April 2013. Vol. 3, No. 1 ISSN 2305-8269
International Journal of Engineering and Applied Sciences

www.eaas-journal.org

GENERALIZATION OF MULTISTAGE CLUSTER SAMPLING USING FINITE POPULATION

L. A. Nafiu[*1], I. O. Oshungade [2] and A. A. Adewara [2]
[1]Department of Mathematics and Statistics, Federal University of Technology, Minna, Nigeria
[2]Department of Statistics, University of Ilorin, Ilorin, Nigeria
[*]E-mail: firdaousmama@gmail.com

ABSTRACT

This paper generalizes the use of multistage cluster sampling design in estimating the population total where all units within the clusters are considered. The focus is on a special design where certain number of visits is considered for estimating the population size and a weighted factor $\frac{N_i}{n_i^2}$ is introduced. The generalized model is: $\hat{Y}_{stagek} = \frac{1}{\gamma}\sum_{i=1}^{n}\{\frac{1}{\gamma_i}\sum_{j=1}^{m_i}(\frac{1}{\gamma_{ij}}\sum_{l=1}^{k_{ij}}\frac{N_i}{n_i^2}y_{ijl})\}$ with its variance given as $\hat{V}(\hat{Y}_{stagek}) = N(N-n)\frac{s^2}{n} + \frac{N}{n}\sum_{i=1}^{n}M_i(M_i-m_i)\frac{s_i^2}{m_i} + \frac{N}{n}\sum_{i=1}^{n}\frac{M_i}{m_i}\sum_{j=1}^{m_i}K_{ij}(K_{ij}-k_{ij})\frac{s_{ij}^2}{k_{ij}}$ where γ, γ_i and γ_{ij} are primary, secondary and tertiary sampling fractions respectively. Eight (8) sets of data were used to justify our model based on the ranking of coefficients of variation criteria. The use of multistage cluster sampling has shown that inclusion of the effect of stage clustering produced better results.

Keywords: Unequal probability sampling, Two-stage sampling, Hansen-Hurwitz estimator and Horvitz-Thompson estimator

INTRODUCTION

Many estimation procedures have been developed in multistage cluster sampling designs. Some of these procedures are very famous for example, Cochran (1977); Kalton (1983); Henry (1990); Thompson (1992); Fink (2002); Okafor (2002); and Tate and Hudgens (2007). Of recent, is the work of Nafiu (2012) on comparison of estimates arising from one-, two- and three- stage; and that of Nafiu *et al.* (2012) on alternative estimation procedure for a three-stage cluster sampling design.

Variability in multistage sampling includes the following:
(i) In one-stage cluster sampling, the estimate varies due to one source: different samples of primary units yield different estimates.
(ii)In two-stage cluster sampling, the estimate varies due to two sources: different samples of primary units and then different samples of secondary units within primary units.
(iii) In three-stage cluster sampling, the estimate varies due to three sources: different samples of primary units, then different samples of secondary units within primary units and then different samples of tertiary units within secondary units.
(iv) In general, if there are k stages of sub sampling, there will be k sources of variability. Thus, variances and variance estimators for multistage cluster sampling with $k-$stage will contain the sum of k components of variability.

AIM AND OBJECTIVES OF THIS STUDY

The aim of this research is to generalize the estimation procedure for multistage sampling scheme. The main objectives are to:
(i) investigate some of the existing estimators used in multistage cluster sampling designs.
(ii) develop new estimator that is more efficient than already existing estimators and generalize it.
(iii) apply this newly generalized estimator to a real life situation. That is, the estimation of population total of diabetic patients in Niger

state for four (4) different years: 2005 – 2008 (four data sets).

MATERIALS AND METHODS

In this section, we derived a generalized form of multistage cluster sampling design given by Nafiu (2012) procedure. The generalization is described as:

1. Select first unit n (i.e. the number of primary units in the sample)
2. Select second unit m_i (i.e. the number of secondary units in the primary unit)
3. Select third unit k_{ij} (i.e. the number of tertiary units in the secondary units of the primary unit)

Let y_{iju} be the value obtained for the uth third-stage units in the jth second-stage units drawn from the ith primary units. The relevant population total for over-all sample in a three-stage is given as follows:

$$Y = \sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{u=1}^{K} y_{iju} \qquad (1)$$

For any estimation $\hat{\Theta}_h$ in the hth cell based on completely arbitrary probabilities of selection, the total variance is then the sum of the variances for all strata. The symbol E is used for the operator of expectation, V for the variance, and $\hat{V}$ for the unbiased estimate of V. We may then write

$$V(\hat{\Theta}_h) = \underset{1}{V}(\underset{>1}{E}(\hat{\Theta}_h)) + \underset{1}{E}(\underset{>1}{V}(\hat{\Theta}_h)) \qquad (2)$$

where ">1" is the symbol to represent all stages of sampling after the first.

The expression (2) may be written into three components as:

$$V(\hat{\Theta}_h) = \underset{1}{V}(\underset{2}{E}(\underset{>2}{E}(\hat{\Theta}_h))) + \underset{1}{E}(\underset{2}{V}(\underset{>2}{E}(\hat{\Theta}_h))) + \underset{1}{E}(\underset{2}{E}(\underset{>2}{V}(\hat{\Theta}_h))) \qquad (3)$$

For instance, the state consists of N number of local government areas out of which a simple random sampling of n number of local government areas is selected. Each local government area consists of M_i number of cities out of which a simple random sampling without replacement of m_i number of cities is selected. Finally, from the selected sample of city containing K_{ij} number of hospitals, k_{ij} number of hospitals is selected at random without replacement and the number of diabetic patients in this hospital is collected.

Then;

$$y = \sum_{i=1}^{n}\sum_{j=1}^{m_i} y_{ij} \qquad (4)$$

An unbiased estimator of the population total at jth secondary unit in the ith primary unit in the sample is:

$$\begin{aligned}\hat{y}_{ij} &= \frac{1}{\gamma_{ij}}\sum_{l=1}^{k_{ij}}\frac{N_i}{n_i^2}y_{ijl} \\ &= \frac{K_{ij}}{k_{ij}}\sum_{l=1}^{k_{ij}}\frac{N_i}{n_i^2}y_{ijl}\end{aligned} \qquad (5)$$

where $\gamma_{ij} = \frac{k_{ij}}{K_{ij}}$ is the known sampling fraction for tertiary units in the jth secondary unit of the ith primary unit.

An unbiased estimator of the population total in the ith primary unit in the sample is:

$$\hat{y}_i = \frac{M_i}{m_i}\sum_{j=1}^{m_i}\hat{y}_{ij} \quad (6)$$

Finally, an unbiased estimator of the population total as given by Nafiu *et al.* (2012) for the diabetic patients undergoing treatment in all the hospitals at the jth secondary unit (city) in the ith primary unit (local government area) is:

$$\hat{Y}_{stagek} = \frac{N}{n}\sum_{i=1}^{n}\{\frac{M_i}{m_i}\sum_{j=1}^{m_i}(\frac{K_{ij}}{k_{ij}}\sum_{l=1}^{k_{ij}}\frac{N_i}{n_i^2}y_{ijl})\} \quad (7)$$

An unbiased estimator of the variance of $\hat{Y}_{stagek}$ given in equation (7) is obtained by replacing the population variances with the sample variances as follows:

$$\hat{V}\left(\hat{Y}_{stagek}\right) = N(N-n)\frac{s_1^2}{n} + \frac{N}{n}\sum_{i=1}^{n}M_i(M_i-m_i)\frac{s_i^2}{m_i} + \frac{N}{n}\sum_{i=1}^{n}\frac{M_i}{m_i}\sum_{j=1}^{m_i}K_{ij}\left(K_{ij}-k_{ij}\right)\frac{s_{ij}^2}{k_{ij}} \quad (8)$$

where
$$s_1^2 = \frac{\sum_{i=1}^{n}(y_i - \frac{\hat{Y}_{stagek}}{n})^2}{n-1}$$
$$s_i^2 = \frac{\sum_{j=1}^{m_i}(y_{ij} - \frac{y_i}{m_i})^2}{m_i - 1}$$
$$s_{ij}^2 = \frac{K_{ij}^2}{k_{ij}^2}\sum_{l=1}^{k_{ij}}\left((\frac{N_i}{n_i^2})^2 - \frac{N_i}{n_i^2}\right)y_{ijl}$$

and $\hat{Y}_{stagek}$ represents one-stage, two-stage or three-stage as the case may be.

EMPIRICAL STUDY

In this section, empirical study was carried out in order to decide about the performance of the generalization of the multistage cluster selection procedure. To carry out the empirical study, the sampling standard errors and the corresponding coefficients of variation of one-stage, two-stage and three-stage cluster designs were obtained for all the cases and the populations.

There are eight (8) categories of data used in this paper. The first four (4) data sets were obtained and used as illustration while the second four (4) data sets used are of secondary type and were collected from Niger State Ministry of Health, Minna, Niger state, Nigeria. We constructed a sampling frame from all diabetic patients with chronic eye disease (Glaucoma and Retinopathy) in the twenty-five (25) Local Government Areas of the state between years 2005 and 2008 as contained in Nafiu *(*2012). The standard errors obtained for the estimated population totals are as shown in Table 2.

ESTIMATED POPULATION TOTALS AND STANDARD ERRORS

Tables 1 and 2 give estimated population totals and their corresponding standard errors using equations (7) and (8) respectively

RANKING OF COEFFICIENTS OF VARIATION FOR THE ESTIMATED POPULATION TOTALS

This is given by the ratio of standard error for the estimated population total to the estimated population total itself and expressed in percentage. That is;

Coefficient of Variation (CV) = se $(\hat{Y})/\hat{Y}$ X 100%

The ranking for coefficient of variations, in ascending order, using one-stage, two-stage and three-stage sampling schemes are given in Table 3 below.

Table 1: Estimated Population Totals Using Multi-Stage Cluster Sampling Schemes

Estimator	Case I	Case II	Case III	Case IV	Population 1	Population 2	Population 3	Population 4
$\hat{Y}_{1NPE}$	421	98966	39	13855	24639	25010	26551	28407
$\hat{Y}_{2NPE}$	417	116761	30	15207	25841	26675	27204	29300
$\hat{Y}_{3NPE}$	492	99136	42	15016	26151	26625	27511	28090

Table 2: Standard Errors for the Estimated Population Totals

Estimator	Case I	Case II	Case III	Case IV	Population 1	Population 2	Population 3	Population 4
$\hat{Y}_{1NPE}$	55.2505	5513.7267	1.1706	1219.5061	106.0996	109.5827	110.4655	114.4640
$\hat{Y}_{2NPE}$	30.2589	1293.0835	1.1105	221.5343	100.6545	103.9577	99.4164	108.5860
$\hat{Y}_{3NPE}$	16.5785	302.2597	1.0538	40.2437	95.4893	98.6229	104.7944	103.0137

Table 3: Ranking of Coefficients of Variation

Estimator	Case I	Case II	Case III	Case IV	Population 1	Population 2	Population 3	Population 4
$\hat{Y}_{1NPE}$	3	3	2	3	3	3	3	3
$\hat{Y}_{2NPE}$	2	2	3	2	2	1	2	1
$\hat{Y}_{3NPE}$	1	1	1	1	1	1	1	1

DISCUSSION OF RESULTS

Table 3 gives the ranking of coefficients of variations, in ascending order, for the estimated population totals and it shows that the higher-stage estimator has the least ranking of coefficient of variation. That is, in general, the coefficients of variation shows that higher-stage sampling scheme performs better than lower-stage sampling scheme.

CONCLUSION AND RECOMMENDATIONS

When an unbiased estimator of high precision and an unbiased sample estimate of its variance is required, the multistage sampling system employing cluster scheme at each stage is particularly appropriate. Higher order multistage cluster sampling design gives the best results. Therefore, it is recommended that higher-stage cluster sampling designs be employed when considering multistage sampling.

REFERENCES

1. Cochran, W.G. 1977. *Sampling Techniques*. Third Edition. John Wiley and Sons: New York.
2. Fink, A. 2002. *How to Sample In Surveys.* Sage Publications: Thousand Oaks, C.A.
3. Henry, G. T. 1990. *Practical Sampling.* Sage Publications: Thousand Oaks, C.A.
4. Kalton, G. 1983. *Introduction to Survey Sampling.* Sage Publications: Thousand Oaks, C.A.
5. Nafiu, L. A. 2012. An Alternate Estimation Method for Multistage Cluster Sampling in Finite Population . Unpublished Ph.D Thesis, University of Ilorin, Ilorin, Nigeria.
6. Nafiu, L. A., Oshungade, I. O. and Adewara, A. A., 2012. "Alternative Estimation Method for a Three-Stage Cluster Sampling in Finite Population". American Journal of Mathematics and Statistics (AJMS). 2 (6): 12 – 17.
7. Okafor, F. 2002. *Sample Survey Theory with Applications.* Afro-Orbis Publications: Nigeria.
8. Tate, J.E. and Hudgens, M.G.2007. "Estimating Population Size with Two-stage and Three-stage sampling designs". American Journal of Epidemiology. 165(11): 1314-1320.
9. Thompson, S. K. 2002. *Sampling*. John Wiley and Sons: New York.