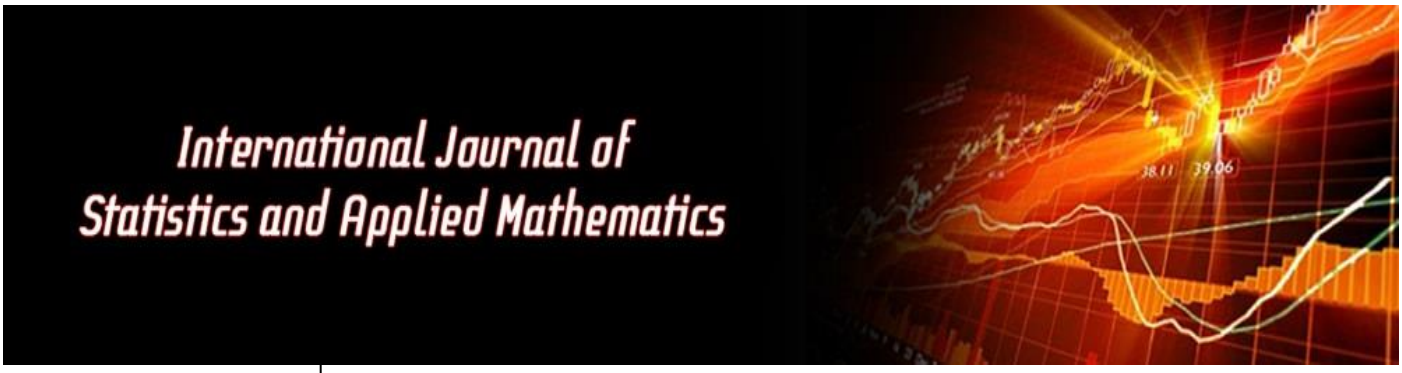
International Journal of Statistics and Applied Mathematics 2019; 4(2): 22-25



ISSN: 2456-1452
Maths 2019; 4(2): 22-25

www.mathsjournal.com
Received: 05-01-2019
Accepted: 08-02-2019

AE Anieting
Department of Mathematics and Statistics, University of Uyo, Nigeria

VO Ezugwu
Department of Mathematics and Statistics, University of Uyo, Nigeria

Correspondence
AE Anieting
Department of Mathematics and Statistics, University of Uyo, Nigeria

Two phase stratified sampling with regression method of estimation using two auxiliary variables

AE Anieting and VO Ezugwu

Abstract
Two phase stratified sampling with regression method of estimation using two auxiliary variables has been proposed. An unbiased estimator for the population mean has also been proposed and the variance of the estimator given. Numerical illustration has been presented to show the efficiency of the estimator. The cost function of the procedure has also been given and the optimum variance derived for a fixed cost and sample size

Keywords: Two phase stratified sampling, regression method, two auxiliary variables

1. Introduction
Sometimes a sampler would select a large sample of units to collect information on certain variables and then select a relative smaller sample to collect information on study variable. This is the problem of two-phase sampling. Two-phase sampling is a powerful technique which was firstly introduced by Neyman (1934) for the stratification purpose. In two-phase sampling, regression and ratio estimation techniques are used to estimate the finite population mean. Regression estimator is considered to be more useful than ratio estimator except when regression line does not pass through origin otherwise these two estimators have almost same significance and analyst has to decide intuitively. Auxiliary information can be used to improve the efficiency of the estimator of a particular population parameter. The effectiveness of the estimation procedure with the use of auxiliary information closely depends upon the method in which the estimator has been proposed. Several authors have worked on different aspects of two-phase sampling using auxiliary information. They include Hansen and Hurwitz (1943) [6], Spurr (1952), Freese (1962), Mohanty (1967) [5], Tripathy (1970), Tripathy(1978), Srivastava and Jhajj (1981) [8], Chand (1975) [2], Sen (1978), Kiregyera (1980, 1984) [3, 4], Sahoo and Sahoo (1993) [9], Sahoo *et al*. (1994), Singh (2001), Uphadhyaya and Singh (2001), Singh *et al*. (2006), Singh *et al*. (2007) [10], Samiddin and Hanif (2007) [7], Singh *et al*. (2011). Using multiple auxiliary variables can improve the efficiency of the estimator during estimation procedure i. e. for instance when $\bar{X}$ is not known, then we may use an additional auxiliary variable z with known population mean $\bar{Z}$. In this case it is assumed that the variable z is also correlated with y than x, hence, the need for this regression estimator in two-phase stratified sampling with two auxiliary variables.

2. Material and Method
There are two types of two-phase sampling design. In the first type, the auxiliary variable does not depend on the measurements, but is purely an indicator variable showing the
stratum to which the variable of interest is to be allocated. This type is termed as two-phase sampling for stratification. In the second type, the relationship between the auxiliary variables and the variables of interest is described by means of ratio or regression. In this case, the design is termed as two-phase sampling with ratio estimators and two-phase sampling with regression estimators.
Pradhan (2000) considered a regression estimator under two phase stratified sampling scheme given by

$\overline{\overline{y_{reg}}} = \sum_h W_h [\bar{y}_h'' + \beta_h (\bar{x}_h' - \bar{x}_h'')]$

Where β_h is the regression coefficient between x and y for the h-th stratum.
Khan and Al-Hossain (2016) gave a difference-type estimator for population mean under two phase sampling as

$t_m = \bar{y} + k_1(\bar{x}' \frac{\bar{Z}}{\bar{z}'} - \bar{x}) + k_2(\bar{Z} \frac{\bar{x}'}{\bar{x}} - \bar{z})$

Where k_1 and k_2 are unknown constant

3. Sampling Design
Let a large sample s_1 of fixed size n' be drawn from a population of size N and is classified into different strata with n_h'units falling in the h^{th} stratum s_{1h} (h = 1, 2, L) where $n' = \sum_{h=1}^{L} n_h'$. In the second phase a SRSWOR subsample s_{2h} of size n_h'' is drawn from n_h' independently for each h and the study variable y is observed. We also assume that at the second phase a constant proportion of units $g_h = \frac{n_h''}{n_h'}$ is sampled from the h-th stratum of the initial sample. Note that the auxiliary variables x and z are observed at the first phase.

4. The proposed Estimator
4.1 The Regression Estimator
The proposed unbiased estimator for population mean of two-phase stratified sampling with two auxiliary variables is given as

$\bar{\bar{y}}_{2reg} = \sum_h^L w_h \left[\bar{y}_h'' + \beta_{yx}(\bar{x}_h' - \bar{x}_h'') + \beta_{yz}(\bar{z}_h' - \bar{z}_h'')\right]$ (4.1)

Where β_{yx} is the regression coefficient between y and x for the h^{th} stratum, while β_{yz} is the regression coefficient between y and z for the h^{th} stratum.

4.2 Proof of Unbiasedness
E $(\bar{\bar{y}}_{2reg}) = E_1\, E_2 E_3\, (\bar{\bar{y}}_{2reg}) = E_1\, E_2\{ \sum_h w_h\, [\bar{y}_h'' + \beta_{yx}\, (\bar{x}_h' - \bar{x}_h'') + \beta_{yz}\, (\bar{z}_h' - \bar{z}_h'')]\}$

$= E_1\, \{ \sum_h W_h[E_2(\bar{y}_h'') + \beta_{yx}\, \bar{x}_h' - \beta_{yx}\, E_2(\bar{x}_h'') + \beta_{yz}\, \bar{z}_h' - E_2(\bar{z}_h'')]\}$

$= E_1\, [\sum_h W_h\{\bar{y}_h' + \beta_{yx}\, \bar{x}_h' - \beta_{yx}\, \bar{x}_h' + \beta_{yz}\, \bar{z}_h' - \beta_{yz}\, \bar{z}_h'\}]$

$= E_1\, [\sum_h W_h\, \bar{y}_h'] = \bar{Y}$

4.3 Variance of the Estimator
V $(\bar{\bar{y}}_{2reg}) = E_1 V_2(\, \bar{\bar{y}}_{2reg}) + V_1 E_2(\, \bar{\bar{y}}_{2reg})$

$E_1 V_2(\, \bar{\bar{y}}_{2reg}) = E_1 V_2(\, \sum_h^L w_h \left[\bar{y}_h'' + \beta_{yx}(\bar{x}_h' - \bar{x}_h'') + \beta_{yz}(\bar{z}_h' - \bar{z}_h'')\right])$

$= E_1\{\sum_h^L w_h^2 V_2(\bar{y}_h'' - \beta_{yx}\, \bar{x}_h'' - \beta_{yz}\, \bar{z}_h'')\}$

$= E_1\left\{w_h^2\left[\frac{\left(\frac{1}{g_h}-1\right)(1-\rho_{yx}^2)S_{yh}'^2}{n'W_h} + \frac{\left(\frac{1}{g_h}-1\right)(1-\rho_{yz}^2)S_{yh}'^2}{n'W_h}\right]\right\} = \sum_h^L \frac{w_h S_{yh}^2}{n'}\left(\frac{1}{g_h} - 1\right)\left[\left(1-\rho_{yx}^2\right) + \left(1-\rho_{yz}^2\right)\right]$

$V_1 E_2(\, \bar{\bar{y}}_{2reg}) = V_1 E_2(\, \sum_h^L w_h \left[\bar{y}_h'' + \beta_{yx}(\bar{x}_h' - \bar{x}_h'') + \beta_{yz}(\bar{z}_h' - \bar{z}_h'')\right])$

$= V_1\, [\sum_h W_h\, \bar{y}_h'] = \left(\frac{1}{n} - \frac{1}{N}\right) S_y^2$

V $(\bar{\bar{y}}_{2reg}) = \left(\frac{1}{n} - \frac{1}{N}\right) S_y^2 + \sum_h^L \frac{w_h S_{yh}^2}{n'}\left(\frac{1}{g_h} - 1\right)\left[\left(1-\rho_{yx}^2\right) + \left(1-\rho_{yz}^2\right)\right]$ (4.2)

4.4 Unbiased Estimator of V $(\bar{\bar{y}}_{2reg})$
The unbiased estimator of V $(\bar{\bar{y}}_{2reg})$ is given as

$\hat{V}\left(\bar{\bar{y}}_{2reg}\right) = \frac{1}{Nn'}\left\{\frac{N-1}{n'-1}\sum_h^L \frac{w_h S_{yh}^2}{n'}\left(\frac{1}{g_h} - 1\right)\left[\left(1-\hat{\rho}_{yx}^2\right) + \left(1-\hat{\rho}_{yz}^2\right)\right] + \frac{N-n'}{n'-1}\sum_h^L \frac{1}{g_h}\sum_{j=1}^{n_h'} y_{hj}^2 \; - \; n'\bar{\bar{y}}_{2reg}^2\right\}$ (4.3)

Where $\hat{\rho}_{yx}\; and\; \hat{\rho}_{yz}$ are estimated correlated coefficient of y and x; and y and z respectively.
Proof

Writing (N-1) $S_y^2 = \sum_h^L \sum_j^{N_h} y_{hj}^2 - N\bar{y}^2$

We can see that it has unbiased estimator

$$N\left[\sum_h^L \frac{W_h}{n_h''} \sum_j^{n_h''} y_{hj}^2 - \langle \bar{\bar{y}}_{2reg}^2 - V(\bar{\bar{y}}_{2reg})\rangle\right]$$

Also, Est. $\sum_h^L W_h S_{yh}^2 \left(\frac{1}{g_h} - 1\right)\left[\frac{(1-\rho_{yx}^2)+(1-\rho_{yz}^2)}{n'}\right] = \sum_h^L w_h s_{yh}^2 \left(\frac{1}{g_h} - 1\right)\left[\frac{(1-\hat{\rho}_{yx}^2)+(1-\hat{\rho}_{yz}^2)}{n'}\right]$

5. Optimum Allocation

Consider the cost function

$$C = c'n' + \sum_h^L c_h'' n_h''$$

Where $c'$is the cost per unit in the first phase sample and c_h'' is the cost per unit in the second phase sample. Since n_h'' is a random variable the expected cost is

$$E(C) = C^* \text{ (say)} = c'n' + \sum_h^L c_h''\, n' g_h W_h \qquad (5.1)$$

Because $n_h'' = n_h' g_h$, E $(n_h'') = g_h$ E $(n_h'') = n' g_h W_h$

It is required to find n' and g_h so as to minimize V $(\bar{\bar{y}}_{2reg})$ for a given expected cost. This is the same as to minimize the product

$$C^*\left[V\left(\bar{\bar{y}}_{2reg}\right) + \frac{S_y^2}{N}\right] = [c' + \sum_h^L c_h''\ g_h W_h] * \left[S_y^2 + \sum_h^L \frac{w_h S_{yh}^2}{n'}\left(\frac{1}{g_h} - 1\right)\left[\left(1-\rho_{yx}^2\right) + \left(1-\rho_{yz}^2\right)\right]\right]$$

With respect to g_h and it exist if and only if

$$\frac{c'}{S_y^2 - \sum_h^L w_h S_{yh}^2 \left[(1-\rho_{yx}^2)+(1-\rho_{yz}^2)\right]} = \frac{c_h'' g_h W_h}{\frac{W_h S_{yh}^2\left[(1-\rho_{yx}^2)+(1-\rho_{yz}^2)\right]}{g_h}}$$

Hence the optimum value of g_h is

$$g_h = \frac{\sqrt{c'A}}{\sqrt{c_h'' B}}$$

Where A = $S_{yh}^2\left[\left(1-\rho_{yx}^2\right) + \left(1-\rho_{yz}^2\right)\right]$ and B = $S_y^2 - \sum_h^L w_h S_{yh}^2\ \left[\left(1-\rho_{yx}^2\right) + \left(1-\rho_{yz}^2\right)\right]$

Substituting the optimum value of n' which is obtain from the expected cost in (5.1) and the optimum value of g_h into the variance expression in (4.2) the optimum variance is obtained as

$$V\left(\bar{\bar{y}}_{2reg}\right)_{opt} = \frac{\left[C'\sqrt{B} + \sum_h^L W_h \sqrt{AC_h''}\right]^2}{C^*} - \frac{S_y^2}{N} \qquad (5.2)$$

6. Numerical Illustration

Consider a hypothetical data given below

Strata	W_h	S_{yh}^2	S_{xh}^2	S_{zh}^2	ρ_{yx}	ρ_{yx}	$\bar{Y}_h$	$\bar{X}_h$	$\bar{Z}_h$
1	0.64	256	54.7	101.1	0.807	0.96	25.2	18.6	18.3
2	0.36	214	51.2	90.5	0.932	0.94	23.4	12.1	16.2

Let the expected cost of the experiment be C*=60 and the cost for each unit of the sample at the second phase be C_h''=0.5 while C' = 0.15. Hence for SRS, a sample of size n=100 is permissible.

Now $S_y^2 = \sum_h^L W_h S_{yh}^2 + \sum_h^L W_h (\bar{Y}_h - \bar{Y})^2$ = 241.692

Hence, V ($\bar{y}_{ran}$) = $\left(\frac{1}{n} - \frac{1}{N}\right) S_y^2$ = 1.76, since N= 375

From (5.2) we have

$$V\left(\bar{\bar{y}}_{2reg}\right)_{opt} = \frac{\left[C'\sqrt{B}+\sum_h^L W_h\sqrt{AC_h''}\right]^2}{C^*} - \frac{S_y^2}{N} = 1.368$$

7. References

1. BK Pradhan. Some problems of Estimation in Multi-phase Sampling, Thesis submitted for the degree of Ph.D degree in Statistics of the Utkal University, Bhubaneswar, 2000.
2. Chand L. Some Ratio-type Estimators based on two or more Auxiliary Variables. Unpublished Ph.D. Dissertation, Iowa State University, Iowa, 1975.
3. Kiregyera B. A Chain Ratio-Type Estimator in Finite Population Double Sampling using two Auxiliary Variables. Metrika. 1980; 27:217-223.
4. Kiregyera B. Regression-type estimators using two auxiliary variables and the model of double sampling from finite populations. Metrica, 1984; 31:215-226.
5. Mohanty S. Combination of regression and ratio estimates. Journal of the Indian Statistical Association. 1997; 5:1–14.
6. MH, Hansen, WN Hurwitz. On the Theory of Sampling from Finite Populations," The Annals of Mathematical Statistics, 1967; 14(4):333-362.
7. M Samiuddin, M Hanif. Estimation of Population Mean in Single Phase and Two-Phase Sampling with or h Applications," ean in Sample Surv without Additional Information," Pakistan Journal of Statistics, 1967; 23:99-118.
8. Srivastava SK, Jhajj HS. A class of estimators of population mean in survey sampling using auxiliary information. Biometrica. 1981; 68:341-343.
9. Sahoo J, LN Sahoo, S Mohanty. A regression approach to estimation in two phase sampling using two auxiliary variables. Current Science. 1993; 65(1):73-75.
10. Singh HP, MR Espejo. Double sampling ratio-product estimator of a finite population mean in sample surveys. J. Appl. Statist. 1963; 34:71-85.
11. Singh R, P Chauhan, N Sawan. A family of estimators for estimating population mean using known correlation coefficient in two phase sampling. Statistics in Transition. 2007; 8(1):89-96,