Three Phase Stratified Sampling with Regression Method of Estimation Using Single Auxiliary Variable

A. E. Anieting & Sesan Ologun

Department of Mathematics and Statistics, University of Uyo, Nigeria

Abstract: **In this paper a three phase stratified sampling is proposed to estimate the population mean of character by a three phase stratified regression estimator and some numerical results are presented to illustrate efficiency of the proposed procedure against possible alternative ones. The expected cost optimum sample sizes on first, second and third phases are also calculated.**

Keywords: **Three Phase Sampling; Stratification; unbiasedness**

I. INTRODUCTION

Multiphase sampling has received much attention from survey Statistician in recent times. In two phase sampling an initial large sample is selected from the population in the first phase to collect information on characteristics such as academic status, income etc, and these information may be used for stratification, or for selection or in estimation procedures. In the second phase, a subsample from the first phase or an independent sample from the population is selected to observe the main character under study. Even though two phase sampling can be extended to three or more phases, few theoretical works have been done regarding such extensions. Pradhan(2013) developed a three phase stratified sampling with ratio method of estimation and proposed an unbiased ratio estimator with some numerical results. Others that worked on three phase sampling include: Fattorini et al (2006), lupke et al (2012) and Mandallazz(2014). Motivated by Pradhan's work the aim of this research is to proposed a three phase stratified sampling with regression method of estimation using single auxiliary variable.

II. Notation

Assume that the population U of size N is divided into L strata. Further for the h^{th} stratum (h = 1,2,….,L) denote

N_h: the number of units for the h^{th} stratum

$n_h^{'}$: the number of units falling into h^{th} stratum after stratifying the first phase sample $n^{'}$ with the help of stratifying variable Z.

$n_h^{''}$: the number of units in a subsample from $n_h^{'}$ for each h in the second phase to observe an auxiliary variable x.

$n_h^{'''}$: the number of units in a subsample drawn from $n_h^{''}$ for each h in the third phase to observed the main character y.

$W_h = \frac{N_h}{N}$;$w_h = \frac{n_h^{'}}{n^{'}}$

$g_h = \frac{n_h^{''}}{n_h^{'}}$: a constant proportion of units sampled from the h^{th} stratum at the second phase, $0<g_h \leq 1$

$t_h = \frac{n_h^{'''}}{n_h^{''}}$: a constant proportion of units sampled from the h^{th} stratum at the third phase, $0<t_h \leq 1$

$\bar{Y}_h = \frac{1}{N_h}\sum_{i=1}^{N_h} y_{hi}$: the mean based on N_h units of y.

$\bar{X}_h = \frac{1}{N_h}\sum_{i=1}^{N_h} x_{hi}$: the mean based on N_h units of x.

$\bar{y}_h^{'}$: the sample mean based on first sample of $n_h^{'}$ in the h^{th} stratum of y.

$\bar{y}_h^{''}$:the sample mean based on second sample of $n_h^{''}$ in the h^{th} stratum of y.

$\bar{y}_h^{'''}$:the sample mean based on first sample of $n_h^{'''}$ in the h^{th} stratum of y.

$\bar{x}_h^{'}$:the sample mean based on first sample of $n_h^{'}$ in the h^{th} stratum of x.

$\bar{x}_h^{''}$: the sample mean based on second sample of $n_h^{''}$ in the h^{th} stratum of x.

$\bar{x}_h^{'''}$: the sample mean based on first sample of $n_h^{'''}$ in the h^{th} stratum of x.

$S_{yh}^2 = \frac{1}{N_h-1}\sum_{i=1}^{N_h}(y_{hi} - \bar{Y}_h)^2$, the mean squared error based on N_h units of y.

$S_{xh}^2 = \frac{1}{N_h-1}\sum_{i=1}^{N_h}(x_{hi} - \bar{X}_h)^2$, the mean squared error based on N_h units of x.

squared error based on the first phase sample of z in theh^{th} stratum.

ρ_{yx} : the correlation coefficient between x and y for the h^{th} stratum.

ρ_{yz} : the correlation coefficient between y and z for the h^{th} stratum.

III. MATERIAL AND METHOD

The method used for this paper in terms of sampling design and the optimum allocation is based on Pradhan (2013)

The Sampling design

Let a large sample s_1 of fixed size n' be drawn from a population of size N and is stratified on the basis of the stratifying variable Z. let n_h' denote the number of units in $s_1(n')$ falling into h^{th} stratum (h = 1, 2, … ,L) with $n'= \sum_{h=1}^{L} n_h'$). A subsamples$s_{2h}(n_h'')$ is drawn from $s_{1h}(n_h')$ independently for each h and an auxiliary x is observed whose frequency distribution is unknown. Further, in the third phase a subsamples$s_{3h}(n_h''')$ is drawn from $s_{2h}(n_h'')$ independently for each h and the character of interest y is observed . SRSWOR is used in selection in all the three phases.

1. Proposed Regression Estimator In Three Phase Sampling With Stratification

Define an estimator of population mean $\bar{Y}$ of y under three phase sampling setting as

$$\bar{\bar{\bar{y}}}_{reg} = \sum_h W_h[\bar{y}_h''' + \beta_h (\bar{x}_h'' - \bar{x}_h''')] \quad (4.1)$$

Where β_h is the regression coefficient between y and x for the h-thstratum

Theorem 4.1. $\bar{\bar{\bar{y}}}_{reg}$ given by (4.1) is approximately an unbiased estimator of $\bar{Y}$

Proof : E $(\bar{\bar{\bar{y}}}_{reg}) = E_1E_2E_3 (\bar{\bar{\bar{y}}}_{reg})$ where E_1, E_2 and E_3 denote expectation operation with respect to first, second and third phase samples respectively.

$= E_1E_2E_3(\bar{\bar{\bar{y}}}_{reg}) = E_1E_2E_3[\sum_h W_h \{\bar{y}_h''' + \beta_h(\bar{x}_h'' - xh'''\}$

$=E_1E_2 [\sum_h W_h \{E_3(\bar{y}_h''') + \beta_h\bar{x}_h'' - \beta_h E_3 (\bar{x}_h''')\}]$

$= E_1E_2 [\sum_h W_h\{\bar{y}_h'' + \beta_h\bar{x}_h'' - \beta_h\bar{x}_h''\}]$

$= E_1E_2 [\sum_h W_h \bar{y}_h''] = E_1[\sum_h W_h E_2(\bar{y}_h'')]$

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

Thus $\bar{\bar{\bar{y}}}_{reg}$ is approximately unbiased estimator of $\bar{Y}$

Theorem 4.2. If a first sample is a random sub-sample of fixed size n' , the second sample is a random stratified sample from the first with fixed g_h ($0<g_h\leq 1$) and the third sample is a random sample from the second with fixed t_h($0<t_h\leq 1$) , then

$$V(\bar{\bar{\bar{y}}}_{reg}) = \left(\frac{1}{n'} - \frac{1}{N}\right) S_y^2 + \sum_h^L \left(\frac{1}{g_h} - 1\right) \frac{W_h S_{yh}^2}{n'} + \sum_h^L \frac{1}{g_h}\left(\frac{1}{t_h} - 11 - \rho h2WhSyh2n'\right) \quad (4.2)$$

Proof : $V(\bar{\bar{\bar{y}}}_{reg}) = E_1E_2V_3(\bar{\bar{\bar{y}}}_{reg}) + E_1V_2E_3(\bar{\bar{\bar{y}}}_{reg}) + V_1E_2E_3(\bar{\bar{\bar{y}}}_{reg})$

now $E_1E_2V_3(\bar{\bar{\bar{y}}}_{reg}) = E_1E_2V_3\{ \sum_h W_h[\bar{y}_h''' + \beta_h (\bar{x}_h'' - \bar{x}_h''')]\}$

$= E_1E_2\{ \sum_h W_h^2 V_3[\bar{y}_h''' - \beta_h (\bar{x}_h''')]\}$

$= E_1E_2 \left\{\sum_h^L W_h^2 \left(\frac{1}{n_h'''} - \frac{1}{n_h''}\right) S_{yh}'^2\right\}$

$=E_1E_2 \left\{\sum_h^L W_h^2 \left(\frac{1}{n'W_h g_h}\right)\left(\frac{1}{t_h} - 1\right)(1-\rho_h^2) S_{yh}'^2\right\}$

$= E_1 \left\{\sum_h^L W_h \left(\frac{1}{n'g_h}\right)\left(\frac{1}{t_h} - 1\right)(1-\rho_h^2) S_{yh}'^2\right\}$

$=\sum_h^L W_h \left(\frac{1}{n'g_h}\right)\left(\frac{1}{t_h} - 1\right)(1-\rho_h^2) S_{yh}^2$

$E_1V_2E_3(\bar{\bar{\bar{y}}}_{reg}) = E_1V_2E_3 \{\sum_h W_h[\bar{y}_h''' + \beta_h (\bar{x}_h'' - \bar{x}_h''')]\}$

$= E_1V_2\{ \sum_h W_h[E_3(\bar{y}_h''') + \beta_h\bar{x}_h'' - \beta_h E_3 (\bar{x}_h''')] \}$

$= E_1V_2\{ \sum_h W_h[\bar{y}_h'' + \beta_h\bar{x}_h'' - \beta_h\bar{x}_h''] \}$

$= E_1V_2 \{ \sum_h W_h \bar{y}_h'' \} = E_1 \{\sum_h W_h^2 \left(\frac{1}{n_h''} - \frac{1}{n_h'}\right) S_{yh}'^2\}$

$= E_1 \left\{\sum_h^L \left(\frac{1}{g_h} - 1\right) \frac{W_h S_{yh}'^2}{n'}\right\}$

$= \sum_h^L \left(\frac{1}{g_h} - 1\right) \frac{W_h S_{yh}^2}{n'}$

$V_1E_2E_3(\bar{\bar{\bar{y}}}_{reg}) = V_1E_2E_3 \{\sum_h W_h[\bar{y}_h''' + \beta_h (\bar{x}_h'' - \bar{x}_h''')]\}$

$= V_1E_2\{ \sum_h W_h[E_3(\bar{y}_h''') + \beta_h\bar{x}_h'' - \beta_h E_3 (\bar{x}_h''')] \}$

$= V_1E_2 \{ \sum_h W_h[\bar{y}_h'' + \beta_h\bar{x}_h'' - \beta_h\bar{x}_h''] \}$

$= V_1\{\sum_h W_h[E_2(\bar{y}_h'')]\} = V_1\{\sum_h W_h \bar{y}_h' \} = \left(\frac{1}{n'} - \frac{1}{N}\right) S_y^2$

Hence,

$$V(\bar{\bar{\bar{y}}}_{reg}) \cong \left(\frac{1}{n'} - \frac{1}{N}\right) S_y^2 + \sum_h^L \left(\frac{1}{g_h} - 1\right) \frac{W_h S_{yh}^2}{n'} + \sum_h^L \frac{1}{g_h}\left(\frac{1}{t_h} - 11 - \rho h2WhSyh2n'\right)$$

Theorem 4.3. An unbiased estimator of V $(\bar{\bar{\bar{y}}}_{reg})$ is given by

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

Proof: Using Est. {.} as an estimator operator, we have the estimator of V $(\bar{\bar{\bar{y}}}_{reg})$ given by

$$\text{Est. } V(\bar{\bar{\bar{y}}}_{reg}) = \hat{V}(\bar{\bar{\bar{y}}}_{reg}) = \text{Est. } \left(\frac{1}{n'} - \frac{1}{N}\right) S_y^2 + \text{Est. } \sum_h^L \left(\frac{1}{g_h} - 1WhSyh2n' + \text{Est.} hL1gh1th - 11 - \rho h2WhSyh2n'\right) \quad (4.4)$$

Now $(N-1)S_y^2 = \sum_{h=1}^{L} \sum_{j=1}^{N_h} y_{hj}^2 - N\bar{Y}^2$

$$= \text{Est.} \sum_{h=1}^{L} \sum_{j=1}^{N_h} y_{hj}^2 - N[\bar{\bar{\bar{y}}}_r^2 - \hat{V}(\bar{\bar{\bar{y}}}_r)] \quad (4.5)$$

$E[\sum_{h=1}^{L} \frac{w_h}{n_h'''} \sum_{j=1}^{n_h'''} y_{hj}^2] = \sum_{h=1}^{L} \frac{w_h}{N_h} \sum_{j=1}^{N_h} y_{hj}^2 = \sum_{h=1}^{L} \frac{W_h}{N_h} \sum_{j=1}^{N_h} y_{hj}^2 = \frac{1}{N} \sum_{h=1}^{L} \sum_{j=1}^{N_h} y_{hj}^2$ hence

Est. $(N-1)S_y^2 = N\ [\sum_{h=1}^{L} \frac{w_h}{n_h'''} \sum_{j=1}^{n_h'''} y_{hj}^2 - \{\bar{\bar{\bar{y}}}_{reg}{}^2 - \hat{V}(\bar{\bar{\bar{y}}}_{reg})\}]$ (4.6)

Est. $\sum_{h=1}^{L}(\frac{1}{g_h} - 1)\frac{W_h S_{yh}^2}{n'} = \sum_{h=1}^{L}(\frac{1}{g_h} - 1)\frac{w_h s_{yh}^2}{n'}$ (4.7)

Est. $\sum_{h}^{L} \frac{1}{g_h}(\frac{1}{t_h} - 1)(1-\rho_h^2)\frac{W_h S_{yh}^2}{n'} = \sum_{h}^{L} \frac{1}{g_h}(\frac{1}{t_h} - 1)(1-\hat{\rho}_h^2)\frac{w_h s_{yh}^2}{n'}$ (4.8)

The result (4.3) is found from (4.6), (4.7) and (4.8)

V. OPTIMUM ALLOCATION

Considering the cost function

$$C = c'n' + \sum_h c_h'' n_h'' + \sum_h c_h''' n_h''' \quad (5.1)$$

Where c' is the cost of observing a unit in the first phase, c_h'' is the cost of observing x-variate on a unit in the h-th stratum in the second phase and c_h''' is the cost of observing y-variate on a unit in the third phase.

$$E(C) = C^* = c'n' + n' \sum_h c_h'' g_h W_h + n' \sum_h c_h''' g_h t_h W_h \quad (5.2)$$

It is required to find $n', g_h \text{ and } t_h$ so as to minimize $V(\bar{\bar{\bar{y}}}_{reg})$ for a given expected cost.

$C^*\ [V + S_y^2/N] = [c' + \sum_h c_h'' g_h W_h + \sum_h c_h''' g_h t_h W_h\ ] \left[S_y^2 + h(1gh-1)WhSyh2+hL1gh1th-11-\rho h2WhSyh2\right.$

$=[c' + \sum_h c_h'' g_h W_h + \sum_h c_h''' g_h t_h W_h\ ]\ [S_y^2 - \sum W_h S_{yh}^2 + \sum_h^L \frac{W_h S_{yh}^2 \rho_h^2}{g_h} + \sum_h^L \frac{W_h S_{yh}^2 (1-\rho_h^2)}{g_h t_h}\ ]$

By applying Cauchy-Schwartz inequality, the minimum value of $C^*\ [V + S_y^2/N]$ occurs if and only if

$$\frac{c'}{S_y^2 - \sum W_h S_{yh}^2} = \frac{c_h'' g_h W_h}{\frac{W_h S_{yh}^2 \rho_h^2}{g_h}} = \frac{c_h''' g_h t_h W_h}{\frac{W_h S_{yh}^2 (1-\rho_h^2)}{g_h t_h}}$$

This gives the minimum values of g_h and t_h as

$$g_h = \frac{S_{yh}\rho_h\sqrt{c'}}{\sqrt{c_h''}\sqrt{S_y^2 - \sum W_h S_{yh}^2}}$$

$$t_h = \frac{\sqrt{c_h''(1-\rho_h^2)}}{\rho_h\sqrt{c_h'''}}$$

Substituting the values of n', g_h and t_h in the variance expression (4.2) , the optimum variance is obtained as

$$V(\bar{\bar{\bar{y}}}_{reg})opt = \frac{c'B + 2\sum_h W_h S_{yh}\sqrt{B}\left(\rho_h\sqrt{c'c_h''} + \sqrt{c_h'''c'A}\right) + \sum_h W_h^2 S_{yh}^2\left(2\rho_h\sqrt{c_h''c_h'''A} + \rho_h^2 c_h'' + Ac_h'''\right)}{C^*} - K \quad (5.3)$$

Where $A = 1 - \rho_h^2$, $B = S_y^2 - \sum W_h S_{yh}^2$, $K = S_y^2/N$

The Derivation of the Optimum Variance

$$n' = \frac{C^*}{c' + \sum_h c_h'' g_h W_h + \sum_h c_h''' g_h t_h W_h} \quad (5.4)$$

Putting the values of g_h and t_h in (5.4) and simplifying, (5.4) becomes

$$n' = \frac{C^*}{\sqrt{B}c' + \sum_h W_h S_{yh}\sqrt{c'}(\rho_h\sqrt{c_h''} + \sqrt{c_h'''A})}$$

From (4.2) the variance of the estimator becomes

$$V(\bar{\bar{\bar{y}}}_{reg}) = \frac{S_y^2 + \sum_h W_h S_{yh}^2\left(\frac{1}{g_h} - 1\right) + \sum_h \frac{1}{g_h}\left(\frac{1}{t_h} - 1\right)(1-\rho_h^2) W_h S_{yh}^2}{n'} - K$$

Substituting the values of $n', g_h \text{ and } t_h$ we have

$$= \frac{S_y^2 + \sum_h W_h S_{yh}\left(\frac{\sqrt{c_h''}B - S_{yh}\rho_h\sqrt{c'}}{\rho_h\sqrt{c'}}\right) + \sum_h W_h S_{yh}\frac{\sqrt{B}\left(\rho_h\sqrt{c_h'''} - \sqrt{c_h''A}\right)\sqrt{A}}{\rho_h\sqrt{c'}}}{n'} - K$$

$$= \frac{S_y^2 + \sum_h W_h S_{yh}\left(\frac{\sqrt{c_h''}B - S_{yh}\rho_h\sqrt{c'}}{\rho_h\sqrt{c'}}\right) + \sum_h W_h S_{yh}\frac{\sqrt{B}\left(\rho_h\sqrt{c_h'''} - \sqrt{c_h''A}\right)\sqrt{A}}{\rho_h\sqrt{c'}}}{\frac{C^*}{\sqrt{B}c' + \sum_h W_h S_{yh}\sqrt{c'}(\rho_h\sqrt{c_h''} + \sqrt{c_h'''A})}} - k \quad (5.5)$$

When (5.5) is simplified we have the optimum variance which is

$$V(\bar{\bar{\bar{y}}}_{reg})opt = \frac{c'B + 2\sum_h W_h S_{yh}\sqrt{B}\left(\rho_h\sqrt{c'c_h''} + \sqrt{c_h'''c'A}\right) + \sum_h W_h^2 S_{yh}^2\left(2\rho_h\sqrt{c_h''c_h'''A} + \rho_h^2 c_h'' + Ac_h'''\right)}{C^*} - K$$

VI. NUMERICAL ILLUSTRATION

The data used for this numerical illustration was used by Pradhan (2013)

The following data come from a particular census in a given year, of all farms in Jefferson County, Iowa. In this example, y_{hj} represents area in acres under corn and x_{hj} as total area in acres in the farm. The population is divided into two strata, the first stratum containing farms of area upto 160 acres and the second stratum containing farms of more than 160 acres. Let the expected cost of the experiment be $C^* = 50$. If the cost

per unit of observation is c = 0.5 , then for SRSWOR a sample size n = 100 is permissible

Table One

Strata	Size	N_h	S_{yh}^2	S_{xh}^2	$\bar{Y}_h$	$\bar{X}_h$	R_h	ρ_h
1	0 - 160	1580	312	2055	19.404	82.56	0.2350	0.61694
2	> 160	430	922	7357	51.626	244.85	0.2109	0.32945

Hence,

V ($\bar{y}_{ran}$) = $\left(\frac{1}{n}-\frac{1}{N}\right)S_y^2$ = 6.6714

i.e $S_y^2 = \sum_h^L W_h S_{yh}^2 + \sum_h^L W_h(\bar{Y}_h - \bar{Y})^2$ = 702.105

For $c' = 0.015, c_h'' = 0.5$ the optimum variance of $\bar{y}_{st}$ in two-phase stratified sampling with regression method of estimation is given by

$$V\ (\bar{\bar{y}}_{reg})_{opt} = \frac{\left[\sqrt{c'(S_y^2-\sum_h^L(1-\rho_h^2))}+\sum_h^L W_h S_{yh}\sqrt{(1-\rho_h^2)c_h''}\right]^2}{C^*} - \frac{S_y^2}{N}$$

= 3.82675

Finally setting $c' = 0.003, c_h'' = 0.015, c_h''' = 0.5$ the optimum variance of $\bar{\bar{\bar{y}}}_{reg}$ in three phase stratified sampling with regression method of estimation is given by

$$V\ (\bar{\bar{\bar{y}}}_{reg})\ opt = \frac{c'B+2\sum_h W_h S_{yh}\sqrt{B}\left(\rho_h\sqrt{c'c_h''}+\sqrt{c_h'''\ c'A}\right)+\sum_h W_h^2 S_{yh}^2\left(2\rho_h\sqrt{c_h''\ c_h'''\ A}+\rho_h^2 c_h''+Ac_h'''\right)}{C^*} - K$$

= 1.4119

Where A =$1-\rho_h^2$, B = $S_y^2 - \sum W_h S_{yh}^2$, K =$\frac{S_y^2}{N}$

Hence, the relative precision of the various methods can be summarized as follows:

Table Two

Sampling Method	Method of Estimation	Relative Precision
Simple Random	Mean Per Unit	100
Stratified Random	Two phase Regression	174.336
Stratified Random	Three Phase Regression	472.5122

The optimum value of the sampling fractions for three phase stratified sampling with regression method of estimation are given by $g_{1(opt)}$ = 0.304, $g_{2(opt)}$ = 0.28, $t_{1(opt)}$ = 0.201, $t_{2(opt)}$ = 0.375

From the expected cost n'= 1156, E (n_1'') = 276, E (n_2'') = 69, E (n_1''') = 56, E (n_2''') = 26

VII. CONCLUSION

From table two above, the relative precision is higher in three phase than in two phase, hence the effectiveness of this scheme is authenticated.

REFERENCES

[1] B.K. PRADHAN, A.K.P.C. SWAIN (2000), Two Phase Stratified Sampling with Ratio and Regression Methods of Estimations, Journal of Science and Technology, Vol. XII, Section B, 52 – 57.

[2] B, K. Pradhan (2013) Three phase stratified sampling with ratio method of estimation, STATISTICA, anno.LXXII, n.2.

[3] Fattorini, Marcheselli M., Pisani C. (2006) A three phase sampling strategy for large scale multiresource forest inventories; JABES, 11: 296-316

[4] Mandallaz (2014) A three phase sampling extension of the generalized regression estimator with partial exhaustive information. Canadian journal of forest research,44(4): 383-388

[5] N lupke, J Hansen,JSaborowski (2012) A three-phase sampling procedure for continuous forest inventory with partial re-measurement and updating of terrestrial sample plot. J. Eur J Forest. Vol 131, issue 6,pp 1979-1990