda_2 S_x^2 + \theta S_{x_2}^2 + \lambda_1 S_x^2 - 2\lambda_1 S_x^2\right) - \frac{1}{\sqrt{\bar{X}}}\left[\left(\lambda_2 S_{xy} + \theta S_{xy2}^2 - \lambda_1 S_{xy}\right)\right]$$

or

$$MSE(t_d) \approx \lambda_3\left(S_y^2 + \frac{S_x^2}{4\bar{X}} - \frac{S_{xy}}{\sqrt{\bar{X}}}\right) + \lambda_1 S_y^2 + \theta\left(S_{y_2}^2 + \frac{S_{x_2}^2}{4\bar{X}} - \frac{S_{xy2}}{\sqrt{\bar{X}}}\right)$$

Let us consider a cost function

$$C = c_1 n_1 + c_2 n_2 + c_3 r_1 + c_4 k$$

Where

c_1 = The unit cost associated with first phase sample, n_1

c_2 = The cost of first attempt on Y with second phase sample, n_2

c_3 = The unit cost for processing the respondent data on Y at the first attempt in r_1

c_4 = The unit cost associated with the sub-sample k of r_2

Since the value of r_1 and k is not known until the first attempt is made, so the expected cost will be used in planning survey. The expected value of r_1 and k are $W_1 n_2$ and $\frac{W_2 n_2}{h}$. Thus the expected cost is given by

$$E(C) = C^* = c_1 n_1 + \left(c_2 + c_3 W_1 + \frac{c_4 W_2}{h}\right)n_2$$

To determine the optimum values of h, n_2, and n_1 that minimize the cost for a fixed variance V_0 we consider the function

$$\phi = C^* + \lambda\left\{MSE(t_d) - V_o\right\}$$

$$\phi = c_1 n_1 + \left(c_2 + c_3 W_1 + \frac{c_4 W_2}{h}\right)n_2 + \lambda\left\{\left(\frac{1}{n_1} - \frac{1}{N}\right)S_y^2 + \left(\frac{1}{n_2} - \frac{1}{n_1}\right)S_r^2 + \left(\frac{W_2(h-1)}{n_2}\right)S_{2r}^2 - V_o\right\}$$

Where

$$S_r^2 = S_y^2 + \frac{S_x^2}{4\bar{X}} - \frac{S_{xy}}{\sqrt{\bar{X}}}$$

$$S_{r_2}^2 = S_{y_2}^2 + \frac{S_{x_2}^2}{4\bar{X}} - \frac{S_{xy2}}{\sqrt{\bar{X}}}$$

Where λ is Lagrange's multiplier.

Using Lagrange's multiplier technique the optimum values h, n_2 and n_1 are

$$\phi = c_1 n_1 + \left(c_2 + c_3 W_1 + \frac{c_4 W_2}{h}\right)n_2 + \lambda\left\{\left(\frac{1}{n_1} - \frac{1}{N}\right)S_y^2 + \left(\frac{1}{n_2} - \frac{1}{n_1}\right)S_r^2 + \left(\frac{W_2(h-1)}{n_2}\right)S_{r_2}^2 - V_o\right\}$$

For optimum valus of λ, differentiate w.r.t. λ and equate to zero.

$$\left(\frac{1}{n_1}\right)\left(S_y^2 - S_r^2\right) + \frac{1}{n^2}[S_r^2 + w_2(h-1)S_r^2] = V_0 + \frac{S_y^2}{N} \qquad (2.3)$$

Now we differentiate w.r.t. h and equate to zero.

$$h^2 = \frac{n_2^2 c_4}{\lambda S_r^2} \tag{2.4}$$

Now we differentiate w.r.t. n_2, we get

$$n_2^2 = \frac{\lambda\left(S_r^2 + W_2(h-1)S_{r_2}^2\right)}{c_2 + c_3W_1 + \frac{c_4W_2}{h}}$$

By putting the value of n_2^2 in (2.4), we get the value of "h"

$$h = \sqrt{\frac{c_4\left(S_r^2 - W_2S_{r_2}^2\right)}{S_{r_2}^2\left(c_2 + c_3W_1\right)}}$$

Differentiate w.r.t. n_1, we get

$$n_1 = \sqrt{\frac{\lambda}{c_1}\left(S_y^2 - S_r^2\right)}$$

Putting the value of n_1 in equation(2.3)

$$\sqrt{\lambda} = \frac{\sqrt{c_1\left(S_y^2 - S_r^2\right)} + \left(\sqrt{c_2 + c_3W_1 + \frac{c_4W_2}{h}}\right)\left(\sqrt{S_r^2 + W_2(h-1)S_{r_2}^2}\right)}{V_o + \frac{S_y^2}{N}}$$

Putting the value of $\sqrt{\lambda}$ in equation (2.3)

$$n_1 = \frac{\left[\sqrt{c_1\left(S_y^2 - S_r^2\right)} + \left(\sqrt{c_2 + c_3W_1 + \frac{c_4W_2}{h}}\right)\left(\sqrt{S_r^2 + W_2(h-1)S_{r_2}^2}\right)\right]\sqrt{S_y^2 - S_r^2}}{\left(V_o + \frac{S_y^2}{N}\right)\sqrt{c_1}}$$

We have

$$n_2 = \sqrt{\frac{\lambda\left(S_r^2 + W_2(h-1)S_{r_2}^2\right)}{c_2 + c_3W_1 + \frac{c_4W_2}{h}}}$$

Replace the value of λ in above

$$n_2 = \frac{\sqrt{S_r^2 + W_2(h-1)S_{r_2}^2}\left[\sqrt{c_1\left(S_y^2 - S_r^2\right)} + \left(\sqrt{c_2 + c_3W_1 + \frac{c_4W_2}{h}}\right)\left(\sqrt{S_r^2 + W_2(h-1)S_{r_2}^2}\right)\right]}{\left(V_o + \frac{S_y^2}{N}\right)\sqrt{c_2 + c_3W_1 + \frac{c_4W_2}{h}}}$$

2.4 Cost Function and Optimum and Values in Hansen Hurwitz Estimator

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

$$Var\left(\bar{y}^*\right) = \lambda_2 S_y^2 + \theta S_{y_2}^2$$

The expected cost function is given by this

$$C_1^* = \left(c_2 + c_3W_1 + \frac{c_4W_2}{h}\right)n_2$$

To determine the optimum values of h, and n_2 that minimize the cost for a fixed variance V_0 we consider the function

$$\phi = C_1^* + \lambda\left\{Var\left(\bar{y}^*\right) - V_o\right\}$$

$$\phi = \left(c_2 + c_3W_1 + \frac{c_4W_2}{h}\right)n_2 + \lambda\left\{\left(\frac{1}{n_2} - \frac{1}{N}\right)S_y^2 + \left(\frac{W_2(h-1)}{n_2}\right)S_{y_2}^2 - V_o\right\}$$

Where λ is Lagrange's multiplier.

Using Lagrange's multiplier technique the optimum values h, n_2 and n_1 are

$$h_{oHH} = \sqrt{\frac{c_4\left(S_y^2 - W_2S_{y_2}^2\right)}{S_{y_2}^2\left(c_2 + c_3W_1\right)}},$$

$$n_{2HH} = \frac{S_y^2 + W_2(h-1)S_{y_2}^2}{\left(V_o + \frac{S_y^2}{N}\right)}$$

2.5 Cost Function and Optimum Values in Tabasum and Khan (2004)

Tabasum and Khan (2004) defined the double sampling ratio estimator as

$$t_{tk} = \bar{y}^*\left(\bar{x}_1/\bar{x}^*\right) \tag{4.4}$$

The approximate mean square error t_{tk} given by

$$MSE\left(t_{tk}\right) \approx \left(\frac{1}{n_1} - \frac{1}{N}\right)S_y^2 + \left(\frac{1}{n_2} - \frac{1}{n_1}\right)S_r^2 + \left(\frac{W_2(h-1)}{n_2}\right)S_{2r}^2$$

The expected cost function is given by

$$C_2^* = c_1n_1 + \left(c_2 + c_3W_1 + \frac{c_4W_2}{h}\right)n_2$$

To determine the optimum value h_{otk}

$$\phi = C_2^* + \lambda\left\{MSE\left(t_{tk}\right) - V_o\right\}$$

$$\phi = c_1n_1 + \left(c_2 + c_3W_1 + \frac{c_4W_2}{h}\right)n_2 + \lambda\left\{\left(\frac{1}{n_1} - \frac{1}{N}\right)S_y^2 + \left(\frac{1}{n_2} - \frac{1}{n_1}\right)S_r^2 + \left(\frac{W_2(h-1)}{n_2}\right)S_{2r}^2 - V_o\right\}$$

Where

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

$$S_{r_2}^2 = S_{y_2}^2 + R^2S_{x_2}^2 - 2RS_{xy_2}$$

Where λ is Lagrange's multiplier.

Using Lagrange's multiplier technique the optimum values h, n_2 and n_1 are

$$\phi = c_1n_1 + \left(c_2 + c_3W_1 + \frac{c_4W_2}{h}\right)n_2 + \lambda\left\{\left(\frac{1}{n_1} - \frac{1}{N}\right)S_y^2 + \left(\frac{1}{n_2} - \frac{1}{n_1}\right)S_r^2 + \left(\frac{W_2(h-1)}{n_2}\right)S_{r_2}^2 - V_o\right\}$$

$$h_{oTK} = \sqrt{\frac{c_4\left(S_r^2 - W_2S_{r_2}^2\right)}{S_{r_2}^2\left(c_2 + c_3W_1\right)}},$$

$$n_{2TK} = \frac{\sqrt{S_r^2 + W_2(h-1)S_{r_2}^2}\left[\sqrt{c_1\left(S_y^2 - S_r^2\right)} + \left(\sqrt{c_2 + c_3W_1 + \frac{c_4W_2}{h}}\right)\left(\sqrt{S_r^2 + W_2(h-1)S_{r_2}^2}\right)\right]}{\left(V_o + \frac{S_y^2}{N}\right)\sqrt{c_2 + c_3W_1 + \frac{c_4W_2}{h}}}$$

$$n_{1TK} = \frac{\left[\sqrt{c_1\left(S_y^2 - S_r^2\right)} + \left(\sqrt{c_2 + c_3W_1 + \frac{c_4W_2}{h}}\right)\left(\sqrt{S_r^2 + W_2(h-1)S_{r_2}^2}\right)\right]\sqrt{S_y^2 - S_r^2}}{\left(V_o + \frac{S_y^2}{N}\right)\sqrt{c_1}}$$

2.6 Empirical Comparison of the Estimators

The expected cost C^* for our proposed estimator t_d and expected cost C_1^* for Hansen Hurwitz estimator $\bar{y}^*$ and C_2^* is the expected cost for t_{tk} are compared by using population of Tabasum and Khan (2004) paper. The parameters of the population are

$N_1 = 500$, $N_2 = 150$, $R = 1.48$, $\rho_1 = 0.81$, $S_x^2 = 350.54$, $S_y^2 = 1213.82$, $S_{xy} = 530.07$, $S_{x_2}^2 = 150.04$, $S_{y_2}^2 = 610.67$, $S_{xy_2} = 253.68$, $\beta_1 = 1.69$, $\beta_2 = 1.69$, $\rho_2 = 0.83$, $\bar{X} = 500$

Table 1 Expected cost for fixed variance

W_1	W_2	c_1	c_2	c_3	c_4	For Fixed Variance $V_o = 5.41$		
						Expected cost C^*	Expected cost C_1^*	Expected cost C_2^*
0.7	0.3	0.1	0.5	1	2	91	265	160
		0.2	0.6	1.4	3	135	361	241
		0.3	0.8	1.6	4	177	448	317
		0.4	0.9	1.9	5	219	531	390

3.0 CONCLUSION

It is observe that the expected cost C^* for our proposed estimator t_d is lesser than and expected cost C_1^* for Hansen Hurwitz estimator $\bar{y}^*$ and C_2^* is the expected cost for t_{tk} Tabasum and Khan (2004).

Acknowledgement

The authors wish to thank the anonymous referee for constructive suggestions regarding improvement of this article.

References

[1] Cochran, W. G. 1977. *Sampling Techniques*. 3rd ed.,New York: John Wiley and Sons.

[2] Hansen, M. H., Hurwitz, W. N., 1946. The Problem of Non Response in Sample Surveys. *J. Amer. Statist. Assoc.* 41: 517–529.

[3] Khare, B. B., Srivastava, S. 1993. Estimation of Population Mean Using Auxiliary Character in the Presence of Non Response: *Nat. Acad. Sci. Lett*. (India). 16(3): 111–114.

[4] Khare, B. B., Srivastava, S. 1995. Study of Conventional and Alternative Two Phase Sampling Ratio, Product and Regression Estimators in the Presence of Non Response. *Proc. Nat. Acad. Sci. (India)*. 65(A), II: 195–203.

[5] Rao, P. S. R. S. 1986. Ratio Estimation with Sub Sampling the Non Respondents: *Surv. Methodol*. 12(2): 217–230.

[6] Singh, H. P., Kumar, S. 2008b. A regression Approach to the Estimation of Finite Population Mean in Presence of Non–response. *Aust. N.Z. J. Stat*. 50(4): 395–408.

[7] Tabasum, R., Khan, I. A. 2004. Double sampling for Ratio Estimation with Non Response. *J. Ind. Soc. Agricult. Statist*. 58(3): 300–306.