Compromise Allocation for Mean Estimation in Stratified Random Sampling Using Auxiliary Attributes When Some Observations are Missing

Sidra Naz
Yousaf Shad Muhammad
Javid Shabbir
Department of Statistics
Quaid-I-Azam University
Islamabad

Abstract

This paper considers the estimation of ratio of population mean when some observations on study variable and auxiliary variables are missing in the case of stratified random sampling. Four estimator are presented and their bias and mean square error are formulated. Here the problem of stratified random sampling in the case of missing observations for nonlinear random cost with certain probability has been formulated .The formulated problem minimize the coefficient of variation and determines the best compromise allocation.

1. Introduction

Ratio of two population mean is conventionally estimated by the ratio of corresponding sample mean, if there some observations are missing than this estimation procedure does not work. The observations are unavailable because of various reasons such as unwillingness of some selected unit to provide the designed information due to unknown factors. In this paper we studied a situation in which sampling is done by stratified random sampling and there are some observations missing on one characteristics at a time. Let us have a data set where (Y_h) is an auxiliary variable in the data and (X_h) are the corresponding auxiliary variables. Then there may be some study variables for X_h is missing or could not be recorded earlier, while the corresponding values of Y_h are available. Similarly this condition holds for Y_h where some values of Y_h are unavailable but there corresponding values of Xh are given.

Toutenburg and Srivastava (1998) consider the estimation of the ratio of population means when some observations are missing. Four estimators were presented and their bias and mean square error properties were studied. HHorng-Jinh Chang Kuo-Chung Huang (2001) proposed several estimators for ratio of population means the presence of missing of some observations. They also make a comparison between different properties of relative mean square errors to study the efficiency for comparison of superiority among them. Further, some distribution of random incompleteness is also considered. Kadilar and Cingi (2003) considered some ratio type estimators and studied their properties in stratified random sampling. Ahmeda Omarand Al-Titib (2005) proposes some general estimators for finite population variance in presence of random non-response using an auxiliary variable. They considered all possible cases of non-response, studied properties of those proposed estimators and compared their performance. We have seen that Toutenburg and Srivastava (1998) considered various ratio type estimators of population mean, under different situations, when some of the observations on either study variable or auxiliary variable or both of the observations are unavailable. In this paper we have assumed the same situation of missing values. We have proposed an estimator which is based on all observations either available or missing when sampling is done by stratified random sampling.

2. Formulation of Problem

Consider a population of N units partitioned into L disjoint groups called strata's with $N_h > 0$ in the h^{th} stratum. h = 1,2, 3, …,L. An independent sample of size n_h is selected by simple random sampling without replacement from each stratum N $=\sum_{h=1}^{L} Nh$. We draw a sample of size n_h from each stratum by SRSWOR such that$\sum_{h=1}^{L} nh = n$. Let $\bar{Y}$and $W_h=N_h/N$ be the population mean, population variance and known stratum weight of hth stratum respectively.

It is assumed that a set of n_h - p_h -q_h complete observations (x_{1h}, y_{1h}),(x_{2h}, y_{2h}),…, ($x_{nh-ph-qh}$,y_{nh}-p_h-q_h) on selected units in sample arecompletely available. Further there are x_1^*,x_2*,…,x_p^*on p_h units in the sample are available but there corresponding Y_h are lost. Similarly a set of q_h observations on Y_h observations are completely available but there corresponding values of auxiliary variables are unavailable. Later on, the quantities p_h and q_h denote the number of incomplete values selected bySRSWOR. In practice, to increase the precision of an estimate, we have to increase samplesize. This action will certainly increase the cost of the survey. If we apply an upper bound on precision or variance (or mean square error, MSE), we can select an optimal sample size by minimizing the survey cost or vice versa. Let us introduce the following means for samples,

$$\bar{y}_{st} = \sum_{h=1}^{L} W_h \bar{Y}_h$$

$$\bar{x}_{st} = \sum_{h=1}^{L} W_h \bar{X}_h$$

$$\bar{x}_{st}{}^* = \sum_{h=1}^{L} W_h \bar{X}_h{}^*$$

$$\bar{y}_{st}{}^* = \sum_{h=1}^{L} W_h \bar{Y}_h{}^*$$

Where$\bar{y}_h{}^*$ and $\bar{x}_h{}^*$ shows the mean of unavailable observations which can be formulated as;

$$\bar{y}_h{}^* = \frac{(n_h - p_h - q_h)\bar{y}_h + q\bar{y}_{h'}{}^*}{n_h - p_h}$$

$$\bar{x}_h{}^* = \frac{(n_h - p_h - q_h)\bar{x}_h + p\bar{x}_{h'}{}^*}{n_h - q_h}$$

The following estimator for the ratio can be formulated as;

$$\bar{y}_{r1} = \frac{\bar{y}_{st}}{\bar{x}_{st}} \bar{X}$$

$$\bar{y}_{r2} = \frac{\bar{y}_{st}{}^*}{\bar{x}_{st}} \bar{X}$$

$$\bar{y}_{r3} = \frac{\bar{y}_{st}}{\bar{x}_{st}{}^*} \bar{X}$$

$$\bar{y}_{r4} = \frac{\bar{y}_{st}{}^*}{\bar{x}_{st}{}^*} \bar{X}$$

The estimator $\bar{y}_{r1}$is based on the complete observations and it is not considering all thepairs of incomplete observations. $\bar{y}_{r2}$and$\bar{y}_{r4}$using incomplete observations only partly.The estimator $\bar{y}_{r4}$considers all the missing or available observations. It is basically the modified term of estimator proposed by Totenberg and Srivastava (1998) when samplings done by SRSWOR.Let us make a comparison between the estimators with respect to the criterion of mean square error. For this purpose, we have the following results derived.

$$MSE(\bar{y}_{r1}) = \sum_{h=1}^{L} W^2{}_h \left(S_{yh}{}^2 + R^2 S_{xh}{}^2 - 2RS_{xyh}\right) f_{p_h+q_h}$$

$$MSE(\bar{y}_{r2}) = \sum_{h=1}^{L} W^2{}_h \left(S_{yh}{}^2 f_{ph} + R^2 S_{xh}{}^2 f_{p_h+q_h} - 2RS_{xyh} f_{ph}\right)$$

$$MSE(\bar{y}_{r3}) = \sum_{h=1}^{L} W^2{}_h \left(S_{yh}{}^2 f_{p_h+q_h} + R^2 S_{xh}{}^2 f_{qh} - 2RS_{xyh} f_{qh}\right)$$

$$MSE(\bar{y}_{r4}) = \sum_{h=1}^{L} W^2{}_h \left(S_{yh}{}^2 f_{qh} + R^2 S_{xh}{}^2 f_{ph} - 2RS_{xyh} f_{ph}\right)$$

where we have, $f_{ph} = E\left(\frac{1}{n_h - p_h}\right) - \frac{1}{N}$, $f_{qh} = E\left(\frac{1}{n_h - qh}\right) - \frac{1}{N}$ and $f_{ph+qh} = E\left(\frac{1}{n_h - p_h - q_h}\right) - \frac{1}{N}$.

3. Allocation Using Lexicographic Goal Programming under Simple and Quadratic cost Functions

In many situations, however, a decision maker may rank his or her goals from the most important (goal 1) to least important (goal m). This is called preemptive goal programming or Lexicographic goal programming and its procedure starts by concentrating on meeting the most important goal as closely as possible, before proceeding to the next higher goal, and so on to the least goal i.e. the objective functions are prioritized such that attainment of first goal is far more important than attainment of second goal which is far more important than attainment of third goal, etc., such that lower order goals are only achieved as long as they do not degrade the solution attained by higher priority goal.

When this is the case, lexicographic goal programming may prove to be a useful tool. Number of unwanted deviations is minimized at each priority level. A goal is set such the increase in variance due to compromise allocation does not exceed the certain quantity called goal variable. The goal variable in the following nonlinear programming formulation is defined as d_j.

3.1 Simple Cost Function

The simple form of cost function is most appropriate to use when a main part of the costis about measurements of all units involved. According to Mandal et al.(2008) the simplecost function for strata can be described as follows;

$$C = c_0 + \sum_{h=1}^{L} c_h n_h$$

Where;

C=Total budget available for sample survey.
c_0=Expected over head coast.
c_h=Measurement cost per unit in h^{th}stratum.
The cost function may be written as:

$$C_0 = C - c_0 = \sum_{h=1}^{L} c_h n_h$$

3.2 Quadratic Cost Function

Let C be the upper limit on the total cost of survey. The problem of optimal sample allocation involves determining the sample size that minimizes the variances under a specific budget C. In each stratum the linear cost function is appropriate when the major item of cost is that of taking the measurement on each unit. Including travel cost between units in a stratum is substantial and mathematical studies indicate that the costs are better represented by expression n_h where t_h is the travel cost. Assuming this nonlinear cost function we have

$$C = c_0 + \sum_{h=1}^{L} c_h + \sum_{h=1}^{L} t_h \sqrt{n_h}$$

The MONLPP of the problem can be written as

Minimize Zj = (CV_j^2)
Subjectto= $\sum_{h=1}^{L} c_h + \sum_{h=1}^{L} t_h \sqrt{n_h} \leq c_0$
And=2$\leq n_h - p_h - q_h \leq N_h$

The application of these models on our proposed study is illustrated by an example solved by using GAMS and R-3.0.1-win.exe. While solving a numerical example we have taken $p_h$and q_h as fixed quantities according to Gorver (2014), so we must have$f_{ph} = E\left(\frac{1}{n_h - p_h}\right) - \frac{1}{N}$, $f_{qh} = E\left(\frac{1}{n_h - qh}\right) - \frac{1}{N}andf_{ph+qh} = E\left(\frac{1}{n_h - p_h - q_h}\right) - \frac{1}{N}$.

4. Numerical Examples

6.1 Data 1 [source; www.agcensus.usda.gov]

Y_1 ; The Quantity of Corn harvested in 2010.

Y_2 ; The Quantity of Soya been harvested in 2010.

X_1 ; The Quantity of Corn harvested in 2009.

X_2 ; The Quantity of Soya been harvested in 2009.

Here,

Here , $\bar{Y}_1$ = 24475.16 and $\bar{Y}_2$ = 5012.424

It is assumed that total cost of survey is $C_0 = 50$

Table 1: Summery statistics for data 1

h	P_{h1}	Q_{h1}	c_{h1}	t_{h1}	$S_{y1h}{}^2$	$S_{x1h}{}^2$	S_{xy1h}	$\bar{Y}_{1h}$	$\bar{X}_{1h}$	$W_{h1}{}^2$	R_1h^2
1	4	5	1	3	46559462	39633300	25326105	10850.0	13855.9	0.091827	0.613183
2	7	4	1	4	58037221	42500981	35470570	20810.91	22933.97	0.111111	0.823424
3	8	6	1	6	45773079	51015409	42218105	31721.94	34737.31	0.132231	0.833925

The data statistics of variable Y_2 is given in the table below ;

Table 2: Summery statistics for data

h	P_{h2}	Q_{h2}	c_{h2}	t_{h2}	$S_{y2h}{}^2$	$S_{x2h}{}^2$	S_{xy2h}	$\bar{Y}_{2h}$	$\bar{X}_{2h}$	$W_{h2}{}^2$	R_2h^2
1	6	4	1	2	1456712	1779537	1561611	3244.1	3332.4	0.091827	0.947707
2	5	5	1	7	2484784	2320859	2088956	4721.030	4571.879	0.111111	1.066311
3	4	7	1	6	3627039.7	3509963.6	644967.7	6753.139	6561.167	0.132231	1.059374

Table 1: Summery statistics for data 1

h	P_{h1}	Q_{h1}	c_{h1}	t_{h1}	$S_{y1h}{}^2$	$S_{x1h}{}^2$	S_{xy1h}	$\bar{Y}_{1h}$	$\bar{X}_{1h}$	$W_{h1}{}^2$	R_1h^2
1	4	5	1	3	46559462	39633300	25326105	10850.0	13855.9	0.091827	0.613183
2	7	4	1	4	58037221	42500981	35470570	20810.91	22933.97	0.111111	0.823424
3	8	6	1	6	45773079	51015409	42218105	31721.94	34737.31	0.132231	0.833925

The data statistics of variable Y_2 is given in the table below ;

Table 2: Summery statistics for data

h	P_{h2}	Q_{h2}	c_{h2}	t_{h2}	$S_{y2h}{}^2$	$S_{x2h}{}^2$	S_{xy2h}	$\bar{Y}_{2h}$	$\bar{X}_{2h}$	$W_{h2}{}^2$	R_2h^2
1	6	4	1	2	1456712	1779537	1561611	3244.1	3332.4	0.091827	0.947707
2	5	5	1	7	2484784	2320859	2088956	4721.030	4571.879	0.111111	1.066311
3	4	7	1	6	3627039.7	3509963.6	644967.7	6753.139	6561.167	0.132231	1.059374

4.1Results

Table 3: Minimized CVs and Optimum Allocation of example for all four Estimators

Minimized CVs by using Simple Cost Function

Estimators	n_1	n_2	n_3	$CV1$	$CV2$
$\bar{y_r}1$	2	2	2	2764500	2764500
$\bar{y_r}2$	9	4	5	31714	33469
$\bar{y_r}3$	4	4	4	578990	647590
$\bar{y_r}4$	5	5	5	821040	821040

Minimized CVs by using Quadratic Cost function

Estimators	n_1	n_2	n_3	$CV1$	$CV2$
$\bar{y_{r1}}$	2	2	2	2764500	2764500
$\bar{y_{r2}}$	5	7	6	523010	523010
$\bar{y_{r3}}$	9	9	12	446450	296456
$\bar{y_{r4}}$	5	14	7	452670	296450

Table (6.1) shows the results of coefficient of variation of all estimators under both simpleand quadratic cost function. Optimum allocation according to Z2 provide smaller CV,s thanZ_1. The values of CV's are larger because of variation in the data set .The CV comparisonshows in Data that the estimator considering missing observations have smaller CV than simple stratified estimator$\bar{y}_{r1}$. Under Quadratic cost function we obtain smaller values ofCVs in comparison of simple cost function. We have represented the optimum allocation forsample sizes for all the estimators using auxiliary variables in stratified random sampling incase of partial missing and complete missing observations on both Y and X.As the samplesize increases for each stratum the value of coefficient of variation decreases.

4.2 Discussion

We have considered the problem of estimating the ratio of population means when observations on some selected units in the sample drawn according to stratified random samplingand each unit is selected from the strata by SRSWOR on either Xh characteristic or Yh characteristic but not are missing at the same time. Accordingly, we have simple estimatorfor the population ratio.The estimator is based on all the complete as well as incomplete pairs of observations. Properties of estimator are analyzed with respect to thebias and mean squared error criteria using the large sample approximation. The problemis represented as multi-objective integer nonlinear programming in which the objective is to minimize the coefficient of variation under simple cost function and quadratic cost function.Also we have compared the results obtained by both cost function and we have noticed what difference has produced by the changing of cost function .We have used thesecondconstraint and restricted constraint to avoid over sampling because we need an integersample size for practical purpose.

4.3Future Studies

This study may further be extended to

- The proposed compromise allocation can be addressed when there are two or more than two auxiliary variables.
- Estimators and proposed allocation can be developed for multivariate stratified sampling , double sampling and two phase sampling.
- This procedure can be used under probabilistic cost function, general cost function and polynomial cost function.

References

Ahmeda., and Abu-Dayyehb, W.(2005). Estimation of finite population variance in presence of random non-response using auxiliary variables. International journal of information and management sciences,16(2):73.s

Cochran, W. G. (1965). Sampling Techniques: 2d Ed. J. Wiley.

Diaz-Garcia, J. A. and Cortez, L. U. (2006). Optimum allocation in multivariate stratified sampling: multi-objective programming. ComunicacionTecnica No. I-06-07/28-03-206 Evans, G. W. (1984). An overview of techniques for solving multiobjectivemathematical programs. Management Science, 30(11):1268-1282.

Ghufran, S., Khowaja, S., and Ahsan, M. (2012). Optimum multivariate stratified sampling designs with travel cost: a multiobjective integer nonlinear programming approach. Communications in Statistics-Simulation and Computation, 41(5):598-610.

Grover, L. K. and Kaur, P. (2014). Exponential ratio type estimators of population mean under non-response. Open Journal of Statistics, 4(01):97.

Iftekhar, S., Haseen, S., Ali, Q. M., and Bari, A. A compromise solutionin multivariate surveys with stochastic random cost function.

Kadilar, C. and Cingi, H. (2003).Ratio estimators in stratified randomsampling. Biometrical Journal, 45(2):218-225.

Khowaja, S., Ghufran, S., and Ahsan, M. (2011). Estimation of population means in multivariate stratified random sampling. Communications in Statistics Simulation and Computation R, 40(5):710-718.

Kokan, A. and Khan, S. (1967). Optimum allocation in multivariate surveys: An analytical solution. Journal of the Royal Statistical Society. SeriesB (Methodological), pages 115-125.

Mahalanobis, P. C. (1944). On large-scale sample surveys. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, pages329-451.

Neyman, J. (1934). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. Journal of the Royal Statistical Society, pages 558-625

Orumie, U. and Ebong, D. (2013).An efficient method of solving lexicographic linear goal programming problem. International Journal of Scientific and Research Publications, 3:1-8.

Raghav, Y. S., Ali, I., and Bari, A. (2013). A compromise allocation in multivariate stratified sampling in presence of non-response. Journal of NonlinearAnalysis and Optimization: Theory & Applications, 5(1):67-80.

SAHIN, S. T. (2011). Determination of sample size selecting from strata under nonlinear cost constraint by using goal programming and kuhn-tucker methods. Gazi University Journal of Science, 24(2):249-262.

Singh, H. P., Tailor, R., and Tailor, R. (2012).Estimation of finite population mean in two-phase sampling with known coefficient of variation of an auxiliary character. Statistical, 72(1):111-126.

Steuer, R. E. (1989). Multiple criteria optimization: theory, computation, and application. Krieger Malabar.

Stuart, A. (1954). A simple presentation of optimum sampling results.Journal of the Royal Statistical Society. Series B (Methodological), pages 239-241.

Toutenburg, H. and Srivastava, V. (1998). Estimationof ratio of population means in survey sampling when some observations are missing. Metrika, 48(3):177-187.

Varshney, R., Ahsan, M., and Khan, M. G. (2011). An optimum multivariate stratified sampling design with non-response: A lexicographic goal programming approach. Journal of Mathematical Modeling and Algorithms, 10(4):393-405.