Available online at www.isas.org.in/jisas
JOURNAL OF THE INDIAN SOCIETY OF
AGRICULTURAL STATISTICS 64(3) 2010 343-347

Domain Estimation in the Presence of Non-Response

U.C. Sud[1*], Hukum Chandra[1] and Raj S. Chhikara[2]
[1]*Indian Agricultural Statistics Research Institute, New Delhi*
[2]*University of Houston-Clear Lake, Houston, Texas, USA*

Received 20 December 2007; Revised 08 July 2009; Accepted 15 January 2010

SUMMARY

The problem of domain estimation in the context of non-response arising out of mail surveys has been considered. Expressions for the unbiased domain estimator, variance of the estimator and unbiased variance estimators are obtained. Optimum values of the sample sizes have been derived under a cost function. The theoretical results are numerically illustrated.

Keywords : Domain estimation, Mails surveys, Non-response.

1. INTRODUCTION

Mail surveys are a useful technique for data collection. They are commonly used in developed countries for data collection purpose. Mail surveys have the advantage that the data can be collected relatively inexpensively. However, non-response is a common problem with such surveys. Non-response can be a serious problem resulting in badly biased estimates. Hansen and Hurwitz (1946) suggested a technique of handling non-response in mail surveys. More recently Fabian and Hyunshik (2000) extended the Hansen and Hurwitz technique to the case where besides the information on character under study, information is also available on auxiliary character. Choudhary *et al*. (2004) used the Hansen and Hurwitz technique in the context of repeat surveys. In this article the theory for use of Hansen and Hurwitz technique for domain estimation has been developed.

2. THEORETICAL FRAMEWORK

Let us consider a population $U = (1, ..., k, ..., N)$ of size N partitioned into D sub-sets $U_1, ... U_d, ..., U_D$ (hereafter we refer them as domains) and let N_d (which is assumed large) be the size of U_d (d = 1, ... , D) such that $U = \bigcup_{d=1}^{D} U_d$ and $N = \sum_{d=1}^{D} N_d$. Let the study variate be denoted by y. Our objective here is to estimate the domain totals of y, $Y_d = \sum_{i=1}^{N_d} y_i$ or the domain means $\bar{Y}_d = N_d^{-1} Y_d$ (d = 1, ..., D). We assume that a sample s of size n is drawn from population according to simple random sampling without replacement (SRSWOR) sampling design and letters/mails containing questionnaires are sent to each unit in the sample. Let s_d denote the part of sample s that happens to fall in U_d, that is, $s_d = s \cap U_d$. Let us denote by n_d the size of s_d such that $s = \bigcup_{d=1}^{D} s_d$ and $n = \sum_{d=1}^{D} n_d$. Note that n_d is random. When the domain sizes are small, n_d may turn out to be very small or it may be equal to '0' in some cases. In such cases small area estimation techniques are needed for reliable estimation at the domain level. However, we do not consider this case here. With the random sample of observations, the

*Corresponding author : U.C. Sud
E-mail address : ucsud@iasri.res.in

statistician's task is to make the best possible estimate for the domain. Let us define by $y_{di} = y_i(i\varepsilon U_d)$ and then the population total of y in domain d can be expressed as $Y_d = \sum_{i=1}^{N_d} y_i = \sum_{i=1}^{N} y_{di}$.

We assume that the population of size N can be divided into two mutually exclusive classes i.e. N_1 and N_2. Here N_1 denotes units of the population, which will respond while N_2 are those units of the population which will not respond at the first attempt. Accordingly, size of the two classes in domain d is denoted by N_{1d} and N_{2d} respectively. Let, out of a sample of n units, n_1 units respond and n_2 units do not respond. Further, n_{1d} and n_{2d} out of n_1 and n_2 units fall in the d-th domain. Let h denote size of sub-sample drawn from the non-response class for collection of information through personal interview, h_d units out of h units fall in the d-th domain. Further, let $n_2 = hf$, obviously $f \geq 1$ in this case. Let $\bar{y}_{1n_d}$ denote the mean of the sample from the response class for the d-th domain while $\bar{y}_{2h_d}$ denote the mean of the sample for the non-response class, where $\bar{y}_{1n_d} = n_1^{-1}\sum_{i=1}^{n_1} y_{1di}$; $\bar{y}_{2h_d} = h^{-1}\sum_{i=1}^{h} y_{2di}$, $y_{2di} = y_{2i}\ (i\varepsilon U_d)$.

With these notations, we propose the estimator for population total of y in domain d (d = 1, ..., D) as

$$\hat{Y}_d = N\frac{n_1\bar{y}_{1n_d} + n_2\bar{y}_{2h_d}}{n} = N\bar{y}_{n_d} \tag{2.1}$$

where $$\bar{y}_{n_d} = \frac{n_1\bar{y}_{1n_d} + n_2\bar{y}_{2h_d}}{n}$$

Theorem. The estimator $\hat{Y}_d$ is unbiased estimator of Y_d with variance

$$V(\hat{Y}_d) = N\frac{(N-n)}{n}\left\{P_d S_d^2 + P_d Q_d \bar{Y}_d^2\right\}$$
$$+(f-1)\frac{NN_2}{n}\left\{P_{2d}S_{2d}^2 + P_{2d}Q_{2d}\bar{Y}_{2d}^2\right\}$$

with $S_d^2 = (N_d - 1)^{-1}\sum_{i=1}^{N_d}(y_i - \bar{Y}_d)^2$

$$S_{2d}^2 = (N_{2d} - 1)^{-1}\sum_{i=1}^{N_{2d}}(y_{2i} - \bar{Y}_{2d})^2$$

$$\bar{Y}_d = N_d^{-1}\sum_{i=1}^{N_d} y_i$$

$$\bar{Y}_{2d} = N_{2d}^{-1}\sum_{i=1}^{N_{2d}} y_i$$

$$P_d = \frac{N_d}{N}\ ;\ P_{2d} = \frac{N_{2d}}{N_2}$$

$Q_d = 1 - P_d$; $Q_{2d} = 1 - P_{2d}$

Then an unbiased variance estimator is given by

$$v(\hat{Y}_d) = N\frac{(N-n)}{(n-1)}\left[\bar{G}_{wd} - \bar{y}_{n_d}^2\right]$$
$$+ N\frac{(N-1)}{(n-1)}(f-1)\frac{n_2}{n}s_{2d}^2 \tag{2.2}$$

where $$\bar{G}_{wd} = \frac{1}{n}\sum_{i=1}^{n_1} y_{1di}^2 + \frac{n_2}{nh}\sum_{i=1}^{h} y_{2di}$$

$$s_{2d}^2 = (h-1)^{-1}\sum_{i=1}^{h}(y_{2di} - \bar{y}_{2h_d})^2$$

Proof. The unbiasedness of $\hat{Y}_d$ can be shown as follows

$$E_1\left[E_2\left\{E_3(\hat{Y}_d)\right\}\right] = NE_1\left\{E_2\left(\frac{n_1\bar{y}_{1n_d} + n_2\bar{y}_{2h_d}}{n}\right)\right\}$$
$$= NE_1\left\{\frac{n_1N_1^{-1}Y_{1d} + n_2N_2^{-1}Y_{2d}}{n}\right\}$$
$$= Y_d$$

where $Y_{1d} = \sum_{i=1}^{N_1} y_{1di}$; $Y_{2d} = \sum_{i=1}^{N_2} y_{2di}$. Here E_1 refers to expectation over all possible samples of size 'n' drawn from a population of size 'N', E_2 is the conditional expectation when samples of size n_1 and n_2 are drawn from a population of size N_1 and N_2 and E_3 refers to expectation over all possible samples of size h drawn from a population of size n_2.

The variance of $\hat{Y}_d = N\bar{y}_{nd}$ is obtained as

$$V(\hat{Y}_d) = V\left(N\bar{y}_{n_d}\right) = E_1E_2\left[V_3\left(N\bar{y}_{n_d}\right)\right]$$
$$+E_1V_2\left[E_3\left(N\bar{y}_{n_d}\right)\right] + V_1\left[E_2E_3\left(N\bar{y}_{n_d}\right)\right]$$

where V_1, V_2, V_3 can be defined on the similar lines as E_1, E_2, E_3. Here,

$$E_1V_2\left[E_3\left(N\bar{y}_{n_d}\right)\right]+V_1\left[E_2E_3\left(N\bar{y}_{n_d}\right)\right]$$

simplify to $N\frac{(N-n)}{n}\left[P_dS_d^2+P_dQ_d\bar{Y}_d^2\right]$ while

$E_1E_2\left[V_3\left(N\bar{y}_{n_d}\right)\right]$ simplify to

$$N\frac{N_2}{n}(f-1)\left[\frac{(N_{2d}-1)}{(N_2-1)}S_{2d}^2+P_{2d}Q_{2d}\bar{Y}_{2d}^2\right]$$

These expressions lead to

$$V(\hat{Y}_d) = N\frac{(N-n)}{n}\left\{P_dS_d^2+P_dQ_d\bar{Y}_d^2\right\} + (f-1)\frac{NN_2}{n}\left\{P_{2d}S_{2d}^2+P_{2d}Q_{2d}\bar{Y}_{2d}^2\right\} \tag{2.3}$$

We now examine unbiasedness of the variance estimator (2.2). That is

$$E_1E_2E_3\left[v(\hat{Y}_d)\right] = E_1E_2E_3\left\{N\frac{(N-n)}{(n-1)}\left[\bar{G}_{wd}-\bar{y}_{n_d}^2\right] + N\frac{(N-1)}{(n-1)}(f-1)\frac{n_2}{n}s_{2d}^2\right\}$$

For this we evaluate following terms

$$E_1E_2E_3(\bar{G}_{wd}) = \frac{1}{N}\sum_{i=1}^{N}y_{di}^2$$

$$E_1E_2E_3\left(\bar{y}_{n_d}^2\right) = \bar{Y}^2+\frac{(N-n)(N_1-1)}{Nn(N-1)}S_1^2 + \frac{(N-n)N_1N_2}{N^2n(N-1)}\left(\bar{Y}_1-\bar{Y}_2\right)^2+\frac{(N-n)(N_2-1)}{Nn(N-1)}S_2^2 + \frac{(f-1)N_2}{nN}S_2^2$$

where $\bar{Y} = \frac{1}{N}\sum_{i=1}^{N}y_{di} = P_d\bar{Y}_d$

$$\bar{Y}_1 = \frac{1}{N_1}\sum_{i=1}^{N_1}y_{di}$$

$$\bar{Y}_2 = \frac{1}{N_2}\sum_{i=1}^{N_2}y_{di}$$

$$S_1^2 = (N_1-1)^{-1}\sum_{i=1}^{N_1}\left(y_{di}-\bar{Y}_1\right)^2$$

$$S_2^2 = (N_2-1)^{-1}\sum_{i=1}^{N_2}\left(y_{di}-\bar{Y}_2\right)^2$$. Also,

$$E_1E_2E_3\,\frac{(f-1)n_2}{n}s_{2d}^2 = \frac{(f-1)N_2}{N}S_2^2$$

By combining the common terms and simplifying we get

$$E_1E_2E_3\left\{N\frac{(N-n)}{(n-1)}\left[\bar{G}_{wd}-\bar{y}_{n_d}^2\right] + N\frac{(N-1)}{(n-1)}(f-1)\frac{n_2}{n}s_{2d}^2\right\}$$ equal to

$$N\frac{(N-n)}{n}\left\{P_dS_d^2+P_dQ_d\bar{Y}_d^2\right\} + (f-1)\frac{NN_2}{n}\left\{P_{2d}S_{2d}^2+P_{2d}Q_{2d}\bar{Y}_{2d}^2\right\}$$

Hence the proof.

3. OPTIMIZATION UNDER A COST FUNCTION

To determine the optimum values of n, say $n_{(d)}$ and f, say $f_{(d)}$, for the given variance we consider the following cost function

$$C_d = c_{0d}n_d + c_{1d}n_{1d} + c_{2d}h_d \tag{3.1}$$

where for the d-th domain, c_{0d} represents the cost per unit of mailing a questionnaire, c_{1d} the per unit cost of processing information in the response class and c_{2d} the cost of interviewing and processing information per unit in the non-response class. The expected cost is given by

$$\frac{n}{N}\left[N_dc_{0d}+N_{1d}c_{1d}+\frac{N_{2d}}{f}c_{2d}\right] \tag{3.2}$$

We minimize the expected cost by fixing the variance as $\phi = C_d + \mu(V_d - V_{0d})$, μ is the Lagrangian multiplier. Here V_{0d} can be determined by fixing the coefficient of variation, say equal to 5%. This gives optimum values as

$$n_{(d)\mathrm{opt}} = \frac{N^2 L_d + \left(f_{opt} - 1\right) L_{2d} N N_2}{0.0025 * Y_d^2 + N L_d} \qquad (3.3)$$

$$f_{(d)\mathrm{opt}} = \sqrt{\frac{M_{1d}\left(N^2 L_d - N_2 L_{2d}\right)}{M_d L_{2d} N_2}} \qquad (3.4)$$

where $L_d = P_d S_d^2 + P_d Q_d \bar{Y}_d^2$

$L_{2d} = P_{2d} S_{2d}^2 + P_{2d} Q_{2d} \bar{Y}_{2d}^2$

$M_d = N_d c_{0d} + c_{1d} N_{1d}$

$M_{1d} = N_{2d} c_{2d}$

When there is no non-response then the variance is

$$V(\hat{Y}_{dc}) = N \frac{(N-n)}{n} L_d \qquad (3.5)$$

where $\hat{Y}_{dc} = N \frac{n_1 \bar{y}_{1n_d} + n_2 \bar{y}_{2n_d}}{n}$. The cost function in this case is given by

$$C_d = c_{2d} n_d \qquad (3.6)$$

while the expected cost is given by $n c_{2d} P_d$. (3.7)

The optimum value of n, say $n_{11\mathrm{dopt}}$, in this case is obtained, as earlier, by fixing the variance, say coefficient of variation equal to 5%, and minimizing the expected cost. Thus, we get

$$n_{11\mathrm{dopt}} = \frac{N^2 L_d}{\left[0.0025 \times Y_d^2 + N L_d\right]} \qquad (3.8)$$

4. NUMERICAL RESULTS AND CONCLUDING REMARKS

The following illustration will give an idea about saving in cost through mail surveys in the context of domain estimation. We assume following values f = 1.5, 2.5; N_d = 100, 150; $N_{1d} = 0.5N_d$, $0.3N_d$; $\frac{S_d^2}{\bar{Y}_d^2}$ = 5,10,15; $\frac{S_{2d}^2}{\bar{Y}_{2d}^2}$ = 2.5,5,10; N_2 = 200; N = 300, and n = 50. Further, we considered the cost function as given in (3.2) and chosen c_{0d} =1, c_{1d} = 4 and c_{2d} = 40 (all values are in rupees). Substituting these values in (3.5) we get the expected cost. For $N_d N^{-1}$ = 1/3, it is equal to Rs 667 while for $N_d N^{-1}$ = 1/2 it comes out to be Rs 1000. Now equating $V(\hat{Y}_d)$ equal to $V(\hat{Y}_{dc})$ and using the assumed values of different parameters, values of sample sizes and the corresponding expected cost of survey were determined in respect of $\hat{Y}_d$. Let the sample size so obtained be denoted by n'. These results are set out in Table 1. A close perusal of Table 1 reveal that there is considerable reduction in cost by using an estimator based on the Hansen and Hurwitz (1946) technique over an estimator based on only interview

Table 1. Sample sizes and the corresponding expected cost of survey which give equal precision of $\hat{Y}_d$ over $\hat{Y}_{dc}$.

f	P_d	P_{2d}	$S_d^2/\bar{Y}_d^2$	$S_{2d}^2/\bar{Y}_{2d}^2$	n'	Expected Cost
1.5	0.33	0.25	5	2.5	90	489.8
1.5	0.33	0.25	10	5	94	509.5
1.5	0.33	0.25	20	10	96	521.0
1.5	0.50	0.375	5	2.5	92	499.8
1.5	0.50	0.375	10	5	95	515.4
1.5	0.50	0.375	20	10	96	524.3
2.5	0.33	0.25	5	2.5	102	374.0
2.5	0.33	0.25	10	5	105	386.4
2.5	0.33	0.25	20	10	107	393.7
2.5	0.50	0.375	5	2.5	104	381.3
2.5	0.50	0.375	10	5	107	390.7
2.5	0.50	0.375	20	10	108	396.1
1.5	0.33	0.35	5	2.5	92	501.5
1.5	0.33	0.35	10	5	96	521.6
1.5	0.33	0.35	20	10	98	533.3
1.5	0.50	0.525	5	2.5	94	510.8
1.5	0.50	0.525	10	5	97	527.2
1.5	0.50	0.525	20	10	99	536.4
2.5	0.33	0.35	5	2.5	108	397.7
2.5	0.33	0.35	10	5	112	410.9
2.5	0.33	0.35	20	10	114	418.6
2.5	0.50	0.525	5	2.5	110	403.5
2.5	0.50	0.525	10	5	113	414.4
2.5	0.50	0.525	20	10	115	420.6

method. The saving in cost increases with increase in f value. Also, the reduction in cost decreases with increase in both $S_d^2 / \bar{Y}_d^2$ and $S_{2d}^2 / \bar{Y}_{2d}^2$. The reduction in cost increases with increase in domain sizes.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the constructive comments from the referee and the editor.

REFERENCES

Choudhary, R.K., Bathla, H.V.L. and Sud, U.C. (2004). On non-response in sampling on two occasions. *J. Ind. Soc. Agril. Statist.*, **58(3)**, 331-343.

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

Fabian, C.O. and Hyunshik, L. (2000). Double sampling for ratio and regression estimation with sub-sampling the non-respondents. *Survey Methodology,* **26(2)**, 183-188.

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

Singh, R. and Mangat, N.P.S. (1996). *Elements of Survey Sampling*. Kluwer Academic publishers.