Modified Two-Stage Cluster Sampling Estimators for Finite Population Total

Mudi, Taiye Adam*[1] and Alhaji, Baba Bukar[2]
[1]Department of Mathematics and Statistics, Kaduna Polytechnic, Kaduna, Nigeria[1]
[2]Department of Mathematical Sciences Nigerian Defence Academy, Kaduna, Nigeria
*corresponding author: adamu110202@gmail.com

Abstract

This work proposes modified versions of the conventional estimators for finite population total under two stage cluster sampling scheme by combining the efficiency gain in Probability Proportional to Size and stratification using both Equal and Unequal probability of selection methods. The population is first stratified into N strata and independent samples of equal size n_h is selected from each stratum using probability Proportional to size (PPS) without replacement in the first stage. In the second stage, m_h and m_{hi} units are selected for equal and unequal probability methods respectively using Simple Random Sampling without replacement (SRSWOR). The variances of the suggested estimators are expressed mathematically.The empirical comparison of the variances, standard errors and the coefficient of variations were used in obtaining the most efficient estimator. It was established that, the proposed estimators were better than the conventional ones. The one that uses unequal probability of selection method performs better amongst the proposed ones and therefore recommended.

Keywords: *Two – Stage Sampling, Stratification, cluster, Probability Proportional to Size*

1. Introduction

In a census, each unit (such as person, household or local government area) is enumerated, whereas in a sample survey, only a sample of units is enumerated and information provided by the sample is used to make estimates relating to all units (Nafiu et. al, 2012). In designing a study, it can be advantageous to sample units in more than one-stage. The criteria for selecting a unit at a given stage typically depend on attributes observed in the previous stages (Kuk, 1988). Two-stage sampling is where the researcher divides the population into clusters, samples the clusters, and then resamples the second time to select the sampling unit of interest.

ISSN NUMBER: 1116-249X

In real life situations, in order to enhance sampling efficiency quite a number of authors have applied two-stage cluster sampling. Some of them include, Fears and Gail (2000), Stehman et al. (2009), Phillips et al. (2008), Horney et al. (2010) as well as Galway et al. (2012). The efficiency of the design when applied to real life situations however depends to great extent on the sampling techniques used in both stages of the design. Lee et al. (2016) suggested a composite estimator to estimate the total when the cluster sizes are different and the population units are unknown in stratified two-stage cluster sampling. The composite estimator they suggested is applicable when the population sizes are not exact but with minor differences and a simple random sampling used to select clusters. However the composite estimator suggested does not solve the problem of estimating the total of size of each cluster hence the need for an estimator to solve the problem

In equal probability sampling, all the population units have equal chances of being selected in the sample regardless of the size of each unit. When units of clusters are of different sizes, it is appropriate to use probability proportional to size (PPS) sampling (see for example Ozturk, (2019) & Micheal et al. (2022)). In this sampling plan, the probability of selection of a cluster element is in proportion to its size or measure of size of the element, so that larger clusters have greater chances of being selected than the smaller clusters, provided the sizes of units of clusters in the population are known and also have positive correlation with the variable under study. The choice of PPS scheme in the first-stage of two-stage sampling under variant cluster sizes has also been supported by Innocenti et al. (2019). Such a procedure of sample selection is also known as unequal probability sampling (Okafor, 2002).

ISSN NUMBER: 1116-249X

Stratification is the process of partitioning the entire population into homogenous subgroupings called strata, with reference to study variables under consideration. That is, homogeneity within a stratum is based on the characteristic under study.

Quite often, strata are available in natural forms. For example, in Agricultural Surveys, geographically contiguous units form a stratum under the assumption that nearby units are likely to be homogenous due to similar agro-climatic conditions as well as similar cultivation practices. However, in other situations strata are formed on the basis of related variables. The essence of stratification is to increase the precision of the estimator.

The study is focused on the modification of the conventional estimators for finite population total under Two stage cluster sampling by combining the efficiency gain in PPS and stratification using the inclusion probability of Horvitz-Thompson Estimator according to Okafor (2002), Wolter (2007) and Enang et al (2016)

2. The New Sampling Design

In this paper, a two-stage cluster sampling where sampling is done among the first stage units by probability proportional to size with stratification and then sampling the same number of units from each selected cluster in the second stage as Scenario 1 and using probability proportional to size with stratification in the first stage and then sampling unequal number of units from each selected clusters as Scenario 2.

2.1 Proposed Modification

A population of N fsu's is stratified into L strata each having N_h clusters such that $N = \sum_{h=1}^{L} N_h$. Each cluster has M_{hi} ssu's; $i = 1,2, \dots, N_h$.

The following notations shall be used in this study

ISSN NUMBER: 1116-249X

N = number of clusters in the first stage unit (fsu's)

n = number of clusters selected from N fsu's

n_h = number of units to be sampled in stratum h

M_{hi} = measure of size for fsu_i in stratum h

N_h = number of clusters in stratum h

$M_h = \sum_{i=1}^{N} M_{hi}$ = total number of clusters in the first stage unit (fsu's)

$m_h =$ equal number of units selected from the selected clusters in the first stage unit.

The inclusion probability of Horvitz-Thompson estimator in the j^{th} ssu's within the i^{th} fsu's according to Okafor (2002), Wolter (2007) and Enang, et al; (2016) can be modified as follows;

Scenario 1

Using Probability Proportional to Size in the First Stage and Simple Random Sampling Without Replacement (SRSWOR) with equal size in the second stage (Equal Probability of Selection Method) (EPSEM))

In the Stage 1 sampling, n_h clusters are selected within each stratum using PPS with replacement. Let π_{hi} be the first order inclusion probability of selecting a cluster that is proportional to the cluster sizes, then

$$\pi_{hi} = \frac{n_h M_{hi}}{M_h} \quad (1)$$

In the Stage 2 sampling, m_h number of units are selected from the clusters for a fixed sample using SRSWOR. Let $\pi_{J/hi}$ be the second order inclusion probability of selection then

$$\pi_{J/hi} = \frac{m_h}{M_{hi}} \quad (2)$$

The overall probability of inclusion of unit in cluster i of stratum h is

$$\pi_{hij} = \pi_{hi} * \pi_{J/hi}$$

ISSN NUMBER: 1116-249X

$$= \frac{n_h M_{hi}}{M_h} * \frac{m_h}{M_{hi}}$$

$$= \frac{n_h m_h}{M_h} \quad (3)$$

The design weight for each unit in cluster i of stratum h is

$$w_{hij} = \frac{1}{\pi_{hij}}$$

Let the population parameter investigated be the population total of Y, Y being a characteristic studied the proposed estimator is

$$\hat{Y}_M = \sum_{h=1}^{L} \sum_{i=1}^{n_h} \sum_{j=1}^{m_h} w_{hij} \hat{Y}_{hij} \quad (4)$$

and based on the assumption that the clusters are sampled with replacement, an estimator of the variance of $\hat{Y}_M$ can be derived as below

Let π_{hi} be the probability of selecting the i^{th} FSU from the h^{th} stratum in the sample, the unbiased estimator of the population total for stratum h is given by

$$\hat{Y}_{M(h)} = \frac{1}{n_h} \sum_{i=1}^{n_h} \frac{M_{hi} \bar{y}_{hi.}}{\pi_{hi}} = \frac{1}{n_h} \sum_{i=1}^{n_h} \frac{\hat{Y}_{hi..}}{\pi_{hi}} \quad (5)$$

Which is shown as

$$E\left(\hat{Y}_{M(h)}\right) = E_1\left[\frac{1}{n_h} \sum_{i=1}^{n_h} \frac{M_{hi}}{\pi_{hi}} E_2(\bar{y}_{hi.})\right] = E_1\left[\frac{1}{n_h} \sum_{i=1}^{n_h} \frac{M_{hi} \bar{Y}_{hi.}}{\pi_{hi}}\right] = E_1\left[\frac{1}{n_h} \sum_{i=1}^{n_h} \frac{Y_{hi.}}{\pi_{hi}}\right] = Y$$

The variance of $\hat{Y}_{M(h)}$ is then derived as follows

$$V\left(\hat{Y}_{M(h)}\right) = V_1\left[E_2\left(\hat{Y}_{M(h)}\right)\right] + E_1\left[V_2\left(\hat{Y}_{M(h)}\right)\right] \quad (6)$$

$$V_1\left[E_2\left(\hat{Y}_{tM(h)}\right)\right] = V_1\left(\frac{1}{n_h} \sum_{i=1}^{n_h} \frac{Y_{hi.}}{\pi_{hi}}\right) = \frac{1}{n_h} \sum_{i=1}^{N} \pi_{hi} \left(\frac{Y_{hi}}{\pi_{hi}} - Y\right)^2 \quad (7)$$

$$E_1\left[V_2\left(\hat{Y}_{M(h)}\right)\right] = E_1\left[\frac{1}{n_h^2} \sum_{i=1}^{n_h} \frac{M_{hi}^2}{\pi_{hi}} V_2(\bar{y}_{hi.})\right] = \frac{1}{n_h} \sum_{i=1}^{n_h} \frac{M_{hi}^2}{\pi_{hi}^2} \frac{(1-f_{2hi})}{m_h} s_{whi}^2 \quad (8)$$

Substituting (8) and (7) in (6) the variance of $\hat{Y}_{M(h)}$ becomes

$$V\left(\hat{Y}_{M(h)}\right) = \frac{1}{n_h} \sum_{i=1}^{N} \pi_{hi} \left(\frac{Y_{hi}}{\pi_{hi}} - Y\right)^2 + \frac{1}{n_h} \sum_{i=1}^{n_h} \frac{M_{hi}^2}{\pi_{hi}^2} \frac{(1-f_{2hi})}{m_h} s_{whi}^2 \quad (9)$$

ISSN NUMBER: 1116-249X

The sample estimator of $V\left(\hat{Y}_{M(h)}\right)$ is

$$\hat{V}\left(\hat{Y}_{M(h)}\right) = \frac{1}{n_h(n_h-1)}\sum_{i=1}^{n_h}\left(\frac{\hat{Y}_{hi}}{\pi_{hi}} - \hat{Y}_{M(h)}\right)^2 \quad (10)$$

The variance estimator for the overall population total is obtained by taking the sum of the variances of the stratum level total. That is

$$\hat{V}\left(\hat{Y}_M\right) = \sum_{h=1}^{L}\frac{1}{n_h(n_h-1)}\sum_{i=1}^{n_h}\left(\frac{\hat{Y}_{hi}}{\pi_{hi}} - \hat{Y}_{M(h)}\right)^2 = \sum_{h=1}^{L}\frac{1}{n_h(n_h-1)}\sum_{i=1}^{n_h}(y_{hi} - \bar{y}_h)^2 \quad (11)$$

Where $y_{hi} = \sum_{j=1}^{m_h} n_h(w_{hij})\hat{Y}_{hij}$, and $\bar{y}_h = \frac{1}{n_h}\sum_{i=1}^{n_h}\hat{Y}_{hi}$ and $w_{hij} = \frac{1}{\pi_{hij}}$

. Theorem 1: $\mathrm{E}\left[\hat{V}\left(\hat{Y}_M\right)\right] = V(\left(\hat{Y}_{M(h)}\right)$

Proof: We show that

$\mathrm{E}\left[\hat{V}\left(\hat{Y}_M\right)\right] = V(\left(\hat{Y}_{M(h)}\right)$

⇨ $n_h(n_h-1)E\left[\hat{V}\left(\hat{Y}_{M(h)}\right)\right] = n_h(n_h-1)\,V\left(\hat{Y}_{M(h)}\right)$

$$n_h(n_h-1)E\left[\hat{V}\left(\hat{Y}_{M(h)}\right)\right] = E_1E_2\left(\sum_{i=1}^{n_h}\frac{\hat{Y}_{hi.}}{\pi_{hi}} - \hat{Y}_{M(h)}\right)^2$$

$$= E_1E_2\left[\sum_{i=1}^{n_h}\frac{\hat{Y}_{hi.}^2}{\pi_{hi}^2} - n_h\,\hat{Y}^2{}_{M(h)}\right] \quad (12)$$

Starting with the first term in the square brackets and remembering that $\hat{Y}_{hi} = M_{hi}\bar{y}_{hi}$

$$E_1E_2\left[\sum_{i=1}^{n_h}\frac{\hat{Y}_{hi}^2}{\pi_{hi}^2}\right] = E_1\left[\sum_{i=1}^{n_h}E_2\left(\frac{\hat{Y}_{hi}^2}{\pi_{hi}^2}\right) =\right]E_1\left[\sum_{i=1}^{n_h}\frac{M_{hi}^2}{\pi_{hi}^2}\left\{\bar{Y}_{hi.} + \frac{(1-f_{2hi})}{m_h}s_{whi}^2\right\}\right]$$

$$= n_h\left[\sum_{i=1}^{N}\frac{Y_{hi}^2}{\pi_{hi}} + \sum_{i=1}^{N}\frac{M_{hi}^2}{\pi_{hi}}\frac{(1-f_{2hi})}{m_h}s_{whi}^2\right] \quad (13)$$

$$E_1E_2\left(n_h\hat{Y}^2{}_{M(h)}\right) = n_hE_1\left[V_2\left(\hat{Y}_{M(h)}\right) + \left\{E_2\left(\hat{Y}_{M(h)}\right)\right\}^2\right]$$

$$= n_hE_1\left[\frac{1}{n_h^2}\sum_{i=1}^{n_h}\frac{M_{hi}^2}{\pi_{hi}}\frac{(1-f_{2hi})}{m_h}s_{whi}^2 + \left(\frac{1}{n_h}\sum_{i=1}^{N}\frac{M_{hi}Y_{hi.}}{\pi_{hi}}\right)^2\right]$$

$$= n_h\left[\frac{1}{n_h}\sum_{i=1}^{N}\frac{M_{hi}^2}{\pi_{hi}}\frac{(1-f_{2hi})}{m_h}s_{whi}^2 + \frac{1}{n_h}\sum_{i=1}^{N}\pi_{hi}\left(\frac{Y_{hi}}{\pi_{hi}} - Y\right)^2 + Y^2\right] \quad (14)$$

ISSN NUMBER: 1116-249X

Subtracting (14) from (13) we have

$$n_h(n_h-1)E\left[\hat{V}\left(\hat{Y}_{M(h)}\right)\right] = (n_h - 1)\left[\sum_{i=1}^{N}\frac{1}{n_h}\sum_{i=1}^{n_h}\frac{M_{hi}^2}{\pi_{hi}}\frac{(1-f_{2hi})}{m_h}s_{whi}^2 + \frac{1}{n_h}\sum_{i=1}^{N}\pi_{hi}\left(\frac{Y_{hi}}{\pi_{hi}} - Y\right)^2\right]$$

$$= n_h(n_h-1)\,V\left(\hat{Y}_{M(h)}\right)$$

Scenario 2

Using Probability Proportional to Size in the First Stage and Simple Random Sampling with unequal size in the second stage (Unequal Probability of Selection Method)

In the Stage 1 sampling, n_h clusters are selected within each stratum using PPS with replacement. Let π_{hi} be the first order inclusion probability of selecting a cluster that is proportional to the cluster sizes, then

$$\pi_{hi} = \frac{n_h M_{hi}}{M_h} \quad (15)$$

In the Stage 2 sampling, m_{hi} number of units are selected from the clusters using SRSWOR. Let $\pi_{j/hi}$ be the second order inclusion probability of selection then

$$\pi_{j/hi} = \frac{\mathrm{m_{hi}}}{M_{hi}} \quad (16)$$

The overall probability of inclusion of unit in cluster i of stratum h is

$$\pi_{hij} = \pi_{hi} * \pi_{j/hi} = \frac{n_h M_{hi}}{M_h} * \frac{\mathrm{m_{hi}}}{M_{hi}} = \frac{n_h \mathrm{m_{hi}}}{M_h} \quad (17)$$

The design weight for each unit in cluster i of stratum h is

$$w_{hij} = \frac{1}{\pi_{hij}}$$

Let the population parameter investigated be the population total of Y, Y being a characteristic studied the proposed estimator is

$$\hat{Y}_{M2} = \sum_{h=1}^{L}\sum_{i=1}^{n_h}\sum_{j=1}^{m_h} w_{hij}\hat{Y}_{hij} \quad (18)$$

ISSN NUMBER: 1116-249X

Where $w_{hij} = \frac{1}{\pi_{hij}} = \frac{M_h}{n_h \mathrm{m_{hi}}}$ and $\hat{Y}_{hij}$ is the j^{th} observation in i^{th} cluster, stratum h

Based on the assumption that the clusters are sampled with replacement, an estimator of the variance of $\hat{Y}_{M2}$ can be derived as below

Let π_{hi} be the probability of selecting the i^{th} FSU from the h^{th} stratum in the sample, the unbiased estimator of the population total for stratum h is given by

$$\hat{Y}_{M2(h)} = \frac{1}{n_h}\sum_{i=1}^{n_h} \frac{M_{hi}\bar{y}_{hi.}}{\pi_{hi}} = \frac{1}{n_h}\sum_{i=1}^{n_h} \frac{\hat{Y}_{hi.}}{\pi_{hi}} \quad (19)$$

Which is shown as

$$E\left(\hat{Y}_{M2(h)}\right) = E_1\left[\frac{1}{n_h}\sum_{i=1}^{n_h} \frac{M_{hi}}{\pi_{hi}} E_2(\bar{y}_{hi.})\right] = E_1\left[\frac{1}{n_h}\sum_{i=1}^{n_h} \frac{M_{hi}\bar{Y}_{hi}}{\pi_{hi}}\right] = E_1\left[\frac{1}{n_h}\sum_{i=1}^{n_h} \frac{Y_{hi.}}{\pi_{hi}}\right] = Y$$

The variance of $\hat{Y}_{M(h)}$ is then derived as follows

$$V\left(\hat{Y}_{M2(h)}\right) = V_1\left[E_2\left(\hat{Y}_{M2(h)}\right)\right] + E_1\left[V_2\left(\hat{Y}_{M2(h)}\right)\right] \quad (20)$$

$$V_1\left[E_2\left(\hat{Y}_{M2(h)}\right)\right] = V_1\left(\frac{1}{n_h}\sum_{i=1}^{n_h} \frac{Y_{hi.}}{\pi_{hi}}\right) = \frac{1}{n_h}\sum_{i=1}^{N} \pi_{hi}\left(\frac{Y_{hi}}{\pi_{hi}} - Y\right)^2 \quad (21)$$

$$E_1\left[V_2\left(\hat{Y}_{M2(h)}\right)\right] = E_1\left[\frac{1}{n_h^2}\sum_{i=1}^{n_h} \frac{M_{hi}^2}{\pi_{hi}} V_2(\bar{y}_h)\right] = \frac{1}{n_h}\sum_{i=1}^{n_h} \frac{M_{hi}^2}{\pi_{hi}^2}\frac{(1-f_{2hi})}{m_{hi}} s_{whi}^2 \quad (22)$$

Substituting (2) and (3) in (1) the variance of $\hat{Y}_{str-pps(h)}$ becomes

$$V\left(\hat{Y}_{M(2h)}\right) = \frac{1}{n_h}\sum_{i=1}^{N} \pi_{hi}\left(\frac{Y_{hi}}{\pi_{hi}} - Y\right)^2 + \frac{1}{n_h}\sum_{i=1}^{n_h} \frac{M_{hi}^2}{\pi_{hi}^2}\frac{(1-f_{2hi})}{m_{hi}} s_{whi}^2 \quad (23)$$

The sample estimator of $V\left(\hat{Y}_{M2(h)}\right)$ is

$$\hat{V}\left(\hat{Y}_{M2(h)}\right) = \frac{1}{n_h(n_h-1)}\sum_{i=1}^{n_h}\left(\frac{\hat{Y}_{hi}}{\pi_{hi}} - \hat{Y}_{M2(h)}\right)^2 \quad (24)$$

The variance estimator for the overall population total is obtained by taking the sum of the variances of the stratum level total. That is

$$\hat{V}\left(\hat{Y}_{M2}\right) = \sum_{h=1}^{L}\frac{1}{n_h(n_h-1)}\sum_{i=1}^{n_h}\left(\frac{\hat{Y}_{hi}}{\pi_{hi}} - \hat{Y}_{M2(h)}\right)^2 = \sum_{h=1}^{L}\frac{1}{n_h(n_h-1)}\sum_{i=1}^{n_h^*}(y_{hi} - \bar{y}_h)^2 \quad (25)$$

Where $y_{hi} = \sum_{j=1}^{m_{hi}} n_h(w_{hij})\hat{Y}_{hij}$, and $\bar{y}_h = \frac{1}{n_h}\sum_{i=1}^{n_h}\hat{Y}_{hi}$ and $w_{hij} = \frac{1}{\pi_{hij}}$

ISSN NUMBER: 1116-249X

Theorem 2: $\mathrm{E}\left[\hat{V}\left(\hat{Y}_{M2}\right)\right] = V\left(\left(\hat{Y}_{M2(h)}\right)\right.$

Proof: We show that

$\mathrm{E}\left[\hat{V}\left(\hat{Y}_{M2(h)}\right)\right] = V\left(\left(\hat{Y}_{M2(h)}\right)\right.$

⇨ $n_h(n_h-1)E\left[\hat{V}\left(\hat{Y}_{M(2h)}\right)\right] = n_h(n_h-1)\,V\left(\hat{Y}_{M2(h)}\right)$

$$n_h(n_h-1)E\left[\hat{V}\left(\hat{Y}_{M2(h)}\right)\right] = E_1E_2\left(\sum_{i=1}^{n_h}\frac{\hat{Y}_{hi}}{\pi_{hi}} - \hat{Y}_{M2(h)}\right)^2$$

$$= E_1E_2\left[\sum_{i=1}^{n_h}\frac{\hat{Y}_{hi}^2}{\pi_{hi}^2} - n_h\,\hat{Y}^2{}_{M2(h)}\right] \qquad (26)$$

Starting with the first term in the square brackets and remembering that $\hat{Y}_{hi} = M_{hi}\bar{y}_{hi}$

$$E_1E_2\left[\sum_{i=1}^{n_h}\frac{\hat{Y}_{hi}^2}{\pi_{hi}^2}\right] = E_1\left[\sum_{i=1}^{n_h}E_2\left(\frac{\hat{Y}_{hi}^2}{\pi_{hi}^2}\right) =\right]E_1\left[\sum_{i=1}^{n_h}\frac{M_{hi}^2}{\pi_{hi}^2}\left\{\bar{Y}_{hi.} + \frac{(1-f_{2hi})}{m_{hi}}s_{whi}^2\right\}\right]$$

$$= n_h\left[\sum_{i=1}^{N}\frac{Y_{hi}^2}{\pi_{hi}} + \sum_{i=1}^{N}\frac{M_{hi}^2}{\pi_{hi}}\frac{(1-f_{2hi})}{m_{hi}}s_{whi}^2\right] \qquad (27)$$

$$E_1E_2\left(n_h\hat{Y}^2{}_{M2(h)}\right) = n_hE_1\left[V_2\left(\hat{Y}_{M2(h)}\right) + \left\{E_2\left(\hat{Y}_{M2(h)}\right)\right\}^2\right]$$

$$= n_hE_1\left[\frac{1}{n_h^2}\sum_{i=1}^{n_h}\frac{M_{hi}^2}{\pi_{hi}}\frac{(1-f_{2hi})}{m_{hi}}s_{whi}^2 + \left(\frac{1}{n_h}\sum_{i=1}^{n_h}\frac{M_{hi}Y_{hi.}}{\pi_{hi}}\right)^2\right]$$

$$== n_h\left[\frac{1}{n_h}\sum_{i=1}^{N}\frac{M_{hi}^2}{\pi_{hi}}\frac{(1-f_{2hi})}{m_h}s_{whi}^2 + \frac{1}{n_h}\sum_{i=1}^{N}\pi_{hi}\left(\frac{Y_{hi}}{\pi_{hi}} - Y\right)^2 + Y^2\right] \qquad (28)$$

Subtracting (28) from (27) we have

$$n_h(n_h-1)E\left[\hat{V}\left(\hat{Y}_{M2(h)}\right)\right] = (n_h - 1)\left[\sum_{i=1}^{N}\frac{1}{n_h}\sum_{i=1}^{n_h}\frac{M_{hi}^2}{\pi_{hi}}\frac{(1-f_{2hi})}{m_{hi}}s_{whi}^2 + \frac{1}{n_h}\sum_{i=1}^{N}\pi_{hi}\left(\frac{Y_{hi}}{\pi_{hi}} - Y\right)^2\right]$$

$$= n_h(n_h-1)\,V\left(\hat{Y}_{M2(h)}\right)$$

2.2. Efficiency Comparison of the Proposed Modified Two Stage Estimators

For effective estimation of total and variance, stratified two stage sampling with Simple Random Sampling is widely used to reduce time and cost. However, the accuracy will not be

ISSN NUMBER: 1116-249X

good if the cluster varies in sizes and a small number of clusters are selected and a large units in each selected clusters are sampled. Hence the need to incorporate Probability Proportional to Size (PPS) with known cluster sizes and the total number of population elements. If the ratio of the stratum population size to the stratum sample size is similar or identical to the ratio of the cluster population size to the cluster sample size that is

$$\frac{N_h}{n_n} = \frac{M_h}{\sum_{i=1}^{n_h} M_{hi}} = \frac{\sum_{i=1}^{N_h} M_{hi}}{\sum_{i=1}^{n_h} M_{hi}}$$

Then the estimated values using $\hat{Y}_M$ and $\hat{Y}_{M2}$ can be used

To check the efficiency of the proposed estimators, let us compare it with the conventional estimators. Let the proposed estimators be $\hat{Y}_M$ and $\hat{Y}_{M2}$ for scenarios 1 and 2 respectively and the conventional estimators be $\hat{Y}_C$,

The relative efficiency of $\hat{Y}_M$ over $\hat{Y}_C$ is

$$RE = \frac{V\left(\hat{Y}_M\right)}{V\left(\hat{Y}_C\right)} ,$$

The relative efficiency of $\hat{Y}_{M2}$ over $\hat{Y}_C$ is

$$RE = \frac{V\left(\hat{Y}_{M2}\right)}{V\left(\hat{Y}_C\right)}$$

If $RE < 1$, then our proposed estimators $\left(\hat{Y}_M\right)$ and $\left(\hat{Y}_{M2}\right)$ are more efficient than the conventional ones $\left(\hat{Y}_C\right)$. Also

The relative efficiency of $\hat{Y}_M$ over $\hat{Y}_{M2}$ is

ISSN NUMBER: 1116-249X

$$RE = \frac{V\left(\hat{Y}_M\right)}{V\left(\hat{Y}_{M2}\right)},$$

If $RE < 1$, then our proposed estimators $\left(\hat{Y}_M\right)$ is more efficient than $\left(\hat{Y}_{M2}\right)$.

3.1 Data Analysis

The proposed estimators were applied to Nigerian Census data of 2006 extracted from the Federal Republic of Nigeria Official Gazette online. The following symbolic Notations were used. The research covers the population of people in the 774 local governments areas across the 36 states and the federal capital territory Abuja. The country is first stratified into 6 geo political zones. In the first stage, a random sample of size n = 18 states is selected using probability proportional to size without replacement with equal number of states in each.

In the second stage, out of the selected states, one third of the local governments is selected and averaged giving m_h= 7 local government per state. This is to ensure equal probability of selection as proposed in Scenario 1. For Scenario 2, one third of size m_{hi} is selected at the second stage. The following notations were used for the analysis

$N =$ Total Number of States in Nigeria
$n =$ Number of States Selected
n_h = Number of states to be sampled in the geo- political zone h
M_{hi} = Number of Local Governments in the geo-political zone h of state i
N_h = Number of states in geo-political zone h
$M = \sum_{i=1}^{N} M_h =$ Total Number of Local Governments in Nigeria.
$M_h = \sum_{i=1}^{N_h} M_{hi}$= Total number of Local Governments in the zone h
$m_h =$ equal number of Local Governments selected from the selected States in the first stage unit of zone h

ISSN NUMBER: 1116-249X

m_{hi} = unequal number of local governments selected from the selected states in the first stage unit of zone h

$y_{hij} =$ the j^{th} population in i^{th} state of zone h

The estimator of the total population and their variances are as follows:

L = number of strata

The analysis was carried out using R Statistical package with the results below.

Table 1: Comparison of Estimators in terms of Variance (V), Standard Error (SE), and Coefficient of Variation for Scenario 1

Estimator	Modified Two Stage	Stratified Two Stage	Two Stage Proportional	Single Stage Estimator	Two Stage Random Estimator
$\hat{Y}$	$\mathbf{149,401,109}$	150,139,116	145,678,791	159,154,534	150,560,098
$V(\hat{Y})$	$\mathbf{2.646808 \times 10^{13}}$	1.509056×10^{14}	1.218261×10^{14}	1.080245×10^{14}	4.999876×10^{13}
$SE(\hat{Y})$	$\mathbf{5,144,714}$	12,284,364.05	11,037,486.13	10,393,483.54	7,070,980.13
$CV(\hat{Y})$	**0.03440**	0.0818	0.0757	0.0653	0.0469

Table 2: Comparison of Estimators in terms of Variance (V), Standard Error (SE), and Coefficient of Variation (CV) for Scenario 2

Estimator	Modified Two Stage	Stratified Two Stage	Two Stage Proportional	Ordinary Estimator	Two Stage Random Estimator
$\hat{Y}$	$\mathbf{152,269,942}$	152,304,960	149,015,814	159,154,534	152,492,500
$V(\hat{Y})$	$\mathbf{2.414846 \times 10^{13}}$	1.497742×10^{14}	1.265733×10^{14}	1.080245×10^{14}	4.727231×10^{13}
$SE(\hat{Y})$	$\mathbf{4,914,109}$	12,238,227	11,250,479.99	10,393,483.54	6,875,486.17
$CV(\hat{Y})$	$\mathbf{0.0323}$	0.0804	0.0755	0.0653	0.0451

ISSN NUMBER: 1116-249X

Table 3: Comparison of the Estimators of the two Scenarios in terms of Variance (V), Standard Error (SE), and Coefficient of Variation (CV

Estimator	Modified Two Stage (Scenario 1)	Modified Two Stage (Scenario 2)
$\hat{Y}$	149,401,109	152,269,942
$V(\hat{Y})$	2.646808×10^{13}	2.414846×10^{13}
$SE(\hat{Y})$	5,144,714	4,914,109
$CV(\hat{Y})$	0.03440	0.0323

4. Discussion of Results

Table 1: In this table, the newly proposed estimator (Modified Two Stage for Scenario 1) has the least for the variance, standard error and coefficient of variation of the population. This shows that the newly proposed estimator performs better.

Table 2: This table also shows the estimated population totals, their variances, standard errors, and coefficient of variations for the census data. The newly proposed estimator (Modified Two Stage for Scenario 2) has the least for the variance, standard error and coefficient of variation. This implies that newly proposed estimator performs better.

Table 3: Here the two newly proposed estimators are compared. The newly proposed estimator in Scenario 2 has the least variance, standard error and coefficient of variation.

Conclusion and Recommendation: When an unbiased estimator of high precision and an unbiased sample estimate of its variance is required for a two-stage sampling design with equal number of samples selected from each clusters, the modified estimators using equal and

ISSN NUMBER: 1116-249X

unequal probability of selection with stratification and probability proportional to size are preferred over the conventional ones. The one that uses unequal probability of selection (Scenario 2) is the better one and therefore recommended.

References

Enang, E.I & Onyishi, I,L (2016): Population Total and Variance Estimation in Multi-Stage Cluster Sampling: A Simple Random Sampling Approach, *International Journal of Applied Science and Mathematical Theory*, 2 (2) 18-30 www.iiardpub.org

Fears, T. R., & Gail, M. H., (2000).: Analysis of a two-stage case-control study with cluster sampling of controls Application to Nonmelanoma skin cancer, *Biometrics.* 56(1), 190–198.

Galway, L. P., Bell, N., Shatari, S. AE. Al., Hagopian, A., Burnham, G., Flaxman, A., Weiss, W. M., Rajaratnam, J. and Takaro, T. K., (2012).. A two-stage cluster sampling method using ridded population data, a GIS and Google Earth TM imagery in population-based mortality survey in Iraq. *International Journal of Health Geographics*, 11(12), pp. 1–9.

Horney, J. J., Dickinson, M., Hsai, J., Williams, A. and Zotti, M., (2010). Two-stage cluster sampling with referral: Improving the efficiency of estimating unmet needs among pregnant and postpartum women after flooding in Northwest Georgia. *Remote Sensing of Environment*, 113(6), pp. 1236–1249.

Innocenti, F., Candel, M. J. J. M., Tan, F. E. S. and van Breukelen, G. J. P., (2019). Relative fficiencies of two-stage sampling schemes for mean estimation in multilevel populations hen cluster size is informative. *Statistics in Medicine*, 38(10), pp. 1817–1834.

Kuk, A (1988): Estimation of Distribution Functions and Medians under Sampling with Unequal Probabilities. *Biometrika.* 75(1), 97-103 https://doi.org/10.1093/biomet/75.1.97

ISSN NUMBER: 1116-249X

Lee, S. E., . Lee P. R and K. I. Shin.,(2016) A composite estimator for stratified two stage cluster sampling. Communications for Statistical Applications and Methods 23(1), (2016), pp. 47-55 http://dx.doi.org/10.5351/CSAM.2016.23.1.047

Lee, S. E., . Lee P. R and K. I. Shin.,(2016) A composite estimator for stratified two stage cluster sampling. Communications for Statistical Applications and Methods 23(1), (2016), pp. 47-55 http://dx.doi.org/10.5351/CSAM.2016.23.1.047

Michael, C.U & Mbanefo, S.M (2022): Two-stage cluster sampling with unequal probability sampling in the first stage and ranked set sampling in the second stage. *Statistic in transition new series. 23(3) 199–214, DOI 10.2478/stattrans-2022-0038*

Nafiu, L.A., Oshungade, I.O.,& Adewara,A.A (2012). An Alternative Estimation Method for Three Stage Cluster Sampling in Finite Population. *American Journal of Mathematics and Statistics,* 2(6): 199-205 DOI: 10.5923/j.ajms.20120206.06

Okafor, F (2002). Sample Survey Theory with Applications. Afro-Orbis Publications: Lagos, Nigeria.

Ozturk, O., (2019). Two-stage cluster samples with ranked set sampling designs *Annals of the Institute of Statistical Mathematics,* 71,. 63–91.

Phillips, A. E., Boily, M. C., Lowndes, C. M., Garnett, G. P., Gurav, K., Ramesh, B. M., Anthony, J., Watts, R., Moses, S. and Alary, M., (2008). Sexual identity and its contribution to MSM risk behaviour in Bangaluru (Bangalore) India: The results of a two-stage cluster sampling survey. *Journal of LGBT Health Research,* 4, pp. 111–126.

ISSN NUMBER: 1116-249X

Stehman, S. V., Wichham, J. D., Fattorini, L., Wade, T. D., Baffetta, F. and Smith, J. H., (2009). Estimating accuracy of land-cover composition from two-stage cluster sampling. Remote Sensing of Environment, 113(6), pp. 1236–1249.

Wolter, K. M. (2007). Introduction to variance estimation. Second Edition Springer-Verlag.

ISSN NUMBER: 1116-249X