ON ESTIMATION IN THREE AND FOUR PHASE SAMPLING

Saman Hanif[1], Nadeem Shafique[2] Butt and Muhammad Qaiser Shahbaz[1]
[1]Department of Statistics, Faculty of Sciences, King Abdul Aziz University, Jeddah, Saudi Arabia
[2]Department of Community Medicine, Faculty of Medicine
Medical College in Rabigh, King Abdul Aziz University, Jeddah, Saudi Arabia
saman.hanif@hotmail.com, nadeemshafique@gmail.com qshahbaz@gmail.com

Multiphase sampling has been widely used in literature when information about sampling frame is not available. In this short note it is shown that the estimation in multiphase sampling can only be based upon two phases. The result has been proved for three and four phases only.

Key Words: *Auxiliary Information, Three and Four Phase Sampling.*

1. INTRODUCTION

Often it happened that the information about sampling frame in a finite population setup is not available and the conventional sampling methods are no more useful. The multiphase sampling is a popular rescue source in these sorts of situations. To set up the seen suppose that it is known in advance that a finite population has *N* differentiable units but the complete listing of these population units is not available and hence the popular designs like simple random sampling, stratified random sampling and others are not applicable. The multiphase sampling design is perhaps the most powerful design that can be efficiently used in situations where sampling frame is not available.

The notations of multiphase sampling are bit different from the conventional single phase sampling in that sample of different sizes are selected in phases and at each phase information is collected about some auxiliary variables that are thought to increase the efficiency of estimation of variable of interest. The information from main variable of interest is collected at the last phase only. We now provide the main theme of multiphase sampling by describing some of the common notations. Let $n_j;\ j=1,2,...,p$; be the size of sample drawn from a population of size *N* at *j*–th phase, $\bar{x}_j;\ j=1,2,...,p$; be the mean of an auxiliary variable based upon the sample at *j*–th phase; that is $\bar{x}_j = n_j^{-1}\sum_{i=1}^{n_j} x_{ij}$ and $\bar{y}_p$ be the mean of main variable of interest based upon the last phase sample. Some other notations are described below:

$$\left.\begin{array}{l} \theta_j = 1/n_j - 1/N\ ; \bar{x}_j = \bar{X} + \bar{e}_{x_j},\ j=1,2,...,p; \\ \bar{y}_q = \bar{Y} + \bar{e}_{y_q}; E\left(\bar{e}_{y_q}^2\right) = \theta_j S_y^2\ ;\ E\left(\bar{e}_{x_j}^2\right) = \theta_j S_x^2; \\ E\left(\bar{e}_{y_q}\bar{e}_{x_j}\right) = \theta_j S_{xy}; \\ E\left(\bar{e}_{x_h} - \bar{e}_{x_j}\right)^2 = \left(\theta_j - \theta_h\right) S_x^2\ ;\ h<j \end{array}\right\} (1.1)$$

where θ_j is sampling fraction at *j*th phase, $\bar{x}_j$ is sample mean of auxiliary variable at *j*th phase, S_x^2 is population variance of auxiliary variable, S_y^2 is variance of study variable and S_{xy} is covariance between *X* and *Y*. We follow the description given in [1] and [2] about the quantities $\bar{e}_{y_q}$ and $\bar{e}_{x_j}$ and assume that these quantities are very small. Several estimators have been proposed for use with two phase sampling only. Following estimator was proposed by [2] in two phase sampling using two auxiliary variables:

$$t_1 = \left[\bar{y}_2 + b_{yx}\left(\bar{x}_1 - \bar{x}_2\right)\right]\frac{\bar{z}_1}{\bar{z}_2} \qquad (1.2)$$

where b_{yx} is regression coefficient between *X* and *Y*. A regression–in–ratio estimator was also proposed by [2] as under:

$$t_2 = \left[\bar{y}_2 + b_{yz}\left(\bar{z}_1 - \bar{z}_2\right)\right]\frac{\bar{X}}{\bar{x}_2} \qquad (1.3)$$

where b_{yz} is regression coefficient between *Y* and *Z*. A chain–ratio type estimators proposed by [3] is are given as:

$$t_3 = y_2\frac{\bar{z}_1}{\bar{z}_2}\cdot\frac{\bar{X}}{\bar{x}_1} \quad and \quad t_4 = y_2\frac{\bar{z}_2}{\bar{z}_1}\cdot\frac{\bar{x}_1}{\bar{X}} \qquad (1.4)$$

Three estimators are proposed by [4, 5] for use in two phase sampling. These estimators are ratio–in–regression and regression–in–ratio type estimators. The first estimator proposed by [4] is a regression–in–ratio type estimator and is given as:

$$t_5 = \frac{\bar{y}_2}{\bar{z}_2}\left[\bar{z}_1 + b_{zx}\left(\bar{X} - \bar{x}_1\right)\right] \qquad (1.5)$$

Another estimator proposed by [5] is a ratio–in–regression and is given as:

$$t_6 = \bar{y}_2 + b_{yz}\left(\frac{\bar{z}_1}{\bar{x}_1}\bar{X} - \bar{z}_2\right) \qquad (1.6)$$

A third estimator proposed by [5] is a regression–in–regression type estimator and is given as:

$$t_7 = \bar{y}_2 + b_{yz}\left\{\left(\bar{z}_1 - \bar{z}_2\right) - b_{zx}\left(\bar{x}_1 - \bar{X}\right)\right\} \qquad (1.7)$$

Several other estimators have been proposed in literature. Some other notable references are of [6], [7, 8], [9] and many others.

2. MAIN RESULTS

In this section we have given some main results for estimation in multiphase sampling. These results are given for three and four phase sampling only. Different estimators are given for three phase sampling.

a. THREE PHASE SAMPLING

In this section the result for three phase sampling has been given. We start with the very simple formation and define the following estimator:

$$\bar{y}'_{3(1)} = \bar{y}_3 + \beta_1\left(\bar{x}_1 - \bar{x}_2\right) + \beta_2\left(\bar{x}_2 - \bar{x}_3\right) \qquad (2.1)$$

where β_1 and β_2 are unknown quantities which minimize the variance of (2.1). Using the notations given in (1.1) the above estimator can be written as:

$$\bar{y}'_{3(1)} = \bar{Y} + \bar{e}_{y_3} + \beta_1\left(\bar{e}_{x_1} - \bar{e}_{x_2}\right) + \beta_2\left(\bar{e}_{x_2} - \bar{e}_{x_3}\right)$$

$$\bar{y}'_{3(1)} - \bar{Y} = \bar{e}_{y_3} + \beta_1\left(\bar{e}_{x_1} - \bar{e}_{x_2}\right) + \beta_2\left(\bar{e}_{x_2} - \bar{e}_{x_3}\right)$$

Squaring and applying the expectation, we get; on use of (1.1); the Mean Square Error of (2.1) as:

$$S = MSE\left(\bar{y}'_{3(1)}\right) = E\left(\bar{y}'_{3(1)} - \bar{Y}\right)^2$$

$$= \theta_3 S_y^2 + \left(\theta_2 - \theta_1\right)\beta_1^2 S_x^2 + \left(\theta_3 - \theta_2\right)\beta_1^2 S_x^2 - 2\left(\theta_2 - \theta_1\right)\beta_1 S_{xy} - 2\left(\theta_3 - \theta_2\right)\beta_2 S_{xy} \quad (2.2)$$

To obtain the optimum values of β_1 and β_2 that minimizes "*S*", we partially differentiate (2.2) with respect to β_1 and β_2 to get:

$$\frac{\partial S}{\partial \beta_1} = 2\left(\theta_2 - \theta_1\right)\beta_1 S_x^2 - 2\left(\theta_2 - \theta_1\right)S_{xy} \quad (2.3)$$

and

$$\frac{\partial S}{\partial \beta_2} = 2\left(\theta_3 - \theta_2\right)\beta_2 S_x^2 - 2\left(\theta_3 - \theta_2\right)S_{xy} \quad (2.4)$$

Equating (2.3) and (2.4) to zero and solving we have:

$$\beta_1 = \beta_2 = S_{xy}/S_x^2 = \rho_{xy} S_y / S_x \quad (2.5)$$

Using (2.5) in (2.1) and (2.2) the estimator and its mean square error become:

$$\bar{y}'_{3(1)} = \bar{y}_3 + \beta\left(\bar{x}_1 - \bar{x}_3\right) \quad (2.6)$$

with

$$MSE\left(\bar{y}'_{3(1)}\right) = \theta_3 S_y^2\left(1 - \rho^2\right) + \theta_1 \rho^2 S_y^2 \quad (2.7)$$

Now let's consider another formation of the estimator in three phase sampling:

$$\bar{y}'_{3(2)} = \bar{y}_3 + \beta_1\left(\bar{x}_1 - \bar{x}_3\right) + \beta_2\left(\bar{x}_2 - \bar{x}_3\right) \quad (2.8)$$

Using notations given in (1.1) in (2.8) we have:

$$\bar{y}'_{3(2)} = \bar{Y} + \bar{e}_{y_3} + \beta_1\left(\bar{e}_{x_1} - \bar{e}_{x_3}\right) + \beta_2\left(\bar{e}_{x_2} - \bar{e}_{x_3}\right)$$

$$\bar{y}'_{3(2)} - \bar{Y} = \bar{e}_{y_3} + \beta_1\left(\bar{e}_{x_1} - \bar{e}_{x_3}\right) + \beta_2\left(\bar{e}_{x_2} - \bar{e}_{x_3}\right)$$

Squaring and applying expectations, the mean square error of (2.8) is given as:

$$S = MSE\left(\bar{y}'_{3(2)}\right) = E\left(\bar{y}'_{3(2)} - \bar{Y}\right)^2$$

$$= \theta_3 S_y^2 + \left(\theta_3 - \theta_1\right)\beta_1^2 S_x^2 + \left(\theta_3 - \theta_2\right)\beta_2^2 S_x^2 - 2\left(\theta_3 - \theta_1\right)\beta_1 S_{xy} - 2\left(\theta_3 - \theta_2\right)\beta_2 S_{xy} + 2\left(\theta_3 - \theta_2\right)\beta_1\beta_2 S_x^2 \quad (2.9)$$

Differentiating (2.9) with respect to β_1 and β_2 and equating the resulting derivatives to zero we have:

$$2\left(\theta_3 - \theta_1\right)\beta_1 S_x^2 - 2\left(\theta_3 - \theta_1\right)S_{xy} + 2\left(\theta_3 - \theta_2\right)\beta_2 S_x^2 = 0 \quad (2.10)$$

and $2\left(\theta_3 - \theta_2\right)\beta_2 S_x^2 - 2\left(\theta_3 - \theta_2\right)S_{xy} + 2\left(\theta_3 - \theta_2\right)\beta_1 S_x^2 = 0$ (2.11)

From (2.11) we have:

$$\beta_1 + \beta_2 = S_{xy}/S_x^2 \quad or \quad \beta_2 = S_{xy}/S_x^2 - \beta_1 \quad (2.12)$$

Using (2.12) in (2.10) we have:

$$2\left(\theta_3 - \theta_1\right)\beta_1 S_x^2 - 2\left(\theta_3 - \theta_1\right)S_{xy} + 2\left(\theta_3 - \theta_2\right)\left(\frac{S_{xy}}{S_x^2} - \beta_1\right)S_x^2 = 0$$

On simplification we have $\beta_1 = S_{xy}/S_x^2$ and hence from (2.12) we have $\beta_2 = 0$. Using these values of β_1 and β_2 in (2.8) and (2.9) the estimator and the mean square error turned out to be:

$$\bar{y}'_{3(2)} = \bar{y}_3 + \beta\left(\bar{x}_1 - \bar{x}_3\right) \quad (2.13)$$

with

$$MSE\left(\bar{y}'_{3(2)}\right) = \theta_3 S_y^2\left(1 - \rho^2\right) + \theta_1 \rho^2 S_y^2 \quad (2.14)$$

The estimator given in (2.13) and (2.14) are same as given in (2.6) and (2.7).

We now give another formation of the estimator following the idea of Kiregyera (1984) as under:

$$\bar{y}'_{3(3)} = \bar{y}_3 + \beta_1\left[\left(\bar{x}_1 - \bar{x}_3\right) + \beta_2\left(\bar{x}_2 - \bar{x}_3\right)\right] \quad (2.15)$$

Using the notations given in (1.1) the estimator (2.15) can be written as:

$$\bar{y}'_{3(3)} = \bar{Y} + \bar{e}_{y_3} + \beta_1\left(\bar{e}_{x_1} - \bar{e}_{x_3}\right) + \beta_1\beta_2\left(\bar{e}_{x_2} - \bar{e}_{x_3}\right)$$

$$\bar{y}'_{3(3)} - \bar{Y} = \bar{e}_{y_3} + \beta_1\left(\bar{e}_{x_1} - \bar{e}_{x_3}\right) + \beta_1\beta_2\left(\bar{e}_{x_2} - \bar{e}_{x_3}\right)$$

Squaring and applying expectations, the mean square error of (2.8) is given as:

$$S = MSE\left(\bar{y}'_{3(3)}\right) = E\left(\bar{y}'_{3(3)} - \bar{Y}\right)^2 = \theta_3 S_y^2 + \left(\theta_3 - \theta_1\right)\beta_1^2 S_x^2 + \left(\theta_3 - \theta_2\right)\beta_1^2\beta_2^2 S_x^2 - 2\left(\theta_3 - \theta_1\right)\beta_1 S_{xy} - 2\left(\theta_3 - \theta_2\right)\beta_1\beta_2 S_{xy} + 2\left(\theta_3 - \theta_2\right)\beta_1^2\beta_2 S_x^2 \quad (2.16)$$

Differentiating (2.16) with respect to β_1 and β_2 and equating the resulting derivatives to zero we have:

$$2\left(\theta_3 - \theta_1\right)\beta_1 S_x^2 + 2\left(\theta_3 - \theta_2\right)\beta_1\beta_2^2 S_x^2 - 2\left(\theta_3 - \theta_1\right)S_{xy} - 2\left(\theta_3 - \theta_2\right)\beta_2 S_{xy} + 4\left(\theta_3 - \theta_2\right)\beta_1\beta_2 S_x^2 = 0 \quad (2.17)$$

$$2\left(\theta_3 - \theta_2\right)\beta_1^2\beta_2 S_x^2 - 2\left(\theta_3 - \theta_2\right)\beta_1 S_{xy} + 2\left(\theta_3 - \theta_2\right)\beta_1^2 S_x^2 = 0 \quad (2.18)$$

From (2.18) we have $\beta_2 = S_{xy}/\left(\beta_1 S_x^2\right) - 1$. Using this value of β_2 in (2.17) we have:

$$2\left(\theta_3-\theta_1\right)\beta_1 S_x^2+2\left(\theta_3-\theta_2\right)\beta_1\left(\frac{S_{xy}}{\beta_1 S_x^2}-1\right)^2 S_x^2$$

$$-2\left(\theta_3-\theta_1\right)S_{xy}-2\left(\theta_3-\theta_2\right)\left(\frac{S_{xy}}{\beta_1 S_x^2}-1\right)S_{xy}$$

$$+4\left(\theta_3-\theta_2\right)\beta_1\left(\frac{S_{xy}}{\beta_1 S_x^2}-1\right)S_x^2=0$$

After some algebra we have $\beta_1=S_{xy}/S_x^2$ and hence $\beta_2=0$. So using these values of β_1 and β_2 in (2.15) and (2.16) the estimator and its mean square error are same as given in (2.13) and (2.14) and hence we conclude that the simple estimator of population mean in three phase sampling is given as in (2.6) with mean square error as given in (2.7).

b. FOUR PHASE SAMPLING

In this section we have given some result about estimation in four phase sampling. Again we start as before and give a very simple estimator as:

$$\bar{y}'_{4(1)}=\bar{y}_4+\beta_1\left(\bar{x}_1-\bar{x}_2\right)+\beta_2\left(\bar{x}_2-\bar{x}_3\right)+\beta_3\left(\bar{x}_3-\bar{x}_4\right) \quad (2.19)$$

Arguing as before and using (1.1) in (2.19) we have:

$$\bar{y}'_{4(1)}-\bar{Y}=\bar{e}_{y_4}+\beta_1\left(\bar{e}_{x_1}-\bar{e}_{x_2}\right)+\beta_2\left(\bar{e}_{x_2}-\bar{e}_{x_3}\right)+\beta_3\left(\bar{e}_{x_3}-\bar{e}_{x_4}\right)$$

Squaring and applying expectation, the mean square error of (2.19) is given as:

$$\begin{aligned}S=MSE\left(\bar{y}'_{4(1)}\right)&=\theta_3 S_y^2+\left(\theta_2-\theta_1\right)\beta_1^2 S_x^2+\left(\theta_3-\theta_2\right)\beta_2^2 S_x^2\\&+\left(\theta_4-\theta_3\right)\beta_3^2 S_x^2-2\left(\theta_2-\theta_1\right)\beta_1 S_{xy}\\&-2\left(\theta_3-\theta_2\right)\beta_2 S_{xy}-2\left(\theta_4-\theta_3\right)\beta_3 S_{xy}\end{aligned} \quad (2.20)$$

Differentiating (2.20) with respect to β_1,β_2 and β_3 and equating the resulting derivatives to zero we have:

$$2\left(\theta_2-\theta_1\right)\beta_1 S_x^2-2\left(\theta_2-\theta_1\right)S_{xy}=0$$

$$2\left(\theta_3-\theta_2\right)\beta_2 S_x^2-2\left(\theta_3-\theta_2\right)S_{xy}=0$$

$$2\left(\theta_4-\theta_3\right)\beta_3 S_x^2-2\left(\theta_4-\theta_3\right)S_{xy}=0$$

Solving above equations we have:

$$\beta_1=\beta_2=\beta_3=S_{xy}/S_x^2=\rho_{xy}S_y/S_x \quad (2.21)$$

Using (2.21) in (2.19) and (2.20), the estimator of population mean along with its mean square error in four phase sampling is given as:

$$\bar{y}'_{4(1)}=\bar{y}_4+\beta\left(\bar{x}_1-\bar{x}_4\right) \quad (2.22)$$

with

$$MSE\left(\bar{y}'_{4(1)}\right)=\theta_4 S_y^2\left(1-\rho^2\right)+\theta_1\rho^2 S_y^2 \quad (2.23)$$

Now we consider following formation of the estimator:

$$\bar{y}'_{4(2)}=\bar{y}_4+\beta_1\left(\bar{x}_1-\bar{x}_4\right)+\beta_2\left(\bar{x}_2-\bar{x}_4\right)+\beta_3\left(\bar{x}_3-\bar{x}_4\right) \quad (2.24)$$

Using (1.1) in (2.24) we have:

$$\bar{y}'_{4(2)}-\bar{Y}=\bar{e}_{y_4}+\beta_1\left(\bar{e}_{x_1}-\bar{e}_{x_4}\right)+\beta_2\left(\bar{e}_{x_2}-\bar{e}_{x_4}\right)+\beta_3\left(\bar{e}_{x_3}-\bar{e}_{x_4}\right)$$

Squaring and applying expectation we have:

$$\begin{aligned}S=MSE\left(\bar{y}'_{4(2)}\right)=E\left(\bar{y}'_{4(2)}-\bar{Y}\right)^2&=\theta_4 S_y^2+\left(\theta_4-\theta_1\right)\beta_1^2 S_x^2\\&+\left(\theta_4-\theta_2\right)\beta_2^2 S_x^2+\left(\theta_4-\theta_3\right)\beta_3^2 S_x^2\\&-2\left(\theta_4-\theta_1\right)\beta_1 S_{xy}-2\left(\theta_4-\theta_2\right)\beta_2 S_{xy}\\&-2\left(\theta_4-\theta_3\right)\beta_3 S_{xy}+2\left(\theta_4-\theta_2\right)\beta_1\beta_2 S_x^2\\&+2\left(\theta_4-\theta_3\right)\beta_1\beta_3 S_x^2+2\left(\theta_4-\theta_3\right)\beta_2\beta_3 S_x^2\end{aligned} \quad (2.25)$$

Differentiating (2.25) with respect to β_1,β_2 and β_3 and equating the resulting derivatives to zero we have:

$$2\left(\theta_4-\theta_1\right)\beta_1 S_x^2+2\left(\theta_4-\theta_2\right)\beta_2 S_x^2+2\left(\theta_4-\theta_3\right)\beta_3 S_x^2=2\left(\theta_4-\theta_1\right)S_{xy}$$

$$2\left(\theta_4-\theta_2\right)\beta_1 S_x^2+2\left(\theta_4-\theta_2\right)\beta_2 S_x^2+2\left(\theta_4-\theta_3\right)\beta_3 S_x^2=2\left(\theta_4-\theta_2\right)S_{xy}$$

$$2\left(\theta_4-\theta_3\right)\beta_1 S_x^2+2\left(\theta_4-\theta_3\right)\beta_2 S_x^2+2\left(\theta_4-\theta_3\right)\beta_3 S_x^2=2\left(\theta_4-\theta_3\right)S_{xy}$$

Solving above three equations simultaneously we have $\beta_1=S_{xy}/S_x^2,\beta_2=0$ and $\beta_3=0$. Using these values in (2.19) the four phase sampling estimator (2.24) becomes

$$\bar{y}'_{4(2)}=\bar{y}_4+\beta_1\left(\bar{x}_1-\bar{x}_4\right). \quad (2.26)$$

Further, by using the optimum values of β_1,β_2 and β_3 in (2.25), the mean square error of (2.26) is

$$MSE\left(\bar{y}'_{4(1)}\right)=\theta_4 S_y^2\left(1-\rho^2\right)+\theta_1\rho^2 S_y^2. \quad (2.27)$$

which again depends upon information of first and last phase only.

3. EMPIRICAL STUDY

In this section we have given a brief empirical study by using an artificial data. Some summary measures for the data are given below

$$N=46,\bar{X}=38.4,\bar{Y}=61.5$$

$$S_x^2=79.5,S_y^2=297.1$$

$$S_{xy}=-120.9,\rho_{xy}=-0.79$$

The following table shows the variance of three phase sampling estimator for various sample sizes at various phases.

	Phase 1	
Phase 3	**n_1 = 20**	**n_1 = 30**
n_3 = 4	70.972	69.111
n_3 = 8	33.835	31.974
n_3 = 12	21.456	19.594
n_3 = 16	15.266	13.405
	Phase 1	
Phase 4	**n_1 = 20**	**n_1 = 30**

n_4 = 3	95.731	93.869
n_4 = 5	56.117	54.256
n_4 = 7	39.140	37.279
n_4 = 10	26.407	24.546

We can see that as the phases increase the mean square error starts increasing as should be the case.

4. CONCLUSION

In this short note we have shown that the classical regression estimators in multiphase sampling depends only upon information of first and last phase. The information gathered at other phases do not contribute at all in regression type estimation in multiphase sampling.

REFERENCES:

1. Cochran, W.G., *Sampling techniques*. 1977, New York: John Wiley & Sons.
2. Mohanty, S., *Combination of regression and ratio estimate.* Jour. Ind. Statist. Asso, 1967. **5**: p. 16-19.
3. Chand, L., *Some ratio-type estimators based on two or more auxiliary variables*. 1975: Unpublished Dissertation, Iowa State University.
4. Kiregyera, B., *A chain ratio-type estimator in finite population double sampling using two auxiliary variables.* Metrika, 1980. **27**(1): p. 217-223.
5. Kiregyera, B., *Regression-type estimators using two auxiliary variables and the model of double sampling from finite populations.* Metrika, 1984. **31**(1): p. 215-226.
6. Sahoo, J. and L.N. Sahoo, *A class of estimators in two-phase sampling using two auxiliary variables.* Jour. Ind. Stat. Assoc, 1993. **31**: p. 107-114.
7. Srivastava, S.K., *A two-phase sampling estimator in sample surveys.* Australian & New Zealand Journal of Statistics, 1970. **12**(1): p. 23-27.
8. Srivastava, S.K. and H.S. Jhajj, *A class of estimators of the population mean in survey sampling using auxiliary information.* Biometrika, 1981. **68**(1): p. 341-343.
9. Roy, D.C., *A regression type estimator in two phase sampling using two auxiliary variables.* Pak J. Statist, 2003. **19**(3): p. 281-290.