Aligarh Journal of Statistics
Vol. 31 (2011), 53-62

GENERALIZED ESTIMATORS FOR POPULATION MEAN WITH SUB SAMPLING THE NON RESPONDENTS

R. R. Sinha and Vinod Kumar

ABSTRACT

This paper considers the problem of estimating the finite population mean by making the use of auxiliary information in presence of non-response. Generalized class of estimators for population mean has been proposed with their bias and mean square error. The minimum value of mean square error of proposed class of estimator has been given. Comparisons have been made with the relevant estimators. An illustration has been given to show the results with an application of using original data.

1. INTRODUCTION

In most of the sample surveys, where the covering area of the population is very large, mail questionnaire is one of the simplest and cheapest methods to collect the information on the character under study. The main drawback of this method is the occurrence of non-response on the units selected in the sample. To reduce the effect of non-response in sample surveys, Hansen and Hurwitz (1946) first suggested the method of sub sampling on the non-responding group and suggested an unbiased estimator for estimating the population mean by using the information available from response and non-response group. It is very common to know that the precision in estimating the population parameters may be increased by using the information on auxiliary characters. Using the known and unknown population mean of auxiliary character (s), several authors Rao (1986, 90), Khare and Srivastava (1993, 95, 96, 2000), Khare and Sinha (2002, 2009) and Singh and Kumar (2009) have proposed the estimators for estimating the population mean in presence of non-response.

In this paper, an attempt is made to estimate the finite population mean by using known mean and mean square error of auxiliary information in presence of non-response. A generalized class of estimators for population mean has been proposed with their bias and mean square error.

The minimum value of mean square error of proposed class of estimator has been given. Comparisons have been made with the relevant estimators. An illustration has been given to show the results with an application of using original data.

2. THE SUGGESTED ESTIMATOR

Let y_l and x_l are the non-negative l^{th} values of the main and auxiliary character under study of the finite population consisting of N distinct units. Let the whole population is divided into N_1 responding and N_2 non-responding softcore units such that $N_1 + N_2 = N$. Let a sample of size $'n'$ be drawn from a population of size N by using simple random sampling without replacement method ($SRSWOR$) and it has been observed that n_1 units respond and n_2 units do not respond. To adjust the effect of non-response, Hansen and Hurwitz (1946) have recommended to draw a sub-sample of size $r(= n_2/k), k \geq 1$ from n_2 units and collected the information on them by taking extra efforts. Let $\bar{y}_1$ and $\bar{y}_{2r}$ are the sample mean of the study character $'y'$ based on n_1 and r units respectively. Using the total information $(n_1 + r)$ units, Hansen and Hurwitz (1946) proposed an unbiased estimator for population mean as

$$\bar{y}^* = \frac{n_1}{n}\bar{y}_1 + \frac{n_2}{n}\bar{y}_{2r} \tag{2.1}$$

and its variance is given by

$$M_0 = Var(\bar{y}^*) = \frac{N-n}{Nn}S_y^2 + \frac{W_2(k-1)}{n}S_{y(2r)}^2 . \tag{2.2}$$

In this paper, we have considered unit non-response i.e. responding and non-responding units are same in the population under study, then similarly we can define

$$\bar{x}^* = \frac{n_1}{n}\bar{x}_1 + \frac{n_2}{n}\bar{x}_{2r} \tag{2.3}$$

and $M_1 = Var(\bar{x}^*) = \frac{N-n}{Nn}S_x^2 + \frac{W_2(k-1)}{n}S_{x(2r)}^2$ (2.4)

where $W_i = \frac{N_i}{N}$; $i = 1,2$. Here $(S_y^2, S_{y(2r)}^2)$ and $(S_x^2, S_{x(2r)}^2)$ denote the population mean square error of study character $'y'$ and auxiliary character $'x'$ for the entire and non-responding group of the population. When the population mean $\bar{X}$ of auxiliary character $'x'$ is known then the ratio, product and regression estimators for estimating the population mean $\bar{Y}$ are given by

$T_1 = \frac{\bar{y}^*}{\bar{x}^*}\bar{X}$ Rao (1986) (2.5)

$T_2 = \frac{\bar{y}^*}{\bar{X}}\bar{x}^*$ Khare and Srivastava (1996) (2.6)

$$T_3 = \overline{y}^* + b^*(\overline{X} - \overline{x}^*) \quad \text{Rao (1990)} \tag{2.7}$$

where $b^* = \dfrac{\hat{S}_{yx}}{\hat{S}_x^2}$, the estimates $\hat{S}_{yx}$ and $\hat{S}_x^2$ are obtained on the basis of the available data under the given sampling design.

Further Khare and Srivastava (1996) extended the problem by transforming the variable and purposed transformed estimator for estimating population mean as

$$T_4 = \overline{y}^* \left(\frac{\overline{x}^* + A}{\overline{X} + A} \right). \tag{2.8}$$

Later on Khare and Srivastava (2000) suggested a generalized estimator for estimating population mean, which is given by

$$T_5 = \overline{y}^* \left(\frac{\overline{x}^*}{\overline{X}} \right)^{\alpha}. \tag{2.9}$$

It can be easily observed that the estimators T_1 and T_2 are the particular members of T_5 for $\alpha = -1$ and $+1$ respectively. It is also observed that the minimum value of mean square error of the estimators T_3, T_4 and T_5 are same and is given by

$$MSE\ (T_i) = Var\ (\overline{y}^*) - \frac{[Cov(\overline{y}^*, \overline{x}^*)]^2}{Var\ (\overline{x}^*)}; \quad i = 3,4,5. \tag{2.10}$$

If the population mean $(\overline{X})$ and mean square error (S_x^2) are known and we have incomplete information on study character as well as on auxiliary character then we propose usual ratio and product estimator for estimating the population mean of study character $'y'$ as

$$T_{\gamma} = \overline{y}^* \ \frac{\overline{X}}{\overline{x}^*} \ \frac{S_x^2}{s_x^{*2}} \tag{2.11}$$

and $T_p = \overline{y}^* \ \dfrac{\overline{x}^*}{\overline{X}} \ \dfrac{s_x^{*2}}{S_x^2}$, (2.12)

where $s_x^{*2} = \dfrac{1}{n-1}\left(\sum^{n_1} x_l + k \sum^{r} x_l^2 - n\overline{x}^{*2} \right)$, which is unbiased for S_x^2 [see Okafor and Lee (2000)].

The bias and mean square error (MSE) of the estimators T_γ and T_p upto the first degree approximation is given by

$$Bias(T_\gamma) = \overline{Y}\left[\frac{M_1}{\overline{X}^2} + \frac{M_2}{(S_x^2)^2} - \frac{C_0}{\overline{X}\overline{Y}} - \frac{C_1}{\overline{Y}S_x^2} + \frac{C_2}{\overline{X}S_x^2}\right], \tag{2.13}$$

$$MSE(T_\gamma) = M_0 + \overline{Y}^2\left[\frac{M_1}{\overline{X}^2} + \frac{M_2}{(S_x^2)^2} - 2\left(\frac{C_0}{\overline{X}\overline{Y}} + \frac{C_1}{\overline{Y}S_x^2} - \frac{C_2}{\overline{X}S_x^2}\right)\right], \tag{2.14}$$

$$Bias(T_p) = \overline{Y}\left[\frac{C_0}{\overline{X}\overline{Y}} + \frac{C_1}{\overline{Y}S_x^2} + \frac{C_2}{\overline{X}S_x^2}\right], \tag{2.15}$$

$$MSE(T_p) = M_0 + \overline{Y}^2\left[\frac{M_1}{\overline{X}^2} + \frac{M_2}{(S_x^2)^2} + 2\left(\frac{C_0}{\overline{X}\overline{Y}} + \frac{C_1}{\overline{Y}S_x^2} + \frac{C_2}{\overline{X}S_x^2}\right)\right]. \tag{2.16}$$

Now following the lines of Srivastava (1967) and Khare and Srivastava (2000), we propose a generalized estimator for estimating the population mean of study character 'y' when $\overline{X}$ and S_x^2 are known in advance and incomplete information on both study and auxiliary character as

$$T_g = \overline{y}^*\left(\frac{\overline{x}^*}{\overline{X}}\right)^a\left(\frac{s_x^{*2}}{S_x^2}\right)^b \tag{2.17}$$

where a and b are suitably chosen constants. It is very clear that the estimator T_γ, T_p are the particular member of T_g for $a = b = -1$ and $a = b = 1$ respectively. The estimator T_5 also becomes a member of our suggested estimator T_g for the values of $a = 1$ and $b = 0$.

3. BIAS AND MEAN SQUARE ERROR (MSE)

To derive the expression of bias and mean square error of proposed generalized estimator T_g upto the first degree of approximation, we assume

$$\overline{y}^* = \overline{Y}(1+\varepsilon_0),\ \overline{x}^* = \overline{X}(1+\varepsilon_1),\ s_x^{*2} = S_x^2(1+\varepsilon_2)$$

such that $E(\varepsilon_0) = E(\varepsilon_1) = E(\varepsilon_2) = 0$.

Using the results discussed by Kandall and Struat (1963) and Kadilar and Cingi (2007), the expressions of bias and mean square error of T_g are given by

$$Bias(T_g) = \bar{Y}\left[\frac{a(a-1)}{2}\frac{M_1}{\bar{X}^2} + \frac{b(b-1)}{2}\frac{M_2}{(S_x^2)^2} + a\frac{C_0}{\bar{X}\bar{Y}} + b\frac{C_1}{\bar{Y}S_x^2} + ab\frac{C_2}{\bar{X}S_x^2}\right], \quad (3.1)$$

$$MSE(T_g) = M_0 + \bar{Y}^2\left[a^2\frac{M_1}{\bar{X}^2} + b^2\frac{M_2}{(S_x^2)^2} + 2a\frac{C_0}{\bar{X}\bar{Y}} + 2b\frac{C_1}{\bar{Y}S_x^2} + 2ab\frac{C_2}{\bar{X}S_x^2}\right]. \quad (3.2)$$

The mean Square error of T_g will be minimum when

$$a = -\frac{\bar{X}}{\bar{Y}}\left[\frac{C_0}{M_1} + \frac{C_0C_2^2 - M_1C_1C_2}{M_2M_1^2 - M_1C_2^2}\right] \quad (3.3)$$

and $$b = \frac{S_x^2}{\bar{Y}}\left[\frac{C_0C_2 - M_1C_1}{M_1M_2 - C_2^2}\right]. \quad (3.4)$$

Sometimes the optimum value of 'a' and 'b' depend upon the unknown parameters, so for practical purpose Reddy (1978) suggested that upto the order (n^{-1}), there would not be any effect if we use prior information available from past data on the required parameters of the population.

However if no information is available from the past data then one may estimate the optimum values of 'a' and 'b' on the basis of sample observations available to the investigator without having any loss in the efficiency [see Srivastava and Jhajj (1983)] of the estimator T_g upto the terms of order (n^{-1}).

The minimum mean square error of T_g is given by

$$MSE(T_g)_{\min.} = M_0 - \frac{M_2C_0^2 + M_1C_1^2 - 2C_0C_1C_2}{M_1M_2 - C_2^2} \quad (3.5)$$

where

$$M_2 = \left(\frac{N-n}{Mn}\right)(S_x^2)^2[\beta_2(x) - 1] + \frac{W_2(k-1)}{\bar{n}}(S_{x(2r)}^2)^2[\beta_{2(2r))}(x) - 1],$$

$$C_0 = \left(\frac{N-n}{Nn}\right)\rho S_x S_y + \frac{W_2(k-1)}{n}\rho_{(2r)} S_{x(2r)} S_{y(2r)},$$

$$C_1 = \left(\frac{N-n}{Nn}\right)\mu_{21} + \frac{W_2(k-1)}{n}\mu_{21(2r)}, \quad C_2 = \left(\frac{N-n}{Nn}\right)\mu_{30} + \frac{W_2(k-1)}{n}\mu_{30(2r)},$$

$$\mu_{pq} = \frac{1}{N}\sum^{N}\left(x_l - \bar{X}\right)^p\left(y_l - \bar{Y}\right)^q, \qquad \beta_2(x) = \frac{\mu_{40}}{(\mu_{20})^2},$$

$$\mu_{pq(2r)} = \frac{1}{N_2}\sum^{N_2}\left(x_l - \overline{X}_{(2r)}\right)^p\left(y_l - \overline{Y}_{(2r)}\right)^q, \quad \beta_{2(2r)}(x) = \frac{\mu_{40(2r)}}{(\mu_{20(2r)})^2},$$

$$\overline{X}_{(2r)} = \frac{\sum^{N_2} x_l}{N_2} \quad \text{and} \quad \overline{Y}_{(2r)} = \frac{\sum^{N_2} y_l}{N_2}.$$

It is also to be observed that if we have known coefficient of variation of auxiliary character then we can use it as ratio and product type estimators for estimating the population mean as

$$T_{\gamma c} = \overline{y}^* \frac{C_x^2}{c_x^{*2}} \tag{3.6}$$

and $T_{pc} = \overline{y}^* \dfrac{c_x^{*2}}{C_x^2}$, (3.7)

where $C_x^2 = \dfrac{S_x^2}{\overline{X}^2}$, $c_x^{*2} = \dfrac{s_x^{*2}}{\overline{x}^{*2}}$.

The bias and mean square error of $T_{\gamma c}$ and T_{pc} can be obtained by substituting $a = 2, b = -1$ and $a = -2, b = 1$ in equations (3.1) and (3.2) respectively. Which are as follows:

$$Bias\ (T_{\gamma c}) = \overline{Y}\left[\frac{M_1}{\overline{X}^2} + \frac{M_2}{(S_x^2)^2} + \frac{2C_0}{\overline{X}\overline{Y}} - \frac{C_1}{\overline{Y}S_x^2} - \frac{2C_2}{\overline{X}S_x^2}\right], \tag{3.8}$$

$$MSE\ (T_{\gamma c}) = M_0 + \overline{Y}^2\left[\frac{4M_1}{\overline{X}^2} + \frac{M_2}{(S_x^2)^2} + \frac{4C_0}{\overline{X}\overline{Y}} - \frac{2C_1}{\overline{Y}S_x^2} - \frac{4C_2}{\overline{X}S_x^2}\right], \tag{3.9}$$

$$Bias\ (T_{pc}) = \overline{Y}\left[\frac{3M_1}{\overline{X}^2} - \frac{2C_0}{\overline{X}\overline{Y}} + \frac{C_1}{\overline{Y}S_x^2} - \frac{2C_2}{\overline{X}S_x^2}\right], \tag{3.10}$$

and $$MSE\ (T_{pc}) = M_0 + \overline{Y}^2\left[\frac{4M_1}{\overline{X}^2} + \frac{M_2}{(S_x^2)^2} - \frac{4C_0}{\overline{X}\overline{Y}} + \frac{2C_1}{\overline{Y}S_x^2} - \frac{4C_2}{\overline{X}S_x^2}\right]. \tag{3.11}$$

However, the minimum mean square error of $T_{\gamma c}$ and T_{pc} upto the first degree of approximation can be obtained, as Khare and Srivastava (2000) suggested, by taking $a = -2\alpha$ and $b = \alpha$ in our proposed generalized class of estimators given in (2.17).

4. COMPARISON OF SUGGESTED ESTIMATOR

In this section, firstly we compare MSE of our proposed class of estimators with the variance of the estimator suggested by Hansen and Hurwitz (1946) and we observe from equation (2.2) and (3.5) that

$$Var(\bar{y}^*) - MSE(T_g)_{\min.} = \frac{C_0^2}{M_1} + \frac{[C_0C_2 - M_1C_1]^2}{M_1[M_1M_2 - C_2^2]} \geq 0. \tag{4.1}$$

Secondly we compare our proposed estimator with the estimator proposed by Khare and Srivastava (2000) and we observe from (2.10) and (3.5) that

$$MSE(T_5)_{\min.} - MSE(T_g)_{\min.} = \frac{[C_0C_2 - M_1C_1]^2}{M_1[M_1M_2 - C_2^2]} \geq 0. \tag{4.2}$$

Finally we draw a conclusion from equations (2.2), (2.10) and (3.5) that

$$MSE(T_g)_{\min.} \leq MSE(T_5)_{\min.} \leq \text{var}(\bar{y}^*). \tag{4.3}$$

However, it should also be noted that mean square error of proposed generalized estimator T_g will be less iff $-\{M_0/\bar{Y}^2\} < \xi < 0$,

where $\xi = \left[a^2\frac{M_1}{\bar{X}^2} + b^2\frac{M_2}{(S_x^2)^2} + 2a\frac{C_0}{\bar{X}\bar{Y}} + 2b\frac{C_1}{\bar{Y}S_x^2} + 2ab\frac{C_2}{\bar{X}S_x^2}\right]$.

5. EMPIRICAL STUDY

One hundred nine village population of urban area under Police Station-Baria, Tahsil-Champa, and Orissa has been taken under consideration from District Census Handbook 1981, Orissa, Published by Government of India.

Data Set -In this data set, the 25% villages (i.e 27 villages) from bottom of the list have been considered as non-responding group of the population. Here we have taken Total Population of the village as study character (y) and number of Agriculture Labours in the village as auxiliary character (x). The population parameters under study are given below

$\bar{Y} = 485.92$	$\bar{X} = 41.2385$	$S_y^2 = 101593.31$	$S_x^2 = 2156.0532$
$S_{y(2r)}^2 = 54070.78$	$S_{x(2r)}^2 = 2314.5871$	$\mu_{30} = 198328.5906$	$\mu_{21} = 370373.422$
$\mu_{30(2r)} = 241364.9$	$\mu_{21(2r)} = 822166.9$	$\beta_2(x) = 7.3648$	$\beta_{2(2r)}(x) = 7.8414$
$\rho = 0.451$	$\rho_{(2r)} = 0.714$	$N = 109$	$n = 35$

To study the performance of the proposed generalized estimators T_g with respect to $\bar{y}^*$ and T_5, we calculated the mean square error and relative efficiency (in %) of different estimators for different value of sub sampling fraction $(1/k)$, which are shown in Table 1. We have computed the relative efficiency in percentage by the formula

$$R.E.=\frac{Var\ (\bar{y}^*)}{Var\ (T_{(i)})}\times 100, i=1,2, \text{ where } T_{(1)}=T_5 \text{ and } T_{(2)}=T_g.$$

For the analysis of Data Set, we assume that $10\%, 20\%, 30\%$ and 50% data have been sub-sampled from non-responding group and information has been collected by extra efforts.

Table 1: Mean square error (.) and $R.E.$ (in %) of the estimators

Estimators	$1/k$			
	1/10	1/5	1/3	1/2
$\bar{y}^*$	100.00 (5426.0327)	100.00 (3506.5202)	100.00 (2738.7152)	100.00 (2354.8127)
T_5	165.95 (3269.7396)	150.63 (2327.8612)	140.41 (1950.5418)	133.68 (1761.4853)
T_g	167.42 (3241.0596)	153.50 (2284.3338)	144.79 (1891.5364)	139.51 (1687.9041)

(*figures in parenthesis show the mean square error of the estimators)

6. CONCLUSION

Table 1 exhibits that the mean square error of both the estimators T_5 and T_g are decreasing when we increase the sub sampling fraction $(1/k)$. It is also clear from the Table 1 that the relative efficiency of estimators T_5 and T_g are decreasing as sub sampling fraction increases. This is due to the faster rate of decreasing of variance of $\bar{y}^*$ compare to mean square error of T_5 and T_g. When we compare the performance of T_5 and T_g with respect to $\bar{y}^*$, we observe from the Table 1 that the proposed estimator T_g is the best at all level of sub sampling fractions. Therefore we may recommend to use the proposed generalized estimator T_g for the further use in future under the conditions mentioned in the manuscript.

Acknowledgement

We wish to thank the editor and the referee for their useful comments and suggestions which significantly improved the presentation of the paper.

REFERENCES

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

Kadilar, C. and Cingi, H. (2007): Improvement in variance estimation using auxiliary information. *Hacttepe Jour. Math. Stat.*, **34**, 2075-2081.

Kendall, M. and Stuart, A. (1963): *The advanced theory of statistics: Distribution theory*, **1**, London:Griffin.

Khare, B. B. and Sinha, R. R. (2002): General class of two phase sampling estimators for the population mean using an auxiliary character in the presence of non-response. *Proceedings of the 5th International Symposium on Optimization and Statistics*, 233-245, Dec. 28-30, 2002, A. M. U., Aligarh, India.

Khare, B. B. and Sinha, R. R. (2009): On class of estimators for population mean using multi-auxiliary character in presence of non-response. *Statistics in Transition-new series*, **10**, 3-14.

Khare, B. B. and Srivastava, S. (1993): Estimation of population mean using auxiliary character in presence of nonresponse. *Nat. Acad. Sc. Letters*, **16**, 111-114.

Khare, B. B. and Srivastava, S. (1995): Study of conventional and alternative two phase sampling ratio, product and regression estimators in presence of non-response. *Proc. Nat. Acad. Sci*., India, **65(a)** II, 195-203.

Khare, B. B. and Srivastava, S. (1996): Transformed product type estimators for population mean in presence of softcore observations. *Proc. Math. Soc. BHU*, **12**, 29-34.

Khare, B. B. and Srivastava, S. (2000): Generalised estimators for population mean in presence of nonresponse. *Internal. J. Math. & Statist. Sci*, **9**, 75-87.

Okafor, F. C. and Lee, H.(2000): Double sampling for ratio and regression estimation with sub sampling the non-respondent. *Survey Methodology*, **26**, 183-188.

Rao, P. S. R. S. (1986): Ratio estimation with subsampling the nonrespondents. *Survey Methodology*, **12**, 217-230.

Rao, P. S. R. S. (1990): Regression estimators with subsampling of nonrespondents. *In-Data Quality Control, Theory and Pragmatics, (Eds.) Gunar E. Liepins and V.R.R. Uppuluri*, *Marcel Dekker*, New York, 191-208.

Reddy, V. N. (1978): A study of use of prior Knowledge on certain population parameters in estimation. *Sankhya*, C, **40**, 29-37.

Singh, H. P. and Kumar, S. (2009): A general class of estimators of population mean in survey sampling using auxiliary information with sub sampling the non-respondents. *Korean J. Appl. Stat.*, **22**, 387-402.

Srivastava, S. K. (1967): An estimator using auxiliary information in sample surveys. *Cal. Statist. Assoc. Bull.*, **16**, 121-132.

Srivastava, S. K. and Jhajj, H. S. (1983): A class of estimators of the population mean using multi-auxiliary information. *Cal. Statist. Assoc. Bull.*, **32**, 47-56

Received : 25-04-2011

Received : 26-09-2011

R. R. Sinha and Vinod Kumar

Department of Mathematics,
Dr. B. R. Ambedkar National Institute of
Technology, Jalandhar, India.
email: raghawraman@gmail.com